Eliciting Categorical Data for Optimal Aggregation
Chien-Ju Ho, Rafael Frongillo, Yiling Chen
Models for collecting and aggregating categorical data on crowdsourcing platforms typically fall into two broad categories: those assuming agents honest and consistent but with heterogeneous error rates, and those assuming agents strategic and seek to maximize their expected reward. The former often leads to tractable aggregation of elicited data, while the latter usually focuses on optimal elicitation and does not consider aggregation. In this paper, we develop a Bayesian model, wherein agents have differing quality of information, but also respond to incentives. Our model generalizes both categories and enables the joint exploration of optimal elicitation and aggregation. This model enables our exploration, both analytically and experimentally, of optimal aggregation of categorical data and optimal multiple-choice interface design.
****
A Locally Adaptive Normal Distribution
Georgios Arvanitidis, Lars K. Hansen, Søren Hauberg
The multivariate normal density is a monotonic function of the distance to the mean, and its ellipsoidal shape is due to the underlying Euclidean metric. We suggest to replace this metric with a locally adaptive, smoothly changing (Riemannian) metric that favors regions of high local density. The resulting locally adaptive normal distribution (LAND) is a generalization of the normal distribution to the "manifold" setting, where data is assumed to lie near a potentially low-dimensional manifold embedded in R^D. The LAND is parametric, depending only on a mean and a covariance, and is the maximum entropy distribution under the given metric. The underlying metric is, however, non-parametric. We develop a maximum likelihood algorithm to infer the distribution parameters that relies on a combination of gradient descent and Monte Carlo integration. We further extend the LAND to mixture models, and provide the corresponding EM algorithm. We demonstrate the efficiency of the LAND to fit non-trivial probability distributions over both synthetic data, and EEG measurements of human sleep.
****
Tagger: Deep Unsupervised Perceptual Grouping
Klaus Greff, Antti Rasmus, Mathias Berglund, Tele Hao, Harri Valpola, Jürgen Schmidhuber
We present a framework for efficient perceptual inference that explicitly reasons about the segmentation of its inputs and features. Rather than being trained for any specific segmentation, our framework learns the grouping process in an unsupervised manner or alongside any supervised task. We enable a neural network to group the representations of different objects in an iterative manner through a differentiable mechanism. We achieve very fast convergence by allowing the system to amortize the joint iterative inference of the groupings and their representations. In contrast to many other recently proposed methods for addressing multi-object scenes, our system does not assume the inputs to be images and can therefore directly handle other modalities. We evaluate our method on multi-digit classification of very cluttered images that require texture segmentation. Remarkably our method achieves improved classification performance over convolutional networks despite being fully connected, by making use of the grouping mechanism. Furthermore, we observe that our system greatly improves upon the semi-supervised result of a baseline Ladder network on our dataset. These results are evidence that grouping is a powerful tool that can help to improve sample efficiency.
****
Online Bayesian Moment Matching for Topic Modeling with Unknown Number of Topics
Wei-Shou Hsu, Pascal Poupart
Latent Dirichlet Allocation (LDA) is a very popular model for topic modeling as well as many other problems with latent groups. It is both simple and effective. When the number of topics (or latent groups) is unknown, the Hierarchical Dirichlet Process (HDP) provides an elegant non-parametric extension; however, it is a complex model and it is difficult to incorporate prior knowledge since the distribution over topics is implicit. We propose two new models that extend LDA

in a simple and intuitive fashion by directly expressing a distribution over the number of topics. We also propose a new online Bayesian moment matching technique to learn the parameters and the number of topics of those models based on streaming data. The approach achieves higher log-likelihood than batch and online HDP with fixed hyperparameters on several corpora.

****

Conditional Generative Moment-Matching Networks

Yong Ren, Jun Zhu, Jialian Li, Yucen Luo

Maximum mean discrepancy (MMD) has been successfully applied to learn deep generative models for characterizing a joint distribution of variables via kernel mean embedding. In this paper, we present conditional generative moment-matching networks (CGMMN), which learn a conditional distribution given some input variables based on a conditional maximum mean discrepancy (CMMD) criterion. The learning is performed by stochastic gradient descent with the gradient calculated by back-propagation. We evaluate CGMMN on a wide range of tasks, including predictive modeling, contextual generation, and Bayesian dark knowledge, which distills knowledge from a Bayesian model by learning a relatively small CGMMN student network. Our results demonstrate competitive performance in all the tasks.

****

Collaborative Recurrent Autoencoder: Recommend while Learning to Fill in the Blanks

Hao Wang, Xingjian SHI, Dit-Yan Yeung

Hybrid methods that utilize both content and rating information are commonly used in many recommender systems. However, most of them use either handcrafted features or the bag-of-words representation as a surrogate for the content information but they are neither effective nor natural enough. To address this problem, we develop a collaborative recurrent autoencoder (CRAE) which is a denoising recurrent autoencoder (DRAE) that models the generation of content sequences in the collaborative filtering (CF) setting. The model generalizes recent advances in recurrent deep learning from i.i.d. input to non-i.i.d. (CF-based) input and provides a new denoising scheme along with a novel learnable pooling scheme for the recurrent autoencoder. To do this, we first develop a hierarchical Bayesian model for the DRAE and then generalize it to the CF setting. The synergy between denoising and CF enables CRAE to make accurate recommendations while learning to fill in the blanks in sequences. Experiments on real-world datasets from different domains (CiteULike and Netflix) show that, by jointly modeling the order-aware generation of sequences for the content information and performing CF for the ratings, CRAE is able to significantly outperform the state of the art on both the recommendation task based on ratings and the sequence generation task based on content information.

****

Bayesian Intermittent Demand Forecasting for Large Inventories

Matthias W. Seeger, David Salinas, Valentin Flunkert

We present a scalable and robust Bayesian method for demand forecasting in the context of a large e-commerce platform, paying special attention to intermittent and bursty target statistics. Inference is approximated by the Newton-Raphson algorithm, reduced to linear-time Kalman smoothing, which allows us to operate on several orders of magnitude larger problems than previous related work. In a study on large real-world sales datasets, our method outperforms competing approaches on fast and medium moving items.

****

Visual Dynamics: Probabilistic Future Frame Synthesis via Cross Convolutional Networks

Tianfan Xue, Jiajun Wu, Katherine Bouman, Bill Freeman

We study the problem of synthesizing a number of likely future frames from a single input image. In contrast to traditional methods, which have tackled this problem in a deterministic or non-parametric way, we propose a novel approach which models future frames in a probabilistic manner. Our proposed method is therefore able to synthesize multiple possible next frames using the same model. Solving this challenging problem involves low- and high-level image and motion understa nding for successful image synthesis. Here, we propose a novel network structure, namely a Cross Convolutional Network, that encodes images as feature maps and motion information as convolutional kernels to aid in synthesizing future frames. In experiments, our model performs well on both synthetic data, such as 2D shapes and animated game sprites, as well as on real-wold video data. We show that our model can also be applied to tasks such as visual analogy-making, and present analysis of the learned network representations.

****

Achieving budget-optimality with adaptive schemes in crowdsourcing

Ashish Khetan, Sewoong Oh

Adaptive schemes, where tasks are assigned based on the data collected thus far, are widely used in practical crowdsourcing systems to efficiently allocate the budget. However, existing theoretical analyses of crowdsourcing systems suggest that the gain of adaptive task assignments is minimal. To bridge this gap, we investigate this question under a strictly more general probabilistic model, which has been recently introduced to model practical crowdsourcing data sets. Under this generalized Dawid-Skene model, we characterize the fundamental trade-off between budget and accuracy, and introduce a novel adaptive scheme that matches this fundamental limit. We further quantify the gain of adaptivity, by comparing the trade-off with the one for non-adaptive schemes, and confirm that the gain is significant and can be made arbitrarily large depending on the distribution of the difficulty level of the tasks at hand.

****

Stochastic Gradient Richardson-Romberg Markov Chain Monte Carlo

Alain Durmus, Umut Simsekli, Eric Moulines, Roland Badeau, Gaël RICHARD

Stochastic Gradient Markov Chain Monte Carlo (SG-MCMC) algorithms have become increasingly popular for Bayesian inference in large-scale applications. Even though these methods have proved useful in several scenarios, their performance is often limited by their bias. In this study, we propose a novel sampling algorithm that aims to reduce the bias of SG-MCMC while keeping the variance at a reasonable level. Our approach is based on a numerical sequence acceleration method, namely the Richardson-Romberg extrapolation, which simply boils down to running almost the same SG-MCMC algorithm twice in parallel with different step sizes. We illustrate our framework on the popular Stochastic Gradient Langevin Dynamics (SGLD) algorithm and propose a novel SG-MCMC algorithm referred to as Stochastic Gradient Richardson-Romberg Langevin Dynamics (SGRRLD). We provide formal theoretical analysis and show that SGRRLD is asymptotically consistent, satisfies a central limit theorem, and its non-asymptotic bias and the mean squared-error can be bounded. Our results show that SGRRLD attains higher rates of convergence than SGLD in both finite-time and asymptotically, and it achieves the theoretical accuracy of the methods that are based on higher-order integrators. We support our findings using both synthetic and real data experiments.

****

Generating Videos with Scene Dynamics

Carl Vondrick, Hamed Pirsiavash, Antonio Torralba

We capitalize on large amounts of unlabeled video in order to learn a model of scene dynamics for both video recognition tasks (e.g. action classification) and video generation tasks (e.g. future prediction). We propose a generative adversarial network for video with a spatio-temporal convolutional architecture that untangles the scene's foreground from the background. Experiments suggest this model can generate tiny videos up to a second at full frame rate better than simple baselines, and we show its utility at predicting plausible futures of static images. Moreover, experiments and visualizations show the model internally learns useful features for recognizing actions with minimal supervision, suggesting scene dynamics are a promising signal for representation learning. We believe generative video models can impact many applications in video understanding and simulation.

****

Approximate maximum entropy principles via Goemans-Williamson with applications to provable variational methods

Andrej Risteski, Yuanzhi Li

The well known maximum-entropy principle due to Jaynes, which states that given mean parameters, the maximum entropy distribution matching them is in an exponential family has been very popular in machine learning due to its “Occam’s razor” interpretation. Unfortunately, calculating the potentials in the maximum entropy distribution is intractable [BGS14]. We provide computationally efficient versions of this principle when the mean parameters are pairwise moments: we design distributions that approximately match given pairwise moments, while having entropy which is comparable to the maximum entropy distribution matching those moments. We additionally provide surprising applications of the approximate maximum entropy principle to designing provable variational methods for partition function calculations for Ising models without any assumptions on the potentials of the model. More precisely, we show that we can get approximation guarantees for the log-partition function comparable to those in the low-temperature limit, which is the setting of optimization of quadratic forms over the hypercube. ([AN06])

****

Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering

Michaël Defferrard, Xavier Bresson, Pierre Vandergheynst

In this work, we are interested in generalizing convolutional neural networks (CNNs) from low-dimensional regular grids, where image, video and speech are represented, to high-dimensional irregular domains, such as social networks, brain connectomes or words’ embedding, represented by graphs. We present a formulation of CNNs in the context of spectral graph theory, which provides the necessary mathematical background and efficient numerical schemes to design fast localized convolutional filters on graphs. Importantly, the proposed technique offers the same linear computational complexity and constant learning complexity as classical CNNs, while being universal to any graph structure. Experiments on MNIST and 20NEWS demonstrate the ability of this novel deep learning system to learn local, stationary, and compositional features on graphs.

****

Fast Distributed Submodular Cover: Public-Private Data Summarization

Baharan Mirzasoleiman, Morteza Zadimoghaddam, Amin Karbasi

In this paper, we introduce the public-private framework of data summarization motivated by privacy concerns in personalized recommender systems and online social services. Such systems have usually access to massive data generated by a large pool of users. A major fraction of the data is public and is visible to (and can be used for) all users. However, each user can also contribute some private data that should not be shared with other users to ensure her privacy. The goal is to provide a succinct summary of massive dataset, ideally as small as possible, from which customized summaries can be built for each user, i.e. it can contain elements from the public data (for diversity) and users' private data (for personalization). To formalize the above challenge, we assume that the scoring function according to which a user evaluates the utility of her summary satisfies submodularity, a widely used notion in data summarization applications. Thus, we model the data summarization targeted to each user as an instance of a submodular cover problem. However, when the data is massive it is infeasible to use the centralized greedy algorithm to find a customized summary even for a single user. Moreover, for a large pool of users, it is too time consuming to find such summaries separately. Instead, we develop a fast distributed algorithm for submodular cover, FASTCOVER, that provides a succinct summary in one shot and for all users. We show that the solution provided by FASTCOVER is competitive with that of the centralized algorithm with the number of rounds that is exponentially smaller than state of the art results. Moreover, we have implemented FASTCOVER with Spark to demonstrate its practical performance on a number of concrete applications, including personalized location recommendation, personalized movie recommendation, and dominating set on tens of millions of data points and varying number of users.

****

Exponential Family Embeddings

Maja Rudolph, Francisco Ruiz, Stephan Mandt, David Blei

Word embeddings are a powerful approach to capturing semantic similarity among terms in a vocabulary. In this paper, we develop exponential family embeddings, which extends the idea of word embeddings to other types of high-dimensional data. As examples, we studied several types of data: neural data with real-valued observations, count data from a market basket analysis, and ratings data from a movie recommendation system. The main idea is that each observation is modeled conditioned on a set of latent embeddings and other observations, called the context, where the way the context is defined depends on the problem. In language the context is the surrounding words; in neuroscience the context is close-by neurons; in market basket data the context is other items in the shopping cart. Each instance of an embedding defines the context, the exponential family of conditional distributions, and how the embedding vectors are shared across data. We infer the embeddings with stochastic gradient descent, with an algorithm that connects closely to generalized linear models. On all three of our applications—neural activity of zebrafish, users’ shopping behavior, and movie ratings—we found that exponential family embedding models are more effective than other dimension reduction methods. They better reconstruct held-out data and find interesting qualitative structure.

****

A Non-parametric Learning Method for Confidently Estimating Patient's Clinical State and Dynamics

William Hoiles, Mihaela van der Schaar

Estimating patient's clinical state from multiple concurrent physiological streams plays an important role in determining if a therapeutic intervention is necessary and for triaging patients in the hospital. In this paper we construct a non-parametric learning algorithm to estimate the clinical state of a patient. The algorithm addresses several known challenges with clinical state estimation such as eliminating bias introduced by therapeutic intervention censoring, increasing the timeliness of state estimation while ensuring a sufficient accuracy, and the ability to detect anomalous clinical states. These benefits are obtained by combining the tools of non-parametric Bayesian inference, permutation testing, and generalizations of the empirical Bernstein inequality. The algorithm is validated using real-world data from a cancer ward in a large academic hospital.

****

Integrated perception with recurrent multi-task neural networks

Hakan Bilen, Andrea Vedaldi

Modern discriminative predictors have been shown to match natural intelligences in specific perceptual tasks in image classification, object and part detection, boundary extraction, etc. However, a major advantage that natural intelligences still have is that they work well for all perceptual problems together, solving them efficiently and coherently in an integrated manner. In order to capture some of these advantages in machine perception, we ask two questions: whether deep neural networks can learn universal image representations, useful not only for a single task but for all of them, and how the solutions to the different tasks can be integrated in this framework. We answer by proposing a new architecture, which we call multinet, in which not only deep image features are shared between tasks, but where tasks can interact in a recurrent manner by encoding the results of their analysis in a common shared representation of the data. In this manner, we show that the performance of individual tasks in standard benchmarks can be improved first by sharing features between them and then, more significantly, by integrating their solutions in the common representation.

****

Dialog-based Language Learning

Jason E. Weston

A long-term goal of machine learning research is to build an intelligent dialog agent. Most research in natural language understanding has focused on learning from fixed training sets of labeled data, with supervision either at the word level (tagging, parsing tasks) or sentence level (question answering, machine translation). This kind of supervision is not realistic of how humans learn, where language is both learned by, and used for, communication. In this work, we study d ialog-based language learning, where supervision is given naturally and implicitly in the response of the dialog partner during the conversation. We study this setup in two domains: the bAbI dataset of (Weston et al., 2015) and large-scale question answering from (Dodge et al., 2015). We evaluate a set of baseline learning strategies on these tasks, and show that a novel model incorporating predictive lookahead is a promising approach for learning from a teacher's response. In particular, a surprising result is that it can learn to answer questions correctly without any reward-based supervision at all.

****

A Theoretically Grounded Application of Dropout in Recurrent Neural Networks

Yarin Gal, Zoubin Ghahramani

Recurrent neural networks (RNNs) stand at the forefront of many recent developments in deep learning. Yet a major difficulty with these models is their tendency to overfit, with dropout shown to fail when applied to recurrent layers. Recent results at the intersection of Bayesian modelling and deep learning offer a Bayesian interpretation of common deep learning techniques such as dropout. This grounding of dropout in approximate Bayesian inference suggests an extension of the theoretical results, offering insights into the use of dropout with RNN models. We apply this new variational inference based dropout technique in LSTM and GRU models, assessing it on language modelling and sentiment analysis tasks. The new approach outperforms existing techniques, and to the best of our knowledge improves on the single model state-of-the-art in language modelling with the Penn Treebank (73.4 test perplexity). This extends our arsenal of variational tools in deep learning.

****

Automatic Neuron Detection in Calcium Imaging Data Using Convolutional Networks

Noah Apthorpe, Alexander Riordan, Robert Aguilar, Jan Homann, Yi Gu, David Tank, H. Sebastian Seung

Calcium imaging is an important technique for monitoring the activity of thousands of neurons simultaneously. As calcium imaging datasets grow in size, automated detection of individual neurons is becoming important. Here we apply a supervised learning approach to this problem and show that convolutional networks can achieve near-human accuracy and superhuman speed. Accuracy is superior to the popular PCA/ICA method based on precision and recall relative to ground truth annotation by a human expert. These results suggest that convolutional networks are an efficient and flexible tool for the analysis of large-scale calcium imaging data.

****

Convolutional Neural Fabrics

Shreyas Saxena, Jakob Verbeek

Despite the success of CNNs, selecting the optimal architecture for a given task remains an open problem. Instead of aiming to select a single optimal architecture, we propose a ``fabric'' that embeds an exponentially large number of architectures. The fabric consists of a 3D trellis that connects response maps at different layers, scales, and channels with a sparse homogeneous local connectivity pattern. The only hyper-parameters of a fabric are the number of channels and layers. While individual architectures can be recovered as paths, the fabric can in addition ensemble all embedded architectures together, sharing their weights where their paths overlap. Parameters can be learned using standard methods based on back-propagation, at a cost that scales linearly in the fabric size. We present benchmark results competitive with the state of the art for image classification on MNIST and CIFAR10, and for semantic segmentation on the Part Labels dataset.

****

Budgeted stream-based active learning via adaptive submodular maximization

Kaito Fujii, Hisashi Kashima

Active learning enables us to reduce the annotation cost by adaptively selecting unlabeled instances to be labeled. For pool-based active learning, several effective methods with theoretical guarantees have been developed through maximizing some utility function satisfying adaptive submodularity. In contrast, there hav

e been few methods for stream-based active learning based on adaptive submodularity. In this paper, we propose a new class of utility functions, policy-adaptive submodular functions, and prove this class includes many existing adaptive submodular functions appearing in real world problems. We provide a general framework based on policy-adaptive submodularity that makes it possible to convert existing pool-based methods to stream-based methods and give theoretical guarantees on their performance. In addition we empirically demonstrate their effectiveness comparing with existing heuristics on common benchmark datasets.
****

An equivalence between high dimensional Bayes optimal inference and M-estimation
Madhu Advani, Surya Ganguli

Due to the computational difficulty of performing MMSE (minimum mean squared error) inference, maximum a posteriori (MAP) is often used as a surrogate. However, the accuracy of MAP is suboptimal for high dimensional inference, where the number of model parameters is of the same order as the number of samples. In this work we demonstrate how MMSE performance is asymptotically achievable via optimization with an appropriately selected convex penalty and regularization function which are a smoothed version of the widely applied MAP algorithm. Our findings provide a new derivation and interpretation for recent optimal M-estimators discovered by El Karoui, et. al. PNAS 2013 as well as extending to non-additive noise models. We demonstrate the performance of these optimal M-estimators with numerical simulations. Overall, at the heart of our work is the revelation of a remarkable equivalence between two seemingly very different computational problems: namely that of high dimensional Bayesian integration, and high dimensional convex optimization. In essence we show that the former computationally difficult integral may be computed by solving the latter, simpler optimization problem.
****

A Sparse Interactive Model for Matrix Completion with Side Information
Jin Lu, Guannan Liang, Jiangwen Sun, Jinbo Bi

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.
****

Bi-Objective Online Matching and Submodular Allocations
Hossein Esfandiari, Nitish Korula, Vahab Mirrokni

Online allocation problems have been widely studied due to their numerous practical applications (particularly to Internet advertising), as well as considerable theoretical interest. The main challenge in such problems is making assignment decisions in the face of uncertainty about future input; effective algorithms need to predict which constraints are most likely to bind, and learn the balance between short-term gain and the value of long-term resource availability. In many important applications, the algorithm designer is faced with multiple objectives to optimize. In particular, in online advertising it is fairly common to optimize multiple metrics, such as clicks, conversions, and impressions, as well as other metrics which may be largely uncorrelated such as ‘share of voice’, and ‘buyer surplus’. While there has been considerable work on multi-objective offline optimization (when the entire input is known in advance), very little is known about the online case, particularly in the case of adversarial input. In this paper, we give the first results for bi-objective online submodular optimization, providing almost matching upper and lower bounds for allocating items to agents with two submodular value functions. We also study practically relevant special cases of this problem related to Internet advertising, and obtain improved results. All our algorithms are nearly best possible, as well as being efficient and easy to implement in practice.
****

Interpretable Distribution Features with Maximum Testing Power
Wittawat Jitkrittum, Zoltán Szabó, Kacper P. Chwialkowski, Arthur Gretton

Two semimetrics on probability distributions are proposed, given as the sum of differences of expectations of analytic functions evaluated at spatial or frequen cy locations (i.e, features). The features are chosen so as to maximize the distinguishability of the distributions, by optimizing a lower bound on test power for a statistical test using these features. The result is a parsimonious and interpretable indication of how and where two distributions differ locally. An empirical estimate of the test power criterion converges with increasing sample size, ensuring the quality of the returned features. In real-world benchmarks on high-dimensional text and image data, linear-time tests using the proposed semimetrics achieve comparable performance to the state-of-the-art quadratic-time maximum mean discrepancy test, while returning human-interpretable features that explain the test results.

****

Finding significant combinations of features in the presence of categorical covariates

Laetitia Papaxanthos, Felipe Llinares-López, Dean Bodenham, Karsten Borgwardt

In high-dimensional settings, where the number of features p is typically much larger than the number of samples n, methods which can systematically examine arbitrary combinations of features, a huge 2^p-dimensional space, have recently begun to be explored. However, none of the current methods is able to assess the association between feature combinations and a target variable while conditioning on a categorical covariate, in order to correct for potential confounding effects. We propose the Fast Automatic Conditional Search (FACS) algorithm, a significant discriminative itemset mining method which conditions on categorical covariates and only scales as O(k log k), where k is the number of states of the categorical covariate. Based on the Cochran-Mantel-Haenszel Test, FACS demonstrates superior speed and statistical power on simulated and real-world datasets compared to the state of the art, opening the door to numerous applications in biomedicine.

****

A Non-convex One-Pass Framework for Generalized Factorization Machine and Rank-One Matrix Sensing

Ming Lin, Jieping Ye

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.

****

Avoiding Imposters and Delinquents: Adversarial Crowdsourcing and Peer Prediction

Jacob Steinhardt, Gregory Valiant, Moses Charikar

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.

****

Threshold Bandits, With and Without Censored Feedback

Jacob D. Abernethy, Kareem Amin, Ruihao Zhu

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 Bayes on Monte Carlo Steroids

Aditya Grover, Stefano Ermon

Variational approaches are often used to approximate intractable posteriors or normalization constants in hierarchical latent variable models. While often effective in practice, it is known that the approximation error can be arbitrarily large. We propose a new class of bounds on the marginal log-likelihood of directed latent variable models. Our approach relies on random projections to simplify the posterior. In contrast to standard variational methods, our bounds are guaranteed to be tight with high probability. We provide a new approach for learning latent variable models based on optimizing our new bounds on the log-likelihood. We demonstrate empirical improvements on benchmark datasets in vision and language for sigmoid belief networks, where a neural network is used to approximate the posterior.

****

Finite-Dimensional BFRY Priors and Variational Bayesian Inference for Power Law Models

Juho Lee, Lancelot F. James, Seungjin Choi

Bayesian nonparametric methods based on the Dirichlet process (DP), gamma process and beta process, have proven effective in capturing aspects of various datasets arising in machine learning. However, it is now recognized that such processes have their limitations in terms of the ability to capture power law behavior. As such there is now considerable interest in models based on the Stable Processs (SP), Generalized Gamma process (GGP) and Stable-beta process (SBP). These models present new challenges in terms of practical statistical implementation. In analogy to tractable processes such as the finite-dimensional Dirichlet process, we describe a class of random processes, we call iid finite-dimensional BFRY processes, that enables one to begin to develop efficient posterior inference algorithms such as variational Bayes that readily scale to massive datasets. For illustrative purposes, we describe a simple variational Bayes algorithm for normalized SP mixture models, and demonstrate its usefulness with experiments on synthetic and real-world datasets.

****

Maximal Sparsity with Deep Networks?

Bo Xin, Yizhou Wang, Wen Gao, David Wipf, Baoyuan Wang

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.

****

Single-Image Depth Perception in the Wild

Weifeng Chen, Zhao Fu, Dawei Yang, Jia Deng

This paper studies single-image depth perception in the wild, i.e., recovering depth from a single image taken in unconstrained settings. We introduce a new dataset “Depth in the Wild” consisting of images in the wild annotated with relative depth between pairs of random points. We also propose a new algorithm that learns to estimate metric depth using annotations of relative depth. Compared to the state of the art, our algorithm is simpler and performs better. Experiments show that our algorithm, combined with existing RGB-D data and our new relative depth annotations, significantly improves single-image depth perception in the wild.

****

Single Pass PCA of Matrix Products

Shanshan Wu, Srinadh Bhojanapalli, Sujay Sanghavi, 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.

****

Optimal Sparse Linear Encoders and Sparse PCA

Malik Magdon-Ismail, Christos Boutsidis

Principal components analysis~(PCA) is the optimal linear encoder of data. Sparse linear encoders (e.g., sparse PCA) produce more interpretable features that can promote better generalization. (\rn{1}) Given a level of sparsity, what is the best approximation to PCA? (\rn{2}) Are there efficient algorithms which can achieve this optimal combinatorial tradeoff? We answer both questions by providing the first polynomial-time algorithms to construct \emph{optimal} sparse linear auto-encoders; additionally, we demonstrate the performance of our algorithms on real data.

****

Measuring the reliability of MCMC inference with bidirectional Monte Carlo
Roger B. Grosse, Siddharth Ancha, Daniel M. Roy
Markov chain Monte Carlo (MCMC) is one of the main workhorses of probabilistic inference, but it is notoriously hard to measure the quality of approximate posterior samples. This challenge is particularly salient in black box inference methods, which can hide details and obscure inference failures. In this work, we extend the recently introduced bidirectional Monte Carlo technique to evaluate MCMC-based posterior inference algorithms. By running annealed importance sampling (AIS) chains both from prior to posterior and vice versa on simulated data, we upper bound in expectation the symmetrized KL divergence between the true posterior distribution and the distribution of approximate samples. We integrate our method into two probabilistic programming languages, WebPPL and Stan, and validate it on several models and datasets. As an example of how our method be used to guide the design of inference algorithms, we apply it to study the effectiveness of different model representations in WebPPL and Stan.
****
Image Restoration Using Very Deep Convolutional Encoder-Decoder Networks with Symmetric Skip Connections
Xiaojiao Mao, Chunhua Shen, Yu-Bin Yang
In this paper, we propose a very deep fully convolutional encoding-decoding framework for image restoration such as denoising and super-resolution. The network is composed of multiple layers of convolution and deconvolution operators, learning end-to-end mappings from corrupted images to the original ones. The convolutional layers act as the feature extractor, which capture the abstraction of image contents while eliminating noises/corruptions. Deconvolutional layers are then used to recover the image details. We propose to symmetrically link convolutional and deconvolutional layers with skip-layer connections, with which the training converges much faster and attains a higher-quality local optimum. First, the skip connections allow the signal to be back-propagated to bottom layers directly, and thus tackles the problem of gradient vanishing, making training deep networks easier and achieving restoration performance gains consequently. Second, these skip connections pass image details from convolutional layers to deconvolutional layers, which is beneficial in recovering the original image. Significantly, with the large capacity, we can handle different levels of noises using a single model. Experimental results show that our network achieves better performance than recent state-of-the-art methods.
****
On Valid Optimal Assignment Kernels and Applications to Graph Classification
Nils M. Kriege, Pierre-Louis Giscard, Richard Wilson
The success of kernel methods has initiated the design of novel positive semidefinite functions, in particular for structured data. A leading design paradigm for this is the convolution kernel, which decomposes structured objects into their parts and sums over all pairs of parts. Assignment kernels, in contrast, are obtained from an optimal bijection between parts, which can provide a more valid notion of similarity. In general however, optimal assignments yield indefinite functions, which complicates their use in kernel methods. We characterize a class of base kernels used to compare parts that guarantees positive semidefinite optimal assignment kernels. These base kernels give rise to hierarchies from which the optimal assignment kernels are computed in linear time by histogram intersection. We apply these results by developing the Weisfeiler-Lehman optimal assignment kernel for graphs. It provides high classification accuracy on widely-used benchmark data sets improving over the original Weisfeiler-Lehman kernel.
****
Multi-view Anomaly Detection via Robust Probabilistic Latent Variable Models
Tomoharu Iwata, Makoto Yamada
We propose probabilistic latent variable models for multi-view anomaly detection, which is the task of finding instances that have inconsistent views given multi-view data. With the proposed model, all views of a non-anomalous instance are assumed to be generated from a single latent vector. On the other hand, an anomalous instance is assumed to have multiple latent vectors, and its different view

s are generated from different latent vectors. By inferring the number of latent vectors used for each instance with Dirichlet process priors, we obtain multi-view anomaly scores. The proposed model can be seen as a robust extension of probabilistic canonical correlation analysis for noisy multi-view data. We present Bayesian inference procedures for the proposed model based on a stochastic EM algorithm. The effectiveness of the proposed model is demonstrated in terms of performance when detecting multi-view anomalies.
****

Optimal Architectures in a Solvable Model of Deep Networks
Jonathan Kadmon, Haim Sompolinsky

Deep neural networks have received a considerable attention due to the success of their training for real world machine learning applications. They are also of great interest to the understanding of sensory processing in cortical sensory hierarchies. The purpose of this work is to advance our theoretical understanding of the computational benefits of these architectures. Using a simple model of clustered noisy inputs and a simple learning rule, we provide analytically derived recursion relations describing the propagation of the signals along the deep network. By analysis of these equations, and defining performance measures, we show that these model networks have optimal depths. We further explore the dependence of the optimal architecture on the system parameters.
****

Efficient state-space modularization for planning: theory, behavioral and neural signatures
Daniel McNamee, Daniel M. Wolpert, Mate Lengyel

Even in state-spaces of modest size, planning is plagued by the “curse of dimensionality”. This problem is particularly acute in human and animal cognition given the limited capacity of working memory, and the time pressures under which planning often occurs in the natural environment. Hierarchically organized modular representations have long been suggested to underlie the capacity of biological systems to efficiently and flexibly plan in complex environments. However, the principles underlying efficient modularization remain obscure, making it difficult to identify its behavioral and neural signatures. Here, we develop a normative theory of efficient state-space representations which partitions an environment into distinct modules by minimizing the average (information theoretic) description length of planning within the environment, thereby optimally trading off the complexity of planning across and within modules. We show that such optimal representations provide a unifying account for a diverse range of hitherto unrelated phenomena at multiple levels of behavior and neural representation.
****

A Communication-Efficient Parallel Algorithm for Decision Tree
Qi Meng, Guolin Ke, Taifeng Wang, Wei Chen, Qiwei Ye, Zhi-Ming Ma, Tie-Yan 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.
****

Supervised Word Mover's Distance
Gao Huang, Chuan Guo, Matt J. Kusner, Yu Sun, Fei Sha, Kilian Q. Weinberger

Accurately measuring the similarity between text documents lies at the core of many real world applications of machine learning. These include web-search ranking, document recommendation, multi-lingual document matching, and article categorization. Recently, a new document metric, the word mover's distance (WMD), has been proposed with unprecedented results on kNN-based document classification. The WMD elevates high quality word embeddings to document metrics by formulating the distance between two documents as an optimal transport problem between the embedded words. However, the document distances are entirely unsupervised and lack a mechanism to incorporate supervision when available. In this paper we propose an efficient technique to learn a supervised metric, which we call the Supervised WMD (S-WMD) metric. Our algorithm learns document distances that measure the underlying semantic differences between documents by leveraging semantic differe nces between individual words discovered during supervised training. This is achieved with an linear transformation of the underlying word embedding space and tailored word-specific weights, learned to minimize the stochastic leave-one-out nearest neighbor classification error on a per-document level. We evaluate our metric on eight real-world text classification tasks on which S-WMD consistently outperforms almost all of our 26 competitive baselines.

****

Fast and accurate spike sorting of high-channel count probes with KiloSort

Marius Pachitariu, Nicholas A. Steinmetz, Shabnam N. Kadir, Matteo Carandini, Kenneth D. Harris

New silicon technology is enabling large-scale electrophysiological recordings in vivo from hundreds to thousands of channels. Interpreting these recordings requires scalable and accurate automated methods for spike sorting, which should minimize the time required for manual curation of the results. Here we introduce KiloSort, a new integrated spike sorting framework that uses template matching both during spike detection and during spike clustering. KiloSort models the electrical voltage as a sum of template waveforms triggered on the spike times, which allows overlapping spikes to be identified and resolved. Unlike previous algorithms that compress the data with PCA, KiloSort operates on the raw data which allows it to construct a more accurate model of the waveforms. Processing times are faster than in previous algorithms thanks to batch-based optimization on GPUs. We compare KiloSort to an established algorithm and show favorable performance, at much reduced processing times. A novel post-clustering merging step based on the continuity of the templates further reduced substantially the number of manual operations required on this data, for the neurons with near-zero error rates, paving the way for fully automated spike sorting of multichannel electrode recordings.

****

Learning brain regions via large-scale online structured sparse dictionary learning

Elvis DOHMATOB, Arthur Mensch, Gael Varoquaux, Bertrand Thirion

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.

****

Improving PAC Exploration Using the Median Of Means

Jason Pazis, Ronald E. Parr, Jonathan P. How

We present the first application of the median of means in a PAC exploration algorithm for MDPs. Using the median of means allows us to significantly reduce the dependence of our bounds on the range of values that the value function can take, while introducing a dependence on the (potentially much smaller) variance of the Bellman operator. Additionally, our algorithm is the first algorithm with PAC bounds that can be applied to MDPs with unbounded rewards.

****

Active Nearest-Neighbor Learning in Metric Spaces

Aryeh Kontorovich, Sivan Sabato, Ruth Urner

We propose a pool-based non-parametric active learning algorithm for general metric spaces, called MArgin Regularized Metric Active Nearest Neighbor (MARMANN), which outputs a nearest-neighbor classifier. We give prediction error guarantees that depend on the noisy-margin properties of the input sample, and are competitive with those obtained by previously proposed passive learners. We prove that the label complexity of MARMANN is significantly lower than that of any passive learner with similar error guarantees. Our algorithm is based on a generalized sample compression scheme and a new label-efficient active model-selection procedure.

****

Learning from Small Sample Sets by Combining Unsupervised Meta-Training with CNNs

Yu-Xiong Wang, Martial Hebert

This work explores CNNs for the recognition of novel categories from few examples. Inspired by the transferability properties of CNNs, we introduce an additional unsupervised meta-training stage that exposes multiple top layer units to a large amount of unlabeled real-world images. By encouraging these units to learn diverse sets of low-density separators across the unlabeled data, we capture a more generic, richer description of the visual world, which decouples these units from ties to a specific set of categories. We propose an unsupervised margin maximization that jointly estimates compact high-density regions and infers low-density separators. The low-density separator (LDS) modules can be plugged into any or all of the top layers of a standard CNN architecture. The resulting CNNs significantly improve the performance in scene classification, fine-grained recognition, and action recognition with small training samples.

****

Learning Bayesian networks with ancestral constraints

Eunice Yuh-Jie Chen, Yujia Shen, Arthur Choi, Adnan Darwiche

We consider the problem of learning Bayesian networks optimally, when subject to background knowledge in the form of ancestral constraints. Our approach is based on a recently proposed framework for optimal structure learning based on non-decomposable scores, which is general enough to accommodate ancestral constraints. The proposed framework exploits oracles for learning structures using decomposable scores, which cannot accommodate ancestral constraints since they are non-decomposable. We show how to empower these oracles by passing them decomposable constraints that they can handle, which are inferred from ancestral constraints that they cannot handle. Empirically, we demonstrate that our approach can be orders-of-magnitude more efficient than alternative frameworks, such as those based on integer linear programming.

****

Exponential expressivity in deep neural networks through transient chaos

Ben Poole, Subhaneil Lahiri, Maithra Raghu, Jascha Sohl-Dickstein, Surya Ganguli

We combine Riemannian geometry with the mean field theory of high dimensional chaos to study the nature of signal propagation in deep neural networks with random weights. Our results reveal a phase transition in the expressivity of random deep networks, with networks in the chaotic phase computing nonlinear functions whose global curvature grows exponentially with depth, but not with width. We prove that this generic class of random functions cannot be efficiently computed by any shallow network, going beyond prior work that restricts their analysis to single functions. Moreover, we formally quantify and demonstrate the long conjectured idea that deep networks can disentangle exponentially curved manifolds in input space into flat manifolds in hidden space. Our theoretical framework for analyzing the expressive power of deep networks is broadly applicable and provides a basis for quantifying previously abstract notions about the geometry of deep functions.

****

MetaGrad: Multiple Learning Rates in Online Learning

Tim van Erven, Wouter M. Koolen

In online convex optimization it is well known that certain subclasses of objective functions are much easier than arbitrary convex functions. We are interested in designing adaptive methods that can automatically get fast rates in as many such subclasses as possible, without any manual tuning. Previous adaptive methods are able to interpolate between strongly convex and general convex functions. We present a new method, MetaGrad, that adapts to a much broader class of functions, including exp-concave and strongly convex functions, but also various types of stochastic and non-stochastic functions without any curvature. For instance, MetaGrad can achieve logarithmic regret on the unregularized hinge loss, even though it has no curvature, if the data come from a favourable probability distribution. MetaGrad's main feature is that it simultaneously considers multiple learning rates. Unlike all previous methods with provable regret guarantees, however, its learning rates are not monotonically decreasing over time and are not tuned based on a theoretically derived bound on the regret. Instead, they are weighted directly proportional to their empirical performance on the data using a til ted exponential weights master algorithm.

****

Learning under uncertainty: a comparison between R-W and Bayesian approach

He Huang, Martin Paulus

Accurately differentiating between what are truly unpredictably random and systematic changes that occur at random can have profound effect on affect and cognition. To examine the underlying computational principles that guide different learning behavior in an uncertain environment, we compared an R-W model and a Bayesian approach in a visual search task with different volatility levels. Both R-W model and the Bayesian approach reflected an individual's estimation of the environmental volatility, and there is a strong correlation between the learning rate in R-W model and the belief of stationarity in the Bayesian approach in different volatility conditions. In a low volatility condition, R-W model indicates that learning rate positively correlates with lose-shift rate, but not choice optimality (inverted U shape). The Bayesian approach indicates that the belief of environmental stationarity positively correlates with choice optimality, but not lose-shift rate (inverted U shape). In addition, we showed that comparing to Expert learners, individuals with high lose-shift rate (sub-optimal learners) had significantly higher learning rate estimated from R-W model and lower belief of stationarity from the Bayesian model.

****

End-to-End Goal-Driven Web Navigation

Rodrigo Nogueira, Kyunghyun Cho

We propose a goal-driven web navigation as a benchmark task for evaluating an agent with abilities to understand natural language and plan on partially observed environments. In this challenging task, an agent navigates through a website, which is represented as a graph consisting of web pages as nodes and hyperlinks as directed edges, to find a web page in which a query appears. The agent is required to have sophisticated high-level reasoning based on natural languages and efficient sequential decision-making capability to succeed. We release a software tool, called WebNav, that automatically transforms a website into this goal-driven web navigation task, and as an example, we make WikiNav, a dataset constructed from the English Wikipedia. We extensively evaluate different variants of neural net based artificial agents on WikiNav and observe that the proposed goal-driven web navigation well reflects the advances in models, making it a suitable benchmark for evaluating future progress. Furthermore, we extend the WikiNav with question-answer pairs from Jeopardy! and test the proposed agent based on recurrent neural networks against strong inverted index based search engines. The artificial agents trained on WikiNav outperforms the engined based approaches, demonstrating the capability of the proposed goal-driven navigation as a good proxy for measuring the progress in real-world tasks such as focused crawling and question-answering.

****

Higher-Order Factorization Machines

Mathieu Blondel, Akinori Fujino, Naonori Ueda, Masakazu Ishihata

Factorization machines (FMs) are a supervised learning approach that can use second-order feature combinations even when the data is very high-dimensional. Unfortunately, despite increasing interest in FMs, there exists to date no efficient training algorithm for higher-order FMs (HOFMs). In this paper, we present the first generic yet efficient algorithms for training arbitrary-order HOFMs. We also present new variants of HOFMs with shared parameters, which greatly reduce model size and prediction times while maintaining similar accuracy. We demonstrate the proposed approaches on four different link prediction tasks.

****

Efficient Second Order Online Learning by Sketching

Haipeng Luo, Alekh Agarwal, Nicolò Cesa-Bianchi, John Langford

We propose Sketched Online Newton (SON), an online second order learning algorithm that enjoys substantially improved regret guarantees for ill-conditioned data. SON is an enhanced version of the Online Newton Step, which, via sketching techniques enjoys a running time linear in the dimension and sketch size. We furth er develop sparse forms of the sketching methods (such as Oja's rule), making the computation linear in the sparsity of features. Together, the algorithm eliminates all computational obstacles in previous second order online learning approaches.

****

Professor Forcing: A New Algorithm for Training Recurrent Networks

Alex M. Lamb, Anirudh Goyal ALIAS PARTH GOYAL, Ying Zhang, Saizheng Zhang, Aaron C. Courville, Yoshua Bengio

The Teacher Forcing algorithm trains recurrent networks by supplying observed sequence values as inputs during training and using the network’s own one-step-ahead predictions to do multi-step sampling. We introduce the Professor Forcing algorithm, which uses adversarial domain adaptation to encourage the dynamics of the recurrent network to be the same when training the network and when sampling from the network over multiple time steps. We apply Professor Forcing to language modeling, vocal synthesis on raw waveforms, handwriting generation, and image generation. Empirically we find that Professor Forcing acts as a regularizer, improving test likelihood on character level Penn Treebank and sequential MNIST. We also find that the model qualitatively improves samples, especially when sampling for a large number of time steps. This is supported by human evaluation of sample quality. Trade-offs between Professor Forcing and Scheduled Sampling are discussed. We produce T-SNEs showing that Professor Forcing successfully makes the dynamics of the network during training and sampling more similar.

****

Deep ADMM-Net for Compressive Sensing MRI

yan yang, Jian Sun, Huibin Li, Zongben Xu

Compressive Sensing (CS) is an effective approach for fast Magnetic Resonance Imaging (MRI). It aims at reconstructing MR image from a small number of under-sampled data in k-space, and accelerating the data acquisition in MRI. To improve the current MRI system in reconstruction accuracy and computational speed, in this paper, we propose a novel deep architecture, dubbed ADMM-Net. ADMM-Net is defined over a data flow graph, which is derived from the iterative procedures in Alternating Direction Method of Multipliers (ADMM) algorithm for optimizing a CS-based MRI model. In the training phase, all parameters of the net, e.g., image transforms, shrinkage functions, etc., are discriminatively trained end-to-end using L-BFGS algorithm. In the testing phase, it has computational overhead similar to ADMM but uses optimized parameters learned from the training data for CS-based reconstruction task. Experiments on MRI image reconstruction under different sampling ratios in k-space demonstrate that it significantly improves the baseline ADMM algorithm and achieves high reconstruction accuracies with fast computational speed.

****

Adaptive Averaging in Accelerated Descent Dynamics

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.

****

Dynamic Mode Decomposition with Reproducing Kernels for Koopman Spectral Analysis

Yoshinobu Kawahara

A spectral analysis of the Koopman operator, which is an infinite dimensional linear operator on an observable, gives a (modal) description of the global behavior of a nonlinear dynamical system without any explicit prior knowledge of its governing equations. In this paper, we consider a spectral analysis of the Koopman operator in a reproducing kernel Hilbert space (RKHS). We propose a modal decomposition algorithm to perform the analysis using finite-length data sequences generated from a nonlinear system. The algorithm is in essence reduced to the calculation of a set of orthogonal bases for the Krylov matrix in RKHS and the eigendecomposition of the projection of the Koopman operator onto the subspace spann ed by the bases. The algorithm returns a decomposition of the dynamics into a finite number of modes, and thus it can be thought of as a feature extraction procedure for a nonlinear dynamical system. Therefore, we further consider applications in machine learning using extracted features with the presented analysis. We illustrate the method on the applications using synthetic and real-world data.

****

Total Variation Classes Beyond 1d: Minimax Rates, and the Limitations of Linear Smoothers

Veeranjaneyulu Sadhanala, Yu-Xiang Wang, Ryan J. Tibshirani

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.

****

Hardness of Online Sleeping Combinatorial Optimization Problems

Satyen Kale, Chansoo Lee, David Pal

We show that several online combinatorial optimization problems that admit efficient no-regret algorithms become computationally hard in the sleeping setting where a subset of actions becomes unavailable in each round. Specifically, we show that the sleeping versions of these problems are at least as hard as PAC learning DNF expressions, a long standing open problem. We show hardness for the sleeping versions of Online Shortest Paths, Online Minimum Spanning Tree, Online k-Subsets, Online k-Truncated Permutations, Online Minimum Cut, and Online Bipartite Matching. The hardness result for the sleeping version of the Online Shortest Paths problem resolves an open problem presented at COLT 2015 [Koolen et al., 2015].

****

Density Estimation via Discrepancy Based Adaptive Sequential Partition

Dangna Li, Kun Yang, Wing Hung Wong

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.

****

Quantized Random Projections and Non-Linear Estimation of Cosine Similarity

Ping Li, Michael Mitzenmacher, Martin Slawski

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.

****

Algorithms and matching lower bounds for approximately-convex optimization

Andrej Risteski, Yuanzhi 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.

****

The Parallel Knowledge Gradient Method for Batch Bayesian Optimization

Jian Wu, Peter Frazier

In many applications of black-box optimization, one can evaluate multiple points simultaneously, e.g. when evaluating the performances of several different neural network architectures in a parallel computing environment. In this paper, we develop a novel batch Bayesian optimization algorithm --- the parallel knowledge gradient method. By construction, this method provides the one-step Bayes optimal batch of points to sample. We provide an efficient strategy for computing this Bayes-optimal batch of points, and we demonstrate that the parallel knowledge gradient method finds global optima significantly faster than previous batch Bayesian optimization algorithms on both synthetic test functions and when tuning hyperparameters of practical machine learning algorithms, especially when functi

on evaluations are noisy.
****

Edge-exchangeable graphs and sparsity
Diana Cai, Trevor Campbell, Tamara Broderick

Many popular network models rely on the assumption of (vertex) exchangeability, in which the distribution of the graph is invariant to relabelings of the vertices. However, the Aldous-Hoover theorem guarantees that these graphs are dense or empty with probability one, whereas many real-world graphs are sparse. We present an alternative notion of exchangeability for random graphs, which we call edge exchangeability, in which the distribution of a graph sequence is invariant to the order of the edges. We demonstrate that edge-exchangeable models, unlike models that are traditionally vertex exchangeable, can exhibit sparsity. To do so, we outline a general framework for graph generative models; by contrast to the pioneering work of Caron and Fox (2015), models within our framework are stationary across steps of the graph sequence. In particular, our model grows the graph by instantiating more latent atoms of a single random measure as the dataset size increases, rather than adding new atoms to the measure.
****

Stochastic Variance Reduction Methods for Saddle-Point Problems
Balamurugan Palaniappan, Francis Bach

We consider convex-concave saddle-point problems where the objective functions may be split in many components, and extend recent stochastic variance reduction methods (such as SVRG or SAGA) to provide the first large-scale linearly convergent algorithms for this class of problems which are common in machine learning. While the algorithmic extension is straightforward, it comes with challenges and opportunities: (a) the convex minimization analysis does not apply and we use the notion of monotone operators to prove convergence, showing in particular that the same algorithm applies to a larger class of problems, such as variational inequalities, (b) there are two notions of splits, in terms of functions, or in terms of partial derivatives, (c) the split does need to be done with convex-concave terms, (d) non-uniform sampling is key to an efficient algorithm, both in theory and practice, and (e) these incremental algorithms can be easily accelerated using a simple extension of the "catalyst" framework, leading to an algorithm which is always superior to accelerated batch algorithms.
****

A Probabilistic Model of Social Decision Making based on Reward Maximization
Koosha Khalvati, Seongmin A. Park, Jean-Claude Dreher, Rajesh PN Rao

A fundamental problem in cognitive neuroscience is how humans make decisions, act, and behave in relation to other humans. Here we adopt the hypothesis that when we are in an interactive social setting, our brains perform Bayesian inference of the intentions and cooperativeness of others using probabilistic representations. We employ the framework of partially observable Markov decision processes (POMDPs) to model human decision making in a social context, focusing specifically on the volunteer's dilemma in a version of the classic Public Goods Game. We show that the POMDP model explains both the behavior of subjects as well as neural activity recorded using fMRI during the game. The decisions of subjects can be modeled across all trials using two interpretable parameters. Furthermore, the expected reward predicted by the model for each subject was correlated with the activation of brain areas related to reward expectation in social interactions. Our results suggest a probabilistic basis for human social decision making within the framework of expected reward maximization.
****

Bootstrap Model Aggregation for Distributed Statistical Learning
JUN HAN, Qiang Liu

In distributed, or privacy-preserving learning, we are often given a set of probabilistic models estimated from different local repositories, and asked to combine them into a single model that gives efficient statistical estimation. A simple method is to linearly average the parameters of the local models, which, however, tends to be degenerate or not applicable on non-convex models, or models with different parameter dimensions. One more practical strategy is to generate boo tstrap samples from the local models, and then learn a joint model based on the combined bootstrap set. Unfortunately, the bootstrap procedure introduces additional noise and can significantly deteriorate the performance. In this work, we propose two variance reduction methods to correct the bootstrap noise, including a weighted M-estimator that is both statistically efficient and practically powerful. Both theoretical and empirical analysis is provided to demonstrate our methods.

****

Unsupervised Learning of 3D Structure from Images

Danilo Jimenez Rezende, S. M. Ali Eslami, Shakir Mohamed, Peter Battaglia, Max Jaderberg, Nicolas Heess

A key goal of computer vision is to recover the underlying 3D structure that gives rise to 2D observations of the world. If endowed with 3D understanding, agents can abstract away from the complexity of the rendering process to form stable, disentangled representations of scene elements. In this paper we learn strong deep generative models of 3D structures, and recover these structures from 2D images via probabilistic inference. We demonstrate high-quality samples and report log-likelihoods on several datasets, including ShapeNet, and establish the first benchmarks in the literature. We also show how these models and their inference networks can be trained jointly, end-to-end, and directly from 2D images without any use of ground-truth 3D labels. This demonstrates for the first time the feasibility of learning to infer 3D representations of the world in a purely unsupervised manner.

****

beta-risk: a New Surrogate Risk for Learning from Weakly Labeled Data

Valentina Zantedeschi, Rémi Emonet, Marc Sebban

During the past few years, the machine learning community has paid attention to developping new methods for learning from weakly labeled data. This field covers different settings like semi-supervised learning, learning with label proportions, multi-instance learning, noise-tolerant learning, etc. This paper presents a generic framework to deal with these weakly labeled scenarios. We introduce the beta-risk as a generalized formulation of the standard empirical risk based on surrogate margin-based loss functions. This risk allows us to express the reliability on the labels and to derive different kinds of learning algorithms. We specifically focus on SVMs and propose a soft margin beta-svm algorithm which behaves better that the state of the art.

****

Learning Supervised PageRank with Gradient-Based and Gradient-Free Optimization Methods

Lev Bogolubsky, Pavel Dvurechenskii, Alexander Gasnikov, Gleb Gusev, Yurii Nesterov, Andrei M. Raigorodskii, Aleksey Tikhonov, Maksim Zhukovskii

In this paper, we consider a non-convex loss-minimization problem of learning Supervised PageRank models, which can account for features of nodes and edges. We propose gradient-based and random gradient-free methods to solve this problem. Our algorithms are based on the concept of an inexact oracle and unlike the state-of-the-art gradient-based method we manage to provide theoretically the convergence rate guarantees for both of them. Finally, we compare the performance of the proposed optimization methods with the state of the art applied to a ranking task.

