Two-Stream Convolutional Networks for Action Recognition in Videos
Karen Simonyan, Andrew Zisserman
We investigate architectures of discriminatively trained deep Convolutional Netw orks (ConvNets) for action recognition in video. The challenge is to capture the complementary information on appearance from still frames and motion between fr ames. We also aim to generalise the best performing hand-crafted features within a data-driven learning framework. Our contribution is three-fold. First, we pro pose a two-stream ConvNet architecture which incorporates spatial and temporal n etworks. Second, we demonstrate that a ConvNet trained on multi-frame dense opti cal flow is able to achieve very good performance in spite of limited training d ata. Finally, we show that multi-task learning, applied to two different action classification datasets, can be used to increase the amount of training data and improve the performance on both. Our architecture is trained and evaluated on t he standard video actions benchmarks of UCF-101 and HMDB-51, where it is competi tive with the state of the art. It also exceeds by a large margin previous attem pts to use deep nets for video classification.
****

Exploiting easy data in online optimization
Amir Sani, Gergely Neu, Alessandro Lazaric
We consider the problem of online optimization, where a learner chooses a decisi on from a given decision set and suffers some loss associated with the decision and the state of the environment. The learner's objective is to minimize its cum ulative regret against the best fixed decision in hindsight. Over the past few d ecades numerous variants have been considered, with many algorithms designed to achieve sub-linear regret in the worst case. However, this level of robustness c omes at a cost. Proposed algorithms are often over-conservative, failing to adap t to the actual complexity of the loss sequence which is often far from the wors t case. In this paper we introduce a general algorithm that, provided with a saf e learning algorithm and an opportunistic benchmark, can effectively combine goo d worst-case guarantees with much improved performance on easy data. We derive g eneral theoretical bounds on the regret of the proposed algorithm and discuss it s implementation in a wide range of applications, notably in the problem of lear ning with shifting experts (a recent COLT open problem). Finally, we provide num erical simulations in the setting of prediction with expert advice with comparis ons to the state of the art.
****

Optimal prior-dependent neural population codes under shared input noise
Agnieszka Grabska-Barwinska, Jonathan W. Pillow
The brain uses population codes to form distributed, noise-tolerant representati ons of sensory and motor variables. Recent work has examined the theoretical opt imality of such codes in order to gain insight into the principles governing pop ulation codes found in the brain. However, the majority of the population coding literature considers either conditionally independent neurons or neurons with n oise governed by a stimulus-independent covariance matrix. Here we analyze popul ation coding under a simple alternative model in which latent input noise" corru pts the stimulus before it is encoded by the population. This provides a conveni ent and tractable description for irreducible uncertainty that cannot be overcom e by adding neurons, and induces stimulus-dependent correlations that mimic cert ain aspects of the correlations observed in real populations. We examine prior-d ependent, Bayesian optimal coding in such populations using exact analyses of ca ses in which the posterior is approximately Gaussian. These analyses extend prev ious results on independent Poisson population codes and yield an analytic expre ssion for squared loss and a tight upper bound for mutual information. We show t hat, for homogeneous populations that tile the input domain, optimal tuning curv e width depends on the prior, the loss function, the resource constraint, and th e amount of input noise. This framework provides a practical testbed for examini ng issues of optimality, noise, correlation, and coding fidelity in realistic ne ural populations."
****

Quantized Kernel Learning for Feature Matching

Danfeng Qin, Xuanli Chen, Matthieu Guillaumin, Luc V. Gool
Matching local visual features is a crucial problem in computer vision and its a ccuracy greatly depends on the choice of similarity measure. As it is generally very difficult to design by hand a similarity or a kernel perfectly adapted to t he data of interest, learning it automatically with as few assumptions as possib le is preferable. However, available techniques for kernel learning suffer from several limitations, such as restrictive parametrization or scalability. In this paper, we introduce a simple and flexible family of non-linear kernels which we refer to as Quantized Kernels (QK). QKs are arbitrary kernels in the index spac e of a data quantizer, i.e., piecewise constant similarities in the original fea ture space. Quantization allows to compress features and keep the learning tract able. As a result, we obtain state-of-the-art matching performance on a standard benchmark dataset with just a few bits to represent each feature dimension. QKs also have explicit non-linear, low-dimensional feature mappings that grant acce ss to Euclidean geometry for uncompressed features.
****
QUIC & DIRTY: A Quadratic Approximation Approach for Dirty Statistical Models
Cho-Jui Hsieh, Inderjit S. Dhillon, Pradeep K. Ravikumar, Stephen Becker, Peder A. Olsen
In this paper, we develop a family of algorithms for optimizing superposition-st ructured” or “dirty” statistical estimators for high-dimensional problems involv ing the minimization of the sum of a smooth loss function with a hybrid regulari zation. Most of the current approaches are first-order methods, including proxim al gradient or Alternating Direction Method of Multipliers (ADMM). We propose a new family of second-order methods where we approximate the loss function using quadratic approximation. The superposition structured regularizer then leads to a subproblem that can be efficiently solved by alternating minimization. We prop ose a general active subspace selection approach to speed up the solver by utili zing the low-dimensional structure given by the regularizers, and provide conver gence guarantees for our algorithm. Empirically, we show that our approach is mo re than 10 times faster than state-of-the-art first-order approaches for the lat ent variable graphical model selection problems and multi-task learning problems when there is more than one regularizer. For these problems, our approach appea rs to be the first algorithm that can extend active subspace ideas to multiple r egularizers."
****
On a Theory of Nonparametric Pairwise Similarity for Clustering: Connecting Clus tering to Classification
Yingzhen Yang, Feng Liang, Shuicheng Yan, Zhangyang Wang, Thomas S. Huang
Pairwise clustering methods partition the data space into clusters by the pairwi se similarity between data points. The success of pairwise clustering largely de pends on the pairwise similarity function defined over the data points, where ke rnel similarity is broadly used. In this paper, we present a novel pairwise clus tering framework by bridging the gap between clustering and multi-class classifi cation. This pairwise clustering framework learns an unsupervised nonparametric classifier from each data partition, and search for the optimal partition of the data by minimizing the generalization error of the learned classifiers associat ed with the data partitions. We consider two nonparametric classifiers in this f ramework, i.e. the nearest neighbor classifier and the plug-in classifier. Model ing the underlying data distribution by nonparametric kernel density estimation, the generalization error bounds for both unsupervised nonparametric classifiers are the sum of nonparametric pairwise similarity terms between the data points for the purpose of clustering. Under uniform distribution, the nonparametric sim ilarity terms induced by both unsupervised classifiers exhibit a well known form of kernel similarity. We also prove that the generalization error bound for the unsupervised plug-in classifier is asymptotically equal to the weighted volume of cluster boundary for Low Density Separation, a widely used criteria for semi-supervised learning and clustering. Based on the derived nonparametric pairwise similarity using the plug-in classifier, we propose a new nonparametric exemplar -based clustering method with enhanced discriminative capability, whose superior

ity is evidenced by the experimental results.
****
Predictive Entropy Search for Efficient Global Optimization of Black-box Functio
ns
José Miguel Hernández-Lobato, Matthew W. Hoffman, Zoubin Ghahramani
We propose a novel information-theoretic approach for Bayesian optimization call
ed Predictive Entropy Search (PES). At each iteration, PES selects the next eval
uation point that maximizes the expected information gained with respect to the
global maximum. PES codifies this intractable acquisition function in terms of t
he expected reduction in the differential entropy of the predictive distribution
. This reformulation allows PES to obtain approximations that are both more accu
rate and efficient than other alternatives such as Entropy Search (ES). Furtherm
ore, PES can easily perform a fully Bayesian treatment of the model hyperparamet
ers while ES cannot. We evaluate PES in both synthetic and real-world applicatio
ns, including optimization problems in machine learning, finance, biotechnology,
 and robotics. We show that the increased accuracy of PES leads to significant g
ains in optimization performance.
****
Discriminative Unsupervised Feature Learning with Convolutional Neural Networks
Alexey Dosovitskiy, Jost Tobias Springenberg, Martin Riedmiller, Thomas Brox
Current methods for training convolutional neural networks depend on large amoun
ts of labeled samples for supervised training. In this paper we present an appro
ach for training a convolutional neural network using only unlabeled data. We tr
ain the network to discriminate between a set of surrogate classes. Each surroga
te class is formed by applying a variety of transformations to a randomly sample
d 'seed' image patch. We find that this simple feature learning algorithm is sur
prisingly successful when applied to visual object recognition. The feature repr
esentation learned by our algorithm achieves classification results matching or
outperforming the current state-of-the-art for unsupervised learning on several
popular datasets (STL-10, CIFAR-10, Caltech-101).
****
Expectation Backpropagation: Parameter-Free Training of Multilayer Neural Networ
ks with Continuous or Discrete Weights
Daniel Soudry, Itay Hubara, Ron Meir
Multilayer Neural Networks (MNNs) are commonly trained using gradient descent-ba
sed methods, such as BackPropagation (BP). Inference in probabilistic graphical
models is often done using variational Bayes methods, such as Expectation Propag
ation (EP). We show how an EP based approach can also be used to train determini
stic MNNs. Specifically, we approximate the posterior of the weights given the d
ata using a “mean-field” factorized distribution, in an online setting. Using on
line EP and the central limit theorem we find an analytical approximation to the
 Bayes update of this posterior, as well as the resulting Bayes estimates of the
 weights and outputs. Despite a different origin, the resulting algorithm, Expec
tation BackPropagation (EBP), is very similar to BP in form and efficiency. Howe
ver, it has several additional advantages: (1) Training is parameter-free, given
 initial conditions (prior) and the MNN architecture. This is useful for large-s
cale problems, where parameter tuning is a major challenge. (2) The weights can
be restricted to have discrete values. This is especially useful for implementin
g trained MNNs in precision limited hardware chips, thus improving their speed a
nd energy efficiency by several orders of magnitude. We test the EBP algorithm n
umerically in eight binary text classification tasks. In all tasks, EBP outperfo
rms: (1) standard BP with the optimal constant learning rate (2) previously repo
rted state of the art. Interestingly, EBP-trained MNNs with binary weights usual
ly perform better than MNNs with continuous (real) weights - if we average the M
NN output using the inferred posterior.
****
Compressive Sensing of Signals from a GMM with Sparse Precision Matrices
Jianbo Yang, Xuejun Liao, Minhua Chen, Lawrence Carin
This paper is concerned with compressive sensing of signals drawn from a Gaussia
n mixture model (GMM) with sparse precision matrices. Previous work has shown: (

i) a signal drawn from a given GMM can be perfectly reconstructed from r noise-f
ree measurements if the (dominant) rank of each covariance matrix is less than r
; (ii) a sparse Gaussian graphical model can be efficiently estimated from fully
-observed training signals using graphical lasso. This paper addresses a problem
 more challenging than both (i) and (ii), by assuming that the GMM is unknown an
d each signal is only partially observed through incomplete linear measurements.
 Under these challenging assumptions, we develop a hierarchical Bayesian method
to simultaneously estimate the GMM and recover the signals using solely the inco
mplete measurements and a Bayesian shrinkage prior that promotes sparsity of the
 Gaussian precision matrices. In addition, we provide theoretical performance bo
unds to relate the reconstruction error to the number of signals for which measu
rements are available, the sparsity level of precision matrices, and the “incomp
leteness” of measurements. The proposed method is demonstrated extensively on co
mpressive sensing of imagery and video, and the results with simulated and hardw
are-acquired real measurements show significant performance improvement over sta
te-of-the-art methods.
****

Recovery of Coherent Data via Low-Rank Dictionary Pursuit
Guangcan Liu, Ping Li
The recently established RPCA method provides a convenient way to restore low-ra
nk matrices from grossly corrupted observations. While elegant in theory and pow
erful in reality, RPCA is not an ultimate solution to the low-rank matrix recove
ry problem. Indeed, its performance may not be perfect even when data are strict
ly low-rank. This is because RPCA ignores clustering structures of the data whic
h are ubiquitous in applications. As the number of cluster grows, the coherence
of data keeps increasing, and accordingly, the recovery performance of RPCA degr
ades. We show that the challenges raised by coherent data (i.e., data with high
coherence) could be alleviated by Low-Rank Representation (LRR)~\cite{tpami2013l
rr}, provided that the dictionary in LRR is configured appropriately. More preci
sely, we mathematically prove that if the dictionary itself is low-rank then LRR
 is immune to the coherence parameter which increases with the underlying cluste
r number. This provides an elementary principle for dealing with coherent data a
nd naturally leads to a practical algorithm for obtaining proper dictionaries in
 unsupervised environments. Experiments on randomly generated matrices and real
motion sequences verify our claims. See the full paper at arXiv:1404.4032.
****

Learning to Discover Efficient Mathematical Identities
Wojciech Zaremba, Karol Kurach, Rob Fergus
In this paper we explore how machine learning techniques can be applied to the d
iscovery of efficient mathematical identities. We introduce an attribute grammar
 framework for representing symbolic expressions. Given a grammar of math operat
ors, we build trees that combine them in different ways, looking for composition
s that are analytically equivalent to a target expression but of lower computati
onal complexity. However, as the space of trees grows exponentially with the com
plexity of the target expression, brute force search is impractical for all but
the simplest of expressions. Consequently, we introduce two novel learning appro
aches that are able to learn from simpler expressions to guide the tree search.
The first of these is a simple n-gram model, the other being a recursive neural-
network. We show how these approaches enable us to derive complex identities, be
yond reach of brute-force search, or human derivation.
****

Global Sensitivity Analysis for MAP Inference in Graphical Models
Jasper De Bock, Cassio P. de Campos, Alessandro Antonucci
We study the sensitivity of a MAP configuration of a discrete probabilistic grap
hical model with respect to perturbations of its parameters. These perturbations
 are global, in the sense that simultaneous perturbations of all the parameters
(or any chosen subset of them) are allowed. Our main contribution is an exact al
gorithm that can check whether the MAP configuration is robust with respect to g
iven perturbations. Its complexity is essentially the same as that of obtaining
the MAP configuration itself, so it can be promptly used with minimal effort. We

use our algorithm to identify the largest global perturbation that does not induce a change in the MAP configuration, and we successfully apply this robustness measure in two practical scenarios: the prediction of facial action units with posed images and the classification of multiple real public data sets. A strong correlation between the proposed robustness measure and accuracy is verified in both scenarios.

****

Recurrent Models of Visual Attention

Volodymyr Mnih, Nicolas Heess, Alex Graves, koray kavukcuoglu

Applying convolutional neural networks to large images is computationally expensive because the amount of computation scales linearly with the number of image pixels. We present a novel recurrent neural network model that is capable of extracting information from an image or video by adaptively selecting a sequence of regions or locations and only processing the selected regions at high resolution. Like convolutional neural networks, the proposed model has a degree of translation invariance built-in, but the amount of computation it performs can be controlled independently of the input image size. While the model is non-differentiable, it can be trained using reinforcement learning methods to learn task-specific policies. We evaluate our model on several image classification tasks, where it significantly outperforms a convolutional neural network baseline on cluttered images, and on a dynamic visual control problem, where it learns to track a simple object without an explicit training signal for doing so.

****

LSDA: Large Scale Detection through Adaptation

Judy Hoffman, Sergio Guadarrama, Eric S. Tzeng, Ronghang Hu, Jeff Donahue, Ross Girshick, Trevor Darrell, Kate Saenko

A major challenge in scaling object detection is the difficulty of obtaining labeled images for large numbers of categories. Recently, deep convolutional neural networks (CNNs) have emerged as clear winners on object classification benchmarks, in part due to training with 1.2M+ labeled classification images. Unfortunately, only a small fraction of those labels are available for the detection task. It is much cheaper and easier to collect large quantities of image-level labels from search engines than it is to collect detection data and label it with precise bounding boxes. In this paper, we propose Large Scale Detection through Adaptation (LSDA), an algorithm which learns the difference between the two tasks and transfers this knowledge to classifiers for categories without bounding box annotated data, turning them into detectors. Our method has the potential to enable detection for the tens of thousands of categories that lack bounding box annotations, yet have plenty of classification data. Evaluation on the ImageNet LSVRC-2013 detection challenge demonstrates the efficacy of our approach. This algorithm enables us to produce a >7.6K detector by using available classification data from leaf nodes in the ImageNet tree. We additionally demonstrate how to modify our architecture to produce a fast detector (running at 2fps for the 7.6K detector). Models and software are available at

****

Analog Memories in a Balanced Rate-Based Network of E-I Neurons

Dylan Festa, Guillaume Hennequin, Mate Lengyel

The persistent and graded activity often observed in cortical circuits is sometimes seen as a signature of autoassociative retrieval of memories stored earlier in synaptic efficacies. However, despite decades of theoretical work on the subject, the mechanisms that support the storage and retrieval of memories remain unclear. Previous proposals concerning the dynamics of memory networks have fallen short of incorporating some key physiological constraints in a unified way. Specifically, some models violate Dale's law (i.e. allow neurons to be both excitatory and inhibitory), while some others restrict the representation of memories to a binary format, or induce recall states in which some neurons fire at rates close to saturation. We propose a novel control-theoretic framework to build functioning attractor networks that satisfy a set of relevant physiological constraints. We directly optimize networks of excitatory and inhibitory neurons to force sets of arbitrary analog patterns to become stable fixed points of the dynamics

. The resulting networks operate in the balanced regime, are robust to corruptio
ns of the memory cue as well as to ongoing noise, and incidentally explain the r
eduction of trial-to-trial variability following stimulus onset that is ubiquito
usly observed in sensory and motor cortices. Our results constitute a step forwa
rd in our understanding of the neural substrate of memory.
****

Efficient Partial Monitoring with Prior Information
Hastagiri P. Vanchinathan, Gábor Bartók, Andreas Krause
Partial monitoring is a general model for online learning with limited feedback:
 a learner chooses actions in a sequential manner while an opponent chooses outc
omes. In every round, the learner suffers some loss and receives some feedback b
ased on the action and the outcome. The goal of the learner is to minimize her c
umulative loss. Applications range from dynamic pricing to label-efficient predi
ction to dueling bandits. In this paper, we assume that we are given some prior
information about the distribution based on which the opponent generates the out
comes. We propose BPM, a family of new efficient algorithms whose core is to tra
ck the outcome distribution with an ellipsoid centered around the estimated dist
ribution. We show that our algorithm provably enjoys near-optimal regret rate fo
r locally observable partial-monitoring problems against stochastic opponents. A
s demonstrated with experiments on synthetic as well as real-world data, the alg
orithm outperforms previous approaches, even for very uninformed priors, with an
 order of magnitude smaller regret and lower running time.
****

Near-optimal Reinforcement Learning in Factored MDPs
Ian Osband, Benjamin Van Roy
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.
****

Robust Kernel Density Estimation by Scaling and Projection in Hilbert Space
Robert A. Vandermeulen, Clayton Scott
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.
****

Extracting Certainty from Uncertainty: Transductive Pairwise Classification from
 Pairwise Similarities
Tianbao Yang, Rong Jin
In this work, we study the problem of transductive pairwise classification from
pairwise similarities~\footnote{The pairwise similarities are usually derived fr
om some side information instead of the underlying class labels.}. The goal of t
ransductive pairwise classification from pairwise similarities is to infer the p
airwise class relationships, to which we refer as pairwise labels, between all e
xamples given a subset of class relationships for a small set of examples, to wh
ich we refer as labeled examples. We propose a very simple yet effective algorit
hm that consists of two simple steps: the first step is to complete the sub-matr
ix corresponding to the labeled examples and the second step is to reconstruct t
he label matrix from the completed sub-matrix and the provided similarity matrix
. Our analysis exhibits that under several mild preconditions we can recover the
 label matrix with a small error, if the top eigen-space that corresponds to the
 largest eigenvalues of the similarity matrix covers well the column space of la
bel matrix and is subject to a low coherence, and the number of observed pairwis
e labels is sufficiently enough. We demonstrate the effectiveness of the propose
d algorithm by several experiments.
****

Diverse Sequential Subset Selection for Supervised Video Summarization
Boqing Gong, Wei-Lun Chao, Kristen Grauman, Fei Sha
Video summarization is a challenging problem with great application potential. W

hereas prior approaches, largely unsupervised in nature, focus on sampling usefu
l frames and assembling them as summaries, we consider video summarization as a
supervised subset selection problem. Our idea is to teach the system to learn fr
om human-created summaries how to select informative and diverse subsets, so as
to best meet evaluation metrics derived from human-perceived quality. To this en
d, we propose the sequential determinantal point process (seqDPP), a probabilist
ic model for diverse sequential subset selection. Our novel seqDPP heeds the inh
erent sequential structures in video data, thus overcoming the deficiency of the
 standard DPP, which treats video frames as randomly permutable items. Meanwhile
, seqDPP retains the power of modeling diverse subsets, essential for summarizat
ion. Our extensive results of summarizing videos from 3 datasets demonstrate the
 superior performance of our method, compared to not only existing unsupervised
methods but also naive applications of the standard DPP model.
****

Simultaneous Model Selection and Optimization through Parameter-free Stochastic
Learning
Francesco Orabona
Stochastic gradient descent algorithms for training linear and kernel predictors
 are gaining more and more importance, thanks to their scalability. While variou
s methods have been proposed to speed up their convergence, the model selection
phase is often ignored. In fact, in theoretical works most of the time assumptio
ns are made, for example, on the prior knowledge of the norm of the optimal solu
tion, while in the practical world validation methods remain the only viable app
roach. In this paper, we propose a new kernel-based stochastic gradient descent
algorithm that performs model selection while training, with no parameters to tu
ne, nor any form of cross-validation. The algorithm builds on recent advancement
 in online learning theory for unconstrained settings, to estimate over time the
 right regularization in a data-dependent way. Optimal rates of convergence are
proved under standard smoothness assumptions on the target function as well as p
reliminary empirical results.
****

On the Number of Linear Regions of Deep Neural Networks
Guido F. Montufar, Razvan Pascanu, Kyunghyun Cho, Yoshua Bengio
We study the complexity of functions computable by deep feedforward neural netwo
rks with piecewise linear activations in terms of the symmetries and the number
of linear regions that they have. Deep networks are able to sequentially map por
tions of each layer's input-space to the same output. In this way, deep models c
ompute functions that react equally to complicated patterns of different inputs.
 The compositional structure of these functions enables them to re-use pieces of
 computation exponentially often in terms of the network's depth. This paper inv
estigates the complexity of such compositional maps and contributes new theoreti
cal results regarding the advantage of depth for neural networks with piecewise
linear activation functions. In particular, our analysis is not specific to a si
ngle family of models, and as an example, we employ it for rectifier and maxout
networks. We improve complexity bounds from pre-existing work and investigate th
e behavior of units in higher layers.
****

Model-based Reinforcement Learning and the Eluder Dimension
Ian Osband, Benjamin Van Roy
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.
****

A Wild Bootstrap for Degenerate Kernel Tests
Kacper P. Chwialkowski, Dino Sejdinovic, Arthur Gretton
A wild bootstrap method for nonparametric hypothesis tests based on kernel distr
ibution embeddings is proposed. This bootstrap method is used to construct prova
bly consistent tests that apply to random processes, for which the naive permuta
tion-based bootstrap fails. It applies to a large group of kernel tests based on

V-statistics, which are degenerate under the null hypothesis, and non-degenerate elsewhere. To illustrate this approach, we construct a two-sample test, an instantaneous independence test and a multiple lag independence test for time series. In experiments, the wild bootstrap gives strong performance on synthetic examples, on audio data, and in performance benchmarking for the Gibbs sampler. The code is available at https://github.com/kacperChwialkowski/wildBootstrap.
****

Extracting Latent Structure From Multiple Interacting Neural Populations
Joao Semedo, Amin Zandvakili, Adam Kohn, Christian K. Machens, Byron M. Yu
Developments in neural recording technology are rapidly enabling the recording of populations of neurons in multiple brain areas simultaneously, as well as the identification of the types of neurons being recorded (e.g., excitatory vs. inhibitory). There is a growing need for statistical methods to study the interaction among multiple, labeled populations of neurons. Rather than attempting to identify direct interactions between neurons (where the number of interactions grows with the number of neurons squared), we propose to extract a smaller number of latent variables from each population and study how the latent variables interact. Specifically, we propose extensions to probabilistic canonical correlation analysis (pCCA) to capture the temporal structure of the latent variables, as well as to distinguish within-population dynamics from across-population interactions (termed Group Latent Auto-Regressive Analysis, gLARA). We then applied these methods to populations of neurons recorded simultaneously in visual areas V1 and V2, and found that gLARA provides a better description of the recordings than pCCA. This work provides a foundation for studying how multiple populations of neurons interact and how this interaction supports brain function.
****

Variational Gaussian Process State-Space Models
Roger Frigola, Yutian Chen, Carl Edward Rasmussen
State-space models have been successfully used for more than fifty years in different areas of science and engineering. We present a procedure for efficient variational Bayesian learning of nonlinear state-space models based on sparse Gaussian processes. The result of learning is a tractable posterior over nonlinear dynamical systems. In comparison to conventional parametric models, we offer the possibility to straightforwardly trade off model capacity and computational cost whilst avoiding overfitting. Our main algorithm uses a hybrid inference approach combining variational Bayes and sequential Monte Carlo. We also present stochastic variational inference and online learning approaches for fast learning with long time series.
****

Multi-View Perceptron: a Deep Model for Learning Face Identity and View Representations
Zhenyao Zhu, Ping Luo, Xiaogang Wang, Xiaoou Tang
Various factors, such as identities, views (poses), and illuminations, are coupled in face images. Disentangling the identity and view representations is a major challenge in face recognition. Existing face recognition systems either use handcrafted features or learn features discriminatively to improve recognition accuracy. This is different from the behavior of human brain. Intriguingly, even without accessing 3D data, human not only can recognize face identity, but can also imagine face images of a person under different viewpoints given a single 2D image, making face perception in the brain robust to view changes. In this sense, human brain has learned and encoded 3D face models from 2D images. To take into account this instinct, this paper proposes a novel deep neural net, named multi-view perceptron (MVP), which can untangle the identity and view features, and infer a full spectrum of multi-view images in the meanwhile, given a single 2D face image. The identity features of MVP achieve superior performance on the Multi PIE dataset. MVP is also capable to interpolate and predict images under viewpoints that are unobserved in the training data.
****

Global Belief Recursive Neural Networks
Romain Paulus, Richard Socher, Christopher D. Manning

Recursive Neural Networks have recently obtained state of the art performance on several natural language processing tasks. However, because of their feedforward architecture they cannot correctly predict phrase or word labels that are determined by context. This is a problem in tasks such as aspect-specific sentiment classification which tries to, for instance, predict that the word Android is positive in the sentence Android beats iOS. We introduce global belief recursive neural networks (GB-RNNs) which are based on the idea of extending purely feedforward neural networks to include one feedbackward step during inference. This allows phrase level predictions and representations to give feedback to words. We show the effectiveness of this model on the task of contextual sentiment analysis. We also show that dropout can improve RNN training and that a combination of unsupervised and supervised word vector representations performs better than either alone. The feedbackward step improves F1 performance by 3% over the standard RNN on this task, obtains state-of-the-art performance on the SemEval 2013 challenge and can accurately predict the sentiment of specific entities.

****

Parallel Sampling of HDPs using Sub-Cluster Splits

Jason Chang, John W. Fisher III

We develop a sampling technique for Hierarchical Dirichlet process models. The parallel algorithm builds upon [Chang & Fisher 2013] by proposing large split and merge moves based on learned sub-clusters. The additional global split and merge moves drastically improve convergence in the experimental results. Furthermore, we discover that cross-validation techniques do not adequately determine convergence, and that previous sampling methods converge slower than were previously expected.

****

Recursive Inversion Models for Permutations

Christopher Meek, Marina Meila

We develop a new exponential family probabilistic model for permutations that can capture hierarchical structure, and that has the well known Mallows and generalized Mallows models as subclasses. We describe how one can do parameter estimation and propose an approach to structure search for this class of models. We provide experimental evidence that this added flexibility both improves predictive performance and enables a deeper understanding of collections of permutations.

****

Active Learning and Best-Response Dynamics

Maria-Florina F. Balcan, Christopher Berlind, Avrim Blum, Emma Cohen, Kaushik Patnaik, Le Song

We consider a setting in which low-power distributed sensors are each making highly noisy measurements of some unknown target function. A center wants to accurately learn this function by querying a small number of sensors, which ordinarily would be impossible due to the high noise rate. The question we address is whether local communication among sensors, together with natural best-response dynamics in an appropriately-de■ned game, can denoise the system without destroying the true signal and allow the center to succeed from only a small number of active queries. We prove positive (and negative) results on the denoising power of several natural dynamics, and also show experimentally that when combined with recent agnostic active learning algorithms, this process can achieve low error from very few queries, performing substantially better than active or passive learning without these denoising dynamics as well as passive learning with denoising.

****

Multilabel Structured Output Learning with Random Spanning Trees of Max-Margin Markov Networks

Mario Marchand, Hongyu Su, Emilie Morvant, Juho Rousu, John S. Shawe-Taylor

We show that the usual score function for conditional Markov networks can be written as the expectation over the scores of their spanning trees. We also show that a small random sample of these output trees can attain a significant fraction of the margin obtained by the complete graph and we provide conditions under which we can perform tractable inference. The experimental results confirm that practical learning is scalable to realistic datasets using this approach.

****
Identifying and attacking the saddle point problem in high-dimensional non-convex optimization
Yann N. Dauphin, Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, Surya Ganguli, Yoshua Bengio
A central challenge to many fields of science and engineering involves minimizing non-convex error functions over continuous, high dimensional spaces. Gradient descent or quasi-Newton methods are almost ubiquitously used to perform such minimizations, and it is often thought that a main source of difficulty for these local methods to find the global minimum is the proliferation of local minima with much higher error than the global minimum. Here we argue, based on results from statistical physics, random matrix theory, neural network theory, and empirical evidence, that a deeper and more profound difficulty originates from the proliferation of saddle points, not local minima, especially in high dimensional problems of practical interest. Such saddle points are surrounded by high error plateaus that can dramatically slow down learning, and give the illusory impression of the existence of a local minimum. Motivated by these arguments, we propose a new approach to second-order optimization, the saddle-free Newton method, that can rapidly escape high dimensional saddle points, unlike gradient descent and quasi-Newton methods. We apply this algorithm to deep or recurrent neural network training, and provide numerical evidence for its superior optimization performance.
****
PEWA: Patch-based Exponentially Weighted Aggregation for image denoising
Charles Kervrann
Patch-based methods have been widely used for noise reduction in recent years. In this paper, we propose a general statistical aggregation method which combines image patches denoised with several commonly-used algorithms. We show that weakly denoised versions of the input image obtained with standard methods, can serve to compute an efficient patch-based aggregated estimd aggregation (EWA) estimator. The resulting approach (PEWA) is based on a MCMC sampling and has a nice statistical foundation while producing denoising results that are comparable to the current state-of-the-art. We demonstrate the performance of the denoising algorithm on real images and we compare the results to several competitive methods.
****
A Safe Screening Rule for Sparse Logistic Regression
Jie Wang, Jiayu Zhou, Jun Liu, Peter Wonka, Jieping Ye
The l1-regularized logistic regression (or sparse logistic regression) is a widely used method for simultaneous classification and feature selection. Although many recent efforts have been devoted to its efficient implementation, its application to high dimensional data still poses significant challenges. In this paper, we present a fast and effective sparse logistic regression screening rule (Slores) to identify the zero components in the solution vector, which may lead to a substantial reduction in the number of features to be entered to the optimization. An appealing feature of Slores is that the data set needs to be scanned only once to run the screening and its computational cost is negligible compared to that of solving the sparse logistic regression problem. Moreover, Slores is independent of solvers for sparse logistic regression, thus Slores can be integrated with any existing solver to improve the efficiency. We have evaluated Slores using high-dimensional data sets from different applications. Extensive experimental results demonstrate that Slores outperforms the existing state-of-the-art screening rules and the efficiency of solving sparse logistic regression is improved by one magnitude in general.
****
Weakly-supervised Discovery of Visual Pattern Configurations
Hyun Oh Song, Yong Jae Lee, Stefanie Jegelka, Trevor Darrell
The prominence of weakly labeled data gives rise to a growing demand for object detection methods that can cope with minimal supervision. We propose an approach that automatically identifies discriminative configurations of visual patterns that are characteristic of a given object class. We formulate the problem as a c

onstrained submodular optimization problem and demonstrate the benefits of the d
iscovered configurations in remedying mislocalizations and finding informative p
ositive and negative training examples. Together, these lead to state-of-the-art
 weakly-supervised detection results on the challenging PASCAL VOC dataset.
****

Deep Networks with Internal Selective Attention through Feedback Connections
Marijn F. Stollenga, Jonathan Masci, Faustino Gomez, Jürgen Schmidhuber
Traditional convolutional neural networks (CNN) are stationary and feedforward.
They neither change their parameters during evaluation nor use feedback from hig
her to lower layers. Real brains, however, do. So does our Deep Attention Select
ive Network (dasNet) architecture. DasNet's feedback structure can dynamically a
lter its convolutional filter sensitivities during classification. It harnesses
the power of sequential processing to improve classification performance, by all
owing the network to iteratively focus its internal attention on some of its con
volutional filters. Feedback is trained through direct policy search in a huge m
illion-dimensional parameter space, through scalable natural evolution strategie
s (SNES). On the CIFAR-10 and CIFAR-100 datasets, dasNet outperforms the previou
s state-of-the-art model on unaugmented datasets.
****

Tight Bounds for Influence in Diffusion Networks and Application to Bond Percola
tion and Epidemiology
Remi Lemonnier, Kevin Scaman, Nicolas Vayatis
In this paper, we derive theoretical bounds for the long-term influence of a nod
e in an Independent Cascade Model (ICM). We relate these bounds to the spectral
radius of a particular matrix and show that the behavior is sub-critical when th
is spectral radius is lower than 1. More specifically, we point out that, in gen
eral networks, the sub-critical regime behaves in O(sqrt(n)) where n is the size
 of the network, and that this upper bound is met for star-shaped networks. We a
pply our results to epidemiology and percolation on arbitrary networks, and deri
ve a bound for the critical value beyond which a giant connected component arise
s. Finally, we show empirically the tightness of our bounds for a large family o
f networks.
****

A Bayesian model for identifying hierarchically organised states in neural popul
ation activity
Patrick Putzky, Florian Franzen, Giacomo Bassetto, Jakob H. Macke
Neural population activity in cortical circuits is not solely driven by external
 inputs, but is also modulated by endogenous states which vary on multiple time-
scales. To understand information processing in cortical circuits, we need to un
derstand the statistical structure of internal states and their interaction with
 sensory inputs. Here, we present a statistical model for extracting hierarchica
lly organised neural population states from multi-channel recordings of neural s
piking activity. Population states are modelled using a hidden Markov decision t
ree with state-dependent tuning parameters and a generalised linear observation
model. We present a variational Bayesian inference algorithm for estimating the
posterior distribution over parameters from neural population recordings. On sim
ulated data, we show that we can identify the underlying sequence of population
states and reconstruct the ground truth parameters. Using population recordings
from visual cortex, we find that a model with two levels of population states ou
tperforms both a one-state and a two-state generalised linear model. Finally, we
 find that modelling of state-dependence also improves the accuracy with which s
ensory stimuli can be decoded from the population response.
****

Deep Convolutional Neural Network for Image Deconvolution
Li Xu, Jimmy SJ Ren, Ce Liu, Jiaya Jia
Many fundamental image-related problems involve deconvolution operators. Real bl
ur degradation seldom complies with an deal linear convolution model due to came
ra noise, saturation, image compression, to name a few. Instead of perfectly mod
eling outliers, which is rather challenging from a generative model perspective,
 we develop a deep convolutional neural network to capture the characteristics o

f degradation. We note directly applying existing deep neural networks does not produce reasonable results. Our solution is to establish the connection between traditional optimization-based schemes and a neural network architecture where a novel, separable structure is introduced as a reliable support for robust deconvolution against artifacts. Our network contains two submodules, both trained in a supervised manner with proper initialization. They yield decent performance on non-blind image deconvolution compared to previous generative-model based methods.
****

Attentional Neural Network: Feature Selection Using Cognitive Feedback
Qian Wang, Jiaxing Zhang, Sen Song, Zheng Zhang
Attentional Neural Network is a new framework that integrates top-down cognitive bias and bottom-up feature extraction in one coherent architecture. The top-down influence is especially effective when dealing with high noise or difficult segmentation problems. Our system is modular and extensible. It is also easy to train and cheap to run, and yet can accommodate complex behaviors. We obtain classification accuracy better than or competitive with state of art results on the MNIST variation dataset, and successfully disentangle overlaid digits with high success rates. We view such a general purpose framework as an essential foundation for a larger system emulating the cognitive abilities of the whole brain.
****

The Blinded Bandit: Learning with Adaptive Feedback
Ofer Dekel, Elad Hazan, Tomer Koren
We study an online learning setting where the player is temporarily deprived of feedback each time it switches to a different action. Such model of \emph{adaptive feedback} naturally occurs in scenarios where the environment reacts to the player's actions and requires some time to recover and stabilize after the algorithm switches actions. This motivates a variant of the multi-armed bandit problem, which we call the \emph{blinded multi-armed bandit}, in which no feedback is given to the algorithm whenever it switches arms. We develop efficient online learning algorithms for this problem and prove that they guarantee the same asymptotic regret as the optimal algorithms for the standard multi-armed bandit problem. This result stands in stark contrast to another recent result, which states that adding a switching cost to the standard multi-armed bandit makes it substantially harder to learn, and provides a direct comparison of how feedback and loss contribute to the difficulty of an online learning problem. We also extend our results to the general prediction framework of bandit linear optimization, again attaining near-optimal regret bounds.
****

Zero-shot recognition with unreliable attributes
Dinesh Jayaraman, Kristen Grauman
In principle, zero-shot learning makes it possible to train an object recognition model simply by specifying the category's attributes. For example, with classifiers for generic attributes like striped and four-legged, one can construct a classifier for the zebra category by enumerating which properties it possesses --- even without providing zebra training images. In practice, however, the standard zero-shot paradigm suffers because attribute predictions in novel images are hard to get right. We propose a novel random forest approach to train zero-shot models that explicitly accounts for the unreliability of attribute predictions. By leveraging statistics about each attribute’s error tendencies, our method obtains more robust discriminative models for the unseen classes. We further devise extensions to handle the few-shot scenario and unreliable attribute descriptions. On three datasets, we demonstrate the benefit for visual category learning with zero or few training examples, a critical domain for rare categories or categories defined on the fly.
****

Communication Efficient Distributed Machine Learning with the Parameter Server
Mu Li, David G. Andersen, Alexander J. Smola, Kai Yu
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.
****

Beyond the Birkhoff Polytope: Convex Relaxations for Vector Permutation Problems
Cong Han Lim, Stephen Wright

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.
****

Deconvolution of High Dimensional Mixtures via Boosting, with Application to Dif
fusion-Weighted MRI of Human Brain
Charles Y. Zheng, Franco Pestilli, Ariel Rokem

