Inferring neural population dynamics from multiple partial recordings of the sam
e neural circuit
Srini Turaga, Lars Buesing, Adam M. Packer, Henry Dalgleish, Noah Pettit, Michae
l Hausser, Jakob H. Macke
Simultaneous recordings of the activity of large neural populations are extremel
y valuable as they can be used to infer the dynamics and interactions of neurons
 in a local circuit, shedding light on the computations performed. It is now pos
sible to measure the activity of hundreds of neurons using 2-photon calcium ima
ging. However, many computations are thought to involve circuits consisting of t
housands of neurons, such as cortical barrels in rodent somatosensory cortex. He
re we contribute a statistical method for stitching" together sequentially image
d sets of neurons into one model by phrasing the problem as fitting a latent dy
namical system with missing observations. This method allows us to substantially
 expand the population-sizes for which population dynamics can be characterized-
--beyond the number of simultaneously imaged neurons. In particular, we demonstr
ate using recordings in mouse somatosensory cortex that this method makes it pos
sible to predict noise correlations between non-simultaneously recorded neuron p
airs."
****

Approximate Gaussian process inference for the drift function in stochastic diff
erential equations
Andreas Ruttor, Philipp Batz, Manfred Opper
We introduce a nonparametric approach for estimating drift functions in systems
of stochastic differential equations from incomplete observations of the state v
ector. Using a Gaussian process prior over the drift as a function of the state
vector, we develop an approximate EM algorithm to deal with the unobserved, late
nt dynamics between observations. The posterior over states is approximated by a
 piecewise linearized process and the MAP estimation of the drift is facilitated
 by a sparse Gaussian process regression.
****

Third-Order Edge Statistics: Contour Continuation, Curvature, and Cortical Conne
ctions
Matthew Lawlor, Steven W. Zucker
Association field models have been used to explain human contour grouping perfor
mance and to explain the mean frequency of long-range horizontal connections acr
oss cortical columns in V1. However, association fields essentially depend on pa
irwise statistics of edges in natural scenes. We develop a spectral test of the
sufficiency of pairwise statistics and show that there is significant higher-ord
er structure. An analysis using a probabilistic spectral embedding reveals curv
ature-dependent components to the association field, and reveals a challenge for
 biological learning algorithms.
****

Transportability from Multiple Environments with Limited Experiments
Elias Bareinboim, Sanghack Lee, Vasant Honavar, Judea Pearl
This paper considers the problem of transferring experimental findings learned f
rom multiple heterogeneous domains to a target environment, in which only limite
d experiments can be performed. We reduce questions of transportability from mul
tiple domains and with limited scope to symbolic derivations in the do-calculus,
 thus extending the treatment of transportability from full experiments introduc
ed in Pearl and Bareinboim (2011). We further provide different graphical and al
gorithmic conditions for computing the transport formula for this setting, that
is, a way of fusing the observational and experimental information scattered thr
oughout different domains to synthesize a consistent estimate of the desired eff
ects.
****

On model selection consistency of penalized M-estimators: a geometric theory
Jason D. Lee, Yuekai Sun, Jonathan E. Taylor
Penalized M-estimators are used in diverse areas of science and engineering to f
it high-dimensional models with some low-dimensional structure. Often, the penal
ties are \emph{geometrically decomposable}, \ie\ can be expressed as a sum of (c

onvex) support functions. We generalize the notion of irrepresentable to geometr
ically decomposable penalties and develop a general framework for establishing c
onsistency and model selection consistency of M-estimators with such penalties.
We then use this framework to derive results for some special cases of interest
in bioinformatics and statistical learning.
****

Robust Bloom Filters for Large MultiLabel Classification Tasks
Moustapha M. Cisse, Nicolas Usunier, Thierry Artières, Patrick Gallinari
This paper presents an approach to multilabel classification (MLC) with a large
 number of labels. Our approach is a reduction to binary classification in which
 label sets are represented by low dimensional binary vectors. This representati
on follows the principle of Bloom filters, a space-efficient data structure orig
inally designed for approximate membership testing. We show that a naive applica
tion of Bloom filters in MLC is not robust to individual binary classifiers' err
ors. We then present an approach that exploits a specific feature of real-world
datasets when the number of labels is large: many labels (almost) never appear t
ogether. Our approch is provably robust, has sublinear training and inference c
omplexity with respect to the number of labels, and compares favorably to state-
of-the-art algorithms on two large scale multilabel datasets.
****

On the Relationship Between Binary Classification, Bipartite Ranking, and Binary
 Class Probability Estimation
Harikrishna Narasimhan, Shivani Agarwal
We investigate the relationship between three fundamental problems in machine le
arning: binary classification, bipartite ranking, and binary class probability e
stimation (CPE). It is known that a good binary CPE model can be used to obtain
a good binary classification model (by thresholding at 0.5), and also to obtain
a good bipartite ranking model (by using the CPE model directly as a ranking mod
el); it is also known that a binary classification model does not necessarily yi
eld a CPE model. However, not much is known about other directions. Formally, th
ese relationships involve regret transfer bounds. In this paper, we introduce th
e notion of weak regret transfer bounds, where the mapping needed to transform a
 model from one problem to another depends on the underlying probability distrib
ution (and in practice, must be estimated from data). We then show that, in this
 weaker sense, a good bipartite ranking model can be used to construct a good cl
assification model (by thresholding at a suitable point), and more surprisingly,
 also to construct a good binary CPE model (by calibrating the scores of the ran
king model).
****

Sequential Transfer in Multi-armed Bandit with Finite Set of Models
Mohammad Gheshlaghi azar, Alessandro Lazaric, Emma Brunskill
Learning from prior tasks and transferring that experience to improve future per
formance is critical for building lifelong learning agents. Although results in
supervised and reinforcement learning show that transfer may significantly impro
ve the learning performance, most of the literature on transfer is focused on ba
tch learning tasks. In this paper we study the problem of sequential transfer in
 online learning, notably in the multi-arm bandit framework, where the objective
 is to minimize the cumulative regret over a sequence of tasks by incrementally
transferring knowledge from prior tasks. We introduce a novel bandit algorithm
based on a method-of-moments approach for the estimation of the possible tasks a
nd derive regret bounds for it.
****

A Graphical Transformation for Belief Propagation: Maximum Weight Matchings and
Odd-Sized Cycles
Jinwoo Shin, Andrew E. Gelfand, Misha Chertkov
Max-product ‘belief propagation’ (BP) is a popular distributed heuristic for fin
ding the Maximum A Posteriori (MAP) assignment in a joint probability distributi
on represented by a Graphical Model (GM). It was recently shown that BP converge
s to the correct MAP assignment for a class of loopy GMs with the following comm
on feature: the Linear Programming (LP) relaxation to the MAP problem is tight (

has no integrality gap). Unfortunately, tightness of the LP relaxation does not, in general, guarantee convergence and correctness of the BP algorithm. The failure of BP in such cases motivates reverse engineering a solution – namely, given a tight LP, can we design a 'good' BP algorithm. In this paper, we design a BP algorithm for the Maximum Weight Matching (MWM) problem over general graphs. We prove that the algorithm converges to the correct optimum if the respective LP relaxation, which may include inequalities associated with non-intersecting odd-sized cycles, is tight. The most significant part of our approach is the introduction of a novel graph transformation designed to force convergence of BP. Our theoretical result suggests an efficient BP-based heuristic for the MWM problem, which consists of making sequential, "cutting plane", modifications to the underlying GM. Our experiments show that this heuristic performs as well as traditional cutting-plane algorithms using LP solvers on MWM problems.

****

A Kernel Test for Three-Variable Interactions

Dino Sejdinovic, Arthur Gretton, Wicher Bergsma

We introduce kernel nonparametric tests for Lancaster three-variable interaction and for total independence, using embeddings of signed measures into a reproducing kernel Hilbert space. The resulting test statistics are straightforward to compute, and are used in powerful three-variable interaction tests, which are consistent against all alternatives for a large family of reproducing kernels. We show the Lancaster test to be sensitive to cases where two independent causes individually have weak influence on a third dependent variable, but their combined effect has a strong influence. This makes the Lancaster test especially suited to finding structure in directed graphical models, where it outperforms competing nonparametric tests in detecting such V-structures.

****

Accelerated Mini-Batch Stochastic Dual Coordinate Ascent

Shai Shalev-Shwartz, Tong Zhang

Stochastic dual coordinate ascent (SDCA) is an effective technique for solving regularized loss minimization problems in machine learning. This paper considers an extension of SDCA under the mini-batch setting that is often used in practice. Our main contribution is to introduce an accelerated mini-batch version of SDCA and prove a fast convergence rate for this method. We discuss an implementation of our method over a parallel computing system, and compare the results to both the vanilla stochastic dual coordinate ascent and to the accelerated deterministic gradient descent method of Nesterov [2007].

****

A Scalable Approach to Probabilistic Latent Space Inference of Large-Scale Networks

Junming Yin, Qirong Ho, Eric P. Xing

We propose a scalable approach for making inference about latent spaces of large networks. With a succinct representation of networks as a bag of triangular motifs, a parsimonious statistical model, and an efficient stochastic variational inference algorithm, we are able to analyze real networks with over a million vertices and hundreds of latent roles on a single machine in a matter of hours, a setting that is out of reach for many existing methods. When compared to the state-of-the-art probabilistic approaches, our method is several orders of magnitude faster, with competitive or improved accuracy for latent space recovery and link prediction.

****

Multi-Prediction Deep Boltzmann Machines

Ian Goodfellow, Mehdi Mirza, Aaron Courville, Yoshua Bengio

We introduce the Multi-Prediction Deep Boltzmann Machine (MP-DBM). The MP-DBM can be seen as a single probabilistic model trained to maximize a variational approximation to the generalized pseudolikelihood, or as a family of recurrent nets that share parameters and approximately solve different inference problems. Prior methods of training DBMs either do not perform well on classification tasks or require an initial learning pass that trains the DBM greedily, one layer at a time. The MP-DBM does not require greedy layerwise pretraining, and outperforms t

he standard DBM at classification, classification with missing inputs, and mean field prediction tasks.
****
Learning and using language via recursive pragmatic reasoning about other agents
Nathaniel J. Smith, Noah Goodman, Michael Frank
Language users are remarkably good at making inferences about speakers' intentions in context, and children learning their native language also display substantial skill in acquiring the meanings of unknown words. These two cases are deeply related: Language users invent new terms in conversation, and language learners learn the literal meanings of words based on their pragmatic inferences about how those words are used. While pragmatic inference and word learning have both been independently characterized in probabilistic terms, no current work unifies these two. We describe a model in which language learners assume that they jointly approximate a shared, external lexicon and reason recursively about the goals of others in using this lexicon. This model captures phenomena in word learning and pragmatic inference; it additionally leads to insights about the emergence of communicative systems in conversation and the mechanisms by which pragmatic inferences become incorporated into word meanings.
****
Reinforcement Learning in Robust Markov Decision Processes
Shiau Hong Lim, Huan Xu, Shie Mannor
An important challenge in Markov decision processes is to ensure robustness with respect to unexpected or adversarial system behavior while taking advantage of well-behaving parts of the system. We consider a problem setting where some unknown parts of the state space can have arbitrary transitions while other parts are purely stochastic. We devise an algorithm that is adaptive to potentially adversarial behavior and show that it achieves similar regret bounds as the purely stochastic case.
****
Regularized Spectral Clustering under the Degree-Corrected Stochastic Blockmodel
Tai Qin, Karl Rohe
Spectral clustering is a fast and popular algorithm for finding clusters in networks. Recently, Chaudhuri et al. and Amini et al. proposed variations on the algorithm that artificially inflate the node degrees for improved statistical performance. The current paper extends the previous theoretical results to the more canonical spectral clustering algorithm in a way that removes any assumption on the minimum degree and provides guidance on the choice of tuning parameter. Moreover, our results show how the star shape" in the eigenvectors--which are consistently observed in empirical networks--can be explained by the Degree-Corrected Stochastic Blockmodel and the Extended Planted Partition model, two statistical model that allow for highly heterogeneous degrees. Throughout, the paper characterizes and justifies several of the variations of the spectral clustering algorithm in terms of these models. "
****
A Novel Two-Step Method for Cross Language Representation Learning
Min Xiao, Yuhong Guo
Cross language text classi■cation is an important learning task in natural language processing. A critical challenge of cross language learning lies in that words of different languages are in disjoint feature spaces. In this paper, we propose a two-step representation learning method to bridge the feature spaces of different languages by exploiting a set of parallel bilingual documents. Speci■cally, we ■rst formulate a matrix completion problem to produce a complete parallel document-term matrix for all documents in two languages, and then induce a cross-lingual document representation by applying latent semantic indexing on the obtained matrix. We use a projected gradient descent algorithm to solve the formulated matrix completion problem with convergence guarantees. The proposed approach is evaluated by conducting a set of experiments with cross language sentiment classi■cation tasks on Amazon product reviews. The experimental results demonstrate that the proposed learning approach outperforms a number of comparison cross language representation learning methods, especially when the number of paralle

l bilingual documents is small.
****
Graphical Models for Inference with Missing Data
Karthika Mohan, Judea Pearl, Jin Tian
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.
****
Convex Tensor Decomposition via Structured Schatten Norm Regularization
Ryota Tomioka, Taiji Suzuki
We propose a new class of structured Schatten norms for tensors that includes two recently proposed norms (overlapped'' and "latent'') for convex-optimization-based tensor decomposition. Based on the properties of the structured Schatten norms, we mathematically analyze the performance of "latent'' approach for tensor decomposition, which was empirically found to perform better than the "overlapped'' approach in some settings. We show theoretically that this is indeed the case. In particular, when the unknown true tensor is low-rank in a specific mode, this approach performs as well as knowing the mode with the smallest rank. Along the way, we show a novel duality result for structures Schatten norms, which is also interesting in the general context of structured sparsity. We confirm through numerical simulations that our theory can precisely predict the scaling behaviour of the mean squared error. "
****
Variational Inference for Mahalanobis Distance Metrics in Gaussian Process Regression
Michalis Titsias RC AUEB, Miguel Lazaro-Gredilla
We introduce a novel variational method that allows to approximately integrate out kernel hyperparameters, such as length-scales, in Gaussian process regression. This approach consists of a novel variant of the variational framework that has been recently developed for the Gaussian process latent variable model which additionally makes use of a standardised representation of the Gaussian process. We consider this technique for learning Mahalanobis distance metrics in a Gaussian process regression setting and provide experimental evaluations and comparisons with existing methods by considering datasets with high-dimensional inputs.
****
Efficient Online Inference for Bayesian Nonparametric Relational Models
Dae Il Kim, Prem K. Gopalan, David Blei, Erik Sudderth
Stochastic block models characterize observed network relationships via latent community memberships. In large social networks, we expect entities to participate in multiple communities, and the number of communities to grow with the network size. We introduce a new model for these phenomena, the hierarchical Dirichlet process relational model, which allows nodes to have mixed membership in an unbounded set of communities. To allow scalable learning, we derive an online stochastic variational inference algorithm. Focusing on assortative models of undirected networks, we also propose an efficient structured mean field variational bound, and online methods for automatically pruning unused communities. Compared to state-of-the-art online learning methods for parametric relational models, we show significantly improved perplexity and link prediction accuracy for sparse networks with tens of thousands of nodes. We also showcase an analysis of LittleSis, a large network of who-knows-who at the heights of business and government.
****
Convergence of Monte Carlo Tree Search in Simultaneous Move Games
Viliam Lisy, Vojta Kovarik, Marc Lanctot, Branislav Bosansky
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 Pass Expectation Propagation Messages

Nicolas Heess, Daniel Tarlow, John Winn

Expectation Propagation (EP) is a popular approximate posterior inference algorithm that often provides a fast and accurate alternative to sampling-based methods. However, while the EP framework in theory allows for complex non-Gaussian factors, there is still a significant practical barrier to using them within EP, because doing so requires the implementation of message update operators, which can be difficult and require hand-crafted approximations. In this work, we study the question of whether it is possible to automatically derive fast and accurate EP updates by learning a discriminative model e.g., a neural network or random forest) to map EP message inputs to EP message outputs. We address the practical concerns that arise in the process, and we provide empirical analysis on several challenging and diverse factors, indicating that there is a space of factors where this approach appears promising.

****

Bayesian Inference and Online Experimental Design for Mapping Neural Microcircuits

Ben Shababo, Brooks Paige, Ari Pakman, Liam Paninski

We develop an inference and optimal design procedure for recovering synaptic weights in neural microcircuits. We base our procedure on data from an experiment in which populations of putative presynaptic neurons can be stimulated while a subthreshold recording is made from a single postsynaptic neuron. We present a realistic statistical model which accounts for the main sources of variability in this experiment and allows for large amounts of information about the biological system to be incorporated if available. We then present a simpler model to facilitate online experimental design which entails the use of efficient Bayesian inference. The optimized approach results in equal quality posterior estimates of the synaptic weights in roughly half the number of experimental trials under experimentally realistic conditions, tested on synthetic data generated from the full model.

****

Action is in the Eye of the Beholder: Eye-gaze Driven Model for Spatio-Temporal Action Localization

Nataliya Shapovalova, Michalis Raptis, Leonid Sigal, Greg Mori

We propose a new weakly-supervised structured learning approach for recognition and spatio-temporal localization of actions in video. As part of the proposed approach we develop a generalization of the Max-Path search algorithm, which allows us to efficiently search over a structured space of multiple spatio-temporal paths, while also allowing to incorporate context information into the model. Instead of using spatial annotations, in the form of bounding boxes, to guide the latent model during training, we utilize human gaze data in the form of a weak supervisory signal. This is achieved by incorporating gaze, along with the classification, into the structured loss within the latent SVM learning framework. Experiments on a challenging benchmark dataset, UCF-Sports, show that our model is more accurate, in terms of classification, and achieves state-of-the-art results in localization. In addition, we show how our model can produce top-down saliency maps conditioned on the classification label and localized latent paths.

****

Integrated Non-Factorized Variational Inference

Shaobo Han, Xuejun Liao, Lawrence Carin

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

****

A Gang of Bandits

Nicolò Cesa-Bianchi, Claudio Gentile, Giovanni Zappella

Multi-armed bandit problems are receiving a great deal of attention because they adequately formalize the exploration-exploitation trade-offs arising in several industrially relevant applications, such as online advertisement and, more generally, recommendation systems. In many cases, however, these applications have

a strong social component, whose integration in the bandit algorithm could lead to a dramatic performance increase. For instance, we may want to serve content to a group of users by taking advantage of an underlying network of social relationships among them. In this paper, we introduce novel algorithmic approaches to the solution of such networked bandit problems. More specifically, we design and analyze a global strategy which allocates a bandit algorithm to each network node (user) and allows it to “share” signals (contexts and payoffs) with the neghboring nodes. We then derive two more scalable variants of this strategy based on different ways of clustering the graph nodes. We experimentally compare the algorithm and its variants to state-of-the-art methods for contextual bandits that do not use the relational information. Our experiments, carried out on synthetic and real-world datasets, show a marked increase in prediction performance obtained by exploiting the network structure.
****

Multiclass Total Variation Clustering
Xavier Bresson, Thomas Laurent, David Uminsky, James von Brecht
Ideas from the image processing literature have recently motivated a new set of clustering algorithms that rely on the concept of total variation. While these algorithms perform well for bi-partitioning tasks, their recursive extensions yield unimpressive results for multiclass clustering tasks. This paper presents a general framework for multiclass total variation clustering that does not rely on recursion. The results greatly outperform previous total variation algorithms and compare well with state-of-the-art NMF approaches.
****

Simultaneous Rectification and Alignment via Robust Recovery of Low-rank Tensors
Xiaoqin Zhang, Di Wang, Zhengyuan Zhou, Yi Ma
In this work, we propose a general method for recovering low-rank three-order tensors, in which the data can be deformed by some unknown transformation and corrupted by arbitrary sparse errors. Since the unfolding matrices of a tensor are interdependent, we introduce auxiliary variables and relax the hard equality constraints by the augmented Lagrange multiplier method. To improve the computational efficiency, we introduce a proximal gradient step to the alternating direction minimization method. We have provided proof for the convergence of the linearized version of the problem which is the inner loop of the overall algorithm. Both simulations and experiments show that our methods are more efficient and effective than previous work. The proposed method can be easily applied to simultaneously rectify and align multiple images or videos frames. In this context, the state-of-the-art algorithms RASL'' and "TILT'' can be viewed as two special cases of our work, and yet each only performs part of the function of our method."
****

BIG & QUIC: Sparse Inverse Covariance Estimation for a Million Variables
Cho-Jui Hsieh, Matyas A. Sustik, Inderjit S. Dhillon, Pradeep K. Ravikumar, Russell Poldrack
The l1-regularized Gaussian maximum likelihood estimator (MLE) has been shown to have strong statistical guarantees in recovering a sparse inverse covariance matrix even under high-dimensional settings. However, it requires solving a difficult non-smooth log-determinant program with number of parameters scaling quadratically with the number of Gaussian variables. State-of-the-art methods thus do not scale to problems with more than 20,000 variables. In this paper, we develop an algorithm BigQUIC, which can solve 1 million dimensional l1-regularized Gaussian MLE problems (which would thus have 1000 billion parameters) using a single machine, with bounded memory. In order to do so, we carefully exploit the underlying structure of the problem. Our innovations include a novel block-coordinate descent method with the blocks chosen via a clustering scheme to minimize repeated computations; and allowing for inexact computation of specific components. In spite of these modifications, we are able to theoretically analyze our procedure and show that BigQUIC can achieve super-linear or even quadratic convergence rates.
****

Robust Multimodal Graph Matching: Sparse Coding Meets Graph Matching

Marcelo Fiori, Pablo Sprechmann, Joshua Vogelstein, Pablo Muse, Guillermo Sapiro
Graph matching is a challenging problem with very important applications in a wide range of fields, from image and video analysis to biological and biomedical problems. We propose a robust graph matching algorithm inspired in sparsity-related techniques. We cast the problem, resembling group or collaborative sparsity formulations, as a non-smooth convex optimization problem that can be efficiently solved using augmented Lagrangian techniques. The method can deal with weighted or unweighted graphs, as well as multimodal data, where different graphs represent different types of data. The proposed approach is also naturally integrated with collaborative graph inference techniques, solving general network inference problems where the observed variables, possibly coming from different modalities, are not in correspondence. The algorithm is tested and compared with state-of-the-art graph matching techniques in both synthetic and real graphs. We also present results on multimodal graphs and applications to collaborative inference of brain connectivity from alignment-free functional magnetic resonance imaging (fMRI) data.
****

Optimal integration of visual speed across different spatiotemporal frequency channels
Matjaz Jogan, Alan A. Stocker
How does the human visual system compute the speed of a coherent motion stimulus that contains motion energy in different spatiotemporal frequency bands? Here we propose that perceived speed is the result of optimal integration of speed information from independent spatiotemporal frequency tuned channels. We formalize this hypothesis with a Bayesian observer model that treats the channel activity as independent cues, which are optimally combined with a prior expectation for slow speeds. We test the model against behavioral data from a 2AFC speed discrimination task with which we measured subjects' perceived speed of drifting sinusoidal gratings with different contrasts and spatial frequencies, and of various combinations of these single gratings. We find that perceived speed of the combined stimuli is independent of the relative phase of the underlying grating components, and that the perceptual biases and discrimination thresholds are always smaller for the combined stimuli, supporting the cue combination hypothesis. The proposed Bayesian model fits the data well, accounting for perceptual biases and thresholds of both simple and combined stimuli. Fits are improved if we assume that the channel responses are subject to divisive normalization, which is in line with physiological evidence. Our results provide an important step toward a more complete model of visual motion perception that can predict perceived speeds for stimuli of arbitrary spatial structure.
****

Translating Embeddings for Modeling Multi-relational Data
Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, Oksana Yakhnenko
We consider the problem of embedding entities and relationships of multi-relational data in low-dimensional vector spaces. Our objective is to propose a canonical model which is easy to train, contains a reduced number of parameters and can scale up to very large databases. Hence, we propose, TransE, a method which models relationships by interpreting them as translations operating on the low-dimensional embeddings of the entities. Despite its simplicity, this assumption proves to be powerful since extensive experiments show that TransE significantly outperforms state-of-the-art methods in link prediction on two knowledge bases. Besides, it can be successfully trained on a large scale data set with 1M entities, 25k relationships and more than 17M training samples.
****

Synthesizing Robust Plans under Incomplete Domain Models
Tuan A. Nguyen, Subbarao Kambhampati, Minh Do
Most current planners assume complete domain models and focus on generating correct plans. Unfortunately, domain modeling is a laborious and error-prone task, thus real world agents have to plan with incomplete domain models. While domain experts cannot guarantee completeness, often they are able to circumscribe the in

completeness of the model by providing annotations as to which parts of the doma in model may be incomplete. In such cases, the goal should be to synthesize plan s that are robust with respect to any known incompleteness of the domain. In thi s paper, we first introduce annotations expressing the knowledge of the domain i ncompleteness and formalize the notion of plan robustness with respect to an inc omplete domain model. We then show an approach to compiling the problem of findi ng robust plans to the conformant probabilistic planning problem, and present ex perimental results with Probabilistic-FF planner.
****

Learning Gaussian Graphical Models with Observed or Latent FVSs
Ying Liu, Alan Willsky
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 regions of interest from biological images with convolutional sparse block coding
Marius Pachitariu, Adam M. Packer, Noah Pettit, Henry Dalgleish, Michael Hausser , Maneesh Sahani
Biological tissue is often composed of cells with similar morphologies replicate d throughout large volumes and many biological applications rely on the accurate identification of these cells and their locations from image data. Here we deve lop a generative model that captures the regularities present in images composed of repeating elements of a few different types. Formally, the model can be desc ribed as convolutional sparse block coding. For inference we use a variant of co nvolutional matching pursuit adapted to block-based representations. We extend t he K-SVD learning algorithm to subspaces by retaining several principal vectors from the SVD decomposition instead of just one. Good models with little cross-ta lk between subspaces can be obtained by learning the blocks incrementally. We pe rform extensive experiments on simulated images and the inference algorithm cons istently recovers a large proportion of the cells with a small number of false p ositives. We fit the convolutional model to noisy GCaMP6 two-photon images of sp iking neurons and to Nissl-stained slices of cortical tissue and show that it re covers cell body locations without supervision. The flexibility of the block-bas ed representation is reflected in the variability of the recovered cell shapes.
****

Training and Analysing Deep Recurrent Neural Networks
Michiel Hermans, Benjamin Schrauwen
Time series often have a temporal hierarchy, with information that is spread out over multiple time scales. Common recurrent neural networks, however, do not ex plicitly accommodate such a hierarchy, and most research on them has been focusi ng on training algorithms rather than on their basic architecture. In this pa- p er we study the effect of a hierarchy of recurrent neural networks on processing time series. Here, each layer is a recurrent network which receives the hidden state of the previous layer as input. This architecture allows us to perform hi- erarchical processing on difficult temporal tasks, and more naturally capture t he structure of time series. We show that they reach state-of-the-art performanc e for recurrent networks in character-level language modelling when trained with sim- ple stochastic gradient descent. We also offer an analysis of the differen t emergent time scales.
****

Low-Rank Matrix and Tensor Completion via Adaptive Sampling
Akshay Krishnamurthy, Aarti Singh
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 Determinantal Point Process Sampling with Application to Clustering

Byungkon Kang
Determinantal Point Process (DPP) has gained much popularity for modeling sets o
f diverse items. The gist of DPP is that the probability of choosing a particula
r set of items is proportional to the determinant of a positive definite matrix
that defines the similarity of those items. However, computing the determinant r
equires time cubic in the number of items, and is hence impractical for large se
ts. In this paper, we address this problem by constructing a rapidly mixing Mark
ov chain, from which we can acquire a sample from the given DPP in sub-cubic tim
e. In addition, we show that this framework can be extended to sampling from ca
rdinality-constrained DPPs. As an application, we show how our sampling algorith
m can be used to provide a fast heuristic for determining the number of clusters
, resulting in better clustering.
****

Matrix factorization with binary components
Martin Slawski, Matthias Hein, Pavlo Lutsik
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.
****

Reshaping Visual Datasets for Domain Adaptation
Boqing Gong, Kristen Grauman, Fei Sha
In visual recognition problems, the common data distribution mismatches between
training and testing make domain adaptation essential. However, image data is di
fficult to manually divide into the discrete domains required by adaptation algo
rithms, and the standard practice of equating datasets with domains is a weak pr
oxy for all the real conditions that alter the statistics in complex ways (light
ing, pose, background, resolution, etc.) We propose an approach to automatically
 discover latent domains in image or video datasets. Our formulation imposes two
 key properties on domains: maximum distinctiveness and maximum learnability. By
 maximum distinctiveness, we require the underlying distributions of the identif
ied domains to be different from each other; by maximum learnability, we ensure
that a strong discriminative model can be learned from the domain. We devise a n
onparametric representation and efficient optimization procedure for distinctive
ness, which, when coupled with our learnability constraint, can successfully dis
cover domains among both training and test data. We extensively evaluate our app
roach on object recognition and human activity recognition tasks.
****

Perfect Associative Learning with Spike-Timing-Dependent Plasticity
Christian Albers, Maren Westkott, Klaus Pawelzik
Recent extensions of the Perceptron, as e.g. the Tempotron, suggest that this th
eoretical concept is highly relevant also for understanding networks of spiking
neurons in the brain. It is not known, however, how the computational power of t
he Perceptron and of its variants might be accomplished by the plasticity mechan
isms of real synapses. Here we prove that spike-timing-dependent plasticity havi
ng an anti-Hebbian form for excitatory synapses as well as a spike-timing-depend
ent plasticity of Hebbian shape for inhibitory synapses are sufficient for reali
zing the original Perceptron Learning Rule if the respective plasticity mechanis
ms act in concert with the hyperpolarisation of the post-synaptic neurons. We al
so show that with these simple yet biologically realistic dynamics Tempotrons ar
e efficiently learned. The proposed mechanism might underly the acquisition of m
appings of spatio-temporal activity patterns in one area of the brain onto other
 spatio-temporal spike patterns in another region and of long term memories in c
ortex. Our results underline that learning processes in realistic networks of sp
iking neurons depend crucially on the interactions of synaptic plasticity mechan
isms with the dynamics of participating neurons.
****

Tracking Time-varying Graphical Structure
Erich Kummerfeld, David Danks
Structure learning algorithms for graphical models have focused almost exclusive

ly on stable environments in which the underlying generative process does not ch
ange; that is, they assume that the generating model is globally stationary. In
real-world environments, however, such changes often occur without warning or si
gnal. Real-world data often come from generating models that are only locally st
ationary. In this paper, we present LoSST, a novel, heuristic structure learning
 algorithm that tracks changes in graphical model structure or parameters in a d
ynamic, real-time manner. We show by simulation that the algorithm performs comp
arably to batch-mode learning when the generating graphical structure is globall
y stationary, and significantly better when it is only locally stationary.
****

Phase Retrieval using Alternating Minimization
Praneeth Netrapalli, Prateek Jain, Sujay Sanghavi
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 Structure Learning of Stochastic And-Or Grammars
Kewei Tu, Maria Pavlovskaia, Song-Chun Zhu
Stochastic And-Or grammars compactly represent both compositionality and reconfi
gurability and have been used to model different types of data such as images an
d events. We present a unified formalization of stochastic And-Or grammars that
is agnostic to the type of the data being modeled, and propose an unsupervised a
pproach to learning the structures as well as the parameters of such grammars. S
tarting from a trivial initial grammar, our approach iteratively induces composi
tions and reconfigurations in a unified manner and optimizes the posterior proba
bility of the grammar. In our empirical evaluation, we applied our approach to l
earning event grammars and image grammars and achieved comparable or better perf
ormance than previous approaches.
****

Learning Multi-level Sparse Representations
Ferran Diego Andilla, Fred A. Hamprecht
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.
****