****

Globally Optimal Training of Generalized Polynomial Neural Networks with Nonlinear Spectral Methods

Antoine Gautier, Quynh N. Nguyen, Matthias Hein

The optimization problem behind neural networks is highly non-convex. Training with stochastic gradient descent and variants requires careful parameter tuning and provides no guarantee to achieve the global optimum. In contrast we show under quite weak assumptions on the data that a particular class of feedforward neural networks can be trained globally optimal with a linear convergence rate. Up to our knowledge this is the first practically feasible method which achieves such a guarantee. While the method can in principle be applied to deep networks, w

e restrict ourselves for simplicity in this paper to one- and two hidden layer networks. Our experiments confirms that these models are already rich enough to achieve good performance on a series of real-world datasets.
****

Optimal Black-Box Reductions Between Optimization Objectives
Zeyuan Allen-Zhu, Elad Hazan

The diverse world of machine learning applications has given rise to a plethora of algorithms and optimization methods, finely tuned to the specific regression or classification task at hand. We reduce the complexity of algorithm design for machine learning by reductions: we develop reductions that take a method developed for one setting and apply it to the entire spectrum of smoothness and strong-convexity in applications. Furthermore, unlike existing results, our new reductions are OPTIMAL and more PRACTICAL. We show how these new reductions give rise to new and faster running times on training linear classifiers for various families of loss functions, and conclude with experiments showing their successes also in practice.
****

Sequential Neural Models with Stochastic Layers
Marco Fraccaro, Søren Kaae Sønderby, Ulrich Paquet, Ole Winther

How can we efficiently propagate uncertainty in a latent state representation with recurrent neural networks? This paper introduces stochastic recurrent neural networks which glue a deterministic recurrent neural network and a state space model together to form a stochastic and sequential neural generative model. The clear separation of deterministic and stochastic layers allows a structured variational inference network to track the factorization of the model’s posterior distribution. By retaining both the nonlinear recursive structure of a recurrent neural network and averaging over the uncertainty in a latent path, like a state space model, we improve the state of the art results on the Blizzard and TIMIT speech modeling data sets by a large margin, while achieving comparable performances to competing methods on polyphonic music modeling.
****

Iterative Refinement of the Approximate Posterior for Directed Belief Networks
Devon Hjelm, Russ R. Salakhutdinov, Kyunghyun Cho, Nebojsa Jojic, Vince Calhoun, Junyoung Chung

Variational methods that rely on a recognition network to approximate the posterior of directed graphical models offer better inference and learning than previous methods. Recent advances that exploit the capacity and flexibility in this approach have expanded what kinds of models can be trained. However, as a proposal for the posterior, the capacity of the recognition network is limited, which can constrain the representational power of the generative model and increase the variance of Monte Carlo estimates. To address these issues, we introduce an iterative refinement procedure for improving the approximate posterior of the recognition network and show that training with the refined posterior is competitive with state-of-the-art methods. The advantages of refinement are further evident in an increased effective sample size, which implies a lower variance of gradient estimates.
****

Stochastic Multiple Choice Learning for Training Diverse Deep Ensembles
Stefan Lee, Senthil Purushwalkam Shiva Prakash, Michael Cogswell, Viresh Ranjan, David Crandall, Dhruv Batra

Many practical perception systems exist within larger processes which often include interactions with users or additional components that are capable of evaluating the quality of predicted solutions. In these contexts, it is beneficial to provide these oracle mechanisms with multiple highly likely hypotheses rather than a single prediction. In this work, we pose the task of producing multiple outputs as a learning problem over an ensemble of deep networks -- introducing a novel stochastic gradient descent based approach to minimize the loss with respect to an oracle. Our method is simple to implement, agnostic to both architecture and loss function, and parameter-free. Our approach achieves lower oracle error compared to existing methods on a wide range of tasks and deep architectures. We

also show qualitatively that solutions produced from our approach often provide interpretable representations of task ambiguity.
****

Learning shape correspondence with anisotropic convolutional neural networks

Davide Boscaini, Jonathan Masci, Emanuele Rodolà, Michael Bronstein

Convolutional neural networks have achieved extraordinary results in many computer vision and pattern recognition applications; however, their adoption in the computer graphics and geometry processing communities is limited due to the non-Euclidean structure of their data. In this paper, we propose Anisotropic Convolutional Neural Network (ACNN), a generalization of classical CNNs to non-Euclidean domains, where classical convolutions are replaced by projections over a set of oriented anisotropic diffusion kernels. We use ACNNs to effectively learn intrinsic dense correspondences between deformable shapes, a fundamental problem in geometry processing, arising in a wide variety of applications. We tested ACNNs performance in very challenging settings, achieving state-of-the-art results on some of the most difficult recent correspondence benchmarks.
****

Learning Tree Structured Potential Games

Vikas Garg, Tommi Jaakkola

Many real phenomena, including behaviors, involve strategic interactions that can be learned from data. We focus on learning tree structured potential games where equilibria are represented by local maxima of an underlying potential function. We cast the learning problem within a max margin setting and show that the problem is NP-hard even when the strategic interactions form a tree. We develop a variant of dual decomposition to estimate the underlying game and demonstrate with synthetic and real decision/voting data that the game theoretic perspective (carving out local maxima) enables meaningful recovery.
****

RETAIN: An Interpretable Predictive Model for Healthcare using Reverse Time Attention Mechanism

Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, Walter Stewart

Accuracy and interpretability are two dominant features of successful predictive models. Typically, a choice must be made in favor of complex black box models such as recurrent neural networks (RNN) for accuracy versus less accurate but more interpretable traditional models such as logistic regression. This tradeoff poses challenges in medicine where both accuracy and interpretability are important. We addressed this challenge by developing the REverse Time AttentIoN model (RETAIN) for application to Electronic Health Records (EHR) data. RETAIN achieves high accuracy while remaining clinically interpretable and is based on a two-level neural attention model that detects influential past visits and significant clinical variables within those visits (e.g. key diagnoses). RETAIN mimics physician practice by attending the EHR data in a reverse time order so that recent clinical visits are likely to receive higher attention. RETAIN was tested on a large health system EHR dataset with 14 million visits completed by 263K patients over an 8 year period and demonstrated predictive accuracy and computational scalability comparable to state-of-the-art methods such as RNN, and ease of interpretability comparable to traditional models.
****

PAC Reinforcement Learning with Rich Observations

Akshay Krishnamurthy, Alekh Agarwal, John Langford

We propose and study a new model for reinforcement learning with rich observations, generalizing contextual bandits to sequential decision making. These models require an agent to take actions based on observations (features) with the goal of achieving long-term performance competitive with a large set of policies. To avoid barriers to sample-efficient learning associated with large observation spaces and general POMDPs, we focus on problems that can be summarized by a small number of hidden states and have long-term rewards that are predictable by a reactive function class. In this setting, we design and analyze a new reinforcement learning algorithm, Least Squares Value Elimination by Exploration. We prove that the algorithm learns near optimal behavior after a number of episodes that is polynomial in all relevant parameters, logarithmic in the number of policies, and independent of the size of the observation space. Our result provides theoretical justification for reinforcement learning with function approximation.

****

Generative Shape Models: Joint Text Recognition and Segmentation with Very Little Training Data

Xinghua Lou, Ken Kansky, Wolfgang Lehrach, CC Laan, Bhaskara Marthi, D. Phoenix, Dileep George

We demonstrate that a generative model for object shapes can achieve state of the art results on challenging scene text recognition tasks, and with orders of magnitude fewer training images than required for competing discriminative methods. In addition to transcribing text from challenging images, our method performs fine-grained instance segmentation of characters. We show that our model is more robust to both affine transformations and non-affine deformations compared to previous approaches.

****

Probabilistic Linear Multistep Methods

Onur Teymur, Kostas Zygalakis, Ben Calderhead

We present a derivation and theoretical investigation of the Adams-Bashforth and Adams-Moulton family of linear multistep methods for solving ordinary differential equations, starting from a Gaussian process (GP) framework. In the limit, this formulation coincides with the classical deterministic methods, which have been used as higher-order initial value problem solvers for over a century. Furthermore, the natural probabilistic framework provided by the GP formulation allows us to derive probabilistic versions of these methods, in the spirit of a number of other probabilistic ODE solvers presented in the recent literature. In contrast to higher-order Runge-Kutta methods, which require multiple intermediate function evaluations per step, Adams family methods make use of previous function evaluations, so that increased accuracy arising from a higher-order multistep approach comes at very little additional computational cost. We show that through a careful choice of covariance function for the GP, the posterior mean and standard deviation over the numerical solution can be made to exactly coincide with the value given by the deterministic method and its local truncation error respectively. We provide a rigorous proof of the convergence of these new methods, as well as an empirical investigation (up to fifth order) demonstrating their convergence rates in practice.

****

Computational and Statistical Tradeoffs in Learning to Rank

Ashish Khetan, Sewoong Oh

For massive and heterogeneous modern data sets, it is of fundamental interest to provide guarantees on the accuracy of estimation when computational resources are limited. In the application of learning to rank, we provide a hierarchy of rank-breaking mechanisms ordered by the complexity in thus generated sketch of the data. This allows the number of data points collected to be gracefully traded off against computational resources available, while guaranteeing the desired level of accuracy. Theoretical guarantees on the proposed generalized rank-breaking implicitly provide such trade-offs, which can be explicitly characterized under certain canonical scenarios on the structure of the data.

****

Split LBI: An Iterative Regularization Path with Structural Sparsity

Chendi Huang, Xinwei Sun, Jiechao Xiong, Yuan Yao

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.

****

Incremental Variational Sparse Gaussian Process Regression

Ching-An Cheng, Byron Boots

Recent work on scaling up Gaussian process regression (GPR) to large datasets ha s primarily focused on sparse GPR, which leverages a small set of basis functions to approximate the full Gaussian process during inference. However, the majority of these approaches are batch methods that operate on the entire training dataset at once, precluding the use of datasets that are streaming or too large to fit into memory. Although previous work has considered incrementally solving variational sparse GPR, most algorithms fail to update the basis functions and therefore perform suboptimally. We propose a novel incremental learning algorithm for variational sparse GPR based on stochastic mirror ascent of probability densities in reproducing kernel Hilbert space. This new formulation allows our algorithm to update basis functions online in accordance with the manifold structure of probability densities for fast convergence. We conduct several experiments and show that our proposed approach achieves better empirical performance in terms of prediction error than the recent state-of-the-art incremental solutions to variational sparse GPR.

****

Sublinear Time Orthogonal Tensor Decomposition
Zhao Song, David Woodruff, Huan 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.

****

Mapping Estimation for Discrete Optimal Transport
Michaël Perrot, Nicolas Courty, Rémi Flamary, Amaury Habrard

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.

****

Greedy Feature Construction
Dino Oglic, Thomas Gärtner

We present an effective method for supervised feature construction. The main goal of the approach is to construct a feature representation for which a set of linear hypotheses is of sufficient capacity -- large enough to contain a satisfactory solution to the considered problem and small enough to allow good generalization from a small number of training examples. We achieve this goal with a greedy procedure that constructs features by empirically fitting squared error residuals. The proposed constructive procedure is consistent and can output a rich set of features. The effectiveness of the approach is evaluated empirically by fitting a linear ridge regression model in the constructed feature space and our empirical results indicate a superior performance of our approach over competing methods.

****

Dynamic Network Surgery for Efficient DNNs
Yiwen Guo, Anbang Yao, Yurong 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-authors prior to requesting a name change in the electronic proceedings.

****

Graph Clustering: Block-models and model free results
Yali Wan, Marina Meila

Clustering graphs under the Stochastic Block Model (SBM) and extensions are well studied. Guarantees of correctness exist under the assumption that the data is sampled from a model. In this paper, we propose a framework, in which we obtain "correctness" guarantees without assuming the data comes from a model. The guarantees we obtain depend instead on the statistics of the data that can be checked. We also show that this framework ties in with the existing model-based framework, and that we can exploit results in model-based recovery, as well as strengthen the results existing in that area of research.

****

CMA-ES with Optimal Covariance Update and Storage Complexity

Oswin Krause, Dídac Rodríguez Arbonès, Christian Igel

The covariance matrix adaptation evolution strategy (CMA-ES) is arguably one of the most powerful real-valued derivative-free optimization algorithms, finding many applications in machine learning. The CMA-ES is a Monte Carlo method, sampling from a sequence of multi-variate Gaussian distributions. Given the function values at the sampled points, updating and storing the covariance matrix dominates the time and space complexity in each iteration of the algorithm. We propose a numerically stable quadratic-time covariance matrix update scheme with minimal memory requirements based on maintaining triangular Cholesky factors. This requires a modification of the cumulative step-size adaption (CSA) mechanism in the CMA-ES, in which we replace the inverse of the square root of the covariance matrix by the inverse of the triangular Cholesky factor. Because the triangular Cholesky factor changes smoothly with the matrix square root, this modification does not change the behavior of the CMA-ES in terms of required objective function evaluations as verified empirically. Thus, the described algorithm can and should replace the standard CMA-ES if updating and storing the covariance matrix matters.

****

Feature selection in functional data classification with recursive maxima hunting

José L. Torrecilla, Alberto Suárez

Dimensionality reduction is one of the key issues in the design of effective machine learning methods for automatic induction. In this work, we introduce recursive maxima hunting (RMH) for variable selection in classification problems with functional data. In this context, variable selection techniques are especially attractive because they reduce the dimensionality, facilitate the interpretation and can improve the accuracy of the predictive models. The method, which is a recursive extension of maxima hunting (MH), performs variable selection by identifying the maxima of a relevance function, which measures the strength of the correlation of the predictor functional variable with the class label. At each stage, the information associated with the selected variable is removed by subtracting the conditional expectation of the process. The results of an extensive empirical evaluation are used to illustrate that, in the problems investigated, RMH has comparable or higher predictive accuracy than standard simensionality reduction techniques, such as PCA and PLS, and state-of-the-art feature selection methods for functional data, such as maxima hunting.

****

Cyclades: Conflict-free Asynchronous Machine Learning

Xinghao Pan, Maximilian Lam, Stephen Tu, Dimitris Papailiopoulos, Ce Zhang, Michael I. Jordan, Kannan Ramchandran, Christopher Ré

We present Cyclades, a general framework for parallelizing stochastic optimization algorithms in a shared memory setting. Cyclades is asynchronous during model updates, and requires no memory locking mechanisms, similar to Hogwild!-type algorithms. Unlike Hogwild!, Cyclades introduces no conflicts during parallel execution, and offers a black-box analysis for provable speedups across a large family of algorithms. Due to its inherent cache locality and conflict-free nature, our multi-core implementation of Cyclades consistently outperforms Hogwild!-type algorithms on sufficiently sparse datasets, leading to up to 40% speedup gains compared to Hogwild!, and up to 5\times gains over asynchronous implementations of variance reduction algorithms.

****

Proximal Stochastic Methods for Nonsmooth Nonconvex Finite-Sum Optimization

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

We analyze stochastic algorithms for optimizing nonconvex, nonsmooth finite-sum problems, where the nonsmooth part is convex. Surprisingly, unlike the smooth case, our knowledge of this fundamental problem is very limited. For example, it is not known whether the proximal stochastic gradient method with constant minibatch converges to a stationary point. To tackle this issue, we develop fast stoc

hastic algorithms that provably converge to a stationary point for constant minibatches. Furthermore, using a variant of these algorithms, we obtain provably faster convergence than batch proximal gradient descent. Our results are based on the recent variance reduction techniques for convex optimization but with a novel analysis for handling nonconvex and nonsmooth functions. We also prove global linear convergence rate for an interesting subclass of nonsmooth nonconvex functions, which subsumes several recent works.
****

Spectral Learning of Dynamic Systems from Nonequilibrium Data
Hao Wu, Frank Noe
Observable operator models (OOMs) and related models are one of the most important and powerful tools for modeling and analyzing stochastic systems. They exactly describe dynamics of finite-rank systems and can be efficiently and consistently estimated through spectral learning under the assumption of identically distributed data. In this paper, we investigate the properties of spectral learning without this assumption due to the requirements of analyzing large-time scale systems, and show that the equilibrium dynamics of a system can be extracted from nonequilibrium observation data by imposing an equilibrium constraint. In addition, we propose a binless extension of spectral learning for continuous data. In comparison with the other continuous-valued spectral algorithms, the binless algorithm can achieve consistent estimation of equilibrium dynamics with only linear complexity.
****

Dimension-Free Iteration Complexity of Finite Sum Optimization Problems
Yossi Arjevani, Ohad Shamir
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.
****

Hierarchical Object Representation for Open-Ended Object Category Learning and Recognition
Seyed Hamidreza Kasaei, Ana Maria Tomé, Luís Seabra Lopes
Most robots lack the ability to learn new objects from past experiences. To migrate a robot to a new environment one must often completely re-generate the knowledge- base that it is running with. Since in open-ended domains the set of categories to be learned is not predefined, it is not feasible to assume that one can pre-program all object categories required by robots. Therefore, autonomous robots must have the ability to continuously execute learning and recognition in a concurrent and interleaved fashion. This paper proposes an open-ended 3D object recognition system which concurrently learns both the object categories and the statistical features for encoding objects. In particular, we propose an extension of Latent Dirichlet Allocation to learn structural semantic features (i.e. topics) from low-level feature co-occurrences for each category independently. Moreover, topics in each category are discovered in an unsupervised fashion and are updated incrementally using new object views. The approach contains similarities with the organization of the visual cortex and builds a hierarchy of increasingly sophisticated representations. Results show the fulfilling performance of this approach on different types of objects. Moreover, this system demonstrates the capability of learning from few training examples and competes with state-of-the-art systems.
****

Active Learning with Oracle Epiphany
Tzu-Kuo Huang, Lihong Li, Ara Vartanian, Saleema Amershi, Jerry Zhu
We present a theoretical analysis of active learning with more realistic interactions with human oracles. Previous empirical studies have shown oracles abstaining on difficult queries until accumulating enough information to make label decisions. We formalize this phenomenon with an “oracle epiphany model” and analyze active learning query complexity under such oracles for both the realizable and the agnos- tic cases. Our analysis shows that active learning is possible with o

racle epiphany, but incurs an additional cost depending on when the epiphany happens. Our results suggest new, principled active learning approaches with realistic oracles.

****

Stochastic Optimization for Large-scale Optimal Transport

Aude Genevay, Marco Cuturi, Gabriel Peyré, Francis Bach

Optimal transport (OT) defines a powerful framework to compare probability distributions in a geometrically faithful way. However, the practical impact of OT is still limited because of its computational burden. We propose a new class of stochastic optimization algorithms to cope with large-scale problems routinely encountered in machine learning applications. These methods are able to manipulate arbitrary distributions (either discrete or continuous) by simply requiring to be able to draw samples from them, which is the typical setup in high-dimensional learning problems. This alleviates the need to discretize these densities, while giving access to provably convergent methods that output the correct distance without discretization error. These algorithms rely on two main ideas: (a) the dual OT problem can be re-cast as the maximization of an expectation; (b) entropic regularization of the primal OT problem results in a smooth dual optimization optimization which can be addressed with algorithms that have a provably faster convergence. We instantiate these ideas in three different computational setups: (i) when comparing a discrete distribution to another, we show that incremental stochastic optimization schemes can beat the current state of the art finite dimensional OT solver (Sinkhorn's algorithm) ; (ii) when comparing a discrete distribution to a continuous density, a re-formulation (semi-discrete) of the dual program is amenable to averaged stochastic gradient descent, leading to better performance than approximately solving the problem by discretization ; (iii) when dealing with two continuous densities, we propose a stochastic gradient descent over a reproducing kernel Hilbert space (RKHS). This is currently the only known method to solve this problem, and is more efficient than discretizing beforehand the two densities. We backup these claims on a set of discrete, semi-discrete and continuous benchmark problems.

****

The Sound of APALM Clapping: Faster Nonsmooth Nonconvex Optimization with Stochastic Asynchronous PALM

Damek Davis, Brent Edmunds, Madeleine Udell

We introduce the Stochastic Asynchronous Proximal Alternating Linearized Minimization (SAPALM) method, a block coordinate stochastic proximal-gradient method for solving nonconvex, nonsmooth optimization problems. SAPALM is the first asynchronous parallel optimization method that provably converges on a large class of nonconvex, nonsmooth problems. We prove that SAPALM matches the best known rates of convergence --- among synchronous or asynchronous methods --- on this problem class. We provide upper bounds on the number of workers for which we can expect to see a linear speedup, which match the best bounds known for less complex problems, and show that in practice SAPALM achieves this linear speedup. We demonstrate state-of-the-art performance on several matrix factorization problems.

****

Coresets for Scalable Bayesian Logistic Regression

Jonathan Huggins, Trevor Campbell, Tamara Broderick

The use of Bayesian methods in large-scale data settings is attractive because of the rich hierarchical models, uncertainty quantification, and prior specification they provide. Standard Bayesian inference algorithms are computationally expensive, however, making their direct application to large datasets difficult or infeasible. Recent work on scaling Bayesian inference has focused on modifying the underlying algorithms to, for example, use only a random data subsample at each iteration. We leverage the insight that data is often redundant to instead obtain a weighted subset of the data (called a coreset) that is much smaller than the original dataset. We can then use this small coreset in any number of existing posterior inference algorithms without modification. In this paper, we develop an efficient coreset construction algorithm for Bayesian logistic regression models. We provide theoretical guarantees on the size and approximation quality o f the coreset -- both for fixed, known datasets, and in expectation for a wide class of data generative models. Crucially, the proposed approach also permits efficient construction of the coreset in both streaming and parallel settings, with minimal additional effort. We demonstrate the efficacy of our approach on a number of synthetic and real-world datasets, and find that, in practice, the size of the coreset is independent of the original dataset size. Furthermore, constructing the coreset takes a negligible amount of time compared to that required to run MCMC on it.

****

Sorting out typicality with the inverse moment matrix SOS polynomial

Edouard Pauwels, Jean B. Lasserre

We study a surprising phenomenon related to the representation of a cloud of data points using polynomials. We start with the previously unnoticed empirical observation that, given a collection (a cloud) of data points, the sublevel sets of a certain distinguished polynomial capture the shape of the cloud very accurately. This distinguished polynomial is a sum-of-squares (SOS) derived in a simple manner from the inverse of the empirical moment matrix. In fact, this SOS polynomial is directly related to orthogonal polynomials and the Christoffel function. This allows to generalize and interpret extremality properties of orthogonal polynomials and to provide a mathematical rationale for the observed phenomenon. Among diverse potential applications, we illustrate the relevance of our results on a network intrusion detection task for which we obtain performances similar to existing dedicated methods reported in the literature.

****

The Multi-fidelity Multi-armed Bandit

Kirthevasan Kandasamy, Gautam Dasarathy, Barnabas Poczos, Jeff Schneider

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.

****

Joint Line Segmentation and Transcription for End-to-End Handwritten Paragraph Recognition

Theodore Bluche

Offline handwriting recognition systems require cropped text line images for both training and recognition. On the one hand, the annotation of position and transcript at line level is costly to obtain. On the other hand, automatic line segmentation algorithms are prone to errors, compromising the subsequent recognition. In this paper, we propose a modification of the popular and efficient Multi-Dimensional Long Short-Term Memory Recurrent Neural Networks (MDLSTM-RNNs) to enable end-to-end processing of handwritten paragraphs. More particularly, we replace the collapse layer transforming the two-dimensional representation into a sequence of predictions by a recurrent version which can select one line at a time. In the proposed model, a neural network performs a kind of implicit line segmentation by computing attention weights on the image representation. The experiments on paragraphs of Rimes and IAM databases yield results that are competitive with those of networks trained at line level, and constitute a significant step towards end-to-end transcription of full documents.

****

k*-Nearest Neighbors: From Global to Local

Oren Anava, Kfir Levy

The weighted k-nearest neighbors algorithm is one of the most fundamental non-parametric methods in pattern recognition and machine learning. The question of setting the optimal number of neighbors as well as the optimal weights has received much attention throughout the years, nevertheless this problem seems to have remained unsettled. In this paper we offer a simple approach to locally weighted regression/classification, where we make the bias-variance tradeoff explicit. Our formulation enables us to phrase a notion of optimal weights, and to efficiently find these weights as well as the optimal number of neighbors efficiently and adaptively, for each data point whose value we wish to estimate. The appli cability of our approach is demonstrated on several datasets, showing superior performance over standard locally weighted methods.

****

Protein contact prediction from amino acid co-evolution using convolutional networks for graph-valued images

Vladimir Golkov, Marcin J. Skwark, Antonij Golkov, Alexey Dosovitskiy, Thomas Brox, Jens Meiler, Daniel Cremers

Proteins are the "building blocks of life", the most abundant organic molecules, and the central focus of most areas of biomedicine. Protein structure is strongly related to protein function, thus structure prediction is a crucial task on the way to solve many biological questions. A contact map is a compact representation of the three-dimensional structure of a protein via the pairwise contacts between the amino acid constituting the protein. We use a convolutional network to calculate protein contact maps from inferred statistical coupling between positions in the protein sequence. The input to the network has an image-like structure amenable to convolutions, but every "pixel" instead of color channels contains a bipartite undirected edge-weighted graph. We propose several methods for treating such "graph-valued images" in a convolutional network. The proposed method outperforms state-of-the-art methods by a large margin. It also allows for a great flexibility with regard to the input data, which makes it useful for studying a wide range of problems.

****

Learnable Visual Markers

Oleg Grinchuk, Vadim Lebedev, Victor Lempitsky

We propose a new approach to designing visual markers (analogous to QR-codes, markers for augmented reality, and robotic fiducial tags) based on the advances in deep generative networks. In our approach, the markers are obtained as color images synthesized by a deep network from input bit strings, whereas another deep network is trained to recover the bit strings back from the photos of these markers. The two networks are trained simultaneously in a joint backpropagation process that takes characteristic photometric and geometric distortions associated with marker fabrication and capture into account. Additionally, a stylization loss based on statistics of activations in a pretrained classification network can be inserted into the learning in order to shift the marker appearance towards some texture prototype. In the experiments, we demonstrate that the markers obtained using our approach are capable of retaining bit strings that are long enough to be practical. The ability to automatically adapt markers according to the usage scenario and the desired capacity as well as the ability to combine information encoding with artistic stylization are the unique properties of our approach. As a byproduct, our approach provides an insight on the structure of patterns that are most suitable for recognition by ConvNets and on their ability to distinguish composite patterns.

****

Finite-Sample Analysis of Fixed-k Nearest Neighbor Density Functional Estimators

Shashank Singh, Barnabas Poczos

We provide finite-sample analysis of a general framework for using k-nearest neighbor statistics to estimate functionals of a nonparametric continuous probability density, including entropies and divergences. Rather than plugging a consistent density estimate (which requires $k \to \infty$ as the sample size $n \to \infty$) into the functional of interest, the estimators we consider fix k and perform a bias correction. This can be more efficient computationally, and, as we show, statistically, leading to faster convergence rates. Our framework unifies several previous estimators, for most of which ours are the first finite sample guarantees.

****

Maximizing Influence in an Ising Network: A Mean-Field Optimal Solution

Christopher Lynn, Daniel D. Lee

Influence maximization in social networks has typically been studied in the context of contagion models and irreversible processes. In this paper, we consider an alternate model that treats individual opinions as spins in an Ising system at dynamic equilibrium. We formalize the \textit{Ising influence maximization} pro blem, which has a natural physical interpretation as maximizing the magnetization given a budget of external magnetic field. Under the mean-field (MF) approximation, we present a gradient ascent algorithm that uses the susceptibility to efficiently calculate local maxima of the magnetization, and we develop a number of sufficient conditions for when the MF magnetization is concave and our algorithm converges to a global optimum. We apply our algorithm on random and real-world networks, demonstrating, remarkably, that the MF optimal external fields (i.e., the external fields which maximize the MF magnetization) exhibit a phase transition from focusing on high-degree individuals at high temperatures to focusing on low-degree individuals at low temperatures. We also establish a number of novel results about the structure of steady-states in the ferromagnetic MF Ising model on general graphs, which are of independent interest.

****

Regret Bounds for Non-decomposable Metrics with Missing Labels
Nagarajan Natarajan, Prateek Jain

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 Concentration Inequalities for Sequential Decision Problems
Shengjia Zhao, Enze Zhou, Ashish Sabharwal, Stefano Ermon

A key challenge in sequential decision problems is to determine how many samples are needed for an agent to make reliable decisions with good probabilistic guarantees.   We introduce Hoeffding-like concentration inequalities that hold for a random, adaptively chosen number of samples. Our inequalities are tight under natural assumptions and can greatly simplify the analysis of common sequential decision problems. In particular, we apply them to sequential hypothesis testing, best arm identification, and sorting. The resulting algorithms rival or exceed the state of the art both theoretically and empirically.

****

Refined Lower Bounds for Adversarial Bandits
Sébastien Gerchinovitz, Tor Lattimore

We provide new lower bounds on the regret that must be suffered by adversarial bandit algorithms. The new results show that recent upper bounds that either (a) hold with high-probability or (b) depend on the total loss of the best arm or (c) depend on the quadratic variation of the losses, are close to tight. Besides this we prove two impossibility results. First, the existence of a single arm that is optimal in every round cannot improve the regret in the worst case. Second, the regret cannot scale with the effective range of the losses. In contrast, both results are possible in the full-information setting.

****

Structure-Blind Signal Recovery
Dmitry Ostrovsky, Zaid Harchaoui, Anatoli Juditsky, Arkadi S. Nemirovski

We consider the problem of recovering a signal observed in Gaussian noise. If the set of signals is convex and compact, and can be specified beforehand, one can use classical linear estimators that achieve a risk within a constant factor of the minimax risk. However, when the set is unspecified, designing an estimator that is blind to the hidden structure of the signal remains a challenging problem. We propose a new family of estimators to recover signals observed in Gaussian noise. Instead of specifying the set where the signal lives, we assume the existence of a well-performing linear estimator. Proposed estimators enjoy exact oracle inequalities and can be efficiently computed through convex optimization. We present several numerical illustrations that show the potential of the approach.

****

Reward Augmented Maximum Likelihood for Neural Structured Prediction
Mohammad Norouzi, Samy Bengio, zhifeng Chen, Navdeep Jaitly, Mike Schuster, Yonghui Wu, Dale Schuurmans

A key problem in structured output prediction is enabling direct optimization of the task reward function that matters for test evaluation. This paper presents a simple and computationally efficient method that incorporates task reward into maximum likelihood training. We establish a connection between maximum likelihood and regularized expected reward, showing that they are approximately equivalent in the vicinity of the optimal solution. Then we show how maximum likelihood can be generalized by optimizing the conditional probability of auxiliary outputs that are sampled proportional to their exponentiated scaled rewards. We apply this framework to optimize edit distance in the output space, by sampling from edited targets. Experiments on speech recognition and machine translation for neural sequence to sequence models show notable improvements over maximum likelihood baseline by simply sampling from target output augmentations.

****

Regularization With Stochastic Transformations and Perturbations for Deep Semi-Supervised Learning

Mehdi Sajjadi, Mehran Javanmardi, Tolga Tasdizen

Effective convolutional neural networks are trained on large sets of labeled data. However, creating large labeled datasets is a very costly and time-consuming task. Semi-supervised learning uses unlabeled data to train a model with higher accuracy when there is a limited set of labeled data available. In this paper, we consider the problem of semi-supervised learning with convolutional neural networks. Techniques such as randomized data augmentation, dropout and random max-pooling provide better generalization and stability for classifiers that are trained using gradient descent. Multiple passes of an individual sample through the network might lead to different predictions due to the non-deterministic behavior of these techniques. We propose an unsupervised loss function that takes advantage of the stochastic nature of these methods and minimizes the difference between the predictions of multiple passes of a training sample through the network. We evaluate the proposed method on several benchmark datasets.

****

An Online Sequence-to-Sequence Model Using Partial Conditioning

Navdeep Jaitly, Quoc V. Le, Oriol Vinyals, Ilya Sutskever, David Sussillo, Samy Bengio

Sequence-to-sequence models have achieved impressive results on various tasks. However, they are unsuitable for tasks that require incremental predictions to be made as more data arrives or tasks that have long input sequences and output sequences. This is because they generate an output sequence conditioned on an entire input sequence. In this paper, we present a Neural Transducer that can make incremental predictions as more input arrives, without redoing the entire computation. Unlike sequence-to-sequence models, the Neural Transducer computes the next-step distribution conditioned on the partially observed input sequence and the partially generated sequence. At each time step, the transducer can decide to emit zero to many output symbols. The data can be processed using an encoder and presented as input to the transducer. The discrete decision to emit a symbol at every time step makes it difficult to learn with conventional backpropagation. It is however possible to train the transducer by using a dynamic programming algorithm to generate target discrete decisions. Our experiments show that the Neural Transducer works well in settings where it is required to produce output predictions as data come in. We also find that the Neural Transducer performs well for long sequences even when attention mechanisms are not used.

****

Interaction Networks for Learning about Objects, Relations and Physics

Peter Battaglia, Razvan Pascanu, Matthew Lai, Danilo Jimenez Rezende, koray kavukcuoglu

Reasoning about objects, relations, and physics is central to human intelligence, and a key goal of artificial intelligence. Here we introduce the interaction network, a model which can reason about how objects in complex systems interact, supporting dynamical predictions, as well as inferences about the abstract properties of the system. Our model takes graphs as input, performs object- and relation-centric reasoning in a way that is analogous to a simulation, and is implemented using deep neural networks. We evaluate its ability to reason about several challenging physical domains: n-body problems, rigid-body collision, and non-rigid dynamics. Our results show it can be trained to accurately simulate the physical trajectories of dozens of objects over thousands of time steps, estimate abstract quantities such as energy, and generalize automatically to systems with different numbers and configurations of objects and relations. Our interaction network implementation is the first general-purpose, learnable physics engine, and a powerful general framework for reasoning about object and relations in a wide variety of complex real-world domains.

****

Bayesian optimization under mixed constraints with a slack-variable augmented Lagrangian

Victor Picheny, Robert B. Gramacy, Stefan Wild, Sebastien Le Digabel

An augmented Lagrangian (AL) can convert a constrained optimization problem into a sequence of simpler (e.g., unconstrained) problems which are then usually solved with local solvers. Recently, surrogate-based Bayesian optimization (BO) sub-solvers have been successfully deployed in the AL framework for a more global search in the presence of inequality constraints; however a drawback was that expected improvement (EI) evaluations relied on Monte Carlo. Here we introduce an alternative slack variable AL, and show that in this formulation the EI may be evaluated with library routines. The slack variables furthermore facilitate equality as well as inequality constraints, and mixtures thereof. We show our new slack "ALBO" compares favorably to the original. Its superiority over conventional alternatives is reinforced on several new mixed constraint examples.

****

Combinatorial Energy Learning for Image Segmentation

Jeremy B. Maitin-Shepard, Viren Jain, Michal Januszewski, Peter Li, Pieter Abbeel

We introduce a new machine learning approach for image segmentation that uses a neural network to model the conditional energy of a segmentation given an image. Our approach, combinatorial energy learning for image segmentation (CELIS) places a particular emphasis on modeling the inherent combinatorial nature of dense image segmentation problems. We propose efficient algorithms for learning deep neural networks to model the energy function, and for local optimization of this energy in the space of supervoxel agglomerations. We extensively evaluate our method on a publicly available 3-D microscopy dataset with 25 billion voxels of ground truth data. On an 11 billion voxel test set, we find that our method improves volumetric reconstruction accuracy by more than 20% as compared to two state-of-the-art baseline methods: graph-based segmentation of the output of a 3-D convolutional neural network trained to predict boundaries, as well as a random forest classifier trained to agglomerate supervoxels that were generated by a 3-D convolutional neural network.

****

Bayesian Optimization for Probabilistic Programs

Tom Rainforth, Tuan Anh Le, Jan-Willem van de Meent, Michael A. Osborne, Frank Wood

We present the first general purpose framework for marginal maximum a posteriori estimation of probabilistic program variables. By using a series of code transformations, the evidence of any probabilistic program, and therefore of any graphical model, can be optimized with respect to an arbitrary subset of its sampled variables. To carry out this optimization, we develop the first Bayesian optimization package to directly exploit the source code of its target, leading to innovations in problem-independent hyperpriors, unbounded optimization, and implicit constraint satisfaction; delivering significant performance improvements over prominent existing packages. We present applications of our method to a number of tasks including engineering design and parameter optimization.

****

Coin Betting and Parameter-Free Online Learning

Francesco Orabona, David Pal

In the recent years, a number of parameter-free algorithms have been developed for online linear optimization over Hilbert spaces and for learning with expert a dvice. These algorithms achieve optimal regret bounds that depend on the unknown competitors, without having to tune the learning rates with oracle choices. We present a new intuitive framework to design parameter-free algorithms for both online linear optimization over Hilbert spaces and for learning with expert advice, based on reductions to betting on outcomes of adversarial coins. We instantiate it using a betting algorithm based on the Krichevsky-Trofimov estimator. The resulting algorithms are simple, with no parameters to be tuned, and they improve or match previous results in terms of regret guarantee and per-round complexity.

****

Learning Deep Embeddings with Histogram Loss

Evgeniya Ustinova, Victor Lempitsky

We suggest a new loss for learning deep embeddings. The key characteristics of the new loss is the absence of tunable parameters and very good results obtained across a range of datasets and problems. The loss is computed by estimating two distribution of similarities for positive (matching) and negative (non-matching) point pairs, and then computing the probability of a positive pair to have a lower similarity score than a negative pair based on these probability estimates. We show that these operations can be performed in a simple and piecewise-differentiable manner using 1D histograms with soft assignment operations. This makes the proposed loss suitable for learning deep embeddings using stochastic optimization. The experiments reveal favourable results compared to recently proposed loss functions.

****

An Efficient Streaming Algorithm for the Submodular Cover Problem

Ashkan Norouzi-Fard, Abbas Bazzi, Ilija Bogunovic, Marwa El Halabi, Ya-Ping Hsieh, Volkan Cevher

We initiate the study of the classical Submodular Cover (SC) problem in the data streaming model which we refer to as the Streaming Submodular Cover (SSC). We show that any single pass streaming algorithm using sublinear memory in the size of the stream will fail to provide any non-trivial approximation guarantees for SSC. Hence, we consider a relaxed version of SSC, where we only seek to find a partial cover. We design the first Efficient bicriteria Submodular Cover Streaming (ESC-Streaming) algorithm for this problem, and provide theoretical guarantees for its performance supported by numerical evidence. Our algorithm finds solutions that are competitive with the near-optimal offline greedy algorithm despite requiring only a single pass over the data stream. In our numerical experiments, we evaluate the performance of ESC-Streaming on active set selection and large-scale graph cover problems.

****

Fundamental Limits of Budget-Fidelity Trade-off in Label Crowdsourcing

Farshad Lahouti, Babak Hassibi

Digital crowdsourcing (CS) is a modern approach to perform certain large projects using small contributions of a large crowd. In CS, a taskmaster typically breaks down the project into small batches of tasks and assigns them to so-called workers with imperfect skill levels. The crowdsourcer then collects and analyzes the results for inference and serving the purpose of the project. In this work, the CS problem, as a human-in-the-loop computation problem, is modeled and analyzed in an information theoretic rate-distortion framework. The purpose is to identify the ultimate fidelity that one can achieve by any form of query from the crowd and any decoding (inference) algorithm with a given budget. The results are established by a joint source channel (de)coding scheme, which represent the query scheme and inference, over parallel noisy channels, which model workers with imperfect skill levels. We also present and analyze a query scheme dubbed k-ary incidence coding and study optimized query pricing in this setting.

****

Beyond Exchangeability: The Chinese Voting Process

Moontae Lee, Seok Hyun Jin, David Mimno

Many online communities present user-contributed responses, such as reviews of products and answers to questions. User-provided helpfulness votes can highlight the most useful responses, but voting is a social process that can gain momentum based on the popularity of responses and the polarity of existing votes. We propose the Chinese Voting Process (CVP) which models the evolution of helpfulness votes as a self-reinforcing process dependent on position and presentation biases. We evaluate this model on Amazon product reviews and more than 80 StackExchange forums, measuring the intrinsic quality of individual responses and behavioral coefficients of different communities.

****

Robust Spectral Detection of Global Structures in the Data by Learning a Regularization

Pan Zhang

Spectral methods are popular in detecting global structures in the given data that can be represented as a matrix. However when the data matrix is sparse or noisy, classic spectral methods usually fail to work, due to localization of eigenvectors (or singular vectors) induced by the sparsity or noise. In this work, we propose a general method to solve the localization problem by learning a regularization matrix from the localized eigenvectors. Using matrix perturbation analysis, we demonstrate that the learned regularizations suppress down the eigenvalues associated with localized eigenvectors and enable us to recover the informative eigenvectors representing the global structure. We show applications of our method in several inference problems: community detection in networks, clustering from pairwise similarities, rank estimation and matrix completion problems. Using extensive experiments, we illustrate that our method solves the localization problem and works down to the theoretical detectability limits in different kinds of synthetic data. This is in contrast with existing spectral algorithms based on data matrix, non-backtracking matrix, Laplacians and those with rank-one regularizations, which perform poorly in the sparse case with noise.

****

Optimal spectral transportation with application to music transcription

Rémi Flamary, Cédric Févotte, Nicolas Courty, Valentin Emiya

Many spectral unmixing methods rely on the non-negative decomposition of spectral data onto a dictionary of spectral templates. In particular, state-of-the-art music transcription systems decompose the spectrogram of the input signal onto a dictionary of representative note spectra. The typical measures of fit used to quantify the adequacy of the decomposition compare the data and template entries frequency-wise. As such, small displacements of energy from a frequency bin to another as well as variations of timber can disproportionally harm the fit. We address these issues by means of optimal transportation and propose a new measure of fit that treats the frequency distributions of energy holistically as opposed to frequency-wise. Building on the harmonic nature of sound, the new measure is invariant to shifts of energy to harmonically-related frequencies, as well as to small and local displacements of energy. Equipped with this new measure of fit, the dictionary of note templates can be considerably simplified to a set of Dirac vectors located at the target fundamental frequencies (musical pitch values). This in turns gives ground to a very fast and simple decomposition algorithm that achieves state-of-the-art performance on real musical data.

****

MoCap-guided Data Augmentation for 3D Pose Estimation in the Wild

Gregory Rogez, Cordelia Schmid

This paper addresses the problem of 3D human pose estimation in the wild. A significant challenge is the lack of training data, i.e., 2D images of humans annotated with 3D poses. Such data is necessary to train state-of-the-art CNN architectures. Here, we propose a solution to generate a large set of photorealistic synthetic images of humans with 3D pose annotations. We introduce an image-based synthesis engine that artificially augments a dataset of real images with 2D human pose annotations using 3D Motion Capture (MoCap) data. Given a candidate 3D pose our algorithm selects for each joint an image whose 2D pose locally matches the projected 3D pose. The selected images are then combined to generate a new synthetic image by stitching local image patches in a kinematically constrained manner. The resulting images are used to train an end-to-end CNN for full-body 3D p ose estimation. We cluster the training data into a large number of pose classes and tackle pose estimation as a K-way classification problem. Such an approach is viable only with large training sets such as ours. Our method outperforms the state of the art in terms of 3D pose estimation in controlled environments (Human3.6M) and shows promising results for in-the-wild images (LSP). This demonstrates that CNNs trained on artificial images generalize well to real images.

****

A Constant-Factor Bi-Criteria Approximation Guarantee for k-means++
Dennis Wei

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.

****

CNNpack: Packing Convolutional Neural Networks in the Frequency Domain
Yunhe Wang, Chang Xu, Shan You, Dacheng Tao, Chao Xu

Deep convolutional neural networks (CNNs) are successfully used in a number of applications. However, their storage and computational requirements have largely prevented their widespread use on mobile devices. Here we present an effective CNN compression approach in the frequency domain, which focuses not only on smaller weights but on all the weights and their underlying connections. By treating convolutional filters as images, we decompose their representations in the frequency domain as common parts (i.e., cluster centers) shared by other similar filters and their individual private parts (i.e., individual residuals). A large number of low-energy frequency coefficients in both parts can be discarded to produce high compression without significantly compromising accuracy. We relax the computational burden of convolution operations in CNNs by linearly combining the convolution responses of discrete cosine transform (DCT) bases. The compression and speed-up ratios of the proposed algorithm are thoroughly analyzed and evaluated on benchmark image datasets to demonstrate its superiority over state-of-the-art methods.

****

Feature-distributed sparse regression: a screen-and-clean approach
Jiyan Yang, Michael W. Mahoney, Michael Saunders, Yuekai Sun

Most existing approaches to distributed sparse regression assume the data is partitioned by samples. However, for high-dimensional data (D >> N), it is more natural to partition the data by features. We propose an algorithm to distributed sparse regression when the data is partitioned by features rather than samples. Our approach allows the user to tailor our general method to various distributed computing platforms by trading-off the total amount of data (in bits) sent over the communication network and the number of rounds of communication. We show that an implementation of our approach is capable of solving L1-regularized L2 regression problems with millions of features in minutes.

****

Generating Images with Perceptual Similarity Metrics based on Deep Networks
Alexey Dosovitskiy, Thomas Brox

We propose a class of loss functions, which we call deep perceptual similarity metrics (DeePSiM), allowing to generate sharp high resolution images from compressed abstract representations. Instead of computing distances in the image space, we compute distances between image features extracted by deep neural networks. This metric reflects perceptual similarity of images much better and, thus, leads to better results. We demonstrate two examples of use cases of the proposed loss: (1) networks that invert the AlexNet convolutional network; (2) a modified version of a variational autoencoder that generates realistic high-resolution random images.

****

Residual Networks Behave Like Ensembles of Relatively Shallow Networks
Andreas Veit, Michael J. Wilber, Serge Belongie

In this work we propose a novel interpretation of residual networks showing that they can be seen as a collection of many paths of differing length. Moreover, r esidual networks seem to enable very deep networks by leveraging only the short paths during training. To support this observation, we rewrite residual networks as an explicit collection of paths. Unlike traditional models, paths through residual networks vary in length. Further, a lesion study reveals that these paths show ensemble-like behavior in the sense that they do not strongly depend on each other. Finally, and most surprising, most paths are shorter than one might expect, and only the short paths are needed during training, as longer paths do not contribute any gradient. For example, most of the gradient in a residual network with 110 layers comes from paths that are only 10-34 layers deep. Our results reveal one of the key characteristics that seem to enable the training of very deep networks: Residual networks avoid the vanishing gradient problem by introducing short paths which can carry gradient throughout the extent of very deep networks.

****

Low-Rank Regression with Tensor Responses

Guillaume Rabusseau, Hachem Kadri

This paper proposes an efficient algorithm (HOLRR) to handle regression tasks where the outputs have a tensor structure. We formulate the regression problem as the minimization of a least square criterion under a multilinear rank constraint, a difficult non convex problem. HOLRR computes efficiently an approximate solution of this problem, with solid theoretical guarantees. A kernel extension is also presented. Experiments on synthetic and real data show that HOLRR computes accurate solutions while being computationally very competitive.

****

Provable Efficient Online Matrix Completion via Non-convex Stochastic Gradient Descent

Chi Jin, Sham M. Kakade, Praneeth Netrapalli

Matrix completion, where we wish to recover a low rank matrix by observing a few entries from it, is a widely studied problem in both theory and practice with wide applications. Most of the provable algorithms so far on this problem have been restricted to the offline setting where they provide an estimate of the unknown matrix using all observations simultaneously. However, in many applications, the online version, where we observe one entry at a time and dynamically update our estimate, is more appealing. While existing algorithms are efficient for the offline setting, they could be highly inefficient for the online setting. In this paper, we propose the first provable, efficient online algorithm for matrix completion. Our algorithm starts from an initial estimate of the matrix and then performs non-convex stochastic gradient descent (SGD). After every observation, it performs a fast update involving only one row of two tall matrices, giving near linear total runtime. Our algorithm can be naturally used in the offline setting as well, where it gives competitive sample complexity and runtime to state of the art algorithms. Our proofs introduce a general framework to show that SGD updates tend to stay away from saddle surfaces and could be of broader interests to other non-convex problems.

****

Local Maxima in the Likelihood of Gaussian Mixture Models: Structural Results and Algorithmic Consequences

Chi Jin, Yuchen Zhang, Sivaraman Balakrishnan, Martin J. Wainwright, Michael I. Jordan

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.

****

Diffusion-Convolutional Neural Networks

James Atwood, Don Towsley

We present diffusion-convolutional neural networks (DCNNs), a new model for graph-structured data. Through the introduction of a diffusion-convolution operation, we show how diffusion-based representations can be learned from graph-structured data and used as an effective basis for node classification. DCNNs have seve ral attractive qualities, including a latent representation for graphical data that is invariant under isomorphism, as well as polynomial-time prediction and learning that can be represented as tensor operations and efficiently implemented on a GPU. Through several experiments with real structured datasets, we demonstrate that DCNNs are able to outperform probabilistic relational models and kernel-on-graph methods at relational node classification tasks.

****

Completely random measures for modelling block-structured sparse networks

Tue Herlau, Mikkel N. Schmidt, Morten Mørup

Statistical methods for network data often parameterize the edge-probability by attributing latent traits such as block structure to the vertices and assume exchangeability in the sense of the Aldous-Hoover representation theorem. These assumptions are however incompatible with traits found in real-world networks such as a power-law degree-distribution. Recently, Caron & Fox (2014) proposed the use of a different notion of exchangeability after Kallenberg (2005) and obtained a network model which permits edge-inhomogeneity, such as a power-law degree-distribution whilst retaining desirable statistical properties. However, this model does not capture latent vertex traits such as block-structure. In this work we re-introduce the use of block-structure for network models obeying Kallenberg’s notion of exchangeability and thereby obtain a collapsed model which both admits the inference of block-structure and edge inhomogeneity. We derive a simple expression for the likelihood and an efficient sampling method. The obtained model is not significantly more difficult to implement than existing approaches to block-modelling and performs well on real network datasets.

****

Pruning Random Forests for Prediction on a Budget

Feng Nan, Joseph Wang, Venkatesh Saligrama

We propose to prune a random forest (RF) for resource-constrained prediction. We first construct a RF and then prune it to optimize expected feature cost & accuracy. We pose pruning RFs as a novel 0-1 integer program with linear constraints that encourages feature re-use. We establish total unimodularity of the constraint set to prove that the corresponding LP relaxation solves the original integer program. We then exploit connections to combinatorial optimization and develop an efficient primal-dual algorithm, scalable to large datasets. In contrast to our bottom-up approach, which benefits from good RF initialization, conventional methods are top-down acquiring features based on their utility value and is generally intractable, requiring heuristics. Empirically, our pruning algorithm outperforms existing state-of-the-art resource-constrained algorithms.

****

Synthesis of MCMC and Belief Propagation

Sung-Soo Ahn, Michael Chertkov, Jinwoo Shin

Markov Chain Monte Carlo (MCMC) and Belief Propagation (BP) are the most popular algorithms for computational inference in Graphical Models (GM). In principle, MCMC is an exact probabilistic method which, however, often suffers from exponentially slow mixing. In contrast, BP is a deterministic method, which is typically fast, empirically very successful, however in general lacking control of accuracy over loopy graphs. In this paper, we introduce MCMC algorithms correcting the approximation error of BP, i.e., we provide a way to compensate for BP errors via a consecutive BP-aware MCMC. Our framework is based on the Loop Calculus (LC) approach which allows to express the BP error as a sum of weighted generalized loops. Although the full series is computationally intractable, it is known that a truncated series, summing up all 2-regular loops, is computable in polynomial-time for planar pair-wise binary GMs and it also provides a highly accurate approximation empirically. Motivated by this, we, first, propose a polynomial-time approximation MCMC scheme for the truncated series of general (non-planar) pair-wise binary models. Our main idea here is to use the Worm algorithm, known to provide fast mixing in other (related) problems, and then design an appropriate rejection scheme to sample 2-regular loops. Furthermore, we also design an efficient rejection-free MCMC scheme for approximating the full series. The main novelty underlying our design is in utilizing the concept of cycle basi

s, which provides an efficient decomposition of the generalized loops. In essence, the proposed MCMC schemes run on transformed GM built upon the non-trivial BP solution, and our experiments show that this synthesis of BP and MCMC outperforms both direct MCMC and bare BP schemes.
****

Neurons Equipped with Intrinsic Plasticity Learn Stimulus Intensity Statistics
Travis Monk, Cristina Savin, Jörg Lücke

Experience constantly shapes neural circuits through a variety of plasticity mechanisms. While the functional roles of some plasticity mechanisms are well-understood, it remains unclear how changes in neural excitability contribute to learning. Here, we develop a normative interpretation of intrinsic plasticity (IP) as a key component of unsupervised learning. We introduce a novel generative mixture model that accounts for the class-specific statistics of stimulus intensities, and we derive a neural circuit that learns the input classes and their intensities. We will analytically show that inference and learning for our generative model can be achieved by a neural circuit with intensity-sensitive neurons equipped with a specific form of IP. Numerical experiments verify our analytical derivations and show robust behavior for artificial and natural stimuli. Our results link IP to non-trivial input statistics, in particular the statistics of stimulus intensities for classes to which a neuron is sensitive. More generally, our work paves the way toward new classification algorithms that are robust to intensity variations.
****

Disease Trajectory Maps
Peter Schulam, Raman Arora

Medical researchers are coming to appreciate that many diseases are in fact complex, heterogeneous syndromes composed of subpopulations that express different variants of a related complication. Longitudinal data extracted from individual electronic health records (EHR) offer an exciting new way to study subtle differences in the way these diseases progress over time. In this paper, we focus on answering two questions that can be asked using these databases of longitudinal EHR data. First, we want to understand whether there are individuals with similar disease trajectories and whether there are a small number of degrees of freedom that account for differences in trajectories across the population. Second, we want to understand how important clinical outcomes are associated with disease trajectories. To answer these questions, we propose the Disease Trajectory Map (DTM), a novel probabilistic model that learns low-dimensional representations of sparse and irregularly sampled longitudinal data. We propose a stochastic variational inference algorithm for learning the DTM that allows the model to scale to large modern medical datasets. To demonstrate the DTM, we analyze data collected on patients with the complex autoimmune disease, scleroderma. We find that DTM learns meaningful representations of disease trajectories and that the representations are significantly associated with important clinical outcomes.
****

Bayesian optimization for automated model selection
Gustavo Malkomes, Charles Schaff, Roman Garnett

Despite the success of kernel-based nonparametric methods, kernel selection still requires considerable expertise, and is often described as a “black art.” We present a sophisticated method for automatically searching for an appropriate kernel from an infinite space of potential choices. Previous efforts in this direction have focused on traversing a kernel grammar, only examining the data via computation of marginal likelihood. Our proposed search method is based on Bayesian optimization in model space, where we reason about model evidence as a function to be maximized. We explicitly reason about the data distribution and how it induces similarity between potential model choices in terms of the explanations they can offer for observed data. In this light, we construct a novel kernel between models to explain a given dataset. Our method is capable of finding a model that explains a given dataset well without any human assistance, often with fewer computations of model evidence than previous approaches, a claim we demonstrate empirically.