Diffusion-weighted magnetic resonance imaging (DWI) and fiber tractography are t
he only methods to measure the structure of the white matter in the living human
brain. The diffusion signal has been modelled as the combined contribution from
many individual fascicles of nerve fibers passing through each location in the
white matter. Typically, this is done via basis pursuit, but estimation of the e
xact directions is limited due to discretization. The difficulties inherent in m
odeling DWI data are shared by many other problems involving fitting non-paramet
ric mixture models. Ekanadaham et al. proposed an approach, continuous basis pur
suit, to overcome discretization error in the 1-dimensional case (e.g., spike-so
rting). Here, we propose a more general algorithm that fits mixture models of an
y dimensionality without discretization. Our algorithm uses the principles of L2
-boost, together with refitting of the weights and pruning of the parameters. Th
e addition of these steps to L2-boost both accelerates the algorithm and assures
its accuracy. We refer to the resulting algorithm as elastic basis pursuit, or
EBP, since it expands and contracts the active set of kernels as needed. We show
that in contrast to existing approaches to fitting mixtures, our boosting frame
work (1) enables the selection of the optimal bias-variance tradeoff along the s
olution path, and (2) scales with high-dimensional problems. In simulations of D
WI, we find that EBP yields better parameter estimates than a non-negative least
squares (NNLS) approach, or the standard model used in DWI, the tensor model, w
hich serves as the basis for diffusion tensor imaging (DTI). We demonstrate the
utility of the method in DWI data acquired in parts of the brain containing cros
sings of multiple fascicles of nerve fibers.
****

On Iterative Hard Thresholding Methods for High-dimensional M-Estimation
Prateek Jain, Ambuj Tewari, Purushottam Kar

The use of M-estimators in generalized linear regression models in high dimensio
nal settings requires risk minimization with hard L_0 constraints. Of the known
methods, the class of projected gradient descent (also known as iterative hard t
hresholding (IHT)) methods is known to offer the fastest and most scalable solut
ions. However, the current state-of-the-art is only able to analyze these method
s in extremely restrictive settings which do not hold in high dimensional statis
tical models. In this work we bridge this gap by providing the first analysis fo
r IHT-style methods in the high dimensional statistical setting. Our bounds are
tight and match known minimax lower bounds. Our results rely on a general analys
is framework that enables us to analyze several popular hard thresholding style
algorithms (such as HTP, CoSaMP, SP) in the high dimensional regression setting.
Finally, we extend our analysis to the problem of low-rank matrix recovery.
****

Bayesian Sampling Using Stochastic Gradient Thermostats
Nan Ding, Youhan Fang, Ryan Babbush, Changyou Chen, Robert D. Skeel, Hartmut Nev
en

Dynamics-based sampling methods, such as Hybrid Monte Carlo (HMC) and Langevin d
ynamics (LD), are commonly used to sample target distributions. Recently, such a
pproaches have been combined with stochastic gradient techniques to increase sam
pling efficiency when dealing with large datasets. An outstanding problem with t
his approach is that the stochastic gradient introduces an unknown amount of noi

se which can prevent proper sampling after discretization. To remedy this proble m, we show that one can leverage a small number of additional variables in order to stabilize momentum fluctuations induced by the unknown noise. Our method is inspired by the idea of a thermostat in statistical physics and is justified by a general theory.
****

Encoding High Dimensional Local Features by Sparse Coding Based Fisher Vectors
Lingqiao Liu, Chunhua Shen, Lei Wang, Anton van den Hengel, Chao Wang
Deriving from the gradient vector of a generative model of local features, Fishe r vector coding (FVC) has been identified as an effective coding method for imag e classification. Most, if not all, FVC implementations employ the Gaussian mixt ure model (GMM) to characterize the generation process of local features. This c hoice has shown to be sufficient for traditional low dimensional local features, e.g., SIFT; and typically, good performance can be achieved with only a few hun dred Gaussian distributions. However, the same number of Gaussians is insufficie nt to model the feature space spanned by higher dimensional local features, whic h have become popular recently. In order to improve the modeling capacity for hi gh dimensional features, it turns out to be inefficient and computationally impr actical to simply increase the number of Gaussians. In this paper, we propose a model in which each local feature is drawn from a Gaussian distribution whose me an vector is sampled from a subspace. With certain approximation, this model can be converted to a sparse coding procedure and the learning/inference problems c an be readily solved by standard sparse coding methods. By calculating the gradi ent vector of the proposed model, we derive a new fisher vector encoding strateg y, termed Sparse Coding based Fisher Vector Coding (SCFVC). Moreover, we adopt t he recently developed Deep Convolutional Neural Network (CNN) descriptor as a hi gh dimensional local feature and implement image classification with the propose d SCFVC. Our experimental evaluations demonstrate that our method not only signi ficantly outperforms the traditional GMM based Fisher vector encoding but also a chieves the state-of-the-art performance in generic object recognition, indoor s cene, and fine-grained image classification problems.
****

Efficient learning by implicit exploration in bandit problems with side observat ions
Tomáš Kocák, Gergely Neu, Michal Valko, Remi Munos
We consider online learning problems under a a partial observability model captu ring situations where the information conveyed to the learner is between full in formation and bandit feedback. In the simplest variant, we assume that in additi on to its own loss, the learner also gets to observe losses of some other action s. The revealed losses depend on the learner's action and a directed observation system chosen by the environment. For this setting, we propose the first algori thm that enjoys near-optimal regret guarantees without having to know the observ ation system before selecting its actions. Along similar lines, we also define a new partial information setting that models online combinatorial optimization p roblems where the feedback received by the learner is between semi-bandit and fu ll feedback. As the predictions of our first algorithm cannot be always computed efficiently in this setting, we propose another algorithm with similar properti es and with the benefit of always being computationally efficient, at the price of a slightly more complicated tuning mechanism. Both algorithms rely on a novel exploration strategy called implicit exploration, which is shown to be more eff icient both computationally and information-theoretically than previously studie d exploration strategies for the problem.
****

An Integer Polynomial Programming Based Framework for Lifted MAP Inference
Somdeb Sarkhel, Deepak Venugopal, Parag Singla, Vibhav G. Gogate
In this paper, we present a new approach for lifted MAP inference in Markov logi c networks (MLNs). The key idea in our approach is to compactly encode the MAP i nference problem as an Integer Polynomial Program (IPP) by schematically applyin g three lifted inference steps to the MLN: lifted decomposition, lifted conditio ning, and partial grounding. Our IPP encoding is lifted in the sense that an int

eger assignment to a variable in the IPP may represent a truth-assignment to mul
tiple indistinguishable ground atoms in the MLN. We show how to solve the IPP by
 first converting it to an Integer Linear Program (ILP) and then solving the lat
ter using state-of-the-art ILP techniques. Experiments on several benchmark MLNs
 show that our new algorithm is substantially superior to ground inference and e
xisting methods in terms of computational efficiency and solution quality.
****

Hardness of parameter estimation in graphical models
Guy Bresler, David Gamarnik, Devavrat Shah
We consider the problem of learning the canonical parameters specifying an undir
ected graphical model (Markov random field) from the mean parameters. For graphi
cal models representing a minimal exponential family, the canonical parameters a
re uniquely determined by the mean parameters, so the problem is feasible in pri
nciple. The goal of this paper is to investigate the computational feasibility o
f this statistical task. Our main result shows that parameter estimation is in g
eneral intractable: no algorithm can learn the canonical parameters of a generic
 pair-wise binary graphical model from the mean parameters in time bounded by a
polynomial in the number of variables (unless RP = NP). Indeed, such a result ha
s been believed to be true (see the monograph by Wainwright and Jordan) but no p
roof was known. Our proof gives a polynomial time reduction from approximating t
he partition function of the hard-core model, known to be hard, to learning appr
oximate parameters. Our reduction entails showing that the marginal polytope bou
ndary has an inherent repulsive property, which validates an optimization proced
ure over the polytope that does not use any knowledge of its structure (as requi
red by the ellipsoid method and others).
****

On the Information Theoretic Limits of Learning Ising Models
Rashish Tandon, Karthikeyan Shanmugam, Pradeep K. Ravikumar, Alexandros G. Dimak
is
We provide a general framework for computing lower-bounds on the sample complexi
ty of recovering the underlying graphs of Ising models, given i.i.d. samples. Wh
ile there have been recent results for specific graph classes, these involve fai
rly extensive technical arguments that are specialized to each specific graph cl
ass. In contrast, we isolate two key graph-structural ingredients that can then
be used to specify sample complexity lower-bounds. Presence of these structural
properties makes the graph class hard to learn. We derive corollaries of our mai
n result that not only recover existing recent results, but also provide lower b
ounds for novel graph classes not considered previously. We also extend our fram
ework to the random graph setting and derive corollaries for Erdos-Renyi graphs
in a certain dense setting.
****

Projecting Markov Random Field Parameters for Fast Mixing
Xianghang Liu, Justin Domke
Markov chain Monte Carlo (MCMC) algorithms are simple and extremely powerful tec
hniques to sample from almost arbitrary distributions. The flaw in practice is t
hat it can take a large and/or unknown amount of time to converge to the station
ary distribution. This paper gives sufficient conditions to guarantee that univa
riate Gibbs sampling on Markov Random Fields (MRFs) will be fast mixing, in a pr
ecise sense. Further, an algorithm is given to project onto this set of fast-mix
ing parameters in the Euclidean norm. Following recent work, we give an example
use of this to project in various divergence measures, comparing of univariate m
arginals obtained by sampling after projection to common variational methods and
 Gibbs sampling on the original parameters.
****

A Boosting Framework on Grounds of Online Learning
Tofigh Naghibi Mohamadpoor, Beat Pfister
By exploiting the duality between boosting and online learning, we present a boo
sting framework which proves to be extremely powerful thanks to employing the va
st knowledge available in the online learning area. Using this framework, we dev
elop various algorithms to address multiple practically and theoretically intere

sting questions including sparse boosting, smooth-distribution boosting, agnosti c learning and, as a by-product, some generalization to double-projection online learning algorithms.
****

Near-Optimal Density Estimation in Near-Linear Time Using Variable-Width Histogr ams
Siu On Chan, Ilias Diakonikolas, Rocco A. Servedio, Xiaorui Sun
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.
****

Efficient Minimax Signal Detection on Graphs
Jing Qian, Venkatesh Saligrama
Several problems such as network intrusion, community detection, and disease out break can be described by observations attributed to nodes or edges of a graph. In these applications presence of intrusion, community or disease outbreak is ch aracterized by novel observations on some unknown connected subgraph. These prob lems can be formulated in terms of optimization of suitable objectives on connec ted subgraphs, a problem which is generally computationally difficult. We overco me the combinatorics of connectivity by embedding connected subgraphs into linea r matrix inequalities (LMI). Computationally efficient tests are then realized b y optimizing convex objective functions subject to these LMI constraints. We pro ve, by means of a novel Euclidean embedding argument, that our tests are minimax optimal for exponential family of distributions on 1-D and 2-D lattices. We sho w that internal conductance of the connected subgraph family plays a fundamental role in characterizing detectability.
****

Magnitude-sensitive preference formation`
Nisheeth Srivastava, Ed Vul, Paul R. Schrater
Our understanding of the neural computations that underlie the ability of animal s to choose among options has advanced through a synthesis of computational mode ling, brain imaging and behavioral choice experiments. Yet, there remains a gulf between theories of preference learning and accounts of the real, economic choi ces that humans face in daily life, choices that are usually between some amount of money and an item. In this paper, we develop a theory of magnitude-sensitive preference learning that permits an agent to rationally infer its preferences f or items compared with money options of different magnitudes. We show how this t heory yields classical and anomalous supply-demand curves and predicts choices f or a large panel of risky lotteries. Accurate replications of such phenomena wit hout recourse to utility functions suggest that the theory proposed is both psyc hologically realistic and econometrically viable.
****

Accelerated Mini-batch Randomized Block Coordinate Descent Method
Tuo Zhao, Mo Yu, Yiming Wang, Raman Arora, Han Liu
We consider regularized empirical risk minimization problems. In particular, we minimize the sum of a smooth empirical risk function and a nonsmooth regularizat ion function. When the regularization function is block separable, we can solve the minimization problems in a randomized block coordinate descent (RBCD) manner . Existing RBCD methods usually decrease the objective value by exploiting the p artial gradient of a randomly selected block of coordinates in each iteration. T hus they need all data to be accessible so that the partial gradient of the bloc k gradient can be exactly obtained. However, such a ``batch setting may be compu tationally expensive in practice. In this paper, we propose a mini-batch randomi zed block coordinate descent (MRBCD) method, which estimates the partial gradien t of the selected block based on a mini-batch of randomly sampled data in each i teration. We further accelerate the MRBCD method by exploiting the semi-stochast ic optimization scheme, which effectively reduces the variance of the partial gr adient estimators. Theoretically, we show that for strongly convex functions, th e MRBCD method attains lower overall iteration complexity than existing RBCD met

hods. As an application, we further trim the MRBCD method to solve the regulariz
ed sparse learning problems. Our numerical experiments shows that the MRBCD meth
od naturally exploits the sparsity structure and achieves better computational p
erformance than existing methods."
****

Multiscale Fields of Patterns
Pedro Felzenszwalb, John G. Oberlin
We describe a framework for defining high-order image models that can be used in
 a variety of applications. The approach involves modeling local patterns in a m
ultiscale representation of an image. Local properties of a coarsened image refl
ect non-local properties of the original image. In the case of binary images loc
al properties are defined by the binary patterns observed over small neighborhoo
ds around each pixel. With the multiscale representation we capture the frequenc
y of patterns observed at different scales of resolution. This framework leads t
o expressive priors that depend on a relatively small number of parameters. For
inference and learning we use an MCMC method for block sampling with very large
blocks. We evaluate the approach with two example applications. One involves con
tour detection. The other involves binary segmentation.
****

How hard is my MDP?" The distribution-norm to the rescue"
Odalric-Ambrym Maillard, Timothy A. Mann, Shie Mannor
Requests for name changes in the electronic proceedings will be accepted with no
 questions asked. However name changes may cause bibliographic tracking issues.
  Authors are asked to consider this carefully and discuss it with their co-auth
ors prior to requesting a name change in the electronic proceedings.
****

Spectral Learning of Mixture of Hidden Markov Models
Cem Subakan, Johannes Traa, Paris Smaragdis
In this paper, we propose a learning approach for the Mixture of Hidden Markov M
odels (MHMM) based on the Method of Moments (MoM). Computational advantages of M
oM make MHMM learning amenable for large data sets. It is not possible to direct
ly learn an MHMM with existing learning approaches, mainly due to a permutation
ambiguity in the estimation process. We show that it is possible to resolve this
 ambiguity using the spectral properties of a global transition matrix even in t
he presence of estimation noise. We demonstrate the validity of our approach on
synthetic and real data.
****

Exploiting Linear Structure Within Convolutional Networks for Efficient Evaluati
on
Emily L. Denton, Wojciech Zaremba, Joan Bruna, Yann LeCun, Rob Fergus
We present techniques for speeding up the test-time evaluation of large convolut
ional networks, designed for object recognition tasks. These models deliver impr
essive accuracy, but each image evaluation requires millions of floating point o
perations, making their deployment on smartphones and Internet-scale clusters pr
oblematic. The computation is dominated by the convolution operations in the low
er layers of the model. We exploit the redundancy present within the convolution
al filters to derive approximations that significantly reduce the required compu
tation. Using large state-of-the-art models, we demonstrate speedups of convolut
ional layers on both CPU and GPU by a factor of 2×, while keeping the accuracy w
ithin 1% of the original model.
****

Tree-structured Gaussian Process Approximations
Thang D. Bui, Richard E. Turner
Gaussian process regression can be accelerated by constructing a small pseudo-da
taset to summarise the observed data. This idea sits at the heart of many approx
imation schemes, but such an approach requires the number of pseudo-datapoints t
o be scaled with the range of the input space if the accuracy of the approximati
on is to be maintained. This presents problems in time-series settings or in spa
tial datasets where large numbers of pseudo-datapoints are required since comput
ation typically scales quadratically with the pseudo-dataset size. In this paper

we devise an approximation whose complexity grows linearly with the number of p
seudo-datapoints. This is achieved by imposing a tree or chain structure on the
pseudo-datapoints and calibrating the approximation using a Kullback-Leibler (KL
) minimisation. Inference and learning can then be performed efficiently using t
he Gaussian belief propagation algorithm. We demonstrate the validity of our app
roach on a set of challenging regression tasks including missing data imputation
for audio and spatial datasets. We trace out the speed-accuracy trade-off for t
he new method and show that the frontier dominates those obtained from a large n
umber of existing approximation techniques.
****

Optimal decision-making with time-varying evidence reliability
Jan Drugowitsch, Ruben Moreno-Bote, Alexandre Pouget
Previous theoretical and experimental work on optimal decision-making was restri
cted to the artificial setting of a reliability of the momentary sensory evidenc
e that remained constant within single trials. The work presented here describes
the computation and characterization of optimal decision-making in the more rea
listic case of an evidence reliability that varies across time even within a tri
al. It shows that, in this case, the optimal behavior is determined by a bound i
n the decision maker's belief that depends only on the current, but not the past
, reliability. We furthermore demonstrate that simpler heuristics fail to match
the optimal performance for certain characteristics of the process that determin
es the time-course of this reliability, causing a drop in reward rate by more th
an 50%.
****

An Autoencoder Approach to Learning Bilingual Word Representations
Sarath Chandar A P, Stanislas Lauly, Hugo Larochelle, Mitesh Khapra, Balaraman R
avindran, Vikas C. Raykar, Amrita Saha
Cross-language learning allows us to use training data from one language to buil
d models for a different language. Many approaches to bilingual learning require
that we have word-level alignment of sentences from parallel corpora. In this w
ork we explore the use of autoencoder-based methods for cross-language learning
of vectorial word representations that are aligned between two languages, while
not relying on word-level alignments. We show that by simply learning to reconst
ruct the bag-of-words representations of aligned sentences, within and between l
anguages, we can in fact learn high-quality representations and do without word
alignments. We empirically investigate the success of our approach on the proble
m of cross-language text classification, where a classifier trained on a given l
anguage (e.g., English) must learn to generalize to a different language (e.g.,
German). In experiments on 3 language pairs, we show that our approach achieves
state-of-the-art performance, outperforming a method exploiting word alignments
and a strong machine translation baseline.
****

Testing Unfaithful Gaussian Graphical Models
De Wen Soh, Sekhar C. Tatikonda
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.
****

Deep Recursive Neural Networks for Compositionality in Language
Ozan Irsoy, Claire Cardie
Recursive neural networks comprise a class of architecture that can operate on s
tructured input. They have been previously successfully applied to model composi
tionality in natural language using parse-tree-based structural representations.
Even though these architectures are deep in structure, they lack the capacity f
or hierarchical representation that exists in conventional deep feed-forward net
works as well as in recently investigated deep recurrent neural networks. In thi
s work we introduce a new architecture --- a deep recursive neural network (deep
RNN) --- constructed by stacking multiple recursive layers. We evaluate the pro
posed model on the task of fine-grained sentiment classification. Our results sh

ow that deep RNNs outperform associated shallow counterparts that employ the sam
e number of parameters. Furthermore, our approach outperforms previous baselines
 on the sentiment analysis task, including a multiplicative RNN variant as well
as the recently introduced paragraph vectors, achieving new state-of-the-art res
ults. We provide exploratory analyses of the effect of multiple layers and show
that they capture different aspects of compositionality in language.
****

Signal Aggregate Constraints in Additive Factorial HMMs, with Application to Ene
rgy Disaggregation

Mingjun Zhong, Nigel Goddard, Charles Sutton

Blind source separation problems are difficult because they are inherently unide
ntifiable, yet the entire goal is to identify meaningful sources. We introduce a
 way of incorporating domain knowledge into this problem, called signal aggregat
e constraints (SACs). SACs encourage the total signal for each of the unknown so
urces to be close to a specified value. This is based on the observation that th
e total signal often varies widely across the unknown sources, and we often have
 a good idea of what total values to expect. We incorporate SACs into an additiv
e factorial hidden Markov model (AFHMM) to formulate the energy disaggregation p
roblems where only one mixture signal is assumed to be observed. A convex quadra
tic program for approximate inference is employed for recovering those source si
gnals. On a real-world energy disaggregation data set, we show that the use of S
ACs dramatically improves the original AFHMM, and significantly improves over a
recent state-of-the art approach.
****

Poisson Process Jumping between an Unknown Number of Rates: Application to Neura
l Spike Data

Florian Stimberg, Andreas Ruttor, Manfred Opper

We introduce a model where the rate of an inhomogeneous Poisson process is modif
ied by a Chinese restaurant process. Applying a MCMC sampler to this model allow
s us to do posterior Bayesian inference about the number of states in Poisson-li
ke data. Our sampler is shown to get accurate results for synthetic data and we
apply it to V1 neuron spike data to find discrete firing rate states depending o
n the orientation of a stimulus.
****

Low-Rank Time-Frequency Synthesis

Cédric Févotte, Matthieu Kowalski

Many single-channel signal decomposition techniques rely on a low-rank factoriza
tion of a time-frequency transform. In particular, nonnegative matrix factorizat
ion (NMF) of the spectrogram -- the (power) magnitude of the short-time Fourier
transform (STFT) -- has been considered in many audio applications. In this sett
ing, NMF with the Itakura-Saito divergence was shown to underly a generative Gau
ssian composite model (GCM) of the STFT, a step forward from more empirical appr
oaches based on ad-hoc transform and divergence specifications. Still, the GCM i
s not yet a generative model of the raw signal itself, but only of its STFT. The
 work presented in this paper fills in this ultimate gap by proposing a novel si
gnal synthesis model with low-rank time-frequency structure. In particular, our
new approach opens doors to multi-resolution representations, that were not poss
ible in the traditional NMF setting. We describe two expectation-maximization al
gorithms for estimation in the new model and report audio signal processing resu
lts with music decomposition and speech enhancement.
****

Spike Frequency Adaptation Implements Anticipative Tracking in Continuous Attrac
tor Neural Networks

Yuanyuan Mi, C. C. Alan Fung, K. Y. Michael Wong, Si Wu

To extract motion information, the brain needs to compensate for time delays tha
t are ubiquitous in neural signal transmission and processing. Here we propose a
 simple yet effective mechanism to implement anticipative tracking in neural sys
tems. The proposed mechanism utilizes the property of spike-frequency adaptation
 (SFA), a feature widely observed in neuronal responses. We employ continuous at
tractor neural networks (CANNs) as the model to describe the tracking behaviors

in neural systems. Incorporating SFA, a CANN exhibits intrinsic mobility, manifested by the ability of the CANN to hold self-sustained travelling waves. In tracking a moving stimulus, the interplay between the external drive and the intrinsic mobility of the network determines the tracking performance. Interestingly, we find that the regime of anticipation effectively coincides with the regime where the intrinsic speed of the travelling wave exceeds that of the external drive. Depending on the SFA amplitudes, the network can achieve either perfect tracking, with zero-lag to the input, or perfect anticipative tracking, with a constant leading time to the input. Our model successfully reproduces experimentally observed anticipative tracking behaviors, and sheds light on our understanding of how the brain processes motion information in a timely manner.
****

Learning to Optimize via Information-Directed Sampling
Daniel Russo, Benjamin Van Roy
We propose information-directed sampling -- a new algorithm for online optimization problems in which a decision-maker must balance between exploration and exploitation while learning from partial feedback. Each action is sampled in a manner that minimizes the ratio between the square of expected single-period regret and a measure of information gain: the mutual information between the optimal action and the next observation. We establish an expected regret bound for information-directed sampling that applies across a very general class of models and scales with the entropy of the optimal action distribution. For the widely studied Bernoulli and linear bandit models, we demonstrate simulation performance surpassing popular approaches, including upper confidence bound algorithms, Thompson sampling, and knowledge gradient. Further, we present simple analytic examples illustrating that information-directed sampling can dramatically outperform upper confidence bound algorithms and Thompson sampling due to the way it measures information gain.
****

Distributed Estimation, Information Loss and Exponential Families
Qiang Liu, Alexander T. Ihler
Distributed learning of probabilistic models from multiple data repositories with minimum communication is increasingly important. We study a simple communication-efficient learning framework that first calculates the local maximum likelihood estimates (MLE) based on the data subsets, and then combines the local MLEs to achieve the best possible approximation to the global MLE, based on the whole dataset jointly. We study the statistical properties of this framework, showing that the loss of efficiency compared to the global setting relates to how much the underlying distribution families deviate from full exponential families, drawing connection to the theory of information loss by Fisher, Rao and Efron. We show that the full-exponential-family-ness" represents the lower bound of the error rate of arbitrary combinations of local MLEs, and is achieved by a KL-divergence-based combination method but not by a more common linear combination method. We also study the empirical properties of the KL and linear combination methods, showing that the KL method significantly outperforms linear combination in practical settings with issues such as model misspecification, non-convexity, and heterogeneous data partitions."
****

A* Sampling
Chris J. Maddison, Daniel Tarlow, Tom Minka
The problem of drawing samples from a discrete distribution can be converted into a discrete optimization problem. In this work, we show how sampling from a continuous distribution can be converted into an optimization problem over continuous space. Central to the method is a stochastic process recently described in mathematical statistics that we call the Gumbel process. We present a new construction of the Gumbel process and A* sampling, a practical generic sampling algorithm that searches for the maximum of a Gumbel process using A* search. We analyze the correctness and convergence time of A* sampling and demonstrate empirically that it makes more efficient use of bound and likelihood evaluations than the most closely related adaptive rejection sampling-based algorithms.

****
Consistent Binary Classification with Generalized Performance Metrics
Oluwasanmi O. Koyejo, Nagarajan Natarajan, Pradeep K. Ravikumar, Inderjit S. Dhillon
Performance metrics for binary classification are designed to capture tradeoffs between four fundamental population quantities: true positives, false positives, true negatives and false negatives. Despite significant interest from theoretical and applied communities, little is known about either optimal classifiers or consistent algorithms for optimizing binary classification performance metrics beyond a few special cases. We consider a fairly large family of performance metrics given by ratios of linear combinations of the four fundamental population quantities. This family includes many well known binary classification metrics such as classification accuracy, AM measure, F-measure and the Jaccard similarity coefficient as special cases. Our analysis identifies the optimal classifiers as the sign of the thresholded conditional probability of the positive class, with a performance metric-dependent threshold. The optimal threshold can be constructed using simple plug-in estimators when the performance metric is a linear combination of the population quantities, but alternative techniques are required for the general case. We propose two algorithms for estimating the optimal classifiers, and prove their statistical consistency. Both algorithms are straightforward modifications of standard approaches to address the key challenge of optimal threshold selection, thus are simple to implement in practice. The first algorithm combines a plug-in estimate of the conditional probability of the positive class with optimal threshold selection. The second algorithm leverages recent work on calibrated asymmetric surrogate losses to construct candidate classifiers. We present empirical comparisons between these algorithms on benchmark datasets.
****
Asymmetric LSH (ALSH) for Sublinear Time Maximum Inner Product Search (MIPS)
Anshumali Shrivastava, Ping Li
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Graphical Models for Recovering Probabilistic and Causal Queries from Missing Data
Karthika Mohan, Judea Pearl
We address the problem of deciding whether a causal or probabilistic query is estimable from data corrupted by missing entries, given a model of missingness process. We extend the results of Mohan et al, 2013 by presenting more general conditions for recovering probabilistic queries of the form P(y|x) and P(y,x) as well as causal queries of the form P(y|do(x)). We show that causal queries may be recoverable even when the factors in their identifying estimands are not recoverable. Specifically, we derive graphical conditions for recovering causal effects of the form P(y|do(x)) when Y and its missingness mechanism are not d-separable. Finally, we apply our results to problems of attrition and characterize the recovery of causal effects from data corrupted by attrition.
****
Restricted Boltzmann machines modeling human choice
Takayuki Osogami, Makoto Otsuka
We extend the multinomial logit model to represent some of the empirical phenomena that are frequently observed in the choices made by humans. These phenomena include the similarity effect, the attraction effect, and the compromise effect. We formally quantify the strength of these phenomena that can be represented by our choice model, which illuminates the flexibility of our choice model. We then show that our choice model can be represented as a restricted Boltzmann machine and that its parameters can be learned effectively from data. Our numerical experiments with real data of human choices suggest that we can train our choice model in such a way that it represents the typical phenomena of choice.
****

On the Statistical Consistency of Plug-in Classifiers for Non-decomposable Performance Measures

Harikrishna Narasimhan, Rohit Vaish, Shivani Agarwal

We study consistency properties of algorithms for non-decomposable performance measures that cannot be expressed as a sum of losses on individual data points, such as the F-measure used in text retrieval and several other performance measures used in class imbalanced settings. While there has been much work on designing algorithms for such performance measures, there is limited understanding of the theoretical properties of these algorithms. Recently, Ye et al. (2012) showed consistency results for two algorithms that optimize the F-measure, but their results apply only to an idealized setting, where precise knowledge of the underlying probability distribution (in the form of the true' posterior class probability) is available to a learning algorithm. In this work, we consider plug-in algorithms that learn a classifier by applying an empirically determined threshold to a suitableestimate' of the class probability, and provide a general methodology to show consistency of these methods for any non-decomposable measure that can be expressed as a continuous function of true positive rate (TPR) and true negative rate (TNR), and for which the Bayes optimal classifier is the class probability function thresholded suitably. We use this template to derive consistency results for plug-in algorithms for the F-measure and for the geometric mean of TPR and precision; to our knowledge, these are the first such results for these measures. In addition, for continuous distributions, we show consistency of plug-in algorithms for any performance measure that is a continuous and monotonically increasing function of TPR and TNR. Experimental results confirm our theoretical findings.

****

Clustering from Labels and Time-Varying Graphs

Shiau Hong Lim, Yudong Chen, Huan Xu

We present a general framework for graph clustering where a label is observed to each pair of nodes. This allows a very rich encoding of various types of pairwise interactions between nodes. We propose a new tractable approach to this problem based on maximum likelihood estimator and convex optimization. We analyze our algorithm under a general generative model, and provide both necessary and sufficient conditions for successful recovery of the underlying clusters. Our theoretical results cover and subsume a wide range of existing graph clustering results including planted partition, weighted clustering and partially observed graphs. Furthermore, the result is applicable to novel settings including time-varying graphs such that new insights can be gained on solving these problems. Our theoretical findings are further supported by empirical results on both synthetic and real data.

****

Unsupervised learning of an efficient short-term memory network

Pietro Vertechi, Wieland Brendel, Christian K. Machens

Learning in recurrent neural networks has been a topic fraught with difficulties and problems. We here report substantial progress in the unsupervised learning of recurrent networks that can keep track of an input signal. Specifically, we show how these networks can learn to efficiently represent their present and past inputs, based on local learning rules only. Our results are based on several key insights. First, we develop a local learning rule for the recurrent weights whose main aim is to drive the network into a regime where, on average, feedforward signal inputs are canceled by recurrent inputs. We show that this learning rule minimizes a cost function. Second, we develop a local learning rule for the feedforward weights that, based on networks in which recurrent inputs already predict feedforward inputs, further minimizes the cost. Third, we show how the learning rules can be modified such that the network can directly encode non-whitened inputs. Fourth, we show that these learning rules can also be applied to a network that feeds a time-delayed version of the network output back into itself. As a consequence, the network starts to efficiently represent both its signal inputs and their history. We develop our main theory for linear networks, but then sketch how the learning rules could be transferred to balanced, spiking networks.

****
Projective dictionary pair learning for pattern classification
Shuhang Gu, Lei Zhang, Wangmeng Zuo, Xiangchu Feng
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Analysis of Learning from Positive and Unlabeled Data
Marthinus C. du Plessis, Gang Niu, Masashi Sugiyama
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 with Fredholm Kernels
Qichao Que, Mikhail Belkin, Yusu Wang
In this paper we propose a framework for supervised and semi-supervised learning based on reformulating the learning problem as a regularized Fredholm integral equation. Our approach fits naturally into the kernel framework and can be interpreted as constructing new data-dependent kernels, which we call Fredholm kernels. We proceed to discuss the noise assumption" for semi-supervised learning and provide evidence evidence both theoretical and experimental that Fredholm kernels can effectively utilize unlabeled data under the noise assumption. We demonstrate that methods based on Fredholm learning show very competitive performance in the standard semi-supervised learning setting."
****
How transferable are features in deep neural networks?
Jason Yosinski, Jeff Clune, Yoshua Bengio, Hod Lipson
Many deep neural networks trained on natural images exhibit a curious phenomenon in common: on the first layer they learn features similar to Gabor filters and color blobs. Such first-layer features appear not to be specific to a particular dataset or task, but general in that they are applicable to many datasets and tasks. Features must eventually transition from general to specific by the last layer of the network, but this transition has not been studied extensively. In this paper we experimentally quantify the generality versus specificity of neurons in each layer of a deep convolutional neural network and report a few surprising results. Transferability is negatively affected by two distinct issues: (1) the specialization of higher layer neurons to their original task at the expense of performance on the target task, which was expected, and (2) optimization difficulties related to splitting networks between co-adapted neurons, which was not expected. In an example network trained on ImageNet, we demonstrate that either of these two issues may dominate, depending on whether features are transferred from the bottom, middle, or top of the network. We also document that the transferability of features decreases as the distance between the base task and target task increases, but that transferring features even from distant tasks can be better than using random features. A final surprising result is that initializing a network with transferred features from almost any number of layers can produce a boost to generalization that lingers even after fine-tuning to the target dataset.
****
Concavity of reweighted Kikuchi approximation
Po-Ling Loh, Andre Wibisono
We analyze a reweighted version of the Kikuchi approximation for estimating the log partition function of a product distribution defined over a region graph. We establish sufficient conditions for the concavity of our reweighted objective function in terms of weight assignments in the Kikuchi expansion, and show that a reweighted version of the sum product algorithm applied to the Kikuchi region graph will produce global optima of the Kikuchi approximation whenever the algorithm converges. When the region graph has two layers, corresponding to a Bethe ap

proximation, we show that our sufficient conditions for concavity are also neces
sary. Finally, we provide an explicit characterization of the polytope of concav
ity in terms of the cycle structure of the region graph. We conclude with simula
tions that demonstrate the advantages of the reweighted Kikuchi approach.
****

The Bayesian Case Model: A Generative Approach for Case-Based Reasoning and Prot
otype Classification
Been Kim, Cynthia Rudin, Julie A. Shah
We present the Bayesian Case Model (BCM), a general framework for Bayesian case-
based reasoning (CBR) and prototype classification and clustering. BCM brings th
e intuitive power of CBR to a Bayesian generative framework. The BCM learns prot
otypes, the ``quintessential observations that best represent clusters in a data
set, by performing joint inference on cluster labels, prototypes and important f
eatures. Simultaneously, BCM pursues sparsity by learning subspaces, the sets of
 features that play important roles in the characterization of the prototypes. T
he prototype and subspace representation provides quantitative benefits in inter
pretability while preserving classification accuracy. Human subject experiments
verify statistically significant improvements to participants' understanding whe
n using explanations produced by BCM, compared to those given by prior art."
****

Advances in Learning Bayesian Networks of Bounded Treewidth
Siqi Nie, Denis D. Maua, Cassio P. de Campos, Qiang Ji
This work presents novel algorithms for learning Bayesian networks of bounded tr
eewidth. Both exact and approximate methods are developed. The exact method comb
ines mixed integer linear programming formulations for structure learning and tr
eewidth computation. The approximate method consists in sampling k-trees (maxima
l graphs of treewidth k), and subsequently selecting, exactly or approximately,
the best structure whose moral graph is a subgraph of that k-tree. The approache
s are empirically compared to each other and to state-of-the-art methods on a co
llection of public data sets with up to 100 variables.
****

Learning From Weakly Supervised Data by The Expectation Loss SVM (e-SVM) algorit
hm
Jun Zhu, Junhua Mao, Alan L. Yuille
In many situations we have some measurement of confidence on positiveness for a
binary label. Thepositiveness" is a continuous value whose range is a bounded in
terval. It quantifies the affiliation of each training data to the positive clas
s. We propose a novel learning algorithm called \emph{expectation loss SVM} (e-S
VM) that is devoted to the problems where only the ``positiveness" instead of a
binary label of each training sample is available. Our e-SVM algorithm can also
be readily extended to learn segment classifiers under weak supervision where th
e exact positiveness value of each training example is unobserved. In experiment
s, we show that the e-SVM algorithm can effectively address the segment proposal
 classification task under both strong supervision (e.g. the pixel-level annotat
ions are available) and the weak supervision (e.g. only bounding-box annotations
 are available), and outperforms the alternative approaches. Besides, we further
 validate this method on two major tasks of computer vision: semantic segmentati
on and object detection. Our method achieves the state-of-the-art object detecti
on performance on PASCAL VOC 2007 dataset."
****

On the Computational Efficiency of Training Neural Networks
Roi Livni, Shai Shalev-Shwartz, Ohad Shamir
It is well-known that neural networks are computationally hard to train. On the
other hand, in practice, modern day neural networks are trained efficiently usin
g SGD and a variety of tricks that include different activation functions (e.g.
ReLU), over-specification (i.e., train networks which are larger than needed), a
nd regularization. In this paper we revisit the computational complexity of trai
ning neural networks from a modern perspective. We provide both positive and neg
ative results, some of them yield new provably efficient and practical algorithm
s for training neural networks.