Estimation, Optimization, and Parallelism when Data is Sparse
John Duchi, Michael I. Jordan, Brendan McMahan
We study stochastic optimization problems when the \emph{data} is sparse, which
 is in a sense dual to the current understanding of high-dimensional statistical
 learning and optimization. We highlight both the difficulties---in terms of inc
reased sample complexity that sparse data necessitates---and the potential benef
its, in terms of allowing parallelism and asynchrony in the design of algorithms
. Concretely, we derive matching upper and lower bounds on the minimax rate for
optimization and learning with sparse data, and we exhibit algorithms achieving
these rates. Our algorithms are adaptive: they achieve the best possible rate fo
r the data observed. We also show how leveraging sparsity leads to (still minima
x optimal) parallel and asynchronous algorithms, providing experimental evidence
 complementing our theoretical results on medium to large-scale learning tasks.
****

Predictive PAC Learning and Process Decompositions
Cosma Shalizi, Aryeh Kontorovich
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 Inference for Logistic-Normal Topic Models
Jianfei Chen, Jun Zhu, Zi Wang, Xun Zheng, Bo Zhang

Logistic-normal topic models can effectively discover correlation structures amo
ng latent topics. However, their inference remains a challenge because of the no
n-conjugacy between the logistic-normal prior and multinomial topic mixing propo
rtions. Existing algorithms either make restricting mean-field assumptions or ar
e not scalable to large-scale applications. This paper presents a partially coll
apsed Gibbs sampling algorithm that approaches the provably correct distribution
 by exploring the ideas of data augmentation. To improve time efficiency, we fur
ther present a parallel implementation that can deal with large-scale applicatio
ns and learn the correlation structures of thousands of topics from millions of
documents. Extensive empirical results demonstrate the promise.
****

A multi-agent control framework for co-adaptation in brain-computer interfaces
Josh S. Merel, Roy Fox, Tony Jebara, Liam Paninski
In a closed-loop brain-computer interface (BCI), adaptive decoders are used to l
earn parameters suited to decoding the user's neural response. Feedback to the u
ser provides information which permits the neural tuning to also adapt. We prese
nt an approach to model this process of co-adaptation between the encoding model
 of the neural signal and the decoding algorithm as a multi-agent formulation of
 the linear quadratic Gaussian (LQG) control problem. In simulation we character
ize how decoding performance improves as the neural encoding and adaptive decode
r optimize, qualitatively resembling experimentally demonstrated closed-loop imp
rovement. We then propose a novel, modified decoder update rule which is aware o
f the fact that the encoder is also changing and show it can improve simulated c
o-adaptation dynamics. Our modeling approach offers promise for gaining insights
 into co-adaptation as well as improving user learning of BCI control in practic
al settings.
****

Conditional Random Fields via Univariate Exponential Families
Eunho Yang, Pradeep K. Ravikumar, Genevera I. Allen, Zhandong 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-auth
ors prior to requesting a name change in the electronic proceedings.
****

Adaptivity to Local Smoothness and Dimension in Kernel Regression
Samory Kpotufe, Vikas Garg
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 Learning with Costly Features and Labels
Navid Zolghadr, Gabor Bartok, Russell Greiner, András György, Csaba Szepesvari
This paper introduces the online probing" problem: In each round, the learner is
 able to purchase the values of a subset of feature values. After the learner us
es this information to come up with a prediction for the given round, he then ha
s the option of paying for seeing the loss that he is evaluated against. Either
way, the learner pays for the imperfections of his predictions and whatever he c
hooses to observe, including the cost of observing the loss function for the giv
en round and the cost of the observed features. We consider two variations of th
is problem, depending on whether the learner can observe the label for free or n
ot. We provide algorithms and upper and lower bounds on the regret for both vari
ants. We show that a positive cost for observing the label significantly increas
es the regret of the problem."
****

An Approximate, Efficient LP Solver for LP Rounding
Srikrishna Sridhar, Stephen Wright, Christopher Re, Ji Liu, Victor Bittorf, Ce Z
hang
Many problems in machine learning can be solved by rounding the solution of an a
ppropriate linear program. We propose a scheme that is based on a quadratic prog

ram relaxation which allows us to use parallel stochastic-coordinate-descent to approximately solve large linear programs efficiently. Our software is an order of magnitude faster than Cplex (a commercial linear programming solver) and yiel ds similar solution quality. Our results include a novel perturbation analysis o f a quadratic-penalty formulation of linear programming and a convergence result , which we use to derive running time and quality guarantees.
****

Regression-tree Tuning in a Streaming Setting
Samory Kpotufe, Francesco Orabona
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.
****

Estimating LASSO Risk and Noise Level
Mohsen Bayati, Murat A. Erdogdu, Andrea Montanari
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-auth ors prior to requesting a name change in the electronic proceedings.
****

Demixing odors - fast inference in olfaction
Agnieszka Grabska-Barwinska, Jeff Beck, Alexandre Pouget, Peter Latham
The olfactory system faces a difficult inference problem: it has to determine wh at odors are present based on the distributed activation of its receptor neurons . Here we derive neural implementations of two approximate inference algorithms that could be used by the brain. One is a variational algorithm (which builds on the work of Beck. et al., 2012), the other is based on sampling. Importantly, w e use a more realistic prior distribution over odors than has been used in the p ast: we use a spike and slab'' prior, for which most odors have zero concentrati on. After mapping the two algorithms onto neural dynamics, we find that both can infer correct odors in less than 100 ms, although it takes ~500 ms to eliminate false positives. Thus, at the behavioral level, the two algorithms make very si milar predictions. However, they make different assumptions about connectivity a nd neural computations, and make different predictions about neural activity. Th us, they should be distinguishable experimentally. If so, that would provide in sight into the mechanisms employed by the olfactory system, and, because the two algorithms use very different coding strategies, that would also provide insigh t into how networks represent probabilities."
****

Zero-Shot Learning Through Cross-Modal Transfer
Richard Socher, Milind Ganjoo, Christopher D. Manning, Andrew Ng
This work introduces a model that can recognize objects in images even if no tra ining data is available for the object class. The only necessary knowledge about unseen categories comes from unsupervised text corpora. Unlike previous zero-s hot learning models, which can only differentiate between unseen classes, our mo del can operate on a mixture of objects, simultaneously obtaining state of the a rt performance on classes with thousands of training images and reasonable perfo rmance on unseen classes. This is achieved by seeing the distributions of words in texts as a semantic space for understanding what objects look like. Our deep learning model does not require any manually defined semantic or visual feature s for either words or images. Images are mapped to be close to semantic word ve ctors corresponding to their classes, and the resulting image embeddings can be used to distinguish whether an image is of a seen or unseen class. Then, a separ ate recognition model can be employed for each type. We demonstrate two strategi es, the first gives high accuracy on unseen classes, while the second is conserv ative in its prediction of novelty and keeps the seen classes' accuracy high.
****

Optimistic policy iteration and natural actor-critic: A unifying view and a non-optimality result

Paul Wagner
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 Inference and Learning in Gaussian Process State-Space Models with Particle MCMC
Roger Frigola, Fredrik Lindsten, Thomas B. Schön, Carl Edward Rasmussen
State-space models are successfully used in many areas of science, engineering and economics to model time series and dynamical systems. We present a fully Bayesian approach to inference and learning in nonlinear nonparametric state-space models. We place a Gaussian process prior over the transition dynamics, resulting in a flexible model able to capture complex dynamical phenomena. However, to enable efficient inference, we marginalize over the dynamics of the model and instead infer directly the joint smoothing distribution through the use of specially tailored Particle Markov Chain Monte Carlo samplers. Once an approximation of the smoothing distribution is computed, the state transition predictive distribution can be formulated analytically. We make use of sparse Gaussian process models to greatly reduce the computational complexity of the approach.
****

Stochastic Gradient Riemannian Langevin Dynamics on the Probability Simplex
Sam Patterson, Yee Whye Teh
In this paper we investigate the use of Langevin Monte Carlo methods on the probability simplex and propose a new method, Stochastic gradient Riemannian Langevin dynamics, which is simple to implement and can be applied online. We apply this method to latent Dirichlet allocation in an online setting, and demonstrate that it achieves substantial performance improvements to the state of the art online variational Bayesian methods.
****

When are Overcomplete Topic Models Identifiable? Uniqueness of Tensor Tucker Decompositions with Structured Sparsity
Anima Anandkumar, Daniel J. Hsu, Majid Janzamin, Sham M. Kakade
Overcomplete latent representations have been very popular for unsupervised feature learning in recent years. In this paper, we specify which overcomplete models can be identified given observable moments of a certain order. We consider probabilistic admixture or topic models in the overcomplete regime, where the number of latent topics can greatly exceed the size of the observed word vocabulary. While general overcomplete topic models are not identifiable, we establish {\em generic} identifiability under a constraint, referred to as {\em topic persistence}. Our sufficient conditions for identifiability involve a novel set of higher order'' expansion conditions on the {\em topic-word matrix} or the {\em population structure} of the model. This set of higher-order expansion conditions allow for overcomplete models, and require the existence of a perfect matching from latent topics to higher order observed words. We establish that random structured topic models are identifiable w.h.p. in the overcomplete regime. Our identifiability results allow for general (non-degenerate) distributions for modeling the topic proportions, and thus, we can handle arbitrarily correlated topics in our framework. Our identifiability results imply uniqueness of a class of tensor decompositions with structured sparsity which is contained in the class of {\em Tucker} decompositions, but is more general than the {\em Candecomp/Parafac} (CP) decomposition."
****

Sign Cauchy Projections and Chi-Square Kernel
Ping Li, Gennady Samorodnitsk, John Hopcroft
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.
****

Transfer Learning in a Transductive Setting
Marcus Rohrbach, Sandra Ebert, Bernt Schiele
Category models for objects or activities typically rely on supervised learning requiring sufficiently large training sets. Transferring knowledge from known categories to novel classes with no or only a few labels however is far less researched even though it is a common scenario. In this work, we extend transfer learning with semi-supervised learning to exploit unlabeled instances of (novel) categories with no or only a few labeled instances. Our proposed approach Propagated Semantic Transfer combines three main ingredients. First, we transfer information from known to novel categories by incorporating external knowledge, such as linguistic or expert-specified information, e.g., by a mid-level layer of semantic attributes. Second, we exploit the manifold structure of novel classes. More specifically we adapt a graph-based learning algorithm - so far only used for semi-supervised learning - to zero-shot and few-shot learning. Third, we improve the local neighborhood in such graph structures by replacing the raw feature-based representation with a mid-level object- or attribute-based representation. We evaluate our approach on three challenging datasets in two different applications, namely on Animals with Attributes and ImageNet for image classification and on MPII Composites for activity recognition. Our approach consistently outperforms state-of-the-art transfer and semi-supervised approaches on all datasets.
****

Solving inverse problem of Markov chain with partial observations
Tetsuro Morimura, Takayuki Osogami, Tsuyoshi Ide
The Markov chain is a convenient tool to represent the dynamics of complex systems such as traffic and social systems, where probabilistic transition takes place between internal states. A Markov chain is characterized by initial-state probabilities and a state-transition probability matrix. In the traditional setting, a major goal is to figure out properties of a Markov chain when those probabilities are known. This paper tackles an inverse version of the problem: we find those probabilities from partial observations at a limited number of states. The observations include the frequency of visiting a state and the rate of reaching a state from another. Practical examples of this task include traffic monitoring systems in cities, where we need to infer the traffic volume on every single link on a road network from a very limited number of observation points. We formulate this task as a regularized optimization problem for probability functions, which is efficiently solved using the notion of natural gradient. Using synthetic and real-world data sets including city traffic monitoring data, we demonstrate the effectiveness of our method.
****

Wavelets on Graphs via Deep Learning
Raif Rustamov, Leonidas J. Guibas
An increasing number of applications require processing of signals defined on weighted graphs. While wavelets provide a flexible tool for signal processing in the classical setting of regular domains, the existing graph wavelet constructions are less flexible -- they are guided solely by the structure of the underlying graph and do not take directly into consideration the particular class of signals to be processed. This paper introduces a machine learning framework for constructing graph wavelets that can sparsely represent a given class of signals. Our construction uses the lifting scheme, and is based on the observation that the recurrent nature of the lifting scheme gives rise to a structure resembling a deep auto-encoder network. Particular properties that the resulting wavelets must satisfy determine the training objective and the structure of the involved neural networks. The training is unsupervised, and is conducted similarly to the greedy pre-training of a stack of auto-encoders. After training is completed, we obtain a linear wavelet transform that can be applied to any graph signal in time and memory linear in the size of the graph. Improved sparsity of our wavelet transform for the test signals is confirmed via experiments both on synthetic and real data.
****

Stochastic Convex Optimization with Multiple Objectives

Mehrdad Mahdavi, Tianbao Yang, Rong Jin
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 Hierarchical Community Discovery
Charles Blundell, Yee Whye Teh
We propose an efficient Bayesian nonparametric model for discovering hierarchical community structure in social networks. Our model is a tree-structured mixture of potentially exponentially many stochastic blockmodels. We describe a family of greedy agglomerative model selection algorithms whose worst case scales quadratically in the number of vertices of the network, but independent of the number of communities. Our algorithms are two orders of magnitude faster than the infinite relational model, achieving comparable or better accuracy.
****

Contrastive Learning Using Spectral Methods
James Y. Zou, Daniel J. Hsu, David C. Parkes, Ryan P. Adams
In many natural settings, the analysis goal is not to characterize a single data set in isolation, but rather to understand the difference between one set of observations and another. For example, given a background corpus of news articles together with writings of a particular author, one may want a topic model that explains word patterns and themes specific to the author. Another example comes from genomics, in which biological signals may be collected from different regions of a genome, and one wants a model that captures the differential statistics observed in these regions. This paper formalizes this notion of contrastive learning for mixture models, and develops spectral algorithms for inferring mixture components specific to a foreground data set when contrasted with a background data set. The method builds on recent moment-based estimators and tensor decompositions for latent variable models, and has the intuitive feature of using background data statistics to appropriately modify moments estimated from foreground data. A key advantage of the method is that the background data need only be coarsely modeled, which is important when the background is too complex, noisy, or not of interest. The method is demonstrated on applications in contrastive topic modeling and genomic sequence analysis.
****

Deep Fisher Networks for Large-Scale Image Classification
Karen Simonyan, Andrea Vedaldi, Andrew Zisserman
As massively parallel computations have become broadly available with modern GPUs, deep architectures trained on very large datasets have risen in popularity. Discriminatively trained convolutional neural networks, in particular, were recently shown to yield state-of-the-art performance in challenging image classification benchmarks such as ImageNet. However, elements of these architectures are similar to standard hand-crafted representations used in computer vision. In this paper, we explore the extent of this analogy, proposing a version of the state-of-the-art Fisher vector image encoding that can be stacked in multiple layers. This architecture significantly improves on standard Fisher vectors, and obtains competitive results with deep convolutional networks at a significantly smaller computational cost. Our hybrid architecture allows us to measure the performance improvement brought by a deeper image classification pipeline, while staying in the realms of conventional SIFT features and FV encodings.
****

Linear Convergence with Condition Number Independent Access of Full Gradients
Lijun Zhang, Mehrdad Mahdavi, Rong Jin
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 Noisy Labels

Nagarajan Natarajan, Inderjit S. Dhillon, Pradeep K. Ravikumar, Ambuj Tewari
In this paper, we theoretically study the problem of binary classification in th
e presence of random classification noise --- the learner, instead of seeing the
 true labels, sees labels that have independently been flipped with some small p
robability. Moreover, random label noise is \emph{class-conditional} --- the fli
p probability depends on the class. We provide two approaches to suitably modify
 any given surrogate loss function. First, we provide a simple unbiased estimato
r of any loss, and obtain performance bounds for empirical risk minimization in
the presence of iid data with noisy labels. If the loss function satisfies a sim
ple symmetry condition, we show that the method leads to an efficient algorithm
for empirical minimization. Second, by leveraging a reduction of risk minimizati
on under noisy labels to classification with weighted 0-1 loss, we suggest the u
se of a simple weighted surrogate loss, for which we are able to obtain strong e
mpirical risk bounds. This approach has a very remarkable consequence --- method
s used in practice such as biased SVM and weighted logistic regression are prova
bly noise-tolerant. On a synthetic non-separable dataset, our methods achieve ov
er 88\% accuracy even when 40\% of the labels are corrupted, and are competitive
 with respect to recently proposed methods for dealing with label noise in sever
al benchmark datasets.
****
Variational Policy Search via Trajectory Optimization
Sergey Levine, Vladlen Koltun
In order to learn effective control policies for dynamical systems, policy searc
h methods must be able to discover successful executions of the desired task. Wh
ile random exploration can work well in simple domains, complex and high-dimensi
onal tasks present a serious challenge, particularly when combined with high-dim
ensional policies that make parameter-space exploration infeasible. We present a
 method that uses trajectory optimization as a powerful exploration strategy tha
t guides the policy search. A variational decomposition of a maximum likelihood
policy objective allows us to use standard trajectory optimization algorithms su
ch as differential dynamic programming, interleaved with standard supervised lea
rning for the policy itself. We demonstrate that the resulting algorithm can out
perform prior methods on two challenging locomotion tasks.
****
Dropout Training as Adaptive Regularization
Stefan Wager, Sida Wang, Percy S. Liang
Requests for name changes in the electronic proceedings will be accepted with no
 questions asked. However name changes may cause bibliographic tracking issues.
  Authors are asked to consider this carefully and discuss it with their co-auth
ors prior to requesting a name change in the electronic proceedings.
****
Prior-free and prior-dependent regret bounds for Thompson Sampling
Sebastien Bubeck, Che-Yu 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-auth
ors prior to requesting a name change in the electronic proceedings.
****
Geometric optimisation on positive definite matrices for elliptically contoured
distributions
Suvrit Sra, Reshad Hosseini
Hermitian positive definite matrices (HPD) recur throughout statistics and machi
ne learning. In this paper we develop \emph{geometric optimisation} for globally
 optimising certain nonconvex loss functions arising in the modelling of data vi
a elliptically contoured distributions (ECDs). We exploit the remarkable structu
re of the convex cone of positive definite matrices which allows one to uncover
hidden geodesic convexity of objective functions that are nonconvex in the ordin
ary Euclidean sense. Going even beyond manifold convexity we show how further me
tric properties of HPD matrices can be exploited to globally optimise several EC
D log-likelihoods that are not even geodesic convex. We present key results that

help recognise this geometric structure, as well as obtain efficient fixed-poin t algorithms to optimise the corresponding objective functions. To our knowledge , ours are the most general results on geometric optimisation of HPD matrices kn own so far. Experiments reveal the benefits of our approach---it avoids any eige nvalue computations which makes it very competitive.
****

Capacity of strong attractor patterns to model behavioural and cognitive prototy pes
Abbas Edalat
We solve the mean field equations for a stochastic Hopfield network with tempera ture (noise) in the presence of strong, i.e., multiply stored patterns, and use this solution to obtain the storage capacity of such a network. Our result provi des for the first time a rigorous solution of the mean field equations for the s tandard Hopfield model and is in contrast to the mathematically unjustifiable re plica technique that has been hitherto used for this derivation. We show that th e critical temperature for stability of a strong pattern is equal to its degree or multiplicity, when sum of the cubes of degrees of all stored patterns is negl igible compared to the network size. In the case of a single strong pattern in t he presence of simple patterns, when the ratio of the number of all stored patte rns and the network size is a positive constant, we obtain the distribution of t he overlaps of the patterns with the mean field and deduce that the storage capa city for retrieving a strong pattern exceeds that for retrieving a simple patter n by a multiplicative factor equal to the square of the degree of the strong pat tern. This square law property provides justification for using strong patterns to model attachment types and behavioural prototypes in psychology and psychothe rapy.
****

Manifold-based Similarity Adaptation for Label Propagation
Masayuki Karasuyama, Hiroshi Mamitsuka
Label propagation is one of the state-of-the-art methods for semi-supervised lea rning, which estimates labels by propagating label information through a graph. Label propagation assumes that data points (nodes) connected in a graph should h ave similar labels. Consequently, the label estimation heavily depends on edge w eights in a graph which represent similarity of each node pair. We propose a met hod for a graph to capture the manifold structure of input features using edge w eights parameterized by a similarity function. In this approach, edge weights re present both similarity and local reconstruction weight simultaneously, both bei ng reasonable for label propagation. For further justification, we provide analy tical considerations including an interpretation as a cross-validation of a prop agation model in the feature space, and an error analysis based on a low dimensi onal manifold model. Experimental results demonstrated the effectiveness of our approach both in synthetic and real datasets.
****

New Subsampling Algorithms for Fast Least Squares Regression
Paramveer Dhillon, Yichao Lu, Dean P. Foster, Lyle Ungar
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 message-passing algorithm for multi-agent trajectory planning
José Bento, Nate Derbinsky, Javier Alonso-Mora, Jonathan S. Yedidia
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.
****

Solving the multi-way matching problem by permutation synchronization
Deepti Pachauri, Risi Kondor, Vikas Singh
The problem of matching not just two, but m different sets of objects to each ot

her arises in a variety of contexts, including finding the correspondence betwee
n feature points across multiple images in computer vision. At present it is usu
ally solved by matching the sets pairwise, in series. In contrast, we propose a
new method, permutation synchronization, which finds all the matchings jointly,
in one shot, via a relaxation to eigenvector decomposition. The resulting algori
thm is both computationally efficient, and, as we demonstrate with theoretical a
rguments as well as experimental results, much more stable to noise than previou
s methods.
****

Auditing: Active Learning with Outcome-Dependent Query Costs
Sivan Sabato, Anand D. Sarwate, Nati Srebro
We propose a learning setting in which unlabeled data is free, and the cost of a
 label depends on its value, which is not known in advance. We study binary clas
sification in an extreme case, where the algorithm only pays for negative labels
. Our motivation are applications such as fraud detection, in which investigatin
g an honest transaction should be avoided if possible. We term the setting audit
ing, and consider the auditing complexity of an algorithm: The number of negativ
e points it labels to learn a hypothesis with low relative error. We design audi
ting algorithms for thresholds on the line and axis-aligned rectangles, and show
 that with these algorithms, the auditing complexity can be significantly lower
than the active label complexity. We discuss a general approach for auditing for
 a general hypothesis class, and describe several interesting directions for fut
ure work.
****

Restricting exchangeable nonparametric distributions
Sinead A. Williamson, Steve N. MacEachern, Eric P. Xing
Distributions over exchangeable matrices with infinitely many columns are useful
 in constructing nonparametric latent variable models. However, the distribution
 implied by such models over the number of features exhibited by each data point
 may be poorly-suited for many modeling tasks. In this paper, we propose a class
 of exchangeable nonparametric priors obtained by restricting the domain of exis
ting models. Such models allow us to specify the distribution over the number of
 features per data point, and can achieve better performance on data sets where
the number of features is not well-modeled by the original distribution.
****

On the Linear Convergence of the Proximal Gradient Method for Trace Norm Regular
ization
Ke Hou, Zirui Zhou, Anthony Man-Cho So, Zhi-Quan Luo
Motivated by various applications in machine learning, the problem of minimizing
 a convex smooth loss function with trace norm regularization has received much
attention lately. Currently, a popular method for solving such problem is the p
roximal gradient method (PGM), which is known to have a sublinear rate of conver
gence. In this paper, we show that for a large class of loss functions, the con
vergence rate of the PGM is in fact linear. Our result is established without a
ny strong convexity assumption on the loss function. A key ingredient in our pr
oof is a new Lipschitzian error bound for the aforementioned trace norm-regulari
zed problem, which may be of independent interest.
****

Eluder Dimension and the Sample Complexity of Optimistic Exploration
Daniel Russo, Benjamin Van Roy
This paper considers the sample complexity of the multi-armed bandit with depend
encies among the arms. Some of the most successful algorithms for this problem u
se the principle of optimism in the face of uncertainty to guide exploration. Th
e clearest example of this is the class of upper confidence bound (UCB) algorith
ms, but recent work has shown that a simple posterior sampling algorithm, someti
mes called Thompson sampling, also shares a close theoretical connection with op
timistic approaches. In this paper, we develop a regret bound that holds for bot
h classes of algorithms. This bound applies broadly and can be specialized to ma
ny model classes. It depends on a new notion we refer to as the eluder dimension
, which measures the degree of dependence among action rewards. Compared to UCB

algorithm regret bounds for specific model classes, our general bound matches th
e best available for linear models and is stronger than the best available for g
eneralized linear models.
****

Efficient Algorithm for Privately Releasing Smooth Queries
Ziteng Wang, Kai Fan, Jiaqi Zhang, Liwei Wang
Requests for name changes in the electronic proceedings will be accepted with no
 questions asked. However name changes may cause bibliographic tracking issues.
  Authors are asked to consider this carefully and discuss it with their co-auth
ors prior to requesting a name change in the electronic proceedings.
****

Buy-in-Bulk Active Learning
Liu Yang, Jaime Carbonell
In many practical applications of active learning, it is more cost-effective to
request labels in large batches, rather than one-at-a-time. This is because the
cost of labeling a large batch of examples at once is often sublinear in the num
ber of examples in the batch. In this work, we study the label complexity of act
ive learning algorithms that request labels in a given number of batches, as wel
l as the tradeoff between the total number of queries and the number of rounds a
llowed. We additionally study the total cost sufficient for learning, for an abs
tract notion of the cost of requesting the labels of a given number of examples
at once. In particular, we find that for sublinear cost functions, it is often d
esirable to request labels in large batches (i.e., buying in bulk); although thi
s may increase the total number of labels requested, it reduces the total cost r
equired for learning.
****

On Poisson Graphical Models
Eunho Yang, Pradeep K. Ravikumar, Genevera I. Allen, Zhandong Liu
Undirected graphical models, such as Gaussian graphical models, Ising, and multi
nomial/categorical graphical models, are widely used in a variety of application
s for modeling distributions over a large number of variables. These standard in
stances, however, are ill-suited to modeling count data, which are increasingly
ubiquitous in big-data settings such as genomic sequencing data, user-ratings da
ta, spatial incidence data, climate studies, and site visits. Existing classes o
f Poisson graphical models, which arise as the joint distributions that correspo
nd to Poisson distributed node-conditional distributions, have a major drawback:
 they can only model negative conditional dependencies for reasons of normalizab
ility given its infinite domain. In this paper, our objective is to modify the P
oisson graphical model distribution so that it can capture a rich dependence str
ucture between count-valued variables. We begin by discussing two strategies for
 truncating the Poisson distribution and show that only one of these leads to a
valid joint distribution; even this model, however, has limitations on the types
 of variables and dependencies that may be modeled. To address this, we propose
two novel variants of the Poisson distribution and their corresponding joint gra
phical model distributions. These models provide a class of Poisson graphical m
odels that can capture both positive and negative conditional dependencies betwe
en count-valued variables. One can learn the graph structure of our model via pe
nalized neighborhood selection, and we demonstrate the performance of our method
s by learning simulated networks as well as a network from microRNA-Sequencing d
ata.
****

On Sampling from the Gibbs Distribution with Random Maximum A-Posteriori Perturb
ations
Tamir Hazan, Subhransu Maji, Tommi Jaakkola
In this paper we describe how MAP inference can be used to sample efficiently fr
om Gibbs distributions. Specifically, we provide means for drawing either appro
ximate or unbiased samples from Gibbs' distributions by introducing low dimensio
nal perturbations and solving the corresponding MAP assignments. Our approach al
so leads to new ways to derive lower bounds on partition functions. We demonstra
te empirically that our method excels in the typical high signal - high coupling

'' regime. The setting results in ragged energy landscapes that are challenging for alternative approaches to sampling and/or lower bounds. "
****
Factorized Asymptotic Bayesian Inference for Latent Feature Models
Kohei Hayashi, Ryohei Fujimaki
This paper extends factorized asymptotic Bayesian (FAB) inference for latent feature models~(LFMs). FAB inference has not been applicable to models, including LFMs, without a specific condition on the Hesqsian matrix of a complete log-likelihood, which is required to derive a factorized information criterion''~(FIC). Our asymptotic analysis of the Hessian matrix of LFMs shows that FIC of LFMs has the same form as those of mixture models. FAB/LFMs have several desirable properties (e.g., automatic hidden states selection and parameter identifiability) and empirically perform better than state-of-the-art Indian Buffet processes in terms of model selection, prediction, and computational efficiency."
****
Minimax Theory for High-dimensional Gaussian Mixtures with Sparse Mean Separation
Martin Azizyan, Aarti Singh, Larry Wasserman
While several papers have investigated computationally and statistically efficient methods for learning Gaussian mixtures, precise minimax bounds for their statistical performance as well as fundamental limits in high-dimensional settings are not well-understood. In this paper, we provide precise information theoretic bounds on the clustering accuracy and sample complexity of learning a mixture of two isotropic Gaussians in high dimensions under small mean separation. If there is a sparse subset of relevant dimensions that determine the mean separation, then the sample complexity only depends on the number of relevant dimensions and mean separation, and can be achieved by a simple computationally efficient procedure. Our results provide the first step of a theoretical basis for recent methods that combine feature selection and clustering.
****
Efficient Optimization for Sparse Gaussian Process Regression
Yanshuai Cao, Marcus A. Brubaker, David J. Fleet, Aaron Hertzmann
We propose an efficient discrete optimization algorithm for selecting a subset of training data to induce sparsity for Gaussian process regression. The algorithm estimates this inducing set and the hyperparameters using a single objective, either the marginal likelihood or a variational free energy. The space and time complexity are linear in the training set size, and the algorithm can be applied to large regression problems on discrete or continuous domains. Empirical evaluation shows state-of-art performance in the discrete case and competitive results in the continuous case.
****
Robust learning of low-dimensional dynamics from large neural ensembles
David Pfau, Eftychios A. Pnevmatikakis, Liam Paninski
Recordings from large populations of neurons make it possible to search for hypothesized low-dimensional dynamics. Finding these dynamics requires models that take into account biophysical constraints and can be fit efficiently and robustly. Here, we present an approach to dimensionality reduction for neural data that is convex, does not make strong assumptions about dynamics, does not require averaging over many trials and is extensible to more complex statistical models that combine local and global influences. The results can be combined with spectral methods to learn dynamical systems models. The basic method can be seen as an extension of PCA to the exponential family using nuclear norm minimization. We evaluate the effectiveness of this method using an exact decomposition of the Bregman divergence that is analogous to variance explained for PCA. We show on model data that the parameters of latent linear dynamical systems can be recovered, and that even if the dynamics are not stationary we can still recover the true latent subspace. We also demonstrate an extension of nuclear norm minimization that can separate sparse local connections from global latent dynamics. Finally, we demonstrate improved prediction on real neural data from monkey motor cortex compared to fitting linear dynamical models without nuclear norm smoothing.