****

Designing smoothing functions for improved worst-case competitive ratio in online optimization

Reza Eghbali, Maryam Fazel

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.

****

Towards Unifying Hamiltonian Monte Carlo and Slice Sampling

Yizhe Zhang, Xiangyu Wang, Changyou Chen, Ricardo Henao, Kai Fan, Lawrence Carin

We unify slice sampling and Hamiltonian Monte Carlo (HMC) sampling, demonstrating their connection via the Hamiltonian-Jacobi equation from Hamiltonian mechanics. This insight enables extension of HMC and slice sampling to a broader family of samplers, called Monomial Gamma Samplers (MGS). We provide a theoretical analysis of the mixing performance of such samplers, proving that in the limit of a single parameter, the MGS draws decorrelated samples from the desired target distribution. We further show that as this parameter tends toward this limit, performance gains are achieved at a cost of increasing numerical difficulty and some practical convergence issues. Our theoretical results are validated with synthetic data and real-world applications.

****

Multi-step learning and underlying structure in statistical models

Maia Fraser

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 non-convex Burer-Monteiro approach works on smooth semidefinite programs

Nicolas Boumal, Vlad Voroninski, Afonso Bandeira

Semidefinite programs (SDP's) can be solved in polynomial time by interior point methods, but scalability can be an issue. To address this shortcoming, over a decade ago, Burer and Monteiro proposed to solve SDP's with few equality constraints via rank-restricted, non-convex surrogates. Remarkably, for some applications, local optimization methods seem to converge to global optima of these non-convex surrogates reliably. Although some theory supports this empirical success, a complete explanation of it remains an open question. In this paper, we consider a class of SDP's which includes applications such as max-cut, community detection in the stochastic block model, robust PCA, phase retrieval and synchronization of rotations. We show that the low-rank Burer-Monteiro formulation of SDP's in that class almost never has any spurious local optima.

****

Minimizing Regret on Reflexive Banach Spaces and Nash Equilibria in Continuous Zero-Sum Games

Maximilian Balandat, Walid Krichene, Claire Tomlin, Alexandre Bayen

We study a general adversarial online learning problem, in which we are given a decision set X' in a reflexive Banach space X and a sequence of reward vectors in the dual space of X. At each iteration, we choose an action from X', based on the observed sequence of previous rewards. Our goal is to minimize regret, defined as the gap between the realized reward and the reward of the best fixed action in hindsight. Using results from infinite dimensional convex analysis, we generalize the method of Dual Averaging (or Follow the Regularized Leader) to our setting and obtain upper bounds on the worst-case regret that generalize many previous results. Under the assumption of uniformly continuous rewards, we obtain explicit regret bounds in a setting where the decision set is the set of probability distributions on a compact metric space S. Importantly, we make no convexity assumptions on either the set S or the reward functions. We also prove a general lower bound on the worst-case regret for any online algorithm. We then apply these results to the problem of learning in repeated two-player zero-sum games on

compact metric spaces. In doing so, we first prove that if both players play a Hannan-consistent strategy, then with probability 1 the empirical distributions of play weakly converge to the set of Nash equilibria of the game. We then show that, under mild assumptions, Dual Averaging on the (infinite-dimensional) space of probability distributions indeed achieves Hannan-consistency.
****

Spatiotemporal Residual Networks for Video Action Recognition
Christoph Feichtenhofer, Axel Pinz, Richard Wildes
Two-stream Convolutional Networks (ConvNets) have shown strong performance for human action recognition in videos. Recently, Residual Networks (ResNets) have arisen as a new technique to train extremely deep architectures. In this paper, we introduce spatiotemporal ResNets as a combination of these two approaches. Our novel architecture generalizes ResNets for the spatiotemporal domain by introducing residual connections in two ways. First, we inject residual connections between the appearance and motion pathways of a two-stream architecture to allow spatiotemporal interaction between the two streams. Second, we transform pretrained image ConvNets into spatiotemporal networks by equipping these with learnable convolutional filters that are initialized as temporal residual connections and operate on adjacent feature maps in time. This approach slowly increases the spatiotemporal receptive field as the depth of the model increases and naturally integrates image ConvNet design principles. The whole model is trained end-to-end to allow hierarchical learning of complex spatiotemporal features. We evaluate our novel spatiotemporal ResNet using two widely used action recognition benchmarks where it exceeds the previous state-of-the-art.
****

Scaling Memory-Augmented Neural Networks with Sparse Reads and Writes
Jack Rae, Jonathan J. Hunt, Ivo Danihelka, Timothy Harley, Andrew W. Senior, Gregory Wayne, Alex Graves, Timothy Lillicrap
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.
****

Neurally-Guided Procedural Models: Amortized Inference for Procedural Graphics Programs using Neural Networks
Daniel Ritchie, Anna Thomas, Pat Hanrahan, Noah Goodman
Probabilistic inference algorithms such as Sequential Monte Carlo (SMC) provide powerful tools for constraining procedural models in computer graphics, but they require many samples to produce desirable results. In this paper, we show how to create procedural models which learn how to satisfy constraints. We augment procedural models with neural networks which control how the model makes random choices based on the output it has generated thus far. We call such models neurally-guided procedural models. As a pre-computation, we train these models to maximize the likelihood of example outputs generated via SMC. They are then used as efficient SMC importance samplers, generating high-quality results with very few samples. We evaluate our method on L-system-like models with image-based constraints. Given a desired quality threshold, neurally-guided models can generate satisfactory results up to 10x faster than unguided models.
****

Reconstructing Parameters of Spreading Models from Partial Observations
Andrey Lokhov
Spreading processes are often modelled as a stochastic dynamics occurring on top of a given network with edge weights corresponding to the transmission probabilities. Knowledge of veracious transmission probabilities is essential for prediction, optimization, and control of diffusion dynamics. Unfortunately, in most cases the transmission rates are unknown and need to be reconstructed from the spreading data. Moreover, in realistic settings it is impossible to monitor the state of each node at every time, and thus the data is highly incomplete. We introduce an efficient dynamic message-passing algorithm, which is able to reconstruct parameters of the spreading model given only partial information on the activat

ion times of nodes in the network. The method is generalizable to a large class of dynamic models, as well to the case of temporal graphs.

****

Tracking the Best Expert in Non-stationary Stochastic Environments

Chen-Yu Wei, Yi-Te Hong, Chi-Jen Lu

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.

****

Statistical Inference for Pairwise Graphical Models Using Score Matching

Ming Yu, Mladen Kolar, Varun Gupta

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 Structured Sparsity in Deep Neural Networks

Wei Wen, Chunpeng Wu, Yandan Wang, Yiran Chen, Hai Li

High demand for computation resources severely hinders deployment of large-scale Deep Neural Networks (DNN) in resource constrained devices. In this work, we propose a Structured Sparsity Learning (SSL) method to regularize the structures (i.e., filters, channels, filter shapes, and layer depth) of DNNs. SSL can: (1) learn a compact structure from a bigger DNN to reduce computation cost; (2) obtain a hardware-friendly structured sparsity of DNN to efficiently accelerate the DNN’s evaluation. Experimental results show that SSL achieves on average 5.1X and 3.1X speedups of convolutional layer computation of AlexNet against CPU and GPU, respectively, with off-the-shelf libraries. These speedups are about twice speedups of non-structured sparsity; (3) regularize the DNN structure to improve classification accuracy. The results show that for CIFAR-10, regularization on layer depth reduces a 20-layer Deep Residual Network (ResNet) to 18 layers while improves the accuracy from 91.25% to 92.60%, which is still higher than that of original ResNet with 32 layers. For AlexNet, SSL reduces the error by ~1%.

****

Efficient Globally Convergent Stochastic Optimization for Canonical Correlation Analysis

Weiran Wang, Jialei Wang, Dan Garber, Dan Garber, Nati Srebro

We study the stochastic optimization of canonical correlation analysis (CCA), whose objective is nonconvex and does not decouple over training samples. Although several stochastic gradient based optimization algorithms have been recently proposed to solve this problem, no global convergence guarantee was provided by any of them. Inspired by the alternating least squares/power iterations formulation of CCA, and the shift-and-invert preconditioning method for PCA, we propose two globally convergent meta-algorithms for CCA, both of which transform the original problem into sequences of least squares problems that need only be solved approximately. We instantiate the meta-algorithms with state-of-the-art SGD methods and obtain time complexities that significantly improve upon that of previous work. Experimental results demonstrate their superior performance.

****

How Deep is the Feature Analysis underlying Rapid Visual Categorization?

Sven Eberhardt, Jonah G. Cader, Thomas Serre

Rapid categorization paradigms have a long history in experimental psychology: Characterized by short presentation times and speeded behavioral responses, these tasks highlight the efficiency with which our visual system processes natural object categories. Previous studies have shown that feed-forward hierarchical models of the visual cortex provide a good fit to human visual decisions. At the same time, recent work in computer vision has demonstrated significant gains in object recognition accuracy with increasingly deep hierarchical architectures. But it is unclear how well these models account for human visual decisions and what they may reveal about the underlying brain processes.   We have conducted a large-scale psychophysics study to assess the correlation between computational models and human behavioral responses on a rapid animal vs. non-animal categorization task. We considered visual representations of varying complexity by analyzing the output of different stages of processing in three state-of-the-art deep networks. We found that recognition accuracy increases with higher stages of visual processing (higher level stages indeed outperforming human participants on the same task) but that human decisions agree best with predictions from intermediate stages.   Overall, these results suggest that human participants may rely on visual features of intermediate complexity and that the complexity of visual representations afforded by modern deep network models may exceed the complexity of those used by human participants during rapid categorization.

****

Regret of Queueing Bandits

Subhashini Krishnasamy, Rajat Sen, Ramesh Johari, Sanjay Shakkottai

We consider a variant of the multiarmed bandit problem where jobs queue for service, and service rates of different servers may be unknown. We study algorithms that minimize queue-regret: the (expected) difference between the queue-lengths obtained by the algorithm, and those obtained by a genie-aided matching algorithm that knows exact service rates. A naive view of this problem would suggest that queue-regret should grow logarithmically: since queue-regret cannot be larger than classical regret, results for the standard MAB problem give algorithms that ensure queue-regret increases no more than logarithmically in time. Our paper shows surprisingly more complex behavior. In particular, the naive intuition is correct as long as the bandit algorithm's queues have relatively long regenerative cycles: in this case queue-regret is similar to cumulative regret, and scales (essentially) logarithmically. However, we show that this "early stage" of the queueing bandit eventually gives way to a "late stage", where the optimal queue-regret scaling is O(1/t). We demonstrate an algorithm that (order-wise) achieves this asymptotic queue-regret, and also exhibits close to optimal switching time from the early stage to the late stage.

****

Dual Space Gradient Descent for Online Learning

Trung Le, Tu Nguyen, Vu Nguyen, Dinh Phung

One crucial goal in kernel online learning is to bound the model size. Common approaches employ budget maintenance procedures to restrict the model sizes using removal, projection, or merging strategies. Although projection and merging, in the literature, are known to be the most effective strategies, they demand extensive computation whilst removal strategy fails to retain information of the removed vectors. An alternative way to address the model size problem is to apply random features to approximate the kernel function. This allows the model to be maintained directly in the random feature space, hence effectively resolve the curse of kernelization. However, this approach still suffers from a serious shortcoming as it needs to use a high dimensional random feature space to achieve a sufficiently accurate kernel approximation. Consequently, it leads to a significant increase in the computational cost. To address all of these aforementioned challenges, we present in this paper the Dual Space Gradient Descent (DualSGD), a novel framework that utilizes random features as an auxiliary space to maintain information from data points removed during budget maintenance. Consequently, our approach permits the budget to be maintained in a simple, direct and elegant way while simultaneously mitigating the impact of the dimensionality issue on learning performance. We further provide convergence analysis and extensively conduct experiments on five real-world datasets to demonstrate the predictive performance and scalability of our proposed method in comparison with the state-of-the-art baselines.

****

Asynchronous Parallel Greedy Coordinate Descent

Yang You, Xiangru Lian, Ji Liu, Hsiang-Fu Yu, Inderjit S. Dhillon, James Demmel, Cho-Jui Hsieh

n this paper, we propose and study an Asynchronous parallel Greedy Coordinate Descent (Asy-GCD) algorithm for minimizing a smooth function with bounded constrai nts. At each iteration, workers asynchronously conduct greedy coordinate descent updates on a block of variables. In the first part of the paper, we analyze the theoretical behavior of Asy-GCD and prove a linear convergence rate. In the second part, we develop an efficient kernel SVM solver based on Asy-GCD in the shared memory multi-core setting. Since our algorithm is fully asynchronous---each core does not need to idle and wait for the other cores---the resulting algorithm enjoys good speedup and outperforms existing multi-core kernel SVM solvers including asynchronous stochastic coordinate descent and multi-core LIBSVM.

****

Catching heuristics are optimal control policies

Boris Belousov, Gerhard Neumann, Constantin A. Rothkopf, Jan R. Peters

Two seemingly contradictory theories attempt to explain how humans move to intercept an airborne ball. One theory posits that humans predict the ball trajectory to optimally plan future actions; the other claims that, instead of performing such complicated computations, humans employ heuristics to reactively choose appropriate actions based on immediate visual feedback. In this paper, we show that interception strategies appearing to be heuristics can be understood as computational solutions to the optimal control problem faced by a ball-catching agent acting under uncertainty. Modeling catching as a continuous partially observable Markov decision process and employing stochastic optimal control theory, we discover that the four main heuristics described in the literature are optimal solutions if the catcher has sufficient time to continuously visually track the ball. Specifically, by varying model parameters such as noise, time to ground contact, and perceptual latency, we show that different strategies arise under different circumstances. The catcher's policy switches between generating reactive and predictive behavior based on the ratio of system to observation noise and the ratio between reaction time and task duration. Thus, we provide a rational account of human ball-catching behavior and a unifying explanation for seemingly contradictory theories of target interception on the basis of stochastic optimal control.

****

Online Pricing with Strategic and Patient Buyers

Michal Feldman, Tomer Koren, Roi Livni, Yishay Mansour, Aviv Zohar

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 a Probabilistic Latent Space of Object Shapes via 3D Generative-Adversarial Modeling

Jiajun Wu, Chengkai Zhang, Tianfan Xue, Bill Freeman, Josh Tenenbaum

We study the problem of 3D object generation. We propose a novel framework, namely 3D Generative Adversarial Network (3D-GAN), which generates 3D objects from a probabilistic space by leveraging recent advances in volumetric convolutional networks and generative adversarial nets. The benefits of our model are three-fold: first, the use of an adversarial criterion, instead of traditional heuristic criteria, enables the generator to capture object structure implicitly and to synthesize high-quality 3D objects; second, the generator establishes a mapping from a low-dimensional probabilistic space to the space of 3D objects, so that we can sample objects without a reference image or CAD models, and explore the 3D object manifold; third, the adversarial discriminator provides a powerful 3D shape descriptor which, learned without supervision, has wide applications in 3D object recognition. Experiments demonstrate that our method generates high-quality 3D objects, and our unsupervisedly learned features achieve impressive performance on 3D object recognition, comparable with those of supervised learning methods.

****

Optimistic Gittins Indices

Eli Gutin, Vivek Farias

Starting with the Thomspon sampling algorithm, recent years have seen a resurgen ce of interest in Bayesian algorithms for the Multi-armed Bandit (MAB) problem. These algorithms seek to exploit prior information on arm biases and while several have been shown to be regret optimal, their design has not emerged from a principled approach. In contrast, if one cared about Bayesian regret discounted over an infinite horizon at a fixed, pre-specified rate, the celebrated Gittins index theorem offers an optimal algorithm. Unfortunately, the Gittins analysis does not appear to carry over to minimizing Bayesian regret over all sufficiently large horizons and computing a Gittins index is onerous relative to essentially any incumbent index scheme for the Bayesian MAB problem.   The present paper proposes a sequence of 'optimistic' approximations to the Gittins index. We show that the use of these approximations in concert with the use of an increasing discount factor appears to offer a compelling alternative to a variety of index schemes proposed for the Bayesian MAB problem in recent years. In addition, we show that the simplest of these approximations yields regret that matches the Lai-Robbins lower bound, including achieving matching constants.

****

Stochastic Gradient Methods for Distributionally Robust Optimization with f-divergences

Hongseok Namkoong, John C. Duchi

We develop efficient solution methods for a robust empirical risk minimization problem designed to give calibrated confidence intervals on performance and provide optimal tradeoffs between bias and variance. Our methods apply to distributionally robust optimization problems proposed by Ben-Tal et al., which put more weight on observations inducing high loss via a worst-case approach over a non-parametric uncertainty set on the underlying data distribution. Our algorithm solves the resulting minimax problems with nearly the same computational cost of stochastic gradient descent through the use of several carefully designed data structures. For a sample of size n, the per-iteration cost of our method scales as O(log n), which allows us to give optimality certificates that distributionally robust optimization provides at little extra cost compared to empirical risk minimization and stochastic gradient methods.

****

Breaking the Bandwidth Barrier: Geometrical Adaptive Entropy Estimation

Weihao Gao, Sewoong Oh, Pramod Viswanath

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.

****

Domain Separation Networks

Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, Dumitru Erhan

The cost of large scale data collection and annotation often makes the application of machine learning algorithms to new tasks or datasets prohibitively expensive. One approach circumventing this cost is training models on synthetic data where annotations are provided automatically. Despite their appeal, such models often fail to generalize from synthetic to real images, necessitating domain adaptation algorithms to manipulate these models before they can be successfully applied. Existing approaches focus either on mapping representations from one domain to the other, or on learning to extract features that are invariant to the domain from which they were extracted. However, by focusing only on creating a mapping or shared representation between the two domains, they ignore the individual characteristics of each domain. We hypothesize that explicitly modeling what is unique to each domain can improve a model's ability to extract domain-invariant features. Inspired by work on private-shared component analysis, we explicitly learn to extract image representations that are partitioned into two subspaces: one component which is private to each domain and one which is shared across domains. Our model is trained to not only perform the task we care about in the source domain, but also to use the partitioned representation to reconstruct the images from both domains. Our novel architecture results in a model that outperform s the state-of-the-art on a range of unsupervised domain adaptation scenarios and additionally produces visualizations of the private and shared representations enabling interpretation of the domain adaptation process.

****

A Probabilistic Programming Approach To Probabilistic Data Analysis
Feras Saad, Vikash K. Mansinghka

Probabilistic techniques are central to data analysis, but different approaches can be challenging to apply, combine, and compare. This paper introduces composable generative population models (CGPMs), a computational abstraction that extends directed graphical models and can be used to describe and compose a broad class of probabilistic data analysis techniques. Examples include discriminative machine learning, hierarchical Bayesian models, multivariate kernel methods, clustering algorithms, and arbitrary probabilistic programs. We demonstrate the integration of CGPMs into BayesDB, a probabilistic programming platform that can express data analysis tasks using a modeling definition language and structured query language. The practical value is illustrated in two ways. First, the paper describes an analysis on a database of Earth satellites, which identifies records that probably violate Kepler’s Third Law by composing causal probabilistic programs with non-parametric Bayes in 50 lines of probabilistic code. Second, it reports the lines of code and accuracy of CGPMs compared with baseline solutions from standard machine learning libraries.

****

Assortment Optimization Under the Mallows model
Antoine Desir, Vineet Goyal, Srikanth Jagabathula, Danny Segev

We consider the assortment optimization problem when customer preferences follow a mixture of Mallows distributions. The assortment optimization problem focuses on determining the revenue/profit maximizing subset of products from a large universe of products; it is an important decision that is commonly faced by retailers in determining what to offer their customers. There are two key challenges: (a) the Mallows distribution lacks a closed-form expression (and requires summing an exponential number of terms) to compute the choice probability and, hence, the expected revenue/profit per customer; and (b) finding the best subset may require an exhaustive search. Our key contributions are an efficiently computable closed-form expression for the choice probability under the Mallows model and a compact mixed integer linear program (MIP) formulation for the assortment problem.

****

An algorithm for L1 nearest neighbor search via monotonic embedding
Xinan Wang, Sanjoy Dasgupta

Fast algorithms for nearest neighbor (NN) search have in large part focused on L2 distance. Here we develop an approach for L1 distance that begins with an explicit and exact embedding of the points into L2. We show how this embedding can efficiently be combined with random projection methods for L2 NN search, such as locality-sensitive hashing or random projection trees. We rigorously establish the correctness of the methodology and show by experimentation that it is competitive in practice with available alternatives.

****

Multi-armed Bandits: Competing with Optimal Sequences
Zohar S. Karnin, Oren Anava

We consider sequential decision making problem in the adversarial setting, where regret is measured with respect to the optimal sequence of actions and the feedback adheres the bandit setting. It is well-known that obtaining sublinear regret in this setting is impossible in general, which arises the question of when can we do better than linear regret? Previous works show that when the environment is guaranteed to vary slowly and furthermore we are given prior knowledge regarding its variation (i.e., a limit on the amount of changes suffered by the environment), then this task is feasible. The caveat however is that such prior knowledge is not likely to be available in practice, which causes the obtained regret bounds to be somewhat irrelevant.   Our main result is a regret guarantee that scales with the variation parameter of the environment, without requiring any pr ior knowledge about it whatsoever. By that, we also resolve an open problem posted by [Gur, Zeevi and Besbes, NIPS' 14]. An important key component in our result is a statistical test for identifying non-stationarity in a sequence of independent random variables. This test either identifies non-stationarity or upper-bounds the absolute deviation of the corresponding sequence of mean values in terms of its total variation. This test is interesting on its own right and has the potential to be found useful in additional settings.

****

NESTT: A Nonconvex Primal-Dual Splitting Method for Distributed and Stochastic Optimization

Davood Hajinezhad, Mingyi Hong, Tuo Zhao, Zhaoran Wang

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.

****

Probing the Compositionality of Intuitive Functions

Eric Schulz, Josh Tenenbaum, David K. Duvenaud, Maarten Speekenbrink, Samuel J. Gershman

How do people learn about complex functional structure? Taking inspiration from other areas of cognitive science, we propose that this is accomplished by harnessing compositionality: complex structure is decomposed into simpler building blocks. We formalize this idea within the framework of Bayesian regression using a grammar over Gaussian process kernels. We show that participants prefer compositional over non-compositional function extrapolations, that samples from the human prior over functions are best described by a compositional model, and that people perceive compositional functions as more predictable than their non-compositional but otherwise similar counterparts. We argue that the compositional nature of intuitive functions is consistent with broad principles of human cognition.

****

Identification and Overidentification of Linear Structural Equation Models

Bryant Chen

In this paper, we address the problems of identifying linear structural equation models and discovering the constraints they imply. We first extend the half-trek criterion to cover a broader class of models and apply our extension to finding testable constraints implied by the model. We then show that any semi-Markovian linear model can be recursively decomposed into simpler sub-models, resulting in improved identification and constraint discovery power. Finally, we show that, unlike the existing methods developed for linear models, the resulting method subsumes the identification and constraint discovery algorithms for non-parametric models.

****

An Architecture for Deep, Hierarchical Generative Models

Philip Bachman

We present an architecture which lets us train deep, directed generative models with many layers of latent variables. We include deterministic paths between all latent variables and the generated output, and provide a richer set of connections between computations for inference and generation, which enables more effective communication of information throughout the model during training. To improve performance on natural images, we incorporate a lightweight autoregressive model in the reconstruction distribution. These techniques permit end-to-end training of models with 10+ layers of latent variables. Experiments show that our approach achieves state-of-the-art performance on standard image modelling benchmarks, can expose latent class structure in the absence of label information, and can provide convincing imputations of occluded regions in natural images.

****

Towards Conceptual Compression

Karol Gregor, Frederic Besse, Danilo Jimenez Rezende, Ivo Danihelka, Daan Wierstra

We introduce convolutional DRAW, a homogeneous deep generative model achieving s tate-of-the-art performance in latent variable image modeling. The algorithm naturally stratifies information into higher and lower level details, creating abstract features and as such addressing one of the fundamentally desired properties of representation learning. Furthermore, the hierarchical ordering of its latents creates the opportunity to selectively store global information about an image, yielding a high quality 'conceptual compression' framework.

****

Exploiting the Structure: Stochastic Gradient Methods Using Raw Clusters

Zeyuan Allen-Zhu, Yang Yuan, Karthik Sridharan

The amount of data available in the world is growing faster than our ability to deal with it. However, if we take advantage of the internal structure, data may become much smaller for machine learning purposes. In this paper we focus on one of the fundamental machine learning tasks, empirical risk minimization (ERM), and provide faster algorithms with the help from the clustering structure of the data. We introduce a simple notion of raw clustering that can be efficiently computed from the data, and propose two algorithms based on clustering information. Our accelerated algorithm ClusterACDM is built on a novel Haar transformation applied to the dual space of the ERM problem, and our variance-reduction based algorithm ClusterSVRG introduces a new gradient estimator using clustering. Our algorithms outperform their classical counterparts ACDM and SVRG respectively.

****

Consistent Kernel Mean Estimation for Functions of Random Variables

Carl-Johann Simon-Gabriel, Adam Scibior, Ilya O. Tolstikhin, Bernhard Schölkopf

We provide a theoretical foundation for non-parametric estimation of functions of random variables using kernel mean embeddings. We show that for any continuous function f, consistent estimators of the mean embedding of a random variable X lead to consistent estimators of the mean embedding of f(X). For Matern kernels and sufficiently smooth functions we also provide rates of convergence. Our results extend to functions of multiple random variables. If the variables are dependent, we require an estimator of the mean embedding of their joint distribution as a starting point; if they are independent, it is sufficient to have separate estimators of the mean embeddings of their marginal distributions. In either case, our results cover both mean embeddings based on i.i.d. samples as well as "reduced set" expansions in terms of dependent expansion points. The latter serves as a justification for using such expansions to limit memory resources when applying the approach as a basis for probabilistic programming.

****

Hierarchical Clustering via Spreading Metrics

Aurko Roy, Sebastian Pokutta

We study the cost function for hierarchical clusterings introduced by [Dasgupta, 2015] where hierarchies are treated as first-class objects rather than deriving their cost from projections into flat clusters. It was also shown in [Dasgupta, 2015] that a top-down algorithm returns a hierarchical clustering of cost at most (O\left(\alphan \log n\right)) times the cost of the optimal hierarchical clustering, where (\alphan) is the approximation ratio of the Sparsest Cut subroutine used. Thus using the best known approximation algorithm for Sparsest Cut due to Arora-Rao-Vazirani, the top down algorithm returns a hierarchical clustering of cost at most (O\left(\log^{3/2} n\right)) times the cost of the optimal solution. We improve this by giving an (O(\log{n}))-approximation algorithm for this problem. Our main technical ingredients are a combinatorial characterization of ultrametrics induced by this cost function, deriving an Integer Linear Programming (ILP) formulation for this family of ultrametrics, and showing how to iteratively round an LP relaxation of this formulation by using the idea of \emph{sphere growing} which has been extensively used in the context of graph partitioning. We also prove that our algorithm returns an (O(\log{n}))-approximate hierarchical clustering for a generalization of this cost function also studied in [Dasgupta, 2015]. Experiments show that the hierarchies found by using the ILP formulation as well as our rounding algorithm often have better projections into flat clusters than the standard linkage based algorithms. We conclude with an inapproximability result for this problem, namely that no polynomial sized LP or

SDP can be used to obtain a constant factor approximation for this problem.

****

Combining Fully Convolutional and Recurrent Neural Networks for 3D Biomedical Image Segmentation

Jianxu Chen, Lin Yang, Yizhe Zhang, Mark Alber, Danny Z. Chen

Segmentation of 3D images is a fundamental problem in biomedical image analysis. Deep learning (DL) approaches have achieved the state-of-the-art segmentation performance. To exploit the 3D contexts using neural networks, known DL segmentation methods, including 3D convolution, 2D convolution on the planes orthogonal to 2D slices, and LSTM in multiple directions, all suffer incompatibility with the highly anisotropic dimensions in common 3D biomedical images. In this paper, we propose a new DL framework for 3D image segmentation, based on a combination of a fully convolutional network (FCN) and a recurrent neural network (RNN), which are responsible for exploiting the intra-slice and inter-slice contexts, respectively. To our best knowledge, this is the first DL framework for 3D image segmentation that explicitly leverages 3D image anisotropism. Evaluating using a dataset from the ISBI Neuronal Structure Segmentation Challenge and in-house image stacks for 3D fungus segmentation, our approach achieves promising results, comparing to the known DL-based 3D segmentation approaches.

****

SDP Relaxation with Randomized Rounding for Energy Disaggregation

Kiarash Shaloudegi, András György, Csaba Szepesvari, Wilsun Xu

We develop a scalable, computationally efficient method for the task of energy disaggregation for home appliance monitoring. In this problem the goal is to estimate the energy consumption of each appliance based on the total energy-consumption signal of a household. The current state of the art models the problem as inference in factorial HMMs, and finds an approximate solution to the resulting quadratic integer program via quadratic programming. Here we take a more principled approach, better suited to integer programming problems, and find an approximate optimum by combining convex semidefinite relaxations with randomized rounding, as well as with a scalable ADMM method that exploits the special structure of the resulting semidefinite program. Simulation results demonstrate the superiority of our methods both in synthetic and real-world datasets.

****

Finite Sample Prediction and Recovery Bounds for Ordinal Embedding

Lalit Jain, Kevin G. Jamieson, Rob Nowak

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.

****

Search Improves Label for Active Learning

Alina Beygelzimer, Daniel J. Hsu, John Langford, Chicheng Zhang

We investigate active learning with access to two distinct oracles: LABEL (which is standard) and SEARCH (which is not). The SEARCH oracle models the situation where a human searches a database to seed or counterexample an existing solution. SEARCH is stronger than LABEL while being natural to implement in many situations. We show that an algorithm using both oracles can provide exponentially large problem-dependent improvements over LABEL alone.

****

A Simple Practical Accelerated Method for Finite Sums

Aaron Defazio

Abstract We describe a novel optimization method for finite sums (such as empirical risk minimization problems) building on the recently introduced SAGA method. Our method achieves an accelerated convergence rate on strongly convex smooth problems. Our method has only one parameter (a step size), and is radically simpler than other accelerated methods for finite sums. Additionally it can be applied when the terms are non-smooth, yielding a method applicable in many areas where operator splitting methods would traditionally be applied.

****

Coupled Generative Adversarial Networks

Ming-Yu Liu, Oncel Tuzel

We propose the coupled generative adversarial nets (CoGAN) framework for generating pairs of corresponding images in two different domains. The framework consists of a pair of generative adversarial nets, each responsible for generating images in one domain. We show that by enforcing a simple weight-sharing constraint, the CoGAN learns to generate pairs of corresponding images without existence of any pairs of corresponding images in the two domains in the training set. In other words, the CoGAN learns a joint distribution of images in the two domains from images drawn separately from the marginal distributions of the individual domains. This is in contrast to the existing multi-modal generative models, which require corresponding images for training. We apply the CoGAN to several pair image generation tasks. For each task, the CoGAN learns to generate convincing pairs of corresponding images. We further demonstrate the applications of the CoGAN framework for the domain adaptation and cross-domain image generation tasks.

****

Minimax Estimation of Maximum Mean Discrepancy with Radial Kernels

Ilya O. Tolstikhin, Bharath K. Sriperumbudur, Bernhard Schölkopf

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.

****

Using Social Dynamics to Make Individual Predictions: Variational Inference with a Stochastic Kinetic Model

Zhen Xu, Wen Dong, Sargur N. Srihari

Social dynamics is concerned primarily with interactions among individuals and the resulting group behaviors, modeling the temporal evolution of social systems via the interactions of individuals within these systems. In particular, the availability of large-scale data from social networks and sensor networks offers an unprecedented opportunity to predict state-changing events at the individual level. Examples of such events include disease transmission, opinion transition in elections, and rumor propagation. Unlike previous research focusing on the collective effects of social systems, this study makes efficient inferences at the individual level. In order to cope with dynamic interactions among a large number of individuals, we introduce the stochastic kinetic model to capture adaptive transition probabilities and propose an efficient variational inference algorithm the complexity of which grows linearly – rather than exponentially– with the number of individuals. To validate this method, we have performed epidemic-dynamics experiments on wireless sensor network data collected from more than ten thousand people over three years. The proposed algorithm was used to track disease transmission and predict the probability of infection for each individual. Our results demonstrate that this method is more efficient than sampling while nonetheless achieving high accuracy.

****

Multiple-Play Bandits in the Position-Based Model

Paul Lagrée, Claire Vernade, Olivier Cappe

Sequentially learning to place items in multi-position displays or lists is a task that can be cast into the multiple-play semi-bandit setting. However, a major concern in this context is when the system cannot decide whether the user feedback for each item is actually exploitable. Indeed, much of the content may have been simply ignored by the user. The present work proposes to exploit available information regarding the display position bias under the so-called Position-based click model (PBM). We first discuss how this model differs from the Cascade model and its variants considered in several recent works on multiple-play bandits. We then provide a novel regret lower bound for this model as well as computationally efficient algorithms that display good empirical and theoretical performance.

****

Learning values across many orders of magnitude

Hado P. van Hasselt, Arthur Guez, Arthur Guez, Matteo Hessel, Volodymyr Mnih, David Silver

Most learning algorithms are not invariant to the scale of the signal that is being approximated. We propose to adaptively normalize the targets used in the learning updates. This is important in value-based reinforcement learning, where the magnitude of appropriate value approximations can change over time when we update the policy of behavior. Our main motivation is prior work on learning to play Atari games, where the rewards were clipped to a predetermined range. This clipping facilitates learning across many different games with a single learning algorithm, but a clipped reward function can result in qualitatively different behavior. Using adaptive normalization we can remove this domain-specific heuristic without diminishing overall performance.

****

Attend, Infer, Repeat: Fast Scene Understanding with Generative Models

S. M. Ali Eslami, Nicolas Heess, Theophane Weber, Yuval Tassa, David Szepesvari, koray kavukcuoglu, Geoffrey E. Hinton

We present a framework for efficient inference in structured image models that explicitly reason about objects. We achieve this by performing probabilistic inference using a recurrent neural network that attends to scene elements and processes them one at a time. Crucially, the model itself learns to choose the appropriate number of inference steps. We use this scheme to learn to perform inference in partially specified 2D models (variable-sized variational auto-encoders) and fully specified 3D models (probabilistic renderers). We show that such models learn to identify multiple objects - counting, locating and classifying the elements of a scene - without any supervision, e.g., decomposing 3D images with various numbers of objects in a single forward pass of a neural network at unprecedented speed. We further show that the networks produce accurate inferences when compared to supervised counterparts, and that their structure leads to improved generalization.

****

Supervised Learning with Tensor Networks

Edwin Stoudenmire, David J. Schwab

Tensor networks are approximations of high-order tensors which are efficient to work with and have been very successful for physics and mathematics applications. We demonstrate how algorithms for optimizing tensor networks can be adapted to supervised learning tasks by using matrix product states (tensor trains) to parameterize non-linear kernel learning models. For the MNIST data set we obtain less than 1% test set classification error. We discuss an interpretation of the additional structure imparted by the tensor network to the learned model.

****

Structured Prediction Theory Based on Factor Graph Complexity

Corinna Cortes, Vitaly Kuznetsov, Mehryar Mohri, Scott Yang

We present a general theoretical analysis of structured prediction with a series of new results. We give new data-dependent margin guarantees for structured prediction for a very wide family of loss functions and a general family of hypotheses, with an arbitrary factor graph decomposition. These are the tightest margin bounds known for both standard multi-class and general structured prediction problems. Our guarantees are expressed in terms of a data-dependent complexity measure, \emph{factor graph complexity}, which we show can be estimated from data and bounded in terms of familiar quantities for several commonly used hypothesis sets, and a sparsity measure for features and graphs. Our proof techniques include generalizations of Talagrand's contraction lemma that can be of independent interest. We further extend our theory by leveraging the principle of Voted Risk Minimization (VRM) and show that learning is possible even with complex factor graphs. We present new learning bounds for this advanced setting, which we use to devise two new algorithms, \emph{Voted Conditional Random Field} (VCRF) and \emph{Voted Structured Boosting} (StructBoost). These algorithms can make use of complex features and factor graphs and yet benefit from favorable learning guarantees. We also report the results of experiments with VCRF on several datasets to validate our theory.

****

The Multiple Quantile Graphical Model

Alnur Ali, J. Zico Kolter, Ryan J. Tibshirani

We introduce the Multiple Quantile Graphical Model (MQGM), which extends the neighborhood selection approach of Meinshausen and Buhlmann for learning sparse graphical models. The latter is defined by the basic subproblem of modeling the conditional mean of one variable as a sparse function of all others. Our approach models a set of conditional quantiles of one variable as a sparse function of all others, and hence offers a much richer, more expressive class of conditional distribution estimates. We establish that, under suitable regularity conditions, the MQGM identifies the exact conditional independencies with probability tending to one as the problem size grows, even outside of the usual homoskedastic Gaussian data model. We develop an efficient algorithm for fitting the MQGM using the alternating direction method of multipliers. We also describe a strategy for sampling from the joint distribution that underlies the MQGM estimate. Lastly, we present detailed experiments that demonstrate the flexibility and effectiveness of the MQGM in modeling hetereoskedastic non-Gaussian data.

****

Orthogonal Random Features

Felix Xinnan X. Yu, Ananda Theertha Suresh, Krzysztof M. Choromanski, Daniel N. Holtmann-Rice, Sanjiv Kumar

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.

****

Coevolutionary Latent Feature Processes for Continuous-Time User-Item Interactions

Yichen Wang, Nan Du, Rakshit Trivedi, Le Song

Matching users to the right items at the right time is a fundamental task in recommendation systems. As users interact with different items over time, users' and items' feature may evolve and co-evolve over time. Traditional models based on static latent features or discretizing time into epochs can become ineffective for capturing the fine-grained temporal dynamics in the user-item interactions. We propose a coevolutionary latent feature process model that accurately captures the coevolving nature of users' and items' feature. To learn parameters, we design an efficient convex optimization algorithm with a novel low rank space sharing constraints. Extensive experiments on diverse real-world datasets demonstrate significant improvements in user behavior prediction compared to state-of-the-arts.

****

Convex Two-Layer Modeling with Latent Structure

Vignesh Ganapathiraman, Xinhua Zhang, Yaoliang Yu, Junfeng Wen

Unsupervised learning of structured predictors has been a long standing pursuit in machine learning. Recently a conditional random field auto-encoder has been proposed in a two-layer setting, allowing latent structured representation to be automatically inferred. Aside from being nonconvex, it also requires the demanding inference of normalization. In this paper, we develop a convex relaxation of two-layer conditional model which captures latent structure and estimates model parameters, jointly and optimally. We further expand its applicability by resorting to a weaker form of inference---maximum a-posteriori. The flexibility of the model is demonstrated on two structures based on total unimodularity---graph matching and linear chain. Experimental results confirm the promise of the method.

****

Online Convex Optimization with Unconstrained Domains and Losses

Ashok Cutkosky, Kwabena A. Boahen

We propose an online convex optimization algorithm (RescaledExp) that achieves optimal regret in the unconstrained setting without prior knowledge of any bounds on the loss functions. We prove a lower bound showing an exponential separation between the regret of existing algorithms that require a known bound on the loss functions and any algorithm that does not require such knowledge. RescaledExp matches this lower bound asymptotically in the number of iterations. RescaledExp is naturally hyperparameter-free and we demonstrate empirically that it matches prior optimization algorithms that require hyperparameter optimization.

****

GAP Safe Screening Rules for Sparse-Group Lasso

Eugene Ndiaye, Olivier Fercoq, Alexandre Gramfort, Joseph Salmon

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 Similarity-Aware Deep Feature Embedding

Chen Huang, Chen Change Loy, Xiaoou Tang

Existing deep embedding methods in vision tasks are capable of learning a compact Euclidean space from images, where Euclidean distances correspond to a similarity metric. To make learning more effective and efficient, hard sample mining is usually employed, with samples identified through computing the Euclidean feature distance. However, the global Euclidean distance cannot faithfully characterize the true feature similarity in a complex visual feature space, where the intraclass distance in a high-density region may be larger than the interclass distance in low-density regions. In this paper, we introduce a Position-Dependent Deep Metric (PDDM) unit, which is capable of learning a similarity metric adaptive to local feature structure. The metric can be used to select genuinely hard samples in a local neighborhood to guide the deep embedding learning in an online and robust manner. The new layer is appealing in that it is pluggable to any convolutional networks and is trained end-to-end. Our local similarity-aware feature embedding not only demonstrates faster convergence and boosted performance on two complex image retrieval datasets, its large margin nature also leads to superior generalization results under the large and open set scenarios of transfer learning and zero-shot learning on ImageNet 2010 and ImageNet-10K datasets.

****

Following the Leader and Fast Rates in Linear Prediction: Curved Constraint Sets and Other Regularities

Ruitong Huang, Tor Lattimore, András György, Csaba Szepesvari

The follow the leader (FTL) algorithm, perhaps the simplest of all online learning algorithms, is known to perform well when the loss functions it is used on are positively curved. In this paper we ask whether there are other "lucky" settings when FTL achieves sublinear, "small" regret. In particular, we study the fundamental problem of linear prediction over a non-empty convex, compact domain. Amongst other results, we prove that the curvature of the boundary of the domain can act as if the losses were curved: In this case, we prove that as long as the mean of the loss vectors have positive lengths bounded away from zero, FTL enjoys a logarithmic growth rate of regret, while, e.g., for polyhedral domains and stochastic data it enjoys finite expected regret. Building on a previously known meta-algorithm, we also get an algorithm that simultaneously enjoys the worst-case guarantees and the bound available for FTL.

****

Learning Multiagent Communication with Backpropagation

Sainbayar Sukhbaatar, arthur szlam, Rob Fergus

Many tasks in AI require the collaboration of multiple agents. Typically, the communication protocol between agents is manually specified and not altered during training. In this paper we explore a simple neural model, called CommNet, that uses continuous communication for fully cooperative tasks. The model consists of multiple agents and the communication between them is learned alongside their policy. We apply this model to a diverse set of tasks, demonstrating the ability of the agents to learn to communicate amongst themselves, yielding improved performance over non-communicative agents and baselines. In some cases, it is possible to interpret the language devised by the agents, revealing simple but effecti ve strategies for solving the task at hand.

****

Sub-sampled Newton Methods with Non-uniform Sampling

Peng Xu, Jiyan Yang, Fred Roosta, Christopher Ré, 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.

****

Examples are not enough, learn to criticize! Criticism for Interpretability

Been Kim, Rajiv Khanna, Oluwasanmi O. Koyejo

Example-based explanations are widely used in the effort to improve the interpretability of highly complex distributions. However, prototypes alone are rarely sufficient to represent the gist of the complexity. In order for users to construct better mental models and understand complex data distributions, we also need {\em criticism} to explain what are \textit{not} captured by prototypes. Motivated by the Bayesian model criticism framework, we develop \texttt{MMD-critic} which efficiently learns prototypes and criticism, designed to aid human interpretability. A human subject pilot study shows that the \texttt{MMD-critic} selects prototypes and criticism that are useful to facilitate human understanding and reasoning. We also evaluate the prototypes selected by \texttt{MMD-critic} via a nearest prototype classifier, showing competitive performance compared to baselines.

****

R-FCN: Object Detection via Region-based Fully Convolutional Networks

Jifeng Dai, Yi Li, Kaiming He, Jian Sun

We present region-based, fully convolutional networks for accurate and efficient object detection. In contrast to previous region-based detectors such as Fast/Faster R-CNN that apply a costly per-region subnetwork hundreds of times, our region-based detector is fully convolutional with almost all computation shared on the entire image. To achieve this goal, we propose position-sensitive score maps to address a dilemma between translation-invariance in image classification and translation-variance in object detection. Our method can thus naturally adopt fully convolutional image classifier backbones, such as the latest Residual Networks (ResNets), for object detection. We show competitive results on the PASCAL VOC datasets (e.g., 83.6% mAP on the 2007 set) with the 101-layer ResNet. Meanwhile, our result is achieved at a test-time speed of 170ms per image, 2.5-20 times faster than the Faster R-CNN counterpart. Code is made publicly available at: https://github.com/daijifeng001/r-fcn.

****

Exploiting Tradeoffs for Exact Recovery in Heterogeneous Stochastic Block Models

Amin Jalali, Qiyang Han, Ioana Dumitriu, Maryam Fazel

The Stochastic Block Model (SBM) is a widely used random graph model for networks with communities. Despite the recent burst of interest in community detection under the SBM from statistical and computational points of view, there are still gaps in understanding the fundamental limits of recovery. In this paper, we consider the SBM in its full generality, where there is no restriction on the number and sizes of communities or how they grow with the number of nodes, as well as on the connectivity probabilities inside or across communities. For such stochastic block models, we provide guarantees for exact recovery via a semidefinite program as well as upper and lower bounds on SBM parameters for exact recoverability. Our results exploit the tradeoffs among the various parameters of heterogenous SBM and provide recovery guarantees for many new interesting SBM configurations.

****

A Powerful Generative Model Using Random Weights for the Deep Image Representation

Kun He, Yan Wang, John Hopcroft

To what extent is the success of deep visualization due to the training? Could we do deep visualization using untrained, random weight networks? To address this issue, we explore new and powerful generative models for three popular deep visualization tasks using untrained, random weight convolutional neural networks. First we invert representations in feature spaces and reconstruct images from white noise inputs. The reconstruction quality is statistically higher than that of the same method applied on well trained networks with the same architecture. Next we synthesize textures using scaled correlations of representations in multiple layers and our results are almost indistinguishable with the original natural texture and the synthesized textures based on the trained network. Third, by recasting the content of an image in the style of various artworks, we create artistic images with high perceptual quality, highly competitive to the prior work of Gatys et al. on pretrained networks. To our knowledge this is the first demonstration of image representations using untrained deep neural networks. Our work provides a new and fascinating tool to study the representation of deep network architecture and sheds light on new understandings on deep visualization. It may possibly lead to a way to compare network architectures without training.

****

Privacy Odometers and Filters: Pay-as-you-Go Composition

Ryan M. Rogers, Aaron Roth, Jonathan Ullman, Salil Vadhan

In this paper we initiate the study of adaptive composition in differential privacy when the length of the composition, and the privacy parameters themselves can be chosen adaptively, as a function of the outcome of previously run analyses. This case is much more delicate than the setting covered by existing composition theorems, in which the algorithms themselves can be chosen adaptively, but the privacy parameters must be fixed up front. Indeed, it isn't even clear how to define differential privacy in the adaptive parameter setting. We proceed by defining two objects which cover the two main use cases of composition theorems. A privacy filter is a stopping time rule that allows an analyst to halt a computation before his pre-specified privacy budget is exceeded. A privacy odometer allows the analyst to track realized privacy loss as he goes, without needing to pre-specify a privacy budget. We show that unlike the case in which privacy parameters are fixed, in the adaptive parameter setting, these two use cases are distinct. We show that there exist privacy filters with bounds comparable (up to constants) with existing privacy composition theorems. We also give a privacy odometer that nearly matches non-adaptive private composition theorems, but is sometimes worse by a small asymptotic factor. Moreover, we show that this is inherent, and that any valid privacy odometer in the adaptive parameter setting must lose this factor, which shows a formal separation between the filter and odometer use-cases.

****

More Supervision, Less Computation: Statistical-Computational Tradeoffs in Weakly Supervised Learning

Xinyang Yi, Zhaoran Wang, Zhuoran Yang, 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.

****

Supervised learning through the lens of compression

Ofir David, Shay Moran, Amir Yehudayoff

This work continues the study of the relationship between sample compression schemes and statistical learning, which has been mostly investigated within the framework of binary classification. We first extend the investigation to multiclass categorization: we prove that in this case learnability is equivalent to compression of logarithmic sample size and that the uniform convergence property implies compression of constant size. We use the compressibility-learnability equivalence to show that (i) for multiclass categorization, PAC and agnostic PAC learnability are equivalent, and (ii) to derive a compactness theorem for learnability. We then consider supervised learning under general loss functions: we show that in this case, in order to maintain the compressibility-learnability equivalence, it is necessary to consider an approximate variant of compression. We use it

to show that PAC and agnostic PAC are not equivalent, even when the loss function has only three values.
****
Sparse Support Recovery with Non-smooth Loss Functions
Kévin Degraux, Gabriel Peyré, Jalal Fadili, Laurent Jacques
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 Operations for Arithmetic Circuits of Probabilistic Models
Yujia Shen, Arthur Choi, Adnan Darwiche
We consider tractable representations of probability distributions and the polytime operations they support. In particular, we consider a recently proposed arithmetic circuit representation, the Probabilistic Sentential Decision Diagram (PSDD). We show that PSDD supports a polytime multiplication operator, while they do not support a polytime operator for summing-out variables. A polytime multiplication operator make PSDDs suitable for a broader class of applications compared to arithmetic circuits, which do not in general support multiplication. As one example, we show that PSDD multiplication leads to a very simple but effective compilation algorithm for probabilistic graphical models: represent each model factor as a PSDD, and then multiply them.
****
Dual Learning for Machine Translation
Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, Wei-Ying Ma
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 Systems of Quadratic Equations via Truncated Generalized Gradient Flow
Gang Wang, Georgios Giannakis
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.
****
Phased LSTM: Accelerating Recurrent Network Training for Long or Event-based Sequences
Daniel Neil, Michael Pfeiffer, Shih-Chii Liu
Recurrent Neural Networks (RNNs) have become the state-of-the-art choice for extracting patterns from temporal sequences. Current RNN models are ill suited to process irregularly sampled data triggered by events generated in continuous time by sensors or other neurons. Such data can occur, for example, when the input comes from novel event-driven artificial sensors which generate sparse, asynchronous streams of events or from multiple conventional sensors with different update intervals. In this work, we introduce the Phased LSTM model, which extends the LSTM unit by adding a new time gate. This gate is controlled by a parametrized oscillation with a frequency range which require updates of the memory cell only during a small percentage of the cycle. Even with the sparse updates imposed by the oscillation, the Phased LSTM network achieves faster convergence than regular LSTMs on tasks which require learning of long sequences.   The model naturally integrates inputs from sensors of arbitrary sampling rates, thereby opening new areas of investigation for processing asynchronous sensory events that carry timing information. It also greatly improves the performance of LSTMs in standard RNN applications, and does so with an order-of-magnitude fewer computes.
****
Only H is left: Near-tight Episodic PAC RL

In many applications such as advertisement placement or automated dialog systems, an intelligent system optimizes performance over a sequence of interactions with each user. Such tasks often involve many states and potentially time-dependent transition dynamics, and can be modeled well as episodic Markov decision processes (MDPs). In this paper, we present a PAC algorithm for reinforcement learning in episodic finite MDPs with time-dependent transitions that acts epsilon-optimal in all but O(S A H^3 / epsilon^2 log(1 / delta)) episodes. Our algorithm has a polynomial computational complexity, and our sample complexity bound accounts for the fact that we may only be able to approximately solve the internal planning problems. In addition, our PAC sample complexity bound has only linear dependency on the number of states S and actions A and strictly improves previous bounds with S^2 dependency in this setting. Compared against other methods for infinite horizon reinforcement learning with linear state space sample complexity our method has much lower dependency on the (effective) horizon. Indeed, our bound is optimal up to a factor of H.

****

Stochastic Three-Composite Convex Minimization

Alp Yurtsever, Bang Cong Vu, Volkan Cevher

We propose a stochastic optimization method for the minimization of the sum of three convex functions, one of which has Lipschitz continuous gradient as well as restricted strong convexity. Our approach is most suitable in the setting where it is computationally advantageous to process smooth term in the decomposition with its stochastic gradient estimate and the other two functions separately with their proximal operators, such as doubly regularized empirical risk minimization problems. We prove the convergence characterization of the proposed algorithm in expectation under the standard assumptions for the stochastic gradient estimate of the smooth term. Our method operates in the primal space and can be considered as a stochastic extension of the three-operator splitting method. Finally, numerical evidence supports the effectiveness of our method in real-world problems.

****

Synthesizing the preferred inputs for neurons in neural networks via deep generator networks

Anh Nguyen, Alexey Dosovitskiy, Jason Yosinski, Thomas Brox, Jeff Clune

Deep neural networks (DNNs) have demonstrated state-of-the-art results on many pattern recognition tasks, especially vision classification problems. Understanding the inner workings of such computational brains is both fascinating basic science that is interesting in its own right---similar to why we study the human brain---and will enable researchers to further improve DNNs. One path to understanding how a neural network functions internally is to study what each of its neurons has learned to detect. One such method is called activation maximization, which synthesizes an input (e.g. an image) that highly activates a neuron. Here we dramatically improve the qualitative state of the art of activation maximization by harnessing a powerful, learned prior: a deep generator network. The algorithm (1) generates qualitatively state-of-the-art synthetic images that look almost real, (2) reveals the features learned by each neuron in an interpretable way, (3) generalizes well to new datasets and somewhat well to different network architectures without requiring the prior to be relearned, and (4) can be considered as a high-quality generative method (in this case, by generating novel, creative, interesting, recognizable images).

****

Bayesian Optimization with a Finite Budget: An Approximate Dynamic Programming Approach

Remi Lam, Karen Willcox, David H. Wolpert

We consider the problem of optimizing an expensive objective function when a finite budget of total evaluations is prescribed. In that context, the optimal solution strategy for Bayesian optimization can be formulated as a dynamic programming instance. This results in a complex problem with uncountable, dimension-increasing state space and an uncountable control space. We show how to approximate the solution of this dynamic programming problem using rollout, and propose roll out heuristics specifically designed for the Bayesian optimization setting. We present numerical experiments showing that the resulting algorithm for optimization with a finite budget outperforms several popular Bayesian optimization algorithms.