****
Scaling-up Importance Sampling for Markov Logic Networks
Deepak Venugopal, Vibhav G. Gogate
Markov Logic Networks (MLNs) are weighted first-order logic templates for genera ting large (ground) Markov networks. Lifted inference algorithms for them bring the power of logical inference to probabilistic inference. These algorithms oper ate as much as possible at the compact first-order level, grounding or propositi onalizing the MLN only as necessary. As a result, lifted inference algorithms ca n be much more scalable than propositional algorithms that operate directly on t he much larger ground network. Unfortunately, existing lifted inference algorith ms suffer from two interrelated problems, which severely affects their scalabili ty in practice. First, for most real-world MLNs having complex structure, they a re unable to exploit symmetries and end up grounding most atoms (the grounding p roblem). Second, they suffer from the evidence problem, which arises because evi dence breaks symmetries, severely diminishing the power of lifted inference. In this paper, we address both problems by presenting a scalable, lifted importance sampling-based approach that never grounds the full MLN. Specifically, we show how to scale up the two main steps in importance sampling: sampling from the pro posal distribution and weight computation. Scalable sampling is achieved by usin g an informed, easy-to-sample proposal distribution derived from a compressed ML N-representation. Fast weight computation is achieved by only visiting a small s ubset of the sampled groundings of each formula instead of all of its possible g roundings. We show that our new algorithm yields an asymptotically unbiased esti mate. Our experiments on several MLNs clearly demonstrate the promise of our app roach.
****
Unsupervised Transcription of Piano Music
Taylor Berg-Kirkpatrick, Jacob Andreas, Dan Klein
We present a new probabilistic model for transcribing piano music from audio to a symbolic form. Our model reflects the process by which discrete musical events give rise to acoustic signals that are then superimposed to produce the observe d data. As a result, the inference procedure for our model naturally resolves th e source separation problem introduced by the the piano's polyphony. In order to adapt to the properties of a new instrument or acoustic environment being trans cribed, we learn recording specific spectral profiles and temporal envelopes in an unsupervised fashion. Our system outperforms the best published approaches on a standard piano transcription task, achieving a 10.6% relative gain in note on set F1 on real piano audio.
****
Multi-Scale Spectral Decomposition of Massive Graphs
Si Si, Donghyuk Shin, Inderjit S. Dhillon, Beresford N. Parlett
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.
****
Dimensionality Reduction with Subspace Structure Preservation
Devansh Arpit, Ifeoma Nwogu, Venu Govindaraju
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.
****
Ranking via Robust Binary Classification
Hyokun Yun, Parameswaran Raman, S. Vishwanathan
We propose RoBiRank, a ranking algorithm that is motivated by observing a close connection between evaluation metrics for learning to rank and loss functions fo r robust classification. The algorithm shows a very competitive performance on s tandard benchmark datasets against other representative algorithms in the litera ture. Further, in large scale problems where explicit feature vectors and scores

are not given, our algorithm can be efficiently parallelized across a large number of machines; for a task that requires 386,133 x 49,824,519 pairwise interactions between items to be ranked, our algorithm finds solutions that are of dramatically higher quality than that can be found by a state-of-the-art competitor algorithm, given the same amount of wall-clock time for computation.
****

Learning the Learning Rate for Prediction with Expert Advice
Wouter M. Koolen, Tim van Erven, Peter Grünwald
Most standard algorithms for prediction with expert advice depend on a parameter called the learning rate. This learning rate needs to be large enough to fit the data well, but small enough to prevent overfitting. For the exponential weights algorithm, a sequence of prior work has established theoretical guarantees for higher and higher data-dependent tunings of the learning rate, which allow for increasingly aggressive learning. But in practice such theoretical tunings often still perform worse (as measured by their regret) than ad hoc tuning with an even higher learning rate. To close the gap between theory and practice we introduce an approach to learn the learning rate. Up to a factor that is at most (poly) logarithmic in the number of experts and the inverse of the learning rate, our method performs as well as if we would know the empirically best learning rate from a large range that includes both conservative small values and values that are much higher than those for which formal guarantees were previously available. Our method employs a grid of learning rates, yet runs in linear time regardless of the size of the grid.
****

Beta-Negative Binomial Process and Exchangeable ■Random Partitions for Mixed-Membership Modeling
Mingyuan Zhou
The beta-negative binomial process (BNBP), an integer-valued stochastic process, is employed to partition a count vector into a latent random count matrix. As the marginal probability distribution of the BNBP that governs the exchangeable random partitions of grouped data has not yet been developed, current inference for the BNBP has to truncate the number of atoms of the beta process. This paper introduces an exchangeable partition probability function to explicitly describe how the BNBP clusters the data points of each group into a random number of exchangeable partitions, which are shared across all the groups. A fully collapsed Gibbs sampler is developed for the BNBP, leading to a novel nonparametric Bayesian topic model that is distinct from existing ones, with simple implementation, fast convergence, good mixing, and state-of-the-art predictive performance.
****

Learning Deep Features for Scene Recognition using Places Database
Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, Aude Oliva
Scene recognition is one of the hallmark tasks of computer vision, allowing definition of a context for object recognition. Whereas the tremendous recent progress in object recognition tasks is due to the availability of large datasets like ImageNet and the rise of Convolutional Neural Networks (CNNs) for learning high-level features, performance at scene recognition has not attained the same level of success. This may be because current deep features trained from ImageNet are not competitive enough for such tasks. Here, we introduce a new scene-centric database called Places with over 7 million labeled pictures of scenes. We propose new methods to compare the density and diversity of image datasets and show that Places is as dense as other scene datasets and has more diversity. Using CNN, we learn deep features for scene recognition tasks, and establish new state-of-the-art results on several scene-centric datasets. A visualization of the CNN layers' responses allows us to show differences in the internal representations of object-centric and scene-centric networks.
****

Efficient Sampling for Learning Sparse Additive Models in High Dimensions
Hemant Tyagi, Bernd Gärtner, Andreas Krause
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.
****

A framework for studying synaptic plasticity with neural spike train data
Scott Linderman, Christopher H. Stock, Ryan P. Adams
Learning and memory in the brain are implemented by complex, time-varying change
s in neural circuitry. The computational rules according to which synaptic weigh
ts change over time are the subject of much research, and are not precisely unde
rstood. Until recently, limitations in experimental methods have made it challen
ging to test hypotheses about synaptic plasticity on a large scale. However, as
such data become available and these barriers are lifted, it becomes necessary t
o develop analysis techniques to validate plasticity models. Here, we present a
highly extensible framework for modeling arbitrary synaptic plasticity rules on
spike train data in populations of interconnected neurons. We treat synaptic wei
ghts as a (potentially nonlinear) dynamical system embedded in a fully-Bayesian
generalized linear model (GLM). In addition, we provide an algorithm for inferri
ng synaptic weight trajectories alongside the parameters of the GLM and of the l
earning rules. Using this method, we perform model comparison of two proposed va
riants of the well-known spike-timing-dependent plasticity (STDP) rule, where no
nlinear effects play a substantial role. On synthetic data generated from the bi
ophysical simulator NEURON, we show that we can recover the weight trajectories,
the pattern of connectivity, and the underlying learning rules.
****

Real-Time Decoding of an Integrate and Fire Encoder
Shreya Saxena, Munther Dahleh
Neuronal encoding models range from the detailed biophysically-based Hodgkin Hux
ley model, to the statistical linear time invariant model specifying firing rate
s in terms of the extrinsic signal. Decoding the former becomes intractable, whi
le the latter does not adequately capture the nonlinearities present in the neur
onal encoding system. For use in practical applications, we wish to record the o
utput of neurons, namely spikes, and decode this signal fast in order to drive a
machine, for example a prosthetic device. Here, we introduce a causal, real-tim
e decoder of the biophysically-based Integrate and Fire encoding neuron model. W
e show that the upper bound of the real-time reconstruction error decreases poly
nomially in time, and that the L2 norm of the error is bounded by a constant tha
t depends on the density of the spikes, as well as the bandwidth and the decay o
f the input signal. We numerically validate the effect of these parameters on th
e reconstruction error.
****

Parallel Direction Method of Multipliers
Huahua Wang, Arindam Banerjee, Zhi-Quan Luo
We consider the problem of minimizing block-separable convex functions subject t
o linear constraints. While the Alternating Direction Method of Multipliers (ADM
M) for two-block linear constraints has been intensively studied both theoretica
lly and empirically, in spite of some preliminary work, effective generalization
s of ADMM to multiple blocks is still unclear. In this paper, we propose a paral
lel randomized block coordinate method named Parallel Direction Method of Multip
liers (PDMM) to solve the optimization problems with multi-block linear constrai
nts. PDMM randomly updates some blocks in parallel, behaving like parallel rando
mized block coordinate descent. We establish the global convergence and the iter
ation complexity for PDMM with constant step size. We also show that PDMM can do
randomized block coordinate descent on overlapping blocks. Experimental results
show that PDMM performs better than state-of-the-arts methods in two applicatio
ns, robust principal component analysis and overlapping group lasso.
****

Spectral Methods for Supervised Topic Models
Yining Wang, Jun Zhu
Supervised topic models simultaneously model the latent topic structure of large
collections of documents and a response variable associated with each document.
Existing inference methods are based on either variational approximation or Mon

te Carlo sampling. This paper presents a novel spectral decomposition algorithm to recover the parameters of supervised latent Dirichlet allocation (sLDA) models. The Spectral-sLDA algorithm is provably correct and computationally efficient. We prove a sample complexity bound and subsequently derive a sufficient condition for the identifiability of sLDA. Thorough experiments on a diverse range of synthetic and real-world datasets verify the theory and demonstrate the practical effectiveness of the algorithm.
****

Exclusive Feature Learning on Arbitrary Structures via $\ell_{1,2}$-norm
Deguang Kong, Ryohei Fujimaki, Ji Liu, Feiping Nie, Chris Ding
Group lasso is widely used to enforce the structural sparsity, which achieves the sparsity at inter-group level. In this paper, we propose a new formulation called ``exclusive group lasso'', which brings out sparsity at intra-group level in the context of feature selection. The proposed exclusive group lasso is applicable on any feature structures, regardless of their overlapping or non-overlapping structures. We give analysis on the properties of exclusive group lasso, and propose an effective iteratively re-weighted algorithm to solve the corresponding optimization problem with rigorous convergence analysis. We show applications of exclusive group lasso for uncorrelated feature selection. Extensive experiments on both synthetic and real-world datasets indicate the good performance of proposed methods.
****

Non-convex Robust PCA
Praneeth Netrapalli, Niranjan U N, Sujay Sanghavi, Animashree Anandkumar, Pratek 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.
****

Expectation-Maximization for Learning Determinantal Point Processes
Jennifer A. Gillenwater, Alex Kulesza, Emily Fox, Ben Taskar
A determinantal point process (DPP) is a probabilistic model of set diversity compactly parameterized by a positive semi-definite kernel matrix. To fit a DPP to a given task, we would like to learn the entries of its kernel matrix by maximizing the log-likelihood of the available data. However, log-likelihood is non-convex in the entries of the kernel matrix, and this learning problem is conjectured to be NP-hard. Thus, previous work has instead focused on more restricted convex learning settings: learning only a single weight for each row of the kernel matrix, or learning weights for a linear combination of DPPs with fixed kernel matrices. In this work we propose a novel algorithm for learning the full kernel matrix. By changing the kernel parameterization from matrix entries to eigenvalues and eigenvectors, and then lower-bounding the likelihood in the manner of expectation-maximization algorithms, we obtain an effective optimization procedure. We test our method on a real-world product recommendation task, and achieve relative gains of up to 16.5% in test log-likelihood compared to the naive approach of maximizing likelihood by projected gradient ascent on the entries of the kernel matrix.
****

Estimation with Norm Regularization
Arindam Banerjee, Sheng Chen, Farideh Fazayeli, Vidyashankar Sivakumar
Analysis of estimation error and associated structured statistical recovery based on norm regularized regression, e.g., Lasso, needs to consider four aspects: the norm, the loss function, the design matrix, and the noise vector. This paper presents generalizations of such estimation error analysis on all four aspects, compared to the existing literature. We characterize the restricted error set, establish relations between error sets for the constrained and regularized problems, and present an estimation error bound applicable to {\em any} norm. Precise characterizations of the bound are presented for a variety of noise vectors, design matrices, including sub-Gaussian, anisotropic, and dependent samples, and lo

ss functions, including least squares and generalized linear models. Gaussian wi dths, as a measure of size of suitable sets, and associated tools play a key rol e in our generalized analysis.
****

Sparse Random Feature Algorithm as Coordinate Descent in Hilbert Space
Ian En-Hsu Yen, Ting-Wei Lin, Shou-De Lin, Pradeep K. Ravikumar, Inderjit S. Dhi llon
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.
****

Nonparametric Bayesian inference on multivariate exponential families
William R. Vega-Brown, Marek Doniec, Nicholas G. Roy
We develop a model by choosing the maximum entropy distribution from the set of models satisfying certain smoothness and independence criteria; we show that inf erence on this model generalizes local kernel estimation to the context of Bayes ian inference on stochastic processes. Our model enables Bayesian inference in c ontexts when standard techniques like Gaussian process inference are too expensi ve to apply. Exact inference on our model is possible for any likelihood functio n from the exponential family. Inference is then highly efficient, requiring onl y O(log N) time and O(N) space at run time. We demonstrate our algorithm on seve ral problems and show quantifiable improvement in both speed and performance rel ative to models based on the Gaussian process.
****

On Communication Cost of Distributed Statistical Estimation and Dimensionality
Ankit Garg, Tengyu Ma, Huy 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-auth ors prior to requesting a name change in the electronic proceedings.
****

A Block-Coordinate Descent Approach for Large-scale Sparse Inverse Covariance Es timation
Eran Treister, Javier S. Turek
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.
****

Local Decorrelation For Improved Pedestrian Detection
Woonhyun Nam, Piotr Dollar, Joon Hee Han
Even with the advent of more sophisticated, data-hungry methods, boosted decisio n trees remain extraordinarily successful for fast rigid object detection, achie ving top accuracy on numerous datasets. While effective, most boosted detectors use decision trees with orthogonal (single feature) splits, and the topology of the resulting decision boundary may not be well matched to the natural topology of the data. Given highly correlated data, decision trees with oblique (multiple feature) splits can be effective. Use of oblique splits, however, comes at cons iderable computational expense. Inspired by recent work on discriminative decorr elation of HOG features, we instead propose an efficient feature transform that removes correlations in local neighborhoods. The result is an overcomplete but l ocally decorrelated representation ideally suited for use with orthogonal decisi on trees. In fact, orthogonal trees with our locally decorrelated features outpe rform oblique trees trained over the original features at a fraction of the comp utational cost. The overall improvement in accuracy is dramatic: on the Caltech Pedestrian Dataset, we reduce false positives nearly tenfold over the previous s tate-of-the-art.
****

Graph Clustering With Missing Data: Convex Algorithms and Analysis

Ramya Korlakai Vinayak, Samet Oymak, Babak Hassibi
We consider the problem of finding clusters in an unweighted graph, when the gra ph is partially observed. We analyze two programs, one which works for dense gra phs and one which works for both sparse and dense graphs, but requires some a pr iori knowledge of the total cluster size, that are based on the convex optimizat ion approach for low-rank matrix recovery using nuclear norm minimization. For t he commonly used Stochastic Block Model, we obtain \emph{explicit} bounds on the parameters of the problem (size and sparsity of clusters, the amount of observe d data) and the regularization parameter characterize the success and failure of the programs. We corroborate our theoretical findings through extensive simulat ions. We also run our algorithm on a real data set obtained from crowdsourcing a n image classification task on the Amazon Mechanical Turk, and observe significa nt performance improvement over traditional methods such as k-means.
****

Spatio-temporal Representations of Uncertainty in Spiking Neural Networks
Cristina Savin, Sophie Denève
It has been long argued that, because of inherent ambiguity and noise, the brain needs to represent uncertainty in the form of probability distributions. The ne ural encoding of such distributions remains however highly controversial. Here w e present a novel circuit model for representing multidimensional real-valued di stributions using a spike based spatio-temporal code. Our model combines the com putational advantages of the currently competing models for probabilistic codes and exhibits realistic neural responses along a variety of classic measures. Fur thermore, the model highlights the challenges associated with interpreting neura l activity in relation to behavioral uncertainty and points to alternative popul ation-level approaches for the experimental validation of distributed representa tions.
****

Hamming Ball Auxiliary Sampling for Factorial Hidden Markov Models
Michalis Titsias RC AUEB, Christopher Yau
We introduce a novel sampling algorithm for Markov chain Monte Carlo-based Bayes ian inference for factorial hidden Markov models. This algorithm is based on an auxiliary variable construction that restricts the model space allowing iterativ e exploration in polynomial time. The sampling approach overcomes limitations wi th common conditional Gibbs samplers that use asymmetric updates and become easi ly trapped in local modes. Instead, our method uses symmetric moves that allows joint updating of the latent sequences and improves mixing. We illustrate the ap plication of the approach with simulated and a real data example.
****

The Infinite Mixture of Infinite Gaussian Mixtures
Halid Z. Yerebakan, Bartek Rajwa, Murat Dundar
Dirichlet process mixture of Gaussians (DPMG) has been used in the literature fo r clustering and density estimation problems. However, many real-world data exhi bit cluster distributions that cannot be captured by a single Gaussian. Modeling such data sets by DPMG creates several extraneous clusters even when clusters a re relatively well-defined. Herein, we present the infinite mixture of infinite Gaussian mixtures (I2GMM) for more flexible modeling of data sets with skewed an d multi-modal cluster distributions. Instead of using a single Gaussian for each cluster as in the standard DPMG model, the generative model of I2GMM uses a sin gle DPMG for each cluster. The individual DPMGs are linked together through cent ering of their base distributions at the atoms of a higher level DP prior. Infer ence is performed by a collapsed Gibbs sampler that also enables partial paralle lization. Experimental results on several artificial and real-world data sets su ggest the proposed I2GMM model can predict clusters more accurately than existin g variational Bayes and Gibbs sampler versions of DPMG.
****

Partition-wise Linear Models
Hidekazu Oiwa, Ryohei Fujimaki
Region-specific linear models are widely used in practical applications because of their non-linear but highly interpretable model representations. One of the k

ey challenges in their use is non-convexity in simultaneous optimization of regi ons and region-specific models. This paper proposes novel convex region-specific linear models, which we refer to as partition-wise linear models. Our key ideas are 1) assigning linear models not to regions but to partitions (region-specifi ers) and representing region-specific linear models by linear combinations of pa rtition-specific models, and 2) optimizing regions via partition selection from a large number of given partition candidates by means of convex structured regul arizations. In addition to providing initialization-free globally-optimal soluti ons, our convex formulation makes it possible to derive a generalization bound a nd to use such advanced optimization techniques as proximal methods and decompos ition of the proximal maps for sparsity-inducing regularizations. Experimental r esults demonstrate that our partition-wise linear models perform better than or are at least competitive with state-of-the-art region-specific or locally linear models.

****

Convex Deep Learning via Normalized Kernels

Özlem Aslan, Xinhua Zhang, Dale Schuurmans

Deep learning has been a long standing pursuit in machine learning, which until recently was hampered by unreliable training methods before the discovery of imp roved heuristics for embedded layer training. A complementary research strategy is to develop alternative modeling architectures that admit efficient training m ethods while expanding the range of representable structures toward deep models. In this paper, we develop a new architecture for nested nonlinearities that all ows arbitrarily deep compositions to be trained to global optimality. The approa ch admits both parametric and nonparametric forms through the use of normalized kernels to represent each latent layer. The outcome is a fully convex formulatio n that is able to capture compositions of trainable nonlinear layers to arbitrar y depth.

****

Discovering Structure in High-Dimensional Data Through Correlation Explanation

Greg Ver Steeg, Aram Galstyan

We introduce a method to learn a hierarchy of successively more abstract represe ntations of complex data based on optimizing an information-theoretic objective. Intuitively, the optimization searches for a set of latent factors that best ex plain the correlations in the data as measured by multivariate mutual informatio n. The method is unsupervised, requires no model assumptions, and scales linearl y with the number of variables which makes it an attractive approach for very hi gh dimensional systems. We demonstrate that Correlation Explanation (CorEx) auto matically discovers meaningful structure for data from diverse sources including personality tests, DNA, and human language.

****

Difference of Convex Functions Programming for Reinforcement Learning

Bilal Piot, Matthieu Geist, Olivier Pietquin

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.

****

Local Linear Convergence of Forward--Backward under Partial Smoothness

Jingwei Liang, Jalal Fadili, Gabriel Peyré

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.

****

Improved Distributed Principal Component Analysis

Yingyu Liang, Maria-Florina F. Balcan, Vandana Kanchanapally, David Woodruff

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.
****
Reputation-based Worker Filtering in Crowdsourcing
Srikanth Jagabathula, Lakshminarayanan Subramanian, Ashwin Venkataraman
In this paper, we study the problem of aggregating noisy labels from crowd worke rs to infer the underlying true labels of binary tasks. Unlike most prior work w hich has examined this problem under the random worker paradigm, we consider a m uch broader class of {\em adversarial} workers with no specific assumptions on t heir labeling strategy. Our key contribution is the design of a computationally efficient reputation algorithm to identify and filter out these adversarial work ers in crowdsourcing systems. Our algorithm uses the concept of optimal semi-mat chings in conjunction with worker penalties based on label disagreements, to ass ign a reputation score for every worker. We provide strong theoretical guarantee s for deterministic adversarial strategies as well as the extreme case of {\em s ophisticated} adversaries where we analyze the worst-case behavior of our algori thm. Finally, we show that our reputation algorithm can significantly improve th e accuracy of existing label aggregation algorithms in real-world crowdsourcing datasets.
****
large scale canonical correlation analysis with iterative least squares
Yichao Lu, Dean P. Foster
Canonical Correlation Analysis (CCA) is a widely used statistical tool with both well established theory and favorable performance for a wide range of machine l earning problems. However, computing CCA for huge datasets can be very slow sinc e it involves implementing QR decomposition or singular value decomposition of h uge matrices. In this paper we introduce L-CCA, an iterative algorithm which can compute CCA fast on huge sparse datasets. Theory on both the asymptotic converg ence and finite time accuracy of L-CCA are established. The experiments also sho w that L-CCA outperform other fast CCA approximation schemes on two real dataset s.
****
Analysis of Variational Bayesian Latent Dirichlet Allocation: Weaker Sparsity Th an MAP
Shinichi Nakajima, Issei Sato, Masashi Sugiyama, Kazuho Watanabe, Hiroko Kobayas hi
Latent Dirichlet allocation (LDA) is a popular generative model of various objec ts such as texts and images, where an object is expressed as a mixture of latent topics. In this paper, we theoretically investigate variational Bayesian (VB) l earning in LDA. More specifically, we analytically derive the leading term of th e VB free energy under an asymptotic setup, and show that there exist transition thresholds in Dirichlet hyperparameters around which the sparsity-inducing beha vior drastically changes. Then we further theoretically reveal the notable pheno menon that VB tends to induce weaker sparsity than MAP in the LDA model, which i s opposed to other models. We experimentally demonstrate the practical validity of our asymptotic theory on real-world Last.FM music data.
****
Iterative Neural Autoregressive Distribution Estimator NADE-k
Tapani Raiko, Yao Li, Kyunghyun Cho, Yoshua Bengio
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.
****
Reducing the Rank in Relational Factorization Models by Including Observable Pat terns
Maximilian Nickel, Xueyan Jiang, Volker Tresp
Tensor factorizations have become popular methods for learning from multi-relati onal data. In this context, the rank of a factorization is an important paramete r that determines runtime as well as generalization ability. To determine condit ions under which factorization is an efficient approach for learning from relati

onal data, we derive upper and lower bounds on the rank required to recover adja
cency tensors. Based on our findings, we propose a novel additive tensor factori
zation model for learning from latent and observable patterns in multi-relationa
l data and present a scalable algorithm for computing the factorization. Experim
entally, we show that the proposed approach does not only improve the predictive
 performance over pure latent variable methods but that it also reduces the requ
ired rank --- and therefore runtime and memory complexity --- significantly.
****

Deterministic Symmetric Positive Semidefinite Matrix Completion
William E. Bishop, Byron M. Yu
We consider the problem of recovering a symmetric, positive semidefinite (SPSD)
matrix from a subset of its entries, possibly corrupted by noise. In contrast to
 previous matrix recovery work, we drop the assumption of a random sampling of e
ntries in favor of a deterministic sampling of principal submatrices of the matr
ix. We develop a set of sufficient conditions for the recovery of a SPSD matrix
from a set of its principal submatrices, present necessity results based on this
 set of conditions and develop an algorithm that can exactly recover a matrix wh
en these conditions are met. The proposed algorithm is naturally generalized to
the problem of noisy matrix recovery, and we provide a worst-case bound on recon
struction error for this scenario. Finally, we demonstrate the algorithm's utili
ty on noiseless and noisy simulated datasets.
****

Distributed Parameter Estimation in Probabilistic Graphical Models
Yariv D. Mizrahi, Misha Denil, Nando de Freitas
This paper presents foundational theoretical results on distributed parameter es
timation for undirected probabilistic graphical models. It introduces a general
condition on composite likelihood decompositions of these models which guarantee
s the global consistency of distributed estimators, provided the local estimator
s are consistent.
****

Two-Layer Feature Reduction for Sparse-Group Lasso via Decomposition of Convex S
ets
Jie Wang, Jieping Ye
Sparse-Group Lasso (SGL) has been shown to be a powerful regression technique fo
r simultaneously discovering group and within-group sparse patterns by using a c
ombination of the l1 and l2 norms. However, in large-scale applications, the com
plexity of the regularizers entails great computational challenges. In this pape
r, we propose a novel two-layer feature reduction method (TLFre) for SGL via a d
ecomposition of its dual feasible set. The two-layer reduction is able to quickl
y identify the inactive groups and the inactive features, respectively, which ar
e guaranteed to be absent from the sparse representation and can be removed from
 the optimization. Existing feature reduction methods are only applicable for sp
arse models with one sparsity-inducing regularizer. To our best knowledge, TLFre
 is the first one that is capable of dealing with multiple sparsity-inducing reg
ularizers. Moreover, TLFre has a very low computational cost and can be integrat
ed with any existing solvers. Experiments on both synthetic and real data sets s
how that TLFre improves the efficiency of SGL by orders of magnitude.
****

The Large Margin Mechanism for Differentially Private Maximization
Kamalika Chaudhuri, Daniel J. Hsu, Shuang Song
A basic problem in the design of privacy-preserving algorithms is the \emph{priv
ate maximization problem}: the goal is to pick an item from a universe that (app
roximately) maximizes a data-dependent function, all under the constraint of dif
ferential privacy. This problem has been used as a sub-routine in many privacy-p
reserving algorithms for statistics and machine learning. Previous algorithms fo
r this problem are either range-dependent---i.e., their utility diminishes with
the size of the universe---or only apply to very restricted function classes. Th
is work provides the first general purpose, range-independent algorithm for priv
ate maximization that guarantees approximate differential privacy. Its applicabi
lity is demonstrated on two fundamental tasks in data mining and machine learnin

g.
****

Causal Inference through a Witness Protection Program
Ricardo Silva, Robin Evans
One of the most fundamental problems in causal inference is the estimation of a causal effect when variables are confounded. This is difficult in an observational study because one has no direct evidence that all confounders have been adjusted for. We introduce a novel approach for estimating causal effects that exploits observational conditional independencies to suggest weak'' paths in a unknown causal graph. The widely used faithfulness condition of Spirtes et al. is relaxed to allow for varying degrees ofpath cancellations'' that will imply conditional independencies but do not rule out the existence of confounding causal paths. The outcome is a posterior distribution over bounds on the average causal effect via a linear programming approach and Bayesian inference. We claim this approach should be used in regular practice to complement other default tools in observational studies.
****

Self-Adaptable Templates for Feature Coding
Xavier Boix, Gemma Roig, Salomon Diether, Luc V. Gool
Hierarchical feed-forward networks have been successfully applied in object recognition. At each level of the hierarchy, features are extracted and encoded, followed by a pooling step. Within this processing pipeline, the common trend is to learn the feature coding templates, often referred as codebook entries, filters, or over-complete basis. Recently, an approach that apparently does not use templates has been shown to obtain very promising results. This is the second-order pooling (O2P). In this paper, we analyze O2P as a coding-pooling scheme. We find that at testing phase, O2P automatically adapts the feature coding templates to the input features, rather than using templates learned during the training phase. From this finding, we are able to bring common concepts of coding-pooling schemes to O2P, such as feature quantization. This allows for significant accuracy improvements of O2P in standard benchmarks of image classification, namely Caltech101 and VOC07.
****

A Framework for Testing Identifiability of Bayesian Models of Perception
Luigi Acerbi, Wei Ji Ma, Sethu Vijayakumar
Bayesian observer models are very effective in describing human performance in perceptual tasks, so much so that they are trusted to faithfully recover hidden mental representations of priors, likelihoods, or loss functions from the data. However, the intrinsic degeneracy of the Bayesian framework, as multiple combinations of elements can yield empirically indistinguishable results, prompts the question of model identifiability. We propose a novel framework for a systematic testing of the identifiability of a significant class of Bayesian observer models, with practical applications for improving experimental design. We examine the theoretical identifiability of the inferred internal representations in two case studies. First, we show which experimental designs work better to remove the underlying degeneracy in a time interval estimation task. Second, we find that the reconstructed representations in a speed perception task under a slow-speed prior are fairly robust.
****

Low Rank Approximation Lower Bounds in Row-Update Streams
David Woodruff
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.
****

Probabilistic ODE Solvers with Runge-Kutta Means
Michael Schober, David K. Duvenaud, Philipp Hennig
Runge-Kutta methods are the classic family of solvers for ordinary differential equations (ODEs), and the basis for the state of the art. Like most numerical me

thods, they return point estimates. We construct a family of probabilistic numer
ical methods that instead return a Gauss-Markov process defining a probability d
istribution over the ODE solution. In contrast to prior work, we construct this
family such that posterior means match the outputs of the Runge-Kutta family exa
ctly, thus inheriting their proven good properties. Remaining degrees of freedom
 not identified by the match to Runge-Kutta are chosen such that the posterior p
robability measure fits the observed structure of the ODE. Our results shed ligh
t on the structure of Runge-Kutta solvers from a new direction, provide a richer
, probabilistic output, have low computational cost, and raise new research ques
tions.
****

Learning a Concept Hierarchy from Multi-labeled Documents
Viet-An Nguyen, Jordan L. Ying, Philip Resnik, Jonathan Chang
While topic models can discover patterns of word usage in large corpora, it is d
ifficult to meld this unsupervised structure with noisy, human-provided labels,
especially when the label space is large. In this paper, we present a model-Labe
l to Hierarchy (L2H)-that can induce a hierarchy of user-generated labels and th
e topics associated with those labels from a set of multi-labeled documents. The
 model is robust enough to account for missing labels from untrained, disparate
annotators and provide an interpretable summary of an otherwise unwieldy label s
et. We show empirically the effectiveness of L2H in predicting held-out words an
d labels for unseen documents.
****

Dependent nonparametric trees for dynamic hierarchical clustering
Kumar Avinava Dubey, Qirong Ho, Sinead A. Williamson, Eric P. Xing
Hierarchical clustering methods offer an intuitive and powerful way to model a w
ide variety of data sets. However, the assumption of a fixed hierarchy is often
overly restrictive when working with data generated over a period of time: We ex
pect both the structure of our hierarchy, and the parameters of the clusters, to
 evolve with time. In this paper, we present a distribution over collections of
time-dependent, infinite-dimensional trees that can be used to model evolving hi
erarchies, and present an efficient and scalable algorithm for performing approx
imate inference in such a model. We demonstrate the efficacy of our model and in
ference algorithm on both synthetic data and real-world document corpora.
****

A Statistical Decision-Theoretic Framework for Social Choice
Hossein Azari Soufiani, David C. Parkes, Lirong Xia
In this paper, we take a statistical decision-theoretic viewpoint on social choi
ce, putting a focus on the decision to be made on behalf of a system of agents.
In our framework, we are given a statistical ranking model, a decision space, an
d a loss function defined on (parameter, decision) pairs, and formulate social c
hoice mechanisms as decision rules that minimize expected loss. This suggests a
general framework for the design and analysis of new social choice mechanisms. W
e compare Bayesian estimators, which minimize Bayesian expected loss, for the Ma
llows model and the Condorcet model respectively, and the Kemeny rule. We consid
er various normative properties, in addition to computational complexity and asy
mptotic behavior. In particular, we show that the Bayesian estimator for the Con
dorcet model satisfies some desired properties such as anonymity, neutrality, an
d monotonicity, can be computed in polynomial time, and is asymptotically differ
ent from the other two rules when the data are generated from the Condorcet mode
l for some ground truth parameter.
****

Generative Adversarial Nets
Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sh
erjil Ozair, Aaron Courville, Yoshua Bengio
We propose a new framework for estimating generative models via adversarial nets
, in which we simultaneously train two models: a generative model G that capture
s the data distribution, and a discriminative model D that estimates the probabi
lity that a sample came from the training data rather than G. The training proce
dure for G is to maximize the probability of D making a mistake. This framework

corresponds to a minimax two-player game. In the space of arbitrary functions G and D, a unique solution exists, with G recovering the training data distributio n and D equal to 1/2 everywhere. In the case where G and D are defined by multil ayer perceptrons, the entire system can be trained with backpropagation. There i s no need for any Markov chains or unrolled approximate inference networks durin g either training or generation of samples. Experiments demonstrate the potentia l of the framework through qualitative and quantitatively evaluation of the gene rated samples.
****

Delay-Tolerant Algorithms for Asynchronous Distributed Online Learning
Brendan McMahan, Matthew Streeter
We analyze new online gradient descent algorithms for distributed systems with l arge delays between gradient computations and the corresponding updates. Using i nsights from adaptive gradient methods, we develop algorithms that adapt not onl y to the sequence of gradients, but also to the precise update delays that occur . We first give an impractical algorithm that achieves a regret bound that preci sely quantifies the impact of the delays. We then analyze AdaptiveRevision, an a lgorithm that is efficiently implementable and achieves comparable guarantees. T he key algorithmic technique is appropriately and efficiently revising the learn ing rate used for previous gradient steps. Experimental results show when the de lays grow large (1000 updates or more), our new algorithms perform significantly better than standard adaptive gradient methods.
****

Sequential Monte Carlo for Graphical Models
Christian Andersson Naesseth, Fredrik Lindsten, Thomas B. Schön
We propose a new framework for how to use sequential Monte Carlo (SMC) algorithm s for inference in probabilistic graphical models (PGM). Via a sequential decomp osition of the PGM we find a sequence of auxiliary distributions defined on a mo notonically increasing sequence of probability spaces. By targeting these auxili ary distributions using SMC we are able to approximate the full joint distributi on defined by the PGM. One of the key merits of the SMC sampler is that it provi des an unbiased estimate of the partition function of the model. We also show ho w it can be used within a particle Markov chain Monte Carlo framework in order t o construct high-dimensional block-sampling algorithms for general PGMs.
****

Learning Time-Varying Coverage Functions
Nan Du, Yingyu Liang, Maria-Florina F. Balcan, Le Song
Coverage functions are an important class of discrete functions that capture law s of diminishing returns. In this paper, we propose a new problem of learning ti me-varying coverage functions which arise naturally from applications in social network analysis, machine learning, and algorithmic game theory. We develop a no vel parametrization of the time-varying coverage function by illustrating the co nnections with counting processes. We present an efficient algorithm to learn th e parameters by maximum likelihood estimation, and provide a rigorous theoretic analysis of its sample complexity. Empirical experiments from information diffus ion in social network analysis demonstrate that with few assumptions about the u nderlying diffusion process, our method performs significantly better than exist ing approaches on both synthetic and real world data.
****

Learning Shuffle Ideals Under Restricted Distributions
Dongqu Chen
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-auth ors prior to requesting a name change in the electronic proceedings.
****

Spectral Clustering of graphs with the Bethe Hessian
Alaa Saade, Florent Krzakala, Lenka Zdeborová
Spectral clustering is a standard approach to label nodes on a graph by studying the (largest or lowest) eigenvalues of a symmetric real matrix such as e.g. the

adjacency or the Laplacian. Recently, it has been argued that using instead a m
ore complicated, non-symmetric and higher dimensional operator, related to the n
on-backtracking walk on the graph, leads to improved performance in detecting cl
usters, and even to optimal performance for the stochastic block model. Here, we
 propose to use instead a simpler object, a symmetric real matrix known as the B
ethe Hessian operator, or deformed Laplacian. We show that this approach combine
s the performances of the non-backtracking operator, thus detecting clusters all
 the way down to the theoretical limit in the stochastic block model, with the c
omputational, theoretical and memory advantages of real symmetric matrices.
****

Optimal Regret Minimization in Posted-Price Auctions with Strategic Buyers
Mehryar Mohri, Andres Munoz
Requests for name changes in the electronic proceedings will be accepted with no
 questions asked. However name changes may cause bibliographic tracking issues.
  Authors are asked to consider this carefully and discuss it with their co-auth
ors prior to requesting a name change in the electronic proceedings.
****

Fundamental Limits of Online and Distributed Algorithms for Statistical Learning
 and Estimation
Ohad Shamir
Many machine learning approaches are characterized by information constraints on
 how they interact with the training data. These include memory and sequential a
ccess constraints (e.g. fast first-order methods to solve stochastic optimizatio
n problems); communication constraints (e.g. distributed learning); partial acce
ss to the underlying data (e.g. missing features and multi-armed bandits) and mo
re. However, currently we have little understanding how such information constra
ints fundamentally affect our performance, independent of the learning problem s
emantics. For example, are there learning problems where any algorithm which has
 small memory footprint (or can use any bounded number of bits from each example
, or has certain communication constraints) will perform worse than what is poss
ible without such constraints? In this paper, we describe how a single set of re
sults implies positive answers to the above, for several different settings.
****

Repeated Contextual Auctions with Strategic Buyers
Kareem Amin, Afshin Rostamizadeh, Umar Syed
Motivated by real-time advertising exchanges, we analyze the problem of pricing
 inventory in a repeated posted-price auction. We consider both the cases of a tr
uthful and surplus-maximizing buyer, where the former makes decisions myopically
 on every round, and the latter may strategically react to our algorithm, forgoi
ng short-term surplus in order to trick the algorithm into setting better prices
 in the future. We further assume a buyer’s valuation of a good is a function of
 a context vector that describes the good being sold. We give the first algorith
m attaining sublinear (O(T^{2/3})) regret in the contextual setting against a su
rplus-maximizing buyer. We also extend this result to repeated second-price auct
ions with multiple buyers.
****

Learning with Pseudo-Ensembles
Philip Bachman, Ouais Alsharif, Doina Precup
We formalize the notion of a pseudo-ensemble, a (possibly infinite) collection o
f child models spawned from a parent model by perturbing it according to some no
ise process. E.g., dropout (Hinton et al, 2012) in a deep neural network trains
a pseudo-ensemble of child subnetworks generated by randomly masking nodes in th
e parent network. We examine the relationship of pseudo-ensembles, which involve
 perturbation in model-space, to standard ensemble methods and existing notions
of robustness, which focus on perturbation in observation-space. We present a no
vel regularizer based on making the behavior of a pseudo-ensemble robust with re
spect to the noise process generating it. In the fully-supervised setting, our r
egularizer matches the performance of dropout. But, unlike dropout, our regulari
zer naturally extends to the semi-supervised setting, where it produces state-of
-the-art results. We provide a case study in which we transform the Recursive Ne

ural Tensor Network of (Socher et al, 2013) into a pseudo-ensemble, which signif
icantly improves its performance on a real-world sentiment analysis benchmark.
****

Top Rank Optimization in Linear Time
Nan Li, Rong Jin, Zhi-Hua Zhou
Bipartite ranking aims to learn a real-valued ranking function that orders posit
ive instances before negative instances. Recent efforts of bipartite ranking are
 focused on optimizing ranking accuracy at the top of the ranked list. Most exis
ting approaches are either to optimize task specific metrics or to extend the ra
nk loss by emphasizing more on the error associated with the top ranked instance
s, leading to a high computational cost that is super-linear in the number of tr
aining instances. We propose a highly efficient approach, titled TopPush, for op
timizing accuracy at the top that has computational complexity linear in the num
ber of training instances. We present a novel analysis that bounds the generaliz
ation error for the top ranked instances for the proposed approach. Empirical st
udy shows that the proposed approach is highly competitive to the state-of-the-a
rt approaches and is 10-100 times faster.
****

Learning Neural Network Policies with Guided Policy Search under Unknown Dynamic
s
Sergey Levine, Pieter Abbeel
We present a policy search method that uses iteratively refitted local linear mo
dels to optimize trajectory distributions for large, continuous problems. These
trajectory distributions can be used within the framework of guided policy searc
h to learn policies with an arbitrary parameterization. Our method fits time-var
ying linear dynamics models to speed up learning, but does not rely on learning
a global model, which can be difficult when the dynamics are complex and discont
inuous. We show that this hybrid approach requires many fewer samples than model
-free methods, and can handle complex, nonsmooth dynamics that can pose a challe
nge for model-based techniques. We present experiments showing that our method c
an be used to learn complex neural network policies that successfully execute si
mulated robotic manipulation tasks in partially observed environments with numer
ous contact discontinuities and underactuation.
****

Optimizing F-Measures by Cost-Sensitive Classification
Shameem Puthiya Parambath, Nicolas Usunier, Yves Grandvalet
We present a theoretical analysis of F-measures for binary, multiclass and multi
label classification. These performance measures are non-linear, but in many sce
narios they are pseudo-linear functions of the per-class false negative/false po
sitive rate. Based on this observation, we present a general reduction of F-meas
ure maximization to cost-sensitive classification with unknown costs. We then pr
opose an algorithm with provable guarantees to obtain an approximately optimal c
lassifier for the F-measure by solving a series of cost-sensitive classification
 problems. The strength of our analysis is to be valid on any dataset and any cl
ass of classifiers, extending the existing theoretical results on F-measures, wh
ich are asymptotic in nature. We present numerical experiments to illustrate the
 relative importance of cost asymmetry and thresholding when learning linear cla
ssifiers on various F-measure optimization tasks.
****