****
Causal Inference on Time Series using Restricted Structural Equation Models
Jonas Peters, Dominik Janzing, Bernhard Schölkopf
Causal inference uses observational data to infer the causal structure of the da ta generating system. We study a class of restricted Structural Equation Models for time series that we call Time Series Models with Independent Noise (TiMINo). These models require independent residual time series, whereas traditional meth ods like Granger causality exploit the variance of residuals. This work contains two main contributions: (1) Theoretical: By restricting the model class (e.g. t o additive noise) we provide more general identifiability results than existing ones. The results cover lagged and instantaneous effects that can be nonlinear a nd unfaithful, and non-instantaneous feedbacks between the time series. (2) Prac tical: If there are no feedback loops between time series, we propose an algorit hm based on non-linear independence tests of time series. When the data are caus ally insufficient, or the data generating process does not satisfy the model ass umptions, this algorithm may still give partial results, but mostly avoids incor rect answers. The Structural Equation Model point of view allows us to extend bo th the theoretical and the algorithmic part to situations in which the time seri es have been measured with different time delays (as may happen for fMRI data, f or example). TiMINo outperforms existing methods on artificial and real data. Co de is provided.
****
Better Approximation and Faster Algorithm Using the Proximal Average
Yao-Liang Yu
It is a common practice to approximate complicated'' functions with more friendl y ones. In large-scale machine learning applications, nonsmooth losses/regulariz ers that entail great computational challenges are usually approximated by smoot h functions. We re-examine this powerful methodology and point out a nonsmooth a pproximation which simply pretends the linearity of the proximal map. The new a pproximation is justified using a recent convex analysis tool---proximal average , and yields a novel proximal gradient algorithm that is strictly better than th e one based on smoothing, without incurring any extra overhead. Numerical experi ments conducted on two important applications, overlapping group lasso and graph -guided fused lasso, corroborate the theoretical claims."
****
Robust Low Rank Kernel Embeddings of Multivariate Distributions
Le Song, Bo Dai
Kernel embedding of distributions has led to many recent advances in machine lea rning. However, latent and low rank structures prevalent in real world distribut ions have rarely been taken into account in this setting. Furthermore, no prior work in kernel embedding literature has addressed the issue of robust embedding when the latent and low rank information are misspecified. In this paper, we pro pose a hierarchical low rank decomposition of kernels embeddings which can explo it such low rank structures in data while being robust to model misspecification . We also illustrate with empirical evidence that the estimated low rank embeddi ngs lead to improved performance in density estimation.
****
Learning the Local Statistics of Optical Flow
Dan Rosenbaum, Daniel Zoran, Yair Weiss
Motivated by recent progress in natural image statistics, we use newly available datasets with ground truth optical flow to learn the local statistics of optic al flow and rigorously compare the learned model to prior models assumed by comp uter vision optical flow algorithms. We find that a Gaussian mixture model with 64 components provides a significantly better model for local flow statistics w hen compared to commonly used models. We investigate the source of the GMMs succ ess and show it is related to an explicit representation of flow boundaries. We also learn a model that jointly models the local intensity pattern and the local optical flow. In accordance with the assumptions often made in computer vision, the model learns that flow boundaries are more likely at intensity boundaries. However, when evaluated on a large dataset, this dependency is very weak and the

benefit of conditioning flow estimation on the local intensity pattern is margi
nal.
****

Fast Algorithms for Gaussian Noise Invariant Independent Component Analysis
James R. Voss, Luis Rademacher, Mikhail Belkin

The performance of standard algorithms for Independent Component Analysis quickl
y deteriorates under the addition of Gaussian noise. This is partially due to a
common first step that typically consists of whitening, i.e., applying Principal
Component Analysis (PCA) and rescaling the components to have identity covarian
ce, which is not invariant under Gaussian noise. In our paper we develop the f
irst practical algorithm for Independent Component Analysis that is provably inv
ariant under Gaussian noise. The two main contributions of this work are as foll
ows: 1. We develop and implement a more efficient version of a Gaussian noise in
variant decorrelation (quasi-orthogonalization) algorithm using Hessians of the
cumulant functions. 2. We propose a very simple and efficient fixed-point GI-ICA
(Gradient Iteration ICA) algorithm, which is compatible with quasi-orthogonaliz
ation, as well as with the usual PCA-based whitening in the noiseless case. The
algorithm is based on a special form of gradient iteration (different from grad
ient descent). We provide an analysis of our algorithm demonstrating fast conv
ergence following from the basic properties of cumulants. We also present a numb
er of experimental comparisons with the existing methods, showing superior resul
ts on noisy data and very competitive performance in the noiseless case.
****

Stochastic Majorization-Minimization Algorithms for Large-Scale Optimization
Julien Mairal

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 Learning in Markov Decision Processes with Adversarially Chosen Transitio
n Probability Distributions
Yasin Abbasi Yadkori, Peter L. Bartlett, Varun Kanade, Yevgeny Seldin, Csaba Sze
pesvari

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 and Generalized Upper Bounds on the Complexity of Policy Iteration
Bruno Scherrer

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.
****

Approximate Inference in Continuous Determinantal Processes
Raja Hafiz Affandi, Emily Fox, Ben Taskar

Determinantal point processes (DPPs) are random point processes well-suited for
modeling repulsion. In machine learning, the focus of DPP-based models has been
on diverse subset selection from a discrete and finite base set. This discrete s
etting admits an efficient algorithm for sampling based on the eigendecompositio
n of the defining kernel matrix. Recently, there has been growing interest in us
ing DPPs defined on continuous spaces. While the discrete-DPP sampler extends fo
rmally to the continuous case, computationally, the steps required cannot be dir
ectly extended except in a few restricted cases. In this paper, we present effic
ient approximate DPP sampling schemes based on Nystrom and random Fourier featur
e approximations that apply to a wide range of kernel functions. We demonstrate
the utility of continuous DPPs in repulsive mixture modeling applications and sy
nthesizing human poses spanning activity spaces.

****
Streaming Variational Bayes
Tamara Broderick, Nicholas Boyd, Andre Wibisono, Ashia C. Wilson, Michael I. Jordan
We present SDA-Bayes, a framework for (S)treaming, (D)istributed, (A)synchronous computation of a Bayesian posterior. The framework makes streaming updates to the estimated posterior according to a user-specified approximation primitive function. We demonstrate the usefulness of our framework, with variational Bayes (VB) as the primitive, by fitting the latent Dirichlet allocation model to two large-scale document collections. We demonstrate the advantages of our algorithm over stochastic variational inference (SVI), both in the single-pass setting SVI was designed for and in the streaming setting, to which SVI does not apply.
****
One-shot learning by inverting a compositional causal process
Brenden M. Lake, Russ R. Salakhutdinov, Josh Tenenbaum
People can learn a new visual class from just one example, yet machine learning algorithms typically require hundreds or thousands of examples to tackle the same problems. Here we present a Hierarchical Bayesian model based on compositionality and causality that can learn a wide range of natural (although simple) visual concepts, generalizing in human-like ways from just one image. We evaluated performance on a challenging one-shot classification task, where our model achieved a human-level error rate while substantially outperforming two deep learning models. We also used a visual Turing test" to show that our model produces human-like performance on other conceptual tasks, including generating new examples and parsing."
****
Large Scale Distributed Sparse Precision Estimation
Huahua Wang, Arindam Banerjee, Cho-Jui Hsieh, Pradeep K. Ravikumar, Inderjit S. Dhillon
We consider the problem of sparse precision matrix estimation in high dimensions using the CLIME estimator, which has several desirable theoretical properties. We present an inexact alternating direction method of multiplier (ADMM) algorithm for CLIME, and establish rates of convergence for both the objective and optimality conditions. Further, we develop a large scale distributed framework for the computations, which scales to millions of dimensions and trillions of parameters, using hundreds of cores. The proposed framework solves CLIME in column-blocks and only involves elementwise operations and parallel matrix multiplications. We evaluate our algorithm on both shared-memory and distributed-memory architectures, which can use block cyclic distribution of data and parameters to achieve load balance and improve the efficiency in the use of memory hierarchies. Experimental results show that our algorithm is substantially more scalable than state-of-the-art methods and scales almost linearly with the number of cores.
****
Online Variational Approximations to non-Exponential Family Change Point Models: With Application to Radar Tracking
Ryan D. Turner, Steven Bottone, Clay J. Stanek
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.
****
RNADE: The real-valued neural autoregressive density-estimator
Benigno Uria, Iain Murray, Hugo Larochelle
We introduce RNADE, a new model for joint density estimation of real-valued vectors. Our model calculates the density of a datapoint as the product of one-dimensional conditionals modeled using mixture density networks with shared parameters. RNADE learns a distributed representation of the data, while having a tractable expression for the calculation of densities. A tractable likelihood allows direct comparison with other methods and training by standard gradient-based optimizers. We compare the performance of RNADE on several datasets of heterogeneous

and perceptual data, finding it outperforms mixture models in all but one case.
****
Estimating the Unseen: Improved Estimators for Entropy and other Properties
Paul Valiant, Gregory Valiant
Recently, [Valiant and Valiant] showed that a class of distributional properties, which includes such practically relevant properties as entropy, the number of distinct elements, and distance metrics between pairs of distributions, can be estimated given a SUBLINEAR sized sample. Specifically, given a sample consisting of independent draws from any distribution over at most n distinct elements, these properties can be estimated accurately using a sample of size O(n / log n).
We propose a novel modification of this approach and show: 1) theoretically, our estimator is optimal (to constant factors, over worst-case instances), and 2) in practice, it performs exceptionally well for a variety of estimation tasks, on a variety of natural distributions, for a wide range of parameters. Perhaps unsurprisingly, the key step in this approach is to first use the sample to characterize the unseen" portion of the distribution. This goes beyond such tools as the Good-Turing frequency estimation scheme, which estimates the total probability mass of the unobserved portion of the distribution: we seek to estimate the "shape"of the unobserved portion of the distribution. This approach is robust, general, and theoretically principled; we expect that it may be fruitfully used as a component within larger machine learning and data analysis systems. "
****
Dynamic Clustering via Asymptotics of the Dependent Dirichlet Process Mixture
Trevor Campbell, Miao Liu, Brian Kulis, Jonathan P. How, Lawrence Carin
This paper presents a novel algorithm, based upon the dependent Dirichlet process mixture model (DDPMM), for clustering batch-sequential data containing an unknown number of evolving clusters. The algorithm is derived via a low-variance asymptotic analysis of the Gibbs sampling algorithm for the DDPMM, and provides a hard clustering with convergence guarantees similar to those of the k-means algorithm. Empirical results from a synthetic test with moving Gaussian clusters and a test with real ADS-B aircraft trajectory data demonstrate that the algorithm requires orders of magnitude less computational time than contemporary probabilistic and hard clustering algorithms, while providing higher accuracy on the examined datasets.
****
Parametric Task Learning
Ichiro Takeuchi, Tatsuya Hongo, Masashi Sugiyama, Shinichi Nakajima
We introduce a novel formulation of multi-task learning (MTL) called parametric task learning (PTL) that can systematically handle infinitely many tasks parameterized by a continuous parameter. Our key finding is that, for a certain class of PTL problems, the path of optimal task-wise solutions can be represented as piecewise-linear functions of the continuous task parameter. Based on this fact, we employ a parametric programming technique to obtain the common shared representation across all the continuously parameterized tasks efficiently. We show that our PTL formulation is useful in various scenarios such as learning under non-stationarity, cost-sensitive learning, and quantile regression, and demonstrate the usefulness of the proposed method experimentally in these scenarios.
****
Generalized Denoising Auto-Encoders as Generative Models
Yoshua Bengio, Li Yao, Guillaume Alain, Pascal Vincent
Recent work has shown how denoising and contractive autoencoders implicitly capture the structure of the data generating density, in the case where the corruption noise is Gaussian, the reconstruction error is the squared error, and the data is continuous-valued. This has led to various proposals for sampling from this implicitly learned density function, using Langevin and Metropolis-Hastings MCMC. However, it remained unclear how to connect the training procedure of regularized auto-encoders to the implicit estimation of the underlying data generating distribution when the data are discrete, or using other forms of corruption process and reconstruction errors. Another issue is the mathematical justification which is only valid in the limit of small corruption noise. We propose here a di

fferent attack on the problem, which deals with all these issues: arbitrary (but noisy enough) corruption, arbitrary reconstruction loss (seen as a log-likelihood), handling both discrete and continuous-valued variables, and removing the bias due to non-infinitesimal corruption noise (or non-infinitesimal contractive penalty).
****

Local Privacy and Minimax Bounds: Sharp Rates for Probability Estimation
John Duchi, Martin J. Wainwright, Michael I. Jordan
We provide a detailed study of the estimation of probability distributions---discrete and continuous---in a stringent setting in which data is kept private even from the statistician. We give sharp minimax rates of convergence for estimation in these locally private settings, exhibiting fundamental tradeoffs between privacy and convergence rate, as well as providing tools to allow movement along the privacy-statistical efficiency continuum. One of the consequences of our results is that Warner's classical work on randomized response is an optimal way to perform survey sampling while maintaining privacy of the respondents.
****

Reward Mapping for Transfer in Long-Lived Agents
Xiaoxiao Guo, Satinder Singh, Richard L. Lewis
We consider how to transfer knowledge from previous tasks to a current task in long-lived and bounded agents that must solve a sequence of MDPs over a finite lifetime. A novel aspect of our transfer approach is that we reuse reward functions. While this may seem counterintuitive, we build on the insight of recent work on the optimal rewards problem that guiding an agent's behavior with reward functions other than the task-specifying reward function can help overcome computational bounds of the agent. Specifically, we use good guidance reward functions learned on previous tasks in the sequence to incrementally train a reward mapping function that maps task-specifying reward functions into good initial guidance reward functions for subsequent tasks. We demonstrate that our approach can substantially improve the agent's performance relative to other approaches, including an approach that transfers policies.
****

Distributed Exploration in Multi-Armed Bandits
Eshcar Hillel, Zohar S. Karnin, Tomer Koren, Ronny Lempel, Oren Somekh
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.
****

It is all in the noise: Efficient multi-task Gaussian process inference with structured residuals
Barbara Rakitsch, Christoph Lippert, Karsten Borgwardt, Oliver Stegle
Multi-task prediction models are widely being used to couple regressors or classification models by sharing information across related tasks. A common pitfall of these models is that they assume that the output tasks are independent conditioned on the inputs. Here, we propose a multi-task Gaussian process approach to model both the relatedness between regressors as well as the task correlations in the residuals, in order to more accurately identify true sharing between regressors. The resulting Gaussian model has a covariance term that is the sum of Kronecker products, for which efficient parameter inference and out of sample prediction are feasible. On both synthetic examples and applications to phenotype prediction in genetics, we find substantial benefits of modeling structured noise compared to established alternatives.
****

Projecting Ising Model Parameters for Fast Mixing
Justin Domke, Xianghang Liu
Inference in general Ising models is difficult, due to high treewidth making tree-based algorithms intractable. Moreover, when interactions are strong, Gibbs sampling may take exponential time to converge to the stationary distribution. We present an algorithm to project Ising model parameters onto a parameter set that

is guaranteed to be fast mixing, under several divergences. We find that Gibbs sampling using the projected parameters is more accurate than with the original parameters when interaction strengths are strong and when limited time is available for sampling.
****

Low-rank matrix reconstruction and clustering via approximate message passing
Ryosuke Matsushita, Toshiyuki Tanaka
We study the problem of reconstructing low-rank matrices from their noisy observations. We formulate the problem in the Bayesian framework, which allows us to exploit structural properties of matrices in addition to low-rankedness, such as sparsity. We propose an efficient approximate message passing algorithm, derived from the belief propagation algorithm, to perform the Bayesian inference for matrix reconstruction. We have also successfully applied the proposed algorithm to a clustering problem, by formulating the problem of clustering as a low-rank matrix reconstruction problem with an additional structural property. Numerical experiments show that the proposed algorithm outperforms Lloyd's K-means algorithm.
****

Inverse Density as an Inverse Problem: the Fredholm Equation Approach
Qichao Que, Mikhail Belkin
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****

Modeling Overlapping Communities with Node Popularities
Prem K. Gopalan, Chong Wang, David Blei
We develop a probabilistic approach for accurate network modeling using node popularities within the framework of the mixed-membership stochastic blockmodel (MMSB). Our model integrates some of the basic properties of nodes in social networks: homophily and preferential connection to popular nodes. We develop a scalable algorithm for posterior inference, based on a novel nonconjugate variant of stochastic variational inference. We evaluate the link prediction accuracy of our algorithm on eight real-world networks with up to 60,000 nodes, and 24 benchmark networks. We demonstrate that our algorithm predicts better than the MMSB. Further, using benchmark networks we show that node popularities are essential to achieving high accuracy in the presence of skewed degree distribution and noisy links---both characteristics of real networks.
****

Reflection methods for user-friendly submodular optimization
Stefanie Jegelka, Francis Bach, Suvrit Sra
Recently, it has become evident that submodularity naturally captures widely occurring concepts in machine learning, signal processing and computer vision. In consequence, there is need for efficient optimization procedures for submodular functions, in particular for minimization problems. While general submodular minimization is challenging, we propose a new approach that exploits existing decomposability of submodular functions. In contrast to previous approaches, our method is neither approximate, nor impractical, nor does it need any cumbersome parameter tuning. Moreover, it is easy to implement and parallelize. A key component of our approach is a formulation of the discrete submodular minimization problem as a continuous best approximation problem. It is solved through a sequence of reflections and its solution can be automatically thresholded to obtain an optimal discrete solution. Our method solves both the continuous and discrete formulations of the problem, and therefore has applications in learning, inference, and reconstruction. In our experiments, we show the benefits of our new algorithms for two image segmentation tasks.
****

Compressive Feature Learning
Hristo S. Paskov, Robert West, John C. Mitchell, Trevor Hastie
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.
****

Sparse nonnegative deconvolution for compressive calcium imaging: algorithms and phase transitions

Eftychios A. Pnevmatikakis, Liam Paninski

We propose a compressed sensing (CS) calcium imaging framework for monitoring la rge neuronal populations, where we image randomized projections of the spatial c alcium concentration at each timestep, instead of measuring the concentration at individual locations. We develop scalable nonnegative deconvolution methods for extracting the neuronal spike time series from such observations. We also addre ss the problem of demixing the spatial locations of the neurons using rank-penal ized matrix factorization methods. By exploiting the sparsity of neural spiking we demonstrate that the number of measurements needed per timestep is significan tly smaller than the total number of neurons, a result that can potentially enab le imaging of larger populations at considerably faster rates compared to tradit ional raster-scanning techniques. Unlike traditional CS setups, our problem invo lves a block-diagonal sensing matrix and a non-orthogonal sparse basis that span s multiple timesteps. We study the effect of these distinctive features in a noi seless setup using recent results relating conic geometry to CS. We provide tigh t approximations to the number of measurements needed for perfect deconvolution for certain classes of spiking processes, and show that this number displays a p hase transition," similar to phenomena observed in more standard CS settings; ho wever, in this case the required measurement rate depends not just on the mean s parsity level but also on other details of the underlying spiking process."
****

Probabilistic Low-Rank Matrix Completion with Adaptive Spectral Regularization A lgorithms

Adrien Todeschini, François Caron, Marie Chavent

We propose a novel class of algorithms for low rank matrix completion. Our appro ach builds on novel penalty functions on the singular values of the low rank mat rix. By exploiting a mixture model representation of this penalty, we show that a suitably chosen set of latent variables enables to derive an Expectation-Maxim ization algorithm to obtain a Maximum A Posteriori estimate of the completed low rank matrix. The resulting algorithm is an iterative soft-thresholded algorithm which iteratively adapts the shrinkage coefficients associated to the singular values. The algorithm is simple to implement and can scale to large matrices. We provide numerical comparisons between our approach and recent alternatives show ing the interest of the proposed approach for low rank matrix completion.
****

Global Solver and Its Efficient Approximation for Variational Bayesian Low-rank Subspace Clustering

Shinichi Nakajima, Akiko Takeda, S. Derin Babacan, Masashi Sugiyama, Ichiro Take uchi

When a probabilistic model and its prior are given, Bayesian learning offers in ference with automatic parameter tuning. However, Bayesian learning is often obs tructed by computational difficulty: the rigorous Bayesian learning is intracta ble in many models, and its variational Bayesian (VB) approximation is prone to suffer from local minima. In this paper, we overcome this difficulty for low-ran k subspace clustering (LRSC) by providing an exact global solver and its efficie nt approximation. LRSC extracts a low-dimensional structure of data by embedding samples into the union of low-dimensional subspaces, and its variational Bayes ian variant has shown good performance. We first prove a key property that the V B-LRSC model is highly redundant. Thanks to this property, the optimization prob lem of VB-LRSC can be separated into small subproblems, each of which has only a small number of unknown variables. Our exact global solver relies on another ke y property that the stationary condition of each subproblem is written as a set of polynomial equations, which is solvable with the homotopy method. For further computational efficiency, we also propose an efficient approximate variant, of

which the stationary condition can be written as a polynomial equation with a s
ingle variable. Experimental results show the usefulness of our approach.
****
Reservoir Boosting : Between Online and Offline Ensemble Learning
Leonidas Lefakis, François Fleuret
We propose to train an ensemble with the help of a reservoir in which the learni
ng algorithm can store a limited number of samples. This novel approach lies in
the area between offline and online ensemble approaches and can be seen either a
s a restriction of the former or an enhancement of the latter. We identify some
basic strategies that can be used to populate this reservoir and present our ma
in contribution, dubbed Greedy Edge Expectation Maximization (GEEM), that mainta
ins the reservoir content in the case of Boosting by viewing the samples through
their projections into the weak classifier response space. We propose an effic
ient algorithmic implementation which makes it tractable in practice, and demons
trate its efficiency experimentally on several compute-vision data-sets, on whic
h it outperforms both online and offline methods in a memory constrained setting
.
****
Faster Ridge Regression via the Subsampled Randomized Hadamard Transform
Yichao Lu, Paramveer Dhillon, Dean P. Foster, Lyle Ungar
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.
****
Convex Relaxations for Permutation Problems
Fajwel Fogel, Rodolphe Jenatton, Francis Bach, Alexandre D'Aspremont
Seriation seeks to reconstruct a linear order between variables using unsorted s
imilarity information. It has direct applications in archeology and shotgun gene
sequencing for example. We prove the equivalence between the seriation and the
combinatorial 2-sum problem (a quadratic minimization problem over permutations)
over a class of similarity matrices. The seriation problem can be solved exactl
y by a spectral algorithm in the noiseless case and we produce a convex relaxati
on for the 2-sum problem to improve the robustness of solutions in a noisy setti
ng. This relaxation also allows us to impose additional structural constraints
on the solution, to solve semi-supervised seriation problems. We present numeric
al experiments on archeological data, Markov chains and gene sequences.
****
Online Learning of Dynamic Parameters in Social Networks
Shahin Shahrampour, Sasha Rakhlin, Ali Jadbabaie
This paper addresses the problem of online learning in a dynamic setting. We con
sider a social network in which each individual observes a private signal about
the underlying state of the world and communicates with her neighbors at each ti
me period. Unlike many existing approaches, the underlying state is dynamic, and
evolves according to a geometric random walk. We view the scenario as an optimi
zation problem where agents aim to learn the true state while suffering the smal
lest possible loss. Based on the decomposition of the global loss function, we i
ntroduce two update mechanisms, each of which generates an estimate of the true
state. We establish a tight bound on the rate of change of the underlying state,
under which individuals can track the parameter with a bounded variance. Then,
we characterize explicit expressions for the steady state mean-square deviation(
MSD) of the estimates from the truth, per individual. We observe that only one o
f the estimators recovers the optimal MSD, which underscores the impact of the o
bjective function decomposition on the learning quality. Finally, we provide an
upper bound on the regret of the proposed methods, measured as an average of err
ors in estimating the parameter in a finite time.
****
Discovering Hidden Variables in Noisy-Or Networks using Quartet Tests
Yacine Jernite, Yonatan Halpern, David Sontag
We give a polynomial-time algorithm for provably learning the structure and para

meters of bipartite noisy-or Bayesian networks of binary variables where the top layer is completely hidden. Unsupervised learning of these models is a form of discrete factor analysis, enabling the discovery of hidden variables and their causal relationships with observed data. We obtain an efficient learning algorithm for a family of Bayesian networks that we call quartet-learnable, meaning that every latent variable has four children that do not have any other parents in common. We show that the existence of such a quartet allows us to uniquely identify each latent variable and to learn all parameters involving that latent variable. Underlying our algorithm are two new techniques for structure learning: a quartet test to determine whether a set of binary variables are singly coupled, and a conditional mutual information test that we use to learn parameters. We also show how to subtract already learned latent variables from the model to create new singly-coupled quartets, which substantially expands the class of structures that we can learn. Finally, we give a proof of the polynomial sample complexity of our learning algorithm, and experimentally compare it to variational EM.

****

Gaussian Process Conditional Copulas with Applications to Financial Time Series

José Miguel Hernández-Lobato, James R. Lloyd, Daniel Hernández-Lobato

The estimation of dependencies between multiple variables is a central problem in the analysis of financial time series. A common approach is to express these dependencies in terms of a copula function. Typically the copula function is assumed to be constant but this may be innacurate when there are covariates that could have a large influence on the dependence structure of the data. To account for this, a Bayesian framework for the estimation of conditional copulas is proposed. In this framework the parameters of a copula are non-linearly related to some arbitrary conditioning variables. We evaluate the ability of our method to predict time-varying dependencies on several equities and currencies and observe consistent performance gains compared to static copula models and other time-varying copula methods.

****

Non-Uniform Camera Shake Removal Using a Spatially-Adaptive Sparse Penalty

Haichao Zhang, David Wipf

Typical blur from camera shake often deviates from the standard uniform convolutional assumption, in part because of problematic rotations which create greater blurring away from some unknown center point. Consequently, successful blind deconvolution for removing shake artifacts requires the estimation of a spatially-varying or non-uniform blur operator. Using ideas from Bayesian inference and convex analysis, this paper derives a non-uniform blind deblurring algorithm with several desirable, yet previously-unexplored attributes. The underlying objective function includes a spatially-adaptive penalty that couples the latent sharp image, non-uniform blur operator, and noise level together. This coupling allows the penalty to automatically adjust its shape based on the estimated degree of local blur and image structure such that regions with large blur or few prominent edges are discounted. Remaining regions with modest blur and revealing edges therefore dominate the overall estimation process without explicitly incorporating structure-selection heuristics. The algorithm can be implemented using an optimization strategy that is virtually parameter free and simpler than existing methods. Detailed theoretical analysis and empirical validation on real images serve to validate the proposed method.

****

Online learning in episodic Markovian decision processes by relative entropy policy search

Alexander Zimin, Gergely Neu

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

****

Bayesian inference for low rank spatiotemporal neural receptive fields

Mijung Park, Jonathan W. Pillow

The receptive field (RF) of a sensory neuron describes how the neuron integrates sensory stimuli over time and space. In typical experiments with naturalistic or flickering spatiotemporal stimuli, RFs are very high-dimensional, due to the large number of coefficients needed to specify an integration profile across time and space. Estimating these coefficients from small amounts of data poses a variety of challenging statistical and computational problems. Here we address these challenges by developing Bayesian reduced rank regression methods for RF estimation. This corresponds to modeling the RF as a sum of several space-time separable (i.e., rank-1) filters, which proves accurate even for neurons with strongly oriented space-time RFs. This approach substantially reduces the number of parameters needed to specify the RF, from 1K-100K down to mere 100s in the examples we consider, and confers substantial benefits in statistical power and computational efficiency. In particular, we introduce a novel prior over low-rank RFs using the restriction of a matrix normal prior to the manifold of low-rank matrices. We then use a localized'' prior over row and column covariances to obtain sparse, smooth, localized estimates of the spatial and temporal RF components. We develop two methods for inference in the resulting hierarchical model: (1) a fully Bayesian method using blocked-Gibbs sampling; and (2) a fast, approximate method that employs alternating coordinate ascent of the conditional marginal likelihood. We develop these methods under Gaussian and Poisson noise models, and show that low-rank estimates substantially outperform full rank estimates in accuracy and speed using neural data from retina and V1."

****

Global MAP-Optimality by Shrinking the Combinatorial Search Area with Convex Relaxation

Bogdan Savchynskyy, Jörg Hendrik Kappes, Paul Swoboda, Christoph Schnörr

We consider energy minimization for undirected graphical models, also known as MAP-inference problem for Markov random fields. Although combinatorial methods, which return a provably optimal integral solution of the problem, made a big progress in the past decade, they are still typically unable to cope with large-scale datasets. On the other hand, large scale datasets are typically defined on sparse graphs, and convex relaxation methods, such as linear programming relaxations often provide good approximations to integral solutions. We propose a novel method of combining combinatorial and convex programming techniques to obtain a global solution of the initial combinatorial problem. Based on the information obtained from the solution of the convex relaxation, our method confines application of the combinatorial solver to a small fraction of the initial graphical model, which allows to optimally solve big problems. We demonstrate the power of our approach on a computer vision energy minimization benchmark.

****

Error-Minimizing Estimates and Universal Entry-Wise Error Bounds for Low-Rank Matrix Completion

Franz Kiraly, Louis Theran

We propose a general framework for reconstructing and denoising single entries of incomplete and noisy entries. We describe: effective algorithms for deciding if and entry can be reconstructed and, if so, for reconstructing and denoising it; and a priori bounds on the error of each entry, individually. In the noiseless case our algorithm is exact. For rank-one matrices, the new algorithm is fast, admits a highly-parallel implementation, and produces an error minimizing estimate that is qualitatively close to our theoretical and the state-of-the-art Nuclear Norm and OptSpace methods.

****

Decision Jungles: Compact and Rich Models for Classification

Jamie Shotton, Toby Sharp, Pushmeet Kohli, Sebastian Nowozin, John Winn, Antonio Criminisi

Randomized decision trees and forests have a rich history in machine learning and have seen considerable success in application, perhaps particularly so for computer vision. However, they face a fundamental limitation: given enough data, the number of nodes in decision trees will grow exponentially with depth. For cert

ain applications, for example on mobile or embedded processors, memory is a limi
ted resource, and so the exponential growth of trees limits their depth, and thu
s their potential accuracy. This paper proposes decision jungles, revisiting the
 idea of ensembles of rooted decision directed acyclic graphs (DAGs), and shows
these to be compact and powerful discriminative models for classification. Unlik
e conventional decision trees that only allow one path to every node, a DAG in a
 decision jungle allows multiple paths from the root to each leaf. We present an
d compare two new node merging algorithms that jointly optimize both the feature
s and the structure of the DAGs efficiently. During training, node splitting and
 node merging are driven by the minimization of exactly the same objective funct
ion, here the weighted sum of entropies at the leaves. Results on varied dataset
s show that, compared to decision forests and several other baselines, decision
jungles require dramatically less memory while considerably improving generaliza
tion.
****

Bayesian Estimation of Latently-grouped Parameters in Undirected Graphical Model
s
Jie Liu, David Page
In large-scale applications of undirected graphical models, such as social netwo
rks and biological networks, similar patterns occur frequently and give rise to
similar parameters. In this situation, it is beneficial to group the parameters
for more efficient learning. We show that even when the grouping is unknown, we
can infer these parameter groups during learning via a Bayesian approach. We imp
ose a Dirichlet process prior on the parameters. Posterior inference usually inv
olves calculating intractable terms, and we propose two approximation algorithms
, namely a Metropolis-Hastings algorithm with auxiliary variables and a Gibbs sa
mpling algorithm with stripped Beta approximation (GibbsSBA). Simulations show t
hat both algorithms outperform conventional maximum likelihood estimation (MLE).
 GibbsSBA's performance is close to Gibbs sampling with exact likelihood calcula
tion. Models learned with Gibbs_SBA also generalize better than the models learn
ed by MLE on real-world Senate voting data.
****

(More) Efficient Reinforcement Learning via Posterior Sampling
Ian Osband, Daniel Russo, 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.
****

Forgetful Bayes and myopic planning: Human learning and decision-making in a ban
dit setting
Shunan Zhang, Angela J. Yu
How humans achieve long-term goals in an uncertain environment, via repeated tri
als and noisy observations, is an important problem in cognitive science. We inv
estigate this behavior in the context of a multi-armed bandit task. We compare h
uman behavior to a variety of models that vary in their representational and com
putational complexity. Our result shows that subjects' choices, on a trial-to-tr
ial basis, are best captured by a forgetful" Bayesian iterative learning model i
n combination with a partially myopic decision policy known as Knowledge Gradien
t. This model accounts for subjects' trial-by-trial choice better than a number
of other previously proposed models, including optimal Bayesian learning and ris
k minimization, epsilon-greedy and win-stay-lose-shift. It has the added benefit
 of being closest in performance to the optimal Bayesian model than all the othe
r heuristic models that have the same computational complexity (all are signific
antly less complex than the optimal model). These results constitute an advancem
ent in the theoretical understanding of how humans negotiate the tension between
 exploration and exploitation in a noisy, imperfectly known environment."
****

Stochastic Optimization of PCA with Capped MSG
Raman Arora, Andy Cotter, Nati Srebro

We study PCA as a stochastic optimization problem and propose a novel stochastic approximation algorithm which we refer to as Matrix Stochastic Gradient'' (MSG), as well as a practical variant, Capped MSG. We study the method both theoretically and empirically. "
****
Embed and Project: Discrete Sampling with Universal Hashing
Stefano Ermon, Carla P. Gomes, Ashish Sabharwal, Bart Selman
We consider the problem of sampling from a probability distribution defined over a high-dimensional discrete set, specified for instance by a graphical model. We propose a sampling algorithm, called PAWS, based on embedding the set into a higher-dimensional space which is then randomly projected using universal hash functions to a lower-dimensional subspace and explored using combinatorial search methods. Our scheme can leverage fast combinatorial optimization tools as a blackbox and, unlike MCMC methods, samples produced are guaranteed to be within an (arbitrarily small) constant factor of the true probability distribution. We demonstrate that by using state-of-the-art combinatorial search tools, PAWS can efficiently sample from Ising grids with strong interactions and from software verification instances, while MCMC and variational methods fail in both cases.
****
Optimal Neural Population Codes for High-dimensional Stimulus Variables
Zhuo Wang, Alan A. Stocker, Daniel D. Lee
How does neural population process sensory information? Optimal coding theories assume that neural tuning curves are adapted to the prior distribution of the stimulus variable. Most of the previous work has discussed optimal solutions for only one-dimensional stimulus variables. Here, we expand some of these ideas and present new solutions that define optimal tuning curves for high-dimensional stimulus variables. We consider solutions for a minimal case where the number of neurons in the population is equal to the number of stimulus dimensions (diffeomorphic). In the case of two-dimensional stimulus variables, we analytically derive optimal solutions for different optimal criteria such as minimal L2 reconstruction error or maximal mutual information. For higher dimensional case, the learning rule to improve the population code is provided.
****
Near-Optimal Entrywise Sampling for Data Matrices
Dimitris Achlioptas, Zohar S. Karnin, Edo Liberty
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
A Comparative Framework for Preconditioned Lasso Algorithms
Fabian L. Wauthier, Nebojsa Jojic, Michael I. Jordan
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Universal models for binary spike patterns using centered Dirichlet processes
Il Memming Park, Evan W. Archer, Kenneth Latimer, Jonathan W. Pillow
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.
****
What Are the Invariant Occlusive Components of Image Patches? A Probabilistic Generative Approach
Zhenwen Dai, Georgios Exarchakis, Jörg Lücke
We study optimal image encoding based on a generative approach with non-linear feature combinations and explicit position encoding. By far most approaches to unsupervised learning learning of visual features, such as sparse coding or ICA, a

ccount for translations by representing the same features at different positions . Some earlier models used a separate encoding of features and their positions t o facilitate invariant data encoding and recognition. All probabilistic generati ve models with explicit position encoding have so far assumed a linear superposi tion of components to encode image patches. Here, we for the first time apply a model with non-linear feature superposition and explicit position encoding. By a voiding linear superpositions, the studied model represents a closer match to co mponent occlusions which are ubiquitous in natural images. In order to account f or occlusions, the non-linear model encodes patches qualitatively very different from linear models by using component representations separated into mask and f eature parameters. We first investigated encodings learned by the model using ar tificial data with mutually occluding components. We find that the model extract s the components, and that it can correctly identify the occlusive components wi th the hidden variables of the model. On natural image patches, the model learns component masks and features for typical image components. By using reverse cor relation, we estimate the receptive fields associated with the model's hidden un its. We find many Gabor-like or globular receptive fields as well as fields sens itive to more complex structures. Our results show that probabilistic models tha t capture occlusions and invariances can be trained efficiently on image patches , and that the resulting encoding represents an alternative model for the neural encoding of images in the primary visual cortex.