****

Gaussian Process Bandit Optimisation with Multi-fidelity Evaluations

Kirthevasan Kandasamy, Gautam Dasarathy, Junier B. Oliva, Jeff Schneider, Barnabas Poczos

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 Parametric Sparse Models for Image Super-Resolution

Yongbo Li, Weisheng Dong, Xuemei Xie, GUANGMING Shi, Xin Li, Donglai Xu

Learning accurate prior knowledge of natural images is of great importance for single image super-resolution (SR). Existing SR methods either learn the prior from the low/high-resolution patch pairs or estimate the prior models from the input low-resolution (LR) image. Specifically, high-frequency details are learned in the former methods. Though effective, they are heuristic and have limitations in dealing with blurred LR images; while the latter suffers from the limitations of frequency aliasing. In this paper, we propose to combine those two lines of ideas for image super-resolution. More specifically, the parametric sparse prior of the desirable high-resolution (HR) image patches are learned from both the input low-resolution (LR) image and a training image dataset. With the learned sparse priors, the sparse codes and thus the HR image patches can be accurately recovered by solving a sparse coding problem. Experimental results show that the proposed SR method outperforms existing state-of-the-art methods in terms of both subjective and objective image qualities.

****

Mutual information for symmetric rank-one matrix estimation: A proof of the replica formula

jean barbier, Mohamad Dia, Nicolas Macris, Florent Krzakala, Thibault Lesieur, Lenka Zdeborová

Factorizing low-rank matrices has many applications in machine learning and statistics. For probabilistic models in the Bayes optimal setting, a general expression for the mutual information has been proposed using heuristic statistical physics computations, and proven in few specific cases. Here, we show how to rigorously prove the conjectured formula for the symmetric rank-one case. This allows to express the minimal mean-square-error and to characterize the detectability phase transitions in a large set of estimation problems ranging from community detection to sparse PCA. We also show that for a large set of parameters, an iterative algorithm called approximate message-passing is Bayes optimal. There exists, however, a gap between what currently known polynomial algorithms can do and what is expected information theoretically. Additionally, the proof technique has an interest of its own and exploits three essential ingredients: the interpolation method introduced in statistical physics by Guerra, the analysis of the approximate message-passing algorithm and the theory of spatial coupling and threshold saturation in coding. Our approach is generic and applicable to other open problems in statistical estimation where heuristic statistical physics predictions are available.

****

Large Margin Discriminant Dimensionality Reduction in Prediction Space

Mohammad Saberian, Jose Costa Pereira, Can Xu, Jian Yang, Nuno Nvasconcelos

In this paper we establish a duality between boosting and SVM, and use this to derive a novel discriminant dimensionality reduction algorithm. In particular, using the multiclass formulation of boosting and SVM we note that both use a combination of mapping and linear classification to maximize the multiclass margin. In SVM this is implemented using a pre-defined mapping (induced by the kernel) and optimizing the linear classifiers. In boosting the linear classifiers are predefined and the mapping (predictor) is learned through combination of weak learners. We argue that the intermediate mapping, e.g. boosting predictor, is preserving the discriminant aspects of the data and by controlling the dimension of this mapping it is possible to achieve discriminant low dimensional representations for the data. We use the aforementioned duality and propose a new method, Large Margin Discriminant Dimensionality Reduction (LADDER) that jointly learns the mapping and the linear classifiers in an efficient manner. This leads to a data-driven mapping which can embed data into any number of dimensions. Experimental results show that this embedding can significantly improve performance on tasks such as hashing and image/scene classification.

****

Fast learning rates with heavy-tailed losses

Vu C. Dinh, Lam S. Ho, Binh Nguyen, Duy Nguyen

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.

****

Dynamic matrix recovery from incomplete observations under an exact low-rank constraint

Liangbei Xu, Mark Davenport

Low-rank matrix factorizations arise in a wide variety of applications -- including recommendation systems, topic models, and source separation, to name just a few. In these and many other applications, it has been widely noted that by incorporating temporal information and allowing for the possibility of time-varying models, significant improvements are possible in practice. However, despite the reported superior empirical performance of these dynamic models over their static counterparts, there is limited theoretical justification for introducing these more complex models. In this paper we aim to address this gap by studying the problem of recovering a dynamically evolving low-rank matrix from incomplete observations. First, we propose the locally weighted matrix smoothing (LOWEMS) framework as one possible approach to dynamic matrix recovery. We then establish error bounds for LOWEMS in both the {\em matrix sensing} and {\em matrix completion} observation models. Our results quantify the potential benefits of exploiting dynamic constraints both in terms of recovery accuracy and sample complexity. To illustrate these benefits we provide both synthetic and real-world experimental results.

****

Tight Complexity Bounds for Optimizing Composite Objectives

Blake E. Woodworth, Nati Srebro

We provide tight upper and lower bounds on the complexity of minimizing the average of m convex functions using gradient and prox oracles of the component functions. We show a significant gap between the complexity of deterministic vs randomized optimization. For smooth functions, we show that accelerated gradient descent (AGD) and an accelerated variant of SVRG are optimal in the deterministic and randomized settings respectively, and that a gradient oracle is sufficient for the optimal rate. For non-smooth functions, having access to prox oracles reduces the complexity and we present optimal methods based on smoothing that improve over methods using just gradient accesses.

****

A forward model at Purkinje cell synapses facilitates cerebellar anticipatory control

Ivan Herreros, Xerxes Arsiwalla, Paul Verschure

How does our motor system solve the problem of anticipatory control in spite of a wide spectrum of response dynamics from different musculo-skeletal systems, transport delays as well as response latencies throughout the central nervous system? To a great extent, our highly-skilled motor responses are a result of a reactive feedback system, originating in the brain-stem and spinal cord, combined with a feed-forward anticipatory system, that is adaptively fine-tuned by sensory experience and originates in the cerebellum. Based on that interaction we design the counterfactual predictive control (CFPC) architecture, an anticipatory adaptive motor control scheme in which a feed-forward module, based on the cerebellum, steers an error feedback controller with counterfactual error signals. Those are signals that trigger reactions as actual errors would, but that do not code for any current of forthcoming errors. In order to determine the optimal learning strategy, we derive a novel learning rule for the feed-forward module that involves an eligibility trace and operates at the synaptic level. In particular, our eligibility trace provides a mechanism beyond co-incidence detection in that it convolves a history of prior synaptic inputs with error signals. In the context of cerebellar physiology, this solution implies that Purkinje cell synapses should generate eligibility traces using a forward model of the system being controlled. From an engineering perspective, CFPC provides a general-purpose anticipatory control architecture equipped with a learning rule that exploits the full dynamics of the closed-loop system.

****

Verification Based Solution for Structured MAB Problems
Zohar S. Karnin

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.

****

SURGE: Surface Regularized Geometry Estimation from a Single Image
Peng Wang, Xiaohui Shen, Bryan Russell, Scott Cohen, Brian Price, Alan L. Yuille

This paper introduces an approach to regularize 2.5D surface normal and depth predictions at each pixel given a single input image. The approach infers and reasons about the underlying 3D planar surfaces depicted in the image to snap predicted normals and depths to inferred planar surfaces, all while maintaining fine detail within objects. Our approach comprises two components: (i) a fourstream convolutional neural network (CNN) where depths, surface normals, and likelihoods of planar region and planar boundary are predicted at each pixel, followed by (ii) a dense conditional random field (DCRF) that integrates the four predictions such that the normals and depths are compatible with each other and regularized by the planar region and planar boundary information. The DCRF is formulated such that gradients can be passed to the surface normal and depth CNNs via backpropagation. In addition, we propose new planar wise metrics to evaluate geometry consistency within planar surfaces, which are more tightly related to dependent 3D editing applications. We show that our regularization yields a 30% relative improvement in planar consistency on the NYU v2 dataset.

****

CliqueCNN: Deep Unsupervised Exemplar Learning
Miguel A. Bautista, Artsiom Sanakoyeu, Ekaterina Tikhoncheva, Bjorn Ommer

Exemplar learning is a powerful paradigm for discovering visual similarities in an unsupervised manner. In this context, however, the recent breakthrough in deep learning could not yet unfold its full potential. With only a single positive sample, a great imbalance between one positive and many negatives, and unreliable relationships between most samples, training of convolutional neural networks is impaired. Given weak estimates of local distance we propose a single optimization problem to extract batches of samples with mutually consistent relations. Conflicting relations are distributed over different batches and similar samples are grouped into compact cliques. Learning exemplar similarities is framed as a sequence of clique categorization tasks. The CNN then consolidates transitivity relations within and between cliques and learns a single representation for all samples without the need for labels. The proposed unsupervised approach has shown competitive performance on detailed posture analysis and object classification.

****

Computing and maximizing influence in linear threshold and triggering models
Justin T. Khim, Varun Jog, Po-Ling Loh

We establish upper and lower bounds for the influence of a set of nodes in certa in types of contagion models. We derive two sets of bounds, the first designed for linear threshold models, and the second more broadly applicable to a general class of triggering models, which subsumes the popular independent cascade models, as well. We quantify the gap between our upper and lower bounds in the case of the linear threshold model and illustrate the gains of our upper bounds for independent cascade models in relation to existing results. Importantly, our lower bounds are monotonic and submodular, implying that a greedy algorithm for influence maximization is guaranteed to produce a maximizer within a (1 - 1/e)-factor of the truth. Although the problem of exact influence computation is NP-hard in general, our bounds may be evaluated efficiently. This leads to an attractive, highly scalable algorithm for influence maximization with rigorous theoretical guarantees.

****

Data Programming: Creating Large Training Sets, Quickly
Alexander J. Ratner, Christopher M. De Sa, Sen Wu, Daniel Selsam, Christopher Ré

Large labeled training sets are the critical building blocks of supervised learning methods and are key enablers of deep learning techniques. For some applications, creating labeled training sets is the most time-consuming and expensive part of applying machine learning. We therefore propose a paradigm for the programmatic creation of training sets called data programming in which users provide a set of labeling functions, which are programs that heuristically label subsets of the data, but that are noisy and may conflict. By viewing these labeling functions as implicitly describing a generative model for this noise, we show that we can recover the parameters of this model to "denoise" the generated training set, and establish theoretically that we can recover the parameters of these generative models in a handful of settings. We then show how to modify a discriminative loss function to make it noise-aware, and demonstrate our method over a range of discriminative models including logistic regression and LSTMs. Experimentally, on the 2014 TAC-KBP Slot Filling challenge, we show that data programming would have led to a new winning score, and also show that applying data programming to an LSTM model leads to a TAC-KBP score almost 6 F1 points over a state-of-the-art LSTM baseline (and into second place in the competition). Additionally, in initial user studies we observed that data programming may be an easier way for non-experts to create machine learning models when training data is limited or unavailable.

****

Flexible Models for Microclustering with Application to Entity Resolution
Brenda Betancourt, Giacomo Zanella, Jeffrey W. Miller, Hanna Wallach, Abbas Zaidi, Rebecca C. Steorts

Most generative models for clustering implicitly assume that the number of data points in each cluster grows linearly with the total number of data points. Finite mixture models, Dirichlet process mixture models, and Pitman--Yor process mixture models make this assumption, as do all other infinitely exchangeable clustering models. However, for some applications, this assumption is inappropriate. For example, when performing entity resolution, the size of each cluster should be unrelated to the size of the data set, and each cluster should contain a negligible fraction of the total number of data points. These applications require models that yield clusters whose sizes grow sublinearly with the size of the data set. We address this requirement by defining the microclustering property and introducing a new class of models that can exhibit this property. We compare models within this class to two commonly used clustering models using four entity-resolution data sets.

****

Blind Regression: Nonparametric Regression for Latent Variable Models via Collaborative Filtering
Dogyoon Song, Christina E. Lee, Yihua Li, Devavrat Shah

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.

****

An ensemble diversity approach to supervised binary hashing

Miguel A. Carreira-Perpinan, Ramin Raziperchikolaei

Binary hashing is a well-known approach for fast approximate nearest-neighbor search in information retrieval. Much work has focused on affinity-based objective functions involving the hash functions or binary codes. These objective functions encode neighborhood information between data points and are often inspired by manifold learning algorithms. They ensure that the hash functions differ from each other through constraints or penalty terms that encourage codes to be orthogonal or dissimilar across bits, but this couples the binary variables and complicates the already difficult optimization. We propose a much simpler approach: we train each hash function (or bit) independently from each other, but introduce diversity among them using techniques from classifier ensembles. Surprisingly, we find that not only is this faster and trivially parallelizable, but it also improves over the more complex, coupled objective function, and achieves state-of-the-art precision and recall in experiments with image retrieval.

****

Learning Influence Functions from Incomplete Observations

Xinran He, Ke Xu, David Kempe, Yan Liu

We study the problem of learning influence functions under incomplete observations of node activations. Incomplete observations are a major concern as most (online and real-world) social networks are not fully observable. We establish both proper and improper PAC learnability of influence functions under randomly missing observations. Proper PAC learnability under the Discrete-Time Linear Threshold (DLT) and Discrete-Time Independent Cascade (DIC) models is established by reducing incomplete observations to complete observations in a modified graph. Our improper PAC learnability result applies for the DLT and DIC models as well as the Continuous-Time Independent Cascade (CIC) model. It is based on a parametrization in terms of reachability features, and also gives rise to an efficient and practical heuristic. Experiments on synthetic and real-world datasets demonstrate the ability of our method to compensate even for a fairly large fraction of missing observations.

****

Backprop KF: Learning Discriminative Deterministic State Estimators

Tuomas Haarnoja, Anurag Ajay, Sergey Levine, Pieter Abbeel

Generative state estimators based on probabilistic filters and smoothers are one of the most popular classes of state estimators for robots and autonomous vehicles. However, generative models have limited capacity to handle rich sensory observations, such as camera images, since they must model the entire distribution over sensor readings. Discriminative models do not suffer from this limitation, but are typically more complex to train as latent variable models for state estimation. We present an alternative approach where the parameters of the latent state distribution are directly optimized as a deterministic computation graph, resulting in a simple and effective gradient descent algorithm for training discriminative state estimators. We show that this procedure can be used to train state estimators that use complex input, such as raw camera images, which must be processed using expressive nonlinear function approximators such as convolutional neural networks. Our model can be viewed as a type of recurrent neural network, and the connection to probabilistic filtering allows us to design a network architecture that is particularly well suited for state estimation. We evaluate our approach on synthetic tracking task with raw image inputs and on the visual odometry task in the KITTI dataset. The results show significant improvement over both standard generative approaches and regular recurrent neural networks.

****

On the Recursive Teaching Dimension of VC Classes

Xi Chen, Xi Chen, Yu Cheng, Bo Tang

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.

****

Generalized Correspondence-LDA Models (GC-LDA) for Identifying Functional Regions in the Brain

Timothy Rubin, Oluwasanmi O. Koyejo, Michael N. Jones, Tal Yarkoni

This paper presents Generalized Correspondence-LDA (GC-LDA), a generalization of the Correspondence-LDA model that allows for variable spatial representations to be associated with topics, and increased flexibility in terms of the strength of the correspondence between data types induced by the model. We present three variants of GC-LDA, each of which associates topics with a different spatial representation, and apply them to a corpus of neuroimaging data. In the context of this dataset, each topic corresponds to a functional brain region, where the region's spatial extent is captured by a probability distribution over neural activity, and the region's cognitive function is captured by a probability distribution over linguistic terms. We illustrate the qualitative improvements offered by GC-LDA in terms of the types of topics extracted with alternative spatial representations, as well as the model's ability to incorporate a-priori knowledge from the neuroimaging literature. We furthermore demonstrate that the novel features of GC-LDA improve predictions for missing data.

****

Fast ε-free Inference of Simulation Models with Bayesian Conditional Density Estimation

George Papamakarios, Iain Murray

Many statistical models can be simulated forwards but have intractable likelihoods. Approximate Bayesian Computation (ABC) methods are used to infer properties of these models from data. Traditionally these methods approximate the posterior over parameters by conditioning on data being inside an ε-ball around the observed data, which is only correct in the limit $\varepsilon \to 0$. Monte Carlo methods can then draw samples from the approximate posterior to approximate predictions or error bars on parameters. These algorithms critically slow down as $\varepsilon \to 0$, and in practice draw samples from a broader distribution than the posterior. We propose a new approach to likelihood-free inference based on Bayesian conditional density estimation. Preliminary inferences based on limited simulation data are used to guide later simulations. In some cases, learning an accurate parametric representation of the entire true posterior distribution requires fewer model simulations than Monte Carlo ABC methods need to produce a single sample from an approximate posterior.

****

Ladder Variational Autoencoders

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

Variational autoencoders are powerful models for unsupervised learning. However deep models with several layers of dependent stochastic variables are difficult to train which limits the improvements obtained using these highly expressive models. We propose a new inference model, the Ladder Variational Autoencoder, that recursively corrects the generative distribution by a data dependent approximate likelihood in a process resembling the recently proposed Ladder Network. We show that this model provides state of the art predictive log-likelihood and tighter log-likelihood lower bound compared to the purely bottom-up inference in layered Variational Autoencoders and other generative models. We provide a detailed analysis of the learned hierarchical latent representation and show that our new inference model is qualitatively different and utilizes a deeper more distributed hierarchy of latent variables. Finally, we observe that batch-normalization and deterministic warm-up (gradually turning on the KL-term) are crucial for training variational models with many stochastic layers.

****

Improved Deep Metric Learning with Multi-class N-pair Loss Objective

Kihyuk Sohn

Deep metric learning has gained much popularity in recent years, following the success of deep learning. However, existing frameworks of deep metric learning based on contrastive loss and triplet loss often suffer from slow convergence, par tially because they employ only one negative example while not interacting with the other negative classes in each update. In this paper, we propose to address this problem with a new metric learning objective called multi-class N-pair loss . The proposed objective function firstly generalizes triplet loss by allowing joint comparison among more than one negative examples – more specifically, N-1 negative examples – and secondly reduces the computational burden of evaluating deep embedding vectors via an efficient batch construction strategy using only N pairs of examples, instead of (N+1)×N. We demonstrate the superiority of our proposed loss to the triplet loss as well as other competing loss functions for a variety of tasks on several visual recognition benchmark, including fine-grained object recognition and verification, image clustering and retrieval, and face verification and identification.

****

Learning Sparse Gaussian Graphical Models with Overlapping Blocks
Mohammad Javad Hosseini, Su-In Lee
We present a novel framework, called GRAB (GRaphical models with overlApping Blocks), to capture densely connected components in a network estimate. GRAB takes as input a data matrix of p variables and n samples, and jointly learns both a network among p variables and densely connected groups of variables (called `blocks'). GRAB has four major novelties as compared to existing network estimation methods: 1) It does not require the blocks to be given a priori. 2) Blocks can overlap. 3) It can jointly learn a network structure and overlapping blocks. 4) It solves a joint optimization problem with the block coordinate descent method that is convex in each step. We show that GRAB reveals the underlying network structure substantially better than four state-of-the-art competitors on synthetic data. When applied to cancer gene expression data, GRAB outperforms its competitors in revealing known functional gene sets and potentially novel genes that drive cancer.

****

Probabilistic Inference with Generating Functions for Poisson Latent Variable Models
Kevin Winner, Daniel R. Sheldon
Graphical models with latent count variables arise in a number of fields. Standard exact inference techniques such as variable elimination and belief propagation do not apply to these models because the latent variables have countably infinite support. As a result, approximations such as truncation or MCMC are employed . We present the first exact inference algorithms for a class of models with latent count variables by developing a novel representation of countably infinite factors as probability generating functions, and then performing variable elimination with generating functions. Our approach is exact, runs in pseudo-polynomial time, and is much faster than existing approximate techniques. It leads to better parameter estimates for problems in population ecology by avoiding error introduced by approximate likelihood computations.

****

Achieving the KS threshold in the general stochastic block model with linearized acyclic belief propagation
Emmanuel Abbe, Colin Sandon
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 Unified Approach for Learning the Parameters of Sum-Product Networks
Han Zhao, Pascal Poupart, Geoffrey J. Gordon
We present a unified approach for learning the parameters of Sum-Product networks (SPNs). We prove that any complete and decomposable SPN is equivalent to a mixture of trees where each tree corresponds to a product of univariate distributions. Based on the mixture model perspective, we characterize the objective function when learning SPNs based on the maximum likelihood estimation (MLE) principle and show that the optimization problem can be formulated as a signomial program

. We construct two parameter learning algorithms for SPNs by using sequential monomial approximations (SMA) and the concave-convex procedure (CCCP), respectively. The two proposed methods naturally admit multiplicative updates, hence effectively avoiding the projection operation. With the help of the unified framework, we also show that, in the case of SPNs, CCCP leads to the same algorithm as Expectation Maximization (EM) despite the fact that they are different in general.
****
The Multiscale Laplacian Graph Kernel
Risi Kondor, Horace Pan
Many real world graphs, such as the graphs of molecules, exhibit structure at multiple different scales, but most existing kernels between graphs are either purely local or purely global in character. In contrast, by building a hierarchy of nested subgraphs, the Multiscale Laplacian Graph kernels (MLG kernels) that we define in this paper can account for structure at a range of different scales. At the heart of the MLG construction is another new graph kernel, called the Feature Space Laplacian Graph kernel (FLG kernel), which has the property that it can lift a base kernel defined on the vertices of two graphs to a kernel between the graphs. The MLG kernel applies such FLG kernels to subgraphs recursively. To make the MLG kernel computationally feasible, we also introduce a randomized projection procedure, similar to the Nystro ■m method, but for RKHS operators.
****
Learning the Number of Neurons in Deep Networks
Jose M. Alvarez, Mathieu Salzmann
Nowadays, the number of layers and of neurons in each layer of a deep network are typically set manually. While very deep and wide networks have proven effective in general, they come at a high memory and computation cost, thus making them impractical for constrained platforms. These networks, however, are known to have many redundant parameters, and could thus, in principle, be replaced by more compact architectures. In this paper, we introduce an approach to automatically determining the number of neurons in each layer of a deep network during learning. To this end, we propose to make use of a group sparsity regularizer on the parameters of the network, where each group is defined to act on a single neuron. Starting from an overcomplete network, we show that our approach can reduce the number of parameters by up to 80\% while retaining or even improving the network accuracy.
****
Deep Alternative Neural Network: Exploring Contexts as Early as Possible for Action Recognition
Jinzhuo Wang, Wenmin Wang, xiongtao Chen, Ronggang Wang, Wen Gao
Contexts are crucial for action recognition in video. Current methods often mine contexts after extracting hierarchical local features and focus on their high-order encodings. This paper instead explores contexts as early as possible and leverages their evolutions for action recognition. In particular, we introduce a novel architecture called deep alternative neural network (DANN) stacking alternative layers. Each alternative layer consists of a volumetric convolutional layer followed by a recurrent layer. The former acts as local feature learner while the latter is used to collect contexts. Compared with feed-forward neural networks, DANN learns contexts of local features from the very beginning. This setting helps to preserve hierarchical context evolutions which we show are essential to recognize similar actions. Besides, we present an adaptive method to determine the temporal size for network input based on optical flow energy, and develop a volumetric pyramid pooling layer to deal with input clips of arbitrary sizes. We demonstrate the advantages of DANN on two benchmarks HMDB51 and UCF101 and report competitive or superior results to the state-of-the-art.
****
Online ICA: Understanding Global Dynamics of Nonconvex Optimization via Diffusion Processes
Chris Junchi Li, Zhaoran Wang, Han Liu
Solving statistical learning problems often involves nonconvex optimization. Despite the empirical success of nonconvex statistical optimization methods, their

global dynamics, especially convergence to the desirable local minima, remain less well understood in theory. In this paper, we propose a new analytic paradigm based on diffusion processes to characterize the global dynamics of nonconvex statistical optimization. As a concrete example, we study stochastic gradient descent (SGD) for the tensor decomposition formulation of independent component analysis. In particular, we cast different phases of SGD into diffusion processes, i.e., solutions to stochastic differential equations. Initialized from an unstable equilibrium, the global dynamics of SGD transit over three consecutive phases: (i) an unstable Ornstein-Uhlenbeck process slowly departing from the initialization, (ii) the solution to an ordinary differential equation, which quickly evolves towards the desirable local minimum, and (iii) a stable Ornstein-Uhlenbeck process oscillating around the desirable local minimum. Our proof techniques are based upon Stroock and Varadhan’s weak convergence of Markov chains to diffusion processes, which are of independent interest.

****

Spatio-Temporal Hilbert Maps for Continuous Occupancy Representation in Dynamic Environments

Ransalu Senanayake, Lionel Ott, Simon O'Callaghan, Fabio T. Ramos

We consider the problem of building continuous occupancy representations in dynamic environments for robotics applications. The problem has hardly been discussed previously due to the complexity of patterns in urban environments, which have both spatial and temporal dependencies. We address the problem as learning a kernel classifier on an efficient feature space. The key novelty of our approach is the incorporation of variations in the time domain into the spatial domain. We propose a method to propagate motion uncertainty into the kernel using a hierarchical model. The main benefit of this approach is that it can directly predict the occupancy state of the map in the future from past observations, being a valuable tool for robot trajectory planning under uncertainty. Our approach preserves the main computational benefits of static Hilbert maps — using stochastic gradient descent for fast optimization of model parameters and incremental updates as new data are captured. Experiments conducted in road intersections of an urban environment demonstrated that spatio-temporal Hilbert maps can accurately model changes in the map while outperforming other techniques on various aspects.

****

CRF-CNN: Modeling Structured Information in Human Pose Estimation

Xiao Chu, Wanli Ouyang, hongsheng Li, Xiaogang Wang

Deep convolutional neural networks (CNN) have achieved great success. On the other hand, modeling structural information has been proved critical in many vision problems. It is of great interest to integrate them effectively. In a classical neural network, there is no message passing between neurons in the same layer. In this paper, we propose a CRF-CNN framework which can simultaneously model structural information in both output and hidden feature layers in a probabilistic way, and it is applied to human pose estimation. A message passing scheme is proposed, so that in various layers each body joint receives messages from all the others in an efficient way. Such message passing can be implemented with convolution between features maps in the same layer, and it is also integrated with feedforward propagation in neural networks. Finally, a neural network implementation of end-to-end learning CRF-CNN is provided. Its effectiveness is demonstrated through experiments on two benchmark datasets.

****

Bayesian latent structure discovery from multi-neuron recordings

Scott Linderman, Ryan P. Adams, Jonathan W. Pillow

Neural circuits contain heterogeneous groups of neurons that differ in type, location, connectivity, and basic response properties. However, traditional methods for dimensionality reduction and clustering are ill-suited to recovering the structure underlying the organization of neural circuits. In particular, they do not take advantage of the rich temporal dependencies in multi-neuron recordings and fail to account for the noise in neural spike trains. Here we describe new tools for inferring latent structure from simultaneously recorded spike train data using a hierarchical extension of a multi-neuron point process model commonly known as the generalized linear model (GLM). Our approach combines the GLM with flexible graph-theoretic priors governing the relationship between latent features and neural connectivity patterns. Fully Bayesian inference via Pólya-gamma augmentation of the resulting model allows us to classify neurons and infer latent dimensions of circuit organization from correlated spike trains. We demonstrate the effectiveness of our method with applications to synthetic data and multi-neuron recordings in primate retina, revealing latent patterns of neural types and locations from spike trains alone.

****

Latent Attention For If-Then Program Synthesis

Chang Liu, Xinyun Chen, Eui Chul Shin, Mingcheng Chen, Dawn Song

Automatic translation from natural language descriptions into programs is a long-standing challenging problem. In this work, we consider a simple yet important sub-problem: translation from textual descriptions to If-Then programs. We devise a novel neural network architecture for this task which we train end-to-end. Specifically, we introduce Latent Attention, which computes multiplicative weights for the words in the description in a two-stage process with the goal of better leveraging the natural language structures that indicate the relevant parts for predicting program elements. Our architecture reduces the error rate by 28.57% compared to prior art. We also propose a one-shot learning scenario of If-Then program synthesis and simulate it with our existing dataset. We demonstrate a variation on the training procedure for this scenario that outperforms the original procedure, significantly closing the gap to the model trained with all data.

****

Understanding Probabilistic Sparse Gaussian Process Approximations

Matthias Bauer, Mark van der Wilk, Carl Edward Rasmussen

Good sparse approximations are essential for practical inference in Gaussian Processes as the computational cost of exact methods is prohibitive for large datasets. The Fully Independent Training Conditional (FITC) and the Variational Free Energy (VFE) approximations are two recent popular methods. Despite superficial similarities, these approximations have surprisingly different theoretical properties and behave differently in practice. We thoroughly investigate the two methods for regression both analytically and through illustrative examples, and draw conclusions to guide practical application.

****

Wasserstein Training of Restricted Boltzmann Machines

Grégoire Montavon, Klaus-Robert Müller, Marco Cuturi

Boltzmann machines are able to learn highly complex, multimodal, structured and multiscale real-world data distributions. Parameters of the model are usually learned by minimizing the Kullback-Leibler (KL) divergence from training samples to the learned model. We propose in this work a novel approach for Boltzmann machine training which assumes that a meaningful metric between observations is known. This metric between observations can then be used to define the Wasserstein distance between the distribution induced by the Boltzmann machine on the one hand, and that given by the training sample on the other hand. We derive a gradient of that distance with respect to the model parameters. Minimization of this new objective leads to generative models with different statistical properties. We demonstrate their practical potential on data completion and denoising, for which the metric between observations plays a crucial role.

****

A primal-dual method for conic constrained distributed optimization problems

Necdet Serhat Aybat, Erfan Yazdandoost Hamedani

We consider cooperative multi-agent consensus optimization problems over an undirected network of agents, where only those agents connected by an edge can directly communicate. The objective is to minimize the sum of agent-specific composite convex functions over agent-specific private conic constraint sets; hence, the optimal consensus decision should lie in the intersection of these private sets. We provide convergence rates in sub-optimality, infeasibility and consensus violation; examine the effect of underlying network topology on the convergence ra tes of the proposed decentralized algorithms; and show how to extend these methods to handle time-varying communication networks.

****

Path-Normalized Optimization of Recurrent Neural Networks with ReLU Activations

Behnam Neyshabur, Yuhuai Wu, Russ R. Salakhutdinov, Nati Srebro

We investigate the parameter-space geometry of recurrent neural networks (RNNs), and develop an adaptation of path-SGD optimization method, attuned to this geometry, that can learn plain RNNs with ReLU activations. On several datasets that require capturing long-term dependency structure, we show that path-SGD can significantly improve trainability of ReLU RNNs compared to RNNs trained with SGD, even with various recently suggested initialization schemes.

****

Communication-Optimal Distributed Clustering

Jiecao Chen, He Sun, David Woodruff, Qin Zhang

Clustering large datasets is a fundamental problem with a number of applications in machine learning. Data is often collected on different sites and clustering needs to be performed in a distributed manner with low communication. We would like the quality of the clustering in the distributed setting to match that in the centralized setting for which all the data resides on a single site. In this work, we study both graph and geometric clustering problems in two distributed models: (1) a point-to-point model, and (2) a model with a broadcast channel. We give protocols in both models which we show are nearly optimal by proving almost matching communication lower bounds. Our work highlights the surprising power of a broadcast channel for clustering problems; roughly speaking, to cluster n points or n vertices in a graph distributed across s servers, for a worst-case partitioning the communication complexity in a point-to-point model is n*s, while in the broadcast model it is n + s. We implement our algorithms and demonstrate this phenomenon on real life datasets, showing that our algorithms are also very efficient in practice.

****

Boosting with Abstention

Corinna Cortes, Giulia DeSalvo, Mehryar Mohri

We present a new boosting algorithm for the key scenario of binary classification with abstention where the algorithm can abstain from predicting the label of a point, at the price of a fixed cost. At each round, our algorithm selects a pair of functions, a base predictor and a base abstention function. We define convex upper bounds for the natural loss function associated to this problem, which we prove to be calibrated with respect to the Bayes solution. Our algorithm benefits from general margin-based learning guarantees which we derive for ensembles of pairs of base predictor and abstention functions, in terms of the Rademacher complexities of the corresponding function classes. We give convergence guarantees for our algorithm along with a linear-time weak-learning algorithm for abstention stumps. We also report the results of several experiments suggesting that our algorithm provides a significant improvement in practice over two confidence-based algorithms.

****

Linear dynamical neural population models through nonlinear embeddings

Yuanjun Gao, Evan W. Archer, Liam Paninski, John P. Cunningham

A body of recent work in modeling neural activity focuses on recovering low- dimensional latent features that capture the statistical structure of large-scale neural populations. Most such approaches have focused on linear generative models, where inference is computationally tractable. Here, we propose fLDS, a general class of nonlinear generative models that permits the firing rate of each neuron to vary as an arbitrary smooth function of a latent, linear dynamical state. This extra flexibility allows the model to capture a richer set of neural variability than a purely linear model, but retains an easily visualizable low-dimensional latent space. To fit this class of non-conjugate models we propose a variational inference scheme, along with a novel approximate posterior capable of capturing rich temporal correlations across time. We show that our techniques permit inference in a wide class of generative models.We also show in application to tw o neural datasets that, compared to state-of-the-art neural population models, fLDS captures a much larger proportion of neural variability with a small number of latent dimensions, providing superior predictive performance and interpretability.

****

Rényi Divergence Variational Inference

Yingzhen Li, Richard E. Turner

This paper introduces the variational Rényi bound (VR) that extends traditional variational inference to Rényi's alpha-divergences. This new family of variational methods unifies a number of existing approaches, and enables a smooth interpolation from the evidence lower-bound to the log (marginal) likelihood that is controlled by the value of alpha that parametrises the divergence. The reparameterization trick, Monte Carlo approximation and stochastic optimisation methods are deployed to obtain a tractable and unified framework for optimisation. We further consider negative alpha values and propose a novel variational inference method as a new special case in the proposed framework. Experiments on Bayesian neural networks and variational auto-encoders demonstrate the wide applicability of the VR bound.

****

Stochastic Gradient Geodesic MCMC Methods

Chang Liu, Jun Zhu, Yang Song

We propose two stochastic gradient MCMC methods for sampling from Bayesian posterior distributions defined on Riemann manifolds with a known geodesic flow, e.g. hyperspheres. Our methods are the first scalable sampling methods on these manifolds, with the aid of stochastic gradients. Novel dynamics are conceived and 2nd-order integrators are developed. By adopting embedding techniques and the geodesic integrator, the methods do not require a global coordinate system of the manifold and do not involve inner iterations. Synthetic experiments show the validity of the method, and its application to the challenging inference for spherical topic models indicate practical usability and efficiency.

****

A posteriori error bounds for joint matrix decomposition problems

Nicolo Colombo, Nikos Vlassis

Joint matrix triangularization is often used for estimating the joint eigenstructure of a set M of matrices, with applications in signal processing and machine learning. We consider the problem of approximate joint matrix triangularization when the matrices in M are jointly diagonalizable and real, but we only observe a set M' of noise perturbed versions of the matrices in M. Our main result is a first-order upper bound on the distance between any approximate joint triangularizer of the matrices in M' and any exact joint triangularizer of the matrices in M. The bound depends only on the observable matrices in M' and the noise level. In particular, it does not depend on optimization specific properties of the triangularizer, such as its proximity to critical points, that are typical of existing bounds in the literature. To our knowledge, this is the first a posteriori bound for joint matrix decomposition. We demonstrate the bound on synthetic data for which the ground truth is known.

****

Global Analysis of Expectation Maximization for Mixtures of Two Gaussians

Ji Xu, Daniel J. Hsu, Arian Maleki

Expectation Maximization (EM) is among the most popular algorithms for estimating parameters of statistical models. However, EM, which is an iterative algorithm based on the maximum likelihood principle, is generally only guaranteed to find stationary points of the likelihood objective, and these points may be far from any maximizer. This article addresses this disconnect between the statistical principles behind EM and its algorithmic properties. Specifically, it provides a global analysis of EM for specific models in which the observations comprise an i.i.d. sample from a mixture of two Gaussians. This is achieved by (i) studying the sequence of parameters from idealized execution of EM in the infinite sample limit, and fully characterizing the limit points of the sequence in terms of the initial parameters; and then (ii) based on this convergence analysis, esta blishing statistical consistency (or lack thereof) for the actual sequence of parameters produced by EM.

****

Stochastic Structured Prediction under Bandit Feedback

Artem Sokolov, Julia Kreutzer, Stefan Riezler, Christopher Lo

Stochastic structured prediction under bandit feedback follows a learning protocol where on each of a sequence of iterations, the learner receives an input, predicts an output structure, and receives partial feedback in form of a task loss evaluation of the predicted structure. We present applications of this learning scenario to convex and non-convex objectives for structured prediction and analyze them as stochastic first-order methods. We present an experimental evaluation on problems of natural language processing over exponential output spaces, and compare convergence speed across different objectives under the practical criterion of optimal task performance on development data and the optimization-theoretic criterion of minimal squared gradient norm. Best results under both criteria are obtained for a non-convex objective for pairwise preference learning under bandit feedback.

****

Estimating the class prior and posterior from noisy positives and unlabeled data

Shantanu Jain, Martha White, Predrag Radivojac

We develop a classification algorithm for estimating posterior distributions from positive-unlabeled data, that is robust to noise in the positive labels and effective for high-dimensional data. In recent years, several algorithms have been proposed to learn from positive-unlabeled data; however, many of these contributions remain theoretical, performing poorly on real high-dimensional data that is typically contaminated with noise. We build on this previous work to develop two practical classification algorithms that explicitly model the noise in the positive labels and utilize univariate transforms built on discriminative classifiers. We prove that these univariate transforms preserve the class prior, enabling estimation in the univariate space and avoiding kernel density estimation for high-dimensional data. The theoretical development and parametric and nonparametric algorithms proposed here constitute an important step towards wide-spread use of robust classification algorithms for positive-unlabeled data.

****

A Minimax Approach to Supervised Learning

Farzan Farnia, David Tse

Given a task of predicting Y from X, a loss function L, and a set of probability distributions Gamma on (X,Y), what is the optimal decision rule minimizing the worst-case expected loss over Gamma? In this paper, we address this question by introducing a generalization of the maximum entropy principle. Applying this principle to sets of distributions with marginal on X constrained to be the empirical marginal, we provide a minimax interpretation of the maximum likelihood problem over generalized linear models as well as some popular regularization schemes. For quadratic and logarithmic loss functions we revisit well-known linear and logistic regression models. Moreover, for the 0-1 loss we derive a classifier which we call the minimax SVM. The minimax SVM minimizes the worst-case expected 0-1 loss over the proposed Gamma by solving a tractable optimization problem. We perform several numerical experiments to show the power of the minimax SVM in outperforming the SVM.

****

Blazing the trails before beating the path: Sample-efficient Monte-Carlo planning

Jean-Bastien Grill, Michal Valko, Remi Munos

We study the sampling-based planning problem in Markov decision processes (MDPs) that we can access only through a generative model, usually referred to as Monte-Carlo planning. Our objective is to return a good estimate of the optimal value function at any state while minimizing the number of calls to the generative model, i.e. the sample complexity. We propose a new algorithm, TrailBlazer, able to handle MDPs with a finite or an infinite number of transitions from state-action to next states. TrailBlazer is an adaptive algorithm that exploits possible structures of the MDP by exploring only a subset of states reachable by following near-optimal policies. We provide bounds on its sample complexity that depend on a measure of the quantity of near-optimal states. The algorithm behavior can be considered as an extension of Monte-Carlo sampling (for estimating an expectation) to problems that alternate maximization (over actions) and expectation (over next states). Finally, another appealing feature of TrailBlazer is that it is simple to implement and computationally efficient.

****

Scaling Factorial Hidden Markov Models: Stochastic Variational Inference without Messages

Yin Cheng Ng, Pawel M. Chilinski, Ricardo Silva

Factorial Hidden Markov Models (FHMMs) are powerful models for sequential data but they do not scale well with long sequences. We propose a scalable inference and learning algorithm for FHMMs that draws on ideas from the stochastic variational inference, neural network and copula literatures. Unlike existing approaches, the proposed algorithm requires no message passing procedure among latent variables and can be distributed to a network of computers to speed up learning. Our experiments corroborate that the proposed algorithm does not introduce further approximation bias compared to the proven structured mean-field algorithm, and achieves better performance with long sequences and large FHMMs.

****

Improved Dropout for Shallow and Deep Learning

Zhe Li, Boqing Gong, Tianbao Yang

Dropout has been witnessed with great success in training deep neural networks by independently   zeroing out the outputs of neurons at random. It has also received a surge of interest for shallow learning, e.g., logistic regression. However, the independent sampling for dropout could be suboptimal for the sake of convergence. In this paper, we propose to use multinomial sampling for dropout, i.e., sampling features or neurons according to a multinomial distribution with different probabilities for different features/neurons. To exhibit the optimal dropout probabilities, we analyze the shallow learning with multinomial dropout and establish the risk bound for stochastic optimization. By minimizing a sampling dependent factor in the risk bound, we obtain a distribution-dependent dropout with sampling probabilities dependent on the second order statistics of the data distribution. To tackle the issue of evolving distribution of neurons in deep learning, we propose an efficient adaptive dropout (named \textbf{evolutional dropout}) that computes the sampling probabilities on-the-fly from a mini-batch of examples. Empirical studies on several benchmark datasets demonstrate that the proposed dropouts achieve not only much faster convergence and but also a smaller testing error than the standard dropout. For example, on the CIFAR-100 data, the evolutional dropout achieves relative improvements over 10\% on the prediction performance and over 50\% on the convergence speed compared to the standard dropout.

****

Clustering Signed Networks with the Geometric Mean of Laplacians

Pedro Mercado, Francesco Tudisco, Matthias Hein

Signed networks allow to model positive and negative relationships. We analyze existing extensions of spectral clustering to signed networks. It turns out that existing approaches do not recover the ground truth clustering in several situations where either the positive or the negative network structures contain no noise. Our analysis shows that these problems arise as existing approaches take some form of arithmetic mean of the Laplacians of the positive and negative part. As a solution we propose to use the geometric mean of the Laplacians of positive and negative part and show that it outperforms the existing approaches. While the geometric mean of matrices is computationally expensive, we show that eigenvectors of the geometric mean can be computed efficiently, leading to a numerical scheme for sparse matrices which is of independent interest.

****

Consistent Estimation of Functions of Data Missing Non-Monotonically and Not at Random

Ilya Shpitser

Missing records are a perennial problem in analysis of complex data of all types, when the target of inference is some function of the full data law. In simple cases, where data is missing at random or completely at random (Rubin, 1976), well-known adjustments exist that result in consistent estimators of target quantities. Assumptions underlying these estimators are generally not realistic in practical missing data problems. Unfortunately, consistent estimators in more complex cases where data is missing not at random, and where no ordering on variables induces monotonicity of missingness status are not known in general, with some notable exceptions (Robins, 1997), (Tchetgen Tchetgen et al, 2016), (Sadinle and Reiter, 2016). In this paper, we propose a general class of consistent estimators for cases where data is missing not at random, and missingness status is non-monotonic. Our estimators, which are generalized inverse probability weighting estimators, make no assumptions on the underlying full data law, but instead place independence restrictions, and certain other fairly mild assumptions, on the distribution of missingness status conditional on the data. The assumptions we place on the distribution of missingness status conditional on the data can be viewed as a version of a conditional Markov random field (MRF) corresponding to a chain graph. Assumptions embedded in our model permit identification from the observed data law, and admit a natural fitting procedure based on the pseudo likelihood approach of (Besag, 1975). We illustrate our approach with a simple simulation study, and an analysis of risk of premature birth in women in Botswana exposed to highly active anti-retroviral therapy.

****

Discriminative Gaifman Models

Mathias Niepert

We present discriminative Gaifman models, a novel family of relational machine learning models. Gaifman models learn feature representations bottom up from representations of locally connected and bounded-size regions of knowledge bases (KBs). Considering local and bounded-size neighborhoods of knowledge bases renders logical inference and learning tractable, mitigates the problem of overfitting, and facilitates weight sharing. Gaifman models sample neighborhoods of knowledge bases so as to make the learned relational models more robust to missing objects and relations which is a common situation in open-world KBs. We present the core ideas of Gaifman models and apply them to large-scale relational learning problems. We also discuss the ways in which Gaifman models relate to some existing relational machine learning approaches.

****

Selective inference for group-sparse linear models

Fan Yang, Rina Foygel Barber, Prateek Jain, John Lafferty

We develop tools for selective inference in the setting of group sparsity, including the construction of confidence intervals and p-values for testing selected groups of variables. Our main technical result gives the precise distribution of the magnitude of the projection of the data onto a given subspace, and enables us to develop inference procedures for a broad class of group-sparse selection methods, including the group lasso, iterative hard thresholding, and forward stepwise regression. We give numerical results to illustrate these tools on simulated data and on health record data.

****

InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, Pieter Abbeel

This paper describes InfoGAN, an information-theoretic extension to the Generative Adversarial Network that is able to learn disentangled representations in a completely unsupervised manner. InfoGAN is a generative adversarial network that also maximizes the mutual information between a small subset of the latent variables and the observation. We derive a lower bound to the mutual information objective that can be optimized efficiently, and show that our training procedure can be interpreted as a variation of the Wake-Sleep algorithm. Specifically, InfoGAN successfully disentangles writing styles from digit shapes on the MNIST datas et, pose from lighting of 3D rendered images, and background digits from the central digit on the SVHN dataset. It also discovers visual concepts that include hair styles, presence/absence of eyeglasses, and emotions on the CelebA face dataset. Experiments show that InfoGAN learns interpretable representations that are competitive with representations learned by existing fully supervised methods.

****

Automated scalable segmentation of neurons from multispectral images

Uygar Sümbül, Douglas Roossien, Dawen Cai, Fei Chen, Nicholas Barry, John P. Cunningham, Edward Boyden, Liam Paninski

Reconstruction of neuroanatomy is a fundamental problem in neuroscience. Stochastic expression of colors in individual cells is a promising tool, although its use in the nervous system has been limited due to various sources of variability in expression. Moreover, the intermingled anatomy of neuronal trees is challenging for existing segmentation algorithms. Here, we propose a method to automate the segmentation of neurons in such (potentially pseudo-colored) images. The method uses spatio-color relations between the voxels, generates supervoxels to reduce the problem size by four orders of magnitude before the final segmentation, and is parallelizable over the supervoxels. To quantify performance and gain insight, we generate simulated images, where the noise level and characteristics, the density of expression, and the number of fluorophore types are variable. We also present segmentations of real Brainbow images of the mouse hippocampus, which reveal many of the dendritic segments.

****

Robustness of classifiers: from adversarial to random noise

Alhussein Fawzi, Seyed-Mohsen Moosavi-Dezfooli, Pascal Frossard

Several recent works have shown that state-of-the-art classifiers are vulnerable to worst-case (i.e., adversarial) perturbations of the datapoints. On the other hand, it has been empirically observed that these same classifiers are relatively robust to random noise. In this paper, we propose to study a semi-random noise regime that generalizes both the random and worst-case noise regimes. We propose the first quantitative analysis of the robustness of nonlinear classifiers in this general noise regime. We establish precise theoretical bounds on the robustness of classifiers in this general regime, which depend on the curvature of the classifier's decision boundary. Our bounds confirm and quantify the empirical observations that classifiers satisfying curvature constraints are robust to random noise. Moreover, we quantify the robustness of classifiers in terms of the subspace dimension in the semi-random noise regime, and show that our bounds remarkably interpolate between the worst-case and random noise regimes. We perform experiments and show that the derived bounds provide very accurate estimates when applied to various state-of-the-art deep neural networks and datasets. This result suggests bounds on the curvature of the classifiers' decision boundaries that we support experimentally, and more generally offers important insights onto the geometry of high dimensional classification problems.

****

Composing graphical models with neural networks for structured representations and fast inference

Matthew J. Johnson, David K. Duvenaud, Alex Wiltschko, Ryan P. Adams, Sandeep R. Datta

We propose a general modeling and inference framework that combines the complementary strengths of probabilistic graphical models and deep learning methods. Our model family composes latent graphical models with neural network observation likelihoods. For inference, we use recognition networks to produce local evidence potentials, then combine them with the model distribution using efficient message-passing algorithms. All components are trained simultaneously with a single stochastic variational inference objective. We illustrate this framework by automatically segmenting and categorizing mouse behavior from raw depth video, and demonstrate several other example models.

****

SoundNet: Learning Sound Representations from Unlabeled Video

Yusuf Aytar, Carl Vondrick, Antonio Torralba

We learn rich natural sound representations by capitalizing on large amounts of unlabeled sound data collected in the wild. We leverage the natural synchronization between vision and sound to learn an acoustic representation using two-million unlabeled videos. Unlabeled video has the advantage that it can be economically acquired at massive scales, yet contains useful signals about natural sound. We propose a student-teacher training procedure which transfers discriminative visual knowledge from well established visual recognition models into the sound modality using unlabeled video as a bridge. Our sound representation yields significant performance improvements over the state-of-the-art results on standard benchmarks for acoustic scene/object classification. Visualizations suggest some high-level semantics automatically emerge in the sound network, even though it is trained without ground truth labels.

****

Dual Decomposed Learning with Factorwise Oracle for Structural SVM of Large Output Domain

Ian En-Hsu Yen, Xiangru Huang, Kai Zhong, Ruohan Zhang, 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.

****

Deep Learning for Predicting Human Strategic Behavior

Jason S. Hartford, James R. Wright, Kevin Leyton-Brown

Predicting the behavior of human participants in strategic settings is an important problem in many domains. Most existing work either assumes that participants are perfectly rational, or attempts to directly model each participant's cognitive processes based on insights from cognitive psychology and experimental economics. In this work, we present an alternative, a deep learning approach that automatically performs cognitive modeling without relying on such expert knowledge. We introduce a novel architecture that allows a single network to generalize across different input and output dimensions by using matrix units rather than scalar units, and show that its performance significantly outperforms that of the previous state of the art, which relies on expert-constructed features.

****

Online and Differentially-Private Tensor Decomposition

Yining Wang, Anima Anandkumar

Tensor decomposition is positioned to be a pervasive tool in the era of big data. In this paper, we resolve many of the key algorithmic questions regarding robustness, memory efficiency, and differential privacy of tensor decomposition. We propose simple variants of the tensor power method which enjoy these strong properties. We propose the first streaming method with a linear memory requirement. Moreover, we present a noise calibrated tensor power method with efficient privacy guarantees. At the heart of all these guarantees lies a careful perturbation analysis derived in this paper which improves up on the existing results significantly.

****

Multivariate tests of association based on univariate tests

Ruth Heller, Yair Heller

For testing two vector random variables for independence, we propose testing whether the distance of one vector from an arbitrary center point is independent from the distance of the other vector from another arbitrary center point by a univariate test. We prove that under minimal assumptions, it is enough to have a consistent univariate independence test on the distances, to guarantee that the power to detect dependence between the random vectors increases to one with sample size. If the univariate test is distribution-free, the multivariate test will also be distribution-free. If we consider multiple center points and aggregate the center-specific univariate tests, the power may be further improved, and the resulting multivariate test may be distribution-free for specific aggregation methods (if the univariate test is distribution-free). We show that certain m

ultivariate tests recently proposed in the literature can be viewed as instances of this general approach. Moreover, we show in experiments that novel tests constructed using our approach can have better power and computational time than competing approaches.
****
Variational Information Maximization for Feature Selection
Shuyang Gao, Greg Ver Steeg, Aram Galstyan
Feature selection is one of the most fundamental problems in machine learning. An extensive body of work on information-theoretic feature selection exists which is based on maximizing mutual information between subsets of features and class labels. Practical methods are forced to rely on approximations due to the difficulty of estimating mutual information. We demonstrate that approximations made by existing methods are based on unrealistic assumptions. We formulate a more flexible and general class of assumptions based on variational distributions and use them to tractably generate lower bounds for mutual information. These bounds define a novel information-theoretic framework for feature selection, which we prove to be optimal under tree graphical models with proper choice of variational distributions. Our experiments demonstrate that the proposed method strongly outperforms existing information-theoretic feature selection approaches.
****
Learning Bound for Parameter Transfer Learning
Wataru Kumagai
We consider a transfer-learning problem by using the parameter transfer approach, where a suitable parameter of feature mapping is learned through one task and applied to another objective task. Then, we introduce the notion of the local stability of parametric feature mapping and parameter transfer learnability, and thereby derive a learning bound for parameter transfer algorithms. As an application of parameter transfer learning, we discuss the performance of sparse coding in self-taught learning. Although self-taught learning algorithms with plentiful unlabeled data often show excellent empirical performance, their theoretical analysis has not been studied. In this paper, we also provide the first theoretical learning bound for self-taught learning.
****
Matrix Completion has No Spurious Local Minimum
Rong Ge, Jason D. Lee, Tengyu Ma
Matrix completion is a basic machine learning problem that has wide applications, especially in collaborative filtering and recommender systems. Simple non-convex optimization algorithms are popular and effective in practice. Despite recent progress in proving various non-convex algorithms converge from a good initial point, it remains unclear why random or arbitrary initialization suffices in practice. We prove that the commonly used non-convex objective function for matrix completion has no spurious local minima --- all local minima must also be global. Therefore, many popular optimization algorithms such as (stochastic) gradient descent can provably solve matrix completion with \textit{arbitrary} initialization in polynomial time.
****
Deep Submodular Functions: Definitions and Learning
Brian W. Dolhansky, Jeff A. Bilmes
We propose and study a new class of submodular functions called deep submodular functions (DSFs). We define DSFs and situate them within the broader context of classes of submodular functions in relationship both to various matroid ranks and sums of concave composed with modular functions (SCMs). Notably, we find that DSFs constitute a strictly broader class than SCMs, thus motivating their use, but that they do not comprise all submodular functions. Interestingly, some DSFs can be seen as special cases of certain deep neural networks (DNNs), hence the name. Finally, we provide a method to learn DSFs in a max-margin framework, and offer preliminary results applying this both to synthetic and real-world data instances.
****
Adaptive optimal training of animal behavior

Ji Hyun Bak, Jung Yoon Choi, Athena Akrami, Ilana Witten, Jonathan W. Pillow

Neuroscience experiments often require training animals to perform tasks designed to elicit various sensory, cognitive, and motor behaviors. Training typically involves a series of gradual adjustments of stimulus conditions and rewards in order to bring about learning. However, training protocols are usually hand-designed, relying on a combination of intuition, guesswork, and trial-and-error, and often require weeks or months to achieve a desired level of task performance. Here we combine ideas from reinforcement learning and adaptive optimal experimental design to formulate methods for adaptive optimal training of animal behavior. Our work addresses two intriguing problems at once: first, it seeks to infer the learning rules underlying an animal's behavioral changes during training; second, it seeks to exploit these rules to select stimuli that will maximize the rate of learning toward a desired objective. We develop and test these methods using data collected from rats during training on a two-interval sensory discrimination task. We show that we can accurately infer the parameters of a policy-gradient-based learning algorithm that describes how the animal's internal model of the task evolves over the course of training. We then formulate a theory for optimal training, which involves selecting sequences of stimuli that will drive the animal's internal policy toward a desired location in the parameter space. Simulations show that our method can in theory provide a substantial speedup over standard training methods. We feel these results will hold considerable theoretical and practical implications both for researchers in reinforcement learning and for experimentalists seeking to train animals.

