Topology Constraints in Graphical Models
Marcelo Fiori, Pablo Musé, Guillermo Sapiro
Graphical models are a very useful tool to describe and understand natural pheno mena, from gene expression to climate change and social interactions. The topolo gical structure of these graphs/networks is a fundamental part of the analysis, and in many cases the main goal of the study. However, little work has been done on incorporating prior topological knowledge onto the estimation of the underly ing graphical models from sample data. In this work we propose extensions to the basic joint regression model for network estimation, which explicitly incorpora te graph-topological constraints into the corresponding optimization approach. T he first proposed extension includes an eigenvector centrality constraint, there by promoting this important prior topological property. The second developed ext ension promotes the formation of certain motifs, triangle-shaped ones in particu lar, which are known to exist for example in genetic regulatory networks. The pr esentation of the underlying formulations, which serve as examples of the introd uction of topological constraints in network estimation, is complemented with ex amples in diverse datasets demonstrating the importance of incorporating such cr itical prior knowledge.
****

Clustering Aggregation as Maximum-Weight Independent Set
Nan Li, Longin Latecki
We formulate clustering aggregation as a special instance of Maximum-Weight Inde pendent Set (MWIS) problem. For a given dataset, an attributed graph is construc ted from the union of the input clusterings generated by different underlying cl ustering algorithms with different parameters. The vertices, which represent the distinct clusters, are weighted by an internal index measuring both cohesion an d separation. The edges connect the vertices whose corresponding clusters overla p. Intuitively, an optimal aggregated clustering can be obtained by selecting an optimal subset of non-overlapping clusters partitioning the dataset together. W e formalize this intuition as the MWIS problem on the attributed graph, i.e., fi nding the heaviest subset of mutually non-adjacent vertices. This MWIS problem exhibits a special structure. Since the clusters of each input clustering form a partition of the dataset, the vertices corresponding to each clustering form a maximal independent set (MIS) in the attributed graph. We propose a variant of s imulated annealing method that takes advantage of this special structure. Our al gorithm starts from each MIS, which is close to a distinct local optimum of the MWIS problem, and utilizes a local search heuristic to explore its neighborhood in order to find the MWIS. Extensive experiments on many challenging datasets sh ow that: 1. our approach to clustering aggregation automatically decides the opt imal number of clusters; 2. it does not require any parameter tuning for the und erlying clustering algorithms; 3. it can combine the advantages of different und erlying clustering algorithms to achieve superior performance; 4. it is robust a gainst moderate or even bad input clusterings.
****

FastEx: Hash Clustering with Exponential Families
Amr Ahmed, Sujith Ravi, Alex Smola, Shravan Narayanamurthy
Requests for name changes in the electronic proceedings will be accepted with 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 Bethe Partition Function of Log-supermodular Graphical Models
Nicholas Ruozzi
Sudderth, Wainwright, and Willsky conjectured that the Bethe approximation corre sponding to any fixed point of the belief propagation algorithm over an attracti ve, pairwise binary graphical model provides a lower bound on the true partition function. In this work, we resolve this conjecture in the affirmative by demons trating that, for any graphical model with binary variables whose potential func tions (not necessarily pairwise) are all log-supermodular, the Bethe partition f unction always lower bounds the true partition function. The proof of this resu

lt follows from a new variant of the “four functions” theorem that may be of ind
ependent interest.
****

Selective Labeling via Error Bound Minimization
Quanquan Gu, Tong Zhang, Jiawei Han, Chris Ding
In many practical machine learning problems, the acquisition of labeled data is
often expensive and/or time consuming. This motivates us to study a problem as f
ollows: given a label budget, how to select data points to label such that the l
earning performance is optimized. We propose a selective labeling method by anal
yzing the generalization error of Laplacian regularized Least Squares (LapRLS).
In particular, we derive a deterministic generalization error bound for LapRLS t
rained on subsampled data, and propose to select a subset of data points to labe
l by minimizing this upper bound. Since the minimization is a combinational prob
lem, we relax it into continuous domain and solve it by projected gradient desce
nt. Experiments on benchmark datasets show that the proposed method outperforms
the state-of-the-art methods.
****