Distributed Power-law Graph Computing: Theoretical and Empirical Analysis
Cong Xie, Ling Yan, Wu-Jun Li, Zhihua Zhang
With the emergence of big graphs in a variety of real applications like social n
etworks, machine learning based on distributed graph-computing~(DGC) frameworks
has attracted much attention from big data machine learning community. In DGC fr
ameworks, the graph partitioning~(GP) strategy plays a key role to affect the pe
rformance, including the workload balance and communication cost. Typically, the
 degree distributions of natural graphs from real applications follow skewed pow
er laws, which makes GP a challenging task. Recently, many methods have been pro
posed to solve the GP problem. However, the existing GP methods cannot achieve s
atisfactory performance for applications with power-law graphs. In this paper, w

e propose a novel vertex-cut method, called \emph{degree-based hashing}~(DBH), f
or GP. DBH makes effective use of the skewed degree distributions for GP. We the
oretically prove that DBH can achieve lower communication cost than existing met
hods and can simultaneously guarantee good workload balance. Furthermore, empiri
cal results on several large power-law graphs also show that DBH can outperform
the state of the art.
****

Parallel Successive Convex Approximation for Nonsmooth Nonconvex Optimization
Meisam Razaviyayn, Mingyi Hong, Zhi-Quan Luo, Jong-Shi Pang
Consider the problem of minimizing the sum of a smooth (possibly non-convex) and
 a convex (possibly nonsmooth) function involving a large number of variables. A
 popular approach to solve this problem is the block coordinate descent (BCD) me
thod whereby at each iteration only one variable block is updated while the rema
ining variables are held fixed. With the recent advances in the developments of
the multi-core parallel processing technology, it is desirable to parallelize th
e BCD method by allowing multiple blocks to be updated simultaneously at each it
eration of the algorithm. In this work, we propose an inexact parallel BCD appro
ach where at each iteration, a subset of the variables is updated in parallel by
 minimizing convex approximations of the original objective function. We investi
gate the convergence of this parallel BCD method for both randomized and cyclic
variable selection rules. We analyze the asymptotic and non-asymptotic convergen
ce behavior of the algorithm for both convex and non-convex objective functions.
 The numerical experiments suggest that for a special case of Lasso minimization
 problem, the cyclic block selection rule can outperform the randomized rule.
****

Active Regression by Stratification
Sivan Sabato, Remi Munos
We propose a new active learning algorithm for parametric linear regression with
 random design. We provide finite sample convergence guarantees for general dist
ributions in the misspecified model. This is the first active learner for this s
etting that provably can improve over passive learning. Unlike other learning se
ttings (such as classification), in regression the passive learning rate of O(1/
epsilon) cannot in general be improved upon. Nonetheless, the so-called `constan
t' in the rate of convergence, which is characterized by a distribution-dependen
t risk, can be improved in many cases. For a given distribution, achieving the o
ptimal risk requires prior knowledge of the distribution. Following the stratifi
cation technique advocated in Monte-Carlo function integration, our active learn
er approaches a the optimal risk using piecewise constant approximations.
****

Distance-Based Network Recovery under Feature Correlation
David Adametz, Volker Roth
We present an inference method for Gaussian graphical models when only pairwise
distances of n objects are observed. Formally, this is a problem of estimating a
n n x n covariance matrix from the Mahalanobis distances dMH(xi, xj), where obje
ct xi lives in a latent feature space. We solve the problem in fully Bayesian fa
shion by integrating over the Matrix-Normal likelihood and a Matrix-Gamma prior;
 the resulting Matrix-T posterior enables network recovery even under strongly c
orrelated features. Hereby, we generalize TiWnet, which assumes Euclidean distan
ces with strict feature independence. In spite of the greatly increased flexibil
ity, our model neither loses statistical power nor entails more computational co
st. We argue that the extension is highly relevant as it yields significantly be
tter results in both synthetic and real-world experiments, which is successfully
 demonstrated for a network of biological pathways in cancer patients.
****

Rounding-based Moves for Metric Labeling
M. Pawan Kumar
Metric labeling is a special case of energy minimization for pairwise Markov ran
dom fields. The energy function consists of arbitrary unary potentials, and pair
wise potentials that are proportional to a given metric distance function over t
he label set. Popular methods for solving metric labeling include (i) move-makin

g algorithms, which iteratively solve a minimum st-cut problem; and (ii) the lin
ear programming (LP) relaxation based approach. In order to convert the fraction
al solution of the LP relaxation to an integer solution, several randomized roun
ding procedures have been developed in the literature. We consider a large class
 of parallel rounding procedures, and design move-making algorithms that closely
 mimic them. We prove that the multiplicative bound of a move-making algorithm e
xactly matches the approximation factor of the corresponding rounding procedure
for any arbitrary distance function. Our analysis includes all known results for
 move-making algorithms as special cases.
****

Transportability from Multiple Environments with Limited Experiments: Completene
ss Results
Elias Bareinboim, Judea Pearl
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.
****

Fast and Robust Least Squares Estimation in Corrupted Linear Models
Brian McWilliams, Gabriel Krummenacher, Mario Lucic, Joachim M. Buhmann
Subsampling methods have been recently proposed to speed up least squares estima
tion in large scale settings. However, these algorithms are typically not robust
 to outliers or corruptions in the observed covariates. The concept of influence
 that was developed for regression diagnostics can be used to detect such corrup
ted observations as shown in this paper. This property of influence -- for which
 we also develop a randomized approximation -- motivates our proposed subsamplin
g algorithm for large scale corrupted linear regression which limits the influen
ce of data points since highly influential points contribute most to the residua
l error. Under a general model of corrupted observations, we show theoretically
and empirically on a variety of simulated and real datasets that our algorithm i
mproves over the current state-of-the-art approximation schemes for ordinary lea
st squares.
****

Incremental Local Gaussian Regression
Franziska Meier, Philipp Hennig, Stefan Schaal
Locally weighted regression (LWR) was created as a nonparametric method that can
 approximate a wide range of functions, is computationally efficient, and can le
arn continually from very large amounts of incrementally collected data. As an i
nteresting feature, LWR can regress on non-stationary functions, a beneficial pr
operty, for instance, in control problems. However, it does not provide a proper
 generative model for function values, and existing algorithms have a variety of
 manual tuning parameters that strongly influence bias, variance and learning sp
eed of the results. Gaussian (process) regression, on the other hand, does provi
de a generative model with rather black-box automatic parameter tuning, but it h
as higher computational cost, especially for big data sets and if a non-stationa
ry model is required. In this paper, we suggest a path from Gaussian (process) r
egression to locally weighted regression, where we retain the best of both appro
aches. Using a localizing function basis and approximate inference techniques, w
e build a Gaussian (process) regression algorithm of increasingly local nature a
nd similar computational complexity to LWR. Empirical evaluations are performed
on several synthetic and real robot datasets of increasing complexity and (big)
data scale, and demonstrate that we consistently achieve on par or superior perf
ormance compared to current state-of-the-art methods while retaining a principle
d approach to fast incremental regression with minimal manual tuning parameters.
****

Controlling privacy in recommender systems
Yu Xin, Tommi Jaakkola
Recommender systems involve an inherent trade-off between accuracy of recommenda
tions and the extent to which users are willing to release information about the
ir preferences. In this paper, we explore a two-tiered notion of privacy where t

here is a small set of public'' users who are willing to share their preferences openly, and a large set ofprivate'' users who require privacy guarantees. We show theoretically and demonstrate empirically that a moderate number of public users with no access to private user information already suffices for reasonable accuracy. Moreover, we introduce a new privacy concept for gleaning relational information from private users while maintaining a first order deniability. We demonstrate gains from controlled access to private user preferences.
****

Localized Data Fusion for Kernel k-Means Clustering with Application to Cancer Biology
Mehmet Gönen, Adam A. Margolin
In many modern applications from, for example, bioinformatics and computer vision, samples have multiple feature representations coming from different data sources. Multiview learning algorithms try to exploit all these available information to obtain a better learner in such scenarios. In this paper, we propose a novel multiple kernel learning algorithm that extends kernel k-means clustering to the multiview setting, which combines kernels calculated on the views in a localized way to better capture sample-specific characteristics of the data. We demonstrate the better performance of our localized data fusion approach on a human colon and rectal cancer data set by clustering patients. Our method finds more relevant prognostic patient groups than global data fusion methods when we evaluate the results with respect to three commonly used clinical biomarkers.
****

Object Localization based on Structural SVM using Privileged Information
Jan Feyereisl, Suha Kwak, Jeany Son, Bohyung Han
We propose a structured prediction algorithm for object localization based on Support Vector Machines (SVMs) using privileged information. Privileged information provides useful high-level knowledge for image understanding and facilitates learning a reliable model even with a small number of training examples. In our setting, we assume that such information is available only at training time since it may be difficult to obtain from visual data accurately without human supervision. Our goal is to improve performance by incorporating privileged information into ordinary learning framework and adjusting model parameters for better generalization. We tackle object localization problem based on a novel structural SVM using privileged information, where an alternating loss-augmented inference procedure is employed to handle the term in the objective function corresponding to privileged information. We apply the proposed algorithm to the Caltech-UCSD Birds 200-2011 dataset, and obtain encouraging results suggesting further investigation into the benefit of privileged information in structured prediction.
****

Robust Logistic Regression and Classification
Jiashi Feng, Huan Xu, Shie Mannor, Shuicheng Yan
We consider logistic regression with arbitrary outliers in the covariate matrix. We propose a new robust logistic regression algorithm, called RoLR, that estimates the parameter through a simple linear programming procedure. We prove that RoLR is robust to a constant fraction of adversarial outliers. To the best of our knowledge, this is the first result on estimating logistic regression model when the covariate matrix is corrupted with any performance guarantees. Besides regression, we apply RoLR to solving binary classification problems where a fraction of training samples are corrupted.
****

Flexible Transfer Learning under Support and Model Shift
Xuezhi Wang, 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.
****

Computing Nash Equilibria in Generalized Interdependent Security Games
Hau Chan, Luis E. Ortiz

We study the computational complexity of computing Nash equilibria in generalize
d interdependent-security (IDS) games. Like traditional IDS games, originally in
troduced by economists and risk-assessment experts Heal and Kunreuther about a d
ecade ago, generalized IDS games model agents' voluntary investment decisions wh
en facing potential direct risk and transfer risk exposure from other agents. A
distinct feature of generalized IDS games, however, is that full investment can
reduce transfer risk. As a result, depending on the transfer-risk reduction leve
l, generalized IDS games may exhibit strategic complementarity (SC) or strategic
 substitutability (SS). We consider three variants of generalized IDS games in w
hich players exhibit only SC, only SS, and both SC+SS. We show that determining
whether there is a pure-strategy Nash equilibrium (PSNE) in SC+SS-type games is
NP-complete, while computing a single PSNE in SC-type games takes worst-case pol
ynomial time. As for the problem of computing all mixed-strategy Nash equilibria
 (MSNE) efficiently, we produce a partial characterization. Whenever each agent
in the game is indiscriminate in terms of the transfer-risk exposure to the othe
r agents, a case that Kearns and Ortiz originally studied in the context of trad
itional IDS games in their NIPS 2003 paper, we can compute all MSNE that satisfy
 some ordering constraints in polynomial time in all three game variants. Yet, t
here is a computational barrier in the general (transfer) case: we show that the
 computational problem is as hard as the Pure-Nash-Extension problem, also origi
nally introduced by Kearns and Ortiz, and that it is NP complete for all three v
ariants. Finally, we experimentally examine and discuss the practical impact tha
t the additional protection from transfer risk allowed in generalized IDS games
has on MSNE by solving several randomly-generated instances of SC+SS-type games
with graph structures taken from several real-world datasets.
****

Multitask learning meets tensor factorization: task imputation via convex optimi
zation

Kishan Wimalawarne, Masashi Sugiyama, Ryota Tomioka

We study a multitask learning problem in which each task is parametrized by a we
ight vector and indexed by a pair of indices, which can be e.g, (consumer, time)
. The weight vectors can be collected into a tensor and the (multilinear-)rank o
f the tensor controls the amount of sharing of information among tasks. Two type
s of convex relaxations have recently been proposed for the tensor multilinear r
ank. However, we argue that both of them are not optimal in the context of multi
task learning in which the dimensions or multilinear rank are typically heteroge
neous. We propose a new norm, which we call the scaled latent trace norm and ana
lyze the excess risk of all the three norms. The results apply to various settin
gs including matrix and tensor completion, multitask learning, and multilinear m
ultitask learning. Both the theory and experiments support the advantage of the
new norm when the tensor is not equal-sized and we do not a priori know which mo
de is low rank.
****

Mind the Nuisance: Gaussian Process Classification using Privileged Noise

Daniel Hernández-lobato, Viktoriia Sharmanska, Kristian Kersting, Christoph H. L
ampert, Novi Quadrianto

The learning with privileged information setting has recently attracted a lot of
 attention within the machine learning community, as it allows the integration o
f additional knowledge into the training process of a classifier, even when this
 comes in the form of a data modality that is not available at test time. Here,
we show that privileged information can naturally be treated as noise in the lat
ent function of a Gaussian process classifier (GPC). That is, in contrast to the
 standard GPC setting, the latent function is not just a nuisance but a feature:
 it becomes a natural measure of confidence about the training data by modulatin
g the slope of the GPC probit likelihood function. Extensive experiments on publ
ic datasets show that the proposed GPC method using privileged noise, called GPC
+, improves over a standard GPC without privileged knowledge, and also over the
current state-of-the-art SVM-based method, SVM+. Moreover, we show that advanced
 neural networks and deep learning methods can be compressed as privileged infor
mation.

****
On Integrated Clustering and Outlier Detection
Lionel Ott, Linsey Pang, Fabio T. Ramos, Sanjay Chawla
We model the joint clustering and outlier detection problem using an extension o f the facility location formulation. The advantages of combining clustering and outlier selection include: (i) the resulting clusters tend to be compact and sem antically coherent (ii) the clusters are more robust against data perturbations and (iii) the outliers are contextualised by the clusters and more interpretable . We provide a practical subgradient-based algorithm for the problem and also st udy the theoretical properties of algorithm in terms of approximation and conver gence. Extensive evaluation on synthetic and real data sets attest to both the q uality and scalability of our proposed method.
****
Latent Support Measure Machines for Bag-of-Words Data Classification
Yuya Yoshikawa, Tomoharu Iwata, Hiroshi Sawada
In many classification problems, the input is represented as a set of features, e.g., the bag-of-words (BoW) representation of documents. Support vector machine s (SVMs) are widely used tools for such classification problems. The performance of the SVMs is generally determined by whether kernel values between data point s can be defined properly. However, SVMs for BoW representations have a major we akness in that the co-occurrence of different but semantically similar words can not be reflected in the kernel calculation. To overcome the weakness, we propose a kernel-based discriminative classifier for BoW data, which we call the latent support measure machine (latent SMM). With the latent SMM, a latent vector is a ssociated with each vocabulary term, and each document is represented as a distr ibution of the latent vectors for words appearing in the document. To represent the distributions efficiently, we use the kernel embeddings of distributions tha t hold high order moment information about distributions. Then the latent SMM fi nds a separating hyperplane that maximizes the margins between distributions of different classes while estimating latent vectors for words to improve the class ification performance. In the experiments, we show that the latent SMM achieves state-of-the-art accuracy for BoW text classification, is robust with respect to its own hyper-parameters, and is useful to visualize words.
****
Sparse Polynomial Learning and Graph Sketching
Murat Kocaoglu, Karthikeyan Shanmugam, Alexandros G. Dimakis, Adam Klivans
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.
****
The Noisy Power Method: A Meta Algorithm with Applications
Moritz Hardt, Eric Price
We provide a new robust convergence analysis of the well-known power method for computing the dominant singular vectors of a matrix that we call noisy power met hod. Our result characterizes the convergence behavior of the algorithm when a l arge amount noise is introduced after each matrix-vector multiplication. The noi sy power method can be seen as a meta-algorithm that has recently found a number of important applications in a broad range of machine learning problems includi ng alternating minimization for matrix completion, streaming principal component analysis (PCA), and privacy-preserving spectral analysis. Our general analysis subsumes several existing ad-hoc convergence bounds and resolves a number of ope n problems in multiple applications. A recent work of Mitliagkas et al.~(NIPS 20 13) gives a space-efficient algorithm for PCA in a streaming model where samples are drawn from a spiked covariance model. We give a simpler and more general an alysis that applies to arbitrary distributions. Moreover, even in the spiked cov ariance model our result gives quantitative improvements in a natural parameter regime. As a second application, we provide an algorithm for differentially priv ate principal component analysis that runs in nearly linear time in the input sp arsity and achieves nearly tight worst-case error bounds. Complementing our wors

t-case bounds, we show that the error dependence of our algorithm on the matrix dimension can be replaced by an essentially tight dependence on the coherence of the matrix. This result resolves the main problem left open by Hardt and Roth (STOC 2013) and leads to strong average-case improvements over the optimal worst-case bound.

****

Robust Tensor Decomposition with Gross Corruption

Quanquan Gu, Huan Gui, Jiawei Han

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.

****

RAAM: The Benefits of Robustness in Approximating Aggregated MDPs in Reinforcement Learning

Marek Petrik, Dharmashankar Subramanian

We describe how to use robust Markov decision processes for value function approximation with state aggregation. The robustness serves to reduce the sensitivity to the approximation error of sub-optimal policies in comparison to classical methods such as fitted value iteration. This results in reducing the bounds on the gamma-discounted infinite horizon performance loss by a factor of 1/(1-gamma) while preserving polynomial-time computational complexity. Our experimental results show that using the robust representation can significantly improve the solution quality with minimal additional computational cost.

****

On Prior Distributions and Approximate Inference for Structured Variables

Oluwasanmi O. Koyejo, Rajiv Khanna, Joydeep Ghosh, Russell Poldrack

We present a general framework for constructing prior distributions with structured variables. The prior is defined as the information projection of a base distribution onto distributions supported on the constraint set of interest. In cases where this projection is intractable, we propose a family of parameterized approximations indexed by subsets of the domain. We further analyze the special case of sparse structure. While the optimal prior is intractable in general, we show that approximate inference using convex subsets is tractable, and is equivalent to maximizing a submodular function subject to cardinality constraints. As a result, inference using greedy forward selection provably achieves within a factor of (1-1/e) of the optimal objective value. Our work is motivated by the predictive modeling of high-dimensional functional neuroimaging data. For this task, we employ the Gaussian base distribution induced by local partial correlations and consider the design of priors to capture the domain knowledge of sparse support. Experimental results on simulated data and high dimensional neuroimaging data show the effectiveness of our approach in terms of support recovery and predictive accuracy.

****

A Residual Bootstrap for High-Dimensional Regression with Near Low-Rank Designs

Miles Lopes

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.

****

Tighten after Relax: Minimax-Optimal Sparse PCA in Polynomial Time

Zhaoran Wang, Huanran Lu, 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.

****

Learning to Search in Branch and Bound Algorithms

He He, Hal Daume III, Jason M. Eisner

Branch-and-bound is a widely used method in combinatorial optimization, includin
g mixed integer programming, structured prediction and MAP inference. While most
 work has been focused on developing problem-specific techniques, little is know
n about how to systematically design the node searching strategy on a branch-and
-bound tree. We address the key challenge of learning an adaptive node searching
 order for any class of problem solvable by branch-and-bound. Our strategies are
 learned by imitation learning. We apply our algorithm to linear programming bas
ed branch-and-bound for solving mixed integer programs (MIP). We compare our met
hod with one of the fastest open-source solvers, SCIP; and a very efficient comm
ercial solver, Gurobi. We demonstrate that our approach achieves better solution
s faster on four MIP libraries.
****

Bayesian Inference for Structured Spike and Slab Priors
Michael R. Andersen, Ole Winther, Lars K. Hansen
Sparse signal recovery addresses the problem of solving underdetermined linear i
nverse problems subject to a sparsity constraint. We propose a novel prior formu
lation, the structured spike and slab prior, which allows to incorporate a prior
i knowledge of the sparsity pattern by imposing a spatial Gaussian process on th
e spike and slab probabilities. Thus, prior information on the structure of the
sparsity pattern can be encoded using generic covariance functions. Furthermore,
 we provide a Bayesian inference scheme for the proposed model based on the expe
ctation propagation framework. Using numerical experiments on synthetic data, we
 demonstrate the benefits of the model.
****

Just-In-Time Learning for Fast and Flexible Inference
S. M. Ali Eslami, Daniel Tarlow, Pushmeet Kohli, John Winn
Much of research in machine learning has centered around the search for inferenc
e algorithms that are both general-purpose and efficient. The problem is extreme
ly challenging and general inference remains computationally expensive. We seek
to address this problem by observing that in most specific applications of a mod
el, we typically only need to perform a small subset of all possible inference c
omputations. Motivated by this, we introduce just-in-time learning, a framework
for fast and flexible inference that learns to speed up inference at run-time. T
hrough a series of experiments, we show how this framework can allow us to combi
ne the flexibility of sampling with the efficiency of deterministic message-pass
ing.
****

Fast Kernel Learning for Multidimensional Pattern Extrapolation
Andrew G. Wilson, Elad Gilboa, Arye Nehorai, John P. Cunningham
The ability to automatically discover patterns and perform extrapolation is an e
ssential quality of intelligent systems. Kernel methods, such as Gaussian proces
ses, have great potential for pattern extrapolation, since the kernel flexibly a
nd interpretably controls the generalisation properties of these methods. Howeve
r, automatically extrapolating large scale multidimensional patterns is in gener
al difficult, and developing Gaussian process models for this purpose involves s
everal challenges. A vast majority of kernels, and kernel learning methods, curr
ently only succeed in smoothing and interpolation. This difficulty is compounded
 by the fact that Gaussian processes are typically only tractable for small data
sets, and scaling an expressive kernel learning approach poses different challen
ges than scaling a standard Gaussian process model. One faces additional computa
tional constraints, and the need to retain significant model structure for expre
ssing the rich information available in a large dataset. In this paper, we propo
se a Gaussian process approach for large scale multidimensional pattern extrapol
ation. We recover sophisticated out of class kernels, perform texture extrapolat
ion, inpainting, and video extrapolation, and long range forecasting of land sur
face temperatures, all on large multidimensional datasets, including a problem w
ith 383,400 training points. The proposed method significantly outperforms alter
native scalable and flexible Gaussian process methods, in speed and accuracy. Mo
reover, we show that a distinct combination of expressive kernels, a fully non-p
arametric representation, and scalable inference which exploits existing model s

tructure, are critical for large scale multidimensional pattern extrapolation.
****

Recursive Context Propagation Network for Semantic Scene Labeling
Abhishek Sharma, Oncel Tuzel, Ming-Yu Liu

We propose a deep feed-forward neural network architecture for pixel-wise semant ic scene labeling. It uses a novel recursive neural network architecture for con text propagation, referred to as rCPN. It first maps the local visual features i nto a semantic space followed by a bottom-up aggregation of local information in to a global representation of the entire image. Then a top-down propagation of t he aggregated information takes place that enhances the contextual information o f each local feature. Therefore, the information from every location in the imag e is propagated to every other location. Experimental results on Stanford backgr ound and SIFT Flow datasets show that the proposed method outperforms previous a pproaches. It is also orders of magnitude faster than previous methods and takes only 0.07 seconds on a GPU for pixel-wise labeling of a 256x256 image starting from raw RGB pixel values, given the super-pixel mask that takes an additional 0 .3 seconds using an off-the-shelf implementation.
****

Spectral Methods meet EM: A Provably Optimal Algorithm for Crowdsourcing
Yuchen Zhang, Xi Chen, Dengyong Zhou, Michael I. Jordan

The Dawid-Skene estimator has been widely used for inferring the true labels fro m the noisy labels provided by non-expert crowdsourcing workers. However, since the estimator maximizes a non-convex log-likelihood function, it is hard to theo retically justify its performance. In this paper, we propose a two-stage efficie nt algorithm for multi-class crowd labeling problems. The first stage uses the s pectral method to obtain an initial estimate of parameters. Then the second stag e refines the estimation by optimizing the objective function of the Dawid-Skene estimator via the EM algorithm. We show that our algorithm achieves the optimal convergence rate up to a logarithmic factor. We conduct extensive experiments o n synthetic and real datasets. Experimental results demonstrate that the propose d algorithm is comparable to the most accurate empirical approach, while outperf orming several other recently proposed methods.
****

Fairness in Multi-Agent Sequential Decision-Making
Chongjie Zhang, Julie A. Shah

We define a fairness solution criterion for multi-agent decision-making problems , where agents have local interests. This new criterion aims to maximize the wor st performance of agents with consideration on the overall performance. We devel op a simple linear programming approach and a more scalable game-theoretic appro ach for computing an optimal fairness policy. This game-theoretic approach formu lates this fairness optimization as a two-player, zero-sum game and employs an i terative algorithm for finding a Nash equilibrium, corresponding to an optimal f airness policy. We scale up this approach by exploiting problem structure and va lue function approximation. Our experiments on resource allocation problems show that this fairness criterion provides a more favorable solution than the utilit arian criterion, and that our game-theoretic approach is significantly faster th an linear programming.
****

Multi-Class Deep Boosting
Vitaly Kuznetsov, Mehryar Mohri, Umar Syed

We present new ensemble learning algorithms for multi-class classification. Our algorithms can use as a base classifier set a family of deep decision trees or o ther rich or complex families and yet benefit from strong generalization guarant ees. We give new data-dependent learning bounds for convex ensembles in the mult i-class classification setting expressed in terms of the Rademacher complexities of the sub-families composing the base classifier set, and the mixture weight a ssigned to each sub-family. These bounds are finer than existing ones both thank s to an improved dependency on the number of classes and, more crucially, by vir tue of a more favorable complexity term expressed as an average of the Rademache r complexities based on the ensemble’s mixture weights. We introduce and discuss

several new multi-class ensemble algorithms benefiting from these guarantees, p
rove positive results for the H-consistency of several of them, and report the r
esults of experiments showing that their performance compares favorably with tha
t of multi-class versions of AdaBoost and Logistic Regression and their L1-regul
arized counterparts.
****

Depth Map Prediction from a Single Image using a Multi-Scale Deep Network
David Eigen, Christian Puhrsch, Rob Fergus
Predicting depth is an essential component in understanding the 3D geometry of a
 scene. While for stereo images local correspondence suffices for estimation, fi
nding depth relations from a single image is less straightforward, requiring int
egration of both global and local information from various cues. Moreover, the t
ask is inherently ambiguous, with a large source of uncertainty coming from the
overall scale. In this paper, we present a new method that addresses this task b
y employing two deep network stacks: one that makes a coarse global prediction b
ased on the entire image, and another that refines this prediction locally. We a
lso apply a scale-invariant error to help measure depth relations rather than sc
ale. By leveraging the raw datasets as large sources of training data, our metho
d achieves state-of-the-art results on both NYU Depth and KITTI, and matches det
ailed depth boundaries without the need for superpixelation.
****

Provable Submodular Minimization using Wolfe's Algorithm
Deeparnab Chakrabarty, Prateek Jain, Pravesh Kothari
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.
****

Online and Stochastic Gradient Methods for Non-decomposable Loss Functions
Purushottam Kar, Harikrishna Narasimhan, Prateek Jain
Modern applications in sensitive domains such as biometrics and medicine frequen
tly require the use of non-decomposable loss functions such as precision@k, F-me
asure etc. Compared to point loss functions such as hinge-loss, these offer much
 more fine grained control over prediction, but at the same time present novel c
hallenges in terms of algorithm design and analysis. In this work we initiate a
study of online learning techniques for such non-decomposable loss functions wit
h an aim to enable incremental learning as well as design scalable solvers for b
atch problems. To this end, we propose an online learning framework for such los
s functions. Our model enjoys several nice properties, chief amongst them being
the existence of efficient online learning algorithms with sublinear regret and
online to batch conversion bounds. Our model is a provable extension of existing
 online learning models for point loss functions. We instantiate two popular los
ses, Prec @k and pAUC, in our model and prove sublinear regret bounds for both o
f them. Our proofs require a novel structural lemma over ranked lists which may
be of independent interest. We then develop scalable stochastic gradient descent
 solvers for non-decomposable loss functions. We show that for a large family of
 loss functions satisfying a certain uniform convergence property (that includes
 Prec @k, pAUC, and F-measure), our methods provably converge to the empirical r
isk minimizer. Such uniform convergence results were not known for these losses
and we establish these using novel proof techniques. We then use extensive exper
imentation on real life and benchmark datasets to establish that our method can
be orders of magnitude faster than a recently proposed cutting plane method.
****

On Model Parallelization and Scheduling Strategies for Distributed Machine Learn
ing
Seunghak Lee, Jin Kyu Kim, Xun Zheng, Qirong Ho, Garth A. Gibson, Eric P. Xing
Distributed machine learning has typically been approached from a data parallel
perspective, where big data are partitioned to multiple workers and an algorithm
 is executed concurrently over different data subsets under various synchronizat
ion schemes to ensure speed-up and/or correctness. A sibling problem that has re

ceived relatively less attention is how to ensure efficient and correct model pa
rallel execution of ML algorithms, where parameters of an ML program are partiti
oned to different workers and undergone concurrent iterative updates. We argue t
hat model and data parallelisms impose rather different challenges for system de
sign, algorithmic adjustment, and theoretical analysis. In this paper, we develo
p a system for model-parallelism, STRADS, that provides a programming abstractio
n for scheduling parameter updates by discovering and leveraging changing struct
ural properties of ML programs. STRADS enables a flexible tradeoff between sched
uling efficiency and fidelity to intrinsic dependencies within the models, and i
mproves memory efficiency of distributed ML. We demonstrate the efficacy of mode
l-parallel algorithms implemented on STRADS versus popular implementations for t
opic modeling, matrix factorization, and Lasso.
****

Shape and Illumination from Shading using the Generic Viewpoint Assumption
Daniel Zoran, Dilip Krishnan, José Bento, Bill Freeman
The Generic Viewpoint Assumption (GVA) states that the position of the viewer or
 the light in a scene is not special. Thus, any estimated parameters from an obs
ervation should be stable under small perturbations such as object, viewpoint or
 light positions. The GVA has been analyzed and quantified in previous works, bu
t has not been put to practical use in actual vision tasks. In this paper, we sh
ow how to utilize the GVA to estimate shape and illumination from a single shadi
ng image, without the use of other priors. We propose a novel linearized Spheric
al Harmonics (SH) shading model which enables us to obtain a computationally eff
icient form of the GVA term. Together with a data term, we build a model whose u
nknowns are shape and SH illumination. The model parameters are estimated using
the Alternating Direction Method of Multipliers embedded in a multi-scale estima
tion framework. In this prior-free framework, we obtain competitive shape and il
lumination estimation results under a variety of models and lighting conditions,
 requiring fewer assumptions than competing methods.
****

Asynchronous Anytime Sequential Monte Carlo
Brooks Paige, Frank Wood, Arnaud Doucet, Yee Whye Teh
We introduce a new sequential Monte Carlo algorithm we call the particle cascade
. The particle cascade is an asynchronous, anytime alternative to traditional se
quential Monte Carlo algorithms that is amenable to parallel and distributed imp
lementations. It uses no barrier synchronizations which leads to improved partic
le throughput and memory efficiency. It is an anytime algorithm in the sense tha
t it can be run forever to emit an unbounded number of particles while keeping w
ithin a fixed memory budget. We prove that the particle cascade provides an unbi
ased marginal likelihood estimator which can be straightforwardly plugged into e
xisting pseudo-marginal methods.
****

Sparse Space-Time Deconvolution for Calcium Image Analysis
Ferran Diego Andilla, Fred A. Hamprecht
We describe a unified formulation and algorithm to find an extremely sparse repr
esentation for Calcium image sequences in terms of cell locations, cell shapes,
spike timings and impulse responses. Solution of a single optimization problem y
ields cell segmentations and activity estimates that are on par with the state o
f the art, without the need for heuristic pre- or postprocessing. Experiments on
 real and synthetic data demonstrate the viability of the proposed method.
****

From Stochastic Mixability to Fast Rates
Nishant A. Mehta, Robert C. Williamson
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.
****

Algorithm selection by rational metareasoning as a model of human strategy selec
tion

Falk Lieder, Dillon Plunkett, Jessica B. Hamrick, Stuart J. Russell, Nicholas Hay, Tom Griffiths
Selecting the right algorithm is an important problem in computer science, because the algorithm often has to exploit the structure of the input to be efficient. The human mind faces the same challenge. Therefore, solutions to the algorithm selection problem can inspire models of human strategy selection and vice versa. Here, we view the algorithm selection problem as a special case of metareasoning and derive a solution that outperforms existing methods in sorting algorithm selection. We apply our theory to model how people choose between cognitive strategies and test its prediction in a behavioral experiment. We find that people quickly learn to adaptively choose between cognitive strategies. People's choices in our experiment are consistent with our model but inconsistent with previous theories of human strategy selection. Rational metareasoning appears to be a promising framework for reverse-engineering how people choose among cognitive strategies and translating the results into better solutions to the algorithm selection problem.
****

PAC-Bayesian AUC classification and scoring
James Ridgway, Pierre Alquier, Nicolas Chopin, Feng Liang
We develop a scoring and classification procedure based on the PAC-Bayesian approach and the AUC (Area Under Curve) criterion. We focus initially on the class of linear score functions. We derive PAC-Bayesian non-asymptotic bounds for two types of prior for the score parameters: a Gaussian prior, and a spike-and-slab prior; the latter makes it possible to perform feature selection. One important advantage of our approach is that it is amenable to powerful Bayesian computational tools. We derive in particular a Sequential Monte Carlo algorithm, as an efficient method which may be used as a gold standard, and an Expectation-Propagation algorithm, as a much faster but approximate method. We also extend our method to a class of non-linear score functions, essentially leading to a nonparametric procedure, by considering a Gaussian process prior.
****

Probabilistic Differential Dynamic Programming
Yunpeng Pan, Evangelos Theodorou
We present a data-driven, probabilistic trajectory optimization framework for systems with unknown dynamics, called Probabilistic Differential Dynamic Programming (PDDP). PDDP takes into account uncertainty explicitly for dynamics models using Gaussian processes (GPs). Based on the second-order local approximation of the value function, PDDP performs Dynamic Programming around a nominal trajectory in Gaussian belief spaces. Different from typical gradient-based policy search methods, PDDP does not require a policy parameterization and learns a locally optimal, time-varying control policy. We demonstrate the effectiveness and efficiency of the proposed algorithm using two nontrivial tasks. Compared with the classical DDP and a state-of-the-art GP-based policy search method, PDDP offers a superior combination of data-efficiency, learning speed, and applicability.
****

Improved Multimodal Deep Learning with Variation of Information
Kihyuk Sohn, Wenling Shang, Honglak Lee
Deep learning has been successfully applied to multimodal representation learning problems, with a common strategy to learning joint representations that are shared across multiple modalities on top of layers of modality-specific networks. Nonetheless, there still remains a question how to learn a good association between data modalities; in particular, a good generative model of multimodal data should be able to reason about missing data modality given the rest of data modalities. In this paper, we propose a novel multimodal representation learning framework that explicitly aims this goal. Rather than learning with maximum likelihood, we train the model to minimize the variation of information. We provide a theoretical insight why the proposed learning objective is sufficient to estimate the data-generating joint distribution of multimodal data. We apply our method to restricted Boltzmann machines and introduce learning methods based on contrastive divergence and multi-prediction training. In addition, we extend to deep net

works with recurrent encoding structure to finetune the whole network. In experi
ments, we demonstrate the state-of-the-art visual recognition performance on MIR
-Flickr database and PASCAL VOC 2007 database with and without text features.
****

On Sparse Gaussian Chain Graph Models
Calvin McCarter, Seyoung Kim
In this paper, we address the problem of learning the structure of Gaussian chai
n graph models in a high-dimensional space. Chain graph models are generalizatio
ns of undirected and directed graphical models that contain a mixed set of direc
ted and undirected edges. While the problem of sparse structure learning has bee
n studied extensively for Gaussian graphical models and more recently for condit
ional Gaussian graphical models (CGGMs), there has been little previous work on
the structure recovery of Gaussian chain graph models. We consider linear regres
sion models and a re-parameterization of the linear regression models using CGGM
s as building blocks of chain graph models. We argue that when the goal is to re
cover model structures, there are many advantages of using CGGMs as chain compon
ent models over linear regression models, including convexity of the optimizatio
n problem, computational efficiency, recovery of structured sparsity, and abilit
y to leverage the model structure for semi-supervised learning. We demonstrate o
ur approach on simulated and genomic datasets.
****

Convolutional Kernel Networks
Julien Mairal, Piotr Koniusz, Zaid Harchaoui, Cordelia Schmid
An important goal in visual recognition is to devise image representations that
are invariant to particular transformations. In this paper, we address this goal
 with a new type of convolutional neural network (CNN) whose invariance is encod
ed by a reproducing kernel. Unlike traditional approaches where neural networks
are learned either to represent data or for solving a classification task, our n
etwork learns to approximate the kernel feature map on training data. Such an ap
proach enjoys several benefits over classical ones. First, by teaching CNNs to b
e invariant, we obtain simple network architectures that achieve a similar accur
acy to more complex ones, while being easy to train and robust to overfitting. S
econd, we bridge a gap between the neural network literature and kernels, which
are natural tools to model invariance. We evaluate our methodology on visual rec
ognition tasks where CNNs have proven to perform well, e.g., digit recognition w
ith the MNIST dataset, and the more challenging CIFAR-10 and STL-10 datasets, wh
ere our accuracy is competitive with the state of the art.
****

Learning Chordal Markov Networks by Dynamic Programming
Kustaa Kangas, Mikko Koivisto, Teppo Niinimäki
We present an algorithm for finding a chordal Markov network that maximizes any
given decomposable scoring function. The algorithm is based on a recursive chara
cterization of clique trees, and it runs in O(4^n) time for n vertices. On an ei
ght-vertex benchmark instance, our implementation turns out to be about ten mill
ion times faster than a recently proposed, constraint satisfaction based algorit
hm (Corander et al., NIPS 2013). Within a few hours, it is able to solve instanc
es up to 18 vertices, and beyond if we restrict the maximum clique size. We also
 study the performance of a recent integer linear programming algorithm (Bartlet
t and Cussens, UAI 2013). Our results suggest that, unless we bound the clique s
izes, currently only the dynamic programming algorithm is guaranteed to solve in
stances with around 15 or more vertices.
****

From MAP to Marginals: Variational Inference in Bayesian Submodular Models
Josip Djolonga, Andreas Krause
Submodular optimization has found many applications in machine learning and beyo
nd. We carry out the first systematic investigation of inference in probabilisti
c models defined through submodular functions, generalizing regular pairwise MRF
s and Determinantal Point Processes. In particular, we present L-Field, a variat
ional approach to general log-submodular and log-supermodular distributions base
d on sub- and supergradients. We obtain both lower and upper bounds on the log-p

artition function, which enables us to compute probability intervals for margina ls, conditionals and marginal likelihoods. We also obtain fully factorized appro ximate posteriors, at the same computational cost as ordinary submodular optimiz ation. Our framework results in convex problems for optimizing over differential s of submodular functions, which we show how to optimally solve. We provide theo retical guarantees of the approximation quality with respect to the curvature of the function. We further establish natural relations between our variational ap proach and the classical mean-field method. Lastly, we empirically demonstrate t he accuracy of our inference scheme on several submodular models.
****

Algorithms for CVaR Optimization in MDPs
Yinlam Chow, Mohammad Ghavamzadeh
In many sequential decision-making problems we may want to manage risk by minimi zing some measure of variability in costs in addition to minimizing a standard c riterion. Conditional value-at-risk (CVaR) is a relatively new risk measure that addresses some of the shortcomings of the well-known variance-related risk meas ures, and because of its computational efficiencies has gained popularity in fin ance and operations research. In this paper, we consider the mean-CVaR optimizat ion problem in MDPs. We first derive a formula for computing the gradient of thi s risk-sensitive objective function. We then devise policy gradient and actor-cr itic algorithms that each uses a specific method to estimate this gradient and u pdates the policy parameters in the descent direction. We establish the converge nce of our algorithms to locally risk-sensitive optimal policies. Finally, we de monstrate the usefulness of our algorithms in an optimal stopping problem.
****