****

Structured Matrix Recovery via the Generalized Dantzig Selector

Sheng Chen, Arindam Banerjee

In recent years, structured matrix recovery problems have gained considerable attention for its real world applications, such as recommender systems and computer vision. Much of the existing work has focused on matrices with low-rank structure, and limited progress has been made on matrices with other types of structure. In this paper we present non-asymptotic analysis for estimation of generally structured matrices via the generalized Dantzig selector based on sub-Gaussian measurements. We show that the estimation error can always be succinctly expressed in terms of a few geometric measures such as Gaussian widths of suitable sets associated with the structure of the underlying true matrix. Further, we derive general bounds on these geometric measures for structures characterized by unitarily invariant norms, a large family covering most matrix norms of practical interest. Examples are provided to illustrate the utility of our theoretical development.

****

Robust k-means: a Theoretical Revisit

ALEXANDROS GEORGOGIANNIS

Over the last years, many variations of the quadratic k-means clustering procedure have been proposed, all aiming to robustify the performance of the algorithm in the presence of outliers. In general terms, two main approaches have been developed: one based on penalized regularization methods, and one based on trimming functions. In this work, we present a theoretical analysis of the robustness and consistency properties of a variant of the classical quadratic k-means algorithm, the robust k-means, which borrows ideas from outlier detection in regression. We show that two outliers in a dataset are enough to breakdown this clustering procedure. However, if we focus on “well-structured” datasets, then robust k-means can recover the underlying cluster structure in spite of the outliers. Finally, we show that, with slight modifications, the most general non-asymptotic results for consistency of quadratic k-means remain valid for this robust variant.

****

Tree-Structured Reinforcement Learning for Sequential Object Localization

Zequn Jie, Xiaodan Liang, Jiashi Feng, Xiaojie Jin, Wen Lu, Shuicheng Yan

Existing object proposal algorithms usually search for possible object regions over multiple locations and scales \emph{ separately}, which ignore the interdependency among different objects and deviate from the human perception procedure.

To incorporate global interdependency between objects into object localization, we propose an effective Tree-structured Reinforcement Learning (Tree-RL) approach to sequentially search for objects by fully exploiting both the current observation and historical search paths. The Tree-RL approach learns multiple searching policies through maximizing the long-term reward that reflects localization accuracies over all the objects. Starting with taking the entire image as a proposal, the Tree-RL approach allows the agent to sequentially discover multiple objects via a tree-structured traversing scheme. Allowing multiple near-optimal policies, Tree-RL offers more diversity in search paths and is able to find multiple objects with a single feed-forward pass. Therefore, Tree-RL can better cover different objects with various scales which is quite appealing in the context of object proposal. Experiments on PASCAL VOC 2007 and 2012 validate the effectiveness of the Tree-RL, which can achieve comparable recalls with current object proposal algorithms via much fewer candidate windows.

****

One-vs-Each Approximation to Softmax for Scalable Estimation of Probabilities
Michalis Titsias RC AUEB

The softmax representation of probabilities for categorical variables plays a prominent role in modern machine learning with numerous applications in areas such as large scale classification, neural language modeling and recommendation systems. However, softmax estimation is very expensive for large scale inference because of the high cost associated with computing the normalizing constant. Here, we introduce an efficient approximation to softmax probabilities which takes the form of a rigorous lower bound on the exact probability. This bound is expressed as a product over pairwise probabilities and it leads to scalable estimation based on stochastic optimization. It allows us to perform doubly stochastic estimation by subsampling both training instances and class labels. We show that the new bound has interesting theoretical properties and we demonstrate its use in classification problems.

****

Poisson-Gamma dynamical systems
Aaron Schein, Hanna Wallach, Mingyuan Zhou

This paper presents a dynamical system based on the Poisson-Gamma construction for sequentially observed multivariate count data. Inherent to the model is a novel Bayesian nonparametric prior that ties and shrinks parameters in a powerful way. We develop theory about the model's infinite limit and its steady-state. The model's inductive bias is demonstrated on a variety of real-world datasets where it is shown to learn interpretable structure and have superior predictive performance.

****

Convergence guarantees for kernel-based quadrature rules in misspecified settings
Motonobu Kanagawa, Bharath K. Sriperumbudur, Kenji Fukumizu

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.

****

Maximization of Approximately Submodular Functions
Thibaut Horel, Yaron Singer

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.

****

Causal meets Submodular: Subset Selection with Directed Information
Yuxun Zhou, Costas J. Spanos

We study causal subset selection with Directed Information as the measure of prediction causality. Two typical tasks, causal sensor placement and covariate selection, are correspondingly formulated into cardinality constrained directed info rmation maximizations. To attack the NP-hard problems, we show that the first problem is submodular while not necessarily monotonic. And the second one is ``nearly'' submodular. To substantiate the idea of approximate submodularity, we introduce a novel quantity, namely submodularity index (SmI), for general set functions. Moreover, we show that based on SmI, greedy algorithm has performance guarantee for the maximization of possibly non-monotonic and non-submodular functions, justifying its usage for a much broader class of problems. We evaluate the theoretical results with several case studies, and also illustrate the application of the subset selection to causal structure learning.

****

Linear Feature Encoding for Reinforcement Learning

Zhao Song, Ronald E. Parr, Xuejun Liao, Lawrence Carin

Feature construction is of vital importance in reinforcement learning, as the quality of a value function or policy is largely determined by the corresponding features. The recent successes of deep reinforcement learning (RL) only increase the importance of understanding feature construction. Typical deep RL approaches use a linear output layer, which means that deep RL can be interpreted as a feature construction/encoding network followed by linear value function approximation. This paper develops and evaluates a theory of linear feature encoding. We extend theoretical results on feature quality for linear value function approximation from the uncontrolled case to the controlled case. We then develop a supervised linear feature encoding method that is motivated by insights from linear value function approximation theory, as well as empirical successes from deep RL. The resulting encoder is a surprisingly effective method for linear value function approximation using raw images as inputs.

****

Mixed Linear Regression with Multiple Components

Kai Zhong, Prateek Jain, 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.

****

Unsupervised Learning of Spoken Language with Visual Context

David Harwath, Antonio Torralba, James Glass

Humans learn to speak before they can read or write, so why can’t computers do the same? In this paper, we present a deep neural network model capable of rudimentary spoken language acquisition using untranscribed audio training data, whose only supervision comes in the form of contextually relevant visual images. We describe the collection of our data comprised of over 120,000 spoken audio captions for the Places image dataset and evaluate our model on an image search and annotation task. We also provide some visualizations which suggest that our model is learning to recognize meaningful words within the caption spectrograms.

****

Crowdsourced Clustering: Querying Edges vs Triangles

Ramya Korlakai Vinayak, Babak Hassibi

We consider the task of clustering items using answers from non-expert crowd workers. In such cases, the workers are often not able to label the items directly, however, it is reasonable to assume that they can compare items and judge whether they are similar or not. An important question is what queries to make, and we compare two types: random edge queries, where a pair of items is revealed, and random triangles, where a triple is. Since it is far too expensive to query all possible edges and/or triangles, we need to work with partial observations subject to a fixed query budget constraint. When a generative model for the data is available (and we consider a few of these) we determine the cost of a query by its entropy; when such models do not exist we use the average response time per query of the workers as a surrogate for the cost. In addition to theoretical justification, through several simulations and experiments on two real data sets on Amazon Mechanical Turk, we empirically demonstrate that, for a fixed budget, triangle queries uniformly outperform edge queries. Even though, in contrast to edg e queries, triangle queries reveal dependent edges, they provide more reliable edges and, for a fixed budget, many more of them. We also provide a sufficient condition on the number of observations, edge densities inside and outside the clusters and the minimum cluster size required for the exact recovery of the true adjacency matrix via triangle queries using a convex optimization-based clustering algorithm.

****

Learning feed-forward one-shot learners

Luca Bertinetto, João F. Henriques, Jack Valmadre, Philip Torr, Andrea Vedaldi

One-shot learning is usually tackled by using generative models or discriminative embeddings. Discriminative methods based on deep learning, which are very effective in other learning scenarios, are ill-suited for one-shot learning as they need large amounts of training data. In this paper, we propose a method to learn the parameters of a deep model in one shot. We construct the learner as a second deep network, called a learnet, which predicts the parameters of a pupil network from a single exemplar. In this manner we obtain an efficient feed-forward one-shot learner, trained end-to-end by minimizing a one-shot classification objective in a learning to learn formulation. In order to make the construction feasible, we propose a number of factorizations of the parameters of the pupil network. We demonstrate encouraging results by learning characters from single exemplars in Omniglot, and by tracking visual objects from a single initial exemplar in the Visual Object Tracking benchmark.

****

Reshaped Wirtinger Flow for Solving Quadratic System of Equations

Huishuai Zhang, Yingbin Liang

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 Poisoning Attacks on Factorization-Based Collaborative Filtering

Bo Li, Yining Wang, Aarti Singh, Yevgeniy Vorobeychik

Recommendation and collaborative filtering systems are important in modern information and e-commerce applications. As these systems are becoming increasingly popular in industry, their outputs could affect business decision making, introducing incentives for an adversarial party to compromise the availability or integrity of such systems. We introduce a data poisoning attack on collaborative filtering systems. We demonstrate how a powerful attacker with full knowledge of the learner can generate malicious data so as to maximize his/her malicious objectives, while at the same time mimicking normal user behaviors to avoid being detected. While the complete knowledge assumption seems extreme, it enables a robust assessment of the vulnerability of collaborative filtering schemes to highly motivated attacks. We present efficient solutions for two popular factorization-based collaborative filtering algorithms: the alternative minimization formulation and the nuclear norm minimization method. Finally, we test the effectiveness of our proposed algorithms on real-world data and discuss potential defensive strategies.

****

PAC-Bayesian Theory Meets Bayesian Inference

Pascal Germain, Francis Bach, Alexandre Lacoste, Simon Lacoste-Julien

We exhibit a strong link between frequentist PAC-Bayesian bounds and the Bayesian marginal likelihood. That is, for the negative log-likelihood loss function, we show that the minimization of PAC-Bayesian generalization bounds maximizes the Bayesian marginal likelihood. This provides an alternative explanation to the Bayesian Occam's razor criteria, under the assumption that the data is generated by an i.i.d. distribution. Moreover, as the negative log-likelihood is an unbounded loss function, we motivate and propose a PAC-Bayesian theorem tailored for the sub-gamma loss family, and we show that our approach is sound on classical Bayesian linear regression tasks.

****

Fast Mixing Markov Chains for Strongly Rayleigh Measures, DPPs, and Constrained Sampling

Chengtao Li, Suvrit Sra, Stefanie Jegelka

We study probability measures induced by set functions with constraints. Such measures arise in a variety of real-world settings, where prior knowledge, resource limitations, or other pragmatic considerations impose constraints. We consider the task of rapidly sampling from such constrained measures, and develop fast Markov chain samplers for them. Our first main result is for MCMC sampling from Strongly Rayleigh (SR) measures, for which we present sharp polynomial bounds on the mixing time. As a corollary, this result yields a fast mixing sampler for Determinantal Point Processes (DPPs), yielding (to our knowledge) the first provably fast MCMC sampler for DPPs since their inception over four decades ago. Beyond SR measures, we develop MCMC samplers for probabilistic models with hard constraints and identify sufficient conditions under which their chains mix rapidly. We illustrate our claims by empirically verifying the dependence of mixing times on the key factors governing our theoretical bounds.

****

Temporal Regularized Matrix Factorization for High-dimensional Time Series Prediction

Hsiang-Fu Yu, Nikhil Rao, Inderjit S. Dhillon

Time series prediction problems are becoming increasingly high-dimensional in modern applications, such as climatology and demand forecasting. For example, in the latter problem, the number of items for which demand needs to be forecast might be as large as 50,000. In addition, the data is generally noisy and full of missing values. Thus, modern applications require methods that are highly scalable, and can deal with noisy data in terms of corruptions or missing values. However, classical time series methods usually fall short of handling these issues. In this paper, we present a temporal regularized matrix factorization (TRMF) framework which supports data-driven temporal learning and forecasting. We develop novel regularization schemes and use scalable matrix factorization methods that are eminently suited for high-dimensional time series data that has many missing values. Our proposed TRMF is highly general, and subsumes many existing approaches for time series analysis. We make interesting connections to graph regularization methods in the context of learning the dependencies in an autoregressive framework. Experimental results show the superiority of TRMF in terms of scalability and prediction quality. In particular, TRMF is two orders of magnitude faster than other methods on a problem of dimension 50,000, and generates better forecasts on real-world datasets such as Wal-mart E-commerce datasets.

****

FPNN: Field Probing Neural Networks for 3D Data

Yangyan Li, Soeren Pirk, Hao Su, Charles R. Qi, Leonidas J. Guibas

Building discriminative representations for 3D data has been an important task in computer graphics and computer vision research. Convolutional Neural Networks (CNNs) have shown to operate on 2D images with great success for a variety of tasks. Lifting convolution operators to 3D (3DCNNs) seems like a plausible and promising next step. Unfortunately, the computational complexity of 3D CNNs grows cubically with respect to voxel resolution. Moreover, since most 3D geometry representations are boundary based, occupied regions do not increase proportionately with the size of the discretization, resulting in wasted computation. In this work, we represent 3D spaces as volumetric fields, and propose a novel design that employs field probing filters to efficiently extract features from them. Each field probing filter is a set of probing points -- sensors that perceive the space. Our learning algorithm optimizes not only the weights associated with the probing points, but also their locations, which deforms the shape of the probing filters and adaptively distributes them in 3D space. The optimized probing points sense the 3D space "intelligently", rather than operating blindly over the entire domain. We show that field probing is significantly more efficient than 3DCNNs, while providing state-of-the-art performance, on classification tasks for 3D object recognition benchmark datasets.

****

Object based Scene Representations using Fisher Scores of Local Subspace Projections
Mandar D. Dixit, Nuno Vasconcelos
Several works have shown that deep CNN classifiers can be easily transferred across datasets, e.g. the transfer of a CNN trained to recognize objects on ImageNET to an object detector on Pascal VOC. Less clear, however, is the ability of CNNs to transfer knowledge across tasks. A common example of such transfer is the problem of scene classification that should leverage localized object detections to recognize holistic visual concepts. While this problem is currently addressed with Fisher vector representations, these are now shown ineffective for the high-dimensional and highly non-linear features extracted by modern CNNs. It is argued that this is mostly due to the reliance on a model, the Gaussian mixture of diagonal covariances, which has a very limited ability to capture the second order statistics of CNN features. This problem is addressed by the adoption of a better model, the mixture of factor analyzers (MFA), which approximates the non-linear data manifold by a collection of local subspaces. The Fisher score with respect to the MFA (MFA-FS) is derived and proposed as an image representation for holistic image classifiers. Extensive experiments show that the MFA-FS has state of the art performance for object-to-scene transfer and this transfer actually outperforms the training of a scene CNN from a large scene dataset. The two representations are also shown to be complementary, in the sense that their combination outperforms each of the representations by itself. When combined, they produce a state of the art scene classifier.
****

Architectural Complexity Measures of Recurrent Neural Networks
Saizheng Zhang, Yuhuai Wu, Tong Che, Zhouhan Lin, Roland Memisevic, Russ R. Salakhutdinov, Yoshua Bengio
In this paper, we systematically analyze the connecting architectures of recurrent neural networks (RNNs). Our main contribution is twofold: first, we present a rigorous graph-theoretic framework describing the connecting architectures of RNNs in general. Second, we propose three architecture complexity measures of RNNs: (a) the recurrent depth, which captures the RNN’s over-time nonlinear complexity, (b) the feedforward depth, which captures the local input-output nonlinearity (similar to the “depth” in feedforward neural networks (FNNs)), and (c) the recurrent skip coefficient which captures how rapidly the information propagates over time. We rigorously prove each measure’s existence and computability. Our experimental results show that RNNs might benefit from larger recurrent depth and feedforward depth. We further demonstrate that increasing recurrent skip coefficient offers performance boosts on long term dependency problems.
****

Interaction Screening: Efficient and Sample-Optimal Learning of Ising Models
Marc Vuffray, Sidhant Misra, Andrey Lokhov, Michael Chertkov
We consider the problem of learning the underlying graph of an unknown Ising model on p spins from a collection of i.i.d. samples generated from the model. We suggest a new estimator that is computationally efficient and requires a number of samples that is near-optimal with respect to previously established information theoretic lower-bound. Our statistical estimator has a physical interpretation in terms of "interaction screening". The estimator is consistent and is efficiently implemented using convex optimization. We prove that with appropriate regularization, the estimator recovers the underlying graph using a number of samples that is logarithmic in the system size p and exponential in the maximum coupling-intensity and maximum node-degree.
****

Community Detection on Evolving Graphs
Aris Anagnostopoulos, Jakub ■■cki, Silvio Lattanzi, Stefano Leonardi, Mohammad Mahdian
Clustering is a fundamental step in many information-retrieval and data-mining applications. Detecting clusters in graphs is also a key tool for finding the community structure in social and behavioral networks. In many of these applications, the input graph evolves over time in a continual and decentralized manner, an

d, to maintain a good clustering, the clustering algorithm needs to repeatedly probe the graph. Furthermore, there are often limitations on the frequency of such probes, either imposed explicitly by the online platform (e.g., in the case of crawling proprietary social networks like twitter) or implicitly because of resource limitations (e.g., in the case of crawling the web). In this paper, we study a model of clustering on evolving graphs that captures this aspect of the problem. Our model is based on the classical stochastic block model, which has been used to assess rigorously the quality of various static clustering methods. In our model, the algorithm is supposed to reconstruct the planted clustering, given the ability to query for small pieces of local information about the graph, at a limited rate. We design and analyze clustering algorithms that work in this model, and show asymptotically tight upper and lower bounds on their accuracy. Finally, we perform simulations, which demonstrate that our main asymptotic results hold true also in practice.

****

Preference Completion from Partial Rankings
Suriya Gunasekar, Oluwasanmi O. Koyejo, 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.

****

“Congruent” and “Opposite” Neurons: Sisters for Multisensory Integration and Segregation
Wen-Hao Zhang, He Wang, K. Y. Michael Wong, Si Wu

Experiments reveal that in the dorsal medial superior temporal (MSTd) and the ventral intraparietal (VIP) areas, where visual and vestibular cues are integrated to infer heading direction, there are two types of neurons with roughly the same number. One is “congruent” cells, whose preferred heading directions are similar in response to visual and vestibular cues; and the other is “opposite” cells, whose preferred heading directions are nearly “opposite” (with an offset of 180 degree) in response to visual vs. vestibular cues. Congruent neurons are known to be responsible for cue integration, but the computational role of opposite neurons remains largely unknown. Here, we propose that opposite neurons may serve to encode the disparity information between cues necessary for multisensory segregation. We build a computational model composed of two reciprocally coupled modules, MSTd and VIP, and each module consists of groups of congruent and opposite neurons. In the model, congruent neurons in two modules are reciprocally connected with each other in the congruent manner, whereas opposite neurons are reciprocally connected in the opposite manner. Mimicking the experimental protocol, our model reproduces the characteristics of congruent and opposite neurons, and demonstrates that in each module, the sisters of congruent and opposite neurons can jointly achieve optimal multisensory information integration and segregation. This study sheds light on our understanding of how the brain implements optimal multisensory integration and segregation concurrently in a distributed manner.

****

A Consistent Regularization Approach for Structured Prediction
Carlo Ciliberto, Lorenzo Rosasco, Alessandro Rudi

We propose and analyze a regularization approach for structured prediction problems. We characterize a large class of loss functions that allows to naturally embed structured outputs in a linear space. We exploit this fact to design learning algorithms using a surrogate loss approach and regularization techniques. We prove universal consistency and finite sample bounds characterizing the generalization properties of the proposed method. Experimental results are provided to demonstrate the practical usefulness of the proposed approach.

****

Fast recovery from a union of subspaces
Chinmay Hegde, Piotr Indyk, Ludwig Schmidt

We address the problem of recovering a high-dimensional but structured vector from linear observations in a general setting where the vector can come from an ar

bitrary union of subspaces. This setup includes well-studied problems such as compressive sensing and low-rank matrix recovery. We show how to design more efficient algorithms for the union-of subspace recovery problem by using approximate projections. Instantiating our general framework for the low-rank matrix recovery problem gives the fastest provable running time for an algorithm with optimal sample complexity. Moreover, we give fast approximate projections for 2D histograms, another well-studied low-dimensional model of data. We complement our theoretical results with experiments demonstrating that our framework also leads to improved time and sample complexity empirically.
****
Improved Techniques for Training GANs
Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, Xi Chen, Xi Chen
We present a variety of new architectural features and training procedures that we apply to the generative adversarial networks (GANs) framework. Using our new techniques, we achieve state-of-the-art results in semi-supervised classification on MNIST, CIFAR-10 and SVHN. The generated images are of high quality as confirmed by a visual Turing test: Our model generates MNIST samples that humans cannot distinguish from real data, and CIFAR-10 samples that yield a human error rate of 21.3%. We also present ImageNet samples with unprecedented resolution and show that our methods enable the model to learn recognizable features of ImageNet classes.
****
Coordinate-wise Power Method
Qi Lei, Kai Zhong, Inderjit S. Dhillon
In this paper, we propose a coordinate-wise version of the power method from an optimization viewpoint. The vanilla power method simultaneously updates all the coordinates of the iterate, which is essential for its convergence analysis. However, different coordinates converge to the optimal value at different speeds. Our proposed algorithm, which we call coordinate-wise power method, is able to select and update the most important k coordinates in O(kn) time at each iteration, where n is the dimension of the matrix and k <= n is the size of the active set. Inspired by the ''greedy'' nature of our method, we further propose a greedy coordinate descent algorithm applied on a non-convex objective function specialized for symmetric matrices. We provide convergence analyses for both methods. Experimental results on both synthetic and real data show that our methods achieve up to 20 times speedup over the basic power method. Meanwhile, due to their coordinate-wise nature, our methods are very suitable for the important case when data cannot fit into memory. Finally, we introduce how the coordinate-wise mechanism could be applied to other iterative methods that are used in machine learning.
****
On Mixtures of Markov Chains
Rishi Gupta, Ravi Kumar, Sergei Vassilvitskii
We study the problem of reconstructing a mixture of Markov chains from the trajectories generated by random walks through the state space. Under mild non-degeneracy conditions, we show that we can uniquely reconstruct the underlying chains by only considering trajectories of length three, which represent triples of states. Our algorithm is spectral in nature, and is easy to implement.
****
Near-Optimal Smoothing of Structured Conditional Probability Matrices
Moein Falahatgar, Mesrob I. Ohannessian, Alon Orlitsky
Utilizing the structure of a probabilistic model can significantly increase its learning speed. Motivated by several recent applications, in particular bigram models in language processing, we consider learning low-rank conditional probability matrices under expected KL-risk. This choice makes smoothing, that is the careful handling of low-probability elements, paramount. We derive an iterative algorithm that extends classical non-negative matrix factorization to naturally incorporate additive smoothing and prove that it converges to the stationary points of a penalized empirical risk. We then derive sample-complexity bounds for the

global minimizer of the penalized risk and show that it is within a small factor of the optimal sample complexity. This framework generalizes to more sophisticated smoothing techniques, including absolute-discounting.

****

Dynamic Filter Networks

Xu Jia, Bert De Brabandere, Tinne Tuytelaars, Luc V. Gool

In a traditional convolutional layer, the learned filters stay fixed after training. In contrast, we introduce a new framework, the Dynamic Filter Network, where filters are generated dynamically conditioned on an input. We show that this architecture is a powerful one, with increased flexibility thanks to its adaptive nature, yet without an excessive increase in the number of model parameters. A wide variety of filtering operation can be learned this way, including local spatial transformations, but also others like selective (de)blurring or adaptive feature extraction. Moreover, multiple such layers can be combined, e.g. in a recurrent architecture. We demonstrate the effectiveness of the dynamic filter network on the tasks of video and stereo prediction, and reach state-of-the-art performance on the moving MNIST dataset with a much smaller model. By visualizing the learned filters, we illustrate that the network has picked up flow information by only looking at unlabelled training data. This suggests that the network can be used to pretrain networks for various supervised tasks in an unsupervised way, like optical flow and depth estimation.

****

Estimating the Size of a Large Network and its Communities from a Random Sample

Lin Chen, Amin Karbasi, Forrest W. Crawford

Most real-world networks are too large to be measured or studied directly and there is substantial interest in estimating global network properties from smaller sub-samples. One of the most important global properties is the number of vertices/nodes in the network. Estimating the number of vertices in a large network is a major challenge in computer science, epidemiology, demography, and intelligence analysis. In this paper we consider a population random graph G = (V;E) from the stochastic block model (SBM) with K communities/blocks. A sample is obtained by randomly choosing a subset W and letting G(W) be the induced subgraph in G of the vertices in W. In addition to G(W), we observe the total degree of each sampled vertex and its block membership. Given this partial information, we propose an efficient PopULation Size Estimation algorithm, called PULSE, that accurately estimates the size of the whole population as well as the size of each community. To support our theoretical analysis, we perform an exhaustive set of experiments to study the effects of sample size, K, and SBM model parameters on the accuracy of the estimates. The experimental results also demonstrate that PULSE significantly outperforms a widely-used method called the network scale-up estimator in a wide variety of scenarios.

****

Generalization of ERM in Stochastic Convex Optimization: The Dimension Strikes Back

Vitaly Feldman

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.

****

Nested Mini-Batch K-Means

James Newling, François Fleuret

A new algorithm is proposed which accelerates the mini-batch k-means algorithm of Sculley (2010) by using the distance bounding approach of Elkan (2003). We argue that, when incorporating distance bounds into a mini-batch algorithm, already used data should preferentially be reused. To this end we propose using nested mini-batches, whereby data in a mini-batch at iteration t is automatically reused at iteration t+1.   Using nested mini-batches presents two difficulties. The first is that unbalanced use of data can bias estimates, which we resolve by ensuring that each data sample contributes exactly once to centroids. The second is in choosing mini-batch sizes, which we address by balancing premature fine-tuning of centroids with redundancy induced slow-down. Experiments show that the resulting nmbatch algorithm is very effective, often arriving within 1\% of the empirical minimum 100 times earlier than the standard mini-batch algorithm.

****

Infinite Hidden Semi-Markov Modulated Interaction Point Process

matt zhang, Peng Lin, Peng Lin, Ting Guo, Yang Wang, Yang Wang, Fang Chen

The correlation between events is ubiquitous and important for temporal events modelling. In many cases, the correlation exists between not only events' emitted observations, but also their arrival times. State space models (e.g., hidden Markov model) and stochastic interaction point process models (e.g., Hawkes process) have been studied extensively yet separately for the two types of correlations in the past. In this paper, we propose a Bayesian nonparametric approach that considers both types of correlations via unifying and generalizing hidden semi-Markov model and interaction point process model. The proposed approach can simultaneously model both the observations and arrival times of temporal events, and determine the number of latent states from data. A Metropolis-within-particle-Gibbs sampler with ancestor resampling is developed for efficient posterior inference. The approach is tested on both synthetic and real-world data with promising outcomes.

****

A Credit Assignment Compiler for Joint Prediction

Kai-Wei Chang, He He, Stephane Ross, Hal Daume III, John Langford

Many machine learning applications involve jointly predicting multiple mutually dependent output variables. Learning to search is a family of methods where the complex decision problem is cast into a sequence of decisions via a search space. Although these methods have shown promise both in theory and in practice, implementing them has been burdensomely awkward. In this paper, we show the search space can be defined by an arbitrary imperative program, turning learning to search into a credit assignment compiler. Altogether with the algorithmic improvements for the compiler, we radically reduce the complexity of programming and the running time. We demonstrate the feasibility of our approach on multiple joint prediction tasks. In all cases, we obtain accuracies as high as alternative approaches, at drastically reduced execution and programming time.

****

Deep Exploration via Bootstrapped DQN

Ian Osband, Charles Blundell, Alexander Pritzel, Benjamin Van Roy

Efficient exploration remains a major challenge for reinforcement learning (RL). Common dithering strategies for exploration, such as epsilon-greedy, do not carry out temporally-extended (or deep) exploration; this can lead to exponentially larger data requirements. However, most algorithms for statistically efficient RL are not computationally tractable in complex environments. Randomized value functions offer a promising approach to efficient exploration with generalization, but existing algorithms are not compatible with nonlinearly parameterized value functions. As a first step towards addressing such contexts we develop bootstrapped DQN. We demonstrate that bootstrapped DQN can combine deep exploration with deep neural networks for exponentially faster learning than any dithering strategy. In the Arcade Learning Environment bootstrapped DQN substantially improves learning speed and cumulative performance across most games.

****

Estimating Nonlinear Neural Response Functions using GP Priors and Kronecker Methods

Cristina Savin, Gasper Tkacik

Jointly characterizing neural responses in terms of several external variables promises novel insights into circuit function, but remains computationally prohibitive in practice. Here we use gaussian process (GP) priors and exploit recent advances in fast GP inference and learning based on Kronecker methods, to efficiently estimate multidimensional nonlinear tuning functions. Our estimator require considerably less data than traditional methods and further provides principled uncertainty estimates. We apply these tools to hippocampal recordings during op en field exploration and use them to characterize the joint dependence of CA1 responses on the position of the animal and several other variables, including the animal's speed, direction of motion, and network oscillations.Our results provide an unprecedentedly detailed quantification of the tuning of hippocampal neurons. The model's generality suggests that our approach can be used to estimate neural response properties in other brain regions.

****

Structured Sparse Regression via Greedy Hard Thresholding

Prateek Jain, Nikhil Rao, 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.

****

A Multi-Batch L-BFGS Method for Machine Learning

Albert S. Berahas, Jorge Nocedal, Martin Takac

The question of how to parallelize the stochastic gradient descent (SGD) method has received much attention in the literature. In this paper, we focus instead on batch methods that use a sizeable fraction of the training set at each iteration to facilitate parallelism, and that employ second-order information. In order to improve the learning process, we follow a multi-batch approach in which the batch changes at each iteration. This can cause difficulties because L-BFGS employs gradient differences to update the Hessian approximations, and when these gradients are computed using different data points the process can be unstable. This paper shows how to perform stable quasi-Newton updating in the multi-batch setting, illustrates the behavior of the algorithm in a distributed computing platform, and studies its convergence properties for both the convex and nonconvex cases.

****

Cooperative Graphical Models

Josip Djolonga, Stefanie Jegelka, Sebastian Tschiatschek, Andreas Krause

We study a rich family of distributions that capture variable interactions significantly more expressive than those representable with low-treewidth or pairwise graphical models, or log-supermodular models. We call these cooperative graphical models. Yet, this family retains structure, which we carefully exploit for efficient inference techniques. Our algorithms combine the polyhedral structure of submodular functions in new ways with variational inference methods to obtain both lower and upper bounds on the partition function. While our fully convex upper bound is minimized as an SDP or via tree-reweighted belief propagation, our lower bound is tightened via belief propagation or mean-field algorithms. The resulting algorithms are easy to implement and, as our experiments show, effectively obtain good bounds and marginals for synthetic and real-world examples.

****

What Makes Objects Similar: A Unified Multi-Metric Learning Approach

Han-Jia Ye, De-Chuan Zhan, Xue-Min Si, Yuan Jiang, Zhi-Hua Zhou

Linkages are essentially determined by similarity measures that may be derived from multiple perspectives. For example, spatial linkages are usually generated based on localities of heterogeneous data, whereas semantic linkages can come from various properties, such as different physical meanings behind social relations. Many existing metric learning models focus on spatial linkages, but leave the rich semantic factors unconsidered. Similarities based on these models are usually overdetermined on linkages. We propose a Unified Multi-Metric Learning (UM2L) framework to exploit multiple types of metrics. In UM2L, a type of combination operator is introduced for distance characterization from multiple perspectives, and thus can introduce flexibilities for representing and utilizing both spatial and semantic linkages. Besides, we propose a uniform solver for UM2L which is guaranteed to converge. Extensive experiments on diverse applications exhibit the superior classification performance and comprehensibility of UM2L. Visualization results also validate its ability on physical meanings discovery.

****

Matching Networks for One Shot Learning

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, koray kavukcuoglu, Daan Wierstra

Learning from a few examples remains a key challenge in machine learning. Despite recent advances in important domains such as vision and language, the standard supervised deep learning paradigm does not offer a satisfactory solution for learning new concepts rapidly from little data. In this work, we employ ideas from metric learning based on deep neural features and from recent advances that augment neural networks with external memories. Our framework learns a network that maps a small labelled support set and an unlabelled example to its label, obviating the need for fine-tuning to adapt to new class types. We then define one-shot learning problems on vision (using Omniglot, ImageNet) and language tasks. Our algorithm improves one-shot accuracy on ImageNet from 82.2% to 87.8% and from 88% accuracy to 95% accuracy on Omniglot compared to competing approaches. We also demonstrate the usefulness of the same model on language modeling by introducing a one-shot task on the Penn Treebank.

****

Gradient-based Sampling: An Adaptive Importance Sampling for Least-squares

Rong Zhu

In modern data analysis, random sampling is an efficient and widely-used strategy to overcome the computational difficulties brought by large sample size. In previous studies, researchers conducted random sampling which is according to the input data but independent on the response variable, however the response variable may also be informative for sampling. In this paper we propose an adaptive sampling called the gradient-based sampling which is dependent on both the input data and the output for fast solving of least-square (LS) problems. We draw the data points by random sampling from the full data according to their gradient values. This sampling is computationally saving, since the running time of computing the sampling probabilities is reduced to O(nd) where n is the full sample size and d is the dimension of the input. Theoretically, we establish an error bound analysis of the general importance sampling with respect to LS solution from full data. The result establishes an improved performance of the use of our gradient-based sampling. Synthetic and real data sets are used to empirically argue that the gradient-based sampling has an obvious advantage over existing sampling methods from two aspects of statistical efficiency and computational saving.

****

Accelerating Stochastic Composition Optimization

Mengdi Wang, Ji Liu, Ethan Fang

Consider the stochastic composition optimization problem where the objective is a composition of two expected-value functions. We propose a new stochastic first-order method, namely the accelerated stochastic compositional proximal gradient (ASC-PG) method, which updates based on queries to the sampling oracle using two different timescales. The ASC-PG is the first proximal gradient method for the stochastic composition problem that can deal with nonsmooth regularization penalty. We show that the ASC-PG exhibits faster convergence than the best known algorithms, and that it achieves the optimal sample-error complexity in several important special cases. We further demonstrate the application of ASC-PG to reinforcement learning and conduct numerical experiments.

****

Variational Inference in Mixed Probabilistic Submodular Models

Josip Djolonga, Sebastian Tschiatschek, Andreas Krause

We consider the problem of variational inference in probabilistic models with both log-submodular and log-supermodular higher-order potentials. These models can represent arbitrary distributions over binary variables, and thus generalize the commonly used pairwise Markov random fields and models with log-supermodular potentials only, for which efficient approximate inference algorithms are known. While inference in the considered models is #P-hard in general, we present efficient approximate algorithms exploiting recent advances in the field of discrete optimization. We demonstrate the effectiveness of our approach in a large set of experiments, where our model allows reasoning about preferences over sets of it ems with complements and substitutes.

****

The Limits of Learning with Missing Data

Brian Bullins, Elad Hazan, Tomer Koren

We study regression and classification in a setting where the learning algorithm is allowed to access only a limited number of attributes per example, known as the limited attribute observation model. In this well-studied model, we provide the first lower bounds giving a limit on the precision attainable by any algorithm for several variants of regression, notably linear regression with the absolute loss and the squared loss, as well as for classification with the hinge loss. We complement these lower bounds with a general purpose algorithm that gives an upper bound on the achievable precision limit in the setting of learning with missing data.

****

Clustering with Same-Cluster Queries

Hassan Ashtiani, Shrinu Kushagra, Shai Ben-David

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.

****

Deconvolving Feedback Loops in Recommender Systems

Ayan Sinha, David F. Gleich, Karthik Ramani

Collaborative filtering is a popular technique to infer users' preferences on new content based on the collective information of all users preferences. Recommender systems then use this information to make personalized suggestions to users. When users accept these recommendations it creates a feedback loop in the recommender system, and these loops iteratively influence the collaborative filtering algorithm's predictions over time. We investigate whether it is possible to identify items affected by these feedback loops. We state sufficient assumptions to deconvolve the feedback loops while keeping the inverse solution tractable. We furthermore develop a metric to unravel the recommender system's influence on the entire user-item rating matrix. We use this metric on synthetic and real-world datasets to (1) identify the extent to which the recommender system affects the final rating matrix, (2) rank frequently recommended items, and (3) distinguish whether a user's rated item was recommended or an intrinsic preference. Our results indicate that it is possible to recover the ratings matrix of intrinsic user preferences using a single snapshot of the ratings matrix without any temporal information.

****

Recovery Guarantee of Non-negative Matrix Factorization via Alternating Updates

Yuanzhi Li, Yingyu Liang, Andrej Risteski

Non-negative matrix factorization is a popular tool for decomposing data into feature and weight matrices under non-negativity constraints. It enjoys practical success but is poorly understood theoretically. This paper proposes an algorithm that alternates between decoding the weights and updating the features, and shows that assuming a generative model of the data, it provably recovers the ground-truth under fairly mild conditions. In particular, its only essential requirement on features is linear independence. Furthermore, the algorithm uses ReLU to exploit the non-negativity for decoding the weights, and thus can tolerate adversarial noise that can potentially be as large as the signal, and can tolerate unbiased noise much larger than the signal. The analysis relies on a carefully designed coupling between two potential functions, which we believe is of independent interest.

****

Threshold Learning for Optimal Decision Making

Nathan F. Lepora

Decision making under uncertainty is commonly modelled as a process of competitive stochastic evidence accumulation to threshold (the drift-diffusion model). However, it is unknown how animals learn these decision thresholds. We examine thr

eshold learning by constructing a reward function that averages over many trials to Wald's cost function that defines decision optimality. These rewards are highly stochastic and hence challenging to optimize, which we address in two ways: first, a simple two-factor reward-modulated learning rule derived from Williams' REINFORCE method for neural networks; and second, Bayesian optimization of the reward function with a Gaussian process. Bayesian optimization converges in fewer trials than REINFORCE but is slower computationally with greater variance. The REINFORCE method is also a better model of acquisition behaviour in animals and a similar learning rule has been proposed for modelling basal ganglia function.
****

Joint M-Best-Diverse Labelings as a Parametric Submodular Minimization

Alexander Kirillov, Alexander Shekhovtsov, 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.
****

Confusions over Time: An Interpretable Bayesian Model to Characterize Trends in Decision Making

Himabindu Lakkaraju, Jure Leskovec

We propose Confusions over Time (CoT), a novel generative framework which facilitates a multi-granular analysis of the decision making process. The CoT not only models the confusions or error properties of individual decision makers and their evolution over time, but also allows us to obtain diagnostic insights into the collective decision making process in an interpretable manner. To this end, the CoT models the confusions of the decision makers and their evolution over time via time-dependent confusion matrices. Interpretable insights are obtained by grouping similar decision makers (and items being judged) into clusters and representing each such cluster with an appropriate prototype and identifying the most important features characterizing the cluster via a subspace feature indicator vector. Experimentation with real world data on bail decisions, asthma treatments, and insurance policy approval decisions demonstrates that CoT can accurately model and explain the confusions of decision makers and their evolution over time.
****

Measuring Neural Net Robustness with Constraints

Osbert Bastani, Yani Ioannou, Leonidas Lampropoulos, Dimitrios Vytiniotis, Aditya Nori, Antonio Criminisi

Despite having high accuracy, neural nets have been shown to be susceptible to adversarial examples, where a small perturbation to an input can cause it to become mislabeled. We propose metrics for measuring the robustness of a neural net and devise a novel algorithm for approximating these metrics based on an encoding of robustness as a linear program. We show how our metrics can be used to evaluate the robustness of deep neural nets with experiments on the MNIST and CIFAR-10 datasets. Our algorithm generates more informative estimates of robustness metrics compared to estimates based on existing algorithms. Furthermore, we show how existing approaches to improving robustness “overfit” to adversarial examples generated using a specific algorithm. Finally, we show that our techniques can be used to additionally improve neural net robustness both according to the metrics that we propose, but also according to previously proposed metrics.
****

The Power of Adaptivity in Identifying Statistical Alternatives

Kevin G. Jamieson, Daniel Haas, Benjamin Recht

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 Skills Adaptive Partitions (ASAP)

Daniel J. Mankowitz, Timothy A. Mann, Shie Mannor

We introduce the Adaptive Skills, Adaptive Partitions (ASAP) framework that (1) learns skills (i.e., temporally extended actions or options) as well as (2) where to apply them. We believe that both (1) and (2) are necessary for a truly general skill learning framework, which is a key building block needed to scale up to lifelong learning agents. The ASAP framework is also able to solve related new tasks simply by adapting where it applies its existing learned skills. We prove that ASAP converges to a local optimum under natural conditions. Finally, our experimental results, which include a RoboCup domain, demonstrate the ability of ASAP to learn where to reuse skills as well as solve multiple tasks with considerably less experience than solving each task from scratch.

****

Riemannian SVRG: Fast Stochastic Optimization on Riemannian Manifolds

Hongyi Zhang, Sashank J. Reddi, Suvrit Sra

We study optimization of finite sums of \emph{geodesically} smooth functions on Riemannian manifolds. Although variance reduction techniques for optimizing finite-sums have witnessed tremendous attention in the recent years, existing work is limited to vector space problems. We introduce \emph{Riemannian SVRG} (\rsvrg), a new variance reduced Riemannian optimization method. We analyze \rsvrg for both geodesically \emph{convex} and \emph{nonconvex} (smooth) functions. Our analysis reveals that \rsvrg inherits advantages of the usual SVRG method, but with factors depending on curvature of the manifold that influence its convergence. To our knowledge, \rsvrg is the first \emph{provably fast} stochastic Riemannian method. Moreover, our paper presents the first non-asymptotic complexity analysis (novel even for the batch setting) for nonconvex Riemannian optimization. Our results have several implications; for instance, they offer a Riemannian perspective on variance reduced PCA, which promises a short, transparent convergence analysis.

****

Hypothesis Testing in Unsupervised Domain Adaptation with Applications in Alzheimer's Disease

Hao Zhou, Vamsi K. Ithapu, Sathya Narayanan Ravi, Vikas Singh, Grace Wahba, Sterling C. Johnson

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.

****

Review Networks for Caption Generation

Zhilin Yang, Ye Yuan, Yuexin Wu, William W. Cohen, Russ R. Salakhutdinov

We propose a novel extension of the encoder-decoder framework, called a review network. The review network is generic and can enhance any existing encoder- decoder model: in this paper, we consider RNN decoders with both CNN and RNN encoders. The review network performs a number of review steps with attention mechanism on the encoder hidden states, and outputs a thought vector after each review step; the thought vectors are used as the input of the attention mechanism in the decoder. We show that conventional encoder-decoders are a special case of our framework. Empirically, we show that our framework improves over state-of- the-art encoder-decoder systems on the tasks of image captioning and source code captioning.

****

Distributed Flexible Nonlinear Tensor Factorization

Shandian Zhe, Kai Zhang, Pengyuan Wang, Kuang-chih Lee, Zenglin Xu, Yuan Qi, Zoubin Ghahramani

Tensor factorization is a powerful tool to analyse multi-way data. Recently proposed nonlinear factorization methods, although capable of capturing complex relationships, are computationally quite expensive and may suffer a severe learning bias in case of extreme data sparsity. Therefore, we propose a distributed, flexible nonlinear tensor factorization model, which avoids the expensive computations and structural restrictions of the Kronecker-product in the existing TGP formulations, allowing an arbitrary subset of tensor entries to be selected for trai ning. Meanwhile, we derive a tractable and tight variational evidence lower bound (ELBO) that enables highly decoupled, parallel computations and high-quality inference. Based on the new bound, we develop a distributed, key-value-free inference algorithm in the MapReduce framework, which can fully exploit the memory cache mechanism in fast MapReduce systems such as Spark. Experiments demonstrate the advantages of our method over several state-of-the-art approaches, in terms of both predictive performance and computational efficiency.

****

Safe Policy Improvement by Minimizing Robust Baseline Regret
Mohammad Ghavamzadeh, Marek Petrik, Yinlam Chow

An important problem in sequential decision-making under uncertainty is to use limited data to compute a safe policy, i.e., a policy that is guaranteed to perform at least as well as a given baseline strategy. In this paper, we develop and analyze a new model-based approach to compute a safe policy when we have access to an inaccurate dynamics model of the system with known accuracy guarantees. Our proposed robust method uses this (inaccurate) model to directly minimize the (negative) regret w.r.t. the baseline policy. Contrary to the existing approaches, minimizing the regret allows one to improve the baseline policy in states with accurate dynamics and seamlessly fall back to the baseline policy, otherwise. We show that our formulation is NP-hard and propose an approximate algorithm. Our empirical results on several domains show that even this relatively simple approximate algorithm can significantly outperform standard approaches.

****

Safe Exploration in Finite Markov Decision Processes with Gaussian Processes
Matteo Turchetta, Felix Berkenkamp, Andreas Krause

In classical reinforcement learning agents accept arbitrary short term loss for long term gain when exploring their environment. This is infeasible for safety critical applications such as robotics, where even a single unsafe action may cause system failure or harm the environment. In this paper, we address the problem of safely exploring finite Markov decision processes (MDP). We define safety in terms of an a priori unknown safety constraint that depends on states and actions and satisfies certain regularity conditions expressed via a Gaussian process prior. We develop a novel algorithm, SAFEMDP, for this task and prove that it completely explores the safely reachable part of the MDP without violating the safety constraint. To achieve this, it cautiously explores safe states and actions in order to gain statistical confidence about the safety of unvisited state-action pairs from noisy observations collected while navigating the environment. Moreover, the algorithm explicitly considers reachability when exploring the MDP, ensuring that it does not get stuck in any state with no safe way out. We demonstrate our method on digital terrain models for the task of exploring an unknown map with a rover.

****

Multimodal Residual Learning for Visual QA
Jin-Hwa Kim, Sang-Woo Lee, Donghyun Kwak, Min-Oh Heo, Jeonghee Kim, Jung-Woo Ha, Byoung-Tak Zhang

Deep neural networks continue to advance the state-of-the-art of image recognition tasks with various methods. However, applications of these methods to multimodality remain limited. We present Multimodal Residual Networks (MRN) for the multimodal residual learning of visual question-answering, which extends the idea of the deep residual learning. Unlike the deep residual learning, MRN effectively learns the joint representation from visual and language information. The main idea is to use element-wise multiplication for the joint residual mappings exploiting the residual learning of the attentional models in recent studies. Various alternative models introduced by multimodality are explored based on our study. We achieve the state-of-the-art results on the Visual QA dataset for both Open-Ended and Multiple-Choice tasks. Moreover, we introduce a novel method to visualize the attention effect of the joint representations for each learning block using back-propagation algorithm, even though the visual features are collapsed without spatial information.

****

Variance Reduction in Stochastic Gradient Langevin Dynamics

Kumar Avinava Dubey, Sashank J. Reddi, Sinead A. Williamson, Barnabas Poczos, Alexander J. Smola, Eric P. Xing

Stochastic gradient-based Monte Carlo methods such as stochastic gradient Langevin dynamics are useful tools for posterior inference on large scale datasets in many machine learning applications. These methods scale to large datasets by using noisy gradients calculated using a mini-batch or subset of the dataset. However, the high variance inherent in these noisy gradients degrades performance and leads to slower mixing. In this paper, we present techniques for reducing variance in stochastic gradient Langevin dynamics, yielding novel stochastic Monte Carlo methods that improve performance by reducing the variance in the stochastic gradient. We show that our proposed method has better theoretical guarantees on convergence rate than stochastic Langevin dynamics. This is complemented by impressive empirical results obtained on a variety of real world datasets, and on four different machine learning tasks (regression, classification, independent component analysis and mixture modeling). These theoretical and empirical contributions combine to make a compelling case for using variance reduction in stochastic Monte Carlo methods.

****

On Regularizing Rademacher Observation Losses

Richard Nock

It has recently been shown that supervised learning linear classifiers with two of the most popular losses, the logistic and square loss, is equivalent to optimizing an equivalent loss over sufficient statistics about the class: Rademacher observations (rados). It has also been shown that learning over rados brings solutions to two prominent problems for which the state of the art of learning from examples can be comparatively inferior and in fact less convenient: protecting and learning from private examples, learning from distributed datasets without entity resolution. Bis repetita placent: the two proofs of equivalence are different and rely on specific properties of the corresponding losses, so whether these can be unified and generalized inevitably comes to mind. This is our first contribution: we show how they can be fit into the same theory for the equivalence between example and rado losses. As a second contribution, we show that the generalization unveils a surprising new connection to regularized learning, and in particular a sufficient condition under which regularizing the loss over examples is equivalent to regularizing the rados (i.e. the data) in the equivalent rado loss, in such a way that an efficient algorithm for one regularized rado loss may be as efficient when changing the regularizer. This is our third contribution: we give a formal boosting algorithm for the regularized exponential rado-loss which boost with any of the ridge, lasso, \slope, l_\infty, or elastic nets, using the same master routine for all. Because the regularized exponential rado-loss is the equivalent of the regularized logistic loss over examples we obtain the first efficient proxy to the minimisation of the regularized logistic loss over examples using such a wide spectrum of regularizers. Experiments with a readily available code display that regularization significantly improves rado-based learning and compares favourably with example-based learning.

****

A scalable end-to-end Gaussian process adapter for irregularly sampled time series classification

Steven Cheng-Xian Li, Benjamin M. Marlin

We present a general framework for classification of sparse and irregularly-sampled time series. The properties of such time series can result in substantial uncertainty about the values of the underlying temporal processes, while making the data difficult to deal with using standard classification methods that assume fixed-dimensional feature spaces. To address these challenges, we propose an uncertainty-aware classification framework based on a special computational layer we refer to as the Gaussian process adapter that can connect irregularly sampled time series data to any black-box classifier learnable using gradient descent. We show how to scale up the required computations based on combining the structured kernel interpolation framework and the Lanczos approximation method, and how to discriminatively train the Gaussian process adapter in combination with a number of classifiers end-to-end using backpropagation.

****

Learning User Perceived Clusters with Feature-Level Supervision

Ting-Yu Cheng, Guiguan Lin, xinyang gong, Kang-Jun Liu, Shan-Hung (Brandon) Wu

Semi-supervised clustering algorithms have been proposed to identify data clusters that align with user perceived ones via the aid of side information such as seeds or pairwise constrains. However, traditional side information is mostly at the instance level and subject to the sampling bias, where non-randomly sampled instances in the supervision can mislead the algorithms to wrong clusters. In this paper, we propose learning from the feature-level supervision. We show that this kind of supervision can be easily obtained in the form of perception vectors in many applications. Then we present novel algorithms, called Perception Embedded (PE) clustering, that exploit the perception vectors as well as traditional side information to find clusters perceived by the user. Extensive experiments are conducted on real datasets and the results demonstrate the effectiveness of PE empirically.

****

Equality of Opportunity in Supervised Learning

Moritz Hardt, Eric Price, Eric Price, Nati Srebro

We propose a criterion for discrimination against a specified sensitive attribute in supervised learning, where the goal is to predict some target based on available features. Assuming data about the predictor, target, and membership in the protected group are available, we show how to optimally adjust any learned predictor so as to remove discrimination according to our definition. Our framework also improves incentives by shifting the cost of poor classification from disadvantaged groups to the decision maker, who can respond by improving the classification accuracy.

****

Kernel Observers: Systems-Theoretic Modeling and Inference of Spatiotemporally Evolving Processes

Hassan A. Kingravi, Harshal R. Maske, Girish Chowdhary

We consider the problem of estimating the latent state of a spatiotemporally evolving continuous function using very few sensor measurements. We show that layering a dynamical systems prior over temporal evolution of weights of a kernel model is a valid approach to spatiotemporal modeling that does not necessarily require the design of complex nonstationary kernels. Furthermore, we show that such a predictive model can be utilized to determine sensing locations that guarantee that the hidden state of the phenomena can be recovered with very few measurements. We provide sufficient conditions on the number and spatial location of samples required to guarantee state recovery, and provide a lower bound on the minimum number of samples required to robustly infer the hidden states. Our approach outperforms existing methods in numerical experiments.

****

Hierarchical Question-Image Co-Attention for Visual Question Answering

Jiasen Lu, Jianwei Yang, Dhruv Batra, Devi Parikh

A number of recent works have proposed attention models for Visual Question Answering (VQA) that generate spatial maps highlighting image regions relevant to answering the question. In this paper, we argue that in addition to modeling "where to look" or visual attention, it is equally important to model "what words to listen to" or question attention. We present a novel co-attention model for VQA that jointly reasons about image and question attention. In addition, our model reasons about the question (and consequently the image via the co-attention mechanism) in a hierarchical fashion via a novel 1-dimensional convolution neural networks (CNN). Our model improves the state-of-the-art on the VQA dataset from 60.3% to 60.5%, and from 61.6% to 63.3% on the COCO-QA dataset. By using ResNet, the performance is further improved to 62.1% for VQA and 65.4% for COCO-QA.