Practical Bayesian Optimization of Machine Learning Algorithms
Jasper Snoek, Hugo Larochelle, Ryan P. Adams
The use of machine learning algorithms frequently involves careful tuning of lea
rning parameters and model hyperparameters. Unfortunately, this tuning is often
a “black art” requiring expert experience, rules of thumb, or sometimes brute-fo
rce search. There is therefore great appeal for automatic approaches that can op
timize the performance of any given learning algorithm to the problem at hand. I
n this work, we consider this problem through the framework of Bayesian optimiza
tion, in which a learning algorithm’s generalization performance is modeled as a
 sample from a Gaussian process (GP). We show that certain choices for the natur
e of the GP, such as the type of kernel and the treatment of its hyperparameters
, can play a crucial role in obtaining a good optimizer that can achieve expert-
level performance. We describe new algorithms that take into account the variabl
e cost (duration) of learning algorithm experiments and that can leverage the pr
esence of multiple cores for parallel experimentation. We show that these propos
ed algorithms improve on previous automatic procedures and can reach or surpass
human expert-level optimization for many algorithms including Latent Dirichlet A
llocation, Structured SVMs and convolutional neural networks.
****

Optimal Neural Tuning Curves for Arbitrary Stimulus Distributions: Discrimax, In
fomax and Minimum L_p Loss
Zhuo Wang, Alan A. Stocker, Daniel D. Lee
Requests for name changes in the electronic proceedings will be accepted with 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.
****

Probabilistic Event Cascades for Alzheimer's disease
Jonathan Huang, Daniel Alexander
Accurate and detailed models of the progression of neurodegenerative diseases su
ch as Alzheimer's (AD) are crucially important for reliable early diagnosis and
 the determination and deployment of effective treatments. In this paper, we int
roduce the ALPACA (Alzheimer's disease Probabilistic Cascades) model, a generati
ve model linking latent Alzheimer's progression dynamics to observable biomarker
 data. In contrast with previous works which model disease progression as a fixe
d ordering of events, we explicitly model the variability over such orderings am
ong patients which is more realistic, particularly for highly detailed disease p
rogression models. We describe efficient learning algorithms for ALPACA and disc
uss promising experimental results on a real cohort of Alzheimer's patients from
 the Alzheimer's Disease Neuroimaging Initiative.
****

Super-Bit Locality-Sensitive Hashing
Jianqiu Ji, Jianmin Li, Shuicheng Yan, Bo Zhang, Qi 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.
****

Bayesian nonparametric models for bipartite graphs
Francois Caron
We develop a novel Bayesian nonparametric model for random bipartite graphs. The model is based on the theory of completely random measures and is able to handle a potentially infinite number of nodes. We show that the model has appealing properties and in particular it may exhibit a power-law behavior. We derive a posterior characterization, an Indian Buffet-like generative process for network growth, and a simple and efficient Gibbs sampler for posterior simulation. Our model is shown to be well fitted to several real-world social networks.
****

Learning Label Trees for Probabilistic Modelling of Implicit Feedback
Andriy Mnih, Yee Teh
User preferences for items can be inferred from either explicit feedback, such as item ratings, or implicit feedback, such as rental histories. Research in collaborative filtering has concentrated on explicit feedback, resulting in the development of accurate and scalable models. However, since explicit feedback is often difficult to collect it is important to develop effective models that take advantage of the more widely available implicit feedback. We introduce a probabilistic approach to collaborative filtering with implicit feedback based on modelling the user's item selection process. In the interests of scalability, we restrict our attention to tree-structured distributions over items and develop a principled and efficient algorithm for learning item trees from data. We also identify a problem with a widely used protocol for evaluating implicit feedback models and propose a way of addressing it using a small quantity of explicit feedback data.
****