Structure Regularization for Structured Prediction
Xu Sun
While there are many studies on weight regularization, the study on structure re gularization is rare. Many existing systems on structured prediction focus on in creasing the level of structural dependencies within the model. However, this tr end could have been misdirected, because our study suggests that complex structu res are actually harmful to generalization ability in structured prediction. To control structure-based overfitting, we propose a structure regularization frame work via \emph{structure decomposition}, which decomposes training samples into mini-samples with simpler structures, deriving a model with better generalizatio n power. We show both theoretically and empirically that structure regularizatio n can effectively control overfitting risk and lead to better accuracy. As a by-product, the proposed method can also substantially accelerate the training spee d. The method and the theoretical results can apply to general graphical models with arbitrary structures. Experiments on well-known tasks demonstrate that our method can easily beat the benchmark systems on those highly-competitive tasks, achieving record-breaking accuracies yet with substantially faster training spee d.
****

Bayes-Adaptive Simulation-based Search with Value Function Approximation
Arthur Guez, Nicolas Heess, David Silver, Peter Dayan
Bayes-adaptive planning offers a principled solution to the exploration-exploita tion trade-off under model uncertainty. It finds the optimal policy in belief sp ace, which explicitly accounts for the expected effect on future rewards of redu ctions in uncertainty. However, the Bayes-adaptive solution is typically intract able in domains with large or continuous state spaces. We present a tractable me thod for approximating the Bayes-adaptive solution by combining simulation-based search with a novel value function approximation technique that generalises ove r belief space. Our method outperforms prior approaches in both discrete bandit tasks and simple continuous navigation and control tasks.
****

Optimal Teaching for Limited-Capacity Human Learners
Kaustubh R. Patil, Jerry Zhu, ■ukasz Kope■, Bradley C. Love
Basic decisions, such as judging a person as a friend or foe, involve categorizi ng novel stimuli. Recent work finds that people’s category judgments are guided

by a small set of examples that are retrieved from memory at decision time. This limited and stochastic retrieval places limits on human performance for probabilistic classification decisions. In light of this capacity limitation, recent work finds that idealizing training items, such that the saliency of ambiguous cases is reduced, improves human performance on novel test items. One shortcoming of previous work in idealization is that category distributions were idealized in an ad hoc or heuristic fashion. In this contribution, we take a first principles approach to constructing idealized training sets. We apply a machine teaching procedure to a cognitive model that is either limited capacity (as humans are) or unlimited capacity (as most machine learning systems are). As predicted, we find that the machine teacher recommends idealized training sets. We also find that human learners perform best when training recommendations from the machine teacher are based on a limited-capacity model. As predicted, to the extent that the learning model used by the machine teacher conforms to the true nature of human learners, the recommendations of the machine teacher prove effective. Our results provide a normative basis (given capacity constraints) for idealization procedures and offer a novel selection procedure for models of human learning.
****

Spectral Methods for Indian Buffet Process Inference
Hsiao-Yu Tung, Alexander J. Smola
The Indian Buffet Process is a versatile statistical tool for modeling distributions over binary matrices. We provide an efficient spectral algorithm as an alternative to costly Variational Bayes and sampling-based algorithms. We derive a novel tensorial characterization of the moments of the Indian Buffet Process proper and for two of its applications. We give a computationally efficient iterative inference algorithm, concentration of measure bounds, and reconstruction guarantees. Our algorithm provides superior accuracy and cheaper computation than comparable Variational Bayesian approach on a number of reference problems.
****

Deep Fragment Embeddings for Bidirectional Image Sentence Mapping
Andrej Karpathy, Armand Joulin, Li F. Fei-Fei
We introduce a model for bidirectional retrieval of images and sentences through a deep, multi-modal embedding of visual and natural language data. Unlike previous models that directly map images or sentences into a common embedding space, our model works on a finer level and embeds fragments of images (objects) and fragments of sentences (typed dependency tree relations) into a common space. We then introduce a structured max-margin objective that allows our model to explicitly associate these fragments across modalities. Extensive experimental evaluation shows that reasoning on both the global level of images and sentences and the finer level of their respective fragments improves performance on image-sentence retrieval tasks. Additionally, our model provides interpretable predictions for the image-sentence retrieval task since the inferred inter-modal alignment of fragments is explicit.
****

Greedy Subspace Clustering
Dohyung Park, Constantine Caramanis, Sujay Sanghavi
We consider the problem of subspace clustering: given points that lie on or near the union of many low-dimensional linear subspaces, recover the subspaces. To this end, one first identifies sets of points close to the same subspace and uses the sets to estimate the subspaces. As the geometric structure of the clusters (linear subspaces) forbids proper performance of general distance based approaches such as K-means, many model-specific methods have been proposed. In this paper, we provide new simple and efficient algorithms for this problem. Our statistical analysis shows that the algorithms are guaranteed exact (perfect) clustering performance under certain conditions on the number of points and the affinity b e- tween subspaces. These conditions are weaker than those considered in the standard statistical literature. Experimental results on synthetic data generated from the standard unions of subspaces model demonstrate our theory. We also show that our algorithm performs competitively against state-of-the-art algorithms on real-world applications such as motion segmentation and face clustering, with m

uch simpler implementation and lower computational cost.
****
Feature Cross-Substitution in Adversarial Classification
Bo Li, Yevgeniy Vorobeychik
The success of machine learning, particularly in supervised settings, has led to numerous attempts to apply it in adversarial settings such as spam and malware detection. The core challenge in this class of applications is that adversaries are not static data generators, but make a deliberate effort to evade the classifiers deployed to detect them. We investigate both the problem of modeling the objectives of such adversaries, as well as the algorithmic problem of accounting for rational, objective-driven adversaries. In particular, we demonstrate severe shortcomings of feature reduction in adversarial settings using several natural adversarial objective functions, an observation that is particularly pronounced when the adversary is able to substitute across similar features (for example, replace words with synonyms or replace letters in words). We offer a simple heuristic method for making learning more robust to feature cross-substitution attacks. We then present a more general approach based on mixed-integer linear programming with constraint generation, which implicitly trades off overfitting and feature selection in an adversarial setting using a sparse regularizer along with an evasion model. Our approach is the first method for combining an adversarial classification algorithm with a very general class of models of adversarial classifier evasion. We show that our algorithmic approach significantly outperforms state-of-the-art alternatives.
****
Distributed Balanced Clustering via Mapping Coresets
Mohammadhossein Bateni, Aditya Bhaskara, Silvio Lattanzi, Vahab Mirrokni
Large-scale clustering of data points in metric spaces is an important problem in mining big data sets. For many applications, we face explicit or implicit size constraints for each cluster which leads to the problem of clustering under capacity constraints or the balanced clustering'' problem. Although the balanced clustering problem has been widely studied, developing a theoretically sound distributed algorithm remains an open problem. In the present paper we develop a general framework based onmapping coresets'' to tackle this issue. For a wide range of clustering objective functions such as k-center, k-median, and k-means, our techniques give distributed algorithms for balanced clustering that match the best known single machine approximation ratios.
****
Stochastic variational inference for hidden Markov models
Nick Foti, Jason Xu, Dillon Laird, Emily Fox
Variational inference algorithms have proven successful for Bayesian analysis in large data settings, with recent advances using stochastic variational inference (SVI). However, such methods have largely been studied in independent or exchangeable data settings. We develop an SVI algorithm to learn the parameters of hidden Markov models (HMMs) in a time-dependent data setting. The challenge in applying stochastic optimization in this setting arises from dependencies in the chain, which must be broken to consider minibatches of observations. We propose an algorithm that harnesses the memory decay of the chain to adaptively bound errors arising from edge effects. We demonstrate the effectiveness of our algorithm on synthetic experiments and a large genomics dataset where a batch algorithm is computationally infeasible.
****
Tight convex relaxations for sparse matrix factorization
Emile Richard, Guillaume R. Obozinski, Jean-Philippe Vert
Based on a new atomic norm, we propose a new convex formulation for sparse matrix factorization problems in which the number of nonzero elements of the factors is assumed fixed and known. The formulation counts sparse PCA with multiple factors, subspace clustering and low-rank sparse bilinear regression as potential applications. We compute slow rates and an upper bound on the statistical dimension of the suggested norm for rank 1 matrices, showing that its statistical dimension is an order of magnitude smaller than the usual l_1-norm, trace norm and the

ir combinations. Even though our convex formulation is in theory hard and does n
ot lead to provably polynomial time algorithmic schemes, we propose an active se
t algorithm leveraging the structure of the convex problem to solve it and show
promising numerical results.
****

Extremal Mechanisms for Local Differential Privacy
Peter Kairouz, Sewoong Oh, Pramod Viswanath
Local differential privacy has recently surfaced as a strong measure of privacy
in contexts where personal information remains private even from data analysts.
Working in a setting where the data providers and data analysts want to maximize
 the utility of statistical inferences performed on the released data, we study
the fundamental tradeoff between local differential privacy and information theo
retic utility functions. We introduce a family of extremal privatization mechani
sms, which we call staircase mechanisms, and prove that it contains the optimal
privatization mechanism that maximizes utility. We further show that for all inf
ormation theoretic utility functions studied in this paper, maximizing utility i
s equivalent to solving a linear program, the outcome of which is the optimal st
aircase mechanism. However, solving this linear program can be computationally e
xpensive since it has a number of variables that is exponential in the data size
. To account for this, we show that two simple staircase mechanisms, the binary
and randomized response mechanisms, are universally optimal in the high and low
privacy regimes, respectively, and well approximate the intermediate regime.
****

Optimization Methods for Sparse Pseudo-Likelihood Graphical Model Selection
Sang Oh, Onkar Dalal, Kshitij Khare, Bala Rajaratnam
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.
****

Unsupervised Deep Haar Scattering on Graphs
Xu Chen, Xiuyuan Cheng, Stephane Mallat
The classification of high-dimensional data defined on graphs is particularly di
fficult when the graph geometry is unknown. We introduce a Haar scattering trans
form on graphs, which computes invariant signal descriptors. It is implemented w
ith a deep cascade of additions, subtractions and absolute values, which iterati
vely compute orthogonal Haar wavelet transforms. Multiscale neighborhoods of unk
nown graphs are estimated by minimizing an average total variation, with a pair
matching algorithm of polynomial complexity. Supervised classification with dime
nsion reduction is tested on data bases of scrambled images, and for signals sam
pled on unknown irregular grids on a sphere.
****

Communication-Efficient Distributed Dual Coordinate Ascent
Martin Jaggi, Virginia Smith, Martin Takac, Jonathan Terhorst, Sanjay Krishnan,
Thomas Hofmann, Michael I. Jordan
Communication remains the most significant bottleneck in the performance of dist
ributed optimization algorithms for large-scale machine learning. In this paper,
 we propose a communication-efficient framework, COCOA, that uses local computat
ion in a primal-dual setting to dramatically reduce the amount of necessary comm
unication. We provide a strong convergence rate analysis for this class of algor
ithms, as well as experiments on real-world distributed datasets with implementa
tions in Spark. In our experiments, we find that as compared to state-of-the-art
 mini-batch versions of SGD and SDCA algorithms, COCOA converges to the same .00
1-accurate solution quality on average 25× as quickly.
****

Learning convolution filters for inverse covariance estimation of neural network
 connectivity
George Mohler
We consider the problem of inferring direct neural network connections from Calc
ium imaging time series. Inverse covariance estimation has proven to be a fast a

nd accurate method for learning macro- and micro-scale network connectivity in t
he brain and in a recent Kaggle Connectomics competition inverse covariance was
the main component of several top ten solutions, including our own and the winni
ng team's algorithm. However, the accuracy of inverse covariance estimation is h
ighly sensitive to signal preprocessing of the Calcium fluorescence time series.
 Furthermore, brute force optimization methods such as grid search and coordinat
e ascent over signal processing parameters is a time intensive process, where le
arning may take several days and parameters that optimize one network may not ge
neralize to networks with different size and parameters. In this paper we show h
ow inverse covariance estimation can be dramatically improved using a simple con
volution filter prior to applying sample covariance. Furthermore, these signal p
rocessing parameters can be learned quickly using a supervised optimization algo
rithm. In particular, we maximize a binomial log-likelihood loss function with r
espect to a convolution filter of the time series and the inverse covariance reg
ularization parameter. Our proposed algorithm is relatively fast on networks the
 size of those in the competition (1000 neurons), producing AUC scores with simi
lar accuracy to the winning solution in training time under 2 hours on a cpu. Pr
ediction on new networks of the same size is carried out in less than 15 minutes
, the time it takes to read in the data and write out the solution.
****

General Stochastic Networks for Classification
Matthias Zöhrer, Franz Pernkopf
We extend generative stochastic networks to supervised learning of representatio
ns. In particular, we introduce a hybrid training objective considering a genera
tive and discriminative cost function governed by a trade-off parameter lambda.
We use a new variant of network training involving noise injection, i.e. walkbac
k training, to jointly optimize multiple network layers. Neither additional regu
larization constraints, such as l1, l2 norms or dropout variants, nor pooling- o
r convolutional layers were added. Nevertheless, we are able to obtain state-of-
the-art performance on the MNIST dataset, without using permutation invariant di
gits and outperform baseline models on sub-variants of the MNIST and rectangles
dataset significantly.
****

Articulated Pose Estimation by a Graphical Model with Image Dependent Pairwise R
elations
Xianjie Chen, Alan L. Yuille
We present a method for estimating articulated human pose from a single static i
mage based on a graphical model with novel pairwise relations that make adaptive
 use of local image measurements. More precisely, we specify a graphical model f
or human pose which exploits the fact the local image measurements can be used b
oth to detect parts (or joints) and also to predict the spatial relationships be
tween them (Image Dependent Pairwise Relations). These spatial relationships are
 represented by a mixture model. We use Deep Convolutional Neural Networks (DCNN
s) to learn conditional probabilities for the presence of parts and their spatia
l relationships within image patches. Hence our model combines the representatio
nal flexibility of graphical models with the efficiency and statistical power of
 DCNNs. Our method significantly outperforms the state of the art methods on the
 LSP and FLIC datasets and also performs very well on the Buffy dataset without
any training.
****

Deep Learning for Real-Time Atari Game Play Using Offline Monte-Carlo Tree Searc
h Planning
Xiaoxiao Guo, Satinder Singh, Honglak Lee, Richard L. Lewis, Xiaoshi Wang
The combination of modern Reinforcement Learning and Deep Learning approaches ho
lds the promise of making significant progress on challenging applications requi
ring both rich perception and policy-selection. The Arcade Learning Environment
(ALE) provides a set of Atari games that represent a useful benchmark set of suc
h applications. A recent breakthrough in combining model-free reinforcement lear
ning with deep learning, called DQN, achieves the best real-time agents thus far
. Planning-based approaches achieve far higher scores than the best model-free a

pproaches, but they exploit information that is not available to human players, and they are orders of magnitude slower than needed for real-time play. Our main goal in this work is to build a better real-time Atari game playing agent than DQN. The central idea is to use the slow planning-based agents to provide traini ng data for a deep-learning architecture capable of real-time play. We proposed new agents based on this idea and show that they outperform DQN.
****

Near-optimal sample compression for nearest neighbors
Lee-Ad Gottlieb, Aryeh Kontorovich, Pinhas Nisnevitch
We present the first sample compression algorithm for nearest neighbors with non -trivial performance guarantees. We complement these guarantees by demonstrating almost matching hardness lower bounds, which show that our bound is nearly opti mal. Our result yields new insight into margin-based nearest neighbor classifica tion in metric spaces and allows us to significantly sharpen and simplify existi ng bounds. Some encouraging empirical results are also presented.
****

Factoring Variations in Natural Images with Deep Gaussian Mixture Models
Aaron van den Oord, Benjamin Schrauwen
Generative models can be seen as the swiss army knives of machine learning, as m any problems can be written probabilistically in terms of the distribution of th e data, including prediction, reconstruction, imputation and simulation. One of the most promising directions for unsupervised learning may lie in Deep Learning methods, given their success in supervised learning. However, one of the curren t problems with deep unsupervised learning methods, is that they often are harde r to scale. As a result there are some easier, more scalable shallow methods, su ch as the Gaussian Mixture Model and the Student-t Mixture Model, that remain su rprisingly competitive. In this paper we propose a new scalable deep generative model for images, called the Deep Gaussian Mixture Model, that is a straightforw ard but powerful generalization of GMMs to multiple layers. The parametrization of a Deep GMM allows it to efficiently capture products of variations in natural images. We propose a new EM-based algorithm that scales well to large datasets, and we show that both the Expectation and the Maximization steps can easily be distributed over multiple machines. In our density estimation experiments we sho w that deeper GMM architectures generalize better than more shallow ones, with r esults in the same ballpark as the state of the art.
****

Automated Variational Inference for Gaussian Process Models
Trung V. Nguyen, Edwin V. Bonilla
We develop an automated variational method for approximate inference in Gaussian process (GP) models whose posteriors are often intractable. Using a mixture of Gaussians as the variational distribution, we show that (i) the variational obje ctive and its gradients can be approximated efficiently via sampling from univar iate Gaussian distributions and (ii) the gradients of the GP hyperparameters can be obtained analytically regardless of the model likelihood. We further propose two instances of the variational distribution whose covariance matrices can be parametrized linearly in the number of observations. These results allow gradien t-based optimization to be done efficiently in a black-box manner. Our approach is thoroughly verified on 5 models using 6 benchmark datasets, performing as wel l as the exact or hard-coded implementations while running orders of magnitude f aster than the alternative MCMC sampling approaches. Our method can be a valuabl e tool for practitioners and researchers to investigate new models with minimal effort in deriving model-specific inference algorithms.
****

Extreme bandits
Alexandra Carpentier, Michal Valko
In many areas of medicine, security, and life sciences, we want to allocate limi ted resources to different sources in order to detect extreme values. In this pa per, we study an efficient way to allocate these resources sequentially under li mited feedback. While sequential design of experiments is well studied in bandit theory, the most commonly optimized property is the regret with respect to the

maximum mean reward. However, in other problems such as network intrusion detect
ion, we are interested in detecting the most extreme value output by the sources
. Therefore, in our work we study extreme regret which measures the efficiency o
f an algorithm compared to the oracle policy selecting the source with the heavi
est tail. We propose the ExtremeHunter algorithm, provide its analysis, and eval
uate it empirically on synthetic and real-world experiments.
****

Learning Mixed Multinomial Logit Model from Ordinal Data
Sewoong Oh, 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-auth
ors prior to requesting a name change in the electronic proceedings.
****

Elementary Estimators for Graphical Models
Eunho Yang, Aurelie C. Lozano, Pradeep K. Ravikumar
Requests for name changes in the electronic proceedings will be accepted with no
 questions asked. However name changes may cause bibliographic tracking issues.
  Authors are asked to consider this carefully and discuss it with their co-auth
ors prior to requesting a name change in the electronic proceedings.
****

Efficient Minimax Strategies for Square Loss Games
Wouter M. Koolen, Alan Malek, 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-auth
ors prior to requesting a name change in the electronic proceedings.
****

Generalized Higher-Order Orthogonal Iteration for Tensor Decomposition and Compl
etion
Yuanyuan Liu, Fanhua Shang, Wei Fan, James Cheng, Hong Cheng
Low-rank tensor estimation has been frequently applied in many real-world proble
ms. Despite successful applications, existing Schatten 1-norm minimization (SNM)
 methods may become very slow or even not applicable for large-scale problems. T
o address this difficulty, we therefore propose an efficient and scalable core t
ensor Schatten 1-norm minimization method for simultaneous tensor decomposition
and completion, with a much lower computational complexity. We first induce the
equivalence relation of Schatten 1-norm of a low-rank tensor and its core tensor
. Then the Schatten 1-norm of the core tensor is used to replace that of the who
le tensor, which leads to a much smaller-scale matrix SNM problem. Finally, an e
fficient algorithm with a rank-increasing scheme is developed to solve the propo
sed problem with a convergence guarantee. Extensive experimental results show th
at our method is usually more accurate than the state-of-the-art methods, and is
 orders of magnitude faster.
****

Submodular meets Structured: Finding Diverse Subsets in Exponentially-Large Stru
ctured Item Sets
Adarsh Prasad, Stefanie Jegelka, Dhruv Batra
To cope with the high level of ambiguity faced in domains such as Computer Visio
n or Natural Language processing, robust prediction methods often search for a d
iverse set of high-quality candidate solutions or proposals. In structured predi
ction problems, this becomes a daunting task, as the solution space (image label
ings, sentence parses, etc.) is exponentially large. We study greedy algorithms
for finding a diverse subset of solutions in structured-output spaces by drawing
 new connections between submodular functions over combinatorial item sets and H
igh-Order Potentials (HOPs) studied for graphical models. Specifically, we show
via examples that when marginal gains of submodular diversity functions allow st
ructured representations, this enables efficient (sub-linear time) approximate m
aximization by reducing the greedy augmentation step to inference in a factor gr
aph with appropriately constructed HOPs. We discuss benefits, tradeoffs, and sho

w that our constructions lead to significantly better proposals.
****
Robust Bayesian Max-Margin Clustering
Changyou Chen, Jun Zhu, Xinhua Zhang
We present max-margin Bayesian clustering (BMC), a general and robust framework that incorporates the max-margin criterion into Bayesian clustering models, as well as two concrete models of BMC to demonstrate its flexibility and effectiveness in dealing with different clustering tasks. The Dirichlet process max-margin Gaussian mixture is a nonparametric Bayesian clustering model that relaxes the underlying Gaussian assumption of Dirichlet process Gaussian mixtures by incorporating max-margin posterior constraints, and is able to infer the number of clusters from data. We further extend the ideas to present max-margin clustering topic model, which can learn the latent topic representation of each document while at the same time cluster documents in the max-margin fashion. Extensive experiments are performed on a number of real datasets, and the results indicate superior clustering performance of our methods compared to related baselines.
****
Approximating Hierarchical MV-sets for Hierarchical Clustering
Assaf Glazer, Omer Weissbrod, Michael Lindenbaum, Shaul Markovitch
The goal of hierarchical clustering is to construct a cluster tree, which can be viewed as the modal structure of a density. For this purpose, we use a convex optimization program that can efficiently estimate a family of hierarchical dense sets in high-dimensional distributions. We further extend existing graph-based methods to approximate the cluster tree of a distribution. By avoiding direct density estimation, our method is able to handle high-dimensional data more efficiently than existing density-based approaches. We present empirical results that demonstrate the superiority of our method over existing ones.
****
Diverse Randomized Agents Vote to Win
Albert Jiang, Leandro Soriano Marcolino, Ariel D. Procaccia, Tuomas Sandholm, Nisarg Shah, Milind Tambe
We investigate the power of voting among diverse, randomized software agents. With teams of computer Go agents in mind, we develop a novel theoretical model of two-stage noisy voting that builds on recent work in machine learning. This model allows us to reason about a collection of agents with different biases (determined by the first-stage noise models), which, furthermore, apply randomized algorithms to evaluate alternatives and produce votes (captured by the second-stage noise models). We analytically demonstrate that a uniform team, consisting of multiple instances of any single agent, must make a significant number of mistakes, whereas a diverse team converges to perfection as the number of agents grows. Our experiments, which pit teams of computer Go agents against strong agents, provide evidence for the effectiveness of voting when agents are diverse.
****
Consistency of Spectral Partitioning of Uniform Hypergraphs under Planted Partition Model
Debarghya Ghoshdastidar, Ambedkar Dukkipati
Spectral graph partitioning methods have received significant attention from both practitioners and theorists in computer science. Some notable studies have been carried out regarding the behavior of these methods for infinitely large sample size (von Luxburg et al., 2008; Rohe et al., 2011), which provide sufficient confidence to practitioners about the effectiveness of these methods. On the other hand, recent developments in computer vision have led to a plethora of applications, where the model deals with multi-way affinity relations and can be posed as uniform hyper-graphs. In this paper, we view these models as random m-uniform hypergraphs and establish the consistency of spectral algorithm in this general setting. We develop a planted partition model or stochastic blockmodel for such problems using higher order tensors, present a spectral technique suited for the purpose and study its large sample behavior. The analysis reveals that the algorithm is consistent for m-uniform hypergraphs for larger values of m, and also the rate of convergence improves for increasing m. Our result provides the first

theoretical evidence that establishes the importance of m-way affinities.
****
An Accelerated Proximal Coordinate Gradient Method
Qihang Lin, Zhaosong Lu, Lin Xiao
We develop an accelerated randomized proximal coordinate gradient (APCG) method, for solving a broad class of composite convex optimization problems. In particular, our method achieves faster linear convergence rates for minimizing strongly convex functions than existing randomized proximal coordinate gradient methods. We show how to apply the APCG method to solve the dual of the regularized empirical risk minimization (ERM) problem, and devise efficient implementations that can avoid full-dimensional vector operations. For ill-conditioned ERM problems, our method obtains improved convergence rates than the state-of-the-art stochastic dual coordinate ascent (SDCA) method.
****
Scalable Non-linear Learning with Adaptive Polynomial Expansions
Alekh Agarwal, Alina Beygelzimer, Daniel J. Hsu, John Langford, Matus J. Telgarsky
Can we effectively learn a nonlinear representation in time comparable to linear learning? We describe a new algorithm that explicitly and adaptively expands higher-order interaction features over base linear representations. The algorithm is designed for extreme computational efficiency, and an extensive experimental study shows that its computation/prediction tradeoff ability compares very favorably against strong baselines.
****
Multi-Step Stochastic ADMM in High Dimensions: Applications to Sparse Optimization and Matrix Decomposition
Hanie Sedghi, Anima Anandkumar, Edmond Jonckheere
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.
****
Stochastic Multi-Armed-Bandit Problem with Non-stationary Rewards
Omar Besbes, Yonatan Gur, Assaf Zeevi
In a multi-armed bandit (MAB) problem a gambler needs to choose at each round of play one of K arms, each characterized by an unknown reward distribution. Reward realizations are only observed when an arm is selected, and the gambler's objective is to maximize his cumulative expected earnings over some given horizon of play T. To do this, the gambler needs to acquire information about arms (exploration) while simultaneously optimizing immediate rewards (exploitation); the price paid due to this trade off is often referred to as the regret, and the main question is how small can this price be as a function of the horizon length T. This problem has been studied extensively when the reward distributions do not change over time; an assumption that supports a sharp characterization of the regret, yet is often violated in practical settings. In this paper, we focus on a MAB formulation which allows for a broad range of temporal uncertainties in the rewards, while still maintaining mathematical tractability. We fully characterize the (regret) complexity of this class of MAB problems by establishing a direct link between the extent of allowable reward variation" and the minimal achievable regret, and by establishing a connection between the adversarial and the stochastic MAB frameworks."
****
Efficient Structured Matrix Rank Minimization
Adams Wei Yu, Wanli Ma, Yaoliang Yu, Jaime Carbonell, Suvrit Sra
We study the problem of finding structured low-rank matrices using nuclear norm regularization where the structure is encoded by a linear map. In contrast to most known approaches for linearly structured rank minimization, we do not (a) use the full SVD; nor (b) resort to augmented Lagrangian techniques; nor (c) solve linear systems per iteration. Instead, we formulate the problem differently so that it is amenable to a generalized conditional gradient method, which results i

n a practical improvement with low per iteration computational cost. Numerical r
esults show that our approach significantly outperforms state-of-the-art competi
tors in terms of running time, while effectively recovering low rank solutions i
n stochastic system realization and spectral compressed sensing problems.
****

Beyond Disagreement-Based Agnostic Active Learning
Chicheng Zhang, Kamalika Chaudhuri
We study agnostic active learning, where the goal is to learn a classifier in a
pre-specified hypothesis class interactively with as few label queries as possib
le, while making no assumptions on the true function generating the labels. The
main algorithms for this problem are {\em{disagreement-based active learning}},
which has a high label requirement, and {\em{margin-based active learning}}, whi
ch only applies to fairly restricted settings. A major challenge is to find an a
lgorithm which achieves better label complexity, is consistent in an agnostic se
tting, and applies to general classification problems. In this paper, we provide
 such an algorithm. Our solution is based on two novel contributions -- a reduct
ion from consistent active learning to confidence-rated prediction with guarante
ed error, and a novel confidence-rated predictor.
****

Optimal Neural Codes for Control and Estimation
Alex K. Susemihl, Ron Meir, Manfred Opper
Agents acting in the natural world aim at selecting appropriate actions based on
 noisy and partial sensory observations. Many behaviors leading to decision maki
ng and action selection in a closed loop setting are naturally phrased within a
control theoretic framework. Within the framework of optimal Control Theory, one
 is usually given a cost function which is minimized by selecting a control law
based on the observations. While in standard control settings the sensors are as
sumed fixed, biological systems often gain from the extra flexibility of optimiz
ing the sensors themselves. However, this sensory adaptation is geared towards c
ontrol rather than perception, as is often assumed. In this work we show that se
nsory adaptation for control differs from sensory adaptation for perception, eve
n for simple control setups. This implies, consistently with recent experimental
 results, that when studying sensory adaptation, it is essential to account for
the task being performed.
****

New Rules for Domain Independent Lifted MAP Inference
Happy Mittal, Prasoon Goyal, Vibhav G. Gogate, Parag Singla
Lifted inference algorithms for probabilistic first-order logic frameworks such
as Markov logic networks (MLNs) have received significant attention in recent ye
ars. These algorithms use so called lifting rules to identify symmetries in the
first-order representation and reduce the inference problem over a large probabi
listic model to an inference problem over a much smaller model. In this paper, w
e present two new lifting rules, which enable fast MAP inference in a large clas
s of MLNs. Our first rule uses the concept of single occurrence equivalence clas
s of logical variables, which we define in the paper. The rule states that the M
AP assignment over an MLN can be recovered from a much smaller MLN, in which eac
h logical variable in each single occurrence equivalence class is replaced by a
constant (i.e., an object in the domain of the variable). Our second rule states
 that we can safely remove a subset of formulas from the MLN if all equivalence
classes of variables in the remaining MLN are single occurrence and all formulas
 in the subset are tautology (i.e., evaluate to true) at extremes (i.e., assignm
ents with identical truth value for groundings of a predicate). We prove that ou
r two new rules are sound and demonstrate via a detailed experimental evaluation
 that our approach is superior in terms of scalability and MAP solution quality
to the state of the art approaches.
****

Sparse Multi-Task Reinforcement Learning
Daniele Calandriello, Alessandro Lazaric, Marcello Restelli
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.
****

The limits of squared Euclidean distance regularization
Michal Derezinski, Manfred K. K. Warmuth
Some of the simplest loss functions considered in Machine Learning are the squar
e loss, the logistic loss and the hinge loss. The most common family of algorith
ms, including Gradient Descent (GD) with and without Weight Decay, always predic
t with a linear combination of the past instances. We give a random construction
 for sets of examples where the target linear weight vector is trivial to learn
but any algorithm from the above family is drastically sub-optimal. Our lower bo
und on the latter algorithms holds even if the algorithms are enhanced with an a
rbitrary kernel function. This type of result was known for the square loss. How
ever, we develop new techniques that let us prove such hardness results for any
 loss function satisfying some minimal requirements on the loss function (includi
ng the three listed above). We also show that algorithms that regularize with th
e squared Euclidean distance are easily confused by random features. Finally, we
 conclude by discussing related open problems regarding feed forward neural netw
orks. We conjecture that our hardness results hold for any training algorithm th
at is based on the squared Euclidean distance regularization (i.e. Back-propagat
ion with the Weight Decay heuristic).
****

Finding a sparse vector in a subspace: Linear sparsity using alternating directi
ons
Qing Qu, Ju Sun, John Wright
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.
****

Scalable Kernel Methods via Doubly Stochastic Gradients
Bo Dai, Bo Xie, Niao He, Yingyu Liang, Anant Raj, Maria-Florina F. Balcan, Le So
ng
The general perception is that kernel methods are not scalable, so neural nets b
ecome the choice for large-scale nonlinear learning problems. Have we tried hard
 enough for kernel methods? In this paper, we propose an approach that scales up
 kernel methods using a novel concept called ``doubly stochastic functional grad
ients''. Based on the fact that many kernel methods can be expressed as convex o
ptimization problems, our approach solves the optimization problems by making tw
o unbiased stochastic approximations to the functional gradient---one using rand
om training points and another using random features associated with the kernel-
--and performing descent steps with this noisy functional gradient. Our algorith
m is simple, need no commit to a preset number of random features, and allows th
e flexibility of the function class to grow as we see more incoming data in the
streaming setting. We demonstrate that a function learned by this procedure afte
r t iterations converges to the optimal function in the reproducing kernel Hilbe
rt space in rate O(1/t), and achieves a generalization bound of O(1/\sqrt{t}). O
ur approach can readily scale kernel methods up to the regimes which are dominat
ed by neural nets. We show competitive performances of our approach as compared
to neural nets in datasets such as 2.3 million energy materials from MolecularSp
ace, 8 million handwritten digits from MNIST, and 1 million photos from ImageNet
 using convolution features.
****

Learning Mixtures of Ranking Models
Pranjal Awasthi, Avrim Blum, Or Sheffet, Aravindan Vijayaraghavan
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.
****

Using Convolutional Neural Networks to Recognize Rhythm ■Stimuli from Electroenc
ephalography Recordings
Sebastian Stober, Daniel J. Cameron, Jessica A. Grahn
Electroencephalography (EEG) recordings of rhythm perception might contain enoug
h information to distinguish different rhythm types/genres or even identify the
rhythms themselves. We apply convolutional neural networks (CNNs) to analyze and
 classify EEG data recorded within a rhythm perception study in Kigali, Rwanda w
hich comprises 12 East African and 12 Western rhythmic stimuli – each presented
in a loop for 32 seconds to 13 participants. We investigate the impact of the da
ta representation and the pre-processing steps for this classification tasks and
 compare different network structures. Using CNNs, we are able to recognize indi
vidual rhythms from the EEG with a mean classification accuracy of 24.4% (chance
 level 4.17%) over all subjects by looking at less than three seconds from a sin
gle channel. Aggregating predictions for multiple channels, a mean accuracy of u
p to 50% can be achieved for individual subjects.
****

Content-based recommendations with Poisson factorization
Prem K. Gopalan, Laurent Charlin, David Blei
We develop collaborative topic Poisson factorization (CTPF), a generative model
of articles and reader preferences. CTPF can be used to build recommender system
s by learning from reader histories and content to recommend personalized articl
es of interest. In detail, CTPF models both reader behavior and article texts wi
th Poisson distributions, connecting the latent topics that represent the texts
with the latent preferences that represent the readers. This provides better rec
ommendations than competing methods and gives an interpretable latent space for
understanding patterns of readership. Further, we exploit stochastic variational
 inference to model massive real-world datasets. For example, we can fit CPTF to
 the full arXiv usage dataset, which contains over 43 million ratings and 42 mil
lion word counts, within a day. We demonstrate empirically that our model outper
forms several baselines, including the previous state-of-the-art approach.
****

Optimizing Energy Production Using Policy Search and Predictive State Representa
tions
Yuri Grinberg, Doina Precup, Michel Gendreau
We consider the challenging practical problem of optimizing the power production
 of a complex of hydroelectric power plants, which involves control over three c
ontinuous action variables, uncertainty in the amount of water inflows and a var
iety of constraints that need to be satisfied. We propose a policy-search-based
approach coupled with predictive modelling to address this problem. This approac
h has some key advantages compared to other alternatives, such as dynamic progra
mming: the policy representation and search algorithm can conveniently incorpora
te domain knowledge; the resulting policies are easy to interpret, and the algor
ithm is naturally parallelizable. Our algorithm obtains a policy which outperfor
ms the solution found by dynamic programming both quantitatively and qualitative
ly.
****

Time--Data Tradeoffs by Aggressive Smoothing
John J. Bruer, Joel A. Tropp, Volkan Cevher, Stephen Becker
This paper proposes a tradeoff between sample complexity and computation time th
at applies to statistical estimators based on convex optimization. As the amount
 of data increases, we can smooth optimization problems more and more aggressive
ly to achieve accurate estimates more quickly. This work provides theoretical an
d experimental evidence of this tradeoff for a class of regularized linear inver
se problems.
****

Shaping Social Activity by Incentivizing Users
Mehrdad Farajtabar, Nan Du, Manuel Gomez Rodriguez, Isabel Valera, Hongyuan Zha,
 Le Song
Events in an online social network can be categorized roughly into endogenous ev
ents, where users just respond to the actions of their neighbors within the netw

ork, or exogenous events, where users take actions due to drives external to the network. How much external drive should be provided to each user, such that the network activity can be steered towards a target state? In this paper, we model social events using multivariate Hawkes processes, which can capture both endogenous and exogenous event intensities, and derive a time dependent linear relation between the intensity of exogenous events and the overall network activity. Exploiting this connection, we develop a convex optimization framework for determining the required level of external drive in order for the network to reach a desired activity level. We experimented with event data gathered from Twitter, and show that our method can steer the activity of the network more accurately than alternatives.

****

Universal Option Models

hengshuai yao, Csaba Szepesvari, Richard S. Sutton, Joseph Modayil, Shalabh Bhatnagar

We consider the problem of learning models of options for real-time abstract planning, in the setting where reward functions can be specified at any time and their expected returns must be efficiently computed. We introduce a new model for an option that is independent of any reward function, called the {\it universal option model (UOM)}. We prove that the UOM of an option can construct a traditional option model given a reward function, and the option-conditional return is computed directly by a single dot-product of the UOM with the reward function. We extend the UOM to linear function approximation, and we show it gives the TD solution of option returns and value functions of policies over options. We provide a stochastic approximation algorithm for incrementally learning UOMs from data and prove its consistency. We demonstrate our method in two domains. The first domain is document recommendation, where each user query defines a new reward function and a document's relevance is the expected return of a simulated random-walk through the document's references. The second domain is a real-time strategy game, where the controller must select the best game unit to accomplish dynamically-specified tasks. Our experiments show that UOMs are substantially more efficient in evaluating option returns and policies than previously known methods.

****

Discovering, Learning and Exploiting Relevance

Cem Tekin, Mihaela van der Schaar

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.

****

Stochastic Network Design in Bidirected Trees

xiaojian wu, Daniel R. Sheldon, Shlomo Zilberstein

We investigate the problem of stochastic network design in bidirected trees. In this problem, an underlying phenomenon (e.g., a behavior, rumor, or disease) starts at multiple sources in a tree and spreads in both directions along its edges. Actions can be taken to increase the probability of propagation on edges, and the goal is to maximize the total amount of spread away from all sources. Our main result is a rounded dynamic programming approach that leads to a fully polynomial-time approximation scheme (FPTAS), that is, an algorithm that can find $(1-\varepsilon)$-optimal solutions for any problem instance in time polynomial in the input size and $1/\varepsilon$. Our algorithm outperforms competing approaches on a motivating problem from computational sustainability to remove barriers in river networks to restore the health of aquatic ecosystems.