****

Double Thompson Sampling for Dueling Bandits

Huasen Wu, Xin 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.

****

A state-space model of cross-region dynamic connectivity in MEG/EEG
Ying Yang, Elissa Aminoff, Michael Tarr, Robert E. Kass

Cross-region dynamic connectivity, which describes spatio-temporal dependence of neural activity among multiple brain regions of interest (ROIs), can provide important information for understanding cognition. For estimating such connectivity, magnetoencephalography (MEG) and electroencephalography (EEG) are well-suited tools because of their millisecond temporal resolution. However, localizing source activity in the brain requires solving an under-determined linear problem. In typical two-step approaches, researchers first solve the linear problem with general priors assuming independence across ROIs, and secondly quantify cross-region connectivity. In this work, we propose a one-step state-space model to improve estimation of dynamic connectivity. The model treats the mean activity in individual ROIs as the state variable, and describes non-stationary dynamic dependence across ROIs using time-varying auto-regression. Compared with a two-step method, which first obtains the commonly used minimum-norm estimates of source activity, and then fits the auto-regressive model, our state-space model yielded smaller estimation errors on simulated data where the model assumptions held. When applied on empirical MEG data from one participant in a scene-processing experiment, our state-space model also demonstrated intriguing preliminary results, indicating leading and lagged linear dependence between the early visual cortex and a higher-level scene-sensitive region, which could reflect feed-forward and feedback information flow within the visual cortex during scene processing.

****

Using Fast Weights to Attend to the Recent Past
Jimmy Ba, Geoffrey E. Hinton, Volodymyr Mnih, Joel Z. Leibo, Catalin Ionescu

Until recently, research on artificial neural networks was largely restricted to systems with only two types of variable: Neural activities that represent the current or recent input and weights that learn to capture regularities among inputs, outputs and payoffs. There is no good reason for this restriction. Synapses have dynamics at many different time-scales and this suggests that artificial neural networks might benefit from variables that change slower than activities but much faster than the standard weights. These ``fast weights'' can be used to store temporary memories of the recent past and they provide a neurally plausible way of implementing the type of attention to the past that has recently proven helpful in sequence-to-sequence models. By using fast weights we can avoid the need to store copies of neural activity patterns.

****

High-Rank Matrix Completion and Clustering under Self-Expressive Models
Ehsan Elhamifar

We propose efficient algorithms for simultaneous clustering and completion of incomplete high-dimensional data that lie in a union of low-dimensional subspaces. We cast the problem as finding a completion of the data matrix so that each point can be reconstructed as a linear or affine combination of a few data points. Since the problem is NP-hard, we propose a lifting framework and reformulate the problem as a group-sparse recovery of each incomplete data point in a dictionary built using incomplete data, subject to rank-one constraints. To solve the problem efficiently, we propose a rank pursuit algorithm and a convex relaxation. The solution of our algorithms recover missing entries and provides a similarity matrix for clustering. Our algorithms can deal with both low-rank and high-rank matrices, does not suffer from initialization, does not need to know dimensions of subspaces and can work with a small number of data points. By extensive experiments on synthetic data and real problems of video motion segmentation and completion of motion capture data, we show that when the data matrix is low-rank, our algorithm performs on par with or better than low-rank matrix completion methods, while for high-rank data matrices, our method significantly outperforms exis ting algorithms.

****

Adaptive Newton Method for Empirical Risk Minimization to Statistical Accuracy

Aryan Mokhtari, Hadi Daneshmand, Aurelien Lucchi, Thomas Hofmann, Alejandro Ribeiro

We consider empirical risk minimization for large-scale datasets. We introduce Ada Newton as an adaptive algorithm that uses Newton's method with adaptive sample sizes. The main idea of Ada Newton is to increase the size of the training set by a factor larger than one in a way that the minimization variable for the current training set is in the local neighborhood of the optimal argument of the next training set. This allows to exploit the quadratic convergence property of Newton's method and reach the statistical accuracy of each training set with only one iteration of Newton's method. We show theoretically that we can iteratively increase the sample size while applying single Newton iterations without line search and staying within the statistical accuracy of the regularized empirical risk. In particular, we can double the size of the training set in each iteration when the number of samples is sufficiently large. Numerical experiments on various datasets confirm the possibility of increasing the sample size by factor 2 at each iteration which implies that Ada Newton achieves the statistical accuracy of the full training set with about two passes over the dataset.

****

Yggdrasil: An Optimized System for Training Deep Decision Trees at Scale

Firas Abuzaid, Joseph K. Bradley, Feynman T. Liang, Andrew Feng, Lee Yang, Matei Zaharia, Ameet S. Talwalkar

Deep distributed decision trees and tree ensembles have grown in importance due to the need to model increasingly large datasets. However, PLANET, the standard distributed tree learning algorithm implemented in systems such as \xgboost and Spark MLlib, scales poorly as data dimensionality and tree depths grow. We present Yggdrasil, a new distributed tree learning method that outperforms existing methods by up to 24x. Unlike PLANET, Yggdrasil is based on vertical partitioning of the data (i.e., partitioning by feature), along with a set of optimized data structures to reduce the CPU and communication costs of training. Yggdrasil (1) trains directly on compressed data for compressible features and labels; (2) introduces efficient data structures for training on uncompressed data; and (3) minimizes communication between nodes by using sparse bitvectors. Moreover, while PLANET approximates split points through feature binning, Yggdrasil does not require binning, and we analytically characterize the impact of this approximation. We evaluate Yggdrasil against the MNIST 8M dataset and a high-dimensional dataset at Yahoo; for both, Yggdrasil is faster by up to an order of magnitude.

****

Adaptive Maximization of Pointwise Submodular Functions With Budget Constraint

Nguyen Cuong, Huan Xu

We study the worst-case adaptive optimization problem with budget constraint that is useful for modeling various practical applications in artificial intelligence and machine learning. We investigate the near-optimality of greedy algorithms for this problem with both modular and non-modular cost functions. In both cases, we prove that two simple greedy algorithms are not near-optimal but the best between them is near-optimal if the utility function satisfies pointwise submodularity and pointwise cost-sensitive submodularity respectively. This implies a combined algorithm that is near-optimal with respect to the optimal algorithm that uses half of the budget. We discuss applications of our theoretical results and also report experiments comparing the greedy algorithms on the active learning problem.

****

Guided Policy Search via Approximate Mirror Descent

William H. Montgomery, Sergey Levine

Guided policy search algorithms can be used to optimize complex nonlinear policies, such as deep neural networks, without directly computing policy gradients in the high-dimensional parameter space. Instead, these methods use supervised learning to train the policy to mimic a “teacher” algorithm, such as a trajectory o ptimizer or a trajectory-centric reinforcement learning method. Guided policy search methods provide asymptotic local convergence guarantees by construction, but it is not clear how much the policy improves within a small, finite number of iterations. We show that guided policy search algorithms can be interpreted as an approximate variant of mirror descent, where the projection onto the constraint manifold is not exact. We derive a new guided policy search algorithm that is simpler and provides appealing improvement and convergence guarantees in simplified convex and linear settings, and show that in the more general nonlinear setting, the error in the projection step can be bounded. We provide empirical results on several simulated robotic manipulation tasks that show that our method is stable and achieves similar or better performance when compared to prior guided policy search methods, with a simpler formulation and fewer hyperparameters.

****

On Graph Reconstruction via Empirical Risk Minimization: Fast Learning Rates and Scalability

Guillaume Papa, Aurélien Bellet, Stephan Clémençon

The problem of predicting connections between a set of data points finds many applications, in systems biology and social network analysis among others. This paper focuses on the \textit{graph reconstruction} problem, where the prediction rule is obtained by minimizing the average error over all n(n-1)/2 possible pairs of the n nodes of a training graph. Our first contribution is to derive learning rates of order O(log n / n) for this problem, significantly improving upon the slow rates of order $O(1/\sqrt{n})$ established in the seminal work of Biau & Bleakley (2006). Strikingly, these fast rates are universal, in contrast to similar results known for other statistical learning problems (e.g., classification, density level set estimation, ranking, clustering) which require strong assumptions on the distribution of the data. Motivated by applications to large graphs, our second contribution deals with the computational complexity of graph reconstruction. Specifically, we investigate to which extent the learning rates can be preserved when replacing the empirical reconstruction risk by a computationally cheaper Monte-Carlo version, obtained by sampling with replacement $B << n^2$ pairs of nodes. Finally, we illustrate our theoretical results by numerical experiments on synthetic and real graphs.

****

Geometric Dirichlet Means Algorithm for topic inference

Mikhail Yurochkin, XuanLong Nguyen

We propose a geometric algorithm for topic learning and inference that is built on the convex geometry of topics arising from the Latent Dirichlet Allocation (LDA) model and its nonparametric extensions. To this end we study the optimization of a geometric loss function, which is a surrogate to the LDA's likelihood. Our method involves a fast optimization based weighted clustering procedure augmented with geometric corrections, which overcomes the computational and statistical inefficiencies encountered by other techniques based on Gibbs sampling and variational inference, while achieving the accuracy comparable to that of a Gibbs sampler. The topic estimates produced by our method are shown to be statistically consistent under some conditions. The algorithm is evaluated with extensive experiments on simulated and real data.

****

Learned Region Sparsity and Diversity Also Predicts Visual Attention

Zijun Wei, Hossein Adeli, Minh Hoai Nguyen, Greg Zelinsky, Dimitris Samaras

Learned region sparsity has achieved state-of-the-art performance in classification tasks by exploiting and integrating a sparse set of local information into global decisions. The underlying mechanism resembles how people sample information from an image with their eye movements when making similar decisions. In this paper we incorporate the biologically plausible mechanism of Inhibition of Return into the learned region sparsity model, thereby imposing diversity on the selected regions. We investigate how these mechanisms of sparsity and diversity relate to visual attention by testing our model on three different types of visual search tasks. We report state-of-the-art results in predicting the locations of human gaze fixations, even though our model is trained only on image-level labels without object location annotations. Notably, the classification performance of the extended model remains the same as the original. This work suggests a new computational perspective on visual attention mechanisms and shows how the inclusion of attention-based mechanisms can improve computer vision techniques.

****

Deep Learning Models of the Retinal Response to Natural Scenes
Lane McIntosh, Niru Maheswaranathan, Aran Nayebi, Surya Ganguli, Stephen Baccus

A central challenge in sensory neuroscience is to understand neural computations and circuit mechanisms that underlie the encoding of ethologically relevant, natural stimuli. In multilayered neural circuits, nonlinear processes such as synaptic transmission and spiking dynamics present a significant obstacle to the creation of accurate computational models of responses to natural stimuli. Here we demonstrate that deep convolutional neural networks (CNNs) capture retinal responses to natural scenes nearly to within the variability of a cell's response, and are markedly more accurate than linear-nonlinear (LN) models and Generalized Linear Models (GLMs). Moreover, we find two additional surprising properties of CNNs: they are less susceptible to overfitting than their LN counterparts when trained on small amounts of data, and generalize better when tested on stimuli drawn from a different distribution (e.g. between natural scenes and white noise). An examination of the learned CNNs reveals several properties. First, a richer set of feature maps is necessary for predicting the responses to natural scenes compared to white noise. Second, temporally precise responses to slowly varying inputs originate from feedforward inhibition, similar to known retinal mechanisms. Third, the injection of latent noise sources in intermediate layers enables our model to capture the sub-Poisson spiking variability observed in retinal ganglion cells. Fourth, augmenting our CNNs with recurrent lateral connections enables them to capture contrast adaptation as an emergent property of accurately describing retinal responses to natural scenes. These methods can be readily generalized to other sensory modalities and stimulus ensembles. Overall, this work demonstrates that CNNs not only accurately capture sensory circuit responses to natural scenes, but also can yield information about the circuit's internal structure and function.

****

Batched Gaussian Process Bandit Optimization via Determinantal Point Processes
Tarun Kathuria, Amit Deshpande, Pushmeet Kohli

Gaussian Process bandit optimization has emerged as a powerful tool for optimizing noisy black box functions. One example in machine learning is hyper-parameter optimization where each evaluation of the target function may require training a model which may involve days or even weeks of computation. Most methods for this so-called “Bayesian optimization” only allow sequential exploration of the parameter space. However, it is often desirable to propose batches or sets of parameter values to explore simultaneously, especially when there are large parallel processing facilities at our disposal. Batch methods require modeling the interaction between the different evaluations in the batch, which can be expensive in complex scenarios. In this paper, we propose a new approach for parallelizing Bayesian optimization by modeling the diversity of a batch via Determinantal point processes (DPPs) whose kernels are learned automatically. This allows us to generalize a previous result as well as prove better regret bounds based on DPP sampling. Our experiments on a variety of synthetic and real-world robotics and hyper-parameter optimization tasks indicate that our DPP-based methods, especially those based on DPP sampling, outperform state-of-the-art methods.

****

Inference by Reparameterization in Neural Population Codes
Rajkumar Vasudeva Raju, Zachary Pitkow

Behavioral experiments on humans and animals suggest that the brain performs probabilistic inference to interpret its environment. Here we present a new general-purpose, biologically-plausible neural implementation of approximate inference. The neural network represents uncertainty using Probabilistic Population Codes (PPCs), which are distributed neural representations that naturally encode probability distributions, and support marginalization and evidence integration in a

biologically-plausible manner. By connecting multiple PPCs together as a probabilistic graphical model, we represent multivariate probability distributions. Approximate inference in graphical models can be accomplished by message-passing algorithms that disseminate local information throughout the graph. An attractive and often accurate example of such an algorithm is Loopy Belief Propagation (LBP), which uses local marginalization and evidence integration operations to perform approximate inference efficiently even for complex models. Unfortunately, a subtle feature of LBP renders it neurally implausible. However, LBP can be elegantly reformulated as a sequence of Tree-based Reparameterizations (TRP) of the graphical model. We re-express the TRP updates as a nonlinear dynamical system with both fast and slow timescales, and show that this produces a neurally plausible solution. By combining all of these ideas, we show that a network of PPCs can represent multivariate probability distributions and implement the TRP updates to perform probabilistic inference. Simulations with Gaussian graphical models demonstrate that the neural network inference quality is comparable to the direct evaluation of LBP and robust to noise, and thus provides a promising mechanism for general probabilistic inference in the population codes of the brain.
****

Blind Attacks on Machine Learners

Alex Beatson, Zhaoran Wang, Han Liu

The importance of studying the robustness of learners to malicious data is well established. While much work has been done establishing both robust estimators and effective data injection attacks when the attacker is omniscient, the ability of an attacker to provably harm learning while having access to little information is largely unstudied. We study the potential of a “blind attacker” to provably limit a learner’s performance by data injection attack without observing the learner’s training set or any parameter of the distribution from which it is drawn. We provide examples of simple yet effective attacks in two settings: firstly, where an “informed learner” knows the strategy chosen by the attacker, and secondly, where a “blind learner” knows only the proportion of malicious data and some family to which the malicious distribution chosen by the attacker belongs. For each attack, we analyze minimax rates of convergence and establish lower bounds on the learner’s minimax risk, exhibiting limits on a learner’s ability to learn under data injection attack even when the attacker is “blind”.
****

Learning Deep Parsimonious Representations

Renjie Liao, Alex Schwing, Richard Zemel, Raquel Urtasun

In this paper we aim at facilitating generalization for deep networks while supporting interpretability of the learned representations. Towards this goal, we propose a clustering based regularization that encourages parsimonious representations. Our k-means style objective is easy to optimize and flexible supporting various forms of clustering, including sample and spatial clustering as well as co-clustering. We demonstrate the effectiveness of our approach on the tasks of unsupervised learning, classification, fine grained categorization and zero-shot learning.
****

Scalable Adaptive Stochastic Optimization Using Random Projections

Gabriel Krummenacher, Brian McWilliams, Yannic Kilcher, Joachim M. Buhmann, Nicolai Meinshausen

Adaptive stochastic gradient methods such as AdaGrad have gained popularity in particular for training deep neural networks. The most commonly used and studied variant maintains a diagonal matrix approximation to second order information by accumulating past gradients which are used to tune the step size adaptively. In certain situations the full-matrix variant of AdaGrad is expected to attain better performance, however in high dimensions it is computationally impractical. We present Ada-LR and RadaGrad two computationally efficient approximations to full-matrix AdaGrad based on randomized dimensionality reduction. They are able to capture dependencies between features and achieve similar performance to full-matrix AdaGrad but at a much smaller computational cost. We show that the regret of Ada-LR is close to the regret of full-matrix AdaGrad which can have an up-to exponentially smaller dependence on the dimension than the diagonal variant. Empirically, we show that Ada-LR and RadaGrad perform similarly to full-matrix AdaGrad. On the task of training convolutional neural networks as well as recurrent neural networks, RadaGrad achieves faster convergence than diagonal AdaGrad.

****

Graphons, mergeons, and so on!

Justin Eldridge, Mikhail Belkin, Yusu Wang

In this work we develop a theory of hierarchical clustering for graphs. Our modelling assumption is that graphs are sampled from a graphon, which is a powerful and general model for generating graphs and analyzing large networks. Graphons are a far richer class of graph models than stochastic blockmodels, the primary setting for recent progress in the statistical theory of graph clustering. We define what it means for an algorithm to produce the ``correct" clustering, give sufficient conditions in which a method is statistically consistent, and provide an explicit algorithm satisfying these properties.

****

Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, Adam T. Kalai

The blind application of machine learning runs the risk of amplifying biases present in data. Such a danger is facing us with word embedding, a popular framework to represent text data as vectors which has been used in many machine learning and natural language processing tasks. We show that even word embeddings trained on Google News articles exhibit female/male gender stereotypes to a disturbing extent. This raises concerns because their widespread use, as we describe, often tends to amplify these biases. Geometrically, gender bias is first shown to be captured by a direction in the word embedding. Second, gender neutral words are shown to be linearly separable from gender definition words in the word embedding. Using these properties, we provide a methodology for modifying an embedding to remove gender stereotypes, such as the association between the words receptionist and female, while maintaining desired associations such as between the words queen and female. Using crowd-worker evaluation as well as standard benchmarks, we empirically demonstrate that our algorithms significantly reduce gender bias in embeddings while preserving the its useful properties such as the ability to cluster related concepts and to solve analogy tasks. The resulting embeddings can be used in applications without amplifying gender bias.

****

Memory-Efficient Backpropagation Through Time

Audrunas Gruslys, Remi Munos, Ivo Danihelka, Marc Lanctot, Alex Graves

We propose a novel approach to reduce memory consumption of the backpropagation through time (BPTT) algorithm when training recurrent neural networks (RNNs). Our approach uses dynamic programming to balance a trade-off between caching of intermediate results and recomputation. The algorithm is capable of tightly fitting within almost any user-set memory budget while finding an optimal execution policy minimizing the computational cost. Computational devices have limited memory capacity and maximizing a computational performance given a fixed memory budget is a practical use-case. We provide asymptotic computational upper bounds for various regimes. The algorithm is particularly effective for long sequences. For sequences of length 1000, our algorithm saves 95\% of memory usage while using only one third more time per iteration than the standard BPTT.

****

Solving Marginal MAP Problems with NP Oracles and Parity Constraints

Yexiang Xue, Zhiyuan Li, Stefano Ermon, Carla P. Gomes, Bart Selman

Arising from many applications at the intersection of decision-making and machine learning, Marginal Maximum A Posteriori (Marginal MAP) problems unify the two main classes of inference, namely maximization (optimization) and marginal inference (counting), and are believed to have higher complexity than both of them. We propose XORMMAP, a novel approach to solve the Marginal MAP problem, which represents the intractable counting subproblem with queries to NP oracles, subject to additional parity constraints. XORMMAP provides a constant factor approximati on to the Marginal MAP problem, by encoding it as a single optimization in a polynomial size of the original problem. We evaluate our approach in several machine learning and decision-making applications, and show that our approach outperforms several state-of-the-art Marginal MAP solvers.

****

Differential Privacy without Sensitivity

Kentaro Minami, HItomi Arai, Issei Sato, Hiroshi Nakagawa

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 Smoothed Online Multi-Task Learning

Keerthiram Murugesan, Hanxiao Liu, Jaime Carbonell, Yiming Yang

This paper addresses the challenge of jointly learning both the per-task model parameters and the inter-task relationships in a multi-task online learning setting. The proposed algorithm features probabilistic interpretation, efficient updating rules and flexible modulation on whether learners focus on their specific task or on jointly address all tasks. The paper also proves a sub-linear regret bound as compared to the best linear predictor in hindsight. Experiments over three multi-task learning benchmark datasets show advantageous performance of the proposed approach over several state-of-the-art online multi-task learning baselines.

****

Efficient and Robust Spiking Neural Circuit for Navigation Inspired by Echolocating Bats

Pulkit Tandon, Yash H. Malviya, Bipin Rajendran

We demonstrate a spiking neural circuit for azimuth angle detection inspired by the echolocation circuits of the Horseshoe bat Rhinolophus ferrumequinum and utilize it to devise a model for navigation and target tracking, capturing several key aspects of information transmission in biology. Our network, using only a simple local-information based sensor implementing the cardioid angular gain function, operates at biological spike rate of 10 Hz. The network tracks large angular targets (60 degrees) within 1 sec with a 10% RMS error. We study the navigational ability of our model for foraging and target localization tasks in a forest of obstacles and show that our network requires less than 200X   spike-triggered decisions, while suffering only a 1% loss in performance compared to a proportional-integral-derivative controller, in the presence of 50% additive noise. Superior performance can be obtained at a higher average spike rate of 100 Hz and 1000 Hz, but even the accelerated networks requires 20X and 10X lesser decisions respectively, demonstrating the superior computational efficiency of bio-inspired information processing systems.

****

Optimal Cluster Recovery in the Labeled Stochastic Block Model

Se-Young Yun, 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.

****

Relevant sparse codes with variational information bottleneck

Matthew Chalk, Olivier Marre, Gasper Tkacik

In many applications, it is desirable to extract only the relevant aspects of data. A principled way to do this is the information bottleneck (IB) method, where one seeks a code that maximises information about a relevance variable, Y, while constraining the information encoded about the original data, X. Unfortunately however, the IB method is computationally demanding when data are high-dimensional and/or non-gaussian. Here we propose an approximate variational scheme for maximising a lower bound on the IB objective, analogous to variational EM. Using this method, we derive an IB algorithm to recover features that are both releva nt and sparse. Finally, we demonstrate how kernelised versions of the algorithm can be used to address a broad range of problems with non-linear relation between X and Y.

****

Learning What and Where to Draw

Scott E. Reed, Zeynep Akata, Santosh Mohan, Samuel Tenka, Bernt Schiele, Honglak Lee

Generative Adversarial Networks (GANs) have recently demonstrated the capability to synthesize compelling real-world images, such as room interiors, album covers, manga, faces, birds, and flowers. While existing models can synthesize images based on global constraints such as a class label or caption, they do not provide control over pose or object location. We propose a new model, the Generative Adversarial What-Where Network (GAWWN), that synthesizes images given instructions describing what content to draw in which location. We show high-quality 128 × 128 image synthesis on the Caltech-UCSD Birds dataset, conditioned on both informal text descriptions and also object location. Our system exposes control over both the bounding box around the bird and its constituent parts. By modeling the conditional distributions over part locations, our system also enables conditioning on arbitrary subsets of parts (e.g. only the beak and tail), yielding an efficient interface for picking part locations.

****

A Bio-inspired Redundant Sensing Architecture

Anh Tuan Nguyen, Jian Xu, Zhi Yang

Sensing is the process of deriving signals from the environment that allows artificial systems to interact with the physical world. The Shannon theorem specifies the maximum rate at which information can be acquired. However, this upper bound is hard to achieve in many man-made systems. The biological visual systems, on the other hand, have highly efficient signal representation and processing mechanisms that allow precise sensing. In this work, we argue that redundancy is one of the critical characteristics for such superior performance. We show architectural advantages by utilizing redundant sensing, including correction of mismatch error and significant precision enhancement. For a proof-of-concept demonstration, we have designed a heuristic-based analog-to-digital converter - a zero-dimensional quantizer. Through Monte Carlo simulation with the error probabilistic distribution as a priori, the performance approaching the Shannon limit is feasible. In actual measurements without knowing the error distribution, we observe at least 2-bit extra precision. The results may also help explain biological processes including the dominance of binocular vision, the functional roles of the fixational eye movements, and the structural mechanisms allowing hyperacuity.

****

Bayesian Optimization with Robust Bayesian Neural Networks

Jost Tobias Springenberg, Aaron Klein, Stefan Falkner, Frank Hutter

Bayesian optimization is a prominent method for optimizing expensive to evaluate black-box functions that is prominently applied to tuning the hyperparameters of machine learning algorithms. Despite its successes, the prototypical Bayesian optimization approach - using Gaussian process models - does not scale well to either many hyperparameters or many function evaluations. Attacking this lack of scalability and flexibility is thus one of the key challenges of the field. We present a general approach for using flexible parametric models (neural networks) for Bayesian optimization, staying as close to a truly Bayesian treatment as possible. We obtain scalability through stochastic gradient Hamiltonian Monte Carlo, whose robustness we improve via a scale adaptation. Experiments including multi-task Bayesian optimization with 21 tasks, parallel optimization of deep neural networks and deep reinforcement learning show the power and flexibility of this approach.

****

Statistical Inference for Cluster Trees

Jisu KIM, Yen-Chi Chen, Sivaraman Balakrishnan, Alessandro Rinaldo, Larry Wasserman

A cluster tree provides an intuitive summary of a density function that reveals essential structure about the high-density clusters. The true cluster tree is estimated from a finite sample from an unknown true density. This paper addresses the basic question of quantifying our uncertainty by assessing the statistical significance of different features of an empirical cluster tree. We first study a variety of metrics that can be used to compare different trees, analyzing their properties and assessing their suitability for our inference task. We then propose methods to construct and summarize confidence sets for the unknown true cluster tree. We introduce a partial ordering on cluster trees which we use to prune some of the statistically insignificant features of the empirical tree, yielding interpretable and parsimonious cluster trees. Finally, we provide a variety of simulations to illustrate our proposed methods and furthermore demonstrate their utility in the analysis of a Graft-versus-Host Disease (GvHD) data set.

****

Combinatorial Multi-Armed Bandit with General Reward Functions

Wei Chen, Wei Hu, Fu Li, Jian Li, Yu Liu, Pinyan Lu

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 Sensor Multiplexing Design through Back-propagation

Ayan Chakrabarti

Recent progress on many imaging and vision tasks has been driven by the use of deep feed-forward neural networks, which are trained by propagating gradients of a loss defined on the final output, back through the network up to the first layer that operates directly on the image. We propose back-propagating one step further---to learn camera sensor designs jointly with networks that carry out inference on the images they capture. In this paper, we specifically consider the design and inference problems in a typical color camera---where the sensor is able to measure only one color channel at each pixel location, and computational inference is required to reconstruct a full color image. We learn the camera sensor's color multiplexing pattern by encoding it as layer whose learnable weights determine which color channel, from among a fixed set, will be measured at each location. These weights are jointly trained with those of a reconstruction network that operates on the corresponding sensor measurements to produce a full color image. Our network achieves significant improvements in accuracy over the traditional Bayer pattern used in most color cameras. It automatically learns to employ a sparse color measurement approach similar to that of a recent design, and moreover, improves upon that design by learning an optimal layout for these measurements.

****

Short-Dot: Computing Large Linear Transforms Distributedly Using Coded Short Dot Products

Sanghamitra Dutta, Viveck Cadambe, Pulkit Grover

Faced with saturation of Moore's law and increasing size and dimension of data, system designers have increasingly resorted to parallel and distributed computing to reduce computation time of machine-learning algorithms. However, distributed computing is often bottle necked by a small fraction of slow processors called "stragglers" that reduce the speed of computation because the fusion node has to wait for all processors to complete their processing. To combat the effect of stragglers, recent literature proposes introducing redundancy in computations across processors, e.g., using repetition-based strategies or erasure codes. The fusion node can exploit this redundancy by completing the computation using outputs from only a subset of the processors, ignoring the stragglers. In this paper, we propose a novel technique - that we call "Short-Dot" - to introduce redundant computations in a coding theory inspired fashion, for computing linear transforms of long vectors. Instead of computing long dot products as required in the original linear transform, we construct a larger number of redundant and short dot products that can be computed more efficiently at individual processors. Further, only a subset of these short dot products are required at the fusion node to finish the computation successfully. We demonstrate through probabilistic analysis as well as experiments on computing clusters that Short-Dot offers significant speed-up compared to existing techniques. We also derive trade-offs between the length of the dot-products and the resilience to stragglers (number of processors required to finish), for any such strategy and compare it to that achieved by our strategy.

****

SEBOOST - Boosting Stochastic Learning Using Subspace Optimization Techniques
Elad Richardson, Rom Herskovitz, Boris Ginsburg, Michael Zibulevsky

We present SEBOOST, a technique for boosting the performance of existing stochastic optimization methods. SEBOOST applies a secondary optimization process in the subspace spanned by the last steps and descent directions. The method was inspired by the SESOP optimization method for large-scale problems, and has been adapted for the stochastic learning framework. It can be applied on top of any existing optimization method with no need to tweak the internal algorithm. We show that the method is able to boost the performance of different algorithms, and make them more robust to changes in their hyper-parameters. As the boosting steps of SEBOOST are applied between large sets of descent steps, the additional subspace optimization hardly increases the overall computational burden. We introduce two hyper-parameters that control the balance between the baseline method and the secondary optimization process. The method was evaluated on several deep learning tasks, demonstrating promising results.

****

VIME: Variational Information Maximizing Exploration
Rein Houthooft, Xi Chen, Xi Chen, Yan Duan, John Schulman, Filip De Turck, Pieter Abbeel

Scalable and effective exploration remains a key challenge in reinforcement learning (RL). While there are methods with optimality guarantees in the setting of discrete state and action spaces, these methods cannot be applied in high-dimensional deep RL scenarios. As such, most contemporary RL relies on simple heuristics such as epsilon-greedy exploration or adding Gaussian noise to the controls. This paper introduces Variational Information Maximizing Exploration (VIME), an exploration strategy based on maximization of information gain about the agent's belief of environment dynamics. We propose a practical implementation, using variational inference in Bayesian neural networks which efficiently handles continuous state and action spaces. VIME modifies the MDP reward function, and can be applied with several different underlying RL algorithms. We demonstrate that VIME achieves significantly better performance compared to heuristic exploration methods across a variety of continuous control tasks and algorithms, including tasks with very sparse rewards.

****

Toward Deeper Understanding of Neural Networks: The Power of Initialization and a Dual View on Expressivity
Amit Daniely, Roy Frostig, Yoram Singer

We develop a general duality between neural networks and compositional kernel Hilbert spaces. We introduce the notion of a computation skeleton, an acyclic graph that succinctly describes both a family of neural networks and a kernel space. Random neural networks are generated from a skeleton through node replication followed by sampling from a normal distribution to assign weights. The kernel space consists of functions that arise by compositions, averaging, and non-linear transformations governed by the skeleton's graph topology and activation functions. We prove that random networks induce representations which approximate the kernel space. In particular, it follows that random weight initialization often yields a favorable starting point for optimization despite the worst-case intractability of training neural networks.

****

Unsupervised Domain Adaptation with Residual Transfer Networks
Mingsheng Long, Han Zhu, Jianmin Wang, Michael I. Jordan

The recent success of deep neural networks relies on massive amounts of labeled data. For a target task where labeled data is unavailable, domain adaptation can transfer a learner from a different source domain. In this paper, we propose a new approach to domain adaptation in deep networks that can jointly learn adaptive classifiers and transferable features from labeled data in the source domain and unlabeled data in the target domain. We relax a shared-classifier assumption made by previous methods and assume that the source classifier and target classifier differ by a residual function. We enable classifier adaptation by plugging several layers into deep network to explicitly learn the residual function with reference to the target classifier. We fuse features of multiple layers with tensor product and embed them into reproducing kernel Hilbert spaces to match distributions for feature adaptation. The adaptation can be achieved in most feed-forward models by extending them with new residual layers and loss functions, which can be trained efficiently via back-propagation. Empirical evidence shows that the new approach outperforms state of the art methods on standard domain adaptation benchmarks.

****

Stochastic Gradient MCMC with Stale Gradients

Changyou Chen, Nan Ding, Chunyuan Li, Yizhe Zhang, Lawrence Carin

Stochastic gradient MCMC (SG-MCMC) has played an important role in large-scale Bayesian learning, with well-developed theoretical convergence properties. In such applications of SG-MCMC, it is becoming increasingly popular to employ distributed systems, where stochastic gradients are computed based on some outdated parameters, yielding what are termed stale gradients. While stale gradients could be directly used in SG-MCMC, their impact on convergence properties has not been well studied. In this paper we develop theory to show that while the bias and MSE of an SG-MCMC algorithm depend on the staleness of stochastic gradients, its estimation variance (relative to the expected estimate, based on a prescribed number of samples) is independent of it. In a simple Bayesian distributed system with SG-MCMC, where stale gradients are computed asynchronously by a set of workers, our theory indicates a linear speedup on the decrease of estimation variance w.r.t. the number of workers. Experiments on synthetic data and deep neural networks validate our theory, demonstrating the effectiveness and scalability of SG-MCMC with stale gradients.

****

Efficient Nonparametric Smoothness Estimation

Shashank Singh, Simon S. Du, Barnabas Poczos

Sobolev quantities (norms, inner products, and distances) of probability density functions are important in the theory of nonparametric statistics, but have rarely been used in practice, partly due to a lack of practical estimators. They also include, as special cases, L^2 quantities which are used in many applications. We propose and analyze a family of estimators for Sobolev quantities of unknown probability density functions. We bound the finite-sample bias and variance of our estimators, finding that they are generally minimax rate-optimal. Our estimators are significantly more computationally tractable than previous estimators, and exhibit a statistical/computational trade-off allowing them to adapt to computational constraints. We also draw theoretical connections to recent work on fast two-sample testing and empirically validate our estimators on synthetic data.

****

Adversarial Multiclass Classification: A Risk Minimization Perspective

Rizal Fathony, Anqi Liu, Kaiser Asif, Brian Ziebart

Recently proposed adversarial classification methods have shown promising results for cost sensitive and multivariate losses. In contrast with empirical risk minimization (ERM) methods, which use convex surrogate losses to approximate the desired non-convex target loss function, adversarial methods minimize non-convex losses by treating the properties of the training data as being uncertain and worst case within a minimax game. Despite this difference in formulation, we recast adversarial classification under zero-one loss as an ERM method with a novel prescribed loss function. We demonstrate a number of theoretical and practical advantages over the very closely related hinge loss ERM methods. This establishes adversarial classification under the zero-one loss as a method that fills the lo ng standing gap in multiclass hinge loss classification, simultaneously guaranteeing Fisher consistency and universal consistency, while also providing dual parameter sparsity and high accuracy predictions in practice.

****

Long-term Causal Effects via Behavioral Game Theory

Panagiotis Toulis, David C. Parkes

Planned experiments are the gold standard in reliably comparing the causal effect of switching from a baseline policy to a new policy. % One critical shortcoming of classical experimental methods, however, is that they typically do not take into account the dynamic nature of response to policy changes. For instance, in an experiment where we seek to understand the effects of a new ad pricing policy on auction revenue, agents may adapt their bidding in response to the experimental pricing changes. Thus, causal effects of the new pricing policy after such adaptation period, the {\em long-term causal effects}, are not captured by the classical methodology even though they clearly are more indicative of the value of the new policy. % Here, we formalize a framework to define and estimate long-term causal effects of   policy changes in multiagent economies. Central to our approach is behavioral game theory, which we leverage   to formulate the ignorability assumptions that are necessary for causal inference. Under such assumptions we estimate long-term causal effects through a latent space approach, where a behavioral model of how agents act conditional on their latent behaviors is combined with a temporal model of how behaviors evolve over time.

****

Sampling for Bayesian Program Learning

Kevin Ellis, Armando Solar-Lezama, Josh Tenenbaum

Towards learning programs from data, we introduce the problem of   sampling programs from posterior distributions conditioned on that   data. Within this setting, we propose an algorithm that uses a   symbolic solver to efficiently sample programs. The proposal   combines constraint-based program synthesis with sampling via random   parity constraints. We give theoretical guarantees on how well the   samples approximate the true posterior, and have empirical results   showing the algorithm is efficient in practice, evaluating our   approach on 22 program learning problems in the domains of text   editing and computer-aided programming.

****

Unifying Count-Based Exploration and Intrinsic Motivation

Marc Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, Remi Munos

We consider an agent's uncertainty about its environment and the problem of generalizing this uncertainty across states. Specifically, we focus on the problem of exploration in non-tabular reinforcement learning. Drawing inspiration from the intrinsic motivation literature, we use density models to measure uncertainty, and propose a novel algorithm for deriving a pseudo-count from an arbitrary density model. This technique enables us to generalize count-based exploration algorithms to the non-tabular case. We apply our ideas to Atari 2600 games, providing sensible pseudo-counts from raw pixels. We transform these pseudo-counts into exploration bonuses and obtain significantly improved exploration in a number of hard games, including the infamously difficult Montezuma's Revenge.

****

Learning HMMs with Nonparametric Emissions via Spectral Decompositions of Continuous Matrices

Kirthevasan Kandasamy, Maruan Al-Shedivat, Eric P. Xing

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.

****

Observational-Interventional Priors for Dose-Response Learning

Ricardo Silva

Controlled interventions provide the most direct source of information for learn ing causal effects. In particular, a dose-response curve can be learned by varying the treatment level and observing the corresponding outcomes. However, interventions can be expensive and time-consuming. Observational data, where the treatment is not controlled by a known mechanism, is sometimes available. Under some strong assumptions, observational data allows for the estimation of dose-response curves. Estimating such curves nonparametrically is hard: sample sizes for controlled interventions may be small, while in the observational case a large number of measured confounders may need to be marginalized. In this paper, we introduce a hierarchical Gaussian process prior that constructs a distribution over the dose-response curve by learning from observational data, and reshapes the distribution with a nonparametric affine transform learned from controlled interventions. This function composition from different sources is shown to speed-up learning, which we demonstrate with a thorough sensitivity analysis and an application to modeling the effect of therapy on cognitive skills of premature infants.

****

Improved Error Bounds for Tree Representations of Metric Spaces

Samir Chowdhury, Facundo Mémoli, Zane T. Smith

Estimating optimal phylogenetic trees or hierarchical clustering trees from metric data is an important problem in evolutionary biology and data analysis. Intuitively, the goodness-of-fit of a metric space to a tree depends on its inherent treeness, as well as other metric properties such as intrinsic dimension. Existing algorithms for embedding metric spaces into tree metrics provide distortion bounds depending on cardinality. Because cardinality is a simple property of any set, we argue that such bounds do not fully capture the rich structure endowed by the metric. We consider an embedding of a metric space into a tree proposed by Gromov. By proving a stability result, we obtain an improved additive distortion bound depending only on the hyperbolicity and doubling dimension of the metric. We observe that Gromov's method is dual to the well-known single linkage hierarchical clustering (SLHC) method. By means of this duality, we are able to transport our results to the setting of SLHC, where such additive distortion bounds were previously unknown.

****

A Bayesian method for reducing bias in neural representational similarity analysis

Mingbo Cai, Nicolas W. Schuck, Jonathan W. Pillow, Yael Niv

In neuroscience, the similarity matrix of neural activity patterns in response to different sensory stimuli or under different cognitive states reflects the structure of neural representational space. Existing methods derive point estimations of neural activity patterns from noisy neural imaging data, and the similarity is calculated from these point estimations. We show that this approach translates structured noise from estimated patterns into spurious bias structure in the resulting similarity matrix, which is especially severe when signal-to-noise ratio is low and experimental conditions cannot be fully randomized in a cognitive task. We propose an alternative Bayesian framework for computing representational similarity in which we treat the covariance structure of neural activity patterns as a hyper-parameter in a generative model of the neural data, and directly estimate this covariance structure from imaging data while marginalizing over the unknown activity patterns. Converting the estimated covariance structure into a correlation matrix offers a much less biased estimate of neural representational similarity. Our method can also simultaneously estimate a signal-to-noise map that informs where the learned representational structure is supported more strongly, and the learned covariance matrix can be used as a structured prior to constrain Bayesian estimation of neural activity patterns. Our code is freely available in Brain Imaging Analysis Kit (Brainiak) (https://github.com/IntelPNI/brainiak), a python toolkit for brain imaging analysis.

****

Multistage Campaigning in Social Networks

Mehrdad Farajtabar, Xiaojing Ye, Sahar Harati, Le Song, Hongyuan Zha

We consider control problems for multi-stage campaigning over social networks. The dynamic programming framework is employed to balance the high present reward

and large penalty on low future outcome in the presence of extensive uncertainties. In particular, we establish theoretical foundations of optimal campaigning over social networks where the user activities are modeled as a multivariate Hawkes process, and we derive a time dependent linear relation between the intensity of exogenous events and several commonly used objective functions of campaigning. We further develop a convex dynamic programming framework for determining the optimal intervention policy that prescribes the required level of external drive at each stage for the desired campaigning result. Experiments on both synthetic data and the real-world MemeTracker dataset show that our algorithm can steer the user activities for optimal campaigning much more accurately than baselines.
****
Conditional Image Generation with PixelCNN Decoders
Aaron van den Oord, Nal Kalchbrenner, Lasse Espeholt, koray kavukcuoglu, Oriol Vinyals, Alex Graves
This work explores conditional image generation with a new image density model based on the PixelCNN architecture. The model can be conditioned on any vector, including descriptive labels or tags, or latent embeddings created by other networks. When conditioned on class labels from the ImageNet database, the model is able to generate diverse, realistic scenes representing distinct animals, objects, landscapes and structures. When conditioned on an embedding produced by a convolutional network given a single image of an unseen face, it generates a variety of new portraits of the same person with different facial expressions, poses and lighting conditions. We also show that conditional PixelCNN can serve as a powerful decoder in an image autoencoder. Additionally, the gated convolutional layers in the proposed model improve the log-likelihood of PixelCNN to match the state-of-the-art performance of PixelRNN on ImageNet, with greatly reduced computational cost.
****
Global Optimality of Local Search for Low Rank Matrix Recovery
Srinadh Bhojanapalli, Behnam Neyshabur, Nati Srebro
We show that there are no spurious local minima in the non-convex factorized parametrization of low-rank matrix recovery from incoherent linear measurements. With noisy measurements we show all local minima are very close to a global optimum. Together with a curvature bound at saddle points, this yields a polynomial time global convergence guarantee for stochastic gradient descent {\em from random initialization}.
****
Tensor Switching Networks
Chuan-Yung Tsai, Andrew M. Saxe, Andrew M. Saxe, David Cox
We present a novel neural network algorithm, the Tensor Switching (TS) network, which generalizes the Rectified Linear Unit (ReLU) nonlinearity to tensor-valued hidden units. The TS network copies its entire input vector to different locations in an expanded representation, with the location determined by its hidden unit activity. In this way, even a simple linear readout from the TS representation can implement a highly expressive deep-network-like function. The TS network hence avoids the vanishing gradient problem by construction, at the cost of larger representation size. We develop several methods to train the TS network, including equivalent kernels for infinitely wide and deep TS networks, a one-pass linear learning algorithm, and two backpropagation-inspired representation learning algorithms. Our experimental results demonstrate that the TS network is indeed more expressive and consistently learns faster than standard ReLU networks.
****
Optimistic Bandit Convex Optimization
Scott Yang, Mehryar Mohri
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.
****
Interpretable Nonlinear Dynamic Modeling of Neural Trajectories

Yuan Zhao, Il Memming Park

A central challenge in neuroscience is understanding how neural system implements computation through its dynamics. We propose a nonlinear time series model aimed at characterizing interpretable dynamics from neural trajectories. Our model assumes low-dimensional continuous dynamics in a finite volume. It incorporates a prior assumption about globally contractional dynamics to avoid overly enthusiastic extrapolation outside of the support of observed trajectories. We show that our model can recover qualitative features of the phase portrait such as attractors, slow points, and bifurcations, while also producing reliable long-term future predictions in a variety of dynamical models and in real neural data.

****

Stein Variational Gradient Descent: A General Purpose Bayesian Inference Algorithm

Qiang Liu, Dilin Wang

We propose a general purpose variational inference algorithm that forms a natural counterpart of gradient descent for optimization. Our method iteratively transports a set of particles to match the target distribution, by applying a form of functional gradient descent that minimizes the KL divergence. Empirical studies are performed on various real world models and datasets, on which our method is competitive with existing state-of-the-art methods. The derivation of our method is based on a new theoretical result that connects the derivative of KL divergence under smooth transforms with Stein’s identity and a recently proposed kernelized Stein discrepancy, which is of independent interest.

****

Learning in Games: Robustness of Fast Convergence

Dylan J. Foster, Zhiyuan Li, Thodoris Lykouris, Karthik Sridharan, Eva Tardos

We show that learning algorithms satisfying a low approximate regret property experience fast convergence to approximate optimality in a large class of repeated games. Our property, which simply requires that each learner has small regret compared to a (1+eps)-multiplicative approximation to the best action in hindsight, is ubiquitous among learning algorithms; it is satisfied even by the vanilla Hedge forecaster. Our results improve upon recent work of Syrgkanis et al. in a number of ways. We require only that players observe payoffs under other players' realized actions, as opposed to expected payoffs. We further show that convergence occurs with high probability, and show convergence under bandit feedback. Finally, we improve upon the speed of convergence by a factor of n, the number of players. Both the scope of settings and the class of algorithms for which our analysis provides fast convergence are considerably broader than in previous work. Our framework applies to dynamic population games via a low approximate regret property for shifting experts. Here we strengthen the results of Lykouris et al. in two ways: We allow players to select learning algorithms from a larger class, which includes a minor variant of the basic Hedge algorithm, and we increase the maximum churn in players for which approximate optimality is achieved. In the bandit setting we present a new algorithm which provides a "small loss"-type bound with improved dependence on the number of actions in utility settings, and is both simple and efficient. This result may be of independent interest.

****

Causal Bandits: Learning Good Interventions via Causal Inference

Finnian Lattimore, Tor Lattimore, Mark D. Reid

We study the problem of using causal models to improve the rate at which good interventions can be learned online in a stochastic environment. Our formalism combines multi-arm bandits and causal inference to model a novel type of bandit feedback that is not exploited by existing approaches. We propose a new algorithm that exploits the causal feedback and prove a bound on its simple regret that is strictly better (in all quantities) than algorithms that do not use the additional causal information.

****

Minimax Optimal Alternating Minimization for Kernel Nonparametric Tensor Learning

Taiji Suzuki, Heishiro Kanagawa, Hayato Kobayashi, Nobuyuki Shimizu, Yukihiro Ta gami

We investigate the statistical performance and computational efficiency of the alternating minimization procedure for nonparametric tensor learning. Tensor modeling has been widely used for capturing the higher order relations between multimodal data sources. In addition to a linear model, a nonlinear tensor model has been received much attention recently because of its high flexibility. We consider an alternating minimization procedure for a general nonlinear model where the true function consists of components in a reproducing kernel Hilbert space (RKHS). In this paper, we show that the alternating minimization method achieves linear convergence as an optimization algorithm and that the generalization error of the resultant estimator yields the minimax optimality. We apply our algorithm to some multitask learning problems and show that the method actually shows favorable performances.

****

Universal Correspondence Network

Christopher B. Choy, JunYoung Gwak, Silvio Savarese, Manmohan Chandraker

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.

****

Phased Exploration with Greedy Exploitation in Stochastic Combinatorial Partial Monitoring Games

Sougata Chaudhuri, 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.

****

Showing versus doing: Teaching by demonstration

Mark K. Ho, Michael Littman, James MacGlashan, Fiery Cushman, Joseph L. Austerweil

People often learn from others' demonstrations, and classic inverse reinforcement learning (IRL) algorithms have brought us closer to realizing this capacity in machines. In contrast, teaching by demonstration has been less well studied computationally. Here, we develop a novel Bayesian model for teaching by demonstration. Stark differences arise when demonstrators are intentionally teaching a task versus simply performing a task. In two experiments, we show that human participants systematically modify their teaching behavior consistent with the predictions of our model. Further, we show that even standard IRL algorithms benefit when learning from behaviors that are intentionally pedagogical. We conclude by discussing IRL algorithms that can take advantage of intentional pedagogy.

****

Learning Transferrable Representations for Unsupervised Domain Adaptation

Ozan Sener, Hyun Oh Song, Ashutosh Saxena, Silvio Savarese

Supervised learning with large scale labelled datasets and deep layered models has caused a paradigm shift in diverse areas in learning and recognition. However, this approach still suffers from generalization issues under the presence of a domain shift between the training and the test data distribution. Since unsupervised domain adaptation algorithms directly address this domain shift problem between a labelled source dataset and an unlabelled target dataset, recent papers have shown promising results by fine-tuning the networks with domain adaptation loss functions which try to align the mismatch between the training and testing data distributions. Nevertheless, these recent deep learning based domain adaptation approaches still suffer from issues such as high sensitivity to the gradient reversal hyperparameters and overfitting during the fine-tuning stage. In this paper, we propose a unified deep learning framework where the representation, cross domain transformation, and target label inference are all jointly optimized in an end-to-end fashion for unsupervised domain adaptation. Our experiments show that the proposed method significantly outperforms state-of-the-art algorith

ms in both object recognition and digit classification experiments by a large margin. We will make our learned models as well as the source code available immediately upon acceptance.
****
On Robustness of Kernel Clustering
Bowei Yan, Purnamrita Sarkar
Clustering is an important unsupervised learning problem in machine learning and statistics. Among many existing algorithms, kernel \km has drawn much research attention due to its ability to find non-linear cluster boundaries and its inherent simplicity. There are two main approaches for kernel k-means: SVD of the kernel matrix and convex relaxations. Despite the attention kernel clustering has received both from theoretical and applied quarters, not much is known about robustness of the methods. In this paper we first introduce a semidefinite programming relaxation for the kernel clustering problem, then prove that under a suitable model specification, both K-SVD and SDP approaches are consistent in the limit, albeit SDP is strongly consistent, i.e. achieves exact recovery, whereas K-SVD is weakly consistent, i.e. the fraction of misclassified nodes vanish. Also the error bounds suggest that SDP is more resilient towards outliers, which we also demonstrate with experiments.
****
Homotopy Smoothing for Non-Smooth Problems with Lower Complexity than $O(1/\epsilon)$
Yi Xu, Yan Yan, Qihang Lin, Tianbao Yang
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 Algorithms for Robust PCA via Gradient Descent
Xinyang Yi, Dohyung Park, Yudong Chen, Constantine Caramanis
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.
****
Dimensionality Reduction of Massive Sparse Datasets Using Coresets
Dan Feldman, Mikhail Volkov, Daniela Rus
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.
****
Doubly Convolutional Neural Networks
Shuangfei Zhai, Yu Cheng, Zhongfei (Mark) Zhang, Weining Lu
Building large models with parameter sharing accounts for most of the success of deep convolutional neural networks (CNNs). In this paper, we propose doubly convolutional neural networks (DCNNs), which significantly improve the performance of CNNs by further exploring this idea. In stead of allocating a set of convolutional filters that are independently learned, a DCNN maintains groups of filters where filters within each group are translated versions of each other. Practically, a DCNN can be easily implemented by a two-step convolution procedure, which is supported by most modern deep learning libraries. We perform extensive experiments on three image classification benchmarks: CIFAR-10, CIFAR-100 and ImageNet, and show that DCNNs consistently outperform other competing architectures. We have also verified that replacing a convolutional layer with a doubly convolutional layer at any depth of a CNN can improve its performance. Moreover, various design choices of DCNNs are demonstrated, which shows that DCNN can serve the dual purpose of building more accurate models and/or reducing the memory footprint without sacrificing the accuracy.
****

Brains on Beats
Umut Güçlü, Jordy Thielen, Michael Hanke, Marcel van Gerven

We developed task-optimized deep neural networks (DNNs) that achieved state-of-the-art performance in different evaluation scenarios for automatic music tagging. These DNNs were subsequently used to probe the neural representations of music. Representational similarity analysis revealed the existence of a representational gradient across the superior temporal gyrus (STG). Anterior STG was shown to be more sensitive to low-level stimulus features encoded in shallow DNN layers whereas posterior STG was shown to be more sensitive to high-level stimulus features encoded in deep DNN layers.

****

Local Minimax Complexity of Stochastic Convex Optimization
sabyasachi chatterjee, John C. Duchi, John Lafferty, Yuancheng Zhu

We extend the traditional worst-case, minimax analysis of stochastic convex optimization by introducing a localized form of minimax complexity for individual functions. Our main result gives function-specific lower and upper bounds on the number of stochastic subgradient evaluations needed to optimize either the function or its ``hardest local alternative'' to a given numerical precision. The bounds are expressed in terms of a localized and computational analogue of the modulus of continuity that is central to statistical minimax analysis. We show how the computational modulus of continuity can be explicitly calculated in concrete cases, and relates to the curvature of the function at the optimum. We also prove a superefficiency result that demonstrates it is a meaningful benchmark, acting as a computational analogue of the Fisher information in statistical estimation. The nature and practical implications of the results are demonstrated in simulations.

****

Kronecker Determinantal Point Processes
Zelda E. Mariet, Suvrit Sra

Determinantal Point Processes (DPPs) are probabilistic models over all subsets a ground set of N items. They have recently gained prominence in several applications that rely on diverse subsets. However, their applicability to large problems is still limited due to O(N^3) complexity of core tasks such as sampling and learning. We enable efficient sampling and learning for DPPs by introducing KronDPP, a DPP model whose kernel matrix decomposes as a tensor product of multiple smaller kernel matrices. This decomposition immediately enables fast exact sampling. But contrary to what one may expect, leveraging the Kronecker product structure for speeding up DPP learning turns out to be more difficult. We overcome this challenge, and derive batch and stochastic optimization algorithms for efficiently learning the parameters of a KronDPP.