****

Correlations strike back (again): the case of associative memory retrieval

Cristina Savin, Peter Dayan, Mate Lengyel

It has long been recognised that statistical dependencies in neuronal activity n eed to be taken into account when decoding stimuli encoded in a neural populatio n. Less studied, though equally pernicious, is the need to take account of depen dencies between synaptic weights when decoding patterns previously encoded in an auto-associative memory. We show that activity-dependent learning generically p roduces such correlations, and failing to take them into account in the dynamics of memory retrieval leads to catastrophically poor recall. We derive optimal ne twork dynamics for recall in the face of synaptic correlations caused by a range of synaptic plasticity rules. These dynamics involve well-studied circuit motif s, such as forms of feedback inhibition and experimentally observed dendritic no nlinearities. We therefore show how addressing the problem of synaptic correlati ons leads to a novel functional account of key biophysical features of the neura l substrate.

****

Understanding Dropout

Pierre Baldi, Peter J. Sadowski

Dropout is a relatively new algorithm for training neural networks which relies on stochastically dropping out'' neurons during training in order to avoid the c o-adaptation of feature detectors. We introduce a general formalism for studying dropout on either units or connections, with arbitrary probability values, and use it to analyze the averaging and regularizing properties of dropout in both l inear and non-linear networks. For deep neural networks, the averaging propertie s of dropout are characterized by three recursive equations, including the appro ximation of expectations by normalized weighted geometric means. We provide esti mates and bounds for these approximations and corroborate the results with simul ations. We also show in simple cases how dropout performs stochastic gradient de scent on a regularized error function."

****

Supervised Sparse Analysis and Synthesis Operators

Pablo Sprechmann, Roee Litman, Tal Ben Yakar, Alexander M. Bronstein, Guillermo Sapiro

In this paper, we propose a new and computationally efficient framework for lear ning sparse models. We formulate a unified approach that contains as particular cases models promoting sparse synthesis and analysis type of priors, and mixture s thereof. The supervised training of the proposed model is formulated as a bile vel optimization problem, in which the operators are optimized to achieve the be

st possible performance on a specific task, e.g., reconstruction or classificati on. By restricting the operators to be shift invariant, our approach can be thou ght as a way of learning analysis+synthesis sparsity-promoting convolutional ope rators. Leveraging recent ideas on fast trainable regressors designed to approxi mate exact sparse codes, we propose a way of constructing feed-forward neural ne tworks capable of approximating the learned models at a fraction of the computat ional cost of exact solvers. In the shift-invariant case, this leads to a princi pled way of constructing task-specific convolutional networks. We illustrate the proposed models on several experiments in music analysis and image processing a pplications.

****

The Pareto Regret Frontier

Wouter M. Koolen

Performance guarantees for online learning algorithms typically take the form of regret bounds, which express that the cumulative loss overhead compared to the best expert in hindsight is small. In the common case of large but structured ex pert sets we typically wish to keep the regret especially small compared to simp le experts, at the cost of modest additional overhead compared to more complex others. We study which such regret trade-offs can be achieved, and how. We anal yse regret w.r.t. each individual expert as a multi-objective criterion in the s imple but fundamental case of absolute loss. We characterise the achievable and Pareto optimal trade-offs, and the corresponding optimal strategies for each sam ple size both exactly for each finite horizon and asymptotically.

****

Approximate Dynamic Programming Finally Performs Well in the Game of Tetris

Victor Gabillon, Mohammad Ghavamzadeh, Bruno Scherrer

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 Feature Selection Dependencies in Multi-task Learning

Daniel Hernández-Lobato, José Miguel Hernández-Lobato

A probabilistic model based on the horseshoe prior is proposed for learning depe ndencies in the process of identifying relevant features for prediction. Exact i nference is intractable in this model. However, expectation propagation offers a n approximate alternative. Because the process of estimating feature selection d ependencies may suffer from over-fitting in the model proposed, additional data from a multi-task learning scenario are considered for induction. The same model can be used in this setting with few modifications. Furthermore, the assumption s made are less restrictive than in other multi-task methods: The different task s must share feature selection dependencies, but can have different relevant fea tures and model coefficients. Experiments with real and synthetic data show that this model performs better than other multi-task alternatives from the literatu re. The experiments also show that the model is able to induce suitable feature selection dependencies for the problems considered, only from the training data.

****

Dimension-Free Exponentiated Gradient

Francesco Orabona

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.

****

Memory Limited, Streaming PCA

Ioannis Mitliagkas, Constantine Caramanis, Prateek Jain

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

****
Σ-Optimality for Active Learning on Gaussian Random Fields
Yifei Ma, Roman Garnett, Jeff Schneider
A common classifier for unlabeled nodes on undirected graphs uses label propagation from the labeled nodes, equivalent to the harmonic predictor on Gaussian random fields (GRFs). For active learning on GRFs, the commonly used V-optimality criterion queries nodes that reduce the L2 (regression) loss. V-optimality satisfies a submodularity property showing that greedy reduction produces a (1 − 1/e) globally optimal solution. However, L2 loss may not characterise the true nature of 0/1 loss in classification problems and thus may not be the best choice for active learning. We consider a new criterion we call Σ-optimality, which queries the node that minimizes the sum of the elements in the predictive covariance. Σ-optimality directly optimizes the risk of the surveying problem, which is to determine the proportion of nodes belonging to one class. In this paper we extend submodularity guarantees from V-optimality to Σ-optimality using properties specific to GRFs. We further show that GRFs satisfy the suppressor-free condition in addition to the conditional independence inherited from Markov random fields. We test Σ-optimality on real-world graphs with both synthetic and real data and show that it outperforms V-optimality and other related methods on classification.
****
Recurrent linear models of simultaneously-recorded neural populations
Marius Pachitariu, Biljana Petreska, Maneesh Sahani
Population neural recordings with long-range temporal structure are often best understood in terms of a shared underlying low-dimensional dynamical process. Advances in recording technology provide access to an ever larger fraction of the population, but the standard computational approaches available to identify the collective dynamics scale poorly with the size of the dataset. Here we describe a new, scalable approach to discovering the low-dimensional dynamics that underlie simultaneously recorded spike trains from a neural population. Our method is based on recurrent linear models (RLMs), and relates closely to timeseries models based on recurrent neural networks. We formulate RLMs for neural data by generalising the Kalman-filter-based likelihood calculation for latent linear dynamical systems (LDS) models to incorporate a generalised-linear observation process. We show that RLMs describe motor-cortical population data better than either directly-coupled generalised-linear models or latent linear dynamical system models with generalised-linear observations. We also introduce the cascaded linear model (CLM) to capture low-dimensional instantaneous correlations in neural populations. The CLM describes the cortical recordings better than either Ising or Gaussian models and, like the RLM, can be fit exactly and quickly. The CLM can also be seen as a generalization of a low-rank Gaussian model, in this case factor analysis. The computational tractability of the RLM and CLM allow both to scale to very high-dimensional neural data.
****
On the Complexity and Approximation of Binary Evidence in Lifted Inference
Guy Van den Broeck, Adnan Darwiche
Lifted inference algorithms exploit symmetries in probabilistic models to speed up inference. They show impressive performance when calculating unconditional probabilities in relational models, but often resort to non-lifted inference when computing conditional probabilities. The reason is that conditioning on evidence breaks many of the model's symmetries, which preempts standard lifting techniques. Recent theoretical results show, for example, that conditioning on evidence which corresponds to binary relations is #P-hard, suggesting that no lifting is to be expected in the worst case. In this paper, we balance this grim result by identifying the Boolean rank of the evidence as a key parameter for characterizing the complexity of conditioning in lifted inference. In particular, we show that conditioning on binary evidence with bounded Boolean rank is efficient. This opens up the possibility of approximating evidence by a low-rank Boolean matrix factorization, which we investigate both theoretically and empirically.
****

Pass-efficient unsupervised feature selection
Crystal Maung, Haim Schweitzer
The goal of unsupervised feature selection is to identify a small number of impo rtant features that can represent the data. We propose a new algorithm, a modifi cation of the classical pivoted QR algorithm of Businger and Golub, that require s a small number of passes over the data. The improvements are based on two idea s: keeping track of multiple features in each pass, and skipping calculations th at can be shown not to affect the final selection. Our algorithm selects the exa ct same features as the classical pivoted QR algorithm, and has the same favorab le numerical stability. We describe experiments on real-world datasets which som etimes show improvements of {\em several orders of magnitude} over the classical algorithm. These results appear to be competitive with recently proposed rando mized algorithms in terms of pass efficiency and run time. On the other hand, th e randomized algorithms may produce better features, at the cost of small probab ility of failure.
****
Adaptive dropout for training deep neural networks
Jimmy Ba, Brendan Frey
Recently, it was shown that by dropping out hidden activities with a probability of 0.5, deep neural networks can perform very well. We describe a model in whic h a binary belief network is overlaid on a neural network and is used to decreas e the information content of its hidden units by selectively setting activities to zero. This ''dropout network can be trained jointly with the neural network b y approximately computing local expectations of binary dropout variables, comput ing derivatives using back-propagation, and using stochastic gradient descent. I nterestingly, experiments show that the learnt dropout network parameters recapi tulate the neural network parameters, suggesting that a good dropout network reg ularizes activities according to magnitude. When evaluated on the MNIST and NORB datasets, we found our method can be used to achieve lower classification error rates than other feather learning methods, including standard dropout, denoisin g auto-encoders, and restricted Boltzmann machines. For example, our model achie ves 5.8% error on the NORB test set, which is better than state-of-the-art resul ts obtained using convolutional architectures. "
****
On the Representational Efficiency of Restricted Boltzmann Machines
James Martens, Arkadev Chattopadhya, Toni Pitassi, Richard Zemel
This paper examines the question: What kinds of distributions can be efficiently represented by Restricted Boltzmann Machines (RBMs)? We characterize the RBM' s unnormalized log-likelihood function as a type of neural network (called an RB M network), and through a series of simulation results relate these networks to types that are better understood. We show the surprising result that RBM networ ks can efficiently compute any function that depends on the number of 1's in the input, such as parity. We also provide the first known example of a particular type of distribution which provably cannot be efficiently represented by an RBM (or equivalently, cannot be efficiently computed by an RBM network), assuming a realistic exponential upper bound on the size of the weights. By formally demo nstrating that a relatively simple distribution cannot be represented efficientl y by an RBM our results provide a new rigorous justification for the use of pote ntially more expressive generative models, such as deeper ones.
****
Robust Spatial Filtering with Beta Divergence
Wojciech Samek, Duncan Blythe, Klaus-Robert Müller, Motoaki Kawanabe
The efficiency of Brain-Computer Interfaces (BCI) largely depends upon a reliabl e extraction of informative features from the high-dimensional EEG signal. A cru cial step in this protocol is the computation of spatial filters. The Common Spa tial Patterns (CSP) algorithm computes filters that maximize the difference in b and power between two conditions, thus it is tailored to extract the relevant in formation in motor imagery experiments. However, CSP is highly sensitive to arti facts in the EEG data, i.e. few outliers may alter the estimate drastically and decrease classification performance. Inspired by concepts from the field of info

rmation geometry we propose a novel approach for robustifying CSP. More precisel y, we formulate CSP as a divergence maximization problem and utilize the propert y of a particular type of divergence, namely beta divergence, for robustifying t he estimation of spatial filters in the presence of artifacts in the data. We de monstrate the usefulness of our method on toy data and on EEG recordings from 80 subjects.

****

DeViSE: A Deep Visual-Semantic Embedding Model

Andrea Frome, Greg S. Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, Tomas Mikolov

Modern visual recognition systems are often limited in their ability to scale to large numbers of object categories. This limitation is in part due to the incre asing difficulty of acquiring sufficient training data in the form of labeled im ages as the number of object categories grows. One remedy is to leverage data fr om other sources -- such as text data -- both to train visual models and to cons train their predictions. In this paper we present a new deep visual-semantic emb edding model trained to identify visual objects using both labeled image data as well as semantic information gleaned from unannotated text. We demonstrate that this model matches state-of-the-art performance on the 1000-class ImageNet obje ct recognition challenge while making more semantically reasonable errors, and a lso show that the semantic information can be exploited to make predictions abou t tens of thousands of image labels not observed during training. Semantic knowl edge improves such zero-shot predictions by up to 65%, achieving hit rates of up to 10% across thousands of novel labels never seen by the visual model.

****

Symbolic Opportunistic Policy Iteration for Factored-Action MDPs

Aswin Raghavan, Roni Khardon, Alan Fern, Prasad Tadepalli

We address the scalability of symbolic planning under uncertainty with factored states and actions. Prior work has focused almost exclusively on factored states but not factored actions, and on value iteration (VI) compared to policy iterat ion (PI). Our ■rst contribution is a novel method for symbolic policy backups vi a the application of constraints, which is used to yield a new ef■cient symbolic imple- mentation of modi■ed PI (MPI) for factored action spaces. While this app roach improves scalability in some cases, naive handling of policy constraints c omes with its own scalability issues. This leads to our second and main contribu tion, symbolic Opportunistic Policy Iteration (OPI), which is a novel convergent al- gorithm lying between VI and MPI. The core idea is a symbolic procedure tha t applies policy constraints only when they reduce the space and time complexity of the update, and otherwise performs full Bellman backups, thus automatically adjusting the backup per state. We also give a memory bounded version of this al gorithm allowing a space-time tradeoff. Empirical results show signi■cantly impr oved scalability over the state-of-the-art.

****

Least Informative Dimensions

Fabian Sinz, Anna Stockl, Jan Grewe, Jan Benda

We present a novel non-parametric method for finding a subspace of stimulus feat ures that contains all information about the response of a system. Our method ge neralizes similar approaches to this problem such as spike triggered average, sp ike triggered covariance, or maximally informative dimensions. Instead of maximi zing the mutual information between features and responses directly, we use inte gral probability metrics in kernel Hilbert spaces to minimize the information be tween uninformative features and the combination of informative features and res ponses. Since estimators of these metrics access the data via kernels, are easy to compute, and exhibit good theoretical convergence properties, our method can easily be generalized to populations of neurons or spike patterns. By using a pa rticular expansion of the mutual information, we can show that the informative f eatures must contain all information if we can make the uninformative features i ndependent of the rest.

****

A memory frontier for complex synapses

Subhaneil Lahiri, Surya Ganguli
An incredible gulf separates theoretical models of synapses, often described sol ely by a single scalar value denoting the size of a postsynaptic potential, from the immense complexity of molecular signaling pathways underlying real synapses . To understand the functional contribution of such molecular complexity to lear ning and memory, it is essential to expand our theoretical conception of a synap se from a single scalar to an entire dynamical system with many internal molecul ar functional states. Moreover, theoretical considerations alone demand such an expansion; network models with scalar synapses assuming finite numbers of distin guishable synaptic strengths have strikingly limited memory capacity. This raise s the fundamental question, how does synaptic complexity give rise to memory? To address this, we develop new mathematical theorems elucidating the relationship between the structural organization and memory properties of complex synapses t hat are themselves molecular networks. Moreover, in proving such theorems, we un cover a framework, based on first passage time theory, to impose an order on the internal states of complex synaptic models, thereby simplifying the relationshi p between synaptic structure and function.
****

Data-driven Distributionally Robust Polynomial Optimization
Martin Mevissen, Emanuele Ragnoli, Jia Yuan Yu
We consider robust optimization for polynomial optimization problems where the u ncertainty set is a set of candidate probability density functions. This set is a ball around a density function estimated from data samples, i.e., it is data-d riven and random. Polynomial optimization problems are inherently hard due to n onconvex objectives and constraints. However, we show that by employing polynom ial and histogram density estimates, we can introduce robustness with respect to distributional uncertainty sets without making the problem harder. We show tha t the solution to the distributionally robust problem is the limit of a sequence of tractable semidefinite programming relaxations. We also give finite-sample consistency guarantees for the data-driven uncertainty sets. Finally, we apply our model and solution method in a water network problem.
****

Learning Stochastic Inverses
Andreas Stuhlmüller, Jacob Taylor, Noah Goodman
We describe a class of algorithms for amortized inference in Bayesian networks. In this setting, we invest computation upfront to support rapid online inference for a wide range of queries. Our approach is based on learning an inverse facto rization of a model's joint distribution: a factorization that turns observation s into root nodes. Our algorithms accumulate information to estimate the local conditional distributions that constitute such a factorization. These stochastic inverses can be used to invert each of the computation steps leading to an obse rvation, sampling backwards in order to quickly find a likely explanation. We sh ow that estimated inverses converge asymptotically in number of (prior or poster ior) training samples. To make use of inverses before convergence, we describe the Inverse MCMC algorithm, which uses stochastic inverses to make block proposa ls for a Metropolis-Hastings sampler. We explore the efficiency of this sampler for a variety of parameter regimes and Bayes nets.
****

Stochastic Ratio Matching of RBMs for Sparse High-Dimensional Inputs
Yann Dauphin, Yoshua Bengio
Sparse high-dimensional data vectors are common in many application domains wher e a very large number of rarely non-zero features can be devised. Unfortunately, this creates a computational bottleneck for unsupervised feature learning algor ithms such as those based on auto-encoders and RBMs, because they involve a reco nstruction step where the whole input vector is predicted from the current featu re values. An algorithm was recently developed to successfully handle the case o f auto-encoders, based on an importance sampling scheme stochastically selectin g which input elements to actually reconstruct during training for each particul ar example. To generalize this idea to RBMs, we propose a stochastic ratio-matc hing algorithm that inherits all the computational advantages and unbiasedness o

f the importance sampling scheme. We show that stochastic ratio matching is a go od estimator, allowing the approach to beat the state-of-the-art on two bag-of-w ord text classification benchmarks (20 Newsgroups and RCV1), while keeping compu tational cost linear in the number of non-zeros.
****

Distributed k-means and k-median Clustering on General Topologies
Maria-Florina F. Balcan, Steven Ehrlich, Yingyu Liang
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-auth ors prior to requesting a name change in the electronic proceedings.
****

Non-strongly-convex smooth stochastic approximation with convergence rate O(1/n)
Francis Bach, Eric Moulines
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.
****

Predicting Parameters in Deep Learning
Misha Denil, Babak Shakibi, Laurent Dinh, Marc'Aurelio Ranzato, Nando de Freitas
We demonstrate that there is significant redundancy in the parameterization of several deep learning models. Given only a few weight values for each feature it is possible to accurately predict the remaining values. Moreover, we show that not only can the parameter values be predicted, but many of them need not be learned at all. We train several different architectures by learning only a small number of weights and predicting the rest. In the best case we are able to predict more than 95% of the weights of a network without any drop in accuracy.
****

Estimation Bias in Multi-Armed Bandit Algorithms for Search Advertising
Min Xu, Tao Qin, Tie-Yan Liu
In search advertising, the search engine needs to select the most profitable adv ertisements to display, which can be formulated as an instance of online learnin g with partial feedback, also known as the stochastic multi-armed bandit (MAB) p roblem. In this paper, we show that the naive application of MAB algorithms to s earch advertising for advertisement selection will produce sample selection bias that harms the search engine by decreasing expected revenue and “estimation of the largest mean” (ELM) bias that harms the advertisers by increasing game-theor etic player-regret. We then propose simple bias-correction methods with benefits to both the search engine and the advertisers.
****

Learning Efficient Random Maximum A-Posteriori Predictors with Non-Decomposable Loss Functions
Tamir Hazan, Subhransu Maji, Joseph Keshet, Tommi Jaakkola
In this work we develop efficient methods for learning random MAP predictors for structured label problems. In particular, we construct posterior distributions over perturbations that can be adjusted via stochastic gradient methods. We show that every smooth posterior distribution would suffice to define a smooth PAC-B ayesian risk bound suitable for gradient methods. In addition, we relate the pos terior distributions to computational properties of the MAP predictors. We sugge st multiplicative posteriors to learn super-modular potential functions that acc ompany specialized MAP predictors such as graph-cuts. We also describe label-aug mented posterior models that can use efficient MAP approximations, such as those arising from linear program relaxations.
****

q-OCSVM: A q-Quantile Estimator for High-Dimensional Distributions
Assaf Glazer, Michael Lindenbaum, Shaul Markovitch
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.
****

Fantope Projection and Selection: A near-optimal convex relaxation of sparse PCA
Vincent Q. Vu, Juhee Cho, Jing Lei, Karl Rohe
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 Template Evaluation with Vector Quantization
Mohammad Amin Sadeghi, David Forsyth
Applying linear templates is an integral part of many object detection systems a
nd accounts for a significant portion of computation time. We describe a method
that achieves a substantial end-to-end speedup over the best current methods, wi
thout loss of accuracy. Our method is a combination of approximating scores by v
ector quantizing feature windows and a number of speedup techniques including ca
scade. Our procedure allows speed and accuracy to be traded off in two ways: by
choosing the number of Vector Quantization levels, and by choosing to rescore wi
ndows or not. Our method can be directly plugged into any recognition system tha
t relies on linear templates. We demonstrate our method to speed up the original
Exemplar SVM detector [1] by an order of magnitude and Deformable Part models [
2] by two orders of magnitude with no loss of accuracy.
****

Sparse Additive Text Models with Low Rank Background
Lei Shi
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.
****

Correlated random features for fast semi-supervised learning
Brian McWilliams, David Balduzzi, Joachim M. Buhmann
This paper presents Correlated Nystrom Views (XNV), a fast semi-supervised algor
ithm for regression and classification. The algorithm draws on two main ideas. F
irst, it generates two views consisting of computationally inexpensive random fe
atures. Second, multiview regression, using Canonical Correlation Analysis (CCA)
on unlabeled data, biases the regression towards useful features. It has been s
hown that CCA regression can substantially reduce variance with a minimal increa
se in bias if the views contains accurate estimators. Recent theoretical and emp
irical work shows that regression with random features closely approximates kern
el regression, implying that the accuracy requirement holds for random views. We
show that XNV consistently outperforms a state-of-the-art algorithm for semi-su
pervised learning: substantially improving predictive performance and reducing t
he variability of performance on a wide variety of real-world datasets, whilst a
lso reducing runtime by orders of magnitude.
****

Variational Planning for Graph-based MDPs
Qiang Cheng, Qiang Liu, Feng Chen, Alexander T. Ihler
Markov Decision Processes (MDPs) are extremely useful for modeling and solving s
equential decision making problems. Graph-based MDPs provide a compact represent
ation for MDPs with large numbers of random variables. However, the complexity o
f exactly solving a graph-based MDP usually grows exponentially in the number of
variables, which limits their application. We present a new variational framewo
rk to describe and solve the planning problem of MDPs, and derive both exact and
approximate planning algorithms. In particular, by exploiting the graph structu
re of graph-based MDPs, we propose a factored variational value iteration algori
thm in which the value function is first approximated by the multiplication of l
ocal-scope value functions, then solved by minimizing a Kullback-Leibler (KL) di
vergence. The KL divergence is optimized using the belief propagation algorithm,

with complexity exponential in only the cluster size of the graph. Experimental comparison on different models shows that our algorithm outperforms existing approximation algorithms at finding good policies.
****

Adaptive Anonymity via b-Matching
Krzysztof M. Choromanski, Tony Jebara, Kui Tang
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****

Statistical Active Learning Algorithms
Maria-Florina F. Balcan, Vitaly Feldman
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****

The Power of Asymmetry in Binary Hashing
Behnam Neyshabur, Nati Srebro, Russ R. Salakhutdinov, Yury Makarychev, Payman Yadollahpour
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 Mixture Modelling and Inference based Thompson Sampling in Monte-Carlo Tree Search
Aijun Bai, Feng Wu, Xiaoping Chen
Monte-Carlo tree search is drawing great interest in the domain of planning under uncertainty, particularly when little or no domain knowledge is available. One of the central problems is the trade-off between exploration and exploitation. In this paper we present a novel Bayesian mixture modelling and inference based Thompson sampling approach to addressing this dilemma. The proposed Dirichlet-NormalGamma MCTS (DNG-MCTS) algorithm represents the uncertainty of the accumulated reward for actions in the MCTS search tree as a mixture of Normal distributions and inferences on it in Bayesian settings by choosing conjugate priors in the form of combinations of Dirichlet and NormalGamma distributions. Thompson sampling is used to select the best action at each decision node. Experimental results show that our proposed algorithm has achieved the state-of-the-art comparing with popular UCT algorithm in the context of online planning for general Markov decision processes.
****

Distributed Submodular Maximization: Identifying Representative Elements in Massive Data
Baharan Mirzasoleiman, Amin Karbasi, Rik Sarkar, Andreas Krause
Many large-scale machine learning problems (such as clustering, non-parametric learning, kernel machines, etc.) require selecting, out of a massive data set, a manageable, representative subset. Such problems can often be reduced to maximizing a submodular set function subject to cardinality constraints. Classical approaches require centralized access to the full data set; but for truly large-scale problems, rendering the data centrally is often impractical. In this paper, we consider the problem of submodular function maximization in a distributed fashion. We develop a simple, two-stage protocol GreeDI, that is easily implemented using MapReduce style computations. We theoretically analyze our approach, and show, that under certain natural conditions, performance close to the (impractical) centralized approach can be achieved. In our extensive experiments, we demonstrate the effectiveness of our approach on several applications, including sparse Gaussian process inference on tens of millions of examples using Hadoop.
****

Analyzing the Harmonic Structure in Graph-Based Learning
Xiao-Ming Wu, Zhenguo Li, Shih-Fu Chang
We show that either explicitly or implicitly, various well-known graph-based mod
els exhibit a common significant \emph{harmonic} structure in its target functio
n -- the value of a vertex is approximately the weighted average of the values o
f its adjacent neighbors. Understanding of such structure and analysis of the lo
ss defined over such structure help reveal important properties of the target fu
nction over a graph. In this paper, we show that the variation of the target fun
ction across a cut can be upper and lower bounded by the ratio of its harmonic l
oss and the cut cost. We use this to develop an analytical tool and analyze 5 po
pular models in graph-based learning: absorbing random walks, partially absorbin
g random walks, hitting times, pseudo-inverse of graph Laplacian, and eigenvecto
rs of the Laplacian matrices. Our analysis well explains several open questions
of these models reported in the literature. Furthermore, it provides theoretical
 justifications and guidelines for their practical use. Simulations on synthetic
 and real datasets support our analysis.
****

Near-optimal Anomaly Detection in Graphs using Lovasz Extended Scan Statistic
James L. Sharpnack, Akshay Krishnamurthy, Aarti Singh
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.
****

Bellman Error Based Feature Generation using Random Projections on Sparse Spaces
Mahdi Milani Fard, Yuri Grinberg, Amir-massoud Farahmand, Joelle Pineau, Doina P
recup
This paper addresses the problem of automatic generation of features for value f
unction approximation in reinforcement learning. Bellman Error Basis Functions
(BEBFs) have been shown to improve the error of policy evaluation with function
approximation, with a convergence rate similar to that of value iteration. We pr
opose a simple, fast and robust algorithm based on random projections, which gen
erates BEBFs for sparse feature spaces. We provide a finite sample analysis of t
he proposed method, and prove that projections logarithmic in the dimension of t
he original space guarantee a contraction in the error. Empirical results demon
strate the strength of this method in domains in which choosing a good state rep
resentation is challenging.
****

Similarity Component Analysis
Soravit Changpinyo, Kuan Liu, Fei Sha
Measuring similarity is crucial to many learning tasks. It is also a richer and
broader notion than what most metric learning algorithms can model. For example,
 similarity can arise from the process of aggregating the decisions of multiple
latent components, where each latent component compares data in its own way by f
ocusing on a different subset of features. In this paper, we propose Similarity
Component Analysis (SCA), a probabilistic graphical model that discovers those l
atent components from data. In SCA, a latent component generates a local similar
ity value, computed with its own metric, independently of other components. The
final similarity measure is then obtained by combining the local similarity valu
es with a (noisy-)OR gate. We derive an EM-based algorithm for fitting the model
 parameters with similarity-annotated data from pairwise comparisons. We validat
e the SCA model on synthetic datasets where SCA discovers the ground-truth about
 the latent components. We also apply SCA to a multiway classification task and
a link prediction task. For both tasks, SCA attains significantly better predict
ion accuracies than competing methods. Moreover, we show how SCA can be instrume
ntal in exploratory analysis of data, where we gain insights about the data by e
xamining patterns hidden in its latent components' local similarity values.
****

Matrix Completion From any Given Set of Observations
Troy Lee, Adi Shraibman

In the matrix completion problem the aim is to recover an unknown real matrix fr
om a subset of its entries. This problem comes up in many application areas, an
d has received a great deal of attention in the context of the netflix prize. A
 central approach to this problem is to output a matrix of lowest possible comp
lexity (e.g. rank or trace norm) that agrees with the partially specified matri
x. The performance of this approach under the assumption that the revealed entr
ies are sampled randomly has received considerable attention. In practice, oft
en the set of revealed entries is not chosen at random and these results do not
apply. We are therefore left with no guarantees on the performance of the algori
thm we are using. We present a means to obtain performance guarantees with resp
ect to any set of initial observations. The first step remains the same: find
a matrix of lowest possible complexity that agrees with the partially specified
matrix. We give a new way to interpret the output of this algorithm by next fi
nding a probability distribution over the non-revealed entries with respect to
which a bound on the generalization error can be proven. The more complex the
set of revealed entries according to a certain measure, the better the bound on
the generalization error.
****