****

Distributed Variational Inference in Sparse Gaussian Process Regression and Latent Variable Models

Yarin Gal, Mark van der Wilk, Carl Edward Rasmussen

Gaussian processes (GPs) are a powerful tool for probabilistic inference over functions. They have been applied to both regression and non-linear dimensionality reduction, and offer desirable properties such as uncertainty estimates, robust

ness to over-fitting, and principled ways for tuning hyper-parameters. However t
he scalability of these models to big datasets remains an active topic of resear
ch. We introduce a novel re-parametrisation of variational inference for sparse
GP regression and latent variable models that allows for an efficient distribute
d algorithm. This is done by exploiting the decoupling of the data given the ind
ucing points to re-formulate the evidence lower bound in a Map-Reduce setting. W
e show that the inference scales well with data and computational resources, whi
le preserving a balanced distribution of the load among the nodes. We further de
monstrate the utility in scaling Gaussian processes to big data. We show that GP
 performance improves with increasing amounts of data in regression (on flight d
ata with 2 million records) and latent variable modelling (on MNIST). The result
s show that GPs perform better than many common models often used for big data.
****

On the Convergence Rate of Decomposable Submodular Function Minimization
Robert Nishihara, Stefanie Jegelka, Michael I. Jordan
Submodular functions describe a variety of discrete problems in machine learning
, signal processing, and computer vision. However, minimizing submodular functio
ns poses a number of algorithmic challenges. Recent work introduced an easy-to-u
se, parallelizable algorithm for minimizing submodular functions that decompose
as the sum of simple" submodular functions. Empirically, this algorithm performs
 extremely well, but no theoretical analysis was given. In this paper, we show t
hat the algorithm converges linearly, and we provide upper and lower bounds on t
he rate of convergence. Our proof relies on the geometry of submodular polyhedra
 and draws on results from spectral graph theory."
****

Efficient Inference of Continuous Markov Random Fields with Polynomial Potential
s
Shenlong Wang, Alex Schwing, Raquel Urtasun
In this paper, we prove that every multivariate polynomial with even degree can
be decomposed into a sum of convex and concave polynomials. Motivated by this pr
operty, we exploit the concave-convex procedure to perform inference on continuo
us Markov random fields with polynomial potentials. In particular, we show that
the concave-convex decomposition of polynomials can be expressed as a sum-of-squ
ares optimization, which can be efficiently solved via semidefinite programming.
 We demonstrate the effectiveness of our approach in the context of 3D reconstru
ction, shape from shading and image denoising, and show that our approach signif
icantly outperforms existing approaches in terms of efficiency as well as the qu
ality of the retrieved solution.
****

Metric Learning for Temporal Sequence Alignment
Damien Garreau, Rémi Lajugie, Sylvain Arlot, Francis Bach
In this paper, we propose to learn a Mahalanobis distance to perform alignment o
f multivariate time series. The learning examples for this task are time series
for which the true alignment is known. We cast the alignment problem as a struct
ured prediction task, and propose realistic losses between alignments for which
the optimization is tractable. We provide experiments on real data in the audio-
to-audio context, where we show that the learning of a similarity measure leads
to improvements in the performance of the alignment task. We also propose to use
 this metric learning framework to perform feature selection and, from basic aud
io features, build a combination of these with better alignment performance.
****

Extended and Unscented Gaussian Processes
Daniel M. Steinberg, Edwin V. Bonilla
We present two new methods for inference in Gaussian process (GP) models with ge
neral nonlinear likelihoods. Inference is based on a variational framework where
 a Gaussian posterior is assumed and the likelihood is linearized about the vari
ational posterior mean using either a Taylor series expansion or statistical lin
earization. We show that the parameter updates obtained by these algorithms are
equivalent to the state update equations in the iterative extended and unscented
 Kalman filters respectively, hence we refer to our algorithms as extended and u

nscented GPs. The unscented GP treats the likelihood as a 'black-box' by not req
uiring its derivative for inference, so it also applies to non-differentiable li
kelihood models. We evaluate the performance of our algorithms on a number of sy
nthetic inversion problems and a binary classification dataset.
****

A State-Space Model for Decoding Auditory Attentional Modulation from MEG in a C
ompeting-Speaker Environment
Sahar Akram, Jonathan Z. Simon, Shihab A. Shamma, Behtash Babadi
Humans are able to segregate auditory objects in a complex acoustic scene, throu
gh an interplay of bottom-up feature extraction and top-down selective attention
 in the brain. The detailed mechanism underlying this process is largely unknown
 and the ability to mimic this procedure is an important problem in artificial i
ntelligence and computational neuroscience. We consider the problem of decoding
the attentional state of a listener in a competing-speaker environment from magn
etoencephalographic (MEG) recordings from the human brain. We develop a behavior
ally inspired state-space model to account for the modulation of the MEG with re
spect to attentional state of the listener. We construct a decoder based on the
maximum a posteriori (MAP) estimate of the state parameters via the Expectation-
Maximization (EM) algorithm. The resulting decoder is able to track the attentio
nal modulation of the listener with multi-second resolution using only the envel
opes of the two speech streams as covariates. We present simulation studies as w
ell as application to real MEG data from two human subjects. Our results reveal
that the proposed decoder provides substantial gains in terms of temporal resolu
tion, complexity, and decoding accuracy.
****

Sensory Integration and Density Estimation
Joseph G. Makin, Philip N. Sabes
The integration of partially redundant information from multiple sensors is a st
andard computational problem for agents interacting with the world. In man and o
ther primates, integration has been shown psychophysically to be nearly optimal
in the sense of error minimization. An influential generalization of this notion
 of optimality is that populations of multisensory neurons should retain all the
 information from their unisensory afferents about the underlying, common stimul
us [1]. More recently, it was shown empirically that a neural network trained to
 perform latent-variable density estimation, with the activities of the unisenso
ry neurons as observed data, satisfies the information-preservation criterion, e
ven though the model architecture was not designed to match the true generative
process for the data [2]. We prove here an analytical connection between these s
eemingly different tasks, density estimation and sensory integration; that the f
ormer implies the latter for the model used in [2]; but that this does not appea
r to be true for all models.
****

Causal Strategic Inference in Networked Microfinance Economies
Mohammad T. Irfan, Luis E. Ortiz
Performing interventions is a major challenge in economic policy-making. We prop
ose \emph{causal strategic inference} as a framework for conducting intervention
s and apply it to large, networked microfinance economies. The basic solution pl
atform consists of modeling a microfinance market as a networked economy, learni
ng the parameters of the model from the real-world microfinance data, and design
ing algorithms for various computational problems in question. We adopt Nash equ
ilibrium as the solution concept for our model. For a special case of our model,
 we show that an equilibrium point always exists and that the equilibrium intere
st rates are unique. For the general case, we give a constructive proof of the e
xistence of an equilibrium point. Our empirical study is based on the microfinan
ce data from Bangladesh and Bolivia, which we use to first learn our models. We
show that causal strategic inference can assist policy-makers by evaluating the
outcomes of various types of interventions, such as removing a loss-making bank
from the market, imposing an interest rate cap, and subsidizing banks.
****

Exact Post Model Selection Inference for Marginal Screening

Jason D. Lee, Jonathan E. Taylor
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.
****

Sequence to Sequence Learning with Neural Networks
Ilya Sutskever, Oriol Vinyals, Quoc V. Le
Deep Neural Networks (DNNs) are powerful models that have achieved excellent performance on difficult learning tasks. Although DNNs work well whenever large labeled training sets are available, they cannot be used to map sequences to sequences. In this paper, we present a general end-to-end approach to sequence learning that makes minimal assumptions on the sequence structure. Our method uses a multilayered Long Short-Term Memory (LSTM) to map the input sequence to a vector of a fixed dimensionality, and then another deep LSTM to decode the target sequence from the vector. Our main result is that on an English to French translation task from the WMT-14 dataset, the translations produced by the LSTM achieve a BLEU score of 34.8 on the entire test set, where the LSTM's BLEU score was penalized on out-of-vocabulary words. Additionally, the LSTM did not have difficulty on long sentences. For comparison, a phrase-based SMT system achieves a BLEU score of 33.3 on the same dataset. When we used the LSTM to rerank the 1000 hypotheses produced by the aforementioned SMT system, its BLEU score increases to 36.5, which is close to the previous state of the art. The LSTM also learned sensible phrase and sentence representations that are sensitive to word order and are relatively invariant to the active and the passive voice. Finally, we found that reversing the order of the words in all source sentences (but not target sentences) improved the LSTM's performance markedly, because doing so introduced many short term dependencies between the source and the target sentence which made the optimization problem easier.
****

Information-based learning by agents in unbounded state spaces
Shariq A. Mobin, James A. Arnemann, Fritz Sommer
The idea that animals might use information-driven planning to explore an unknown environment and build an internal model of it has been proposed for quite some time. Recent work has demonstrated that agents using this principle can efficiently learn models of probabilistic environments with discrete, bounded state spaces. However, animals and robots are commonly confronted with unbounded environments. To address this more challenging situation, we study information-based learning strategies of agents in unbounded state spaces using non-parametric Bayesian models. Specifically, we demonstrate that the Chinese Restaurant Process (CRP) model is able to solve this problem and that an Empirical Bayes version is able to efficiently explore bounded and unbounded worlds by relying on little prior information.
****

Multi-Resolution Cascades for Multiclass Object Detection
Mohammad Saberian, Nuno Vasconcelos
An algorithm for learning fast multiclass object detection cascades is introduced. It produces multi-resolution (MRes) cascades, whose early stages are binary target vs. non-target detectors that eliminate false positives, late stages multiclass classifiers that finely discriminate target classes, and middle stages have intermediate numbers of classes, determined in a data-driven manner. This MRes structure is achieved with a new structurally biased boosting algorithm (SBBoost). SBBost extends previous multiclass boosting approaches, whose boosting mechanisms are shown to implement two complementary data-driven biases: 1) the standard bias towards examples difficult to classify, and 2) a bias towards difficult classes. It is shown that structural biases can be implemented by generalizing this class-based bias, so as to encourage the desired MRes structure. This is accomplished through a generalized definition of multiclass margin, which includes a set of bias parameters. SBBoost is a boosting algorithm for maximization of this margin. It can also be interpreted as standard multiclass boosting algorithm

augmented with margin thresholds or a cost-sensitive boosting algorithm with cos ts defined by the bias parameters. A stage adaptive bias policy is then introduc ed to determine bias parameters in a data driven manner. This is shown to produc e MRes cascades that have high detection rate and are computationally efficient. Experiments on multiclass object detection show improved performance over previ ous solutions.

****

Generalized Dantzig Selector: Application to the k-support norm

Soumyadeep Chatterjee, Sheng Chen, Arindam Banerjee

We propose a Generalized Dantzig Selector (GDS) for linear models, in which any norm encoding the parameter structure can be leveraged for estimation. We invest igate both computational and statistical aspects of the GDS. Based on conjugate proximal operator, a flexible inexact ADMM framework is designed for solving GDS . Thereafter, non-asymptotic high-probability bounds are established on the esti mation error, which rely on Gaussian widths of the unit norm ball and the error set. Further, we consider a non-trivial example of the GDS using k-support norm. We derive an efficient method to compute the proximal operator for k-support no rm since existing methods are inapplicable in this setting. For statistical anal ysis, we provide upper bounds for the Gaussian widths needed in the GDS analysis , yielding the first statistical recovery guarantee for estimation with the k-su pport norm. The experimental results confirm our theoretical analysis.

****

Conditional Swap Regret and Conditional Correlated Equilibrium

Mehryar Mohri, Scott Yang

We introduce a natural extension of the notion of swap regret, conditional swap regret, that allows for action modifications conditioned on the player’s action history. We prove a series of new results for conditional swap regret minimizati on. We present algorithms for minimizing conditional swap regret with bounded co nditioning history. We further extend these results to the case where conditiona l swaps are considered only for a subset of actions. We also define a new notion of equilibrium, conditional correlated equilibrium, that is tightly connected t o the notion of conditional swap regret: when all players follow conditional swa p regret minimization strategies, then the empirical distribution approaches thi s equilibrium. Finally, we extend our results to the multi-armed bandit scenario .

****

Gaussian Process Volatility Model

Yue Wu, José Miguel Hernández-Lobato, Zoubin Ghahramani

The prediction of time-changing variances is an important task in the modeling o f financial data. Standard econometric models are often limited as they assume r igid functional relationships for the evolution of the variance. Moreover, funct ional parameters are usually learned by maximum likelihood, which can lead to ov erfitting. To address these problems we introduce GP-Vol, a novel non-parametric model for time-changing variances based on Gaussian Processes. This new model c an capture highly flexible functional relationships for the variances. Furthermo re, we introduce a new online algorithm for fast inference in GP-Vol. This metho d is much faster than current offline inference procedures and it avoids overfit ting problems by following a fully Bayesian approach. Experiments with financial data show that GP-Vol performs significantly better than current standard alter natives.

****

Fast Sampling-Based Inference in Balanced Neuronal Networks

Guillaume Hennequin, Laurence Aitchison, Mate Lengyel

Multiple lines of evidence support the notion that the brain performs probabilis tic inference in multiple cognitive domains, including perception and decision m aking. There is also evidence that probabilistic inference may be implemented in the brain through the (quasi-)stochastic activity of neural circuits, producing samples from the appropriate posterior distributions, effectively implementing a Markov chain Monte Carlo algorithm. However, time becomes a fundamental bottle neck in such sampling-based probabilistic representations: the quality of infere

nces depends on how fast the neural circuit generates new, uncorrelated samples from its stationary distribution (the posterior). We explore this bottleneck in a simple, linear-Gaussian latent variable model, in which posterior sampling can be achieved by stochastic neural networks with linear dynamics. The well-known Langevin sampling (LS) recipe, so far the only sampling algorithm for continuous variables of which a neural implementation has been suggested, naturally fits into this dynamical framework. However, we first show analytically and through simulations that the symmetry of the synaptic weight matrix implied by LS yields critically slow mixing when the posterior is high-dimensional. Next, using methods from control theory, we construct and inspect networks that are optimally fast, and hence orders of magnitude faster than LS, while being far more biologically plausible. In these networks, strong -- but transient -- selective amplification of external noise generates the spatially correlated activity fluctuations prescribed by the posterior. Intriguingly, although a detailed balance of excitation and inhibition is dynamically maintained, detailed balance of Markov chain steps in the resulting sampler is violated, consistent with recent findings on how statistical irreversibility can overcome the speed limitation of random walks in other domains.

****

Semi-Separable Hamiltonian Monte Carlo for Inference in Bayesian Hierarchical Models

Yichuan Zhang, Charles Sutton

Sampling from hierarchical Bayesian models is often difficult for MCMC methods, because of the strong correlations between the model parameters and the hyperparameters. Recent Riemannian manifold Hamiltonian Monte Carlo (RMHMC) methods have significant potential advantages in this setting, but are computationally expensive. We introduce a new RMHMC method, which we call semi-separable Hamiltonian Monte Carlo, which uses a specially designed mass matrix that allows the joint Hamiltonian over model parameters and hyperparameters to decompose into two simpler Hamiltonians. This structure is exploited by a new integrator which we call the alternating blockwise leapfrog algorithm. The resulting method can mix faster than simpler Gibbs sampling while being simpler and more efficient than previous instances of RMHMC.

****

Predicting Useful Neighborhoods for Lazy Local Learning

Aron Yu, Kristen Grauman

Lazy local learning methods train a classifier on the fly" at test time, using only a subset of the training instances that are most relevant to the novel test example. The goal is to tailor the classifier to the properties of the data surrounding the test example. Existing methods assume that the instances most useful for building the local model are strictly those closest to the test example. However, this fails to account for the fact that the success of the resulting classifier depends on the full distribution of selected training instances. Rather than simply gather the test example's nearest neighbors, we propose to predict the subset of training data that is jointly relevant to training its local model. We develop an approach to discover patterns between queries and their "good" neighborhoods using large-scale multi-label classification with compressed sensing. Given a novel test point, we estimate both the composition and size of the training subset likely to yield an accurate local model. We demonstrate the approach on image classification tasks on SUN and aPascal and show it outperforms traditional global and local approaches."

****

(Almost) No Label No Cry

Giorgio Patrini, Richard Nock, Paul Rivera, Tiberio Caetano

In Learning with Label Proportions (LLP), the objective is to learn a supervised classifier when, instead of labels, only label proportions for bags of observations are known. This setting has broad practical relevance, in particular for privacy preserving data processing. We first show that the mean operator, a statistic which aggregates all labels, is minimally sufficient for the minimization of many proper scoring losses with linear (or kernelized) classifiers without usin

g labels. We provide a fast learning algorithm that estimates the mean operator via a manifold regularizer with guaranteed approximation bounds. Then, we present an iterative learning algorithm that uses this as initialization. We ground this algorithm in Rademacher-style generalization bounds that fit the LLP setting, introducing a generalization of Rademacher complexity and a Label Proportion Complexity measure. This latter algorithm optimizes tractable bounds for the corresponding bag-empirical risk. Experiments are provided on fourteen domains, whose size ranges up to 300K observations. They display that our algorithms are scalable and tend to consistently outperform the state of the art in LLP. Moreover, in many cases, our algorithms compete with or are just percents of AUC away from the Oracle that learns knowing all labels. On the largest domains, half a dozen proportions can suffice, i.e. roughly 40K times less than the total number of labels.

****

Learning Mixtures of Submodular Functions for Image Collection Summarization

Sebastian Tschiatschek, Rishabh K. Iyer, Haochen Wei, Jeff A. Bilmes

We address the problem of image collection summarization by learning mixtures of submodular functions. We argue that submodularity is very natural to this problem, and we show that a number of previously used scoring functions are submodular — a property not explicitly mentioned in these publications. We provide classes of submodular functions capturing the necessary properties of summaries, namely coverage, likelihood, and diversity. To learn mixtures of these submodular functions as scoring functions, we formulate summarization as a supervised learning problem using large-margin structured prediction. Furthermore, we introduce a novel evaluation metric, which we call V-ROUGE, for automatic summary scoring. While a similar metric called ROUGE has been successfully applied to document summarization [14], no such metric was known for quantifying the quality of image collection summaries. We provide a new dataset consisting of 14 real-world image collections along with many human-generated ground truth summaries collected using mechanical turk. We also extensively compare our method with previously explored methods for this problem and show that our learning approach outperforms all competitors on this new dataset. This paper provides, to our knowledge, the first systematic approach for quantifying the problem of image collection summarization, along with a new dataset of image collections and human summaries.

****

Distributed Bayesian Posterior Sampling via Moment Sharing

Minjie Xu, Balaji Lakshminarayanan, Yee Whye Teh, Jun Zhu, Bo Zhang

We propose a distributed Markov chain Monte Carlo (MCMC) inference algorithm for large scale Bayesian posterior simulation. We assume that the dataset is partitioned and stored across nodes of a cluster. Our procedure involves an independent MCMC posterior sampler at each node based on its local partition of the data. Moment statistics of the local posteriors are collected from each sampler and propagated across the cluster using expectation propagation message passing with low communication costs. The moment sharing scheme improves posterior estimation quality by enforcing agreement among the samplers. We demonstrate the speed and inference quality of our method with empirical studies on Bayesian logistic regression and sparse linear regression with a spike-and-slab prior.

****

Proximal Quasi-Newton for Computationally Intensive L1-regularized M-estimators

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

We consider the class of optimization problems arising from computationally intensive L1-regularized M-estimators, where the function or gradient values are very expensive to compute. A particular instance of interest is the L1-regularized MLE for learning Conditional Random Fields (CRFs), which are a popular class of statistical models for varied structured prediction problems such as sequence labeling, alignment, and classification with label taxonomy. L1-regularized MLEs for CRFs are particularly expensive to optimize since computing the gradient values requires an expensive inference step. In this work, we propose the use of a carefully constructed proximal quasi-Newton algorithm for such computationally intensive M-estimation problems, where we employ an aggressive active set selectio

n technique. In a key contribution of the paper, we show that our proximal quasi -Newton algorithm is provably super-linearly convergent, even in the absence of strong convexity, by leveraging a restricted variant of strong convexity. In our experiments, the proposed algorithm converges considerably faster than current state-of-the-art on the problems of sequence labeling and hierarchical classific ation.

****

Fast Multivariate Spatio-temporal Analysis via Low Rank Tensor Learning
Mohammad Taha Bahadori, Qi (Rose) Yu, Yan Liu

Accurate and efficient analysis of multivariate spatio-temporal data is critical in climatology, geology, and sociology applications. Existing models usually as sume simple inter-dependence among variables, space, and time, and are computati onally expensive. We propose a unified low rank tensor learning framework for mu ltivariate spatio-temporal analysis, which can conveniently incorporate differen t properties in spatio-temporal data, such as spatial clustering and shared stru cture among variables. We demonstrate how the general framework can be applied t o cokriging and forecasting tasks, and develop an efficient greedy algorithm to solve the resulting optimization problem with convergence guarantee. We conduct experiments on both synthetic datasets and real application datasets to demonstr ate that our method is not only significantly faster than existing methods but a lso achieves lower estimation error.

****

Multi-scale Graphical Models for Spatio-Temporal Processes
firdaus janoos, Huseyin Denli, Niranjan Subrahmanya

Learning the dependency structure between spatially distributed observations of a spatio-temporal process is an important problem in many fields such as geology , geophysics, atmospheric sciences, oceanography, etc. . However, estimation of such systems is complicated by the fact that they exhibit dynamics at multiple s cales of space and time arising due to a combination of diffusion and convection /advection. As we show, time-series graphical models based on vector auto-regres sive processes are inef■cient in capturing such multi-scale structure. In this p aper, we present a hierarchical graphical model with physically derived priors t hat better represents the multi-scale character of these dynamical systems. We a lso propose algorithms to ef■ciently estimate the interaction structure from dat a. We demonstrate results on a general class of problems arising in exploration geophysics by discovering graphical structure that is physically meaningful and provide evidence of its advantages over alternative approaches.

****

Bregman Alternating Direction Method of Multipliers
Huahua Wang, Arindam Banerjee

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.

****

Bounded Regret for Finite-Armed Structured Bandits
Tor Lattimore, Remi Munos

We study a new type of K-armed bandit problem where the expected return of one a rm may depend on the returns of other arms. We present a new algorithm for this general class of problems and show that under certain circumstances it is possib le to achieve finite expected cumulative regret. We also give problem-dependent lower bounds on the cumulative regret showing that at least in special cases the new algorithm is nearly optimal.

****

Exponential Concentration of a Density Functional Estimator
Shashank Singh, Barnabas Poczos

We analyse a plug-in estimator for a large class of integral functionals of one or more continuous probability densities. This class includes important families of entropy, divergence, mutual information, and their conditional versions. For densities on the d-dimensional unit cube [0,1]^d that lie in a beta-Holder smoo

thness class, we prove our estimator converges at the rate O(n^(1/(beta+d))). Furthermore, we prove that the estimator obeys an exponential concentration inequality about its mean, whereas most previous related results have bounded only expected error of estimators. Finally, we demonstrate our bounds to the case of conditional Renyi mutual information.
****

Decomposing Parameter Estimation Problems
Khaled S. Refaat, Arthur Choi, Adnan Darwiche
We propose a technique for decomposing the parameter learning problem in Bayesian networks into independent learning problems. Our technique applies to incomplete datasets and exploits variables that are either hidden or observed in the given dataset. We show empirically that the proposed technique can lead to orders-of-magnitude savings in learning time. We explain, analytically and empirically, the reasons behind our reported savings, and compare the proposed technique to related ones that are sometimes used by inference algorithms.
****

Convex Optimization Procedure for Clustering: Theoretical Revisit
Changbo Zhu, Huan Xu, Chenlei Leng, Shuicheng Yan
In this paper, we present theoretical analysis of SON~-~a convex optimization procedure for clustering using a sum-of-norms (SON) regularization recently proposed in \cite{ICML2011Hocking_419,SON, Lindsten650707, pelckmans2005convex}. In particular, we show if the samples are drawn from two cubes, each being one cluster, then SON can provably identify the cluster membership provided that the distance between the two cubes is larger than a threshold which (linearly) depends on the size of the cube and the ratio of numbers of samples in each cluster. To the best of our knowledge, this paper is the first to provide a rigorous analysis to understand why and when SON works. We believe this may provide important insights to develop novel convex optimization based algorithms for clustering.
****

Submodular Attribute Selection for Action Recognition in Video
Jingjing Zheng, Zhuolin Jiang, Rama Chellappa, Jonathon P. Phillips
In real-world action recognition problems, low-level features cannot adequately characterize the rich spatial-temporal structures in action videos. In this work, we encode actions based on attributes that describes actions as high-level concepts: \textit{e.g.}, jump forward and motion in the air. We base our analysis on two types of action attributes. One type of action attributes is generated by humans. The second type is data-driven attributes, which is learned from data using dictionary learning methods. Attribute-based representation may exhibit high variance due to noisy and redundant attributes. We propose a discriminative and compact attribute-based representation by selecting a subset of discriminative attributes from a large attribute set. Three attribute selection criteria are proposed and formulated as a submodular optimization problem. A greedy optimization algorithm is presented and guaranteed to be at least (1-1/e)-approximation to the optimum. Experimental results on the Olympic Sports and UCF101 datasets demonstrate that the proposed attribute-based representation can significantly boost the performance of action recognition algorithms and outperform most recently proposed recognition approaches.
****

Quantized Estimation of Gaussian Sequence Models in Euclidean Balls
Yuancheng Zhu, John Lafferty
A central result in statistical theory is Pinsker's theorem, which characterizes the minimax rate in the normal means model of nonparametric estimation. In this paper, we present an extension to Pinsker's theorem where estimation is carried out under storage or communication constraints. In particular, we place limits on the number of bits used to encode an estimator, and analyze the excess risk in terms of this constraint, the signal size, and the noise level. We give sharp upper and lower bounds for the case of a Euclidean ball, which establishes the Pareto-optimal minimax tradeoff between storage and risk in this setting.
****

Large-Margin Convex Polytope Machine

Alex Kantchelian, Michael C. Tschantz, Ling Huang, Peter L. Bartlett, Anthony D. Joseph, J. D. Tygar
We present the Convex Polytope Machine (CPM), a novel non-linear learning algorithm for large-scale binary classification tasks. The CPM finds a large margin convex polytope separator which encloses one class. We develop a stochastic gradient descent based algorithm that is amenable to massive datasets, and augment it with a heuristic procedure to avoid sub-optimal local minima. Our experimental evaluations of the CPM on large-scale datasets from distinct domains (MNIST handwritten digit recognition, text topic, and web security) demonstrate that the CPM trains models faster, sometimes several orders of magnitude, than state-of-the-art similar approaches and kernel-SVM methods while achieving comparable or better classification performance. Our empirical results suggest that, unlike prior similar approaches, we do not need to control the number of sub-classifiers (sides of the polytope) to avoid overfitting.
****

A provable SVD-based algorithm for learning topics in dominant admixture corpus
Trapit Bansal, Chiranjib Bhattacharyya, Ravindran Kannan
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.
****

Divide-and-Conquer Learning by Anchoring a Conical Hull
Tianyi Zhou, Jeff A. Bilmes, Carlos Guestrin
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.
****

Discriminative Metric Learning by Neighborhood Gerrymandering
Shubhendu Trivedi, David Mcallester, Greg Shakhnarovich
We formulate the problem of metric learning for k nearest neighbor classification as a large margin structured prediction problem, with a latent variable representing the choice of neighbors and the task loss directly corresponding to classification error. We describe an efficient algorithm for exact loss augmented inference,and a fast gradient descent algorithm for learning in this model. The objective drives the metric to establish neighborhood boundaries that benefit the true class labels for the training points. Our approach, reminiscent of gerrymandering (redrawing of political boundaries to provide advantage to certain parties), is more direct in its handling of optimizing classification accuracy than those previously proposed. In experiments on a variety of data sets our method is shown to achieve excellent results compared to current state of the art in metric learning.
****

Learning on graphs using Orthonormal Representation is Statistically Consistent
Rakesh Shivanna, Chiranjib Bhattacharyya
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 Unsupervised Manifold Alignment
Zhen Cui, Hong Chang, Shiguang Shan, Xilin Chen
In this paper, we propose a generalized Unsupervised Manifold Alignment (GUMA) method to build the connections between different but correlated datasets without any known correspondences. Based on the assumption that datasets of the same theme usually have similar manifold structures, GUMA is formulated into an explicit integer optimization problem considering the structure matching and preserving criteria, as well as the feature comparability of the corresponding points in the mutual embedding space. The main benefits of this model include: (1) simultan

eous discovery and alignment of manifold structures; (2) fully unsupervised matc
hing without any pre-specified correspondences; (3) efficient iterative alignmen
t without computations in all permutation cases. Experimental results on dataset
 matching and real-world applications demonstrate the effectiveness and the prac
ticability of our manifold alignment method.
****

Grouping-Based Low-Rank Trajectory Completion and 3D Reconstruction
Katerina Fragkiadaki, Marta Salas, Pablo Arbelaez, Jitendra Malik
Extracting 3D shape of deforming objects in monocular videos, a task known as no
n-rigid structure-from-motion (NRSfM), has so far been studied only on synthetic
 datasets and controlled environments. Typically, the objects to reconstruct are
 pre-segmented, they exhibit limited rotations and occlusions, or full-length tr
ajectories are assumed. In order to integrate NRSfM into current video analysis
pipelines, one needs to consider as input realistic -thus incomplete- tracking,
and perform spatio-temporal grouping to segment the objects from their surroundi
ngs. Furthermore, NRSfM needs to be robust to noise in both segmentation and tra
cking, e.g., drifting, segmentation leaking'', optical flowbleeding'' etc. In th
is paper, we make a first attempt towards this goal, and propose a method that c
ombines dense optical flow tracking, motion trajectory clustering and NRSfM for
3D reconstruction of objects in videos. For each trajectory cluster, we compute
multiple reconstructions by minimizing the reprojection error and the rank of th
e 3D shape under different rank bounds of the trajectory matrix. We show that de
nse 3D shape is extracted and trajectories are completed across occlusions and l
ow textured regions, even under mild relative motion between the object and the
camera. We achieve competitive results on a public NRSfM benchmark while using f
ixed parameters across all sequences and handling incomplete trajectories, in co
ntrast to existing approaches. We further test our approach on popular video seg
mentation datasets. To the best of our knowledge, our method is the first to ext
ract dense object models from realistic videos, such as those found in Youtube o
r Hollywood movies, without object-specific priors.
****

A statistical model for tensor PCA
Emile Richard, Andrea Montanari
We consider the Principal Component Analysis problem for large tensors of arbitr
ary order k under a single-spike (or rank-one plus noise) model. On the one hand
, we use information theory, and recent results in probability theory to establi
sh necessary and sufficient conditions under which the principal component can b
e estimated using unbounded computational resources. It turns out that this is p
ossible as soon as the signal-to-noise ratio beta becomes larger than C\sqrt{k l
og k} (and in particular beta can remain bounded has the problem dimensions incr
ease). On the other hand, we analyze several polynomial-time estimation algorith
ms, based on tensor unfolding, power iteration and message passing ideas from gr
aphical models. We show that, unless the signal-to-noise ratio diverges in the s
ystem dimensions, none of these approaches succeeds. This is possibly related to
 a fundamental limitation of computationally tractable estimators for this probl
em. For moderate dimensions, we propose an hybrid approach that uses unfolding t
ogether with power iteration, and show that it outperforms significantly baselin
e methods. Finally, we consider the case in which additional side information is
 available about the unknown signal. We characterize the amount of side informat
ion that allow the iterative algorithms to converge to a good estimate.
****

Making Pairwise Binary Graphical Models Attractive
Nicholas Ruozzi, Tony Jebara
Computing the partition function (i.e., the normalizing constant) of a given pai
rwise binary graphical model is NP-hard in general. As a result, the partition f
unction is typically estimated by approximate inference algorithms such as belie
f propagation (BP) and tree-reweighted belief propagation (TRBP). The former pro
vides reasonable estimates in practice but has convergence issues. The later has
 better convergence properties but typically provides poorer estimates. In this
work, we propose a novel scheme that has better convergence properties than BP a

nd provably provides better partition function estimates in many instances than TRBP. In particular, given an arbitrary pairwise binary graphical model, we construct a specific ``attractive'' 2-cover. We explore the properties of this special cover and show that it can be used to construct an algorithm with the desired properties.
****

Subspace Embeddings for the Polynomial Kernel
Haim Avron, Huy Nguyen, David Woodruff
Sketching is a powerful dimensionality reduction tool for accelerating statistical learning algorithms. However, its applicability has been limited to a certain extent since the crucial ingredient, the so-called oblivious subspace embedding, can only be applied to data spaces with an explicit representation as the column span or row span of a matrix, while in many settings learning is done in a high-dimensional space implicitly defined by the data matrix via a kernel transformation. We propose the first {\em fast} oblivious subspace embeddings that are able to embed a space induced by a non-linear kernel {\em without} explicitly mapping the data to the high-dimensional space. In particular, we propose an embedding for mappings induced by the polynomial kernel. Using the subspace embeddings, we obtain the fastest known algorithms for computing an implicit low rank approximation of the higher-dimension mapping of the data matrix, and for computing an approximate kernel PCA of the data, as well as doing approximate kernel principal component regression.
****

On the relations of LFPs & Neural Spike Trains
David E. Carlson, Jana Schaich Borg, Kafui Dzirasa, Lawrence Carin
One of the goals of neuroscience is to identify neural networks that correlate with important behaviors, environments, or genotypes. This work proposes a strategy for identifying neural networks characterized by time- and frequency-dependent connectivity patterns, using convolutional dictionary learning that links spike-train data to local field potentials (LFPs) across multiple areas of the brain. Analytical contributions are: (i) modeling dynamic relationships between LFPs and spikes; (ii) describing the relationships between spikes and LFPs, by analyzing the ability to predict LFP data from one region based on spiking information from across the brain; and (iii) development of a clustering methodology that allows inference of similarities in neurons from multiple regions. Results are based on data sets in which spike and LFP data are recorded simultaneously from up to 16 brain regions in a mouse.
****

Online Optimization for Max-Norm Regularization
Jie Shen, Huan Xu, Ping Li
Max-norm regularizer has been extensively studied in the last decade as it promotes an effective low rank estimation of the underlying data. However, max-norm regularized problems are typically formulated and solved in a batch manner, which prevents it from processing big data due to possible memory bottleneck. In this paper, we propose an online algorithm for solving max-norm regularized problems that is scalable to large problems. Particularly, we consider the matrix decomposition problem as an example, although our analysis can also be applied in other problems such as matrix completion. The key technique in our algorithm is to reformulate the max-norm into a matrix factorization form, consisting of a basis component and a coefficients one. In this way, we can solve the optimal basis and coefficients alternatively. We prove that the basis produced by our algorithm converges to a stationary point asymptotically. Experiments demonstrate encouraging results for the effectiveness and robustness of our algorithm. See the full paper at arXiv:1406.3190.
****

Cone-Constrained Principal Component Analysis
Yash Deshpande, Andrea Montanari, Emile Richard
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.
****

Fast Training of Pose Detectors in the Fourier Domain
João F. Henriques, Pedro Martins, Rui F. Caseiro, Jorge Batista
In many datasets, the samples are related by a known image transformation, such as rotation, or a repeatable non-rigid deformation. This applies to both datasets with the same objects under different viewpoints, and datasets augmented with virtual samples. Such datasets possess a high degree of redundancy, because geometrically-induced transformations should preserve intrinsic properties of the objects. Likewise, ensembles of classifiers used for pose estimation should also share many characteristics, since they are related by a geometric transformation. By assuming that this transformation is norm-preserving and cyclic, we propose a closed-form solution in the Fourier domain that can eliminate most redundancies. It can leverage off-the-shelf solvers with no modification (e.g. libsvm), and train several pose classifiers simultaneously at no extra cost. Our experiments show that training a sliding-window object detector and pose estimator can be sped up by orders of magnitude, for transformations as diverse as planar rotation, the walking motion of pedestrians, and out-of-plane rotations of cars.
****

Optimistic Planning in Markov Decision Processes Using a Generative Model
Balázs Szörényi, Gunnar Kedenburg, Remi Munos
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****

Bayesian Nonlinear Support Vector Machines and Discriminative Factor Modeling
Ricardo Henao, Xin Yuan, Lawrence Carin
A new Bayesian formulation is developed for nonlinear support vector machines (SVMs), based on a Gaussian process and with the SVM hinge loss expressed as a scaled mixture of normals. We then integrate the Bayesian SVM into a factor model, in which feature learning and nonlinear classifier design are performed jointly; almost all previous work on such discriminative feature learning has assumed a linear classifier. Inference is performed with expectation conditional maximization (ECM) and Markov Chain Monte Carlo (MCMC). An extensive set of experiments demonstrate the utility of using a nonlinear Bayesian SVM within discriminative feature learning and factor modeling, from the standpoints of accuracy and interpretability
****

Feedback Detection for Live Predictors
Stefan Wager, Nick Chamandy, Omkar Muralidharan, Amir Najmi
A predictor that is deployed in a live production system may perturb the features it uses to make predictions. Such a feedback loop can occur, for example, when a model that predicts a certain type of behavior ends up causing the behavior it predicts, thus creating a self-fulfilling prophecy. In this paper we analyze predictor feedback detection as a causal inference problem, and introduce a local randomization scheme that can be used to detect non-linear feedback in real-world problems. We conduct a pilot study for our proposed methodology using a predictive system currently deployed as a part of a search engine.
****

Do Convnets Learn Correspondence?
Jonathan L. Long, Ning Zhang, Trevor Darrell
Convolutional neural nets (convnets) trained from massive labeled datasets have substantially improved the state-of-the-art in image classification and object detection. However, visual understanding requires establishing correspondence on a finer level than object category. Given their large pooling regions and training from whole-image labels, it is not clear that convnets derive their success from an accurate correspondence model which could be used for precise localization. In this paper, we study the effectiveness of convnet activation features for tasks requiring correspondence. We present evidence that convnet features locali

ze at a much finer scale than their receptive field sizes, that they can be used to perform intraclass aligment as well as conventional hand-engineered features, and that they outperform conventional features in keypoint prediction on objects from PASCAL VOC 2011.
****
Convolutional Neural Network Architectures for Matching Natural Language Sentences
Baotian Hu, Zhengdong Lu, Hang Li, Qingcai Chen
Semantic matching is of central importance to many natural language tasks \cite{bordes2014semantic,RetrievalQA}. A successful matching algorithm needs to adequately model the internal structures of language objects and the interaction between them. As a step toward this goal, we propose convolutional neural network models for matching two sentences, by adapting the convolutional strategy in vision and speech. The proposed models not only nicely represent the hierarchical structures of sentences with their layer-by-layer composition and pooling, but also capture the rich matching patterns at different levels. Our models are rather generic, requiring no prior knowledge on language, and can hence be applied to matching tasks of different nature and in different languages. The empirical study on a variety of matching tasks demonstrates the efficacy of the proposed model on a variety of matching tasks and its superiority to competitor models.
****
Conditional Random Field Autoencoders for Unsupervised Structured Prediction
Waleed Ammar, Chris Dyer, Noah A. Smith
We introduce a framework for unsupervised learning of structured predictors with overlapping, global features. Each input's latent representation is predicted conditional on the observed data using a feature-rich conditional random field (CRF). Then a reconstruction of the input is (re)generated, conditional on the latent structure, using a generative model which factorizes similarly to the CRF. The autoencoder formulation enables efficient exact inference without resorting to unrealistic independence assumptions or restricting the kinds of features that can be used. We illustrate insightful connections to traditional autoencoders, posterior regularization and multi-view learning. Finally, we show competitive results with instantiations of the framework for two canonical tasks in natural language processing: part-of-speech induction and bitext word alignment, and show that training our model can be substantially more efficient than comparable feature-rich baselines.
****
Optimal rates for k-NN density and mode estimation
Sanjoy Dasgupta, Samory Kpotufe
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.
****
Coresets for k-Segmentation of Streaming Data
Guy Rosman, Mikhail Volkov, Dan Feldman, John W. Fisher III, Daniela Rus
Life-logging video streams, financial time series, and Twitter tweets are a few examples of high-dimensional signals over practically unbounded time. We consider the problem of computing optimal segmentation of such signals by k-piecewise linear function, using only one pass over the data by maintaining a coreset for the signal. The coreset enables fast further analysis such as automatic summarization and analysis of such signals. A coreset (core-set) is a compact representation of the data seen so far, which approximates the data well for a specific task -- in our case, segmentation of the stream. We show that, perhaps surprisingly, the segmentation problem admits coresets of cardinality only linear in the number of segments k, independently of both the dimension d of the signal, and its number n of points. More precisely, we construct a representation of size O(klog n /eps^2) that provides a (1+eps)-approximation for the sum of squared distances to any given k-piecewise linear function. Moreover, such coresets can be constructed in a parallel streaming approach. Our results rely on a novel eduction of

statistical estimations to problems in computational geometry. We empirically e
valuate our algorithms on very large synthetic and real data sets from GPS, vide
o and financial domains, using 255 machines in Amazon cloud.
****

Message Passing Inference for Large Scale Graphical Models with High Order Poten
tials
Jian Zhang, Alex Schwing, Raquel Urtasun
To keep up with the Big Data challenge, parallelized algorithms based on dual de
composition have been proposed to perform inference in Markov random fields. Des
pite this parallelization, current algorithms struggle when the energy has high
order terms and the graph is densely connected. In this paper we propose a parti
tioning strategy followed by a message passing algorithm which is able to exploi
t pre-computations. It only updates the high-order factors when passing messages
across machines. We demonstrate the effectiveness of our approach on the task o
f joint layout and semantic segmentation estimation from single images, and show
that our approach is orders of magnitude faster than current methods.
****

Weighted importance sampling for off-policy learning with linear function approx
imation
A. Rupam Mahmood, Hado P. van Hasselt, Richard S. Sutton
Importance sampling is an essential component of off-policy model-free reinforce
ment learning algorithms. However, its most effective variant, \emph{weighted} i
mportance sampling, does not carry over easily to function approximation and, be
cause of this, it is not utilized in existing off-policy learning algorithms. In
this paper, we take two steps toward bridging this gap. First, we show that wei
ghted importance sampling can be viewed as a special case of weighting the error
of individual training samples, and that this weighting has theoretical and emp
irical benefits similar to those of weighted importance sampling. Second, we sho
w that these benefits extend to a new weighted-importance-sampling version of of
f-policy LSTD(lambda). We show empirically that our new WIS-LSTD(lambda) algorit
hm can result in much more rapid and reliable convergence than conventional off-
policy LSTD(lambda) (Yu 2010, Bertsekas & Yu 2009).
****

Incremental Clustering: The Case for Extra Clusters
Margareta Ackerman, Sanjoy Dasgupta
The explosion in the amount of data available for analysis often necessitates a
transition from batch to incremental clustering methods, which process one eleme
nt at a time and typically store only a small subset of the data. In this paper,
we initiate the formal analysis of incremental clustering methods focusing on t
he types of cluster structure that they are able to detect. We find that the inc
remental setting is strictly weaker than the batch model, proving that a fundame
ntal class of cluster structures that can readily be detected in the batch setti
ng is impossible to identify using any incremental method. Furthermore, we show
how the limitations of incremental clustering can be overcome by allowing additi
onal clusters.
****

Positive Curvature and Hamiltonian Monte Carlo
Christof Seiler, Simon Rubinstein-Salzedo, Susan Holmes
The Jacobi metric introduced in mathematical physics can be used to analyze Hami
ltonian Monte Carlo (HMC). In a geometrical setting, each step of HMC correspond
s to a geodesic on a Riemannian manifold with a Jacobi metric. Our calculation o
f the sectional curvature of this HMC manifold allows us to see that it is posit
ive in cases such as sampling from a high dimensional multivariate Gaussian. We
show that positive curvature can be used to prove theoretical concentration resu
lts for HMC Markov chains.
****

Inferring sparse representations of continuous signals with continuous orthogona
l matching pursuit
Karin C. Knudson, Jacob Yates, Alexander Huk, Jonathan W. Pillow
Many signals, such as spike trains recorded in multi-channel electrophysiologica

l recordings, may be represented as the sparse sum of translated and scaled copi es of waveforms whose timing and amplitudes are of interest. From the aggregate signal, one may seek to estimate the identities, amplitudes, and translations of the waveforms that compose the signal. Here we present a fast method for recove ring these identities, amplitudes, and translations. The method involves greedil y selecting component waveforms and then refining estimates of their amplitudes and translations, moving iteratively between these steps in a process analogous to the well-known Orthogonal Matching Pursuit (OMP) algorithm. Our approach for modeling translations borrows from Continuous Basis Pursuit (CBP), which we exte nd in several ways: by selecting a subspace that optimally captures translated c opies of the waveforms, replacing the convex optimization problem with a greedy approach, and moving to the Fourier domain to more precisely estimate time shift s. We test the resulting method, which we call Continuous Orthogonal Matching Pu rsuit (COMP), on simulated and neural data, where it shows gains over CBP in bot h speed and accuracy.