****

Normalized Spectral Map Synchronization
Yanyao Shen, Qixing Huang, Nati Srebro, Sujay Sanghavi

The algorithmic advancement of synchronizing maps is important in order to solve a wide range of practice problems with possible large-scale dataset. In this paper, we provide theoretical justifications for spectral techniques for the map synchronization problem, i.e., it takes as input a collection of objects and noisy maps estimated between pairs of objects, and outputs clean maps between all pairs of objects. We show that a simple normalized spectral method that projects the blocks of the top eigenvectors of a data matrix to the map space leads to surprisingly good results. As the noise is modelled naturally as random permutation matrix, this algorithm NormSpecSync leads to competing theoretical guarantees as state-of-the-art convex optimization techniques, yet it is much more efficient. We demonstrate the usefulness of our algorithm in a couple of applications, where it is optimal in both complexity and exactness among existing methods.

****

Error Analysis of Generalized Nyström Kernel Regression
Hong Chen, Haifeng Xia, Heng Huang, Weidong Cai

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.

****

Regularized Nonlinear Acceleration

Damien Scieur, Alexandre d'Aspremont, Francis Bach

We describe a convergence acceleration technique for generic optimization problems. Our scheme computes estimates of the optimum from a nonlinear average of the iterates produced by any optimization method. The weights in this average are computed via a simple and small linear system, whose solution can be updated online. This acceleration scheme runs in parallel to the base algorithm, providing improved estimates of the solution on the fly, while the original optimization method is running. Numerical experiments are detailed on classical classification problems.

****

Pairwise Choice Markov Chains

Stephen Ragain, Johan Ugander

As datasets capturing human choices grow in richness and scale, particularly in online domains, there is an increasing need for choice models flexible enough to handle data that violate traditional choice-theoretic axioms such as regularity, stochastic transitivity, or Luce's choice axiom. In this work we introduce the Pairwise Choice Markov Chain (PCMC) model of discrete choice, an inferentially tractable model that does not assume these traditional axioms while still satisfying the foundational axiom of uniform expansion, which can be viewed as a weaker version of Luce's axiom. We show that the PCMC model significantly outperforms the Multinomial Logit (MNL) model in prediction tasks on two empirical data sets known to exhibit violations of Luce's axiom. Our analysis also synthesizes several recent observations connecting the Multinomial Logit model and Markov chains; the PCMC model retains the Multinomial Logit model as a special case.

****

Stochastic Variational Deep Kernel Learning

Andrew G. Wilson, Zhiting Hu, Russ R. Salakhutdinov, Eric P. Xing

Deep kernel learning combines the non-parametric flexibility of kernel methods with the inductive biases of deep learning architectures. We propose a novel deep kernel learning model and stochastic variational inference procedure which generalizes deep kernel learning approaches to enable classification, multi-task learning, additive covariance structures, and stochastic gradient training. Specifically, we apply additive base kernels to subsets of output features from deep neural architectures, and jointly learn the parameters of the base kernels and deep network through a Gaussian process marginal likelihood objective. Within this framework, we derive an efficient form of stochastic variational inference which leverages local kernel interpolation, inducing points, and structure exploiting algebra. We show improved performance over stand alone deep networks, SVMs, and state of the art scalable Gaussian processes on several classification benchmarks, including an airline delay dataset containing 6 million training points, CIFAR, and ImageNet.

****

Theoretical Comparisons of Positive-Unlabeled Learning against Positive-Negative Learning

Gang Niu, Marthinus Christoffel du Plessis, Tomoya Sakai, Yao Ma, Masashi Sugiyama

In PU learning, a binary classifier is trained from positive (P) and unlabeled (U) data without negative (N) data. Although N data is missing, it sometimes outperforms PN learning (i.e., ordinary supervised learning). Hitherto, neither theoretical nor experimental analysis has been given to explain this phenomenon. In this paper, we theoretically compare PU (and NU) learning against PN learning based on the upper bounds on estimation errors. We find simple conditions when PU and NU learning are likely to outperform PN learning, and we prove that, in terms of the upper bounds, either PU or NU learning (depending on the class-prior probability and the sizes of P and N data) given infinite U data will improve on PN learning. Our theoretical findings well agree with the experimental results on artificial and benchmark data even when the experimental setup does not match the theoretical assumptions exactly.

****

A Non-generative Framework and Convex Relaxations for Unsupervised Learning
Elad Hazan, Tengyu Ma

We give a novel formal theoretical framework for unsupervised learning with two distinctive characteristics. First, it does not assume any generative model and based on a worst-case performance metric. Second, it is comparative, namely performance is measured with respect to a given hypothesis class. This allows to avoid known computational hardness results and improper algorithms based on convex relaxations. We show how several families of unsupervised learning models, which were previously only analyzed under probabilistic assumptions and are otherwise provably intractable, can be efficiently learned in our framework by convex optimization.

****

DISCO Nets : DISsimilarity COefficients Networks
Diane Bouchacourt, Pawan K. Mudigonda, Sebastian Nowozin

We present a new type of probabilistic model which we call DISsimilarity COefficient Networks (DISCO Nets). DISCO Nets allow us to efficiently sample from a posterior distribution parametrised by a neural network. During training, DISCO Nets are learned by minimising the dissimilarity coefficient between the true distribution and the estimated distribution. This allows us to tailor the training to the loss related to the task at hand. We empirically show that (i) by modeling uncertainty on the output value, DISCO Nets outperform equivalent non-probabilistic predictive networks and (ii) DISCO Nets accurately model the uncertainty of the output, outperforming existing probabilistic models based on deep neural networks.

****

Learning to Poke by Poking: Experiential Learning of Intuitive Physics
Pulkit Agrawal, Ashvin V. Nair, Pieter Abbeel, Jitendra Malik, Sergey Levine

We investigate an experiential learning paradigm for acquiring an internal model of intuitive physics. Our model is evaluated on a real-world robotic manipulation task that requires displacing objects to target locations by poking. The robot gathered over 400 hours of experience by executing more than 50K pokes on different objects. We propose a novel approach based on deep neural networks for modeling the dynamics of robot's interactions directly from images, by jointly estimating forward and inverse models of dynamics. The inverse model objective provides supervision to construct informative visual features, which the forward model can then predict and in turn regularize the feature space for the inverse model. The interplay between these two objectives creates useful, accurate models that can then be used for multi-step decision making. This formulation has the additional benefit that it is possible to learn forward models in an abstract feature space and thus alleviate the need of predicting pixels. Our experiments show that this joint modeling approach outperforms alternative methods. We also demonstrate that active data collection using the learned model further improves performance.

****

Value Iteration Networks
Aviv Tamar, YI WU, Garrett Thomas, Sergey Levine, Pieter Abbeel

We introduce the value iteration network (VIN): a fully differentiable neural network with a `planning module' embedded within. VINs can learn to plan, and are suitable for predicting outcomes that involve planning-based reasoning, such as policies for reinforcement learning. Key to our approach is a novel differentiable approximation of the value-iteration algorithm, which can be represented as a convolutional neural network, and trained end-to-end using standard backpropagation. We evaluate VIN based policies on discrete and continuous path-planning domains, and on a natural-language based search task. We show that by learning an explicit planning computation, VIN policies generalize better to new, unseen domains.

****

Training and Evaluating Multimodal Word Embeddings with Large-scale Web Annotated Images
Junhua Mao, Jiajing Xu, Kevin Jing, Alan L. Yuille
In this paper, we focus on training and evaluating effective word embeddings with both text and visual information. More specifically, we introduce a large-scale dataset with 300 million sentences describing over 40 million images crawled and downloaded from publicly available Pins (i.e. an image with sentence descriptions uploaded by users) on Pinterest. This dataset is more than 200 times larger than MS COCO, the standard large-scale image dataset with sentence descriptions. In addition, we construct an evaluation dataset to directly assess the effectiveness of word embeddings in terms of finding semantically similar or related words and phrases. The word/phrase pairs in this evaluation dataset are collected from the click data with millions of users in an image search system, thus contain rich semantic relationships. Based on these datasets, we propose and compare several Recurrent Neural Networks (RNNs) based multimodal (text and image) models. Experiments show that our model benefits from incorporating the visual information into the word embeddings, and a weight sharing strategy is crucial for learning such multimodal embeddings. The project page is: http://www.stat.ucla.edu/~junhua.mao/multimodal_embedding.html (The datasets introduced in this work will be gradually released on the project page.).
****
Simple and Efficient Weighted Minwise Hashing
Anshumali Shrivastava
Weighted minwise hashing (WMH) is one of the fundamental subroutine, required by many celebrated approximation algorithms, commonly adopted in industrial practice for large -scale search and learning. The resource bottleneck with WMH is the computation of multiple (typically a few hundreds to thousands) independent hashes of the data. We propose a simple rejection type sampling scheme based on a carefully designed red-green map, where we show that the number of rejected sample has exactly the same distribution as weighted minwise sampling. The running time of our method, for many practical datasets, is an order of magnitude smaller than existing methods. Experimental evaluations, on real datasets, show that for computing 500 WMH, our proposal can be 60000x faster than the Ioffe's method without losing any accuracy. Our method is also around 100x faster than approximate heuristics capitalizing on the efficient ``densified" one permutation hashing schemes~\cite{Proc:OneHashLSHICML14,Proc:ShrivastavaUAI14}. Given the simplicity of our approach and its significant advantages, we hope that it will replace existing implementations in practice.
****
Cooperative Inverse Reinforcement Learning
Dylan Hadfield-Menell, Stuart J. Russell, Pieter Abbeel, Anca Dragan
For an autonomous system to be helpful to humans and to pose no unwarranted risks, it needs to align its values with those of the humans in its environment in such a way that its actions contribute to the maximization of value for the humans. We propose a formal definition of the value alignment problem as cooperative inverse reinforcement learning (CIRL). A CIRL problem is a cooperative, partial- information game with two agents, human and robot; both are rewarded according to the human’s reward function, but the robot does not initially know what this is. In contrast to classical IRL, where the human is assumed to act optimally in isolation, optimal CIRL solutions produce behaviors such as active teaching, active learning, and communicative actions that are more effective in achieving value alignment. We show that computing optimal joint policies in CIRL games can be reduced to solving a POMDP, prove that optimality in isolation is suboptimal in CIRL, and derive an approximate CIRL algorithm.
****
Safe and Efficient Off-Policy Reinforcement Learning
Remi Munos, Tom Stepleton, Anna Harutyunyan, Marc Bellemare
In this work, we take a fresh look at some old and new algorithms for off-policy, return-based reinforcement learning. Expressing these in a common form, we derive a novel algorithm, Retrace(lambda), with three desired properties: (1) it ha

s low variance; (2) it safely uses samples collected from any behaviour policy, whatever its degree of "off-policyness"; and (3) it is efficient as it makes the best use of samples collected from near on-policy behaviour policies. We analyse the contractive nature of the related operator under both off-policy policy evaluation and control settings and derive online sample-based algorithms. We believe this is the first return-based off-policy control algorithm converging a.s. to Q* without the GLIE assumption (Greedy in the Limit with Infinite Exploration). As a corollary, we prove the convergence of Watkins' Q(lambda), which was an open problem since 1989. We illustrate the benefits of Retrace(lambda) on a standard suite of Atari 2600 games.

****

LightRNN: Memory and Computation-Efficient Recurrent Neural Networks

Xiang Li, Tao Qin, Jian Yang, Tie-Yan 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.

****

Deep Learning Games

Dale Schuurmans, Martin A. Zinkevich

We investigate a reduction of supervised learning to game playing that reveals new connections and learning methods. For convex one-layer problems, we demonstrate an equivalence between global minimizers of the training problem and Nash equilibria in a simple game. We then show how the game can be extended to general acyclic neural networks with differentiable convex gates, establishing a bijection between the Nash equilibria and critical (or KKT) points of the deep learning problem. Based on these connections we investigate alternative learning methods, and find that regret matching can achieve competitive training performance while producing sparser models than current deep learning approaches.

****

Strategic Attentive Writer for Learning Macro-Actions

Alexander Vezhnevets, Volodymyr Mnih, Simon Osindero, Alex Graves, Oriol Vinyals, John Agapiou, koray kavukcuoglu

We present a novel deep recurrent neural network architecture that learns to build implicit plans in an end-to-end manner purely by interacting with an environment in reinforcement learning setting. The network builds an internal plan, which is continuously updated upon observation of the next input from the environment. It can also partition this internal representation into contiguous sub-sequences by learning for how long the plan can be committed to -- i.e. followed without replaning. Combining these properties, the proposed model, dubbed STRategic Attentive Writer (STRAW) can learn high-level, temporally abstracted macro-actions of varying lengths that are solely learnt from data without any prior information. These macro-actions enable both structured exploration and economic computation. We experimentally demonstrate that STRAW delivers strong improvements on several ATARI games by employing temporally extended planning strategies (e.g. Ms. Pacman and Frostbite). It is at the same time a general algorithm that can be applied on any sequence data. To that end, we also show that when trained on text prediction task, STRAW naturally predicts frequent n-grams (instead of macro-actions), demonstrating the generality of the approach.

****

Clustering with Bregman Divergences: an Asymptotic Analysis

Chaoyue Liu, Mikhail Belkin

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.

****

Swapout: Learning an ensemble of deep architectures

Saurabh Singh, Derek Hoiem, David Forsyth

We describe Swapout, a new stochastic training method, that outperforms ResNets of identical network structure yielding impressive results on CIFAR-10 and CIFAR-100. Swapout samples from a rich set of architectures including dropout, stochastic depth and residual architectures as special cases. When viewed as a regularization method swapout not only inhibits co-adaptation of units in a layer, similar to dropout, but also across network layers. We conjecture that swapout achieves strong regularization by implicitly tying the parameters across layers. When viewed as an ensemble training method, it samples a much richer set of architectures than existing methods such as dropout or stochastic depth. We propose a parameterization that reveals connections to exiting architectures and suggests a much richer set of architectures to be explored. We show that our formulation suggests an efficient training method and validate our conclusions on CIFAR-10 and CIFAR-100 matching state of the art accuracy. Remarkably, our 32 layer wider model performs similar to a 1001 layer ResNet model.

****

Stochastic Online AUC Maximization

Yiming Ying, Longyin Wen, Siwei Lyu

Area under ROC (AUC) is a metric which is widely used for measuring the classification performance for imbalanced data. It is of theoretical and practical interest to develop online learning algorithms that maximizes AUC for large-scale data. A specific challenge in developing online AUC maximization algorithm is that the learning objective function is usually defined over a pair of training examples of opposite classes, and existing methods achieves on-line processing with higher space and time complexity. In this work, we propose a new stochastic online algorithm for AUC maximization. In particular, we show that AUC optimization can be equivalently formulated as a convex-concave saddle point problem. From this saddle representation, a stochastic online algorithm (SOLAM) is proposed which has time and space complexity of one datum. We establish theoretical convergence of SOLAM with high probability and demonstrate its effectiveness and efficiency on standard benchmark datasets.

****

Optimizing affinity-based binary hashing using auxiliary coordinates

Ramin Raziperchikolaei, Miguel A. Carreira-Perpinan

In supervised binary hashing, one wants to learn a function that maps a high-dimensional feature vector to a vector of binary codes, for application to fast image retrieval. This typically results in a difficult optimization problem, nonconvex and nonsmooth, because of the discrete variables involved. Much work has simply relaxed the problem during training, solving a continuous optimization, and truncating the codes a posteriori. This gives reasonable results but is quite suboptimal. Recent work has tried to optimize the objective directly over the binary codes and achieved better results, but the hash function was still learned a posteriori, which remains suboptimal. We propose a general framework for learning hash functions using affinity-based loss functions that uses auxiliary coordinates. This closes the loop and optimizes jointly over the hash functions and the binary codes so that they gradually match each other. The resulting algorithm can be seen as an iterated version of the procedure of optimizing first over the codes and then learning the hash function. Compared to this, our optimization is guaranteed to obtain better hash functions while being not much slower, as demonstrated experimentally in various supervised datasets. In addition, our framework facilitates the design of optimization algorithms for arbitrary types of loss and hash functions.

****

Sample Complexity of Automated Mechanism Design

Maria-Florina F. Balcan, Tuomas Sandholm, Ellen Vitercik

The design of revenue-maximizing combinatorial auctions, i.e. multi item auctions over bundles of goods, is one of the most fundamental problems in computational economics, unsolved even for two bidders and two items for sale. In the traditional economic models, it is assumed that the bidders' valuations are drawn from an underlying distribution and that the auction designer has perfect knowledge of this distribution. Despite this strong and oftentimes unrealistic assumption, it is remarkable that the revenue-maximizing combinatorial auction remains unkn own. In recent years, automated mechanism design has emerged as one of the most practical and promising approaches to designing high-revenue combinatorial auctions. The most scalable automated mechanism design algorithms take as input samples from the bidders' valuation distribution and then search for a high-revenue auction in a rich auction class. In this work, we provide the first sample complexity analysis for the standard hierarchy of deterministic combinatorial auction classes used in automated mechanism design. In particular, we provide tight sample complexity bounds on the number of samples needed to guarantee that the empirical revenue of the designed mechanism on the samples is close to its expected revenue on the underlying, unknown distribution over bidder valuations, for each of the auction classes in the hierarchy. In addition to helping set automated mechanism design on firm foundations, our results also push the boundaries of learning theory. In particular, the hypothesis functions used in our contexts are defined through multi stage combinatorial optimization procedures, rather than simple decision boundaries, as are common in machine learning.

****

LazySVD: Even Faster SVD Decomposition Yet Without Agonizing Pain

Zeyuan Allen-Zhu, Yuanzhi 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.

****

A Probabilistic Framework for Deep Learning

Ankit B. Patel, Minh Tan Nguyen, Richard Baraniuk

We develop a probabilistic framework for deep learning based on the Deep Rendering Mixture Model (DRMM), a new generative probabilistic model that explicitly capture variations in data due to latent task nuisance variables. We demonstrate that max-sum inference in the DRMM yields an algorithm that exactly reproduces the operations in deep convolutional neural networks (DCNs), providing a first principles derivation. Our framework provides new insights into the successes and shortcomings of DCNs as well as a principled route to their improvement. DRMM training via the Expectation-Maximization (EM) algorithm is a powerful alternative to DCN back-propagation, and initial training results are promising. Classification based on the DRMM and other variants outperforms DCNs in supervised digit classification, training 2-3x faster while achieving similar accuracy. Moreover, the DRMM is applicable to semi-supervised and unsupervised learning tasks, achieving results that are state-of-the-art in several categories on the MNIST benchmark and comparable to state of the art on the CIFAR10 benchmark.

****

Without-Replacement Sampling for Stochastic Gradient Methods

Ohad Shamir

Stochastic gradient methods for machine learning and optimization problems are usually analyzed assuming data points are sampled with replacement. In contrast, sampling without replacement is far less understood, yet in practice it is very common, often easier to implement, and usually performs better. In this paper, we provide competitive convergence guarantees for without-replacement sampling under several scenarios, focusing on the natural regime of few passes over the data. Moreover, we describe a useful application of these results in the context of distributed optimization with randomly-partitioned data, yielding a nearly-optimal algorithm for regularized least squares (in terms of both communication complexity and runtime complexity) under broad parameter regimes. Our proof techniques combine ideas from stochastic optimization, adversarial online learning and transductive learning theory, and can potentially be applied to other stochastic optimization and learning problems.

****

Learning to Communicate with Deep Multi-Agent Reinforcement Learning

Jakob Foerster, Ioannis Alexandros Assael, Nando de Freitas, Shimon Whiteson

We consider the problem of multiple agents sensing and acting in environments with the goal of maximising their shared utility. In these environments, agents mu st learn communication protocols in order to share information that is needed to solve the tasks. By embracing deep neural networks, we are able to demonstrate end-to-end learning of protocols in complex environments inspired by communication riddles and multi-agent computer vision problems with partial observability. We propose two approaches for learning in these domains: Reinforced Inter-Agent Learning (RIAL) and Differentiable Inter-Agent Learning (DIAL). The former uses deep Q-learning, while the latter exploits the fact that, during learning, agents can backpropagate error derivatives through (noisy) communication channels. Hence, this approach uses centralised learning but decentralised execution. Our experiments introduce new environments for studying the learning of communication protocols and present a set of engineering innovations that are essential for success in these domains.

****

Understanding the Effective Receptive Field in Deep Convolutional Neural Networks

Wenjie Luo, Yujia Li, Raquel Urtasun, Richard Zemel

We study characteristics of receptive fields of units in deep convolutional networks. The receptive field size is a crucial issue in many visual tasks, as the output must respond to large enough areas in the image to capture information about large objects. We introduce the notion of an effective receptive field size, and show that it both has a Gaussian distribution and   only occupies a fraction of the full theoretical receptive field size. We analyze the effective receptive field in several architecture designs, and the effect of sub-sampling, skip connections, dropout and nonlinear activations on it. This leads to suggestions for ways to address its tendency to be too small.

****

Barzilai-Borwein Step Size for Stochastic Gradient Descent

Conghui Tan, Shiqian Ma, Yu-Hong Dai, Yuqiu Qian

One of the major issues in stochastic gradient descent (SGD) methods is how to choose an appropriate step size while running the algorithm. Since the traditional line search technique does not apply for stochastic optimization methods, the common practice in SGD is either to use a diminishing step size, or to tune a step size by hand, which can be time consuming in practice. In this paper, we propose to use the Barzilai-Borwein (BB) method to automatically compute step sizes for SGD and its variant: stochastic variance reduced gradient (SVRG) method, which leads to two algorithms: SGD-BB and SVRG-BB. We prove that SVRG-BB converges linearly for strongly convex objective functions. As a by-product, we prove the linear convergence result of SVRG with Option I proposed in [10], whose convergence result has been missing in the literature. Numerical experiments on standard data sets show that the performance of SGD-BB and SVRG-BB is comparable to and sometimes even better than SGD and SVRG with best-tuned step sizes, and is superior to some advanced SGD variants.

****

The Power of Optimization from Samples

Eric Balkanski, Aviad Rubinstein, Yaron Singer

We consider the problem of optimization from samples of monotone submodular functions with bounded curvature. In numerous applications, the function optimized is not known a priori, but instead learned from data. What are the guarantees we have when optimizing functions from sampled data? In this paper we show that for any monotone submodular function with curvature c there is a (1 - c)/(1 + c - c^2) approximation algorithm for maximization under cardinality constraints when polynomially-many samples are drawn from the uniform distribution over feasible sets. Moreover, we show that this algorithm is optimal. That is, for any c < 1, there exists a submodular function with curvature c for which no algorithm can achieve a better approximation. The curvature assumption is crucial as for general monotone submodular functions no algorithm can obtain a constant-factor approximation for maximization under a cardinality constraint when observing polynomially-many samples drawn from any distribution over feasible sets, even when the function is statistically learnable.

****

New Liftable Classes for First-Order Probabilistic Inference
Seyed Mehran Kazemi, Angelika Kimmig, Guy Van den Broeck, David Poole

Statistical relational models provide compact encodings of probabilistic dependencies in relational domains, but result in highly intractable graphical models. The goal of lifted inference is to carry out probabilistic inference without needing to reason about each individual separately, by instead treating exchangeable, undistinguished objects as a whole. In this paper, we study the domain recursion inference rule, which, despite its central role in early theoretical results on domain-lifted inference, has later been believed redundant. We show that this rule is more powerful than expected, and in fact significantly extends the range of models for which lifted inference runs in time polynomial in the number of individuals in the domain. This includes an open problem called S4, the symmetric transitivity model, and a first-order logic encoding of the birthday paradox. We further identify new classes S2FO2 and S2RU of domain-liftable theories, which respectively subsume FO2 and recursively unary theories, the largest classes of domain-liftable theories known so far, and show that using domain recursion can achieve exponential speedup even in theories that cannot fully be lifted with the existing set of inference rules.

****

Optimal Tagging with Markov Chain Optimization
Nir Rosenfeld, Amir Globerson

Many information systems use tags and keywords to describe and annotate content. These allow for efficient organization and categorization of items, as well as facilitate relevant search queries. As such, the selected set of tags for an item can have a considerable effect on the volume of traffic that eventually reaches an item. In tagging systems where tags are exclusively chosen by an item's owner, who in turn is interested in maximizing traffic, a principled approach for assigning tags can prove valuable. In this paper we introduce the problem of optimal tagging, where the task is to choose a subset of tags for a new item such that the probability of browsing users reaching that item is maximized. We formulate the problem by modeling traffic using a Markov chain, and asking how transitions in this chain should be modified to maximize traffic into a certain state of interest. The resulting optimization problem involves maximizing a certain function over subsets, under a cardinality constraint. We show that the optimization problem is NP-hard, but has a (1-1/e)-approximation via a simple greedy algorithm due to monotonicity and submodularity. Furthermore, the structure of the problem allows for an efficient computation of the greedy step. To demonstrate the effectiveness of our method, we perform experiments on three tagging datasets, and show that the greedy algorithm outperforms other baselines.

****

Fast and Flexible Monotonic Functions with Ensembles of Lattices
Mahdi Milani Fard, Kevin Canini, Andrew Cotter, Jan Pfeifer, Maya Gupta

For many machine learning problems, there are some inputs that are known to be positively (or negatively) related to the output, and in such cases training the model to respect that monotonic relationship can provide regularization, and makes the model more interpretable. However, flexible monotonic functions are computationally challenging to learn beyond a few features. We break through this barrier by learning ensembles of monotonic calibrated interpolated look-up tables (lattices). A key contribution is an automated algorithm for selecting feature subsets for the ensemble base models. We demonstrate that compared to random forests, these ensembles produce similar or better accuracy, while providing guaranteed monotonicity consistent with prior knowledge, smaller model size and faster evaluation.

****

A scaled Bregman theorem with applications
Richard Nock, Aditya Menon, Cheng Soon Ong

Bregman divergences play a central role in the design and analysis of a range of machine learning algorithms through a handful of popular theorems. We present a new theorem which shows that ``Bregman distortions'' (employing a potentially non-convex generator) may be exactly re-written as a scaled Bregman divergence co

mputed over transformed data. This property can be viewed from the standpoints of geometry (a scaled isometry with adaptive metrics) or convex optimization (relating generalized perspective transforms). Admissible distortions include {geodesic distances} on curved manifolds and projections or gauge-normalisation. Our theorem allows one to leverage to the wealth and convenience of Bregman divergences when analysing algorithms relying on the aforementioned Bregman distortions. We illustrate this with three novel applications of our theorem: a reduction from multi-class density ratio to class-probability estimation, a new adaptive projection free yet norm-enforcing dual norm mirror descent algorithm, and a reduction from clustering on flat manifolds to clustering on curved manifolds. Experiments on each of these domains validate the analyses and suggest that the scaled Bregman theorem might be a worthy addition to the popular handful of Bregman divergence properties that have been pervasive in machine learning.
****

The Product Cut
Thomas Laurent, James von Brecht, Xavier Bresson, arthur szlam
We introduce a theoretical and algorithmic framework for multi-way graph partitioning that relies on a multiplicative cut-based objective. We refer to this objective as the Product Cut. We provide a detailed investigation of the mathematical properties of this objective and an effective algorithm for its optimization. The proposed model has strong mathematical underpinnings, and the corresponding algorithm achieves state-of-the-art performance on benchmark data sets.
****

Learning from Rational Behavior: Predicting Solutions to Unknown Linear Programs
Shahin Jabbari, Ryan M. Rogers, Aaron Roth, Steven Z. Wu
We define and study the problem of predicting the solution to a linear program (LP) given only partial information about its objective and constraints. This generalizes the problem of learning to predict the purchasing behavior of a rational agent who has an unknown objective function, that has been studied under the name “Learning from Revealed Preferences". We give mistake bound learning algorithms in two settings: in the first, the objective of the LP is known to the learner but there is an arbitrary, fixed set of constraints which are unknown. Each example is defined by an additional known constraint and the goal of the learner is to predict the optimal solution of the LP given the union of the known and unknown constraints. This models the problem of predicting the behavior of a rational agent whose goals are known, but whose resources are unknown. In the second setting, the objective of the LP is unknown, and changing in a controlled way. The constraints of the LP may also change every day, but are known. An example is given by a set of constraints and partial information about the objective, and the task of the learner is again to predict the optimal solution of the partially known LP.
****

Unified Methods for Exploiting Piecewise Linear Structure in Convex Optimization
Tyler B. Johnson, Carlos Guestrin
We develop methods for rapidly identifying important components of a convex optimization problem for the purpose of achieving fast convergence times. By considering a novel problem formulation—the minimization of a sum of piecewise functions—we describe a principled and general mechanism for exploiting piecewise linear structure in convex optimization. This result leads to a theoretically justified working set algorithm and a novel screening test, which generalize and improve upon many prior results on exploiting structure in convex optimization. In empirical comparisons, we study the scalability of our methods. We find that screening scales surprisingly poorly with the size of the problem, while our working set algorithm convincingly outperforms alternative approaches.
****

Large-Scale Price Optimization via Network Flow
Shinji Ito, Ryohei Fujimaki
This paper deals with price optimization, which is to find the best pricing strategy that maximizes revenue or profit, on the basis of demand forecasting models. Though recent advances in regression technologies have made it possible to rev

eal price-demand relationship of a number of multiple products, most existing price optimization methods, such as mixed integer programming formulation, cannot handle tens or hundreds of products because of their high computational costs. To cope with this problem, this paper proposes a novel approach based on network flow algorithms. We reveal a connection between supermodularity of the revenue and cross elasticity of demand. On the basis of this connection, we propose an efficient algorithm that employs network flow algorithms. The proposed algorithm can handle hundreds or thousands of products, and returns an exact optimal solution under an assumption regarding cross elasticity of demand. Even in case in which the assumption does not hold, the proposed algorithm can efficiently find approximate solutions as good as can other state-of-the-art methods, as empirical results show.

****

Generative Adversarial Imitation Learning

Jonathan Ho, Stefano Ermon

Consider learning a policy from example expert behavior, without interaction with the expert or access to a reinforcement signal. One approach is to recover the expert's cost function with inverse reinforcement learning, then extract a policy from that cost function with reinforcement learning. This approach is indirect and can be slow. We propose a new general framework for directly extracting a policy from data as if it were obtained by reinforcement learning following inverse reinforcement learning. We show that a certain instantiation of our framework draws an analogy between imitation learning and generative adversarial networks, from which we derive a model-free imitation learning algorithm that obtains significant performance gains over existing model-free methods in imitating complex behaviors in large, high-dimensional environments.

****

Truncated Variance Reduction: A Unified Approach to Bayesian Optimization and Level-Set Estimation

Ilija Bogunovic, Jonathan Scarlett, Andreas Krause, Volkan Cevher

We present a new algorithm, truncated variance reduction (TruVaR), that treats Bayesian optimization (BO) and level-set estimation (LSE) with Gaussian processes in a unified fashion. The algorithm greedily shrinks a sum of truncated variances within a set of potential maximizers (BO) or unclassified points (LSE), which is updated based on confidence bounds. TruVaR is effective in several important settings that are typically non-trivial to incorporate into myopic algorithms, including pointwise costs and heteroscedastic noise. We provide a general theoretical guarantee for TruVaR covering these aspects, and use it to recover and strengthen existing results on BO and LSE. Moreover, we provide a new result for a setting where one can select from a number of noise levels having associated costs. We demonstrate the effectiveness of the algorithm on both synthetic and real-world data sets.

****

f-GAN: Training Generative Neural Samplers using Variational Divergence Minimization

Sebastian Nowozin, Botond Cseke, 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.

****

Nearly Isometric Embedding by Relaxation

James McQueen, Marina Meila, Dominique Joncas

Many manifold learning algorithms aim to create embeddings with low or no distortion (i.e. isometric). If the data has intrinsic dimension d, it is often impossible to obtain an isometric embedding in d dimensions, but possible in s > d dimensions. Yet, most geometry preserving algorithms cannot do the latter. This paper proposes an embedding algorithm that overcomes this problem. The algorithm directly computes, for any data embedding Y, a distortion loss(Y), and iteratively updates Y in order to decrease it. The distortion measure we propose is based o n the push-forward Riemannian metric associated with the coordinates Y. The experiments confirm the superiority of our algorithm in obtaining low distortion embeddings.

****

DECOrrelated feature space partitioning for distributed sparse regression

Xiangyu Wang, David B. Dunson, Chenlei Leng

Fitting statistical models is computationally challenging when the sample size or the dimension of the dataset is huge. An attractive approach for down-scaling the problem size is to first partition the dataset into subsets and then fit using distributed algorithms. The dataset can be partitioned either horizontally (in the sample space) or vertically (in the feature space). While the majority of the literature focuses on sample space partitioning, feature space partitioning is more effective when p >> n. Existing methods for partitioning features, however, are either vulnerable to high correlations or inefficient in reducing the model dimension. In this paper, we solve these problems through a new embarrassingly parallel framework named DECO for distributed variable selection and parameter estimation. In DECO, variables are first partitioned and allocated to m distributed workers. The decorrelated subset data within each worker are then fitted via any algorithm designed for high-dimensional problems. We show that by incorporating the decorrelation step, DECO can achieve consistent variable selection and parameter estimation on each subset with (almost) no assumptions. In addition, the convergence rate is nearly minimax optimal for both sparse and weakly sparse models and does NOT depend on the partition number m. Extensive numerical experiments are provided to illustrate the performance of the new framework.

****

Incremental Boosting Convolutional Neural Network for Facial Action Unit Recognition

Shizhong Han, Zibo Meng, AHMED-SHEHAB KHAN, Yan Tong

Recognizing facial action units (AUs) from spontaneous facial expressions is still a challenging problem. Most recently, CNNs have shown promise on facial AU recognition. However, the learned CNNs are often overfitted and do not generalize well to unseen subjects due to limited AU-coded training images. We proposed a novel Incremental Boosting CNN (IB-CNN) to integrate boosting into the CNN via an incremental boosting layer that selects discriminative neurons from the lower layer and is incrementally updated on successive mini-batches. In addition, a novel loss function that accounts for errors from both the incremental boosted classifier and individual weak classifiers was proposed to fine-tune the IB-CNN. Experimental results on four benchmark AU databases have demonstrated that the IB-CNN yields significant improvement over the traditional CNN and the boosting CNN without incremental learning, as well as outperforming the state-of-the-art CNN-based methods in AU recognition. The improvement is more impressive for the AUs that have the lowest frequencies in the databases.

****

An urn model for majority voting in classification ensembles

Victor Soto, Alberto Suárez, Gonzalo Martinez-Muñoz

In this work we analyze the class prediction of parallel randomized ensembles by majority voting as an urn model. For a given test instance, the ensemble can be viewed as an urn of marbles of different colors. A marble represents an individual classifier. Its color represents the class label prediction of the corresponding classifier. The sequential querying of classifiers in the ensemble can be seen as draws without replacement from the urn. An analysis of this classical urn model based on the hypergeometric distribution makes it possible to estimate the confidence on the outcome of majority voting when only a fraction of the individual predictions is known. These estimates can be used to speed up the prediction by the ensemble. Specifically, the aggregation of votes can be halted when the confidence in the final prediction is sufficiently high. If one assumes a uniform prior for the distribution of possible votes the analysis is shown to be equivalent to a previous one based on Dirichlet distributions. The advantage of the current approach is that prior knowledge on the possible vote outcomes can be readily incorporated in a Bayesian framework. We show how incorporating this type of problem-specific knowledge into the statistical analysis of majority voting leads to faster classification by the ensemble and allows us to estimate the expected average speed-up beforehand.

****

Unsupervised Feature Extraction by Time-Contrastive Learning and Nonlinear ICA

Aapo Hyvarinen, Hiroshi Morioka

Nonlinear independent component analysis (ICA) provides an appealing framework for unsupervised feature learning, but the models proposed so far are not identifiable. Here, we first propose a new intuitive principle of unsupervised deep learning from time series which uses the nonstationary structure of the data. Our learning principle, time-contrastive learning (TCL), finds a representation which allows optimal discrimination of time segments (windows). Surprisingly, we show how TCL can be related to a nonlinear ICA model, when ICA is redefined to include temporal nonstationarities. In particular, we show that TCL combined with linear ICA estimates the nonlinear ICA model up to point-wise transformations of the sources, and this solution is unique --- thus providing the first identifiability result for nonlinear ICA which is rigorous, constructive, as well as very general.

****

Leveraging Sparsity for Efficient Submodular Data Summarization

Erik Lindgren, Shanshan Wu, Alexandros G. Dimakis

The facility location problem is widely used for summarizing large datasets and has additional applications in sensor placement, image retrieval, and clustering. One difficulty of this problem is that submodular optimization algorithms require the calculation of pairwise benefits for all items in the dataset. This is infeasible for large problems, so recent work proposed to only calculate nearest neighbor benefits. One limitation is that several strong assumptions were invoked to obtain provable approximation guarantees. In this paper we establish that these extra assumptions are not necessary—solving the sparsified problem will be almost optimal under the standard assumptions of the problem. We then analyze a different method of sparsification that is a better model for methods such as Locality Sensitive Hashing to accelerate the nearest neighbor computations and extend the use of the problem to a broader family of similarities. We validate our approach by demonstrating that it rapidly generates interpretable summaries.

****

Mistake Bounds for Binary Matrix Completion

Mark Herbster, Stephen Pasteris, Massimiliano Pontil

We study the problem of completing a binary matrix in an online learning setting. On each trial we predict a matrix entry and then receive the true entry. We propose a Matrix Exponentiated Gradient algorithm [1] to solve this problem. We provide a mistake bound for the algorithm, which scales with the margin complexity [2, 3] of the underlying matrix. The bound suggests an interpretation where each row of the matrix is a prediction task over a finite set of objects, the columns. Using this we show that the algorithm makes a number of mistakes which is comparable up to a logarithmic factor to the number of mistakes made by the Kernel Perceptron with an optimal kernel in hindsight. We discuss applications of the algorithm to predicting as well as the best biclustering and to the problem of predicting the labeling of a graph without knowing the graph in advance.

****

Quantum Perceptron Models

Ashish Kapoor, Nathan Wiebe, Krysta Svore

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.

****

Direct Feedback Alignment Provides Learning in Deep Neural Networks

Arild Nøkland

Artificial neural networks are most commonly trained with the back-propagation algorithm, where the gradient for learning is provided by back-propagating the er ror, layer by layer, from the output layer to the hidden layers. A recently discovered method called feedback-alignment shows that the weights used for propagating the error backward don't have to be symmetric with the weights used for propagation the activation forward. In fact, random feedback weights work evenly well, because the network learns how to make the feedback useful. In this work, the feedback alignment principle is used for training hidden layers more independently from the rest of the network, and from a zero initial condition. The error is propagated through fixed random feedback connections directly from the output layer to each hidden layer. This simple method is able to achieve zero training error even in convolutional networks and very deep networks, completely without error back-propagation. The method is a step towards biologically plausible machine learning because the error signal is almost local, and no symmetric or reciprocal weights are required. Experiments show that the test performance on MNIST and CIFAR is almost as good as those obtained with back-propagation for fully connected networks. If combined with dropout, the method achieves 1.45% error on the permutation invariant MNIST task.

****

Average-case hardness of RIP certification

Tengyao Wang, Quentin Berthet, Yaniv Plan

The restricted isometry property (RIP) for design matrices gives guarantees for optimal recovery in sparse linear models. It is of high interest in compressed sensing and statistical learning. This property is particularly important for computationally efficient recovery methods. As a consequence, even though it is in general NP-hard to check that RIP holds, there have been substantial efforts to find tractable proxies for it. These would allow the construction of RIP matrices and the polynomial-time verification of RIP given an arbitrary matrix. We consider the framework of average-case certifiers, that never wrongly declare that a matrix is RIP, while being often correct for random instances. While there are such functions which are tractable in a suboptimal parameter regime, we show that this is a computationally hard task in any better regime. Our results are based on a new, weaker assumption on the problem of detecting dense subgraphs.

****

Efficient High-Order Interaction-Aware Feature Selection Based on Conditional Mutual Information

Alexander Shishkin, Anastasia Bezzubtseva, Alexey Drutsa, Ilia Shishkov, Ekaterina Gladkikh, Gleb Gusev, Pavel Serdyukov

This study introduces a novel feature selection approach CMICOT, which is a further evolution of filter methods with sequential forward selection (SFS) whose scoring functions are based on conditional mutual information (MI). We state and study a novel saddle point (max-min) optimization problem to build a scoring function that is able to identify joint interactions between several features. This method fills the gap of MI-based SFS techniques with high-order dependencies. In this high-dimensional case, the estimation of MI has prohibitively high sample complexity. We mitigate this cost using a greedy approximation and binary representatives what makes our technique able to be effectively used. The superiority of our approach is demonstrated by comparison with recently proposed interaction-aware filters and several interaction-agnostic state-of-the-art ones on ten publicly available benchmark datasets.

****

Fast and Provably Good Seedings for k-Means

Olivier Bachem, Mario Lucic, Hamed Hassani, Andreas Krause

Seeding - the task of finding initial cluster centers - is critical in obtaining high-quality clusterings for k-Means. However, k-means++ seeding, the state of the art algorithm, does not scale well to massive datasets as it is inherently sequential and requires k full passes through the data. It was recently shown that Markov chain Monte Carlo sampling can be used to efficiently approximate the seeding step of k-means++. However, this result requires assumptions on the data generating distribution. We propose a simple yet fast seeding algorithm that produces provably good clusterings even without assumptions on the data. Our analysis shows that the algorithm allows for a favourable trade-off between solution

quality and computational cost, speeding up k-means++ seeding by up to several orders of magnitude. We validate our theoretical results in extensive experiments on a variety of real-world data sets.
****

Anchor-Free Correlated Topic Modeling: Identifiability and Algorithm

Kejun Huang, Xiao Fu, Nikolaos D. Sidiropoulos

In topic modeling, many algorithms that guarantee identifiability of the topics have been developed under the premise that there exist anchor words -- i.e., words that only appear (with positive probability) in one topic. Follow-up work has resorted to three or higher-order statistics of the data corpus to relax the anchor word assumption. Reliable estimates of higher-order statistics are hard to obtain, however, and the identification of topics under those models hinges on uncorrelatedness of the topics, which can be unrealistic. This paper revisits topic modeling based on second-order moments, and proposes an anchor-free topic mining framework. The proposed approach guarantees the identification of the topics under a much milder condition compared to the anchor-word assumption, thereby exhibiting much better robustness in practice. The associated algorithm only involves one eigen-decomposition and a few small linear programs. This makes it easy to implement and scale up to very large problem instances. Experiments using the TDT2 and Reuters-21578 corpus demonstrate that the proposed anchor-free approach exhibits very favorable performance (measured using coherence, similarity count, and clustering accuracy metrics) compared to the prior art.
****

High Dimensional Structured Superposition Models

Qilong Gu, Arindam Banerjee

High dimensional superposition models characterize observations using parameters which can be written as a sum of multiple component parameters, each with its own structure, e.g., sum of low rank and sparse matrices. In this paper, we consider general superposition models which allow sum of any number of component parameters, and each component structure can be characterized by any norm. We present a simple estimator for such models, give a geometric condition under which the components can be accurately estimated, characterize sample complexity of the estimator, and give non-asymptotic bounds on the componentwise estimation error. We use tools from empirical processes and generic chaining for the statistical analysis, and our results, which substantially generalize prior work on superposition models, are in terms of Gaussian widths of suitable spherical caps.
****

A Bandit Framework for Strategic Regression

Yang Liu, Yiling 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-authors prior to requesting a name change in the electronic proceedings.
****

Linear Relaxations for Finding Diverse Elements in Metric Spaces

Aditya Bhaskara, Mehrdad Ghadiri, Vahab Mirrokni, Ola Svensson

Choosing a diverse subset of a large collection of points in a metric space is a fundamental problem, with applications in feature selection, recommender systems, web search, data summarization, etc. Various notions of diversity have been proposed, tailored to different applications. The general algorithmic goal is to find a subset of points that maximize diversity, while obeying a cardinality (or more generally, matroid) constraint. The goal of this paper is to develop a novel linear programming (LP) framework that allows us to design approximation algorithms for such problems. We study an objective known as {\em sum-min} diversity, which is known to be effective in many applications, and give the first constant factor approximation algorithm. Our LP framework allows us to easily incorporate additional constraints, as well as secondary objectives. We also prove a hardness result for two natural diversity objectives, under the so-called {\em planted clique} assumption. Finally, we study the empirical performance of our algorithm on several standard datasets. We first study the approximation quality of the algorithm by comparing with the LP objective. Then, we compare the quality of the solutions produced by our method with other popular diversity maximization algorithms.

****

Binarized Neural Networks
Itay Hubara, Matthieu Courbariaux, Daniel Soudry, Ran El-Yaniv, Yoshua Bengio

We introduce a method to train Binarized Neural Networks (BNNs) - neural networks with binary weights and activations at run-time. At train-time the binary weights and activations are used for computing the parameter gradients. During the forward pass, BNNs drastically reduce memory size and accesses, and replace most arithmetic operations with bit-wise operations, which is expected to substantially improve power-efficiency. To validate the effectiveness of BNNs, we conducted two sets of experiments on the Torch7 and Theano frameworks. On both, BNNs achieved nearly state-of-the-art results over the MNIST, CIFAR-10 and SVHN datasets. We also report our preliminary results on the challenging ImageNet dataset. Last but not least, we wrote a binary matrix multiplication GPU kernel with which it is possible to run our MNIST BNN 7 times faster than with an unoptimized GPU kernel, without suffering any loss in classification accuracy. The code for training and running our BNNs is available on-line.

****

Learning a Metric Embedding for Face Recognition using the Multibatch Method
Oren Tadmor, Tal Rosenwein, Shai Shalev-Shwartz, Yonatan Wexler, Amnon Shashua

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.

****

Operator Variational Inference
Rajesh Ranganath, Dustin Tran, Jaan Altosaar, David Blei

Variational inference is an umbrella term for algorithms which cast Bayesian inference as optimization. Classically, variational inference uses the Kullback-Leibler divergence to define the optimization. Though this divergence has been widely used, the resultant posterior approximation can suffer from undesirable statistical properties. To address this, we reexamine variational inference from its roots as an optimization problem. We use operators, or functions of functions, to design variational objectives. As one example, we design a variational objective with a Langevin-Stein operator. We develop a black box algorithm, operator variational inference (OPVI), for optimizing any operator objective. Importantly, operators enable us to make explicit the statistical and computational tradeoffs for variational inference. We can characterize different properties of variational objectives, such as objectives that admit data subsampling---allowing inference to scale to massive data---as well as objectives that admit variational programs---a rich class of posterior approximations that does not require a tractable density. We illustrate the benefits of OPVI on a mixture model and a generative model of images.

****

Unsupervised Learning for Physical Interaction through Video Prediction
Chelsea Finn, Ian Goodfellow, Sergey Levine

A core challenge for an agent learning to interact with the world is to predict how its actions affect objects in its environment. Many existing methods for learning the dynamics of physical interactions require labeled object information. However, to scale real-world interaction learning to a variety of scenes and objects, acquiring labeled data becomes increasingly impractical. To learn about physical object motion without labels, we develop an action-conditioned video prediction model that explicitly models pixel motion, by predicting a distribution over pixel motion from previous frames. Because our model explicitly predicts motion, it is partially invariant to object appearance, enabling it to generalize to previously unseen objects. To explore video prediction for real-world interactive agents, we also introduce a dataset of 59,000 robot interactions involving pushing motions, including a test set with novel objects. In this dataset, accura te prediction of videos conditioned on the robot's future actions amounts to learning a "visual imagination" of different futures based on different courses of action. Our experiments show that our proposed method produces more accurate video predictions both quantitatively and qualitatively, when compared to prior methods.

****

Full-Capacity Unitary Recurrent Neural Networks

Scott Wisdom, Thomas Powers, John Hershey, Jonathan Le Roux, Les Atlas

Recurrent neural networks are powerful models for processing sequential data, but they are generally plagued by vanishing and exploding gradient problems. Unitary recurrent neural networks (uRNNs), which use unitary recurrence matrices, have recently been proposed as a means to avoid these issues. However, in previous experiments, the recurrence matrices were restricted to be a product of parameterized unitary matrices, and an open question remains: when does such a parameterization fail to represent all unitary matrices, and how does this restricted representational capacity limit what can be learned? To address this question, we propose full-capacity uRNNs that optimize their recurrence matrix over all unitary matrices, leading to significantly improved performance over uRNNs that use a restricted-capacity recurrence matrix. Our contribution consists of two main components. First, we provide a theoretical argument to determine if a unitary parameterization has restricted capacity. Using this argument, we show that a recently proposed unitary parameterization has restricted capacity for hidden state dimension greater than 7. Second,we show how a complete, full-capacity unitary recurrence matrix can be optimized over the differentiable manifold of unitary matrices. The resulting multiplicative gradient step is very simple and does not require gradient clipping or learning rate adaptation. We confirm the utility of our claims by empirically evaluating our new full-capacity uRNNs on both synthetic and natural data, achieving superior performance compared to both LSTMs and the original restricted-capacity uRNNs.

****

Linear-Memory and Decomposition-Invariant Linearly Convergent Conditional Gradient Algorithm for Structured Polytopes

Dan Garber, Dan Garber, Ofer Meshi

Recently, several works have shown that natural modifications of the classical conditional gradient method (aka Frank-Wolfe algorithm) for constrained convex optimization, provably converge with a linear rate when the feasible set is a polytope, and the objective is smooth and strongly-convex. However, all of these results suffer from two significant shortcomings: i) large memory requirement due to the need to store an explicit convex decomposition of the current iterate, and as a consequence, large running-time overhead per iteration ii) the worst case convergence rate depends unfavorably on the dimension In this work we present a new conditional gradient variant and a corresponding analysis that improves on both of the above shortcomings. In particular, both memory and computation overheads are only linear in the dimension, and in addition, in case the optimal solution is sparse, the new convergence rate replaces a factor which is at least linear in the dimension in previous works, with a linear dependence on the number of non-zeros in the optimal solution At the heart of our method, and corresponding analysis, is a novel way to compute decomposition-invariant away-steps. While our theoretical guarantees do not apply to any polytope, they apply to several important structured polytopes that capture central concepts such as paths in graphs, perfect matchings in bipartite graphs, marginal distributions that arise in structured prediction tasks, and more. Our theoretical findings are complemented by empirical evidence that shows that our method delivers state-of-the-art performance.

****

Combining Adversarial Guarantees and Stochastic Fast Rates in Online Learning

Wouter M. Koolen, Peter Grünwald, Tim van Erven

We consider online learning algorithms that guarantee worst-case regret rates in adversarial environments (so they can be deployed safely and will perform robustly), yet adapt optimally to favorable stochastic environments (so they will per form well in a variety of settings of practical importance). We quantify the friendliness of stochastic environments by means of the well-known Bernstein (a.k.a. generalized Tsybakov margin) condition. For two recent algorithms (Squint for the Hedge setting and MetaGrad for online convex optimization) we show that the particular form of their data-dependent individual-sequence regret guarantees implies that they adapt automatically to the Bernstein parameters of the stochastic environment. We prove that these algorithms attain fast rates in their respective settings both in expectation and with high probability.

****

A Comprehensive Linear Speedup Analysis for Asynchronous Stochastic Parallel Optimization from Zeroth-Order to First-Order

Xiangru Lian, Huan Zhang, Cho-Jui Hsieh, Yijun Huang, Ji Liu

Asynchronous parallel optimization received substantial successes and extensive attention recently. One of core theoretical questions is how much speedup (or benefit) the asynchronous parallelization can bring to us. This paper provides a comprehensive and generic analysis to study the speedup property for a broad range of asynchronous parallel stochastic algorithms from the zeroth order to the first order methods. Our result recovers or improves existing analysis on special cases, provides more insights for understanding the asynchronous parallel behaviors, and suggests a novel asynchronous parallel zeroth order method for the first time. Our experiments provide novel applications of the proposed asynchronous parallel zeroth order method on hyper parameter tuning and model blending problems.

****

Noise-Tolerant Life-Long Matrix Completion via Adaptive Sampling

Maria-Florina F. Balcan, Hongyang 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.

****

Satisfying Real-world Goals with Dataset Constraints

Gabriel Goh, Andrew Cotter, Maya Gupta, Michael P. Friedlander

The goal of minimizing misclassification error on a training set is often just one of several real-world goals that might be defined on different datasets. For example, one may require a classifier to also make positive predictions at some specified rate for some subpopulation (fairness), or to achieve a specified empirical recall. Other real-world goals include reducing churn with respect to a previously deployed model, or stabilizing online training. In this paper we propose handling multiple goals on multiple datasets by training with dataset constraints, using the ramp penalty to accurately quantify costs, and present an efficient algorithm to approximately optimize the resulting non-convex constrained optimization problem. Experiments on both benchmark and real-world industry datasets demonstrate the effectiveness of our approach.

****

Launch and Iterate: Reducing Prediction Churn

Mahdi Milani Fard, Quentin Cormier, Kevin Canini, Maya Gupta

Practical applications of machine learning often involve successive training iterations with changes to features and training examples. Ideally, changes in the output of any new model should only be improvements (wins) over the previous iteration, but in practice the predictions may change neutrally for many examples, resulting in extra net-zero wins and losses, referred to as unnecessary churn. These changes in the predictions are problematic for usability for some applications, and make it harder and more expensive to measure if a change is statistically significant positive. In this paper, we formulate the problem and present a stabilization operator to regularize a classifier towards a previous classifier. We use a Markov chain Monte Carlo stabilization operator to produce a model with more consistent predictions without adversely affecting accuracy. We investigate the properties of the proposal with theoretical analysis. Experiments on benchmark datasets for different classification algorithms demonstrate the method and the resulting reduction in churn.

****

Constraints Based Convex Belief Propagation

Y■aniv Tenzer, Alex Schwing, Kevin Gimpel, Tamir Hazan

Inference in Markov random fields subject to consistency structure is a fundamental problem that arises in many real-life applications. In order to enforce consistency, classical approaches utilize consistency potentials or encode constraints over feasible instances. Unfortunately this comes at the price of a serious computational bottleneck. In this paper we suggest to tackle consistency by incorporating constraints on beliefs. This permits derivation of a closed-form message-passing algorithm which we refer to as the Constraints Based Convex Belief Propagation (CBCBP). Experiments show that CBCBP outperforms the standard approach while being at least an order of magnitude faster.