A Deep Architecture for Matching Short Texts
Zhengdong Lu, Hang Li
Many machine learning problems can be interpreted as learning for matching two t
ypes of objects (e.g., images and captions, users and products, queries and docu
ments). The matching level of two objects is usually measured as the inner produ
ct in a certain feature space, while the modeling effort focuses on mapping of o
bjects from the original space to the feature space. This schema, although prove
n successful on a range of matching tasks, is insufficient for capturing the ric
h structure in the matching process of more complicated objects. In this paper,
 we propose a new deep architecture to more effectively model the complicated ma
tching relations between two objects from heterogeneous domains. More specifical
ly, we apply this model to matching tasks in natural language, e.g., finding sen
sible responses for a tweet, or relevant answers to a given question. This new
architecture naturally combines the localness and hierarchy intrinsic to the nat
ural language problems, and therefore greatly improves upon the state-of-the-art
 models.
****

Sensor Selection in High-Dimensional Gaussian Trees with Nuisances
Daniel S. Levine, Jonathan P. How
We consider the sensor selection problem on multivariate Gaussian distributions
where only a \emph{subset} of latent variables is of inferential interest. For
pairs of vertices connected by a unique path in the graph, we show that there ex
ist decompositions of nonlocal mutual information into local information measure
s that can be computed efficiently from the output of message passing algorithms
. We integrate these decompositions into a computationally efficient greedy sel
ector where the computational expense of quantification can be distributed acros
s nodes in the network. Experimental results demonstrate the comparative effici
ency of our algorithms for sensor selection in high-dimensional distributions. W
e additionally derive an online-computable performance bound based on augmentati
ons of the relevant latent variable set that, when such a valid augmentation exi
sts, is applicable for \emph{any} distribution with nuisances.
****

The Total Variation on Hypergraphs - Learning on Hypergraphs Revisited
Matthias Hein, Simon Setzer, Leonardo Jost, Syama Sundar Rangapuram
Hypergraphs allow to encode higher-order relationships in data and are thus a ve
ry flexible modeling tool. Current learning methods are either based on approxim
ations of the hypergraphs via graphs or on tensor methods which are only applic
able under special conditions. In this paper we present a new learning framework
 on hypergraphs which fully uses the hypergraph structure. The key element is a
 family of regularization functionals based on the total variation on hypergrap
hs.
****

Lasso Screening Rules via Dual Polytope Projection
Jie Wang, Jiayu Zhou, Peter Wonka, Jieping Ye
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****

Multiscale Dictionary Learning for Estimating Conditional Distributions
Francesca Petralia, Joshua T. Vogelstein, David B. Dunson
Nonparametric estimation of the conditional distribution of a response given high-dimensional features is a challenging problem. It is important to allow not only the mean but also the variance and shape of the response density to change flexibly with features, which are massive-dimensional. We propose a multiscale dictionary learning model, which expresses the conditional response density as a convex combination of dictionary densities, with the densities used and their weights dependent on the path through a tree decomposition of the feature space. A fast graph partitioning algorithm is applied to obtain the tree decomposition, with Bayesian methods then used to adaptively prune and average over different sub-trees in a soft probabilistic manner. The algorithm scales efficiently to approximately one million features. State of the art predictive performance is demonstrated for toy examples and two neuroscience applications including up to a million features.
****

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

Online Learning of Nonparametric Mixture Models via Sequential Variational Approximation
Dahua Lin
Reliance on computationally expensive algorithms for inference has been limiting the use of Bayesian nonparametric models in large scale applications. To tackle this problem, we propose a Bayesian learning algorithm for DP mixture models. Instead of following the conventional paradigm -- random initialization plus iterative update, we take an progressive approach. Starting with a given prior, our method recursively transforms it into an approximate posterior through sequential variational approximation. In this process, new components will be incorporated on the fly when needed. The algorithm can reliably estimate a DP mixture model in one pass, making it particularly suited for applications with massive data. Experiments on both synthetic data and real datasets demonstrate remarkable improvement on efficiency -- orders of magnitude speed-up compared to the state-of-the-art.
****

Exact and Stable Recovery of Pairwise Interaction Tensors
Shouyuan Chen, Michael R. Lyu, Irwin King, Zenglin Xu
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****

A* Lasso for Learning a Sparse Bayesian Network Structure for Continuous Variables
Jing Xiang, Seyoung Kim
We address the problem of learning a sparse Bayesian network structure for continuous variables in a high-dimensional space. The constraint that the estimated Bayesian network structure must be a directed acyclic graph (DAG) makes the problem challenging because of the huge search space of network structures. Most pr

evious methods were based on a two-stage approach that prunes the search space i
n the first stage and then searches for a network structure that satisfies the D
AG constraint in the second stage. Although this approach is effective in a low-
dimensional setting, it is difficult to ensure that the correct network structur
e is not pruned in the first stage in a high-dimensional setting. In this paper
, we propose a single-stage method, called A* lasso, that recovers the optimal
sparse Bayesian network structure by solving a single optimization problem with
A* search algorithm that uses lasso in its scoring system. Our approach substant
ially improves the computational efficiency of the well-known exact methods bas
ed on dynamic programming. We also present a heuristic scheme that further impro
ves the efficiency of A* lasso without significantly compromising the quality of
 solutions and demonstrate this on benchmark Bayesian networks and real data.
****

High-Dimensional Gaussian Process Bandits
Josip Djolonga, Andreas Krause, Volkan Cevher
Many applications in machine learning require optimizing unknown functions defin
ed over a high-dimensional space from noisy samples that are expensive to obtain
. We address this notoriously hard challenge, under the assumptions that the fun
ction varies only along some low-dimensional subspace and is smooth (i.e., it ha
s a low norm in a Reproducible Kernel Hilbert Space). In particular, we present
the SI-BO algorithm, which leverages recent low-rank matrix recovery techniques
to learn the underlying subspace of the unknown function and applies Gaussian Pr
ocess Upper Confidence sampling for optimization of the function. We carefully c
alibrate the exploration–exploitation tradeoff by allocating sampling budget to
subspace estimation and function optimization, and obtain the first subexponenti
al cumulative regret bounds and convergence rates for Bayesian optimization in h
igh-dimensions under noisy observations. Numerical results demonstrate the effec
tiveness of our approach in difficult scenarios.
****

Generalizing Analytic Shrinkage for Arbitrary Covariance Structures
Daniel Bartz, Klaus-Robert Müller
Analytic shrinkage is a statistical technique that offers a fast alternative to
cross-validation for the regularization of covariance matrices and has appealing
 consistency properties. We show that the proof of consistency implies bounds on
 the growth rates of eigenvalues and their dispersion, which are often violated
in data. We prove consistency under assumptions which do not restrict the covari
ance structure and therefore better match real world data. In addition, we propo
se an extension of analytic shrinkage --orthogonal complement shrinkage-- which
adapts to the covariance structure. Finally we demonstrate the superior performa
nce of our novel approach on data from the domains of finance, spoken letter and
 optical character recognition, and neuroscience.
****

Higher Order Priors for Joint Intrinsic Image, Objects, and Attributes Estimatio
n
Vibhav Vineet, Carsten Rother, Philip Torr
Many methods have been proposed to recover the intrinsic scene properties such a
s shape, reflectance and illumination from a single image. However, most of thes
e models have been applied on laboratory datasets. In this work we explore the s
ynergy effects between intrinsic scene properties recovered from an image, and t
he objects and attributes present in the scene. We cast the problem in a joint e
nergy minimization framework; thus our model is able to encode the strong correl
ations between intrinsic properties (reflectance, shape, illumination), objects
(table, tv-monitor), and materials (wooden, plastic) in a given scene. We tested
 our approach on the NYU and Pascal datasets, and observe both qualitative and q
uantitative improvements in the overall accuracy.
****

Online Robust PCA via Stochastic Optimization
Jiashi Feng, Huan Xu, Shuicheng Yan
Robust PCA methods are typically based on batch optimization and have to load al
l the samples into memory. This prevents them from efficiently processing big da

ta. In this paper, we develop an Online Robust Principal Component Analysis (OR
-PCA) that processes one sample per time instance and hence its memory cost is i
ndependent of the data size, significantly enhancing the computation and storag
e efficiency. The proposed method is based on stochastic optimization of an equi
valent reformulation of the batch RPCA method. Indeed, we show that OR-PCA provi
des a sequence of subspace estimations converging to the optimum of its batch co
unterpart and hence is provably robust to sparse corruption. Moreover, OR-PCA c
an naturally be applied for tracking dynamic subspace. Comprehensive simulations
 on subspace recovering and tracking demonstrate the robustness and efficiency a
dvantages of the OR-PCA over online PCA and batch RPCA methods.
****

Compete to Compute
Rupesh K. Srivastava, Jonathan Masci, Sohrob Kazerounian, Faustino Gomez, Jürgen
 Schmidhuber
Local competition among neighboring neurons is common in biological neural netwo
rks (NNs). We apply the concept to gradient-based, backprop-trained artificial m
ultilayer NNs. NNs with competing linear units tend to outperform those with non
-competing nonlinear units, and avoid catastrophic forgetting when training sets
 change over time.
****

Heterogeneous-Neighborhood-based Multi-Task Local Learning Algorithms
Yu Zhang
All the existing multi-task local learning methods are defined on homogeneous ne
ighborhood which consists of all data points from only one task. In this paper,
different from existing methods, we propose local learning methods for multi-tas
k classification and regression problems based on heterogeneous neighborhood whi
ch is defined on data points from all tasks. Specifically, we extend the k-neare
st-neighbor classifier by formulating the decision function for each data point
as a weighted voting among the neighbors from all tasks where the weights are ta
sk-specific. By defining a regularizer to enforce the task-specific weight matri
x to approach a symmetric one, a regularized objective function is proposed and
an efficient coordinate descent method is developed to solve it. For regression
problems, we extend the kernel regression to multi-task setting in a similar way
 to the classification case. Experiments on some toy data and real-world dataset
s demonstrate the effectiveness of our proposed methods.
****

Scalable Influence Estimation in Continuous-Time Diffusion Networks
Nan Du, Le Song, Manuel Gomez Rodriguez, Hongyuan Zha
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.
****

More data speeds up training time in learning halfspaces over sparse vectors
Amit Daniely, Nati Linial, Shai Shalev-Shwartz
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.
****

Top-Down Regularization of Deep Belief Networks
Hanlin Goh, Nicolas Thome, Matthieu Cord, Joo-Hwee Lim
Designing a principled and effective algorithm for learning deep architectures i
s a challenging problem. The current approach involves two training phases: a fu
lly unsupervised learning followed by a strongly discriminative optimization. We
 suggest a deep learning strategy that bridges the gap between the two phases, r
esulting in a three-phase learning procedure. We propose to implement the scheme
 using a method to regularize deep belief networks with top-down information. Th
e network is constructed from building blocks of restricted Boltzmann machines l
earned by combining bottom-up and top-down sampled signals. A global optimizatio

n procedure that merges samples from a forward bottom-up pass and a top-down pas s is used. Experiments on the MNIST dataset show improvements over the existing algorithms for deep belief networks. Object recognition results on the Caltech-1 01 dataset also yield competitive results.
****
Polar Operators for Structured Sparse Estimation
Xinhua Zhang, Yao-Liang Yu, Dale Schuurmans
Structured sparse estimation has become an important technique in many areas of data analysis. Unfortunately, these estimators normally create computational dif ficulties that entail sophisticated algorithms. Our first contribution is to unc over a rich class of structured sparse regularizers whose polar operator can be evaluated efficiently. With such an operator, a simple conditional gradient meth od can then be developed that, when combined with smoothing and local optimizati on, significantly reduces training time vs. the state of the art. We also demons trate a new reduction of polar to proximal maps that enables more efficient late nt fused lasso.
****
Learning with Invariance via Linear Functionals on Reproducing Kernel Hilbert Sp ace
Xinhua Zhang, Wee Sun Lee, Yee Whye Teh
Incorporating invariance information is important for many learning problems. To exploit invariances, most existing methods resort to approximations that either lead to expensive optimization problems such as semi-definite programming, or r ely on separation oracles to retain tractability. Some methods further limit th e space of functions and settle for non-convex models. In this paper, we propos e a framework for learning in reproducing kernel Hilbert spaces (RKHS) using loc al invariances that explicitly characterize the behavior of the target function around data instances. These invariances are \emph{compactly} encoded as linear functionals whose value are penalized by some loss function. Based on a repres enter theorem that we establish, our formulation can be efficiently optimized vi a a convex program. For the representer theorem to hold, the linear functionals are required to be bounded in the RKHS, and we show that this is true for a vari ety of commonly used RKHS and invariances. Experiments on learning with unlabele d data and transform invariances show that the proposed method yields better or similar results compared with the state of the art.
****
Real-Time Inference for a Gamma Process Model of Neural Spiking
David E. Carlson, Vinayak Rao, Joshua T. Vogelstein, Lawrence Carin
With simultaneous measurements from ever increasing populations of neurons, ther e is a growing need for sophisticated tools to recover signals from individual n eurons. In electrophysiology experiments, this classically proceeds in a two-ste p process: (i) threshold the waveforms to detect putative spikes and (ii) cluste r the waveforms into single units (neurons). We extend previous Bayesian nonpara met- ric models of neural spiking to jointly detect and cluster neurons using a Gamma process model. Importantly, we develop an online approximate inference sch eme enabling real-time analysis, with performance exceeding the previous state-o f-the- art. Via exploratory data analysis—using data with partial ground truth a s well as two novel data sets—we find several features of our model collectively contribute to our improved performance including: (i) accounting for colored no ise, (ii) de- tecting overlapping spikes, (iii) tracking waveform dynamics, and (iv) using mul- tiple channels. We hope to enable novel experiments simultaneous ly measuring many thousands of neurons and possibly adapting stimuli dynamically to probe ever deeper into the mysteries of the brain.
****
Direct 0-1 Loss Minimization and Margin Maximization with Boosting
Shaodan Zhai, Tian Xia, Ming Tan, Shaojun Wang
We propose a boosting method, DirectBoost, a greedy coordinate descent algorithm that builds an ensemble classifier of weak classifiers through directly minimiz ing empirical classification error over labeled training examples; once the tra ining classification error is reduced to a local coordinatewise minimum, DirectB

oost runs a greedy coordinate ascent algorithm that continuously adds weak class
ifiers to maximize any targeted arbitrarily defined margins until reaching a loc
al coordinatewise maximum of the margins in a certain sense. Experimental result
s on a collection of machine-learning benchmark datasets show that DirectBoost g
ives consistently better results than AdaBoost, LogitBoost, LPBoost with column
generation and BrownBoost, and is noise tolerant when it maximizes an n'th order
 bottom sample margin.
****

Marginals-to-Models Reducibility
Tim Roughgarden, Michael Kearns
We consider a number of classical and new computational problems regarding margi
nal distributions, and inference in models specifying a full joint distribution.
 We prove general and efficient reductions between a number of these problems, w
hich demonstrate that algorithmic progress in inference automatically yields pro
gress for “pure data” problems. Our main technique involves formulating the prob
lems as linear programs, and proving that the dual separation oracle for the Ell
ipsoid Method is provided by the target problem. This technique may be of indepe
ndent interest in probabilistic inference.
****

Sketching Structured Matrices for Faster Nonlinear Regression
Haim Avron, Vikas Sindhwani, 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.
****

Variance Reduction for Stochastic Gradient Optimization
Chong Wang, Xi Chen, Alexander J. Smola, Eric P. Xing
Stochastic gradient optimization is a class of widely used algorithms for traini
ng machine learning models. To optimize an objective, it uses the noisy gradient
 computed from the random data samples instead of the true gradient computed fro
m the entire dataset. However, when the variance of the noisy gradient is large,
 the algorithm might spend much time bouncing around, leading to slower converge
nce and worse performance. In this paper, we develop a general approach of using
 control variate for variance reduction in stochastic gradient. Data statistics
such as low-order moments (pre-computed or estimated online) is used to form the
 control variate. We demonstrate how to construct the control variate for two pr
actical problems using stochastic gradient optimization. One is convex---the MAP
 estimation for logistic regression, and the other is non-convex---stochastic va
riational inference for latent Dirichlet allocation. On both problems, our appro
ach shows faster convergence and better performance than the classical approach.
****

On Decomposing the Proximal Map
Yao-Liang Yu
The proximal map is the key step in gradient-type algorithms, which have become
prevalent in large-scale high-dimensional problems. For simple functions this pr
oximal map is available in closed-form while for more complicated functions it c
an become highly nontrivial. Motivated by the need of combining regularizers to
simultaneously induce different types of structures, this paper initiates a syst
ematic investigation of when the proximal map of a sum of functions decomposes i
nto the composition of the proximal maps of the individual summands. We not only
 unify a few known results scattered in the literature but also discover several
 new decompositions obtained almost effortlessly from our theory.
****

Documents as multiple overlapping windows into grids of counts
Alessandro Perina, Nebojsa Jojic, Manuele Bicego, Andrzej Truski
In text analysis documents are represented as disorganized bags of words, models
 of count features are typically based on mixing a small number of topics \cite{
lda,sam}. Recently, it has been observed that for many text corpora documents ev
olve into one another in a smooth way, with some features dropping and new ones

being introduced. The counting grid \cite{cgUai} models this spatial metaphor literally: it is multidimensional grid of word distributions learned in such a way that a document's own distribution of features can be modeled as the sum of the histograms found in a window into the grid. The major drawback of this method is that it is essentially a mixture and all the content much be generated by a single contiguous area on the grid. This may be problematic especially for lower dimensional grids. In this paper, we overcome to this issue with the \emph{Componential Counting Grid} which brings the componential nature of topic models to the basic counting grid. We also introduce a generative kernel based on the document's grid usage and a visualization strategy useful for understanding large text corpora. We evaluate our approach on document classification and multimodal retrieval obtaining state of the art results on standard benchmarks.
****

Robust Sparse Principal Component Regression under the High Dimensional Elliptical Model
Fang Han, 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.
****

Optimizing Instructional Policies
Robert V. Lindsey, Michael C. Mozer, William J. Huggins, Harold Pashler
Psychologists are interested in developing instructional policies that boost student learning. An instructional policy specifies the manner and content of instruction. For example, in the domain of concept learning, a policy might specify the nature of exemplars chosen over a training sequence. Traditional psychological studies compare several hand-selected policies, e.g., contrasting a policy that selects only difficult-to-classify exemplars with a policy that gradually progresses over the training sequence from easy exemplars to more difficult (known as {\em fading}). We propose an alternative to the traditional methodology in which we define a parameterized space of policies and search this space to identify the optimum policy. For example, in concept learning, policies might be described by a fading function that specifies exemplar difficulty over time. We propose an experimental technique for searching policy spaces using Gaussian process surrogate-based optimization and a generative model of student performance. Instead of evaluating a few experimental conditions each with many human subjects, as the traditional methodology does, our technique evaluates many experimental conditions each with a few subjects. Even though individual subjects provide only a noisy estimate of the population mean, the optimization method allows us to determine the shape of the policy space and identify the global optimum, and is as efficient in its subject budget as a traditional A-B comparison. We evaluate the method via two behavioral studies, and suggest that the method has broad applicability to optimization problems involving humans in domains beyond the educational arena.
****

Adaptive Market Making via Online Learning
Jacob Abernethy, Satyen Kale
We consider the design of strategies for \emph{market making} in a market like a stock, commodity, or currency exchange. In order to obtain profit guarantees for a market maker one typically requires very particular stochastic assumptions on the sequence of price fluctuations of the asset in question. We propose a class of spread-based market making strategies whose performance can be controlled even under worst-case (adversarial) settings. We prove structural properties of these strategies which allows us to design a master algorithm which obtains low regret relative to the best such strategy in hindsight. We run a set of experiments showing favorable performance on real-world price data.
****

Learning Prices for Repeated Auctions with Strategic Buyers
Kareem Amin, Afshin Rostamizadeh, Umar Syed

Inspired by real-time ad exchanges for online display advertising, we consider t
he problem of inferring a buyer's value distribution for a good when the buyer i
s repeatedly interacting with a seller through a posted-price mechanism. We mod
el the buyer as a strategic agent, whose goal is to maximize her long-term surpl
us, and we are interested in mechanisms that maximize the seller's long-term rev
enue. We present seller algorithms that are no-regret when the buyer discounts h
er future surplus --- i.e. the buyer prefers showing advertisements to users soo
ner rather than later. We also give a lower bound on regret that increases as th
e buyer's discounting weakens and shows, in particular, that any seller algorith
m will suffer linear regret if there is no discounting.
****

Multilinear Dynamical Systems for Tensor Time Series
Mark Rogers, Lei Li, Stuart J. Russell
Many scientific data occur as sequences of multidimensional arrays called tensor
s. How can hidden, evolving trends in such data be extracted while preserving t
he tensor structure? The model that is traditionally used is the linear dynamic
al system (LDS), which treats the observation at each time slice as a vector. I
n this paper, we propose the multilinear dynamical system (MLDS) for modeling te
nsor time series and an expectation-maximization (EM) algorithm to estimate the
parameters. The MLDS models each time slice of the tensor time series as the mu
ltilinear projection of a corresponding member of a sequence of latent, low-dime
nsional tensors. Compared to the LDS with an equal number of parameters, the ML
DS achieves higher prediction accuracy and marginal likelihood for both simulate
d and real datasets.
****

Computing the Stationary Distribution Locally
Christina E. Lee, Asuman Ozdaglar, 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.
****

Latent Maximum Margin Clustering
Guang-Tong Zhou, Tian Lan, Arash Vahdat, Greg Mori
We present a maximum margin framework that clusters data using latent variables.
Using latent representations enables our framework to model unobserved informat
ion embedded in the data. We implement our idea by large margin learning, and de
velop an alternating descent algorithm to effectively solve the resultant non-co
nvex optimization problem. We instantiate our latent maximum margin clustering f
ramework with tag-based video clustering tasks, where each video is represented
by a latent tag model describing the presence or absence of video tags. Experime
ntal results obtained on three standard datasets show that the proposed method o
utperforms non-latent maximum margin clustering as well as conventional clusteri
ng approaches.
****

Hierarchical Modular Optimization of Convolutional Networks Achieves Representat
ions Similar to Macaque IT and Human Ventral Stream
Daniel L. Yamins, Ha Hong, Charles Cadieu, James J. DiCarlo
Humans recognize visually-presented objects rapidly and accurately. To understan
d this ability, we seek to construct models of the ventral stream, the series of
cortical areas thought to subserve object recognition. One tool to assess the q
uality of a model of the ventral stream is the Representation Dissimilarity Matr
ix (RDM), which uses a set of visual stimuli and measures the distances produced
in either the brain (i.e. fMRI voxel responses, neural firing rates) or in mode
ls (features). Previous work has shown that all known models of the ventral stre
am fail to capture the RDM pattern observed in either IT cortex, the highest ven
tral area, or in the human ventral stream. In this work, we construct models of
the ventral stream using a novel optimization procedure for category-level objec
t recognition problems, and produce RDMs resembling both macaque IT and human ve
ntral stream. The model, while novel in the optimization procedure, further deve

lops a long-standing functional hypothesis that the ventral visual stream is a h
ierarchically arranged series of processing stages optimized for visual object r
ecognition.
****

Online PCA for Contaminated Data
Jiashi Feng, Huan Xu, Shie Mannor, Shuicheng Yan
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.
****

Distributed Representations of Words and Phrases and their Compositionality
Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, Jeff Dean
The recently introduced continuous Skip-gram model is an efficient method for le
arning high-quality distributed vector representations that capture a large numb
er of precise syntactic and semantic word relationships. In this paper we prese
nt several improvements that make the Skip-gram model more expressive and enable
 it to learn higher quality vectors more rapidly. We show that by subsampling f
requent words we obtain significant speedup, and also learn higher quality repr
esentations as measured by our tasks. We also introduce Negative Sampling, a sim
plified variant of Noise Contrastive Estimation (NCE) that learns more accurate
vectors for frequent words compared to the hierarchical softmax. An inherent l
imitation of word representations is their indifference to word order and their
inability to represent idiomatic phrases. For example, the meanings of Canada''
 and "Air'' cannot be easily combined to obtain "Air Canada''. Motivated by thi
s example, we present a simple and efficient method for finding phrases, and sho
w that their vector representations can be accurately learned by the Skip-gram m
odel. "
****

Learning Multiple Models via Regularized Weighting
Daniel Vainsencher, Shie Mannor, Huan Xu
We consider the general problem of Multiple Model Learning (MML) from data, from
 the statistical and algorithmic perspectives; this problem includes clustering,
 multiple regression and subspace clustering as special cases. A common approach
 to solving new MML problems is to generalize Lloyd's algorithm for clustering
(or Expectation-Maximization for soft clustering). However this approach is unfo
rtunately sensitive to outliers and large noise: a single exceptional point may
take over one of the models. We propose a different general formulation that s
eeks for each model a distribution over data points; the weights are regularized
 to be sufficiently spread out. This enhances robustness by making assumptions o
n class balance. We further provide generalization bounds and explain how the ne
w iterations may be computed efficiently. We demonstrate the robustness benefits
 of our approach with some experimental results and prove for the important cas
e of clustering that our approach has a non-trivial breakdown point, i.e., is gu
aranteed to be robust to a fixed percentage of adversarial unbounded outliers.
****

Discriminative Transfer Learning with Tree-based Priors
Nitish Srivastava, Russ R. Salakhutdinov
This paper proposes a way of improving classification performance for classes wh
ich have very few training examples. The key idea is to discover classes which a
re similar and transfer knowledge among them. Our method organizes the classes i
nto a tree hierarchy. The tree structure can be used to impose a generative prio
r over classification parameters. We show that these priors can be combined with
 discriminative models such as deep neural networks. Our method benefits from th
e power of discriminative training of deep neural networks, at the same time usi
ng tree-based generative priors over classification parameters. We also propose
an algorithm for learning the underlying tree structure. This gives the model so
me flexibility to tune the tree so that the tree is pertinent to task being solv
ed. We show that the model can transfer knowledge across related classes using f
ixed semantic trees. Moreover, it can learn new meaningful trees usually leading

to improved performance. Our method achieves state-of-the-art classification results on the CIFAR-100 image data set and the MIR Flickr multimodal data set.
****

Machine Teaching for Bayesian Learners in the Exponential Family
Jerry Zhu
What if there is a teacher who knows the learning goal and wants to design good training data for a machine learner? We propose an optimal teaching framework aimed at learners who employ Bayesian models. Our framework is expressed as an optimization problem over teaching examples that balance the future loss of the learner and the effort of the teacher. This optimization problem is in general hard. In the case where the learner employs conjugate exponential family models, we present an approximate algorithm for finding the optimal teaching set. Our algorithm optimizes the aggregate sufficient statistics, then unpacks them into actual teaching examples. We give several examples to illustrate our framework.
****

Online Learning with Switching Costs and Other Adaptive Adversaries
Nicolò Cesa-Bianchi, Ofer Dekel, Ohad Shamir
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****

From Bandits to Experts: A Tale of Domination and Independence
Noga Alon, Nicolò Cesa-Bianchi, Claudio Gentile, Yishay Mansour
We consider the partial observability model for multi-armed bandits, introduced by Mannor and Shamir (2011). Our main result is a characterization of regret in the directed observability model in terms of the dominating and independence numbers of the observability graph. We also show that in the undirected case, the learner can achieve optimal regret without even accessing the observability graph before selecting an action. Both results are shown using variants of the Exp3 algorithm operating on the observability graph in a time-efficient manner.
****

Which Space Partitioning Tree to Use for Search?
Parikshit Ram, Alexander Gray
We consider the task of nearest-neighbor search with the class of binary-space-partitioning trees, which includes kd-trees, principal axis trees and random projection trees, and try to rigorously answer the question which tree to use for nearest-neighbor search?'' To this end, we present the theoretical results which imply that trees with better vector quantization performance have better search performance guarantees. We also explore another factor affecting the search performance -- margins of the partitions in these trees. We demonstrate, both theoretically and empirically, that large margin partitions can improve the search performance of a space-partitioning tree. "
****

Small-Variance Asymptotics for Hidden Markov Models
Anirban Roychowdhury, Ke Jiang, Brian Kulis
Small-variance asymptotics provide an emerging technique for obtaining scalable combinatorial algorithms from rich probabilistic models. We present a small-variance asymptotic analysis of the Hidden Markov Model and its infinite-state Bayesian nonparametric extension. Starting with the standard HMM, we first derive a “hard” inference algorithm analogous to k-means that arises when particular variances in the model tend to zero. This analysis is then extended to the Bayesian nonparametric case, yielding a simple, scalable, and flexible algorithm for discrete-state sequence data with a non-fixed number of states. We also derive the corresponding combinatorial objective functions arising from our analysis, which involve a k-means-like term along with penalties based on state transitions and the number of states. A key property of such algorithms is that — particularly in the nonparametric setting — standard probabilistic inference algorithms lack scalability and are heavily dependent on good initialization. A number of results on synthetic and real data sets demonstrate the advantages of the proposed frame

work.
****
Sparse Overlapping Sets Lasso for Multitask Learning and its Application to fMRI Analysis
Nikhil Rao, Christopher Cox, Rob Nowak, Timothy T. Rogers
Multitask learning can be effective when features useful in one task are also useful for other tasks, and the group lasso is a standard method for selecting a common subset of features. In this paper, we are interested in a less restrictive form of multitask learning, wherein (1) the available features can be organized into subsets according to a notion of similarity and (2) features useful in one task are similar, but not necessarily identical, to the features best suited for other tasks. The main contribution of this paper is a new procedure called {\em Sparse Overlapping Sets (SOS) lasso}, a convex optimization that automatically selects similar features for related learning tasks. Error bounds are derived for SOSlasso and its consistency is established for squared error loss. In particular, SOSlasso is motivated by multi-subject fMRI studies in which functional activity is classified using brain voxels as features. Experiments with real and synthetic data demonstrate the advantages of SOSlasso compared to the lasso and group lasso.
****
Submodular Optimization with Submodular Cover and Submodular Knapsack Constraints
Rishabh K. Iyer, Jeff A. Bilmes
We investigate two new optimization problems — minimizing a submodular function subject to a submodular lower bound constraint (submodular cover) and maximizing a submodular function subject to a submodular upper bound constraint (submodular knapsack). We are motivated by a number of real-world applications in machine learning including sensor placement and data subset selection, which require maximizing a certain submodular function (like coverage or diversity) while simultaneously minimizing another (like cooperative cost). These problems are often posed as minimizing the difference between submodular functions [9, 23] which is in the worst case inapproximable. We show, however, that by phrasing these problems as constrained optimization, which is more natural for many applications, we achieve a number of bounded approximation guarantees. We also show that both these problems are closely related and, an approximation algorithm solving one can be used to obtain an approximation guarantee for the other. We provide hardness results for both problems thus showing that our approximation factors are tight up to log-factors. Finally, we empirically demonstrate the performance and good scalability properties of our algorithms.
****
Model Selection for High-Dimensional Regression under the Generalized Irrepresentability Condition
Adel Javanmard, 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.
****
Scalable kernels for graphs with continuous attributes
Aasa Feragen, Niklas Kasenburg, Jens Petersen, Marleen de Bruijne, Karsten Borgwardt
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Bayesian optimization explains human active search
Ali Borji, Laurent Itti
Many real-world problems have complicated objective functions. To optimize such functions, humans utilize sophisticated sequential decision-making strategies. M

any optimization algorithms have also been developed for this same purpose, but how do they compare to humans in terms of both performance and behavior? We try to unravel the general underlying algorithm people may be using while searching for the maximum of an invisible 1D function. Subjects click on a blank screen an d are shown the ordinate of the function at each clicked abscissa location. Thei r task is to find the function’s maximum in as few clicks as possible. Subjects win if they get close enough to the maximum location. Analysis over 23 non-maths undergraduates, optimizing 25 functions from different families, shows that hum ans outperform 24 well-known optimization algorithms. Bayesian Optimization base d on Gaussian Processes, which exploit all the x values tried and all the f(x) v alues obtained so far to pick the next x, predicts human performance and searche d locations better. In 6 follow-up controlled experiments over 76 subjects, cove ring interpolation, extrapolation, and optimization tasks, we further confirm th at Gaussian Processes provide a general and unified theoretical account to expla in passive and active function learning and search in humans.
****

B-test: A Non-parametric, Low Variance Kernel Two-sample Test
Wojciech Zaremba, Arthur Gretton, Matthew Blaschko
We propose a family of maximum mean discrepancy (MMD) kernel two-sample tests th at have low sample complexity and are consistent. The test has a hyperparameter that allows one to control the tradeoff between sample complexity and computatio nal time. Our family of tests, which we denote as B-tests, is both computational ly and statistically efficient, combining favorable properties of previously pro posed MMD two-sample tests. It does so by better leveraging samples to produce low variance estimates in the finite sample case, while avoiding a quadratic num ber of kernel evaluations and complex null-hypothesis approximation as would be required by tests relying on one sample U-statistics. The B-test uses a smaller than quadratic number of kernel evaluations and avoids completely the computatio nal burden of complex null-hypothesis approximation while maintaining consistenc y and probabilistically conservative thresholds on Type I error. Finally, recent results of combining multiple kernels transfer seamlessly to our hypothesis tes t, allowing a further increase in discriminative power and decrease in sample co mplexity.
****

Moment-based Uniform Deviation Bounds for k-means and Friends
Matus J. Telgarsky, Sanjoy Dasgupta
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.
****

Convex Calibrated Surrogates for Low-Rank Loss Matrices with Applications to Sub set Ranking Losses
Harish G. Ramaswamy, Shivani Agarwal, Ambuj Tewari
The design of convex, calibrated surrogate losses, whose minimization entails co nsistency with respect to a desired target loss, is an important concept to have emerged in the theory of machine learning in recent years. We give an explicit construction of a convex least-squares type surrogate loss that can be designed to be calibrated for any multiclass learning problem for which the target loss m atrix has a low-rank structure; the surrogate loss operates on a surrogate targe t space of dimension at most the rank of the target loss. We use this result to design convex calibrated surrogates for a variety of subset ranking problems, wi th target losses including the precision@q, expected rank utility, mean average precision, and pairwise disagreement.
****

Auxiliary-variable Exact Hamiltonian Monte Carlo Samplers for Binary Distributio ns
Ari Pakman, Liam Paninski
We present a new approach to sample from generic binary distributions, based on an exact Hamiltonian Monte Carlo algorithm applied to a piecewise continuous au

gmentation of the binary distribution of interest. An extension of this idea to distributions over mixtures of binary and continuous variables allows us to sample from posteriors of linear and probit regression models with spike-and-slab priors and truncated parameters. We illustrate the advantages of these algorithms in several examples in which they outperform the Metropolis or Gibbs samplers.