****

Zeta Hull Pursuits: Learning Nonconvex Data Hulls

Yuanjun Xiong, Wei Liu, Deli Zhao, Xiaoou Tang

Selecting a small informative subset from a given dataset, also called column sa mpling, has drawn much attention in machine learning. For incorporating structur ed data information into column sampling, research efforts were devoted to the c ases where data points are fitted with clusters, simplices, or general convex hu lls. This paper aims to study nonconvex hull learning which has rarely been inve stigated in the literature. In order to learn data-adaptive nonconvex hulls, a n ovel approach is proposed based on a graph-theoretic measure that leverages grap h cycles to characterize the structural complexities of input data points. Emplo ying this measure, we present a greedy algorithmic framework, dubbed Zeta Hulls, to perform structured column sampling. The process of pursuing a Zeta hull invo lves the computation of matrix inverse. To accelerate the matrix inversion compu tation and reduce its space complexity as well, we exploit a low-rank approximat ion to the graph adjacency matrix by using an efficient anchor graph technique. Extensive experimental results show that data representation learned by Zeta Hul ls can achieve state-of-the-art accuracy in text and image classification tasks.

****

Near-Optimal-Sample Estimators for Spherical Gaussian Mixtures

Ananda Theertha Suresh, Alon Orlitsky, Jayadev Acharya, Ashkan Jafarpour

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.

****

Provable Tensor Factorization with Missing Data

Prateek Jain, Sewoong Oh

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.

****

Learning Generative Models with Visual Attention

Charlie Tang, Nitish Srivastava, Russ R. Salakhutdinov

Attention has long been proposed by psychologists to be important for efficientl y dealing with the massive amounts of sensory stimulus in the neocortex. Inspire d by the attention models in visual neuroscience and the need for object-centere d data for generative models, we propose a deep-learning based generative framew ork using attention. The attentional mechanism propagates signals from the regio n of interest in a scene to an aligned canonical representation for generative m odeling. By ignoring scene background clutter, the generative model can concentr ate its resources on the object of interest. A convolutional neural net is emplo yed to provide good initializations during posterior inference which uses Hamilt onian Monte Carlo. Upon learning images of faces, our model can robustly attend

to the face region of novel test subjects. More importantly, our model can learn generative models of new faces from a novel dataset of large images where the face locations are not known.

****

Bandit Convex Optimization: Towards Tight Bounds

Elad Hazan, Kfir Levy

Bandit Convex Optimization (BCO) is a fundamental framework for decision making under uncertainty, which generalizes many problems from the realm of online and statistical learning. While the special case of linear cost functions is well understood, a gap on the attainable regret for BCO with nonlinear losses remains an important open question. In this paper we take a step towards understanding the best attainable regret bounds for BCO: we give an efficient and near-optimal regret algorithm for BCO with strongly-convex and smooth loss functions. In contrast to previous works on BCO that use time invariant exploration schemes, our method employs an exploration scheme that shrinks with time.

****

A Latent Source Model for Online Collaborative Filtering

Guy Bresler, George H. Chen, 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.

****

Self-Paced Learning with Diversity

Lu Jiang, Deyu Meng, Shoou-I Yu, Zhenzhong Lan, Shiguang Shan, Alexander Hauptmann

Self-paced learning (SPL) is a recently proposed learning regime inspired by the learning process of humans and animals that gradually incorporates easy to more complex samples into training. Existing methods are limited in that they ignore an important aspect in learning: diversity. To incorporate this information, we propose an approach called self-paced learning with diversity (SPLD) which formalizes the preference for both easy and diverse samples into a general regularizer. This regularization term is independent of the learning objective, and thus can be easily generalized into various learning tasks. Albeit non-convex, the optimization of the variables included in this SPLD regularization term for sample selection can be globally solved in linearithmic time. We demonstrate that our method significantly outperforms the conventional SPL on three real-world datasets. Specifically, SPLD achieves the best MAP so far reported in literature on the Hollywood2 and Olympic Sports datasets.

****

Inference by Learning: Speeding-up Graphical Model Optimization via a Coarse-to-Fine Cascade of Pruning Classifiers

Bruno Conejo, Nikos Komodakis, Sebastien Leprince, Jean Philippe Avouac

We propose a general and versatile framework that significantly speeds-up graphical model optimization while maintaining an excellent solution accuracy. The proposed approach, refereed as Inference by Learning or IbyL, relies on a multi-scale pruning scheme that progressively reduces the solution space by use of a coarse-to-fine cascade of learnt classifiers. We thoroughly experiment with classic computer vision related MRF problems, where our novel framework constantly yields a significant time speed-up (with respect to the most efficient inference methods) and obtains a more accurate solution than directly optimizing the MRF. We make our code available on-line.

****

Capturing Semantically Meaningful Word Dependencies with an Admixture of Poisson MRFs

David I. Inouye, 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.

****
Kernel Mean Estimation via Spectral Filtering
Krikamol Muandet, Bharath Sriperumbudur, Bernhard Schölkopf
The problem of estimating the kernel mean in a reproducing kernel Hilbert space (RKHS) is central to kernel methods in that it is used by classical approaches (e.g., when centering a kernel PCA matrix), and it also forms the core inference step of modern kernel methods (e.g., kernel-based non-parametric tests) that rel y on embedding probability distributions in RKHSs. Previous work [1] has shown t hat shrinkage can help in constructing “better” estimators of the kernel mean th an the empirical estimator. The present paper studies the consistency and admiss ibility of the estimators in [1], and proposes a wider class of shrinkage estima tors that improve upon the empirical estimator by considering appropriate basis functions. Using the kernel PCA basis, we show that some of these estimators can be constructed using spectral filtering algorithms which are shown to be consis tent under some technical assumptions. Our theoretical analysis also reveals a f undamental connection to the kernel-based supervised learning framework. The pro posed estimators are simple to implement and perform well in practice.
****
A Dual Algorithm for Olfactory Computation in the Locust Brain
Sina Tootoonian, Mate Lengyel
We study the early locust olfactory system in an attempt to explain its well-cha racterized structure and dynamics. We first propose its computational function a s recovery of high-dimensional sparse olfactory signals from a small number of m easurements. Detailed experimental knowledge about this system rules out standar d algorithmic solutions to this problem. Instead, we show that solving a dual fo rmulation of the corresponding optimisation problem yields structure and dynamic s in good agreement with biological data. Further biological constraints lead us to a reduced form of this dual formulation in which the system uses independent component analysis to continuously adapt to its olfactory environment to allow accurate sparse recovery. Our work demonstrates the challenges and rewards of at tempting detailed understanding of experimentally well-characterized systems.
****
Online combinatorial optimization with stochastic decision sets and adversarial losses
Gergely Neu, Michal Valko
Most work on sequential learning assumes a fixed set of actions that are availab le all the time. However, in practice, actions can consist of picking subsets of readings from sensors that may break from time to time, road segments that can be blocked or goods that are out of stock. In this paper we study learning algor ithms that are able to deal with stochastic availability of such unreliable comp osite actions. We propose and analyze algorithms based on the Follow-The-Perturb ed-Leader prediction method for several learning settings differing in the feedb ack provided to the learner. Our algorithms rely on a novel loss estimation tech nique that we call Counting Asleep Times. We deliver regret bounds for our algor ithms for the previously studied full information and (semi-)bandit settings, as well as a natural middle point between the two that we call the restricted info rmation setting. A special consequence of our results is a significant improveme nt of the best known performance guarantees achieved by an efficient algorithm f or the sleeping bandit problem with stochastic availability. Finally, we evaluat e our algorithms empirically and show their improvement over the known approache s.
****
Learning Multiple Tasks in Parallel with a Shared Annotator
Haim Cohen, Koby Crammer
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.
****
A Complete Variational Tracker

Ryan D. Turner, Steven Bottone, Bhargav Avasarala
We introduce a novel probabilistic tracking algorithm that incorporates combinat
orial data association constraints and model-based track management using variat
ional Bayes. We use a Bethe entropy approximation to incorporate data associatio
n constraints that are often ignored in previous probabilistic tracking algorith
ms. Noteworthy aspects of our method include a model-based mechanism to replace
heuristic logic typically used to initiate and destroy tracks, and an assignment
 posterior with linear computation cost in window length as opposed to the expon
ential scaling of previous MAP-based approaches. We demonstrate the applicabilit
y of our method on radar tracking and computer vision problems.
****

Low-dimensional models of neural population activity in sensory cortical circuit
s
Evan W. Archer, Urs Koster, Jonathan W. Pillow, Jakob H. Macke
Neural responses in visual cortex are influenced by visual stimuli and by ongoin
g spiking activity in local circuits. An important challenge in computational ne
uroscience is to develop models that can account for both of these features in l
arge multi-neuron recordings and to reveal how stimulus representations interact
 with and depend on cortical dynamics. Here we introduce a statistical model of
neural population activity that integrates a nonlinear receptive field model wit
h a latent dynamical model of ongoing cortical activity. This model captures the
 temporal dynamics, effective network connectivity in large population recording
s, and correlations due to shared stimulus drive as well as common noise. Moreov
er, because the nonlinear stimulus inputs are mixed by the ongoing dynamics, the
 model can account for a relatively large number of idiosyncratic receptive fiel
d shapes with a small number of nonlinear inputs to a low-dimensional latent dyn
amical model. We introduce a fast estimation method using online expectation max
imization with Laplace approximations. Inference scales linearly in both populat
ion size and recording duration. We apply this model to multi-channel recordings
 from primary visual cortex and show that it accounts for a large number of indi
vidual neural receptive fields using a small number of nonlinear inputs and a lo
w-dimensional dynamical model.
****

Spectral k-Support Norm Regularization
Andrew M. McDonald, Massimiliano Pontil, Dimitris Stamos
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.
****

Learning Optimal Commitment to Overcome Insecurity
Avrim Blum, Nika Haghtalab, Ariel D. Procaccia
Game-theoretic algorithms for physical security have made an impressive real-wor
ld impact. These algorithms compute an optimal strategy for the defender to comm
it to in a Stackelberg game, where the attacker observes the defender's strategy
 and best-responds. In order to build the game model, though, the payoffs of pot
ential attackers for various outcomes must be estimated; inaccurate estimates ca
n lead to significant inefficiencies. We design an algorithm that optimizes the
defender's strategy with no prior information, by observing the attacker's respo
nses to randomized deployments of resources and learning his priorities. In cont
rast to previous work, our algorithm requires a number of queries that is polyno
mial in the representation of the game.
****

A Drifting-Games Analysis for Online Learning and Applications to Boosting
Haipeng Luo, Robert E. Schapire
We provide a general mechanism to design online learning algorithms based on a m
inimax analysis within a drifting-games framework. Different online learning set
tings (Hedge, multi-armed bandit problems and online convex optimization) are st
udied by converting into various kinds of drifting games. The original minimax a
nalysis for drifting games is then used and generalized by applying a series of

relaxations, starting from choosing a convex surrogate of the 0-1 loss function. With different choices of surrogates, we not only recover existing algorithms, but also propose new algorithms that are totally parameter-free and enjoy other useful properties. Moreover, our drifting-games framework naturally allows us to study high probability bounds without resorting to any concentration results, and also a generalized notion of regret that measures how good the algorithm is compared to all but the top small fraction of candidates. Finally, we translate our new Hedge algorithm into a new adaptive boosting algorithm that is computationally faster as shown in experiments, since it ignores a large number of examples on each round.

****

Dynamic Rank Factor Model for Text Streams

Shaobo Han, Lin Du, Esther Salazar, Lawrence Carin

We propose a semi-parametric and dynamic rank factor model for topic modeling, capable of (1) discovering topic prevalence over time, and (2) learning contemporary multi-scale dependence structures, providing topic and word correlations as a byproduct. The high-dimensional and time-evolving ordinal/rank observations (such as word counts), after an arbitrary monotone transformation, are well accommodated through an underlying dynamic sparse factor model. The framework naturally admits heavy-tailed innovations, capable of inferring abrupt temporal jumps in the importance of topics. Posterior inference is performed through straightforward Gibbs sampling, based on the forward-filtering backward-sampling algorithm. Moreover, an efficient data subsampling scheme is leveraged to speed up inference on massive datasets. The modeling framework is illustrated on two real datasets: the US State of the Union Address and the JSTOR collection from Science.

****

Consistency of weighted majority votes

Daniel Berend, Aryeh Kontorovich

We revisit from a statistical learning perspective the classical decision-theoretic problem of weighted expert voting. In particular, we examine the consistency (both asymptotic and finitary) of the optimal Nitzan-Paroush weighted majority and related rules. In the case of known expert competence levels, we give sharp error estimates for the optimal rule. When the competence levels are unknown, they must be empirically estimated. We provide frequentist and Bayesian analyses for this situation. Some of our proof techniques are non-standard and may be of independent interest. The bounds we derive are nearly optimal, and several challenging open problems are posed. Experimental results are provided to illustrate the theory.

****

Modeling Deep Temporal Dependencies with Recurrent Grammar Cells""

Vincent Michalski, Roland Memisevic, Kishore Konda

We propose modeling time series by representing the transformations that take a frame at time t to a frame at time t+1. To this end we show how a bi-linear model of transformations, such as a gated autoencoder, can be turned into a recurrent network, by training it to predict future frames from the current one and the inferred transformation using backprop-through-time. We also show how stacking multiple layers of gating units in a recurrent pyramid makes it possible to represent the ”syntax” of complicated time series, and that it can outperform standard recurrent neural networks in terms of prediction accuracy on a variety of tasks.

****

Augur: Data-Parallel Probabilistic Modeling

Jean-Baptiste Tristan, Daniel Huang, Joseph Tassarotti, Adam C. Pocock, Stephen Green, Guy L. Steele

Implementing inference procedures for each new probabilistic model is time-consuming and error-prone. Probabilistic programming addresses this problem by allowing a user to specify the model and then automatically generating the inference procedure. To make this practical it is important to generate high performance inference code. In turn, on modern architectures, high performance requires parallel execution. In this paper we present Augur, a probabilistic modeling language

and compiler for Bayesian networks designed to make effective use of data-parall
el architectures such as GPUs. We show that the compiler can generate data-paral
lel inference code scalable to thousands of GPU cores by making use of the condi
tional independence relationships in the Bayesian network.
****

Augmentative Message Passing for Traveling Salesman Problem and Graph Partitioni
ng

Siamak Ravanbakhsh, Reihaneh Rabbany, Russell Greiner

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.
****

Mondrian Forests: Efficient Online Random Forests

Balaji Lakshminarayanan, Daniel M. Roy, Yee Whye Teh

Ensembles of randomized decision trees, usually referred to as random forests, a
re widely used for classification and regression tasks in machine learning and s
tatistics. Random forests achieve competitive predictive performance and are com
putationally efficient to train and test, making them excellent candidates for r
eal-world prediction tasks. The most popular random forest variants (such as Bre
iman's random forest and extremely randomized trees) operate on batches of train
ing data. Online methods are now in greater demand. Existing online random fores
ts, however, require more training data than their batch counterpart to achieve
comparable predictive performance. In this work, we use Mondrian processes (Roy
and Teh, 2009) to construct ensembles of random decision trees we call Mondrian
forests. Mondrian forests can be grown in an incremental/online fashion and rema
rkably, the distribution of online Mondrian forests is the same as that of batch
 Mondrian forests. Mondrian forests achieve competitive predictive performance c
omparable with existing online random forests and periodically re-trained batch
random forests, while being more than an order of magnitude faster, thus represe
nting a better computation vs accuracy tradeoff.
****

A Probabilistic Framework for Multimodal Retrieval using Integrative Indian Buff
et Process

Bahadir Ozdemir, Larry S. Davis

We propose a multimodal retrieval procedure based on latent feature models. The
procedure consists of a nonparametric Bayesian framework for learning underlying
 semantically meaningful abstract features in a multimodal dataset, a probabilis
tic retrieval model that allows cross-modal queries and an extension model for r
elevance feedback. Experiments on two multimodal datasets, PASCAL-Sentence and S
UN-Attribute, demonstrate the effectiveness of the proposed retrieval procedure
in comparison to the state-of-the-art algorithms for learning binary codes.
****

A Multi-World Approach to Question Answering about Real-World Scenes based on Un
certain Input

Mateusz Malinowski, Mario Fritz

We propose a method for automatically answering questions about images by bringi
ng together recent advances from natural language processing and computer vision
. We combine discrete reasoning with uncertain predictions by a multi-world appr
oach that represents uncertainty about the perceived world in a bayesian framewo
rk. Our approach can handle human questions of high complexity about realistic s
cenes and replies with range of answer like counts, object classes, instances an
d lists of them. The system is directly trained from question-answer pairs. We e
stablish a first benchmark for this task that can be seen as a modern attempt at
 a visual turing test.
****

Semi-supervised Learning with Deep Generative Models

Durk P. Kingma, Shakir Mohamed, Danilo Jimenez Rezende, Max Welling

The ever-increasing size of modern data sets combined with the difficulty of obt
aining label information has made semi-supervised learning one of the problems o

f significant practical importance in modern data analysis. We revisit the appro
ach to semi-supervised learning with generative models and develop new models th
at allow for effective generalisation from small labelled data sets to large unl
abelled ones. Generative approaches have thus far been either inflexible, ineffi
cient or non-scalable. We show that deep generative models and approximate Bayes
ian inference exploiting recent advances in variational methods can be used to p
rovide significant improvements, making generative approaches highly competitive
 for semi-supervised learning.
****

Biclustering Using Message Passing
Luke O'Connor, Soheil Feizi
Biclustering is the analog of clustering on a bipartite graph. Existent methods
 infer biclusters through local search strategies that find one cluster at a time
; a common technique is to update the row memberships based on the current colum
n memberships, and vice versa. We propose a biclustering algorithm that maximize
s a global objective function using message passing. Our objective function clos
ely approximates a general likelihood function, separating a cluster size penalt
y term into row- and column-count penalties. Because we use a global optimizatio
n framework, our approach excels at resolving the overlaps between biclusters, w
hich are important features of biclusters in practice. Moreover, Expectation-Max
imization can be used to learn the model parameters if they are unknown. In simu
lations, we find that our method outperforms two of the best existing biclusteri
ng algorithms, ISA and LAS, when the planted clusters overlap. Applied to three
gene expression datasets, our method finds coregulated gene clusters that have h
igh quality in terms of cluster size and density.
****

Stochastic Proximal Gradient Descent with Acceleration Techniques
Atsushi Nitanda
Proximal gradient descent (PGD) and stochastic proximal gradient descent (SPGD)
are popular methods for solving regularized risk minimization problems in machin
e learning and statistics. In this paper, we propose and analyze an accelerated
variant of these methods in the mini-batch setting. This method incorporates two
 acceleration techniques: one is Nesterov's acceleration method, and the other i
s a variance reduction for the stochastic gradient. Accelerated proximal gradien
t descent (APG) and proximal stochastic variance reduction gradient (Prox-SVRG)
are in a trade-off relationship. We show that our method, with the appropriate m
ini-batch size, achieves lower overall complexity than both APG and Prox-SVRG.
****

DFacTo: Distributed Factorization of Tensors
Joon Hee Choi, S. Vishwanathan
We present a technique for significantly speeding up Alternating Least Squares (
ALS) and Gradient Descent (GD), two widely used algorithms for tensor factorizat
ion. By exploiting properties of the Khatri-Rao product, we show how to efficien
tly address a computationally challenging sub-step of both algorithms. Our algor
ithm, DFacTo, only requires two matrix-vector products and is easy to paralleliz
e. DFacTo is not only scalable but also on average 4 to 10 times faster than com
peting algorithms on a variety of datasets. For instance, DFacTo only takes 480
seconds on 4 machines to perform one iteration of the ALS algorithm and 1,143 se
conds to perform one iteration of the GD algorithm on a 6.5 million x 2.5 millio
n x 1.5 million dimensional tensor with 1.2 billion non-zero entries.
****

Robust Classification Under Sample Selection Bias
Anqi Liu, Brian Ziebart
In many important machine learning applications, the source distribution used to
 estimate a probabilistic classifier differs from the target distribution on whi
ch the classifier will be used to make predictions. Due to its asymptotic proper
ties, sample-reweighted loss minimization is a commonly employed technique to de
al with this difference. However, given finite amounts of labeled source data, t
his technique suffers from significant estimation errors in settings with large
sample selection bias. We develop a framework for robustly learning a probabilis

tic classifier that adapts to different sample selection biases using a minimax estimation formulation. Our approach requires only accurate estimates of statistics under the source distribution and is otherwise as robust as possible to unknown properties of the conditional label distribution, except when explicit generalization assumptions are incorporated. We demonstrate the behavior and effectiveness of our approach on synthetic and UCI binary classification tasks.
****

Deep Joint Task Learning for Generic Object Extraction
Xiaolong Wang, Liliang Zhang, Liang Lin, Zhujin Liang, Wangmeng Zuo
This paper investigates how to extract objects-of-interest without relying on hand-craft features and sliding windows approaches, that aims to jointly solve two sub-tasks: (i) rapidly localizing salient objects from images, and (ii) accurately segmenting the objects based on the localizations. We present a general joint task learning framework, in which each task (either object localization or object segmentation) is tackled via a multi-layer convolutional neural network, and the two networks work collaboratively to boost performance. In particular, we propose to incorporate latent variables bridging the two networks in a joint optimization manner. The first network directly predicts the positions and scales of salient objects from raw images, and the latent variables adjust the object localizations to feed the second network that produces pixelwise object masks. An EM-type method is then studied for the joint optimization, iterating with two steps: (i) by using the two networks, it estimates the latent variables by employing an MCMC-based sampling method; (ii) it optimizes the parameters of the two networks unitedly via back propagation, with the fixed latent variables. Extensive experiments demonstrate that our joint learning framework significantly outperforms other state-of-the-art approaches in both accuracy and efficiency (e.g., 1000 times faster than competing approaches).
****

Automatic Discovery of Cognitive Skills to Improve the Prediction of Student Learning
Robert V. Lindsey, Mohammad Khajah, Michael C. Mozer
To master a discipline such as algebra or physics, students must acquire a set of cognitive skills. Traditionally, educators and domain experts manually determine what these skills are and then select practice exercises to hone a particular skill. We propose a technique that uses student performance data to automatically discover the skills needed in a discipline. The technique assigns a latent skill to each exercise such that a student's expected accuracy on a sequence of same-skill exercises improves monotonically with practice. Rather than discarding the skills identified by experts, our technique incorporates a nonparametric prior over the exercise-skill assignments that is based on the expert-provided skills and a weighted Chinese restaurant process. We test our technique on datasets from five different intelligent tutoring systems designed for students ranging in age from middle school through college. We obtain two surprising results. First, in three of the five datasets, the skills inferred by our technique support significantly improved predictions of student performance over the expert-provided skills. Second, the expert-provided skills have little value: our technique predicts student performance nearly as well when it ignores the domain expertise as when it attempts to leverage it. We discuss explanations for these surprising results and also the relationship of our skill-discovery technique to alternative approaches.
****

General Table Completion using a Bayesian Nonparametric Model
Isabel Valera, Zoubin Ghahramani
Even though heterogeneous databases can be found in a broad variety of applications, there exists a lack of tools for estimating missing data in such databases. In this paper, we provide an efficient and robust table completion tool, based on a Bayesian nonparametric latent feature model. In particular, we propose a general observation model for the Indian buffet process (IBP) adapted to mixed continuous (real-valued and positive real-valued) and discrete (categorical, ordinal and count) observations. Then, we propose an inference algorithm that scales l

inearly with the number of observations. Finally, our experiments over five real databases show that the proposed approach provides more robust and accurate estimates than the standard IBP and the Bayesian probabilistic matrix factorization with Gaussian observations.
****

A Representation Theory for Ranking Functions
Harsh H. Pareek, Pradeep K. Ravikumar
This paper presents a representation theory for permutation-valued functions, which in their general form can also be called listwise ranking functions. Pointwise ranking functions assign a score to each object independently, without taking into account the other objects under consideration; whereas listwise loss functions evaluate the set of scores assigned to all objects as a whole. In many supervised learning to rank tasks, it might be of interest to use listwise ranking functions instead; in particular, the Bayes Optimal ranking functions might themselves be listwise, especially if the loss function is listwise. A key caveat to using listwise ranking functions has been the lack of an appropriate representation theory for such functions. We show that a natural symmetricity assumption that we call exchangeability allows us to explicitly characterize the set of such exchangeable listwise ranking functions. Our analysis draws from the theories of tensor analysis, functional analysis and De Finetti theorems. We also present experiments using a novel reranking method motivated by our representation theory.
****

Multivariate Regression with Calibration
Han Liu, Lie Wang, Tuo Zhao
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.
****

Blossom Tree Graphical Models
Zhe Liu, John Lafferty
We combine the ideas behind trees and Gaussian graphical models to form a new nonparametric family of graphical models. Our approach is to attach nonparanormal blossoms", with arbitrary graphs, to a collection of nonparametric trees. The tree edges are chosen to connect variables that most violate joint Gaussianity. The non-tree edges are partitioned into disjoint groups, and assigned to tree nodes using a nonparametric partial correlation statistic. A nonparanormal blossom is then "grown" for each group using established methods based on the graphical lasso. The result is a factorization with respect to the union of the tree branches and blossoms, defining a high-dimensional joint density that can be efficiently estimated and evaluated on test points. Theoretical properties and experiments with simulated and real data demonstrate the effectiveness of blossom trees."
****

Scalable Methods for Nonnegative Matrix Factorizations of Near-separable Tall-and-skinny Matrices
Austin R. Benson, Jason D. Lee, Bartek Rajwa, David F. Gleich
Numerous algorithms are used for nonnegative matrix factorization under the assumption that the matrix is nearly separable. In this paper, we show how to make these algorithms scalable for data matrices that have many more rows than columns, so-called tall-and-skinny matrices." One key component to these improved methods is an orthogonal matrix transformation that preserves the separability of the NMF problem. Our final methods need to read the data matrix only once and are suitable for streaming, multi-core, and MapReduce architectures. We demonstrate the efficacy of these algorithms on terabyte-sized matrices from scientific computing and bioinformatics."
****

Rates of Convergence for Nearest Neighbor Classification
Kamalika Chaudhuri, Sanjoy Dasgupta
We analyze the behavior of nearest neighbor classification in metric spaces and

provide finite-sample, distribution-dependent rates of convergence under minimal assumptions. These are more general than existing bounds, and enable us, as a by-product, to establish the universal consistency of nearest neighbor in a broader range of data spaces than was previously known. We illustrate our upper and lower bounds by introducing a new smoothness class customized for nearest neighbor classification. We find, for instance, that under the Tsybakov margin condition the convergence rate of nearest neighbor matches recently established lower bounds for nonparametric classification.

****

Inferring synaptic conductances from spike trains with a biophysically inspired point process model

Kenneth W. Latimer, E.J. Chichilnisky, Fred Rieke, Jonathan W. Pillow

A popular approach to neural characterization describes neural responses in terms of a cascade of linear and nonlinear stages: a linear filter to describe stimulus integration, followed by a nonlinear function to convert the filter output to spike rate. However, real neurons respond to stimuli in a manner that depends on the nonlinear integration of excitatory and inhibitory synaptic inputs. Here we introduce a biophysically inspired point process model that explicitly incorporates stimulus-induced changes in synaptic conductance in a dynamical model of neuronal membrane potential. Our work makes two important contributions. First, on a theoretical level, it offers a novel interpretation of the popular generalized linear model (GLM) for neural spike trains. We show that the classic GLM is a special case of our conductance-based model in which the stimulus linearly modulates excitatory and inhibitory conductances in an equal and opposite “push-pull” fashion. Our model can therefore be viewed as a direct extension of the GLM in which we relax these constraints; the resulting model can exhibit shunting as well as hyperpolarizing inhibition, and time-varying changes in both gain and membrane time constant. Second, on a practical level, we show that our model provides a tractable model of spike responses in early sensory neurons that is both more accurate and more interpretable than the GLM. Most importantly, we show that we can accurately infer intracellular synaptic conductances from extracellularly recorded spike trains. We validate these estimates using direct intracellular measurements of excitatory and inhibitory conductances in parasol retinal ganglion cells. We show that the model fit to extracellular spike trains can predict excitatory and inhibitory conductances elicited by novel stimuli with nearly the same accuracy as a model trained directly with intracellular conductances.

****

Scalable Inference for Neuronal Connectivity from Calcium Imaging

Alyson K. Fletcher, Sundeep Rangan

Fluorescent calcium imaging provides a potentially powerful tool for inferring connectivity in neural circuits with up to thousands of neurons. However, a key challenge in using calcium imaging for connectivity detection is that current systems often have a temporal response and frame rate that can be orders of magnitude slower than the underlying neural spiking process. Bayesian inference based on expectation-maximization (EM) have been proposed to overcome these limitations, but they are often computationally demanding since the E-step in the EM procedure typically involves state estimation in a high-dimensional nonlinear dynamical system. In this work, we propose a computationally fast method for the state estimation based on a hybrid of loopy belief propagation and approximate message passing (AMP). The key insight is that a neural system as viewed through calcium imaging can be factorized into simple scalar dynamical systems for each neuron with linear interconnections between the neurons. Using the structure, the updates in the proposed hybrid AMP methodology can be computed by a set of one-dimensional state estimation procedures and linear transforms with the connectivity matrix. This yields a computationally scalable method for inferring connectivity of large neural circuits. Simulations of the method on realistic neural networks demonstrate good accuracy with computation times that are potentially significantly faster than current approaches based on Markov Chain Monte Carlo methods.