****

Data driven estimation of Laplace-Beltrami operator

Frederic Chazal, Ilaria Giulini, Bertrand Michel

Approximations of Laplace-Beltrami operators on manifolds through graph Laplacians have become popular tools in data analysis and machine learning. These discretized operators usually depend on bandwidth parameters whose tuning remains a theoretical and practical problem. In this paper, we address this problem for the unormalized graph Laplacian by establishing an oracle inequality that opens the door to a well-founded data-driven procedure for the bandwidth selection. Our approach relies on recent results by Lacour and Massart (2015) on the so-called Lepski's method.

****

The Robustness of Estimator Composition

Pingfan Tang, Jeff M. Phillips

We formalize notions of robustness for composite estimators via the notion of a breakdown point. A composite estimator successively applies two (or more) estimators: on data decomposed into disjoint parts, it applies the first estimator on each part, then the second estimator on the outputs of the first estimator. And so on, if the composition is of more than two estimators. Informally, the breakdown point is the minimum fraction of data points which if significantly modified will also significantly modify the output of the estimator, so it is typically desirable to have a large breakdown point. Our main result shows that, under mild conditions on the individual estimators, the breakdown point of the composite estimator is the product of the breakdown points of the individual estimators. We also demonstrate several scenarios, ranging from regression to statistical testing, where this analysis is easy to apply, useful in understanding worst case robustness, and sheds powerful insights onto the associated data analysis.

****

Active Learning from Imperfect Labelers

Songbai Yan, Kamalika Chaudhuri, Tara Javidi

We study active learning where the labeler can not only return incorrect labels but also abstain from labeling. We consider different noise and abstention conditions of the labeler. We propose an algorithm which utilizes abstention responses, and analyze its statistical consistency and query complexity under fairly natural assumptions on the noise and abstention rate of the labeler. This algorithm is adaptive in a sense that it can automatically request less queries with a more informed or less noisy labeler. We couple our algorithm with lower bounds to show that under some technical conditions, it achieves nearly optimal query complexity.

****

Improved Variational Inference with Inverse Autoregressive Flow

Durk P. Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, Max Welling

The framework of normalizing flows provides a general strategy for flexible variational inference of posteriors over latent variables. We propose a new type of normalizing flow, inverse autoregressive flow (IAF), that, in contrast to earlier published flows, scales well to high-dimensional latent spaces. The proposed f low consists of a chain of invertible transformations, where each transformation is based on an autoregressive neural network. In experiments, we show that IAF significantly improves upon diagonal Gaussian approximate posteriors. In addition, we demonstrate that a novel type of variational autoencoder, coupled with IAF, is competitive with neural autoregressive models in terms of attained log-likelihood on natural images, while allowing significantly faster synthesis.

****

Select-and-Sample for Spike-and-Slab Sparse Coding

Abdul-Saboor Sheikh, Jörg Lücke

Probabilistic inference serves as a popular model for neural processing. It is still unclear, however, how approximate probabilistic inference can be accurate and scalable to very high-dimensional continuous latent spaces. Especially as typical posteriors for sensory data can be expected to exhibit complex latent dependencies including multiple modes. Here, we study an approach that can efficiently be scaled while maintaining a richly structured posterior approximation under these conditions. As example model we use spike-and-slab sparse coding for V1 processing, and combine latent subspace selection with Gibbs sampling (select-and-sample). Unlike factored variational approaches, the method can maintain large numbers of posterior modes and complex latent dependencies. Unlike pure sampling, the method is scalable to very high-dimensional latent spaces. Among all sparse coding approaches with non-trivial posterior approximations (MAP or ICA-like models), we report the largest-scale results. In applications we firstly verify the approach by showing competitiveness in standard denoising benchmarks. Secondly, we use its scalability to, for the first time, study highly-overcomplete settings for V1 encoding using sophisticated posterior representations. More generally, our study shows that very accurate probabilistic inference for multi-modal posteriors with complex dependencies is tractable, functionally desirable and consistent with models for neural inference.

****

Learning Infinite RBMs with Frank-Wolfe

Wei Ping, Qiang Liu, Alexander T. Ihler

In this work, we propose an infinite restricted Boltzmann machine (RBM), whose maximum likelihood estimation (MLE) corresponds to a constrained convex optimization. We consider the Frank-Wolfe algorithm to solve the program, which provides a sparse solution that can be interpreted as inserting a hidden unit at each iteration, so that the optimization process takes the form of a sequence of finite models of increasing complexity. As a side benefit, this can be used to easily and efficiently identify an appropriate number of hidden units during the optimization. The resulting model can also be used as an initialization for typical state-of-the-art RBM training algorithms such as contrastive divergence, leading to models with consistently higher test likelihood than random initialization.

****

Faster Projection-free Convex Optimization over the Spectrahedron

Dan Garber, Dan Garber

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.

****

Improved Regret Bounds for Oracle-Based Adversarial Contextual Bandits

Vasilis Syrgkanis, Haipeng Luo, Akshay Krishnamurthy, 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.

****

Joint quantile regression in vector-valued RKHSs

Maxime Sangnier, Olivier Fercoq, Florence d'Alché-Buc

Addressing the will to give a more complete picture than an average relationship provided by standard regression, a novel framework for estimating and predictin

g simultaneously several conditional quantiles is introduced. The proposed methodology leverages kernel-based multi-task learning to curb the embarrassing phenomenon of quantile crossing, with a one-step estimation procedure and no post-processing. Moreover, this framework comes along with theoretical guarantees and an efficient coordinate descent learning algorithm. Numerical experiments on benchmark and real datasets highlight the enhancements of our approach regarding the prediction error, the crossing occurrences and the training time.
****
Kernel Bayesian Inference with Posterior Regularization
Yang Song, Jun Zhu, Yong Ren
We propose a vector-valued regression problem whose solution is equivalent to the reproducing kernel Hilbert space (RKHS) embedding of the Bayesian posterior distribution. This equivalence provides a new understanding of kernel Bayesian inference. Moreover, the optimization problem induces a new regularization for the posterior embedding estimator, which is faster and has comparable performance to the squared regularization in kernel Bayes' rule. This regularization coincides with a former thresholding approach used in kernel POMDPs whose consistency remains to be established. Our theoretical work solves this open problem and provides consistency analysis in regression settings. Based on our optimizational formulation, we propose a flexible Bayesian posterior regularization framework which for the first time enables us to put regularization at the distribution level. We apply this method to nonparametric state-space filtering tasks with extremely nonlinear dynamics and show performance gains over all other baselines.
****
Scaled Least Squares Estimator for GLMs in Large-Scale Problems
Murat A. Erdogdu, Lee H. Dicker, Mohsen Bayati
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.
****
Contextual semibandits via supervised learning oracles
Akshay Krishnamurthy, Alekh Agarwal, Miro Dudik
We study an online decision making problem where on each round a learner chooses a list of items based on some side information, receives a scalar feedback value for each individual item, and a reward that is linearly related to this feedback. These problems, known as contextual semibandits, arise in crowdsourcing, recommendation, and many other domains. This paper reduces contextual semibandits to supervised learning, allowing us to leverage powerful supervised learning methods in this partial-feedback setting. Our first reduction applies when the mapping from feedback to reward is known and leads to a computationally efficient algorithm with near-optimal regret. We show that this algorithm outperforms state-of-the-art approaches on real-world learning-to-rank datasets, demonstrating the advantage of oracle-based algorithms. Our second reduction applies to the previously unstudied setting when the linear mapping from feedback to reward is unknown. Our regret guarantees are superior to prior techniques that ignore the feedback.
****
Learning Treewidth-Bounded Bayesian Networks with Thousands of Variables
Mauro Scanagatta, Giorgio Corani, Cassio P. de Campos, Marco Zaffalon
We present a method for learning treewidth-bounded Bayesian networks from data sets containing thousands of variables. Bounding the treewidth of a Bayesian network greatly reduces the complexity of inferences. Yet, being a global property of the graph, it considerably increases the difficulty of the learning process. Our novel algorithm accomplishes this task, scaling both to large domains and to large treewidths. Our novel approach consistently outperforms the state of the art on experiments with up to thousands of variables.
****
Unsupervised Learning from Noisy Networks with Applications to Hi-C Data
Bo Wang, Junjie Zhu, Armin Pourshafeie, Oana Ursu, Serafim Batzoglou, Anshul Kun

daje
Complex networks play an important role in a plethora of disciplines in natural sciences. Cleaning up noisy observed networks, poses an important challenge in network analysis Existing methods utilize labeled data to alleviate the noise effect in the network. However, labeled data is usually expensive to collect while unlabeled data can be gathered cheaply. In this paper, we propose an optimization framework to mine useful structures from noisy networks in an unsupervised manner. The key feature of our optimization framework is its ability to utilize local structures as well as global patterns in the network. We extend our method to incorporate multi-resolution networks in order to add further resistance to high-levels of noise. We also generalize our framework to utilize partial labels to enhance the performance. We specifically focus our method on multi-resolution Hi-C data by recovering clusters of genomic regions that co-localize in 3D space. Additionally, we use Capture-C-generated partial labels to further denoise the Hi-C network. We empirically demonstrate the effectiveness of our framework in denoising the network and improving community detection results.
****

Scan Order in Gibbs Sampling: Models in Which it Matters and Bounds on How Much
Bryan D. He, Christopher M. De Sa, Ioannis Mitliagkas, Christopher Ré
Gibbs sampling is a Markov Chain Monte Carlo sampling technique that iteratively samples variables from their conditional distributions. There are two common scan orders for the variables: random scan and systematic scan. Due to the benefits of locality in hardware, systematic scan is commonly used, even though most statistical guarantees are only for random scan. While it has been conjectured that the mixing times of random scan and systematic scan do not differ by more than a logarithmic factor, we show by counterexample that this is not the case, and we prove that that the mixing times do not differ by more than a polynomial factor under mild conditions. To prove these relative bounds, we introduce a method of augmenting the state space to study systematic scan using conductance.
****

Deep Neural Networks with Inexact Matching for Person Re-Identification
Arulkumar Subramaniam, Moitreya Chatterjee, Anurag Mittal
Person Re-Identification is the task of matching images of a person across multiple camera views. Almost all prior approaches address this challenge by attempting to learn the possible transformations that relate the different views of a person from a training corpora. Then, they utilize these transformation patterns for matching a query image to those in a gallery image bank at test time. This necessitates learning good feature representations of the images and having a robust feature matching technique. Deep learning approaches, such as Convolutional Neural Networks (CNN), simultaneously do both and have shown great promise recently. In this work, we propose two CNN-based architectures for Person Re-Identification. In the first, given a pair of images, we extract feature maps from these images via multiple stages of convolution and pooling. A novel inexact matching technique then matches pixels in the first representation with those of the second. Furthermore, we search across a wider region in the second representation for matching. Our novel matching technique allows us to tackle the challenges posed by large viewpoint variations, illumination changes or partial occlusions. Our approach shows a promising performance and requires only about half the parameters as a current state-of-the-art technique. Nonetheless, it also suffers from false matches at times. In order to mitigate this issue, we propose a fused architecture that combines our inexact matching pipeline with a state-of-the-art exact matching technique. We observe substantial gains with the fused model over the current state-of-the-art on multiple challenging datasets of varying sizes, with gains of up to about 21%.
****

Efficient Neural Codes under Metabolic Constraints
Zhuo Wang, Xue-Xin Wei, Alan A. Stocker, Daniel D. Lee
Neural codes are inevitably shaped by various kinds of biological constraints, \emph{e.g.} noise and metabolic cost. Here we formulate a coding framework which explicitly deals with noise and the metabolic costs associated with the neural r

epresentation of information, and analytically derive the optimal neural code for monotonic response functions and arbitrary stimulus distributions. For a single neuron, the theory predicts a family of optimal response functions depending on the metabolic budget and noise characteristics. Interestingly, the well-known histogram equalization solution can be viewed as a special case when metabolic resources are unlimited. For a pair of neurons, our theory suggests that under more severe metabolic constraints, ON-OFF coding is an increasingly more efficient coding scheme compared to ON-ON or OFF-OFF. The advantage could be as large as one-fold, substantially larger than the previous estimation. Some of these predictions could be generalized to the case of large neural populations. In particular, these analytical results may provide a theoretical basis for the predominant segregation into ON- and OFF-cells in early visual processing areas. Overall, we provide a unified framework for optimal neural codes with monotonic tuning curves in the brain, and makes predictions that can be directly tested with physiology experiments.

****

Learning Kernels with Random Features

Aman Sinha, John C. Duchi

Randomized features provide a computationally efficient way to approximate kernel machines in machine learning tasks. However, such methods require a user-defined kernel as input. We extend the randomized-feature approach to the task of learning a kernel (via its associated random features). Specifically, we present an efficient optimization problem that learns a kernel in a supervised manner. We prove the consistency of the estimated kernel as well as generalization bounds for the class of estimators induced by the optimized kernel, and we experimentally evaluate our technique on several datasets. Our approach is efficient and highly scalable, and we attain competitive results with a fraction of the training cost of other techniques.

****

Combinatorial semi-bandit with known covariance

Rémy Degenne, Vianney Perchet

The combinatorial stochastic semi-bandit problem is an extension of the classical multi-armed bandit problem in which an algorithm pulls more than one arm at each stage and the rewards of all pulled arms are revealed. One difference with the single arm variant is that the dependency structure of the arms is crucial. Previous works on this setting either used a worst-case approach or imposed independence of the arms. We introduce a way to quantify the dependency structure of the problem and design an algorithm that adapts to it. The algorithm is based on linear regression and the analysis uses techniques from the linear bandit literature. By comparing its performance to a new lower bound, we prove that it is optimal, up to a poly-logarithmic factor in the number of arms pulled.

****

Perspective Transformer Nets: Learning Single-View 3D Object Reconstruction without 3D Supervision

Xinchen Yan, Jimei Yang, Ersin Yumer, Yijie Guo, Honglak Lee

Understanding the 3D world is a fundamental problem in computer vision. However, learning a good representation of 3D objects is still an open problem due to the high dimensionality of the data and many factors of variation involved. In this work, we investigate the task of single-view 3D object reconstruction from a learning agent's perspective. We formulate the learning process as an interaction between 3D and 2D representations and propose an encoder-decoder network with a novel projection loss defined by the projective transformation. More importantly, the projection loss enables the unsupervised learning using 2D observation without explicit 3D supervision. We demonstrate the ability of the model in generating 3D volume from a single 2D image with three sets of experiments: (1) learning from single-class objects; (2) learning from multi-class objects and (3) testing on novel object classes. Results show superior performance and better generalization ability for 3D object reconstruction when the projection loss is involved.

****

Exact Recovery of Hard Thresholding Pursuit
Xiaotong Yuan, Ping Li, Tong 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.

****

Parameter Learning for Log-supermodular Distributions
Tatiana Shpakova, Francis Bach

We consider log-supermodular models on binary variables, which are probabilistic models with negative log-densities which are submodular. These models provide probabilistic interpretations of common combinatorial optimization tasks such as image segmentation. In this paper, we focus primarily on parameter estimation in the models from known upper-bounds on the intractable log-partition function. We show that the bound based on separable optimization on the base polytope of the submodular function is always inferior to a bound based on ``perturb-and-MAP'' ideas. Then, to learn parameters, given that our approximation of the log-partition function is an expectation (over our own randomization), we use a stochastic subgradient technique to maximize a lower-bound on the log-likelihood. This can also be extended to conditional maximum likelihood. We illustrate our new results in a set of experiments in binary image denoising, where we highlight the flexibility of a probabilistic model to learn with missing data.

****

A Multi-step Inertial Forward-Backward Splitting Method for Non-convex Optimization
Jingwei Liang, Jalal Fadili, Gabriel Peyré

In this paper, we propose a multi-step inertial Forward--Backward splitting algorithm for minimizing the sum of two non-necessarily convex functions, one of which is proper lower semi-continuous while the other is differentiable with a Lipschitz continuous gradient. We first prove global convergence of the scheme with the help of the Kurdyka–■ojasiewicz property. Then, when the non-smooth part is also partly smooth relative to a smooth submanifold, we establish finite identification of the latter and provide sharp local linear convergence analysis. The proposed method is illustrated on a few problems arising from statistics and machine learning.

****

Optimal Binary Classifier Aggregation for General Losses
Akshay Balsubramani, Yoav S. Freund

We address the problem of aggregating an ensemble of predictors with known loss bounds in a semi-supervised binary classification setting, to minimize prediction loss incurred on the unlabeled data. We find the minimax optimal predictions for a very general class of loss functions including all convex and many non-convex losses, extending a recent analysis of the problem for misclassification error. The result is a family of semi-supervised ensemble aggregation algorithms which are as efficient as linear learning by convex optimization, but are minimax optimal without any relaxations. Their decision rules take a form familiar in decision theory -- applying sigmoid functions to a notion of ensemble margin -- without the assumptions typically made in margin-based learning.

****

Dense Associative Memory for Pattern Recognition
Dmitry Krotov, John J. Hopfield

A model of associative memory is studied, which stores and reliably retrieves many more patterns than the number of neurons in the network. We propose a simple duality between this dense associative memory and neural networks commonly used in deep learning. On the associative memory side of this duality, a family of models that smoothly interpolates between two limiting cases can be constructed. One limit is referred to as the feature-matching mode of pattern recognition, and the other one as the prototype regime. On the deep learning side of the duality, this family corresponds to feedforward neural networks with one hidden layer and various activation functions, which transmit the activities of the visible

neurons to the hidden layer. This family of activation functions includes logistics, rectified linear units, and rectified polynomials of higher degrees. The proposed duality makes it possible to apply energy-based intuition from associative memory to analyze computational properties of neural networks with unusual activation functions - the higher rectified polynomials which until now have not been used in deep learning. The utility of the dense memories is illustrated for two test cases: the logical gate XOR and the recognition of handwritten digits from the MNIST data set.
****

Fairness in Learning: Classic and Contextual Bandits
Matthew Joseph, Michael Kearns, Jamie H. Morgenstern, Aaron Roth
We introduce the study of fairness in multi-armed bandit problems. Our fairness definition demands that, given a pool of applicants, a worse applicant is never favored over a better one, despite a learning algorithm’s uncertainty over the true payoffs. In the classic stochastic bandits problem we provide a provably fair algorithm based on “chained” confidence intervals, and prove a cumulative regret bound with a cubic dependence on the number of arms. We further show that any fair algorithm must have such a dependence, providing a strong separation between fair and unfair learning that extends to the general contextual case. In the general contextual case, we prove a tight connection between fairness and the KWIK (Knows What It Knows) learning model: a KWIK algorithm for a class of functions can be transformed into a provably fair contextual bandit algorithm and vice versa. This tight connection allows us to provide a provably fair algorithm for the linear contextual bandit problem with a polynomial dependence on the dimension, and to show (for a different class of functions) a worst-case exponential gap in regret between fair and non-fair learning algorithms.
****

Variational Autoencoder for Deep Learning of Images, Labels and Captions
Yunchen Pu, Zhe Gan, Ricardo Henao, Xin Yuan, Chunyuan Li, Andrew Stevens, Lawrence Carin
A novel variational autoencoder is developed to model images, as well as associated labels or captions. The Deep Generative Deconvolutional Network (DGDN) is used as a decoder of the latent image features, and a deep Convolutional Neural Network (CNN) is used as an image encoder; the CNN is used to approximate a distribution for the latent DGDN features/code. The latent code is also linked to generative models for labels (Bayesian support vector machine) or captions (recurrent neural network). When predicting a label/caption for a new image at test, averaging is performed across the distribution of latent codes; this is computationally efficient as a consequence of the learned CNN-based encoder. Since the framework is capable of modeling the image in the presence/absence of associated labels/captions, a new semi-supervised setting is manifested for CNN learning with images; the framework even allows unsupervised CNN learning, based on images alone.
****

Weight Normalization: A Simple Reparameterization to Accelerate Training of Deep Neural Networks
Tim Salimans, Durk P. Kingma
We present weight normalization: a reparameterization of the weight vectors in a neural network that decouples the length of those weight vectors from their direction. By reparameterizing the weights in this way we improve the conditioning of the optimization problem and we speed up convergence of stochastic gradient descent. Our reparameterization is inspired by batch normalization but does not introduce any dependencies between the examples in a minibatch. This means that our method can also be applied successfully to recurrent models such as LSTMs and to noise-sensitive applications such as deep reinforcement learning or generative models, for which batch normalization is less well suited. Although our method is much simpler, it still provides much of the speed-up of full batch normalization. In addition, the computational overhead of our method is lower, permitting more optimization steps to be taken in the same amount of time. We demonstrate the usefulness of our method on applications in supervised image recognition, g

enerative modelling, and deep reinforcement learning.

****

Learning Additive Exponential Family Graphical Models via $\ell_{2,1}$-norm Regularized M-Estimation

Xiaotong Yuan, Ping Li, Tong Zhang, Qingshan Liu, Guangcan 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.

****

Disentangling factors of variation in deep representation using adversarial training

Michael F. Mathieu, Junbo Jake Zhao, Junbo Zhao, Aditya Ramesh, Pablo Sprechmann, Yann LeCun

We propose a deep generative model for learning to distill the hidden factors of variation within a set of labeled observations into two complementary codes. One code describes the factors of variation relevant to solving a specified task. The other code describes the remaining factors of variation that are irrelevant to solving this task. The only available source of supervision during the training process comes from our ability to distinguish among different observations belonging to the same category. Concrete examples include multiple images of the same object from different viewpoints, or multiple speech samples from the same speaker. In both of these instances, the factors of variation irrelevant to classification are implicitly expressed by intra-class variabilities, such as the relative position of an object in an image, or the linguistic content of an utterance. Most existing approaches for solving this problem rely heavily on having access to pairs of observations only sharing a single factor of variation, e.g. different objects observed in the exact same conditions. This assumption is often not encountered in realistic settings where data acquisition is not controlled and labels for the uninformative components are not available. In this work, we propose to overcome this limitation by augmenting deep convolutional autoencoders with a form of adversarial training. Both factors of variation are implicitly captured in the organization of the learned embedding space, and can be used for solving single-image analogies. Experimental results on synthetic and real datasets show that the proposed method is capable of disentangling the influences of style and content factors using a flexible representation, as well as generalizing to unseen styles or content classes.

****

Gaussian Processes for Survival Analysis

Tamara Fernandez, Nicolas Rivera, Yee Whye Teh

We introduce a semi-parametric Bayesian model for survival analysis. The model is centred on a parametric baseline hazard, and uses a Gaussian process to model variations away from it nonparametrically, as well as dependence on covariates. As opposed to many other methods in survival analysis, our framework does not impose unnecessary constraints in the hazard rate or in the survival function. Furthermore, our model handles left, right and interval censoring mechanisms common in survival analysis. We propose a MCMC algorithm to perform inference and an approximation scheme based on random Fourier features to make computations faster. We report experimental results on synthetic and real data, showing that our model performs better than competing models such as Cox proportional hazards, ANOVA-DDP and random survival forests.

****

Correlated-PCA: Principal Components' Analysis when Data and Noise are Correlated

Namrata Vaswani, Han Guo

Given a matrix of observed data, Principal Components Analysis (PCA) computes a small number of orthogonal directions that contain most of its variability. Provably accurate solutions for PCA have been in use for a long time. However, to the best of our knowledge, all existing theoretical guarantees for it assume that the data and the corrupting noise are mutually independent, or at least uncorre lated. This is valid in practice often, but not always. In this paper, we study the PCA problem in the setting where the data and noise can be correlated. Such noise is often also referred to as ``data-dependent noise". We obtain a correctness result for the standard eigenvalue decomposition (EVD) based solution to PCA under simple assumptions on the data-noise correlation. We also develop and analyze a generalization of EVD, cluster-EVD, that improves upon EVD in certain regimes.

****

On Explore-Then-Commit strategies

Aurelien Garivier, Tor Lattimore, Emilie Kaufmann

We study the problem of minimising regret in two-armed bandit problems with Gaussian rewards. Our objective is to use this simple setting to illustrate that strategies based on an exploration phase (up to a stopping time) followed by exploitation are necessarily suboptimal. The results hold regardless of whether or not the difference in means between the two arms is known. Besides the main message, we also refine existing deviation inequalities, which allow us to design fully sequential strategies with finite-time regret guarantees that are (a) asymptotically optimal as the horizon grows and (b) order-optimal in the minimax sense. Furthermore we provide empirical evidence that the theory also holds in practice and discuss extensions to non-gaussian and multiple-armed case.

****

Adaptive Neural Compilation

Rudy R. Bunel, Alban Desmaison, Pawan K. Mudigonda, Pushmeet Kohli, Philip Torr

This paper proposes an adaptive neural-compilation framework to address the problem of learning efficient program. Traditional code optimisation strategies used in compilers are based on applying pre-specified set of transformations that make the code faster to execute without changing its semantics. In contrast, our work involves adapting programs to make them more efficient while considering correctness only on a target input distribution. Our approach is inspired by the recent works on differentiable representations of programs. We show that it is possible to compile programs written in a low-level language to a differentiable representation. We also show how programs in this representation can be optimised to make them efficient on a target distribution of inputs. Experimental results demonstrate that our approach enables learning specifically-tuned algorithms for given data distributions with a high success rate.

****

Graphical Time Warping for Joint Alignment of Multiple Curves

Yizhi Wang, David J. Miller, Kira Poskanzer, Yue Wang, Lin Tian, Guoqiang Yu

Dynamic time warping (DTW) is a fundamental technique in time series analysis for comparing one curve to another using a flexible time-warping function. However, it was designed to compare a single pair of curves. In many applications, such as in metabolomics and image series analysis, alignment is simultaneously needed for multiple pairs. Because the underlying warping functions are often related, independent application of DTW to each pair is a sub-optimal solution. Yet, it is largely unknown how to efficiently conduct a joint alignment with all warping functions simultaneously considered, since any given warping function is constrained by the others and dynamic programming cannot be applied. In this paper, we show that the joint alignment problem can be transformed into a network flow problem and thus can be exactly and efficiently solved by the max flow algorithm, with a guarantee of global optimality. We name the proposed approach graphical time warping (GTW), emphasizing the graphical nature of the solution and that the dependency structure of the warping functions can be represented by a graph. Modifications of DTW, such as windowing and weighting, are readily derivable within GTW. We also discuss optimal tuning of parameters and hyperparameters in GTW. We illustrate the power of GTW using both synthetic data and a real case study of an astrocyte calcium movie.

****

PerforatedCNNs: Acceleration through Elimination of Redundant Convolutions

Mikhail Figurnov, Aizhan Ibraimova, Dmitry P. Vetrov, Pushmeet Kohli

We propose a novel approach to reduce the computational cost of evaluation of co nvolutional neural networks, a factor that has hindered their deployment in low-power devices such as mobile phones. Inspired by the loop perforation technique from source code optimization, we speed up the bottleneck convolutional layers by skipping their evaluation in some of the spatial positions. We propose and analyze several strategies of choosing these positions. We demonstrate that perforation can accelerate modern convolutional networks such as AlexNet and VGG-16 by a factor of 2x - 4x. Additionally, we show that perforation is complementary to the recently proposed acceleration method of Zhang et al.

****

DeepMath - Deep Sequence Models for Premise Selection
Geoffrey Irving, Christian Szegedy, Alexander A. Alemi, Niklas Een, Francois Chollet, Josef Urban

We study the effectiveness of neural sequence models for premise selection in automated theorem proving, a key bottleneck for progress in formalized mathematics. We propose a two stage approach for this task that yields good results for the premise selection task on the Mizar corpus while avoiding the hand-engineered features of existing state-of-the-art models. To our knowledge, this is the first time deep learning has been applied theorem proving on a large scale.

****

A Pseudo-Bayesian Algorithm for Robust PCA
Tae-Hyun Oh, Yasuyuki Matsushita, In Kweon, David Wipf

Commonly used in many applications, robust PCA represents an algorithmic attempt to reduce the sensitivity of classical PCA to outliers. The basic idea is to learn a decomposition of some data matrix of interest into low rank and sparse components, the latter representing unwanted outliers. Although the resulting problem is typically NP-hard, convex relaxations provide a computationally-expedient alternative with theoretical support. However, in practical regimes performance guarantees break down and a variety of non-convex alternatives, including Bayesian-inspired models, have been proposed to boost estimation quality. Unfortunately though, without additional a priori knowledge none of these methods can significantly expand the critical operational range such that exact principal subspace recovery is possible. Into this mix we propose a novel pseudo-Bayesian algorithm that explicitly compensates for design weaknesses in many existing non-convex approaches leading to state-of-the-art performance with a sound analytical foundation.

****

The Forget-me-not Process
Kieran Milan, Joel Veness, James Kirkpatrick, Michael Bowling, Anna Koop, Demis Hassabis

We introduce the Forget-me-not Process, an efficient, non-parametric meta-algorithm for online probabilistic sequence prediction for piecewise stationary, repeating sources. Our method works by taking a Bayesian approach to partition a stream of data into postulated task-specific segments, while simultaneously building a model for each task. We provide regret guarantees with respect to piecewise stationary data sources under the logarithmic loss, and validate the method empirically across a range of sequence prediction and task identification problems.

****

Unsupervised Risk Estimation Using Only Conditional Independence Structure
Jacob Steinhardt, Percy S. Liang

We show how to estimate a model’s test error from unlabeled data, on distributions very different from the training distribution, while assuming only that certain conditional independencies are preserved between train and test. We do not need to assume that the optimal predictor is the same between train and test, or that the true distribution lies in any parametric family. We can also efficiently compute gradients of the estimated error and hence perform unsupervised discriminative learning. Our technical tool is the method of moments, which allows us to exploit conditional independencies in the absence of a fully-specified model. Our framework encompasses a large family of losses including the log and exponential loss, and extends to structured output settings such as conditional random fields.

****

Deep Learning without Poor Local Minima

Kenji Kawaguchi

In this paper, we prove a conjecture published in 1989 and also partially address an open problem announced at the Conference on Learning Theory (COLT) 2015. For an expected loss function of a deep nonlinear neural network, we prove the following statements under the independence assumption adopted from recent work: 1) the function is non-convex and non-concave, 2) every local minimum is a global minimum, 3) every critical point that is not a global minimum is a saddle point, and 4) the property of saddle points differs for shallow networks (with three layers) and deeper networks (with more than three layers). Moreover, we prove that the same four statements hold for deep linear neural networks with any depth, any widths and no unrealistic assumptions. As a result, we present an instance, for which we can answer to the following question: how difficult to directly train a deep model in theory? It is more difficult than the classical machine learning models (because of the non-convexity), but not too difficult (because of the nonexistence of poor local minima and the property of the saddle points). We note that even though we have advanced the theoretical foundations of deep learning, there is still a gap between theory and practice.

****

Linear Contextual Bandits with Knapsacks

Shipra Agrawal, Nikhil Devanur

We consider the linear contextual bandit problem with resource consumption, in addition to reward generation. In each round, the outcome of pulling an arm is a reward as well as a vector of resource consumptions. The expected values of these outcomes depend linearly on the context of that arm. The budget/capacity constraints require that the sum of these vectors doesn't exceed the budget in each dimension. The objective is once again to maximize the total reward. This problem turns out to be a common generalization of classic linear contextual bandits (linContextual), bandits with knapsacks (BwK), and the online stochastic packing problem (OSPP). We present algorithms with near-optimal regret bounds for this problem. Our bounds compare favorably to results on the unstructured version of the problem, where the relation between the contexts and the outcomes could be arbitrary, but the algorithm only competes against a fixed set of policies accessible through an optimization oracle. We combine techniques from the work on linContextual, BwK and OSPP in a nontrivial manner while also tackling new difficulties that are not present in any of these special cases.

****

High resolution neural connectivity from incomplete tracing data using nonnegative spline regression

Kameron D. Harris, Stefan Mihalas, Eric Shea-Brown

Whole-brain neural connectivity data are now available from viral tracing experiments, which reveal the connections between a source injection site and elsewhere in the brain. These hold the promise of revealing spatial patterns of connectivity throughout the mammalian brain. To achieve this goal, we seek to fit a weighted, nonnegative adjacency matrix among 100 µm brain “voxels” using viral tracer data. Despite a multi-year experimental effort, injections provide incomplete coverage, and the number of voxels in our data is orders of magnitude larger than the number of injections, making the problem severely underdetermined. Furthermore, projection data are missing within the injection site because local connections there are not separable from the injection signal. We use a novel machine-learning algorithm to meet these challenges and develop a spatially explicit, voxel-scale connectivity map of the mouse visual system. Our method combines three features: a matrix completion loss for missing data, a smoothing spline penalty to regularize the problem, and (optionally) a low rank factorization. We demonstrate the consistency of our estimator using synthetic data and then apply it to newly available Allen Mouse Brain Connectivity Atlas data for the visual system. Our algorithm is significantly more predictive than current state of the art approaches which assume regions to be homogeneous. We demonstrate the efficacy of a low rank version on visual cortex data and discuss the possibility of extend ing this to a whole-brain connectivity matrix at the voxel scale.

****

Learning and Forecasting Opinion Dynamics in Social Networks

Abir De, Isabel Valera, Niloy Ganguly, Sourangshu Bhattacharya, Manuel Gomez Rodriguez

Social media and social networking sites have become a global pinboard for exposition and discussion of news, topics, and ideas, where social media users often update their opinions about a particular topic by learning from the opinions shared by their friends. In this context, can we learn a data-driven model of opinion dynamics that is able to accurately forecast users' opinions? In this paper, we introduce SLANT, a probabilistic modeling framework of opinion dynamics, which represents users' opinions over time by means of marked jump diffusion stochastic differential equations, and allows for efficient model simulation and parameter estimation from historical fine grained event data. We then leverage our framework to derive a set of efficient predictive formulas for opinion forecasting and identify conditions under which opinions converge to a steady state. Experiments on data gathered from Twitter show that our model provides a good fit to the data and our formulas achieve more accurate forecasting than alternatives.

****

Lifelong Learning with Weighted Majority Votes

Anastasia Pentina, Ruth Urner

Better understanding of the potential benefits of information transfer and representation learning is an important step towards the goal of building intelligent systems that are able to persist in the world and learn over time. In this work, we consider a setting where the learner encounters a stream of tasks but is able to retain only limited information from each encountered task, such as a learned predictor. In contrast to most previous works analyzing this scenario, we do not make any distributional assumptions on the task generating process. Instead, we formulate a complexity measure that captures the diversity of the observed tasks. We provide a lifelong learning algorithm with error guarantees for every observed task (rather than on average). We show sample complexity reductions in comparison to solving every task in isolation in terms of our task complexity measure. Further, our algorithmic framework can naturally be viewed as learning a representation from encountered tasks with a neural network.

****

Depth from a Single Image by Harmonizing Overcomplete Local Network Predictions

Ayan Chakrabarti, Jingyu Shao, Greg Shakhnarovich

A single color image can contain many cues informative towards different aspects of local geometric structure. We approach the problem of monocular depth estimation by using a neural network to produce a mid-level representation that summarizes these cues. This network is trained to characterize local scene geometry by predicting, at every image location, depth derivatives of different orders, orientations and scales. However, instead of a single estimate for each derivative, the network outputs probability distributions that allow it to express confidence about some coefficients, and ambiguity about others. Scene depth is then estimated by harmonizing this overcomplete set of network predictions, using a globalization procedure that finds a single consistent depth map that best matches all the local derivative distributions. We demonstrate the efficacy of this approach through evaluation on the NYU v2 depth data set.

****

Ancestral Causal Inference

Sara Magliacane, Tom Claassen, Joris M. Mooij

Constraint-based causal discovery from limited data is a notoriously difficult challenge due to the many borderline independence test decisions. Several approaches to improve the reliability of the predictions by exploiting redundancy in the independence information have been proposed recently. Though promising, existing approaches can still be greatly improved in terms of accuracy and scalability. We present a novel method that reduces the combinatorial explosion of the search space by using a more coarse-grained representation of causal information, drastically reducing computation time. Additionally, we propose a method to score

causal predictions based on their confidence. Crucially, our implementation also allows one to easily combine observational and interventional data and to incorporate various types of available background knowledge. We prove soundness and asymptotic consistency of our method and demonstrate that it can outperform the state-of-the-art on synthetic data, achieving a speedup of several orders of magnitude. We illustrate its practical feasibility by applying it on a challenging protein data set.
****
Hierarchical Deep Reinforcement Learning: Integrating Temporal Abstraction and Intrinsic Motivation
Tejas D. Kulkarni, Karthik Narasimhan, Ardavan Saeedi, Josh Tenenbaum
Learning goal-directed behavior in environments with sparse feedback is a major challenge for reinforcement learning algorithms. One of the key difficulties is insufficient exploration, resulting in an agent being unable to learn robust policies. Intrinsically motivated agents can explore new behavior for their own sake rather than to directly solve external goals. Such intrinsic behaviors could eventually help the agent solve tasks posed by the environment. We present hierarchical-DQN (h-DQN), a framework to integrate hierarchical action-value functions, operating at different temporal scales, with goal-driven intrinsically motivated deep reinforcement learning. A top-level q-value function learns a policy over intrinsic goals, while a lower-level function learns a policy over atomic actions to satisfy the given goals. h-DQN allows for flexible goal specifications, such as functions over entities and relations. This provides an efficient space for exploration in complicated environments. We demonstrate the strength of our approach on two problems with very sparse and delayed feedback: (1) a complex discrete stochastic decision process with stochastic transitions, and (2) the classic ATARI game -- `Montezuma's Revenge'.
****
Agnostic Estimation for Misspecified Phase Retrieval Models
Matey Neykov, Zhaoran Wang, 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.
****
Can Peripheral Representations Improve Clutter Metrics on Complex Scenes?
Arturo Deza, Miguel Eckstein
Previous studies have proposed image-based clutter measures that correlate with human search times and/or eye movements. However, most models do not take into account the fact that the effects of clutter interact with the foveated nature of the human visual system: visual clutter further from the fovea has an increasing detrimental influence on perception. Here, we introduce a new foveated clutter model to predict the detrimental effects in target search utilizing a forced fixation search task. We use Feature Congestion (Rosenholtz et al.) as our non foveated clutter model, and we stack a peripheral architecture on top of Feature Congestion for our foveated model. We introduce the Peripheral Integration Feature Congestion (PIFC) coefficient, as a fundamental ingredient of our model that modulates clutter as a non-linear gain contingent on eccentricity. We finally show that Foveated Feature Congestion (FFC) clutter scores ($r(44) = -0.82 \pm 0.04$, $p < 0.0001$) correlate better with target detection (hit rate) than regular Feature Congestion ($r(44) = -0.19 \pm 0.13$, $p = 0.0774$) in forced fixation search; and we extend foveation to other clutter models showing stronger correlations in all cases. Thus, our model allows us to enrich clutter perception research by computing fixation specific clutter maps. Code for building peripheral representations is available.
****
Proximal Deep Structured Models
Shenlong Wang, Sanja Fidler, Raquel Urtasun
Many problems in real-world applications involve predicting continuous-valued random variables that are statistically related. In this paper, we propose a powe

rful deep structured model that is able to learn complex non-linear functions which encode the dependencies between continuous output variables. We show that inference in our model using proximal methods can be efficiently solved as a feed-foward pass of a special type of deep recurrent neural network. We demonstrate the effectiveness of our approach in the tasks of image denoising, depth refinement and optical flow estimation.

****

SPALS: Fast Alternating Least Squares via Implicit Leverage Scores Sampling
Dehua Cheng, Richard Peng, Yan Liu, Ioakeim Perros

Tensor CANDECOMP/PARAFAC (CP) decomposition is a powerful but computationally challenging tool in modern data analytics. In this paper, we show ways of sampling intermediate steps of alternating minimization algorithms for computing low rank tensor CP decompositions, leading to the sparse alternating least squares (SPALS) method. Specifically, we sample the the Khatri-Rao product, which arises as an intermediate object during the iterations of alternating least squares. This product captures the interactions between different tensor modes, and form the main computational bottleneck for solving many tensor related tasks. By exploiting the spectral structures of the matrix Khatri-Rao product, we provide efficient access to its statistical leverage scores. When applied to the tensor CP decomposition, our method leads to the first algorithm that runs in sublinear time per-iteration and approximates the output of deterministic alternating least squares algorithms. Empirical evaluations of this approach show significantly speedups over existing randomized and deterministic routines for performing CP decomposition. On a tensor of the size 2.4m by 6.6m by 92k with over 2 billion nonzeros formed by Amazon product reviews, our routine converges in two minutes to the same error as deterministic ALS.

****

On Multiplicative Integration with Recurrent Neural Networks
Yuhuai Wu, Saizheng Zhang, Ying Zhang, Yoshua Bengio, Russ R. Salakhutdinov

We introduce a general simple structural design called “Multiplicative Integration” (MI) to improve recurrent neural networks (RNNs). MI changes the way of how the information flow gets integrated in the computational building block of an RNN, while introducing almost no extra parameters. The new structure can be easily embedded into many popular RNN models, including LSTMs and GRUs. We empirically analyze its learning behaviour and conduct evaluations on several tasks using different RNN models. Our experimental results demonstrate that Multiplicative Integration can provide a substantial performance boost over many of the existing RNN models.

****

The Generalized Reparameterization Gradient
Francisco R. Ruiz, Michalis Titsias RC AUEB, David Blei

The reparameterization gradient has become a widely used method to obtain Monte Carlo gradients to optimize the variational objective. However, this technique does not easily apply to commonly used distributions such as beta or gamma without further approximations, and most practical applications of the reparameterization gradient fit Gaussian distributions. In this paper, we introduce the generalized reparameterization gradient, a method that extends the reparameterization gradient to a wider class of variational distributions. Generalized reparameterizations use invertible transformations of the latent variables which lead to transformed distributions that weakly depend on the variational parameters. This results in new Monte Carlo gradients that combine reparameterization gradients and score function gradients. We demonstrate our approach on variational inference for two complex probabilistic models. The generalized reparameterization is effective: even a single sample from the variational distribution is enough to obtain a low-variance gradient.

****

Semiparametric Differential Graph Models
Pan Xu, Quanquan Gu

In many cases of network analysis, it is more attractive to study how a network varies under different conditions than an individual static network. We propose a novel graphical model, namely Latent Differential Graph Model, where the networks under two different conditions are represented by two semiparametric elliptical distributions respectively, and the variation of these two networks (i.e., differential graph) is characterized by the difference between their latent precision matrices. We propose an estimator for the differential graph based on quasi likelihood maximization with nonconvex regularization. We show that our estimator attains a faster statistical rate in parameter estimation than the state-of-the-art methods, and enjoys oracle property under mild conditions. Thorough experiments on both synthetic and real world data support our theory.

****

Neural Universal Discrete Denoiser

Taesup Moon, Seonwoo Min, Byunghan Lee, Sungroh Yoon

We present a new framework of applying deep neural networks (DNN) to devise a universal discrete denoiser. Unlike other approaches that utilize supervised learning for denoising, we do not require any additional training data. In such setting, while the ground-truth label, i.e., the clean data, is not available, we devise ``pseudo-labels'' and a novel objective function such that DNN can be trained in a same way as supervised learning to become a discrete denoiser. We experimentally show that our resulting algorithm, dubbed as Neural DUDE, significantly outperforms the previous state-of-the-art in several applications with a systematic rule of choosing the hyperparameter, which is an attractive feature in practice.

****

Testing for Differences in Gaussian Graphical Models: Applications to Brain Connectivity

Eugene Belilovsky, Gaël Varoquaux, Matthew B. Blaschko

Functional brain networks are well described and estimated from data with Gaussian Graphical Models (GGMs), e.g.\ using sparse inverse covariance estimators. Comparing functional connectivity of subjects in two populations calls for comparing these estimated GGMs. Our goal is to identify differences in GGMs known to have similar structure. We characterize the uncertainty of differences with confidence intervals obtained using a parametric distribution on parameters of a sparse estimator. Sparse penalties enable statistical guarantees and interpretable models even in high-dimensional and low-sample settings. Characterizing the distributions of sparse models is inherently challenging as the penalties produce a biased estimator. Recent work invokes the sparsity assumptions to effectively remove the bias from a sparse estimator such as the lasso. These distributions can be used to give confidence intervals on edges in GGMs, and by extension their differences. However, in the case of comparing GGMs, these estimators do not make use of any assumed joint structure among the GGMs. Inspired by priors from brain functional connectivity we derive the distribution of parameter differences under a joint penalty when parameters are known to be sparse in the difference. This leads us to introduce the debiased multi-task fused lasso, whose distribution can be characterized in an efficient manner. We then show how the debiased lasso and multi-task fused lasso can be used to obtain confidence intervals on edge differences in GGMs. We validate the techniques proposed on a set of synthetic examples as well as neuro-imaging dataset created for the study of autism.

****

Learning to learn by gradient descent by gradient descent

Marcin Andrychowicz, Misha Denil, Sergio Gómez, Matthew W. Hoffman, David Pfau, Tom Schaul, Brendan Shillingford, Nando de Freitas

The move from hand-designed features to learned features in machine learning has been wildly successful. In spite of this, optimization algorithms are still designed by hand. In this paper we show how the design of an optimization algorithm can be cast as a learning problem, allowing the algorithm to learn to exploit structure in the problems of interest in an automatic way. Our learned algorithms, implemented by LSTMs, outperform generic, hand-designed competitors on the tasks for which they are trained, and also generalize well to new tasks with similar structure. We demonstrate this on a number of tasks, including simple convex problems, training neural networks, and styling images with neural art.

****
Mixed vine copulas as joint models of spike counts and local field potentials
Arno Onken, Stefano Panzeri
Concurrent measurements of neural activity at multiple scales, sometimes performed with multimodal techniques, become increasingly important for studying brain function. However, statistical methods for their concurrent analysis are currently lacking. Here we introduce such techniques in a framework based on vine copulas with mixed margins to construct multivariate stochastic models. These models can describe detailed mixed interactions between discrete variables such as neural spike counts, and continuous variables such as local field potentials. We propose efficient methods for likelihood calculation, inference, sampling and mutual information estimation within this framework. We test our methods on simulated data and demonstrate applicability on mixed data generated by a biologically realistic neural network. Our methods hold the promise to considerably improve statistical analysis of neural data recorded simultaneously at different scales.
****
Can Active Memory Replace Attention?
■ukasz Kaiser, Samy Bengio
Several mechanisms to focus attention of a neural network on selected parts of its input or memory have been used successfully in deep learning models in recent years. Attention has improved image classification, image captioning, speech recognition, generative models, and learning algorithmic tasks, but it had probably the largest impact on neural machine translation. Recently, similar improvements have been obtained using alternative mechanisms that do not focus on a single part of a memory but operate on all of it in parallel, in a uniform way. Such mechanism, which we call active memory, improved over attention in algorithmic tasks, image processing, and in generative modelling. So far, however, active memory has not improved over attention for most natural language processing tasks, in particular for machine translation. We analyze this shortcoming in this paper and propose an extended model of active memory that matches existing attention models on neural machine translation and generalizes better to longer sentences. We investigate this model and explain why previous active memory models did not succeed. Finally, we discuss when active memory brings most benefits and where attention can be a better choice.
****
Fast Active Set Methods for Online Spike Inference from Calcium Imaging
Johannes Friedrich, Liam Paninski
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.
****
Human Decision-Making under Limited Time
Pedro A. Ortega, Alan A. Stocker
Abstract Subjective expected utility theory assumes that decision-makers possess unlimited computational resources to reason about their choices; however, virtually all decisions in everyday life are made under resource constraints---i.e. decision-makers are bounded in their rationality. Here we experimentally tested the predictions made by a formalization of bounded rationality based on ideas from statistical mechanics and information-theory. We systematically tested human subjects in their ability to solve combinatorial puzzles under different time limitations. We found that our bounded-rational model accounts well for the data. The decomposition of the fitted model parameter into the subjects' expected utility function and resource parameter provide interesting insight into the subjects' information capacity limits. Our results confirm that humans gradually fall back on their learned prior choice patterns when confronted with increasing resource limitations.
****
End-to-End Kernel Learning with Supervised Convolutional Kernel Networks
Julien Mairal

In this paper, we introduce a new image representation based on a multilayer kernel machine. Unlike traditional kernel methods where data representation is decoupled from the prediction task, we learn how to shape the kernel with supervision. We proceed by first proposing improvements of the recently-introduced convolutional kernel networks (CKNs) in the context of unsupervised learning; then, we derive backpropagation rules to take advantage of labeled training data. The resulting model is a new type of convolutional neural network, where optimizing the filters at each layer is equivalent to learning a linear subspace in a reproducing kernel Hilbert space (RKHS). We show that our method achieves reasonably competitive performance for image classification on some standard ``deep learning'' datasets such as CIFAR-10 and SVHN, and also for image super-resolution, demonstrating the applicability of our approach to a large variety of image-related tasks.

****

Dueling Bandits: Beyond Condorcet Winners to General Tournament Solutions
Siddartha Y. Ramamohan, Arun Rajkumar, Shivani Agarwal, Shivani Agarwal

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.

****

Visual Question Answering with Question Representation Update (QRU)
Ruiyu Li, Jiaya Jia

Our method aims at reasoning over natural language questions and visual images. Given a natural language question about an image, our model updates the question representation iteratively by selecting image regions relevant to the query and learns to give the correct answer. Our model contains several reasoning layers, exploiting complex visual relations in the visual question answering (VQA) task. The proposed network is end-to-end trainable through back-propagation, where its weights are initialized using pre-trained convolutional neural network (CNN) and gated recurrent unit (GRU). Our method is evaluated on challenging datasets of COCO-QA and VQA and yields state-of-the-art performance.

****

Optimal Learning for Multi-pass Stochastic Gradient Methods
Junhong Lin, Lorenzo Rosasco

We analyze the learning properties of the stochastic gradient method when multiple passes over the data and mini-batches are allowed. In particular, we consider the square loss and show that   for a universal step-size choice, the number of passes acts as a regularization parameter, and optimal finite sample bounds can be achieved by early-stopping. Moreover, we show that larger step-sizes are allowed when considering mini-batches. Our analysis is based on a unifying approach, encompassing both batch and stochastic gradient methods as special cases.

****

General Tensor Spectral Co-clustering for Higher-Order Data
Tao Wu, Austin R. Benson, David F. Gleich

Spectral clustering and co-clustering are well-known techniques in data analysis, and recent work has extended spectral clustering to square, symmetric tensors and hypermatrices derived from a network. We develop a new tensor spectral co-clustering method that simultaneously clusters the rows, columns, and slices of a nonnegative three-mode tensor and generalizes to tensors with any number of modes. The algorithm is based on a new random walk model which we call the super-spacey random surfer. We show that our method out-performs state-of-the-art co-clustering methods on several synthetic datasets with ground truth clusters and then use the algorithm to analyze several real-world datasets.

****

Balancing Suspense and Surprise: Timely Decision Making with Endogenous Information Acquisition
Ahmed M. Alaa, Mihaela van der Schaar

We develop a Bayesian model for decision-making under time pressure with endogenous information acquisition. In our model, the decision-maker decides when to ob serve (costly) information by sampling an underlying continuous-time stochastic process (time series) that conveys information about the potential occurrence/non-occurrence of an adverse event which will terminate the decision-making process. In her attempt to predict the occurrence of the adverse event, the decision-maker follows a policy that determines when to acquire information from the time series (continuation), and when to stop acquiring information and make a final prediction (stopping). We show that the optimal policy has a "rendezvous" structure, i.e. a structure in which whenever a new information sample is gathered from the time series, the optimal "date" for acquiring the next sample becomes computable. The optimal interval between two information samples balances a trade-off between the decision maker’s "surprise", i.e. the drift in her posterior belief after observing new information, and "suspense", i.e. the probability that the adverse event occurs in the time interval between two information samples. Moreover, we characterize the continuation and stopping regions in the decision-maker’s state-space, and show that they depend not only on the decision-maker’s beliefs, but also on the "context", i.e. the current realization of the time series.

****

Generating Long-term Trajectories Using Deep Hierarchical Networks

Stephan Zheng, Yisong Yue, Jennifer Hobbs

We study the problem of modeling spatiotemporal trajectories over long time horizons using expert demonstrations. For instance, in sports, agents often choose action sequences with long-term goals in mind, such as achieving a certain strategic position. Conventional policy learning approaches, such as those based on Markov decision processes, generally fail at learning cohesive long-term behavior in such high-dimensional state spaces, and are only effective when fairly myopic decision-making yields the desired behavior. The key difficulty is that conventional models are ``single-scale'' and only learn a single state-action policy. We instead propose a hierarchical policy class that automatically reasons about both long-term and short-term goals, which we instantiate as a hierarchical neural network. We showcase our approach in a case study on learning to imitate demonstrated basketball trajectories, and show that it generates significantly more realistic trajectories compared to non-hierarchical baselines as judged by professional sports analysts.

****

Natural-Parameter Networks: A Class of Probabilistic Neural Networks

Hao Wang, Xingjian SHI, Dit-Yan Yeung

Neural networks (NN) have achieved state-of-the-art performance in various applications. Unfortunately in applications where training data is insufficient, they are often prone to overfitting. One effective way to alleviate this problem is to exploit the Bayesian approach by using Bayesian neural networks (BNN). Another shortcoming of NN is the lack of flexibility to customize different distributions for the weights and neurons according to the data, as is often done in probabilistic graphical models. To address these problems, we propose a class of probabilistic neural networks, dubbed natural-parameter networks (NPN), as a novel and lightweight Bayesian treatment of NN. NPN allows the usage of arbitrary exponential-family distributions to model the weights and neurons. Different from traditional NN and BNN, NPN takes distributions as input and goes through layers of transformation before producing distributions to match the target output distributions. As a Bayesian treatment, efficient backpropagation (BP) is performed to learn the natural parameters for the distributions over both the weights and neurons. The output distributions of each layer, as byproducts, may be used as second-order representations for the associated tasks such as link prediction. Experiments on real-world datasets show that NPN can achieve state-of-the-art performance.

****

Minimizing Quadratic Functions in Constant Time

Kohei Hayashi, 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-auth

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