****

Spectral methods for neural characterization using generalized quadratic models
Il Memming Park, Evan W. Archer, Nicholas Priebe, Jonathan W. Pillow

We describe a set of fast, tractable methods for characterizing neural responses to high-dimensional sensory stimuli using a model we refer to as the generalized quadratic model (GQM). The GQM consists of a low-rank quadratic form followed by a point nonlinearity and exponential-family noise. The quadratic form characterizes the neuron's stimulus selectivity in terms of a set linear receptive fields followed by a quadratic combination rule, and the invertible nonlinearity maps this output to the desired response range. Special cases of the GQM include the 2nd-order Volterra model (Marmarelis and Marmarelis 1978, Koh and Powers 1985) and the elliptical Linear-Nonlinear-Poisson model (Park and Pillow 2011). Here we show that for canonical form" GQMs, spectral decomposition of the first two response-weighted moments yields approximate maximum-likelihood estimators via a quantity called the expected log-likelihood. The resulting theory generalizes moment-based estimators such as the spike-triggered covariance, and, in the Gaussian noise case, provides closed-form estimators under a large class of non-Gaussian stimulus distributions. We show that these estimators are fast and provide highly accurate estimates with far lower computational cost than full maximum likelihood. Moreover, the GQM provides a natural framework for combining multi-dimensional stimulus sensitivity and spike-history dependencies within a single model. We show applications to both analog and spiking data using intracellular recordings of V1 membrane potential and extracellular recordings of retinal spike trains."

****

A Latent Source Model for Nonparametric Time Series Classification
George H. Chen, Stanislav Nikolov, Devavrat Shah

For classifying time series, a nearest-neighbor approach is widely used in practice with performance often competitive with or better than more elaborate methods such as neural networks, decision trees, and support vector machines. We develop theoretical justification for the effectiveness of nearest-neighbor-like classification of time series. Our guiding hypothesis is that in many applications, such as forecasting which topics will become trends on Twitter, there aren't actually that many prototypical time series to begin with, relative to the number of time series we have access to, e.g., topics become trends on Twitter only in a few distinct manners whereas we can collect massive amounts of Twitter data. To operationalize this hypothesis, we propose a latent source model for time series, which naturally leads to a weighted majority voting" classification rule that can be approximated by a nearest-neighbor classifier. We establish nonasymptotic performance guarantees of both weighted majority voting and nearest-neighbor classification under our model accounting for how much of the time series we observe and the model complexity. Experimental results on synthetic data show weighted majority voting achieving the same misclassification rate as nearest-neighbor classification while observing less of the time series. We then use weighted majority to forecast which news topics on Twitter become trends, where we are able to detect such "trending topics" in advance of Twitter 79% of the time, with a mean early advantage of 1 hour and 26 minutes, a true positive rate of 95%, and a false positive rate of 4%."

****

PAC-Bayes-Empirical-Bernstein Inequality
Ilya O. Tolstikhin, Yevgeny Seldin

We present PAC-Bayes-Empirical-Bernstein inequality. The inequality is based on combination of PAC-Bayesian bounding technique with Empirical Bernstein bound. It allows to take advantage of small empirical variance and is especially useful

in regression. We show that when the empirical variance is significantly smaller than the empirical loss PAC-Bayes-Empirical-Bernstein inequality is significantly tighter than PAC-Bayes-kl inequality of Seeger (2002) and otherwise it is comparable. PAC-Bayes-Empirical-Bernstein inequality is an interesting example of application of PAC-Bayesian bounding technique to self-bounding functions. We provide empirical comparison of PAC-Bayes-Empirical-Bernstein inequality with PAC-Bayes-kl inequality on a synthetic example and several UCI datasets.
****

Convex Two-Layer Modeling
Özlem Aslan, Hao Cheng, Xinhua Zhang, Dale Schuurmans
Latent variable prediction models, such as multi-layer networks, impose auxiliary latent variables between inputs and outputs to allow automatic inference of implicit features useful for prediction. Unfortunately, such models are difficult to train because inference over latent variables must be performed concurrently with parameter optimization---creating a highly non-convex problem. Instead of proposing another local training method, we develop a convex relaxation of hidden-layer conditional models that admits global training. Our approach extends current convex modeling approaches to handle two nested nonlinearities separated by a non-trivial adaptive latent layer. The resulting methods are able to acquire two-layer models that cannot be represented by any single-layer model over the same features, while improving training quality over local heuristics.
****

The Randomized Dependence Coefficient
David Lopez-Paz, Philipp Hennig, Bernhard Schölkopf
We introduce the Randomized Dependence Coefficient (RDC), a measure of non-linear dependence between random variables of arbitrary dimension based on the Hirschfeld-Gebelein-Rényi Maximum Correlation Coefficient. RDC is defined in terms of correlation of random non-linear copula projections; it is invariant with respect to marginal distribution transformations, has low computational cost and is easy to implement: just five lines of R code, included at the end of the paper.
****

Sparse Inverse Covariance Estimation with Calibration
Tuo Zhao, 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.
****

Thompson Sampling for 1-Dimensional Exponential Family Bandits
Nathaniel Korda, Emilie Kaufmann, 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.
****

Accelerating Stochastic Gradient Descent using Predictive Variance Reduction
Rie Johnson, Tong Zhang
Stochastic gradient descent is popular for large scale optimization but has slow convergence asymptotically due to the inherent variance. To remedy this problem, we introduce an explicit variance reduction method for stochastic gradient descent which we call stochastic variance reduced gradient (SVRG). For smooth and strongly convex functions, we prove that this method enjoys the same fast convergence rate as those of stochastic dual coordinate ascent (SDCA) and Stochastic Average Gradient (SAG). However, our analysis is significantly simpler and more intuitive. Moreover, unlike SDCA or SAG, our method does not require the storage of gradients, and thus is more easily applicable to complex problems such as some structured prediction problems and neural network learning.
****

Multisensory Encoding, Decoding, and Identification
Aurel A. Lazar, Yevgeniy Slutskiy

We investigate a spiking neuron model of multisensory integration. Multiple stim
uli from different sensory modalities are encoded by a single neural circuit com
prised of a multisensory bank of receptive fields in cascade with a population o
f biophysical spike generators. We demonstrate that stimuli of different dimensi
ons can be faithfully multiplexed and encoded in the spike domain and derive tra
ctable algorithms for decoding each stimulus from the common pool of spikes. We
also show that the identification of multisensory processing in a single neuron
is dual to the recovery of stimuli encoded with a population of multisensory neu
rons, and prove that only a projection of the circuit onto input stimuli can be
identified. We provide an example of multisensory integration using natural audi
o and video and discuss the performance of the proposed decoding and identificat
ion algorithms.
****

Learning invariant representations and applications to face verification
Qianli Liao, Joel Z. Leibo, Tomaso Poggio
One approach to computer object recognition and modeling the brain's ventral str
eam involves unsupervised learning of representations that are invariant to comm
on transformations. However, applications of these ideas have usually been limit
ed to 2D affine transformations, e.g., translation and scaling, since they are e
asiest to solve via convolution. In accord with a recent theory of transformatio
n-invariance, we propose a model that, while capturing other common convolutiona
l networks as special cases, can also be used with arbitrary identity-preserving
 transformations. The model's wiring can be learned from videos of transforming
objects---or any other grouping of images into sets by their depicted object. Th
rough a series of successively more complex empirical tests, we study the invari
ance/discriminability properties of this model with respect to different transfo
rmations. First, we empirically confirm theoretical predictions for the case of
2D affine transformations. Next, we apply the model to non-affine transformation
s: as expected, it performs well on face verification tasks requiring invariance
 to the relatively smooth transformations of 3D rotation-in-depth and changes in
 illumination direction. Surprisingly, it can also tolerate clutter transformati
ons'' which map an image of a face on one background to an image of the same fac
e on a different background. Motivated by these empirical findings, we tested th
e same model on face verification benchmark tasks from the computer vision liter
ature: Labeled Faces in the Wild, PubFig and a new dataset we gathered---achievi
ng strong performance in these highly unconstrained cases as well."
****

Optimistic Concurrency Control for Distributed Unsupervised Learning
Xinghao Pan, Joseph E. Gonzalez, Stefanie Jegelka, Tamara Broderick, Michael I.
Jordan
Research on distributed machine learning algorithms has focused primarily on one
 of two extremes---algorithms that obey strict concurrency constraints or algori
thms that obey few or no such constraints. We consider an intermediate alternat
ive in which algorithms optimistically assume that conflicts are unlikely and if
 conflicts do arise a conflict-resolution protocol is invoked. We view this opti
mistic concurrency control'' paradigm as particularly appropriate for large-scal
e machine learning algorithms, particularly in the unsupervised setting. We dem
onstrate our approach in three problem areas: clustering, feature learning and o
nline facility location. We evaluate our methods via large-scale experiments i
n a cluster computing environment. "
****

Statistical analysis of coupled time series with Kernel Cross-Spectral Density o
perators.
Michel Besserve, Nikos K. Logothetis, Bernhard Schölkopf
Many applications require the analysis of complex interactions between time seri
es. These interactions can be non-linear and involve vector valued as well as co
mplex data structures such as graphs or strings. Here we provide a general frame
work for the statistical analysis of these interactions when random variables ar
e sampled from stationary time-series of arbitrary objects. To achieve this goal
 we analyze the properties of the kernel cross-spectral density operator induced

by positive definite kernels on arbitrary input domains. This framework enables us to develop an independence test between time series as well as a similarity measure to compare different types of coupling. The performance of our test is compared to the HSIC test using i.i.d. assumptions, showing improvement in terms of detection errors as well as the suitability of this approach for testing dependency in complex dynamical systems. Finally, we use this approach to characterize complex interactions in electrophysiological neural time series.
****

Sinkhorn Distances: Lightspeed Computation of Optimal Transport
Marco Cuturi
Optimal transportation distances are a fundamental family of parameterized distances for histograms in the probability simplex. Despite their appealing theoretical properties, excellent performance and intuitive formulation, their computation involves the resolution of a linear program whose cost is prohibitive whenever the histograms' dimension exceeds a few hundreds. We propose in this work a new family of optimal transportation distances that look at transportation problems from a maximum-entropy perspective. We smooth the classical optimal transportation problem with an entropic regularization term, and show that the resulting optimum is also a distance which can be computed through Sinkhorn's matrix scaling algorithm at a speed that is several orders of magnitude faster than that of transportation solvers. We also report improved performance on the MNIST benchmark problem over competing distances.
****

Nonparametric Multi-group Membership Model for Dynamic Networks
Myunghwan Kim, Jure Leskovec
Relational data—like graphs, networks, and matrices—is often dynamic, where the relational structure evolves over time. A fundamental problem in the analysis of time-varying network data is to extract a summary of the common structure and the dynamics of underlying relations between entities. Here we build on the intuition that changes in the network structure are driven by the dynamics at the level of groups of nodes. We propose a nonparametric multi-group membership model for dynamic networks. Our model contains three main components. We model the birth and death of groups with respect to the dynamics of the network structure via a distance dependent Indian Buffet Process. We capture the evolution of individual node group memberships via a Factorial Hidden Markov model. And, we explain the dynamics of the network structure by explicitly modeling the connectivity structure. We demonstrate our model’s capability of identifying the dynamics of latent groups in a number of different types of network data. Experimental results show our model achieves higher predictive performance on the future network forecasting and missing link prediction.
****

EDML for Learning Parameters in Directed and Undirected Graphical Models
Khaled S. Refaat, Arthur Choi, Adnan Darwiche
EDML is a recently proposed algorithm for learning parameters in Bayesian networks. It was originally derived in terms of approximate inference on a meta-network, which underlies the Bayesian approach to parameter estimation. While this initial derivation helped discover EDML in the first place and provided a concrete context for identifying some of its properties (e.g., in contrast to EM), the formal setting was somewhat tedious in the number of concepts it drew on. In this paper, we propose a greatly simplified perspective on EDML, which casts it as a general approach to continuous optimization. The new perspective has several advantages. First, it makes immediate some results that were non-trivial to prove initially. Second, it facilitates the design of EDML algorithms for new graphical models, leading to a new algorithm for learning parameters in Markov networks. We derive this algorithm in this paper, and show, empirically, that it can sometimes learn better estimates from complete data, several times faster than commonly used optimization methods, such as conjugate gradient and L-BFGS.
****

Flexible sampling of discrete data correlations without the marginal distributions

Alfredo Kalaitzis, Ricardo Silva
Learning the joint dependence of discrete variables is a fundamental problem in machine learning, with many applications including prediction, clustering and dimensionality reduction. More recently, the framework of copula modeling has gained popularity due to its modular parametrization of joint distributions. Among other properties, copulas provide a recipe for combining flexible models for univariate marginal distributions with parametric families suitable for potentially high dimensional dependence structures. More radically, the extended rank likelihood approach of Hoff (2007) bypasses learning marginal models completely when such information is ancillary to the learning task at hand as in, e.g., standard dimensionality reduction problems or copula parameter estimation. The main idea is to represent data by their observable rank statistics, ignoring any other information from the marginals. Inference is typically done in a Bayesian framework with Gaussian copulas, and it is complicated by the fact this implies sampling within a space where the number of constraints increase quadratically with the number of data points. The result is slow mixing when using off-the-shelf Gibbs sampling. We present an efficient algorithm based on recent advances on constrained Hamiltonian Markov chain Monte Carlo that is simple to implement and does not require paying for a quadratic cost in sample size.
****

Designed Measurements for Vector Count Data
Liming Wang, David E. Carlson, Miguel Rodrigues, David Wilcox, Robert Calderbank, Lawrence Carin
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****

Reasoning With Neural Tensor Networks for Knowledge Base Completion
Richard Socher, Danqi Chen, Christopher D. Manning, Andrew Ng
A common problem in knowledge representation and related fields is reasoning over a large joint knowledge graph, represented as triples of a relation between two entities. The goal of this paper is to develop a more powerful neural network model suitable for inference over these relationships. Previous models suffer from weak interaction between entities or simple linear projection of the vector space. We address these problems by introducing a neural tensor network (NTN) model which allow the entities and relations to interact multiplicatively. Additionally, we observe that such knowledge base models can be further improved by representing each entity as the average of vectors for the words in the entity name, giving an additional dimension of similarity by which entities can share statistical strength. We assess the model by considering the problem of predicting additional true relations between entities given a partial knowledge base. Our model outperforms previous models and can classify unseen relationships in WordNet and FreeBase with an accuracy of 86.2% and 90.0%, respectively.
****

Deep content-based music recommendation
Aaron van den Oord, Sander Dieleman, Benjamin Schrauwen
Automatic music recommendation has become an increasingly relevant problem in recent years, since a lot of music is now sold and consumed digitally. Most recommender systems rely on collaborative filtering. However, this approach suffers from the cold start problem: it fails when no usage data is available, so it is not effective for recommending new and unpopular songs. In this paper, we propose to use a latent factor model for recommendation, and predict the latent factors from music audio when they cannot be obtained from usage data. We compare a traditional approach using a bag-of-words representation of the audio signals with deep convolutional neural networks, and evaluate the predictions quantitatively and qualitatively on the Million Song Dataset. We show that using predicted latent factors produces sensible recommendations, despite the fact that there is a large semantic gap between the characteristics of a song that affect user preference and the corresponding audio signal. We also show that recent advances in deep

learning translate very well to the music recommendation setting, with deep convolutional neural networks significantly outperforming the traditional approach.
****
What do row and column marginals reveal about your dataset?
Behzad Golshan, John Byers, Evimaria Terzi
Numerous datasets ranging from group memberships within social networks to purchase histories on e-commerce sites are represented by binary matrices. While this data is often either proprietary or sensitive, aggregated data, notably row and column marginals, is often viewed as much less sensitive, and may be furnished for analysis. Here, we investigate how these data can be exploited to make inferences about the underlying matrix H. Instead of assuming a generative model for H, we view the input marginals as constraints on the dataspace of possible realizations of H and compute the probability density function of particular entries H(i,j) of interest. We do this, for all the cells of H simultaneously, without generating realizations but rather via implicitly sampling the datasets that satisfy the input marginals. The end result is an efficient algorithm with running time equal to the time required by standard sampling techniques to generate a single dataset from the same dataspace. Our experimental evaluation demonstrates the efficiency and the efficacy of our framework in multiple settings.
****
Analyzing Hogwild Parallel Gaussian Gibbs Sampling
Matthew J. Johnson, James Saunderson, Alan Willsky
Sampling inference methods are computationally difficult to scale for many models in part because global dependencies can reduce opportunities for parallel computation. Without strict conditional independence structure among variables, standard Gibbs sampling theory requires sample updates to be performed sequentially, even if dependence between most variables is not strong. Empirical work has shown that some models can be sampled effectively by going Hogwild'' and simply running Gibbs updates in parallel with only periodic global communication, but the successes and limitations of such a strategy are not well understood. As a step towards such an understanding, we study the Hogwild Gibbs sampling strategy in the context of Gaussian distributions. We develop a framework which provides convergence conditions and error bounds along with simple proofs and connections to methods in numerical linear algebra. In particular, we show that if the Gaussian precision matrix is generalized diagonally dominant, then any Hogwild Gibbs sampler, with any update schedule or allocation of variables to processors, yields a stable sampling process with the correct sample mean. "
****
Latent Structured Active Learning
Wenjie Luo, Alex Schwing, Raquel Urtasun
In this paper we present active learning algorithms in the context of structured prediction problems. To reduce the amount of labeling necessary to learn good models, our algorithms only label subsets of the output. To this end, we query examples using entropies of local marginals, which are a good surrogate for uncertainty. We demonstrate the effectiveness of our approach in the task of 3D layout prediction from single images, and show that good models are learned when labeling only a handful of random variables. In particular, the same performance as using the full training set can be obtained while only labeling ~10\% of the random variables.
****
Confidence Intervals and Hypothesis Testing for High-Dimensional Statistical Models
Adel Javanmard, Andrea Montanari
Fitting high-dimensional statistical models often requires the use of non-linear parameter estimation procedures. As a consequence, it is generally impossible to obtain an exact characterization of the probability distribution of the parameter estimates. This in turn implies that it is extremely challenging to quantify the uncertainty' associated with a certain parameter estimate. Concretely, no commonly accepted procedure exists for computing classical measures of uncertainty and statistical significance as confidence intervals or p-values. We consider

here a broad class of regression problems, and propose an efficient algorithm for constructing confidence intervals and p-values. The resulting confidence intervals have nearly optimal size. When testing for the null hypothesis that a certain parameter is vanishing, our method has nearly optimal power. Our approach is based on constructing ade-biased' version of regularized M-estimators. The new construction improves over recent work in the field in that it does not assume a special structure on the design matrix. Furthermore, proofs are remarkably simple. We test our method on a diabetes prediction problem.
****

Stochastic blockmodel approximation of a graphon: Theory and consistent estimation
Edo M. Airoldi, Thiago B. Costa, Stanley H. Chan
Given a convergent sequence of graphs, there exists a limit object called the graphon from which random graphs are generated. This nonparametric perspective of random graphs opens the door to study graphs beyond the traditional parametric models, but at the same time also poses the challenging question of how to estimate the graphon underlying observed graphs. In this paper, we propose a computationally efficient algorithm to estimate a graphon from a set of observed graphs generated from it. We show that, by approximating the graphon with stochastic block models, the graphon can be consistently estimated, that is, the estimation error vanishes as the size of the graph approaches infinity.
****

More Effective Distributed ML via a Stale Synchronous Parallel Parameter Server
Qirong Ho, James Cipar, Henggang Cui, Seunghak Lee, Jin Kyu Kim, Phillip B. Gibbons, Garth A. Gibson, Greg Ganger, Eric P. Xing
We propose a parameter server system for distributed ML, which follows a Stale Synchronous Parallel (SSP) model of computation that maximizes the time computational workers spend doing useful work on ML algorithms, while still providing correctness guarantees. The parameter server provides an easy-to-use shared interface for read/write access to an ML model's values (parameters and variables), and the SSP model allows distributed workers to read older, stale versions of these values from a local cache, instead of waiting to get them from a central storage. This significantly increases the proportion of time workers spend computing, as opposed to waiting. Furthermore, the SSP model ensures ML algorithm correctness by limiting the maximum age of the stale values. We provide a proof of correctness under SSP, as well as empirical results demonstrating that the SSP model achieves faster algorithm convergence on several different ML problems, compared to fully-synchronous and asynchronous schemes.
****

On Algorithms for Sparse Multi-factor NMF
Siwei Lyu, Xin Wang
Nonnegative matrix factorization (NMF) is a popular data analysis method, the objective of which is to decompose a matrix with all nonnegative components into the product of two other nonnegative matrices. In this work, we describe a new simple and efficient algorithm for multi-factor nonnegative matrix factorization problem ({mfNMF}), which generalizes the original NMF problem to more than two factors. Furthermore, we extend the mfNMF algorithm to incorporate a regularizer based on Dirichlet distribution over normalized columns to encourage sparsity in the obtained factors. Our sparse NMF algorithm affords a closed form and an intuitive interpretation, and is more efficient in comparison with previous works that use fix point iterations. We demonstrate the effectiveness and efficiency of our algorithms on both synthetic and real data sets.
****

Efficient Exploration and Value Function Generalization in Deterministic Systems
Zheng Wen, Benjamin Van Roy
We consider the problem of reinforcement learning over episodes of a finite-horizon deterministic system and as a solution propose optimistic constraint propagation (OCP), an algorithm designed to synthesize efficient exploration and value function generalization. We establish that when the true value function lies within the given hypothesis class, OCP selects optimal actions over all but at mos

t K episodes, where K is the eluder dimension of the given hypothesis class. We establish further efficiency and asymptotic performance guarantees that apply ev en if the true value function does not lie in the given hypothesis space, for t he special case where the hypothesis space is the span of pre-specified indicato r functions over disjoint sets.
****

Parallel Sampling of DP Mixture Models using Sub-Cluster Splits
Jason Chang, John W. Fisher III
We present a novel MCMC sampler for Dirichlet process mixture models that can be used for conjugate or non-conjugate prior distributions. The proposed sampler c an be massively parallelized to achieve significant computational gains. A non-e rgodic restricted Gibbs iteration is mixed with split/merge proposals to produce a valid sampler. Each regular cluster is augmented with two sub-clusters to con struct likely split moves. Unlike many previous parallel samplers, the proposed sampler accurately enforces the correct stationary distribution of the Markov ch ain without the need for approximate models. Empirical results illustrate that t he new sampler exhibits better convergence properties than current methods.
****

Context-sensitive active sensing in humans
Sheeraz Ahmad, He Huang, Angela J. Yu
Humans and animals readily utilize active sensing, or the use of self-motion, to focus sensory and cognitive resources on the behaviorally most relevant stimuli and events in the environment. Understanding the computational basis of natural active sensing is important both for advancing brain sciences and for developin g more powerful artificial systems. Recently, a goal-directed, context-sensitive , Bayesian control strategy for active sensing, termed C-DAC (Context-Dependent Active Controller), was proposed (Ahmad & Yu, 2013). In contrast to previously p roposed algorithms for human active vision, which tend to optimize abstract stat istical objectives and therefore cannot adapt to changing behavioral context or task goals, C-DAC directly minimizes behavioral costs and thus, automatically ad apts itself to different task conditions. However, C-DAC is limited as a model o f human active sensing, given its computational/representational requirements, e specially for more complex, real-world situations. Here, we propose a myopic app roximation to C-DAC, which also takes behavioral costs into account, but achieve s a significant reduction in complexity by looking only one step ahead. We also present data from a human active visual search experiment, and compare the perfo rmance of the various models against human behavior. We find that C-DAC and its myopic variant both achieve better fit to human data than Infomax (Butko & Movel lan, 2010), which maximizes expected cumulative future information gain. In summ ary, this work provides novel experimental results that differentiate theoretica l models for human active sensing, as well as a novel active sensing algorithm t hat retains the context-sensitivity of the optimal controller while achieving si gnificant computational savings.
****

On the Sample Complexity of Subspace Learning
Alessandro Rudi, Guillermo D. Canas, Lorenzo Rosasco
A large number of algorithms in machine learning, from principal component analy sis (PCA), and its non-linear (kernel) extensions, to more recent spectral embed ding and support estimation methods, rely on estimating a linear subspace from s amples. In this paper we introduce a general formulation of this problem and derive novel learning error estimates. Our results rely on natural assumptions on the spectral properties of the covariance operator associated to the data d istribution, and hold for a wide class of metrics between subspaces. As special cases, we discuss sharp error estimates for the reconstruction properties of PC A and spectral support estimation. Key to our analysis is an operator theoretic approach that has broad applicability to spectral learning methods.
****

Non-Linear Domain Adaptation with Boosting
Carlos J. Becker, Christos M. Christoudias, Pascal Fua
A common assumption in machine vision is that the training and test samples are

drawn from the same distribution. However, there are many problems when this ass
umption is grossly violated, as in bio-medical applications where different acqu
isitions can generate drastic variations in the appearance of the data due to ch
anging experimental conditions. This problem is accentuated with 3D data, for wh
ich annotation is very time-consuming, limiting the amount of data that can be l
abeled in new acquisitions for training. In this paper we present a multi-task l
earning algorithm for domain adaptation based on boosting. Unlike previous appro
aches that learn task-specific decision boundaries, our method learns a single d
ecision boundary in a shared feature space, common to all tasks. We use the boos
ting-trick to learn a non-linear mapping of the observations in each task, with
no need for specific a-priori knowledge of its global analytical form. This yiel
ds a more parameter-free domain adaptation approach that successfully leverages
learning on new tasks where labeled data is scarce. We evaluate our approach on
two challenging bio-medical datasets and achieve a significant improvement over
the state-of-the-art.
****

Learning Trajectory Preferences for Manipulators via Iterative Improvement
Ashesh Jain, Brian Wojcik, Thorsten Joachims, Ashutosh Saxena
We consider the problem of learning good trajectories for manipulation tasks. Th
is is challenging because the criterion defining a good trajectory varies with u
sers, tasks and environments. In this paper, we propose a co-active online lea
rning framework for teaching robots the preferences of its users for object man
ipulation tasks. The key novelty of our approach lies in the type of feedback ex
pected from the user: the human user does not need to demonstrate optimal trajec
tories as training data, but merely needs to iteratively provide trajectories th
at slightly improve over the trajectory currently proposed by the system. We arg
ue that this co-active preference feedback can be more easily elicited from the
user than demonstrations of optimal trajectories, which are often challenging a
nd non-intuitive to provide on high degrees of freedom manipulators. Nevertheles
s, theoretical regret bounds of our algorithm match the asymptotic rates of opti
mal trajectory algorithms. We also formulate a score function to capture the con
textual information and demonstrate the generalizability of our algorithm on a v
ariety of household tasks, for whom, the preferences were not only influenced by
the object being manipulated but also by the surrounding environment.
****

Learning Chordal Markov Networks by Constraint Satisfaction
Jukka Corander, Tomi Janhunen, Jussi Rintanen, Henrik Nyman, Johan Pensar
We investigate the problem of learning the structure of a Markov network from da
ta. It is shown that the structure of such networks can be described in terms of
constraints which enables the use of existing solver technology with optimizati
on capabilities to compute optimal networks starting from initial scores compute
d from the data. To achieve efficient encodings, we develop a novel characteriza
tion of Markov network structure using a balancing condition on the separators b
etween cliques forming the network. The resulting translations into propositiona
l satisfiability and its extensions such as maximum satisfiability, satisfiabili
ty modulo theories, and answer set programming, enable us to prove the optimalit
y of networks which have been previously found by stochastic search.
****

Curvature and Optimal Algorithms for Learning and Minimizing Submodular Function
s
Rishabh K. Iyer, Stefanie Jegelka, Jeff A. Bilmes
We investigate three related and important problems connected to machine learnin
g, namely approximating a submodular function everywhere, learning a submodular
function (in a PAC like setting [26]), and constrained minimization of submodula
r functions. In all three problems, we provide improved bounds which depend on t
he “curvature” of a submodular function and improve on the previously known best
results for these problems [9, 3, 7, 25] when the function is not too curved –
a property which is true of many real-world submodular functions. In the former
two problems, we obtain these bounds through a generic black-box transformation
(which can potentially work for any algorithm), while in the case of submodular

minimization, we propose a framework of algorithms which depend on choosing an a
ppropriate surrogate for the submodular function. In all these cases, we provide
 almost matching lower bounds. While improved curvature-dependent bounds were sh
own for monotone submodular maximization [4, 27], the existence of similar impro
ved bounds for the aforementioned problems has been open. We resolve this questi
on in this paper by showing that the same notion of curvature provides these imp
roved results. Empirical experiments add further support to our claims.
****

A New Convex Relaxation for Tensor Completion
Bernardino Romera-Paredes, Massimiliano Pontil
We study the problem of learning a tensor from a set of linear measurements. A p
rominent methodology for this problem is based on the extension of trace norm re
gularization, which has been used extensively for learning low rank matrices, to
 the tensor setting. In this paper, we highlight some limitations of this appro
ach and propose an alternative convex relaxation on the Euclidean unit ball. We
 then describe a technique to solve the associated regularization problem, which
 builds upon the alternating direction method of multipliers. Experiments on one
 synthetic dataset and two real datasets indicate that the proposed method impro
ves significantly over tensor trace norm regularization in terms of estimation e
rror, while remaining computationally tractable.
****

DESPOT: Online POMDP Planning with Regularization
Adhiraj Somani, Nan Ye, David Hsu, Wee Sun Lee
POMDPs provide a principled framework for planning under uncertainty, but are co
mputationally intractable, due to the “curse of dimensionality” and the “curse o
f history”. This paper presents an online lookahead search algorithm that allevi
ates these difficulties by limiting the search to a set of sampled scenarios. Th
e execution of all policies on the sampled scenarios is summarized using a Deter
minized Sparse Partially Observable Tree (DESPOT), which is a sparsely sampled b
elief tree. Our algorithm, named Regularized DESPOT (R-DESPOT), searches the DES
POT for a policy that optimally balances the size of the policy and the accuracy
 on its value estimate obtained through sampling. We give an output-sensitive pe
rformance bound for all policies derived from the DESPOT, and show that R-DESPOT
 works well if a small optimal policy exists. We also give an anytime approximat
ion to R-DESPOT. Experiments show strong results, compared with two of the faste
st online POMDP algorithms.
****

Speeding up Permutation Testing in Neuroimaging
Chris Hinrichs, Vamsi K. Ithapu, Qinyuan Sun, Sterling C. Johnson, Vikas Singh
Multiple hypothesis testing is a significant problem in nearly all neuroimaging
studies. In order to correct for this phenomena, we require a reliable estimate
of the Family-Wise Error Rate (FWER). The well known Bonferroni correction metho
d, while being simple to implement, is quite conservative, and can substantially
 under-power a study because it ignores dependencies between test statistics. Pe
rmutation testing, on the other hand, is an exact, non parametric method of esti
mating the FWER for a given α threshold, but for acceptably low thresholds the c
omputational burden can be prohibitive. In this paper, we observe that permutati
on testing in fact amounts to populating the columns of a very large matrix P. B
y analyzing the spectrum of this matrix, under certain conditions, we see that P
 has a low-rank plus a low-variance residual decomposition which makes it suitab
le for highly sub–sampled – on the order of 0.5% – matrix completion methods. Th
us, we propose a novel permutation testing methodology which offers a large spee
dup, without sacrificing the fidelity of the estimated FWER. Our valuations on f
our different neuroimaging datasets show that a computational speedup factor of
roughly 50× can be achieved while recovering the FWER distribution up to very hi
gh accuracy. Further, we show that the estimated α threshold is also recovered f
aithfully, and is stable.
****

Neural representation of action sequences: how far can a simple snippet-matching
 model take us?