****

Minimax-optimal Inference from Partial Rankings

Bruce Hajek, Sewoong Oh, Jiaming Xu
This paper studies the problem of rank aggregation under the Plackett-Luce model . The goal is to infer a global ranking and related scores of the items, based o n partial rankings provided by multiple users over multiple subsets of items. A question of particular interest is how to optimally assign items to users for ra nking and how many item assignments are needed to achieve a target estimation er ror. Without any assumptions on how the items are assigned to users, we derive a n oracle lower bound and the Cram\'er-Rao lower bound of the estimation error. W e prove an upper bound on the estimation error achieved by the maximum likelihoo d estimator, and show that both the upper bound and the Cram\'er-Rao lower bound inversely depend on the spectral gap of the Laplacian of an appropriately defin ed comparison graph. Since random comparison graphs are known to have large spec tral gaps, this suggests the use of random assignments when we have the control. Precisely, the matching oracle lower bound and the upper bound on the estimatio n error imply that the maximum likelihood estimator together with a random assig nment is minimax-optimal up to a logarithmic factor. We further analyze a popula r rank-breaking scheme that decompose partial rankings into pairwise comparisons . We show that even if one applies the mismatched maximum likelihood estimator t hat assumes independence (on pairwise comparisons that are now dependent due to rank-breaking), minimax optimal performance is still achieved up to a logarithmi c factor.
****
Neurons as Monte Carlo Samplers: Bayesian ■Inference and Learning in Spiking Net works
Yanping Huang, Rajesh PN Rao
We propose a two-layer spiking network capable of performing approximate inferen ce and learning for a hidden Markov model. The lower layer sensory neurons detec t noisy measurements of hidden world states. The higher layer neurons with recur rent connections infer a posterior distribution over world states from spike tra ins generated by sensory neurons. We show how such a neuronal network with synap tic plasticity can implement a form of Bayesian inference similar to Monte Carlo methods such as particle filtering. Each spike in the population of inference n eurons represents a sample of a particular hidden world state. The spiking activ ity across the neural population approximates the posterior distribution of hidd en state. The model provides a functional explanation for the Poisson-like noise commonly observed in cortical responses. Uncertainties in spike times provide t he necessary variability for sampling during inference. Unlike previous models, the hidden world state is not observed by the sensory neurons, and the temporal dynamics of the hidden state is unknown. We demonstrate how this network can seq uentially learn the hidden Markov model using a spike-timing dependent Hebbian l earning rule and achieve power-law convergence rates.
****
Simple MAP Inference via Low-Rank Relaxations
Roy Frostig, Sida Wang, Percy S. Liang, Christopher D. Manning
We focus on the problem of maximum a posteriori (MAP) inference in Markov random fields with binary variables and pairwise interactions. For this common subclas s of inference tasks, we consider low-rank relaxations that interpolate between the discrete problem and its full-rank semidefinite relaxation, followed by rand omized rounding. We develop new theoretical bounds studying the effect of rank, showing that as the rank grows, the relaxed objective increases but saturates, a nd that the fraction in objective value retained by the rounded discrete solutio n decreases. In practice, we show two algorithms for optimizing the low-rank obj ectives which are simple to implement, enjoy ties to the underlying theory, and outperform existing approaches on benchmark MAP inference tasks.
****
Fast Prediction for Large-Scale Kernel Machines
Cho-Jui Hsieh, Si Si, Inderjit S. Dhillon
Kernel machines such as kernel SVM and kernel ridge regression usually construct high quality models; however, their use in real-world applications remains limi ted due to the high prediction cost. In this paper, we present two novel insight

s for improving the prediction efficiency of kernel machines. First, we show tha t by adding “pseudo landmark points” to the classical Nystr¨om kernel approximat ion in an elegant way, we can significantly reduce the prediction error without much additional prediction cost. Second, we provide a new theoretical analysis o n bounding the error of the solution computed by using Nystr¨om kernel approxima tion method, and show that the error is related to the weighted kmeans objective function where the weights are given by the model computed from the original ke rnel. This theoretical insight suggests a new landmark point selection technique for the situation where we have knowledge of the original model. Based on these two insights, we provide a divide-and-conquer framework for improving the predi ction speed. First, we divide the whole problem into smaller local subproblems t o reduce the problem size. In the second phase, we develop a kernel approximatio n based fast prediction approach within each subproblem. We apply our algorithm to real world large-scale classification and regression datasets, and show that the proposed algorithm is consistently and significantly better than other compe titors. For example, on the Covertype classification problem, in terms of predic tion time, our algorithm achieves more than 10000 times speedup over the full ke rnel SVM, and a two-fold speedup over the state-of-the-art LDKL approach, while obtaining much higher prediction accuracy than LDKL (95.2% vs. 89.53%).
****

Median Selection Subset Aggregation for Parallel Inference
Xiangyu Wang, Peichao Peng, David B. Dunson
For massive data sets, efficient computation commonly relies on distributed algo rithms that store and process subsets of the data on different machines, minimiz ing communication costs. Our focus is on regression and classification problems involving many features. A variety of distributed algorithms have been proposed in this context, but challenges arise in defining an algorithm with low communic ation, theoretical guarantees and excellent practical performance in general set tings. We propose a MEdian Selection Subset AGgregation Estimator (message) algo rithm, which attempts to solve these problems. The algorithm applies feature sel ection in parallel for each subset using Lasso or another method, calculates the `median' feature inclusion index, estimates coefficients for the selected featu res in parallel for each subset, and then averages these estimates. The algorith m is simple, involves very minimal communication, scales efficiently in both sam ple and feature size, and has theoretical guarantees. In particular, we show mod el selection consistency and coefficient estimation efficiency. Extensive experi ments show excellent performance in variable selection, estimation, prediction, and computation time relative to usual competitors.
****

Design Principles of the Hippocampal Cognitive Map
Kimberly L. Stachenfeld, Matthew Botvinick, Samuel J. Gershman
Hippocampal place fields have been shown to reflect behaviorally relevant aspect s of space. For instance, place fields tend to be skewed along commonly traveled directions, they cluster around rewarded locations, and they are constrained by the geometric structure of the environment. We hypothesize a set of design prin ciples for the hippocampal cognitive map that explain how place fields represent space in a way that facilitates navigation and reinforcement learning. In parti cular, we suggest that place fields encode not just information about the curren t location, but also predictions about future locations under the current transi tion distribution. Under this model, a variety of place field phenomena arise na turally from the structure of rewards, barriers, and directional biases as refle cted in the transition policy. Furthermore, we demonstrate that this representat ion of space can support efficient reinforcement learning. We also propose that grid cells compute the eigendecomposition of place fields in part because is use ful for segmenting an enclosure along natural boundaries. When applied recursive ly, this segmentation can be used to discover a hierarchical decomposition of sp ace. Thus, grid cells might be involved in computing subgoals for hierarchical r einforcement learning.
****

Smoothed Gradients for Stochastic Variational Inference

Stephan Mandt, David Blei
Stochastic variational inference (SVI) lets us scale up Bayesian computation to massive data. It uses stochastic optimization to fit a variational distribution, following easy-to-compute noisy natural gradients. As with most traditional stochastic optimization methods, SVI takes precautions to use unbiased stochastic gradients whose expectations are equal to the true gradients. In this paper, we explore the idea of following biased stochastic gradients in SVI. Our method replaces the natural gradient with a similarly constructed vector that uses a fixed-window moving average of some of its previous terms. We will demonstrate the many advantages of this technique. First, its computational cost is the same as for SVI and storage requirements only multiply by a constant factor. Second, it enjoys significant variance reduction over the unbiased estimates, smaller bias than averaged gradients, and leads to smaller mean-squared error against the full gradient. We test our method on latent Dirichlet allocation with three large corpora.
****

A Multiplicative Model for Learning Distributed Text-Based Attribute Representations
Ryan Kiros, Richard Zemel, Russ R. Salakhutdinov
In this paper we propose a general framework for learning distributed representations of attributes: characteristics of text whose representations can be jointly learned with word embeddings. Attributes can correspond to a wide variety of concepts, such as document indicators (to learn sentence vectors), language indicators (to learn distributed language representations), meta-data and side information (such as the age, gender and industry of a blogger) or representations of authors. We describe a third-order model where word context and attribute vectors interact multiplicatively to predict the next word in a sequence. This leads to the notion of conditional word similarity: how meanings of words change when conditioned on different attributes. We perform several experimental tasks including sentiment classification, cross-lingual document classification, and blog authorship attribution. We also qualitatively evaluate conditional word neighbours and attribute-conditioned text generation.
****

Feedforward Learning of Mixture Models
Matthew Lawlor, Steven W. Zucker
We develop a biologically-plausible learning rule that provably converges to the class means of general mixture models. This rule generalizes the classical BCM neural rule within a tensor framework, substantially increasing the generality of the learning problem it solves. It achieves this by incorporating triplets of samples from the mixtures, which provides a novel information processing interpretation to spike-timing-dependent plasticity. We provide both proofs of convergence, and a close fit to experimental data on STDP.
****

Searching for Higgs Boson Decay Modes with Deep Learning
Peter J. Sadowski, Daniel Whiteson, Pierre Baldi
Particle colliders enable us to probe the fundamental nature of matter by observing exotic particles produced by high-energy collisions. Because the experimental measurements from these collisions are necessarily incomplete and imprecise, machine learning algorithms play a major role in the analysis of experimental data. The high-energy physics community typically relies on standardized machine learning software packages for this analysis, and devotes substantial effort towards improving statistical power by hand crafting high-level features derived from the raw collider measurements. In this paper, we train artificial neural networks to detect the decay of the Higgs boson to tau leptons on a dataset of 82 million simulated collision events. We demonstrate that deep neural network architectures are particularly well-suited for this task with the ability to automatically discover high-level features from the data and increase discovery significance.
****

Decoupled Variational Gaussian Inference

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

On Multiplicative Multitask Feature Learning
Xin Wang, Jinbo Bi, Shipeng Yu, Jiangwen Sun
We investigate a general framework of multiplicative multitask feature learning which decomposes each task's model parameters into a multiplication of two components. One of the components is used across all tasks and the other component is task-specific. Several previous methods have been proposed as special cases of our framework. We study the theoretical properties of this framework when different regularization conditions are applied to the two decomposed components. We prove that this framework is mathematically equivalent to the widely used multitask feature learning methods that are based on a joint regularization of all model parameters, but with a more general form of regularizers. Further, an analytical formula is derived for the across-task component as related to the task-specific component for all these regularizers, leading to a better understanding of the shrinkage effect. Study of this framework motivates new multitask learning algorithms. We propose two new learning formulations by varying the parameters in the proposed framework. Empirical studies have revealed the relative advantages of the two new formulations by comparing with the state of the art, which provides instructive insights into the feature learning problem with multiple tasks.
****

Learning Distributed Representations for Structured Output Prediction
Vivek Srikumar, Christopher D. Manning
In recent years, distributed representations of inputs have led to performance gains in many applications by allowing statistical information to be shared across inputs. However, the predicted outputs (labels, and more generally structures) are still treated as discrete objects even though outputs are often not discrete units of meaning. In this paper, we present a new formulation for structured prediction where we represent individual labels in a structure as dense vectors and allow semantically similar labels to share parameters. We extend this representation to larger structures by defining compositionality using tensor products to give a natural generalization of standard structured prediction approaches. We define a learning objective for jointly learning the model parameters and the label vectors and propose an alternating minimization algorithm for learning. We show that our formulation outperforms structural SVM baselines in two tasks: multiclass document classification and part-of-speech tagging.
****

Large-scale L-BFGS using MapReduce
Weizhu Chen, Zhenghao Wang, Jingren Zhou
L-BFGS has been applied as an effective parameter estimation method for various machine learning algorithms since 1980s. With an increasing demand to deal with massive instances and variables, it is important to scale up and parallelize L-BFGS effectively in a distributed system. In this paper, we study the problem of parallelizing the L-BFGS algorithm in large clusters of tens of thousands of shared-nothing commodity machines. First, we show that a naive implementation of L-BFGS using Map-Reduce requires either a significant amount of memory or a large number of map-reduce steps with negative performance impact. Second, we propose a new L-BFGS algorithm, called Vector-free L-BFGS, which avoids the expensive dot product operations in the two loop recursion and greatly improves computation efficiency with a great degree of parallelism. The algorithm scales very well and enables a variety of machine learning algorithms to handle a massive number of variables over large datasets. We prove the mathematical equivalence of the new Vector-free L-BFGS and demonstrate its excellent performance and scalability using real-world machine learning problems with billions of variables in production clusters.
****

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

Randomized Experimental Design for Causal Graph Discovery
Huining Hu, Zhentao Li, Adrian R. Vetta
We examine the number of controlled experiments required to discover a causal graph. Hauser and Buhlmann showed that the number of experiments required is logarithmic in the cardinality of maximum undirected clique in the essential graph. Their lower bounds, however, assume that the experiment designer cannot use randomization in selecting the experiments. We show that significant improvements are possible with the aid of randomization – in an adversarial (worst-case) setting, the designer can then recover the causal graph using at most O(log log n) experiments in expectation. This bound cannot be improved; we show it is tight for some causal graphs. We then show that in a non-adversarial (average-case) setting, even larger improvements are possible: if the causal graph is chosen uniformly at random under a Erdös-Rényi model then the expected number of experiments to discover the causal graph is constant. Finally, we present computer simulations to complement our theoretic results. Our work exploits a structural characterization of essential graphs by Andersson et al. Their characterization is based upon a set of orientation forcing operations. Our results show a distinction between which forcing operations are most important in worst-case and average-case settings.
****

Combinatorial Pure Exploration of Multi-Armed Bandits
Shouyuan Chen, Tian Lin, Irwin King, Michael R. Lyu, Wei Chen
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****

Deep Learning Face Representation by Joint Identification-Verification
Yi Sun, Yuheng Chen, Xiaogang Wang, Xiaoou Tang
The key challenge of face recognition is to develop effective feature representations for reducing intra-personal variations while enlarging inter-personal differences. In this paper, we show that it can be well solved with deep learning and using both face identification and verification signals as supervision. The Deep IDentification-verification features (DeepID2) are learned with carefully designed deep convolutional networks. The face identification task increases the inter-personal variations by drawing DeepID2 features extracted from different identities apart, while the face verification task reduces the intra-personal variations by pulling DeepID2 features extracted from the same identity together, both of which are essential to face recognition. The learned DeepID2 features can be well generalized to new identities unseen in the training data. On the challenging LFW dataset, 99.15% face verification accuracy is achieved. Compared with the best previous deep learning result on LFW, the error rate has been significantly reduced by 67%.
****

Joint Training of a Convolutional Network and a Graphical Model for Human Pose Estimation
Jonathan J. Tompson, Arjun Jain, Yann LeCun, Christoph Bregler
This paper proposes a new hybrid architecture that consists of a deep Convolutional Network and a Markov Random Field. We show how this architecture is successfully applied to the challenging problem of articulated human pose estimation in monocular images. The architecture can exploit structural domain constraints such as geometric relationships between body joint locations. We show that joint training of these two model paradigms improves performance and allows us to signif

icantly outperform existing state-of-the-art techniques.
****
Permutation Diffusion Maps (PDM) with Application to the Image Association Probl em in Computer Vision
Deepti Pachauri, Risi Kondor, Gautam Sargur, Vikas Singh
Consistently matching keypoints across images, and the related problem of findin g clusters of nearby images, are critical components of various tasks in Compute r Vision, including Structure from Motion (SfM). Unfortunately, occlusion and la rge repetitive structures tend to mislead most currently used matching algorithm s, leading to characteristic pathologies in the final output. In this paper we i ntroduce a new method, Permutations Diffusion Maps (PDM), to solve the matching problem, as well as a related new affinity measure, derived using ideas from har monic analysis on the symmetric group. We show that just by using it as a prepro cessing step to existing SfM pipelines, PDM can greatly improve reconstruction q uality on difficult datasets.
****
Structure learning of antiferromagnetic Ising models
Guy Bresler, David Gamarnik, 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-auth ors prior to requesting a name change in the electronic proceedings.
****
Clustered factor analysis of multineuronal spike data
Lars Buesing, Timothy A. Machado, John P. Cunningham, Liam Paninski
High-dimensional, simultaneous recordings of neural spiking activity are often e xplored, analyzed and visualized with the help of latent variable or factor mode ls. Such models are however ill-equipped to extract structure beyond shared, dis tributed aspects of firing activity across multiple cells. Here, we extend unstr uctured factor models by proposing a model that discovers subpopulations or grou ps of cells from the pool of recorded neurons. The model combines aspects of mix ture of factor analyzer models for capturing clustering structure, and aspects o f latent dynamical system models for capturing temporal dependencies. In the res ulting model, we infer the subpopulations and the latent factors from data using variational inference and model parameters are estimated by Expectation Maximiz ation (EM). We also address the crucial problem of initializing parameters for E M by extending a sparse subspace clustering algorithm to integer-valued spike co unt observations. We illustrate the merits of the proposed model by applying it to calcium-imaging data from spinal cord neurons, and we show that it uncovers m eaningful clustering structure in the data.
****
Mode Estimation for High Dimensional Discrete Tree Graphical Models
Chao Chen, Han Liu, Dimitris Metaxas, Tianqi Zhao
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.
****
Sampling for Inference in Probabilistic Models with Fast Bayesian Quadrature
Tom Gunter, Michael A. Osborne, Roman Garnett, Philipp Hennig, Stephen J. Robert s
We propose a novel sampling framework for inference in probabilistic models: an active learning approach that converges more quickly (in wall-clock time) than M arkov chain Monte Carlo (MCMC) benchmarks. The central challenge in probabilisti c inference is numerical integration, to average over ensembles of models or unk nown (hyper-)parameters (for example to compute marginal likelihood or a partiti on function). MCMC has provided approaches to numerical integration that deliver state-of-the-art inference, but can suffer from sample inefficiency and poor co nvergence diagnostics. Bayesian quadrature techniques offer a model-based soluti on to such problems, but their uptake has been hindered by prohibitive computati

on costs. We introduce a warped model for probabilistic integrands (likelihoods) that are known to be non-negative, permitting a cheap active learning scheme to optimally select sample locations. Our algorithm is demonstrated to offer faster convergence (in seconds) relative to simple Monte Carlo and annealed importance sampling on both synthetic and real-world examples.
****

Do Deep Nets Really Need to be Deep?
Jimmy Ba, Rich Caruana
Currently, deep neural networks are the state of the art on problems such as speech recognition and computer vision. In this paper we empirically demonstrate that shallow feed-forward nets can learn the complex functions previously learned by deep nets and achieve accuracies previously only achievable with deep models. Moreover, in some cases the shallow nets can learn these deep functions using the same number of parameters as the original deep models. On the TIMIT phoneme recognition and CIFAR-10 image recognition tasks, shallow nets can be trained that perform similarly to complex, well-engineered, deeper convolutional models.
****

A Unified Semantic Embedding: Relating Taxonomies and Attributes
Sung Ju Hwang, Leonid Sigal
We propose a method that learns a discriminative yet semantic space for object categorization, where we also embed auxiliary semantic entities such as supercategories and attributes. Contrary to prior work which only utilized them as side information, we explicitly embed the semantic entities into the same space where we embed categories, which enables us to represent a category as their linear combination. By exploiting such a unified model for semantics, we enforce each category to be generated as a sparse combination of a supercategory + attributes, with an additional exclusive regularization to learn discriminative composition. The proposed reconstructive regularization guides the discriminative learning process to learn a better generalizing model, as well as generates compact semantic description of each category, which enables humans to analyze what has been learned.
****

Parallel Double Greedy Submodular Maximization
Xinghao Pan, Stefanie Jegelka, Joseph E. Gonzalez, Joseph K. Bradley, Michael I. Jordan
Many machine learning problems can be reduced to the maximization of submodular functions. Although well understood in the serial setting, the parallel maximization of submodular functions remains an open area of research with recent results only addressing monotone functions. The optimal algorithm for maximizing the more general class of non-monotone submodular functions was introduced by Buchbinder et al. and follows a strongly serial double-greedy logic and program analysis. In this work, we propose two methods to parallelize the double-greedy algorithm. The first, coordination-free approach emphasizes speed at the cost of a weaker approximation guarantee. The second, concurrency control approach guarantees a tight 1/2-approximation, at the quantifiable cost of additional coordination and reduced parallelism. As a consequence we explore the trade off space between guaranteed performance and objective optimality. We implement and evaluate both algorithms on multi-core hardware and billion edge graphs, demonstrating both the scalability and tradeoffs of each approach.
****

Efficient Optimization for Average Precision SVM
Pritish Mohapatra, C.V. Jawahar, M. Pawan Kumar
The accuracy of information retrieval systems is often measured using average precision (AP). Given a set of positive (relevant) and negative (non-relevant) samples, the parameters of a retrieval system can be estimated using the AP-SVM framework, which minimizes a regularized convex upper bound on the empirical AP loss. However, the high computational complexity of loss-augmented inference, which is required for learning an AP-SVM, prohibits its use with large training datasets. To alleviate this deficiency, we propose three complementary approaches. The first approach guarantees an asymptotic decrease in the computational complexi

ty of loss-augmented inference by exploiting the problem structure. The second approach takes advantage of the fact that we do not require a full ranking during loss-augmented inference. This helps us to avoid the expensive step of sorting the negative samples according to their individual scores. The third approach approximates the AP loss over all samples by the AP loss over difficult samples (for example, those that are incorrectly classified by a binary SVM), while ensuring the correct classification of the remaining samples. Using the PASCAL VOC action classification and object detection datasets, we show that our approaches provide significant speed-ups during training without degrading the test accuracy of AP-SVM.

****

SAGA: A Fast Incremental Gradient Method With Support for Non-Strongly Convex Composite Objectives

Aaron Defazio, Francis Bach, Simon Lacoste-Julien

In this work we introduce a new fast incremental gradient method SAGA, in the spirit of SAG, SDCA, MISO and SVRG. SAGA improves on the theory behind SAG and SVRG, with better theoretical convergence rates, and support for composite objectives where a proximal operator is used on the regulariser. Unlike SDCA, SAGA supports non-strongly convex problems directly, and is adaptive to any inherent strong convexity of the problem. We give experimental results showing the effectiveness of our method.

****

A Differential Equation for Modeling Nesterov’s Accelerated Gradient Method: Theory and Insights

Weijie Su, Stephen Boyd, Emmanuel Candes

We derive a second-order ordinary differential equation (ODE), which is the limit of Nesterov’s accelerated gradient method. This ODE exhibits approximate equivalence to Nesterov’s scheme and thus can serve as a tool for analysis. We show that the continuous time ODE allows for a better understanding of Nesterov’s scheme. As a byproduct, we obtain a family of schemes with similar convergence rates. The ODE interpretation also suggests restarting Nesterov’s scheme leading to an algorithm, which can be rigorously proven to converge at a linear rate whenever the objective is strongly convex.

****

Sparse PCA with Oracle Property

Quanquan Gu, 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.

****

SerialRank: Spectral Ranking using Seriation

Fajwel Fogel, Alexandre d'Aspremont, Milan Vojnovic

We describe a seriation algorithm for ranking a set of n items given pairwise comparisons between these items. Intuitively, the algorithm assigns similar rankings to items that compare similarly with all others. It does so by constructing a similarity matrix from pairwise comparisons, using seriation methods to reorder this matrix and construct a ranking. We first show that this spectral seriation algorithm recovers the true ranking when all pairwise comparisons are observed and consistent with a total order. We then show that ranking reconstruction is still exact even when some pairwise comparisons are corrupted or missing, and that seriation based spectral ranking is more robust to noise than other scoring methods. An additional benefit of the seriation formulation is that it allows us to solve semi-supervised ranking problems. Experiments on both synthetic and real datasets demonstrate that seriation based spectral ranking achieves competitive and in some cases superior performance compared to classical ranking methods.

****

Pre-training of Recurrent Neural Networks via Linear Autoencoders

Luca Pasa, Alessandro Sperduti

We propose a pre-training technique for recurrent neural networks based on linea

r autoencoder networks for sequences, i.e. linear dynamical systems modelling th
e target sequences. We start by giving a closed form solution for the definition
 of the optimal weights of a linear autoencoder given a training set of sequence
s. This solution, however, is computationally very demanding, so we suggest a pr
ocedure to get an approximate solution for a given number of hidden units. The w
eights obtained for the linear autoencoder are then used as initial weights for
the input-to-hidden connections of a recurrent neural network, which is then tra
ined on the desired task. Using four well known datasets of sequences of polypho
nic music, we show that the proposed pre-training approach is highly effective,
since it allows to largely improve the state of the art results on all the consi
dered datasets.
****

Stochastic Gradient Descent, Weighted Sampling, and the Randomized Kaczmarz algo
rithm
Deanna Needell, Rachel Ward, Nati Srebro
We improve a recent gurantee of Bach and Moulines on the linear convergence of S
GD for smooth and strongly convex objectives, reducing a quadratic dependence on
 the strong convexity to a linear dependence. Furthermore, we show how reweighti
ng the sampling distribution (i.e. importance sampling) is necessary in order to
 further improve convergence, and obtain a linear dependence on average smoothne
ss, dominating previous results, and more broadly discus how importance sampling
 for SGD can improve convergence also in other scenarios. Our results are based
on a connection we make between SGD and the randomized Kaczmarz algorithm, which
 allows us to transfer ideas between the separate bodies of literature studying
each of the two methods.
****

Best-Arm Identification in Linear Bandits
Marta Soare, Alessandro Lazaric, Remi Munos
Requests for name changes in the electronic proceedings will be accepted with no
 questions asked. However name changes may cause bibliographic tracking issues.
  Authors are asked to consider this carefully and discuss it with their co-auth
ors prior to requesting a name change in the electronic proceedings.
****

Multivariate f-divergence Estimation With Confidence
Kevin Moon, Alfred Hero
The problem of f-divergence estimation is important in the fields of machine lea
rning, information theory, and statistics. While several divergence estimators e
xist, relatively few have known convergence properties. In particular, even for
those estimators whose MSE convergence rates are known, the asymptotic distribut
ions are unknown. We establish the asymptotic normality of a recently proposed e
nsemble estimator of f-divergence between two distributions from a finite number
 of samples. This estimator has MSE convergence rate of O(1/T), is simple to imp
lement, and performs well in high dimensions. This theory enables us to perform
divergence-based inference tasks such as testing equality of pairs of distributi
ons based on empirical samples. We experimentally validate our theoretical resul
ts and, as an illustration, use them to empirically bound the best achievable cl
assification error.
****

Online Decision-Making in General Combinatorial Spaces
Arun Rajkumar, Shivani Agarwal
We study online combinatorial decision problems, where one must make sequential
decisions in some combinatorial space without knowing in advance the cost of dec
isions on each trial; the goal is to minimize the total regret over some sequenc
e of trials relative to the best fixed decision in hindsight. Such problems have
 been studied mostly in settings where decisions are represented by Boolean vect
ors and costs are linear in this representation. Here we study a general setting
 where costs may be linear in any suitable low-dimensional vector representation
 of elements of the decision space. We give a general algorithm for such problem
s that we call low-dimensional online mirror descent (LDOMD); the algorithm gene
ralizes both the Component Hedge algorithm of Koolen et al. (2010), and a recent

algorithm of Suehiro et al. (2012). Our study offers a unification and generalization of previous work, and emphasizes the role of the convex polytope arising from the vector representation of the decision space; while Boolean representations lead to 0-1 polytopes, more general vector representations lead to more general polytopes. We study several examples of both types of polytopes. Finally, we demonstrate the benefit of having a general framework for such problems via an application to an online transportation problem; the associated transportation polytopes generalize the Birkhoff polytope of doubly stochastic matrices, and the resulting algorithm generalizes the PermELearn algorithm of Helmbold and Warmuth (2009).
****

Altitude Training: Strong Bounds for Single-Layer Dropout
Stefan Wager, William Fithian, Sida Wang, Percy S. Liang
Dropout training, originally designed for deep neural networks, has been successful on high-dimensional single-layer natural language tasks. This paper proposes a theoretical explanation for this phenomenon: we show that, under a generative Poisson topic model with long documents, dropout training improves the exponent in the generalization bound for empirical risk minimization. Dropout achieves this gain much like a marathon runner who practices at altitude: once a classifier learns to perform reasonably well on training examples that have been artificially corrupted by dropout, it will do very well on the uncorrupted test set. We also show that, under similar conditions, dropout preserves the Bayes decision boundary and should therefore induce minimal bias in high dimensions.
****

Probabilistic low-rank matrix completion on finite alphabets
Jean Lafond, Olga Klopp, Eric Moulines, Joseph Salmon
The task of reconstructing a matrix given a sample of observed entries is known as the \emph{matrix completion problem}. Such a consideration arises in a wide variety of problems, including recommender systems, collaborative filtering, dimensionality reduction, image processing, quantum physics or multi-class classification to name a few. Most works have focused on recovering an unknown real-valued low-rank matrix from randomly sub-sampling its entries. Here, we investigate the case where the observations take a finite numbers of values, corresponding for examples to ratings in recommender systems or labels in multi-class classification. We also consider a general sampling scheme (non-necessarily uniform) over the matrix entries. The performance of a nuclear-norm penalized estimator is analyzed theoretically. More precisely, we derive bounds for the Kullback-Leibler divergence between the true and estimated distributions. In practice, we have also proposed an efficient algorithm based on lifted coordinate gradient descent in order to tackle potentially high dimensional settings.
****

Tight Continuous Relaxation of the Balanced k-Cut Problem
Syama Sundar Rangapuram, Pramod Kaushik Mudrakarta, Matthias Hein
Spectral Clustering as a relaxation of the normalized/ratio cut has become one of the standard graph-based clustering methods. Existing methods for the computation of multiple clusters, corresponding to a balanced k-cut of the graph, are either based on greedy techniques or heuristics which have weak connection to the original motivation of minimizing the normalized cut. In this paper we propose a new tight continuous relaxation for any balanced k-cut problem and show that a related recently proposed relaxation is in most cases loose leading to poor performance in practice. For the optimization of our tight continuous relaxation we propose a new algorithm for the hard sum-of-ratios minimization problem which achieves monotonic descent. Extensive comparisons show that our method beats all existing approaches for ratio cut and other balanced k-cut criteria.
****

Discrete Graph Hashing
Wei Liu, Cun Mu, Sanjiv Kumar, Shih-Fu Chang
Hashing has emerged as a popular technique for fast nearest neighbor search in gigantic databases. In particular, learning based hashing has received considerable attention due to its appealing storage and search efficiency. However, the pe

rformance of most unsupervised learning based hashing methods deteriorates rapid ly as the hash code length increases. We argue that the degraded performance is due to inferior optimization procedures used to achieve discrete binary codes. T his paper presents a graph-based unsupervised hashing model to preserve the neig hborhood structure of massive data in a discrete code space. We cast the graph h ashing problem into a discrete optimization framework which directly learns the binary codes. A tractable alternating maximization algorithm is then proposed to explicitly deal with the discrete constraints, yielding high-quality codes to w ell capture the local neighborhoods. Extensive experiments performed on four lar ge datasets with up to one million samples show that our discrete optimization b ased graph hashing method obtains superior search accuracy over state-of-the-art unsupervised hashing methods, especially for longer codes.

****

Orbit Regularization

Renato Negrinho, Andre Martins

We propose a general framework for regularization based on group majorization. I n this framework, a group is defined to act on the parameter space and an orbit is fixed; to control complexity, the model parameters are confined to lie in the convex hull of this orbit (the orbitope). Common regularizers are recovered as particular cases, and a connection is revealed between the recent sorted 1 -norm and the hyperoctahedral group. We derive the properties a group must satisfy fo r being amenable to optimization with conditional and projected gradient algorit hms. Finally, we suggest a continuation strategy for orbit exploration, presenti ng simulation results for the symmetric and hyperoctahedral groups.

****

A Synaptical Story of Persistent Activity with Graded Lifetime in a Neural Syste m

Yuanyuan Mi, Luozheng Li, Dahui Wang, Si Wu

Persistent activity refers to the phenomenon that cortical neurons keep firing e ven after the stimulus triggering the initial neuronal responses is moved. Persi stent activity is widely believed to be the substrate for a neural system retain ing a memory trace of the stimulus information. In a conventional view, persiste nt activity is regarded as an attractor of the network dynamics, but it faces a challenge of how to be closed properly. Here, in contrast to the view of attract or, we consider that the stimulus information is encoded in a marginally unstabl e state of the network which decays very slowly and exhibits persistent firing f or a prolonged duration. We propose a simple yet effective mechanism to achieve this goal, which utilizes the property of short-term plasticity (STP) of neurona l synapses. STP has two forms, short-term depression (STD) and short-term facili tation (STF), which have opposite effects on retaining neuronal responses. We fi nd that by properly combining STF and STD, a neural system can hold persistent a ctivity of graded lifetime, and that persistent activity fades away naturally wi thout relying on an external drive. The implications of these results on neural information representation are discussed.

****

Log-Hilbert-Schmidt metric between positive definite operators on Hilbert spaces

Minh Ha Quang, Marco San Biagio, Vittorio Murino

This paper introduces a novel mathematical and computational framework, namely { \it Log-Hilbert-Schmidt metric} between positive definite operators on a Hilbert space. This is a generalization of the Log-Euclidean metric on the Riemannian m anifold of positive definite matrices to the infinite-dimensional setting. The g eneral framework is applied in particular to compute distances between covarianc e operators on a Reproducing Kernel Hilbert Space (RKHS), for which we obtain ex plicit formulas via the corresponding Gram matrices. Empirically, we apply our f ormulation to the task of multi-category image classification, where each image is represented by an infinite-dimensional RKHS covariance operator. On several c hallenging datasets, our method significantly outperforms approaches based on co variance matrices computed directly on the original input features, including th ose using the Log-Euclidean metric, Stein and Jeffreys divergences, achieving ne w state of the art results.

****
Constant Nullspace Strong Convexity and Fast Convergence of Proximal Methods under High-Dimensional Settings
Ian En-Hsu Yen, Cho-Jui Hsieh, Pradeep K. Ravikumar, Inderjit S. Dhillon
State of the art statistical estimators for high-dimensional problems take the form of regularized, and hence non-smooth, convex programs. A key facet of theses tatistical estimation problems is that these are typically not strongly convex under a high-dimensional sampling regime when the Hessian matrix becomes rank-deficient. Under vanilla convexity however, proximal optimization methods attain only a sublinear rate. In this paper, we investigate a novel variant of strong convexity, which we call Constant Nullspace Strong Convexity (CNSC), where we require that the objective function be strongly convex only over a constant subspace. As we show, the CNSC condition is naturally satisfied by high-dimensional statistical estimators. We then analyze the behavior of proximal methods under this CNSC condition: we show global linear convergence of Proximal Gradient and local quadratic convergence of Proximal Newton Method, when the regularization function comprising the statistical estimator is decomposable. We corroborate our theory via numerical experiments, and show a qualitative difference in the convergence rates of the proximal algorithms when the loss function does satisfy the CNSC condition.
****
Sparse Bayesian structure learning with “dependent relevance determination” priors
Anqi Wu, Mijung Park, Oluwasanmi O. Koyejo, Jonathan W. Pillow
In many problem settings, parameter vectors are not merely sparse, but dependent in such a way that non-zero coefficients tend to cluster together. We refer to this form of dependency as “region sparsity”. Classical sparse regression methods, such as the lasso and automatic relevance determination (ARD), model parameters as independent a priori, and therefore do not exploit such dependencies. Here we introduce a hierarchical model for smooth, region-sparse weight vectors and tensors in a linear regression setting. Our approach represents a hierarchical extension of the relevance determination framework, where we add a transformed Gaussian process to model the dependencies between the prior variances of regression weights. We combine this with a structured model of the prior variances of Fourier coefficients, which eliminates unnecessary high frequencies. The resulting prior encourages weights to be region-sparse in two different bases simultaneously. We develop efficient approximate inference methods and show substantial improvements over comparable methods (e.g., group lasso and smooth RVM) for both simulated and real datasets from brain imaging.
****
Deep Symmetry Networks
Robert Gens, Pedro M. Domingos
The chief difficulty in object recognition is that objects' classes are obscured by a large number of extraneous sources of variability, such as pose and part deformation. These sources of variation can be represented by symmetry groups, sets of composable transformations that preserve object identity. Convolutional neural networks (convnets) achieve a degree of translational invariance by computing feature maps over the translation group, but cannot handle other groups. As a result, these groups' effects have to be approximated by small translations, which often requires augmenting datasets and leads to high sample complexity. In this paper, we introduce deep symmetry networks (symnets), a generalization of convnets that forms feature maps over arbitrary symmetry groups. Symnets use kernel-based interpolation to tractably tie parameters and pool over symmetry spaces of any dimension. Like convnets, they are trained with backpropagation. The composition of feature transformations through the layers of a symnet provides a new approach to deep learning. Experiments on NORB and MNIST-rot show that symnets over the affine group greatly reduce sample complexity relative to convnets by better capturing the symmetries in the data.
****
Covariance shrinkage for autocorrelated data

Daniel Bartz, Klaus-Robert Müller
The accurate estimation of covariance matrices is essential for many signal proc
essing and machine learning algorithms. In high dimensional settings the sample
covariance is known to perform poorly, hence regularization strategies such as a
nalytic shrinkage of Ledoit/Wolf are applied. In the standard setting, i.i.d. da
ta is assumed, however, in practice, time series typically exhibit strong autoco
rrelation structure, which introduces a pronounced estimation bias. Recent work
by Sancetta has extended the shrinkage framework beyond i.i.d. data. We contribu
te in this work by showing that the Sancetta estimator, while being consistent i
n the high-dimensional limit, suffers from a high bias in finite sample sizes. W
e propose an alternative estimator, which is (1) unbiased, (2) less sensitive to
 hyperparameter choice and (3) yields superior performance in simulations on toy
 data and on a real world data set from an EEG-based Brain-Computer-Interfacing
experiment.
****
Scale Adaptive Blind Deblurring
Haichao Zhang, Jianchao Yang
The presence of noise and small scale structures usually leads to large kernel e
stimation errors in blind image deblurring empirically, if not a total failure.
We present a scale space perspective on blind deblurring algorithms, and introdu
ce a cascaded scale space formulation for blind deblurring. This new formulation
 suggests a natural approach robust to noise and small scale structures through
tying the estimation across multiple scales and balancing the contributions of d
ifferent scales automatically by learning from data. The proposed formulation al
so allows to handle non-uniform blur with a straightforward extension. Experimen
ts are conducted on both benchmark dataset and real-world images to validate the
 effectiveness of the proposed method. One surprising finding based on our appro
ach is that blur kernel estimation is not necessarily best at the finest scale.
****
Parallel Feature Selection Inspired by Group Testing
Yingbo Zhou, Utkarsh Porwal, Ce Zhang, Hung Q. Ngo, XuanLong Nguyen, Christopher
 Ré, Venu Govindaraju
This paper presents a parallel feature selection method for classification that
scales up to very high dimensions and large data sizes. Our original method is i
nspired by group testing theory, under which the feature selection procedure con
sists of a collection of randomized tests to be performed in parallel. Each test
 corresponds to a subset of features, for which a scoring function may be applie
d to measure the relevance of the features in a classification task. We develop
a general theory providing sufficient conditions under which true features are g
uaranteed to be correctly identified. Superior performance of our method is demo
nstrated on a challenging relation extraction task from a very large data set th
at have both redundant features and sample size in the order of millions. We pre
sent comprehensive comparisons with state-of-the-art feature selection methods o
n a range of data sets, for which our method exhibits competitive performance in
 terms of running time and accuracy. Moreover, it also yields substantial speedu
p when used as a pre-processing step for most other existing methods.
****
Streaming, Memory Limited Algorithms for Community Detection
Se-Young Yun, marc lelarge, Alexandre Proutiere
In this paper, we consider sparse networks consisting of a finite number of non-
overlapping communities, i.e. disjoint clusters, so that there is higher density
 within clusters than across clusters. Both the intra- and inter-cluster edge de
nsities vanish when the size of the graph grows large, making the cluster recons
truction problem nosier and hence difficult to solve. We are interested in scena
rios where the network size is very large, so that the adjacency matrix of the g
raph is hard to manipulate and store. The data stream model in which columns of
the adjacency matrix are revealed sequentially constitutes a natural framework i
n this setting. For this model, we develop two novel clustering algorithms that
extract the clusters asymptotically accurately. The first algorithm is {\it offl
ine}, as it needs to store and keep the assignments of nodes to clusters, and re

quires a memory that scales linearly with the network size. The second algorithm is {\it online}, as it may classify a node when the corresponding column is revealed and then discard this information. This algorithm requires a memory growing sub-linearly with the network size. To construct these efficient streaming memory-limited clustering algorithms, we first address the problem of clustering with partial information, where only a small proportion of the columns of the adjacency matrix is observed and develop, for this setting, a new spectral algorithm which is of independent interest.
****
Analysis of Brain States from Multi-Region LFP Time-Series
Kyle R. Ulrich, David E. Carlson, Wenzhao Lian, Jana S. Borg, Kafui Dzirasa, Lawrence Carin
The local field potential (LFP) is a source of information about the broad patterns of brain activity, and the frequencies present in these time-series measurements are often highly correlated between regions. It is believed that these regions may jointly constitute a ``brain state,'' relating to cognition and behavior. An infinite hidden Markov model (iHMM) is proposed to model the evolution of brain states, based on electrophysiological LFP data measured at multiple brain regions. A brain state influences the spectral content of each region in the measured LFP. A new state-dependent tensor factorization is employed across brain regions, and the spectral properties of the LFPs are characterized in terms of Gaussian processes (GPs). The LFPs are modeled as a mixture of GPs, with state- and region-dependent mixture weights, and with the spectral content of the data encoded in GP spectral mixture covariance kernels. The model is able to estimate the number of brain states and the number of mixture components in the mixture of GPs. A new variational Bayesian split-merge algorithm is employed for inference. The model infers state changes as a function of external covariates in two novel electrophysiological datasets, using LFP data recorded simultaneously from multiple brain regions in mice; the results are validated and interpreted by subject-matter experts.
****
Clamping Variables and Approximate Inference
Adrian Weller, Tony Jebara
It was recently proved using graph covers (Ruozzi, 2012) that the Bethe partition function is upper bounded by the true partition function for a binary pairwise model that is attractive. Here we provide a new, arguably simpler proof from first principles. We make use of the idea of clamping a variable to a particular value. For an attractive model, we show that summing over the Bethe partition functions for each sub-model obtained after clamping any variable can only raise (and hence improve) the approximation. In fact, we derive a stronger result that may have other useful implications. Repeatedly clamping until we obtain a model with no cycles, where the Bethe approximation is exact, yields the result. We also provide a related lower bound on a broad class of approximate partition functions of general pairwise multi-label models that depends only on the topology. We demonstrate that clamping a few wisely chosen variables can be of practical value by dramatically reducing approximation error.
****
Neural Word Embedding as Implicit Matrix Factorization
Omer Levy, Yoav Goldberg
We analyze skip-gram with negative-sampling (SGNS), a word embedding method introduced by Mikolov et al., and show that it is implicitly factorizing a word-context matrix, whose cells are the pointwise mutual information (PMI) of the respective word and context pairs, shifted by a global constant. We find that another embedding method, NCE, is implicitly factorizing a similar matrix, where each cell is the (shifted) log conditional probability of a word given its context. We show that using a sparse Shifted Positive PMI word-context matrix to represent words improves results on two word similarity tasks and one of two analogy tasks. When dense low-dimensional vectors are preferred, exact factorization with SVD can achieve solutions that are at least as good as SGNS's solutions for word similarity tasks. On analogy questions SGNS remains superior to SVD. We conjecture

that this stems from the weighted nature of SGNS's factorization.
****
Constrained convex minimization via model-based excessive gap
Quoc Tran-Dinh, Volkan Cevher
We introduce a model-based excessive gap technique to analyze first-order primal - dual methods for constrained convex minimization. As a result, we construct fi rst- order primal-dual methods with optimal convergence rates on the primal obje c- tive residual and the primal feasibility gap of their iterates separately. Th rough a dual smoothing and prox-center selection strategy, our framework subsume s the augmented Lagrangian, alternating direction, and dual fast-gradient method s as special cases, where our rates apply.
****
A Filtering Approach to Stochastic Variational Inference
Neil Houlsby, David Blei
Stochastic variational inference (SVI) uses stochastic optimization to scale up Bayesian computation to massive data. We present an alternative perspective on S VI as approximate parallel coordinate ascent. SVI trades-off bias and variance t o step close to the unknown true coordinate optimum given by batch variational B ayes (VB). We define a model to automate this process. The model infers the loca tion of the next VB optimum from a sequence of noisy realizations. As a conseque nce of this construction, we update the variational parameters using Bayes rule, rather than a hand-crafted optimization schedule. When our model is a Kalman fi lter this procedure can recover the original SVI algorithm and SVI with adaptive steps. We may also encode additional assumptions in the model, such as heavy-ta iled noise. By doing so, our algorithm outperforms the original SVI schedule and a state-of-the-art adaptive SVI algorithm in two diverse domains.
****