Cheston Tan, Jedediah M. Singer, Thomas Serre, David Sheinberg, Tomaso Poggio
The macaque Superior Temporal Sulcus (STS) is a brain area that receives and int
egrates inputs from both the ventral and dorsal visual processing streams (thoug
ht to specialize in form and motion processing respectively). For the processing
 of articulated actions, prior work has shown that even a small population of ST
S neurons contains sufficient information for the decoding of actor invariant to
 action, action invariant to actor, as well as the specific conjunction of actor
 and action. This paper addresses two questions. First, what are the invariance
properties of individual neural representations (rather than the population repr
esentation) in STS? Second, what are the neural encoding mechanisms that can pro
duce such individual neural representations from streams of pixel images? We fin
d that a baseline model, one that simply computes a linear weighted sum of ventr
al and dorsal responses to short action “snippets”, produces surprisingly good f
its to the neural data. Interestingly, even using inputs from a single stream, b
oth actor-invariance and action-invariance can be produced simply by having diff
erent linear weights.
****

Modeling Clutter Perception using Parametric Proto-object Partitioning
Chen-Ping Yu, Wen-Yu Hua, Dimitris Samaras, Greg Zelinsky
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.
****

Relevance Topic Model for Unstructured Social Group Activity Recognition
Fang Zhao, Yongzhen Huang, Liang Wang, Tieniu Tan
Unstructured social group activity recognition in web videos is a challenging ta
sk due to 1) the semantic gap between class labels and low-level visual features
 and 2) the lack of labeled training data. To tackle this problem, we propose a
relevance topic model" for jointly learning meaningful mid-level representations
 upon bag-of-words (BoW) video representations and a classifier with sparse weig
hts. In our approach, sparse Bayesian learning is incorporated into an undirecte
d topic model (i.e., Replicated Softmax) to discover topics which are relevant t
o video classes and suitable for prediction. Rectified linear units are utilized
 to increase the expressive power of topics so as to explain better video data c
ontaining complex contents and make variational inference tractable for the prop
osed model. An efficient variational EM algorithm is presented for model paramet
er estimation and inference. Experimental results on the Unstructured Social Act
ivity Attribute dataset show that our model achieves state of the art performanc
e and outperforms other supervised topic model in terms of classification accura
cy, particularly in the case of a very small number of labeled training videos."
****

Generalized Random Utility Models with Multiple Types
Hossein Azari Soufiani, Hansheng Diao, Zhenyu Lai, David C. Parkes
We propose a model for demand estimation in multi-agent, differentiated product
settings and present an estimation algorithm that uses reversible jump MCMC tech
niques to classify agents' types. Our model extends the popular setup in Berry,
Levinsohn and Pakes (1995) to allow for the data-driven classification of agents
' types using agent-level data. We focus on applications involving data on agent
s' ranking over alternatives, and present theoretical conditions that establish
the identifiability of the model and uni-modality of the likelihood/posterior. R
esults on both real and simulated data provide support for the scalability of ou
r approach.
****

(Nearly) Optimal Algorithms for Private Online Learning in Full-information and
Bandit Settings
Abhradeep Guha Thakurta, Adam Smith
Requests for name changes in the electronic proceedings will be accepted with no
 questions asked. However name changes may cause bibliographic tracking issues.
  Authors are asked to consider this carefully and discuss it with their co-auth

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

The Fast Convergence of Incremental PCA
Akshay Balsubramani, Sanjoy Dasgupta, Yoav Freund
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.
****

Point Based Value Iteration with Optimal Belief Compression for Dec-POMDPs
Liam C. MacDermed, Charles L. Isbell
This paper presents four major results towards solving decentralized partially o bservable Markov decision problems (DecPOMDPs) culminating in an algorithm that outperforms all existing algorithms on all but one standard infinite-horizon ben chmark problems. (1) We give an integer program that solves collaborative Bayesi an games (CBGs). The program is notable because its linear relaxation is very of ten integral. (2) We show that a DecPOMDP with bounded belief can be converted t o a POMDP (albeit with actions exponential in the number of beliefs). These acti ons correspond to strategies of a CBG. (3) We present a method to transform any DecPOMDP into a DecPOMDP with bounded beliefs (the number of beliefs is a free p arameter) using optimal (not lossless) belief compression. (4) We show that the combination of these results opens the door for new classes of DecPOMDP algorith ms based on previous POMDP algorithms. We choose one such algorithm, point-based valued iteration, and modify it to produce the first tractable value iteration method for DecPOMDPs which outperforms existing algorithms.
****

Summary Statistics for Partitionings and Feature Allocations
Isik B. Fidaner, Taylan Cemgil
Infinite mixture models are commonly used for clustering. One can sample from th e posterior of mixture assignments by Monte Carlo methods or find its maximum a posteriori solution by optimization. However, in some problems the posterior is diffuse and it is hard to interpret the sampled partitionings. In this paper, we introduce novel statistics based on block sizes for representing sample sets of partitionings and feature allocations. We develop an element-based definition o f entropy to quantify segmentation among their elements. Then we propose a simpl e algorithm called entropy agglomeration (EA) to summarize and visualize this in formation. Experiments on various infinite mixture posteriors as well as a featu re allocation dataset demonstrate that the proposed statistics are useful in pra ctice.
****

Learning Hidden Markov Models from Non-sequence Data via Tensor Decomposition
Tzu-Kuo Huang, Jeff Schneider
Learning dynamic models from observed data has been a central issue in many scie ntific studies or engineering tasks. The usual setting is that data are collect ed sequentially from trajectories of some dynamical system operation. In quit e a few modern scientific modeling tasks, however, it turns out that reliable s equential data are rather difficult to gather, whereas out-of-order snapshots ar e much easier to obtain. Examples include the modeling of galaxies, chronic dise ases such Alzheimer's, or certain biological processes. Existing methods for le arning dynamic model from non-sequence data are mostly based on Expectation-Maxi mization, which involves non-convex optimization and is thus hard to analyze. I nspired by recent advances in spectral learning methods, we propose to study thi s problem from a different perspective: moment matching and spectral decomposit ion. Under that framework, we identify reasonable assumptions on the generative process of non-sequence data, and propose learning algorithms based on the ten sor decomposition method \cite{anandkumar2012tensor} to \textit{provably} recov er first-order Markov models and hidden Markov models. To the best of our knowle dge, this is the first formal guarantee on learning from non-sequence data. Prel iminary simulation results confirm our theoretical findings.
****

On Flat versus Hierarchical Classification in Large-Scale Taxonomies
Rohit Babbar, Ioannis Partalas, Eric Gaussier, Massih R. Amini
We study in this paper flat and hierarchical classification strategies in the co ntext of large-scale taxonomies. To this end, we first propose a multiclass, hie rarchical data dependent bound on the generalization error of classifiers deploy ed in large-scale taxonomies. This bound provides an explanation to several empi rical results reported in the literature, related to the performance of flat and hierarchical classifiers. We then introduce another type of bounds targeting th e approximation error of a family of classifiers, and derive from it features us ed in a meta-classifier to decide which nodes to prune (or flatten) in a large-s cale taxonomy. We finally illustrate the theoretical developments through severa l experiments conducted on two widely used taxonomies.
****
Scoring Workers in Crowdsourcing: How Many Control Questions are Enough?
Qiang Liu, Alexander T. Ihler, Mark Steyvers
We study the problem of estimating continuous quantities, such as prices, probab ilities, and point spreads, using a crowdsourcing approach. A challenging aspec t of combining the crowd's answers is that workers' reliabilities and biases are usually unknown and highly diverse. Control items with known answers can be us ed to evaluate workers' performance, and hence improve the combined results on t he target items with unknown answers. This raises the problem of how many contr ol items to use when the total number of items each workers can answer is limite d: more control items evaluates the workers better, but leaves fewer resources f or the target items that are of direct interest, and vice versa. We give theoret ical results for this problem under different scenarios, and provide a simple ru le of thumb for crowdsourcing practitioners. As a byproduct, we also provide th eoretical analysis of the accuracy of different consensus methods.
****
Cluster Trees on Manifolds
Sivaraman Balakrishnan, Srivatsan Narayanan, Alessandro Rinaldo, Aarti Singh, La rry Wasserman
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.
****
Bayesian inference as iterated random functions with applications to sequential inference in graphical models
Arash Amini, XuanLong Nguyen
We propose a general formalism of iterated random functions with semigroup prope rty, under which exact and approximate Bayesian posterior updates can be viewed as specific instances. A convergence theory for iterated random functions is pre sented. As an application of the general theory we analyze convergence behaviors of exact and approximate message-passing algorithms that arise in a sequential change point detection problem formulated via a latent variable directed graphic al model. The sequential inference algorithm and its supporting theory are illu strated by simulated examples.
****
Recurrent networks of coupled Winner-Take-All oscillators for solving constraint satisfaction problems
Hesham Mostafa, Lorenz. K. Mueller, Giacomo Indiveri
We present a recurrent neuronal network, modeled as a continuous-time dynamical system, that can solve constraint satisfaction problems. Discrete variables are represented by coupled Winner-Take-All (WTA) networks, and their values are enco ded in localized patterns of oscillations that are learned by the recurrent weig hts in these networks. Constraints over the variables are encoded in the network connectivity. Although there are no sources of noise, the network can escape fr om local optima in its search for solutions that satisfy all constraints by modi fying the effective network connectivity through oscillations. If there is no so lution that satisfies all constraints, the network state changes in a pseudo-ran

dom manner and its trajectory approximates a sampling procedure that selects a v
ariable assignment with a probability that increases with the fraction of constr
aints satisfied by this assignment. External evidence, or input to the network,
can force variables to specific values. When new inputs are applied, the network
 re-evaluates the entire set of variables in its search for the states that sati
sfy the maximum number of constraints, while being consistent with the external
input. Our results demonstrate that the proposed network architecture can perfor
m a deterministic search for the optimal solution to problems with non-convex co
st functions. The network is inspired by canonical microcircuit models of the co
rtex and suggests possible dynamical mechanisms to solve constraint satisfaction
 problems that can be present in biological networks, or implemented in neuromor
phic electronic circuits.
****

Rapid Distance-Based Outlier Detection via Sampling
Mahito Sugiyama, Karsten Borgwardt
Distance-based approaches to outlier detection are popular in data mining, as th
ey do not require to model the underlying probability distribution, which is par
ticularly challenging for high-dimensional data. We present an empirical compari
son of various approaches to distance-based outlier detection across a large num
ber of datasets. We report the surprising observation that a simple, sampling-ba
sed scheme outperforms state-of-the-art techniques in terms of both efficiency a
nd effectiveness. To better understand this phenomenon, we provide a theoretical
 analysis why the sampling-based approach outperforms alternative methods based
on k-nearest neighbor search.
****

Visual Concept Learning: Combining Machine Vision and Bayesian Generalization on
 Concept Hierarchies
Yangqing Jia, Joshua T. Abbott, Joseph L. Austerweil, Tom Griffiths, Trevor Darr
ell
Learning a visual concept from a small number of positive examples is a signific
ant challenge for machine learning algorithms. Current methods typically fail to
 find the appropriate level of generalization in a concept hierarchy for a given
 set of visual examples. Recent work in cognitive science on Bayesian models of
generalization addresses this challenge, but prior results assumed that objects
were perfectly recognized. We present an algorithm for learning visual concepts
 directly from images, using probabilistic predictions generated by visual clas
sifiers as the input to a Bayesian generalization model. As no existing challeng
e data tests this paradigm, we collect and make available a new, large-scale dat
aset for visual concept learning using the ImageNet hierarchy as the source of p
ossible concepts, with human annotators to provide ground truth labels as to whe
ther a new image is an instance of each concept using a paradigm similar to that
 used in experiments studying word learning in children. We compare the perform
ance of our system to several baseline algorithms, and show a significant advant
age results from combining visual classifiers with the ability to identify an ap
propriate level of abstraction using Bayesian generalization.
****

Memoized Online Variational Inference for Dirichlet Process Mixture Models
Michael C. Hughes, Erik Sudderth
Variational inference algorithms provide the most effective framework for large-
scale training of Bayesian nonparametric models. Stochastic online approaches a
re promising, but are sensitive to the chosen learning rate and often converge t
o poor local optima. We present a new algorithm, memoized online variational in
ference, which scales to very large (yet finite) datasets while avoiding the com
plexities of stochastic gradient. Our algorithm maintains finite-dimensional su
fficient statistics from batches of the full dataset, requiring some additional
memory but still scaling to millions of examples. Exploiting nested families of
 variational bounds for infinite nonparametric models, we develop principled bir
th and merge moves allowing non-local optimization. Births adaptively add compo
nents to the model to escape local optima, while merges remove redundancy and im
prove speed. Using Dirichlet process mixture models for image clustering and de

noising, we demonstrate major improvements in robustness and accuracy.
****
Locally Adaptive Bayesian Multivariate Time Series
Daniele Durante, Bruno Scarpa, David B. Dunson
In modeling multivariate time series, it is important to allow time-varying smoothness in the mean and covariance process. In particular, there may be certain time intervals exhibiting rapid changes and others in which changes are slow. If such locally adaptive smoothness is not accounted for, one can obtain misleading inferences and predictions, with over-smoothing across erratic time intervals and under-smoothing across times exhibiting slow variation. This can lead to miscalibration of predictive intervals, which can be substantially too narrow or wide depending on the time. We propose a continuous multivariate stochastic process for time series having locally varying smoothness in both the mean and covariance matrix. This process is constructed utilizing latent dictionary functions in time, which are given nested Gaussian process priors and linearly related to the observed data through a sparse mapping. Using a differential equation representation, we bypass usual computational bottlenecks in obtaining MCMC and online algorithms for approximate Bayesian inference. The performance is assessed in simulations and illustrated in a financial application.
****
When in Doubt, SWAP: High-Dimensional Sparse Recovery from Correlated Measurements
Divyanshu Vats, Richard Baraniuk
We consider the problem of accurately estimating a high-dimensional sparse vector using a small number of linear measurements that are contaminated by noise. It is well known that standard computationally tractable sparse recovery algorithms, such as the Lasso, OMP, and their various extensions, perform poorly when the measurement matrix contains highly correlated columns. We develop a simple greedy algorithm, called SWAP, that iteratively swaps variables until a desired loss function cannot be decreased any further. SWAP is surprisingly effective in handling measurement matrices with high correlations. We prove that SWAP can be easily used as a wrapper around standard sparse recovery algorithms for improved performance. We theoretically quantify the statistical guarantees of SWAP and complement our analysis with numerical results on synthetic and real data.
****
Information-theoretic lower bounds for distributed statistical estimation with communication constraints
Yuchen Zhang, John Duchi, Michael I. Jordan, Martin J. Wainwright
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 Stochastic Feedforward Neural Networks
Charlie Tang, Russ R. Salakhutdinov
Multilayer perceptrons (MLPs) or neural networks are popular models used for nonlinear regression and classification tasks. As regressors, MLPs model the conditional distribution of the predictor variables Y given the input variables X. However, this predictive distribution is assumed to be unimodal (e.g. Gaussian). For tasks such as structured prediction problems, the conditional distribution should be multimodal, forming one-to-many mappings. By using stochastic hidden variables rather than deterministic ones, Sigmoid Belief Nets (SBNs) can induce a rich multimodal distribution in the output space. However, previously proposed learning algorithms for SBNs are very slow and do not work well for real-valued data. In this paper, we propose a stochastic feedforward network with hidden layers having \emph{both deterministic and stochastic} variables. A new Generalized EM training procedure using importance sampling allows us to efficiently learn complicated conditional distributions. We demonstrate the superiority of our model to conditional Restricted Boltzmann Machines and Mixture Density Networks on synthetic datasets and on modeling facial expressions. Moreover, we show that laten

t features of our model improves classification and provide additional qualitati ve results on color images.

****

Robust Transfer Principal Component Analysis with Rank Constraints

Yuhong Guo

Principal component analysis (PCA), a well-established technique for data analys is and processing, provides a convenient form of dimensionality reduction that i s effective for cleaning small Gaussian noises presented in the data. However, t he applicability of standard principal component analysis in real scenarios is l imited by its sensitivity to large errors. In this paper, we tackle the challeng e problem of recovering data corrupted with errors of high magnitude by developi ng a novel robust transfer principal component analysis method. Our method is ba sed on the assumption that useful information for the recovery of a corrupted da ta matrix can be gained from an uncorrupted related data matrix. Speci■cally, we formulate the data recovery problem as a joint robust principal component analy sis problem on the two data matrices, with shared common principal components ac ross matrices and individual principal components speci■c to each data matrix. T he formulated optimization problem is a minimization problem over a convex objec tive function but with non-convex rank constraints. We develop an ef■cient proxi mal projected gradient descent algorithm to solve the proposed optimization prob lem with convergence guarantees. Our empirical results over image denoising task s show the proposed method can effectively recover images with random large erro rs, and signi■cantly outperform both standard PCA and robust PCA.

****

A Determinantal Point Process Latent Variable Model for Inhibition in Neural Spi king Data

Jasper Snoek, Richard Zemel, Ryan P. Adams

Point processes are popular models of neural spiking behavior as they provide a statistical distribution over temporal sequences of spikes and help to reveal th e complexities underlying a series of recorded action potentials. However, the most common neural point process models, the Poisson process and the gamma renew al process, do not capture interactions and correlations that are critical to mo deling populations of neurons. We develop a novel model based on a determinanta l point process over latent embeddings of neurons that effectively captures and helps visualize complex inhibitory and competitive interaction. We show that th is model is a natural extension of the popular generalized linear model to sets of interacting neurons. The model is extended to incorporate gain control or d ivisive normalization, and the modulation of neural spiking based on periodic ph enomena. Applied to neural spike recordings from the rat hippocampus, we see th at the model captures inhibitory relationships, a dichotomy of classes of neuron s, and a periodic modulation by the theta rhythm known to be present in the data .

****

Reciprocally Coupled Local Estimators Implement Bayesian Information Integration Distributively

Wen-Hao Zhang, Si Wu

Psychophysical experiments have demonstrated that the brain integrates informati on from multiple sensory cues in a near Bayesian optimal manner. The present stu dy proposes a novel mechanism to achieve this. We consider two reciprocally conn ected networks, mimicking the integration of heading direction information betwe en the dorsal medial superior temporal (MSTd) and the ventral intraparietal (VIP) areas. Each network serves as a local estimator and receives an independent cu e, either the visual or the vestibular, as direct input for the external stimulu s. We find that positive reciprocal interactions can improve the decoding accura cy of each individual network as if it implements Bayesian inference from two cu es. Our model successfully explains the experimental finding that both MSTd and VIP achieve Bayesian multisensory integration, though each of them only receives a single cue as direct external input. Our result suggests that the brain may i mplement optimal information integration distributively at each local estimator through the reciprocal connections between cortical regions.

****
Structured Learning via Logistic Regression
Justin Domke
A successful approach to structured learning is to write the learning objective as a joint function of linear parameters and inference messages, and iterate between updates to each. This paper observes that if the inference problem is “smoothed” through the addition of entropy terms, for fixed messages, the learning objective reduces to a traditional (non-structured) logistic regression problem with respect to parameters. In these logistic regression problems, each training example has a bias term determined by the current set of messages. Based on this insight, the structured energy function can be extended from linear factors to any function class where an “oracle” exists to minimize a logistic loss.
****
Learning word embeddings efficiently with noise-contrastive estimation
Andriy Mnih, Koray Kavukcuoglu
Continuous-valued word embeddings learned by neural language models have recently been shown to capture semantic and syntactic information about words very well, setting performance records on several word similarity tasks. The best results are obtained by learning high-dimensional embeddings from very large quantities of data, which makes scalability of the training method a critical factor. We propose a simple and scalable new approach to learning word embeddings based on training log-bilinear models with noise-contrastive estimation. Our approach is simpler, faster, and produces better results than the current state-of-the art method of Mikolov et al. (2013a). We achieve results comparable to the best ones reported, which were obtained on a cluster, using four times less data and more than an order of magnitude less computing time. We also investigate several model types and find that the embeddings learned by the simpler models perform at least as well as those learned by the more complex ones.
****
Generalized Method-of-Moments for Rank Aggregation
Hossein Azari Soufiani, William Chen, David C. Parkes, Lirong Xia
In this paper we propose a class of efficient Generalized Method-of-Moments(GMM) algorithms for computing parameters of the Plackett-Luce model, where the data consists of full rankings over alternatives. Our technique is based on breaking the full rankings into pairwise comparisons, and then computing parameters that satisfy a set of generalized moment conditions. We identify conditions for the output of GMM to be unique, and identify a general class of consistent and inconsistent breakings. We then show by theory and experiments that our algorithms run significantly faster than the classical Minorize-Maximization (MM) algorithm, while achieving competitive statistical efficiency.
****
Reconciling "priors" & "priors" without prejudice?
Remi Gribonval, Pierre Machart
There are two major routes to address linear inverse problems. Whereas regularization-based approaches build estimators as solutions of penalized regression optimization problems, Bayesian estimators rely on the posterior distribution of the unknown, given some assumed family of priors. While these may seem radically different approaches, recent results have shown that, in the context of additive white Gaussian denoising, the Bayesian conditional mean estimator is always the solution of a penalized regression problem. The contribution of this paper is twofold. First, we extend the additive white Gaussian denoising results to general linear inverse problems with colored Gaussian noise. Second, we characterize conditions under which the penalty function associated to the conditional mean estimator can satisfy certain popular properties such as convexity, separability, and smoothness. This sheds light on some tradeoff between computational efficiency and estimation accuracy in sparse regularization, and draws some connections between Bayesian estimation and proximal optimization.
****
Learning a Deep Compact Image Representation for Visual Tracking
Naiyan Wang, Dit-Yan Yeung

In this paper, we study the challenging problem of tracking the trajectory of a moving object in a video with possibly very complex background. In contrast to most existing trackers which only learn the appearance of the tracked object online, we take a different approach, inspired by recent advances in deep learning architectures, by putting more emphasis on the (unsupervised) feature learning problem. Specifically, by using auxiliary natural images, we train a stacked denoising autoencoder offline to learn generic image features that are more robust against variations. This is then followed by knowledge transfer from offline training to the online tracking process. Online tracking involves a classification neural network which is constructed from the encoder part of the trained autoencoder as a feature extractor and an additional classification layer. Both the feature extractor and the classifier can be further tuned to adapt to appearance changes of the moving object. Comparison with the state-of-the-art trackers on some challenging benchmark video sequences shows that our deep learning tracker is very efficient as well as more accurate.

****

Trading Computation for Communication: Distributed Stochastic Dual Coordinate Ascent

Tianbao Yang

We present and study a distributed optimization algorithm by employing a stochastic dual coordinate ascent method. Stochastic dual coordinate ascent methods enjoy strong theoretical guarantees and often have better performances than stochastic gradient descent methods in optimizing regularized loss minimization problems. It still lacks of efforts in studying them in a distributed framework. We make a progress along the line by presenting a distributed stochastic dual coordinate ascent algorithm in a star network, with an analysis of the tradeoff between computation and communication. We verify our analysis by experiments on real data sets. Moreover, we compare the proposed algorithm with distributed stochastic gradient descent methods and distributed alternating direction methods of multipliers for optimizing SVMs in the same distributed framework, and observe competitive performances.

****

Projected Natural Actor-Critic

Philip S. Thomas, William C. Dabney, Stephen Giguere, Sridhar Mahadevan

Natural actor-critics are a popular class of policy search algorithms for finding locally optimal policies for Markov decision processes. In this paper we address a drawback of natural actor-critics that limits their real-world applicability - their lack of safety guarantees. We present a principled algorithm for performing natural gradient descent over a constrained domain. In the context of reinforcement learning, this allows for natural actor-critic algorithms that are guaranteed to remain within a known safe region of policy space. While deriving our class of constrained natural actor-critic algorithms, which we call Projected Natural Actor-Critics (PNACs), we also elucidate the relationship between natural gradient descent and mirror descent.

****

Minimax Optimal Algorithms for Unconstrained Linear Optimization

Brendan McMahan, Jacob Abernethy

We design and analyze minimax-optimal algorithms for online linear optimization games where the player's choice is unconstrained. The player strives to minimize regret, the difference between his loss and the loss of a post-hoc benchmark strategy. The standard benchmark is the loss of the best strategy chosen from a bounded comparator set, whereas we consider a broad range of benchmark functions. We consider the problem as a sequential multi-stage zero-sum game, and we give a thorough analysis of the minimax behavior of the game, providing characterizations for the value of the game, as well as both the player's and the adversary's optimal strategy. We show how these objects can be computed efficiently under certain circumstances, and by selecting an appropriate benchmark, we construct a novel hedging strategy for an unconstrained betting game.

****

Policy Shaping: Integrating Human Feedback with Reinforcement Learning
Shane Griffith, Kaushik Subramanian, Jonathan Scholz, Charles L. Isbell, Andrea L. Thomaz
A long term goal of Interactive Reinforcement Learning is to incorporate non-expert human feedback to solve complex tasks. State-of-the-art methods have approached this problem by mapping human information to reward and value signals to indicate preferences and then iterating over them to compute the necessary control policy. In this paper we argue for an alternate, more effective characterization of human feedback: Policy Shaping. We introduce Advise, a Bayesian approach that attempts to maximize the information gained from human feedback by utilizing it as direct labels on the policy. We compare Advise to state-of-the-art approaches and highlight scenarios where it outperforms them and importantly is robust to infrequent and inconsistent human feedback.
****
Regret based Robust Solutions for Uncertain Markov Decision Processes
Asrar Ahmed, Pradeep Varakantham, Yossiri Adulyasak, Patrick Jaillet
In this paper, we seek robust policies for uncertain Markov Decision Processes (MDPs). Most robust optimization approaches for these problems have focussed on the computation of {\em maximin} policies which maximize the value corresponding to the worst realization of the uncertainty. Recent work has proposed {\em minimax} regret as a suitable alternative to the {\em maximin} objective for robust optimization. However, existing algorithms for handling {\em minimax} regret are restricted to models with uncertainty over rewards only. We provide algorithms that employ sampling to improve across multiple dimensions: (a) Handle uncertainties over both transition and reward models; (b) Dependence of model uncertainties across state, action pairs and decision epochs; (c) Scalability and quality bounds. Finally, to demonstrate the empirical effectiveness of our sampling approaches, we provide comparisons against benchmark algorithms on two domains from literature. We also provide a Sample Average Approximation (SAA) analysis to compute a posteriori error bounds.
****
Understanding variable importances in forests of randomized trees
Gilles Louppe, Louis Wehenkel, Antonio Sutera, Pierre Geurts
Despite growing interest and practical use in various scientific areas, variable importances derived from tree-based ensemble methods are not well understood from a theoretical point of view. In this work we characterize the Mean Decrease Impurity (MDI) variable importances as measured by an ensemble of totally randomized trees in asymptotic sample and ensemble size conditions. We derive a three-level decomposition of the information jointly provided by all input variables about the output in terms of i) the MDI importance of each input variable, ii) the degree of interaction of a given input variable with the other input variables, iii) the different interaction terms of a given degree. We then show that this MDI importance of a variable is equal to zero if and only if the variable is irrelevant and that the MDI importance of a relevant variable is invariant with respect to the removal or the addition of irrelevant variables. We illustrate these properties on a simple example and discuss how they may change in the case of non-totally randomized trees such as Random Forests and Extra-Trees.
****
Linear decision rule as aspiration for simple decision heuristics
Özgür ■im■ek
Many attempts to understand the success of simple decision heuristics have examined heuristics as an approximation to a linear decision rule. This research has identified three environmental structures that aid heuristics: dominance, cumulative dominance, and noncompensatoriness. Here, we further develop these ideas and examine their empirical relevance in 51 natural environments. We find that all three structures are prevalent, making it possible for some simple rules to reach the accuracy levels of the linear decision rule using less information.
****
Adaptive Multi-Column Deep Neural Networks with Application to Robust Image Denoising

Forest Agostinelli, Michael R. Anderson, Honglak Lee
Stacked sparse denoising auto-encoders (SSDAs) have recently been shown to be su ccessful at removing noise from corrupted images. However, like most denoising t echniques, the SSDA is not robust to variation in noise types beyond what it has seen during training. We present the multi-column stacked sparse denoising auto encoder, a novel technique of combining multiple SSDAs into a multi-column SSDA (MC-SSDA) by combining the outputs of each SSDA. We eliminate the need to determ ine the type of noise, let alone its statistics, at test time. We show that good denoising performance can be achieved with a single system on a variety of diff erent noise types, including ones not seen in the training set. Additionally, we experimentally demonstrate the efficacy of MC-SSDA denoising by achieving MNIST digit error rates on denoised images at close to that of the uncorrupted images .
****
Probabilistic Movement Primitives
Alexandros Paraschos, Christian Daniel, Jan R. Peters, Gerhard Neumann
Movement Primitives (MP) are a well-established approach for representing modula r and re-usable robot movement generators. Many state-of-the-art robot learning successes are based MPs, due to their compact representation of the inherently c ontinuous and high dimensional robot movements. A major goal in robot learning i s to combine multiple MPs as building blocks in a modular control architecture t o solve complex tasks. To this effect, a MP representation has to allow for blen ding between motions, adapting to altered task variables, and co-activating mult iple MPs in parallel. We present a probabilistic formulation of the MP concept t hat maintains a distribution over trajectories. Our probabilistic approach allow s for the derivation of new operations which are essential for implementing all aforementioned properties in one framework. In order to use such a trajectory di stribution for robot movement control, we analytically derive a stochastic feedb ack controller which reproduces the given trajectory distribution. We evaluate a nd compare our approach to existing methods on several simulated as well as real robot scenarios.
****
Speedup Matrix Completion with Side Information: Application to Multi-Label Lear ning
Miao Xu, Rong Jin, Zhi-Hua Zhou
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.
****
Message Passing Inference with Chemical Reaction Networks
Nils E. Napp, Ryan P. Adams
Recent work on molecular programming has explored new possibilities for computat ional abstractions with biomolecules, including logic gates, neural networks, an d linear systems. In the future such abstractions might enable nanoscale device s that can sense and control the world at a molecular scale. Just as in macrosc ale robotics, it is critical that such devices can learn about their environment and reason under uncertainty. At this small scale, systems are typically modele d as chemical reaction networks. In this work, we develop a procedure that can t ake arbitrary probabilistic graphical models, represented as factor graphs over discrete random variables, and compile them into chemical reaction networks that implement inference. In particular, we show that marginalization based on sum-product message passing can be implemented in terms of reactions between chemica l species whose concentrations represent probabilities. We show algebraically t hat the steady state concentration of these species correspond to the marginal d istributions of the random variables in the graph and validate the results in si mulations. As with standard sum-product inference, this procedure yields exact results for tree-structured graphs, and approximate solutions for loopy graphs.
****
Binary to Bushy: Bayesian Hierarchical Clustering with the Beta Coalescent

Yuening Hu, Jordan L. Ying, Hal Daume III, Z. Irene Ying
Discovering hierarchical regularities in data is a key problem in interacting with large datasets, modeling cognition, and encoding knowledge. A previous Bayesian solution---Kingman's coalescent---provides a convenient probabilistic model for data represented as a binary tree. Unfortunately, this is inappropriate for data better described by bushier trees. We generalize an existing belief propagation framework of Kingman's coalescent to the beta coalescent, which models a wider range of tree structures. Because of the complex combinatorial search over possible structures, we develop new sampling schemes using sequential Monte Carlo and Dirichlet process mixture models, which render inference efficient and tractable. We present results on both synthetic and real data that show the beta coalescent outperforms Kingman's coalescent on real datasets and is qualitatively better at capturing data in bushy hierarchies.
****

A Stability-based Validation Procedure for Differentially Private Machine Learning
Kamalika Chaudhuri, Staal A. Vinterbo
Differential privacy is a cryptographically motivated definition of privacy which has gained considerable attention in the algorithms, machine-learning and data-mining communities. While there has been an explosion of work on differentially private machine learning algorithms, a major barrier to achieving end-to-end differential privacy in practical machine learning applications is the lack of an effective procedure for differentially private parameter tuning, or, determining the parameter value, such as a bin size in a histogram, or a regularization parameter, that is suitable for a particular application. In this paper, we introduce a generic validation procedure for differentially private machine learning algorithms that apply when a certain stability condition holds on the training algorithm and the validation performance metric. The training data size and the privacy budget used for training in our procedure is independent of the number of parameter values searched over. We apply our generic procedure to two fundamental tasks in statistics and machine-learning -- training a regularized linear classifier and building a histogram density estimator that result in end-to-end differentially private solutions for these problems.
****

Unsupervised Spectral Learning of Finite State Transducers
Raphael Bailly, Xavier Carreras, Ariadna Quattoni
Finite-State Transducers (FST) are a standard tool for modeling paired input-output sequences and are used in numerous applications, ranging from computational biology to natural language processing. Recently Balle et al. presented a spectral algorithm for learning FST from samples of aligned input-output sequences. In this paper we address the more realistic, yet challenging setting where the alignments are unknown to the learning algorithm. We frame FST learning as finding a low rank Hankel matrix satisfying constraints derived from observable statistics. Under this formulation, we provide identifiability results for FST distributions. Then, following previous work on rank minimization, we propose a regularized convex relaxation of this objective which is based on minimizing a nuclear norm penalty subject to linear constraints and can be solved efficiently.
****

One-shot learning and big data with n=2
Lee H. Dicker, Dean P. Foster
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****

Mapping paradigm ontologies to and from the brain
Yannick Schwartz, Bertrand Thirion, Gael Varoquaux
Imaging neuroscience links brain activation maps to behavior and cognition via correlational studies. Due to the nature of the individual experiments, based on eliciting neural response from a small number of stimuli, this link is incomplet

e, and unidirectional from the causal point of view. To come to conclusions on t
he function implied by the activation of brain regions, it is necessary to comb
ine a wide exploration of the various brain functions and some inversion of the
statistical inference. Here we introduce a methodology for accumulating knowledg
e towards a bidirectional link between observed brain activity and the correspon
ding function. We rely on a large corpus of imaging studies and a predictive eng
ine. Technically, the challenges are to find commonality between the studies wit
hout denaturing the richness of the corpus. The key elements that we contribute
are labeling the tasks performed with a cognitive ontology, and modeling the lon
g tail of rare paradigms in the corpus. To our knowledge, our approach is the fi
rst demonstration of predicting the cognitive content of completely new brain im
ages. To that end, we propose a method that predicts the experimental paradigms
across different studies.
****

Density estimation from unweighted k-nearest neighbor graphs: a roadmap
Ulrike Von Luxburg, Morteza Alamgir
Consider an unweighted k-nearest neighbor graph on n points that have been sam
pled i.i.d. from some unknown density p on R^d. We prove how one can estimate th
e density p just from the unweighted adjacency matrix of the graph, without know
ing the points themselves or their distance or similarity scores. The key insigh
ts are that local differences in link numbers can be used to estimate some local
 function of p, and that integrating this function along shortest paths leads to
 an estimate of the underlying density.
****

Actor-Critic Algorithms for Risk-Sensitive MDPs
Prashanth L.A., Mohammad Ghavamzadeh
In many sequential decision-making problems we may want to manage risk by minimi
zing some measure of variability in rewards in addition to maximizing a standard
 criterion. Variance related risk measures are among the most common risk-sensit
ive criteria in finance and operations research. However, optimizing many such c
riteria is known to be a hard problem. In this paper, we consider both discounte
d and average reward Markov decision processes. For each formulation, we first d
efine a measure of variability for a policy, which in turn gives us a set of ris
k-sensitive criteria to optimize. For each of these criteria, we derive a formul
a for computing its gradient. We then devise actor-critic algorithms for estimat
ing the gradient and updating the policy parameters in the ascent direction. We
establish the convergence of our algorithms to locally risk-sensitive optimal po
licies. Finally, we demonstrate the usefulness of our algorithms in a traffic si
gnal control application.
****

Probabilistic Principal Geodesic Analysis
Miaomiao Zhang, Tom Fletcher
Principal geodesic analysis (PGA) is a generalization of principal component ana
lysis (PCA) for dimensionality reduction of data on a Riemannian manifold. Curre
ntly PGA is defined as a geometric fit to the data, rather than as a probabilist
ic model. Inspired by probabilistic PCA, we present a latent variable model for
PGA that provides a probabilistic framework for factor analysis on manifolds. To
 compute maximum likelihood estimates of the parameters in our model, we develop
 a Monte Carlo Expectation Maximization algorithm, where the expectation is appr
oximated by Hamiltonian Monte Carlo sampling of the latent variables. We demonst
rate the ability of our method to recover the ground truth parameters in simulat
ed sphere data, as well as its effectiveness in analyzing shape variability of a
 corpus callosum data set from human brain images.
****

k-Prototype Learning for 3D Rigid Structures
Hu Ding, Ronald Berezney, Jinhui Xu
Requests for name changes in the electronic proceedings will be accepted with no
 questions asked. However name changes may cause bibliographic tracking issues.
  Authors are asked to consider this carefully and discuss it with their co-auth
ors prior to requesting a name change in the electronic proceedings.

****
Learning Adaptive Value of Information for Structured Prediction
David J. Weiss, Ben Taskar
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.
****
Spike train entropy-rate estimation using hierarchical Dirichlet process priors
Karin C. Knudson, Jonathan W. Pillow
Entropy rate quantifies the amount of disorder in a stochastic process. For spiking neurons, the entropy rate places an upper bound on the rate at which the spike train can convey stimulus information, and a large literature has focused on the problem of estimating entropy rate from spike train data. Here we present Bayes Least Squares and Empirical Bayesian entropy rate estimators for binary spike trains using Hierarchical Dirichlet Process (HDP) priors. Our estimator leverages the fact that the entropy rate of an ergodic Markov Chain with known transition probabilities can be calculated analytically, and many stochastic processes that are non-Markovian can still be well approximated by Markov processes of sufficient depth. Choosing an appropriate depth of Markov model presents challenges due to possibly long time dependencies and short data sequences: a deeper model can better account for long time-dependencies, but is more difficult to infer from limited data. Our approach mitigates this difficulty by using a hierarchical prior to share statistical power across Markov chains of different depths. We present both a fully Bayesian and empirical Bayes entropy rate estimator based on this model, and demonstrate their performance on simulated and real neural spike train data.
****
Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima
Po-Ling Loh, Martin J. Wainwright
We establish theoretical results concerning all local optima of various regularized M-estimators, where both loss and penalty functions are allowed to be nonconvex. Our results show that as long as the loss function satisfies restricted strong convexity and the penalty function satisfies suitable regularity conditions, any local optimum of the composite objective function lies within statistical precision of the true parameter vector. Our theory covers a broad class of nonconvex objective functions, including corrected versions of the Lasso for errors-in-variables linear models; regression in generalized linear models using nonconvex regularizers such as SCAD and MCP; and graph and inverse covariance matrix estimation. On the optimization side, we show that a simple adaptation of composite gradient descent may be used to compute a global optimum up to the statistical precision epsilon in log(1/epsilon) iterations, which is the fastest possible rate of any first-order method. We provide a variety of simulations to illustrate the sharpness of our theoretical predictions.
****
Robust Data-Driven Dynamic Programming
Grani Adiwena Hanasusanto, Daniel Kuhn
In stochastic optimal control the distribution of the exogenous noise is typically unknown and must be inferred from limited data before dynamic programming (DP)-based solution schemes can be applied. If the conditional expectations in the DP recursions are estimated via kernel regression, however, the historical sample paths enter the solution procedure directly as they determine the evaluation points of the cost-to-go functions. The resulting data-driven DP scheme is asymptotically consistent and admits efficient computational solution when combined with parametric value function approximations. If training data is sparse, however, the estimated cost-to-go functions display a high variability and an optimistic bias, while the corresponding control policies perform poorly in out-of-sample tests. To mitigate these small sample effects, we propose a robust data-driven DP scheme, which replaces the expectations in the DP recursions with worst-case

expectations over a set of distributions close to the best estimate. We show tha
t the arising min-max problems in the DP recursions reduce to tractable conic pr
ograms. We also demonstrate that this robust algorithm dominates state-of-the-ar
t benchmark algorithms in out-of-sample tests across several application domains
.
****

Provable Subspace Clustering: When LRR meets SSC
Yu-Xiang Wang, Huan Xu, Chenlei Leng
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.
****

Optimization, Learning, and Games with Predictable Sequences
Sasha Rakhlin, Karthik Sridharan
We provide several applications of Optimistic Mirror Descent, an online learning
 algorithm based on the idea of predictable sequences. First, we recover the Mir
ror-Prox algorithm, prove an extension to Holder-smooth functions, and apply the
 results to saddle-point type problems. Second, we prove that a version of Optim
istic Mirror Descent (which has a close relation to the Exponential Weights algo
rithm) can be used by two strongly-uncoupled players in a finite zero-sum matrix
 game to converge to the minimax equilibrium at the rate of O(log T / T). This a
ddresses a question of Daskalakis et al, 2011. Further, we consider a partial in
formation version of the problem. We then apply the results to approximate conve
x programming and show a simple algorithm for the approximate Max-Flow problem.
****

Beyond Pairwise: Provably Fast Algorithms for Approximate k-Way Similarity Se
arch
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-auth
ors prior to requesting a name change in the electronic proceedings.
****

Firing rate predictions in optimal balanced networks
David G. Barrett, Sophie Denève, Christian K. Machens
How are firing rates in a spiking network related to neural input, connectivity
and network function? This is an important problem because firing rates are one
of the most important measures of network activity, in both the study of neural
computation and neural network dynamics. However, it is a difficult problem, bec
ause the spiking mechanism of individual neurons is highly non-linear, and these
 individual neurons interact strongly through connectivity. We develop a new tec
hnique for calculating firing rates in optimal balanced networks. These are part
icularly interesting networks because they provide an optimal spike-based signal
 representation while producing cortex-like spiking activity through a dynamic b
alance of excitation and inhibition. We can calculate firing rates by treating b
alanced network dynamics as an algorithm for optimizing signal representation. W
e identify this algorithm and then calculate firing rates by finding the solutio
n to the algorithm. Our firing rate calculation relates network firing rates dir
ectly to network input, connectivity and function. This allows us to explain the
 function and underlying mechanism of tuning curves in a variety of systems.
****

Multi-Task Bayesian Optimization
Kevin Swersky, Jasper Snoek, Ryan P. Adams
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 multiple samples to learn mixture models

Jason D. Lee, Ran Gilad-Bachrach, Rich Caruana
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.
****

First-order Decomposition Trees
Nima Taghipour, Jesse Davis, Hendrik Blockeel
Lifting attempts to speedup probabilistic inference by exploiting symmetries in the model. Exact lifted inference methods, like their propositional counterparts, work by recursively decomposing the model and the problem. In the propositional case, there exist formal structures, such as decomposition trees (dtrees), that represent such a decomposition and allow us to determine the complexity of inference a priori. However, there is currently no equivalent structure nor analogous complexity results for lifted inference. In this paper, we introduce FO-dtrees, which upgrade propositional dtrees to the first-order level. We show how these trees can characterize a lifted inference solution for a probabilistic logical model (in terms of a sequence of lifted operations), and make a theoretical analysis of the complexity of lifted inference in terms of the novel notion of lifted width for the tree.
****

Noise-Enhanced Associative Memories
Amin Karbasi, Amir Hesam Salavati, Amin Shokrollahi, Lav R. Varshney
Recent advances in associative memory design through structured pattern sets and graph-based inference algorithms have allowed reliable learning and recall of an exponential number of patterns. Although these designs correct external errors in recall, they assume neurons that compute noiselessly, in contrast to the highly variable neurons in hippocampus and olfactory cortex. Here we consider associative memories with noisy internal computations and analytically characterize performance. As long as the internal noise level is below a specified threshold, the error probability in the recall phase can be made exceedingly small. More surprisingly, we show that internal noise actually improves the performance of the recall phase. Computational experiments lend additional support to our theoretical analysis. This work suggests a functional benefit to noisy neurons in biological neuronal networks.
****

Adaptive Submodular Maximization in Bandit Setting
Victor Gabillon, Branislav Kveton, Zheng Wen, Brian Eriksson, S. Muthukrishnan
Maximization of submodular functions has wide applications in machine learning and artificial intelligence. Adaptive submodular maximization has been traditionally studied under the assumption that the model of the world, the expected gain of choosing an item given previously selected items and their states, is known. In this paper, we study the scenario where the expected gain is initially unknown and it is learned by interacting repeatedly with the optimized function. We propose an efficient algorithm for solving our problem and prove that its expected cumulative regret increases logarithmically with time. Our regret bound captures the inherent property of submodular maximization, earlier mistakes are more costly than later ones. We refer to our approach as Optimistic Adaptive Submodular Maximization (OASM) because it trades off exploration and exploitation based on the optimism in the face of uncertainty principle. We evaluate our method on a preference elicitation problem and show that non-trivial K-step policies can be learned from just a few hundred interactions with the problem.
****

Approximate inference in latent Gaussian-Markov models from continuous time observations
Botond Cseke, Manfred Opper, Guido Sanguinetti
We propose an approximate inference algorithm for continuous time Gaussian-Markov process models with both discrete and continuous time likelihoods. We show that the continuous time limit of the expectation propagation algorithm exists and results in a hybrid fixed point iteration consisting of (1) expectation propagat

ion updates for the discrete time terms and (2) variational updates for the cont
inuous time term. We introduce corrections methods that improve on the marginals
 of the approximation. This approach extends the classical Kalman-Bucy smoothing
 procedure to non-Gaussian observations, enabling continuous-time inference in a
 variety of models, including spiking neuronal models (state-space models with p
oint process observations) and box likelihood models. Experimental results on re
al and simulated data demonstrate high distributional accuracy and significant c
omputational savings compared to discrete-time approaches in a neural applicatio
n.
****

Lexical and Hierarchical Topic Regression
Viet-An Nguyen, Jordan L. Ying, Philip Resnik
Inspired by a two-level theory that unifies agenda setting and ideological fram
ing, we propose supervised hierarchical latent Dirichlet allocation (SHLDA) whi
ch jointly captures documents' multi-level topic structure and their polar resp
onse variables. Our model extends the nested Chinese restaurant process to disc
over a tree-structured topic hierarchy and uses both per-topic hierarchical and
 per-word lexical regression parameters to model the response variables. Experi
ments in a political domain and on sentiment analysis tasks show that SHLDA imp
roves predictive accuracy while adding a new dimension of insight into how topi
cs under discussion are framed.
****

Adaptive Step-Size for Policy Gradient Methods
Matteo Pirotta, Marcello Restelli, Luca Bascetta
In the last decade, policy gradient methods have significantly grown in populari
ty in the reinforcement--learning field. In particular, they have been largely e
mployed in motor control and robotic applications, thanks to their ability to co
pe with continuous state and action domains and partial observable problems. Pol
icy gradient researches have been mainly focused on the identification of effect
ive gradient directions and the proposal of efficient estimation algorithms. Non
etheless, the performance of policy gradient methods is determined not only by t
he gradient direction, since convergence properties are strongly influenced by t
he choice of the step size: small values imply slow convergence rate, while larg
e values may lead to oscillations or even divergence of the policy parameters. S
tep--size value is usually chosen by hand tuning and still little attention has
been paid to its automatic selection. In this paper, we propose to determine the
 learning rate by maximizing a lower bound to the expected performance gain. Foc
using on Gaussian policies, we derive a lower bound that is second--order polyno
mial of the step size, and we show how a simplified version of such lower bound
can be maximized when the gradient is estimated from trajectory samples. The pro
perties of the proposed approach are empirically evaluated in a linear--quadrati
c regulator problem.
****

Mixed Optimization for Smooth Functions
Mehrdad Mahdavi, Lijun Zhang, Rong Jin
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 Neural Networks for Object Detection
Christian Szegedy, Alexander Toshev, Dumitru Erhan
Deep Neural Networks (DNNs) have recently shown outstanding performance on the t
ask of whole image classification. In this paper we go one step further and addr
ess the problem of object detection -- not only classifying but also precisely l
ocalizing objects of various classes using DNNs. We present a simple and yet pow
erful formulation of object detection as a regression to object masks. We define
 a multi-scale inference procedure which is able to produce a high-resolution ob
ject detection at a low cost by a few network applications. The approach achieve
s state-of-the-art performance on Pascal 2007 VOC.

****
A simple example of Dirichlet process mixture inconsistency for the number of co
mponents
Jeffrey W. Miller, Matthew T. Harrison
For data assumed to come from a finite mixture with an unknown number of compone
nts, it has become common to use Dirichlet process mixtures (DPMs) not only for
density estimation, but also for inferences about the number of components. The
typical approach is to use the posterior distribution on the number of component
s occurring so far --- that is, the posterior on the number of clusters in the o
bserved data. However, it turns out that this posterior is not consistent --- it
does not converge to the true number of components. In this note, we give an el
ementary demonstration of this inconsistency in what is perhaps the simplest pos
sible setting: a DPM with normal components of unit variance, applied to data fr
om a mixture" with one standard normal component. Further, we find that this exa
mple exhibits severe inconsistency: instead of going to 1, the posterior probabi
lity that there is one cluster goes to 0."
****
Learning Kernels Using Local Rademacher Complexity
Corinna Cortes, Marius Kloft, Mehryar Mohri
We use the notion of local Rademacher complexity to design new algorithms for le
arning kernels. Our algorithms thereby benefit from the sharper learning bounds
based on that notion which, under certain general conditions, guarantee a faster
convergence rate. We devise two new learning kernel algorithms: one based on a
convex optimization problem for which we give an efficient solution using exist
ing learning kernel techniques, and another one that can be formulated as a DC-p
rogramming problem for which we describe a solution in detail. We also report th
e results of experiments with both algorithms in both binary and multi-class cla
ssification tasks.
****
Bayesian entropy estimation for binary spike train data using parametric prior k
nowledge
Evan W. Archer, Il Memming Park, Jonathan W. Pillow
Shannon's entropy is a basic quantity in information theory, and a fundamental
building block for the analysis of neural codes. Estimating the entropy of a d
iscrete distribution from samples is an important and difficult problem that has
received considerable attention in statistics and theoretical neuroscience.
However, neural responses have characteristic statistical structure that gene
ric entropy estimators fail to exploit. For example, existing Bayesian entropy
estimators make the naive assumption that all spike words are equally likely
a priori, which makes for an inefficient allocation of prior probability mass i
n cases where spikes are sparse. Here we develop Bayesian estimators for the
entropy of binary spike trains using priors designed to flexibly exploit the
statistical structure of simultaneously-recorded spike responses. We define tw
o prior distributions over spike words using mixtures of Dirichlet distributio
ns centered on simple parametric models. The parametric model captures high-le
vel statistical features of the data, such as the average spike count in a spi
ke word, which allows the posterior over entropy to concentrate more rapidly
than with standard estimators (e.g., in cases where the probability of spiking
differs strongly from 0.5). Conversely, the Dirichlet distributions assign pri
or mass to distributions far from the parametric model, ensuring consistent est
imates for arbitrary distributions. We devise a compact representation of the
data and prior that allow for computationally efficient implementations of B
ayesian least squares and empirical Bayes entropy estimators with large numbers
of neurons. We apply these estimators to simulated and real neural data and
show that they substantially outperform traditional methods.
****
Mid-level Visual Element Discovery as Discriminative Mode Seeking
Carl Doersch, Abhinav Gupta, Alexei A. Efros
Recent work on mid-level visual representations aims to capture information at t
he level of complexity higher than typical visual words", but lower than full-bl

own semantic objects. Several approaches have been proposed to discover mid-leve
l visual elements, that are both 1) representative, i.e. frequently occurring wi
thin a visual dataset, and 2) visually discriminative. However, the current appr
oaches are rather ad hoc and difficult to analyze and evaluate. In this work, we
 pose visual element discovery as discriminative mode seeking, drawing connectio
ns to the the well-known and well-studied mean-shift algorithm. Given a weakly-l
abeled image collection, our method discovers visually-coherent patch clusters t
hat are maximally discriminative with respect to the labels. One advantage of ou
r formulation is that it requires only a single pass through the data. We also p
ropose the Purity-Coverage plot as a principled way of experimentally analyzing
and evaluating different visual discovery approaches, and compare our method aga
inst prior work on the Paris Street View dataset. We also evaluate our method on
 the task of scene classification, demonstrating state-of-the-art performance on
 the MIT Scene-67 dataset."
****

Approximate Bayesian Image Interpretation using Generative Probabilistic Graphic
s Programs

Vikash K. Mansinghka, Tejas D. Kulkarni, Yura N. Perov, Josh Tenenbaum

The idea of computer vision as the Bayesian inverse problem to computer graphics
 has a long history and an appealing elegance, but it has proved difficult to di
rectly implement. Instead, most vision tasks are approached via complex bottom-u
p processing pipelines. Here we show that it is possible to write short, simple
probabilistic graphics programs that define flexible generative models and to au
tomatically invert them to interpret real-world images. Generative probabilistic
 graphics programs consist of a stochastic scene generator, a renderer based on
graphics software, a stochastic likelihood model linking the renderer's output a
nd the data, and latent variables that adjust the fidelity of the renderer and t
he tolerance of the likelihood model. Representations and algorithms from comput
er graphics, originally designed to produce high-quality images, are instead use
d as the deterministic backbone for highly approximate and stochastic generative
 models. This formulation combines probabilistic programming, computer graphics,
 and approximate Bayesian computation, and depends only on general-purpose, auto
matic inference techniques. We describe two applications: reading sequences of d
egraded and adversarially obscured alphanumeric characters, and inferring 3D roa
d models from vehicle-mounted camera images. Each of the probabilistic graphics
programs we present relies on under 20 lines of probabilistic code, and supports
 accurate, approximately Bayesian inferences about ambiguous real-world images.
****

Annealing between distributions by averaging moments

Roger B. Grosse, Chris J. Maddison, Russ R. Salakhutdinov

Many powerful Monte Carlo techniques for estimating partition functions, such as
 annealed importance sampling (AIS), are based on sampling from a sequence of in
termediate distributions which interpolate between a tractable initial distribut
ion and an intractable target distribution. The near-universal practice is to us
e geometric averages of the initial and target distributions, but alternative pa
ths can perform substantially better. We present a novel sequence of intermediat
e distributions for exponential families: averaging the moments of the initial a
nd target distributions. We derive an asymptotically optimal piecewise linear sc
hedule for the moments path and show that it performs at least as well as geomet
ric averages with a linear schedule. Moment averaging performs well empirically
 at estimating partition functions of restricted Boltzmann machines (RBMs), whic
h form the building blocks of many deep learning models, including Deep Belief N
etworks and Deep Boltzmann Machines.
****

Blind Calibration in Compressed Sensing using Message Passing Algorithms

Christophe Schulke, Francesco Caltagirone, Florent Krzakala, Lenka Zdeborová

Compressed sensing (CS) is a concept that allows to acquire compressible signals
 with a small number of measurements. As such, it is very attractive for hardwar
e implementations. Therefore, correct calibration of the hardware is a central i
ssue. In this paper we study the so-called blind calibration, i.e. when the trai

ning signals that are available to perform the calibration are sparse but unknown. We extend the approximate message passing (AMP) algorithm used in CS to the case of blind calibration. In the calibration-AMP, both the gains on the sensors and the elements of the signals are treated as unknowns. Our algorithm is also applicable to settings in which the sensors distort the measurements in other ways than multiplication by a gain, unlike previously suggested blind calibration algorithms based on convex relaxations. We study numerically the phase diagram of the blind calibration problem, and show that even in cases where convex relaxation is possible, our algorithm requires a smaller number of measurements and/or signals in order to perform well.
****

Active Learning for Probabilistic Hypotheses Using the Maximum Gibbs Error Criterion
Nguyen Viet Cuong, Wee Sun Lee, Nan Ye, Kian Ming A. Chai, Hai Leong Chieu
We introduce a new objective function for pool-based Bayesian active learning with probabilistic hypotheses. This objective function, called the policy Gibbs error, is the expected error rate of a random classifier drawn from the prior distribution on the examples adaptively selected by the active learning policy. Exact maximization of the policy Gibbs error is hard, so we propose a greedy strategy that maximizes the Gibbs error at each iteration, where the Gibbs error on an instance is the expected error of a random classifier selected from the posterior label distribution on that instance. We apply this maximum Gibbs error criterion to three active learning scenarios: non-adaptive, adaptive, and batch active learning. In each scenario, we prove that the criterion achieves near-maximal policy Gibbs error when constrained to a fixed budget. For practical implementations, we provide approximations to the maximum Gibbs error criterion for Bayesian conditional random fields and transductive Naive Bayes. Our experimental results on a named entity recognition task and a text classification task show that the maximum Gibbs error criterion is an effective active learning criterion for noisy models.
****

Two-Target Algorithms for Infinite-Armed Bandits with Bernoulli Rewards
Thomas Bonald, Alexandre Proutiere
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****

Learning to Prune in Metric and Non-Metric Spaces
Leonid Boytsov, Bilegsaikhan Naidan
Our focus is on approximate nearest neighbor retrieval in metric and non-metric spaces. We employ a VP-tree and explore two simple yet effective learning-to prune approaches: density estimation through sampling and “stretching” of the triangle inequality. Both methods are evaluated using data sets with metric (Euclidean) and non-metric (KL-divergence and Itakura-Saito) distance functions. Conditions on spaces where the VP-tree is applicable are discussed. The VP-tree with a learned pruner is compared against the recently proposed state-of-the-art approaches: the bbtree, the multi-probe locality sensitive hashing (LSH), and permutation methods. Our method was competitive against state-of-the-art methods and, in most cases, was more efficient for the same rank approximation quality.
****

Learning from Limited Demonstrations
Beomjoon Kim, Amir-massoud Farahmand, Joelle Pineau, Doina Precup
We propose an approach to learning from demonstration (LfD) which leverages expert data, even if the expert examples are very few or inaccurate. We achieve this by integrating LfD in an approximate policy iteration algorithm. The key idea of our approach is that expert examples are used to generate linear constraints on the optimization, in a similar fashion to large-margin classification. We prove an upper bound on the true Bellman error of the approximation computed by the algorithm at each iteration. We show empirically that the algorithm outperfo

rms both pure policy iteration, as well as DAgger (a state-of-art LfD algorithm) and supervised learning in a variety of scenarios, including when very few and/or imperfect demonstrations are available. Our experiments include simulations as well as a real robotic navigation task.
****

Aggregating Optimistic Planning Trees for Solving Markov Decision Processes
Gunnar Kedenburg, Raphael Fonteneau, Remi Munos
This paper addresses the problem of online planning in Markov Decision Processes using only a generative model. We propose a new algorithm which is based on the construction of a forest of single successor state planning trees. For every explored state-action, such a tree contains exactly one successor state, drawn from the generative model. The trees are built using a planning algorithm which follows the optimism in the face of uncertainty principle, in assuming the most favorable outcome in the absence of further information. In the decision making step of the algorithm, the individual trees are combined. We discuss the approach, prove that our proposed algorithm is consistent, and empirically show that it performs better than a related algorithm which additionally assumes the knowledge of all transition distributions.
****

Action from Still Image Dataset and Inverse Optimal Control to Learn Task Specific Visual Scanpaths
Stefan Mathe, Cristian Sminchisescu
Human eye movements provide a rich source of information into the human visual processing. The complex interplay between the task and the visual stimulus is believed to determine human eye movements, yet it is not fully understood. This has precluded the development of reliable dynamic eye movement prediction systems. Our work makes three contributions towards addressing this problem. First, we complement one of the largest and most challenging static computer vision datasets, VOC 2012 Actions, with human eye movement annotations collected under the task constraints of action and context recognition. Our dataset is unique among eyetracking datasets for still images in terms of its large scale (over 1 million fixations, 9157 images), task control and action from a single image emphasis. Second, we introduce models to automatically discover areas of interest (AOI) and introduce novel dynamic consistency metrics, based on them. Our method can automatically determine the number and spatial support of the AOIs, in addition to their locations. Based on such encodings, we show that, on unconstrained read-world stimuli, task instructions have significant influence on visual behavior. Finally, we leverage our large scale dataset in conjunction with powerful machine learning techniques and computer vision features, to introduce novel dynamic eye movement prediction methods which learn task-sensitive reward functions from eye movement data and efficiently integrate these rewards to plan future saccades based on inverse optimal control. We show that the propose methodology achieves state of the art scanpath modeling results.
****

How to Hedge an Option Against an Adversary: Black-Scholes Pricing is Minimax Optimal
Jacob Abernethy, Peter L. Bartlett, Rafael Frongillo, Andre Wibisono
We consider a popular problem in finance, option pricing, through the lens of an online learning game between Nature and an Investor. In the Black-Scholes option pricing model from 1973, the Investor can continuously hedge the risk of an option by trading the underlying asset, assuming that the asset's price fluctuates according to Geometric Brownian Motion (GBM). We consider a worst-case model, in which Nature chooses a sequence of price fluctuations under a cumulative quadratic volatility constraint, and the Investor can make a sequence of hedging decisions. Our main result is to show that the value of our proposed game, which is the regret'' of hedging strategy, converges to the Black-Scholes option price. We use significantly weaker assumptions than previous work---for instance, we allow large jumps in the asset price---and show that the Black-Scholes hedging strategy is near-optimal for the Investor even in this non-stochastic framework."
****