Factoring nonnegative matrices with linear programs
Ben Recht, Christopher Re, Joel Tropp, Victor Bittorf
This paper describes a new approach for computing nonnegative matrix factorizations (NMFs) with linear programming. The key idea is a data-driven model for the factorization, in which the most salient features in the data are used to express the remaining features. More precisely, given a data matrix X, the algorithm identifies a matrix C that satisfies X = CX and some linear constraints. The matrix C selects features, which are then used to compute a low-rank NMF of X. A theoretical analysis demonstrates that this approach has the same type of guarantees as the recent NMF algorithm of Arora et al.~(2012). In contrast with this earlier work, the proposed method has (1) better noise tolerance, (2) extends to more general noise models, and (3) leads to efficient, scalable algorithms. Experiments with synthetic and real datasets provide evidence that the new approach is also superior in practice. An optimized C++ implementation of the new algorithm can factor a multi-Gigabyte matrix in a matter of minutes.
****

Privacy Aware Learning
Martin J. Wainwright, Michael Jordan, John C. Duchi
We study statistical risk minimization problems under a version of privacy in which the data is kept confidential even from the learner. In this local privacy framework, we show sharp upper and lower bounds on the convergence rates of statistical estimation procedures. As a consequence, we exhibit a precise tradeoff between the amount of privacy the data preserves and the utility, measured by convergence rate, of any statistical estimator.
****

Truncation-free Online Variational Inference for Bayesian Nonparametric Models
Chong Wang, David Blei
We present a truncation-free online variational inference algorithm for Bayesian nonparametric models. Unlike traditional (online) variational inference algorit

hms that require truncations for the model or the variational distribution, our method adapts model complexity on the fly. Our experiments for Dirichlet process mixture models and hierarchical Dirichlet process topic models on two large-scale data sets show better performance than previous online variational inference algorithms.
****

Provable ICA with Unknown Gaussian Noise, with Implications for Gaussian Mixtures and Autoencoders
Sanjeev Arora, Rong Ge, Ankur Moitra, Sushant Sachdeva
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****

Learning Image Descriptors with the Boosting-Trick
Tomasz Trzcinski, Mario Christoudias, Vincent Lepetit, Pascal Fua
In this paper we apply boosting to learn complex non-linear local visual feature representations, drawing inspiration from its successful application to visual object detection. The main goal of local feature descriptors is to distinctively represent a salient image region while remaining invariant to viewpoint and illumination changes. This representation can be improved using machine learning, however, past approaches have been mostly limited to learning linear feature mappings in either the original input or a kernelized input feature space. While kernelized methods have proven somewhat effective for learning non-linear local feature descriptors, they rely heavily on the choice of an appropriate kernel function whose selection is often difficult and non-intuitive. We propose to use the boosting-trick to obtain a non-linear mapping of the input to a high-dimensional feature space. The non-linear feature mapping obtained with the boosting-trick is highly intuitive. We employ gradient-based weak learners resulting in a learned descriptor that closely resembles the well-known SIFT. As demonstrated in our experiments, the resulting descriptor can be learned directly from intensity patches achieving state-of-the-art performance.
****

A latent factor model for highly multi-relational data
Rodolphe Jenatton, Nicolas Roux, Antoine Bordes, Guillaume R. Obozinski
Many data such as social networks, movie preferences or knowledge bases are multi-relational, in that they describe multiple relationships between entities. While there is a large body of work focused on modeling these data, few considered modeling these multiple types of relationships jointly. Further, existing approaches tend to breakdown when the number of these types grows. In this paper, we propose a method for modeling large multi-relational datasets, with possibly thousands of relations. Our model is based on a bilinear structure, which captures the various orders of interaction of the data, but also shares sparse latent factors across different relations. We illustrate the performance of our approach on standard tensor-factorization datasets where we attain, or outperform, state-of-the-art results. Finally, a NLP application demonstrates our scalability and the ability of our model to learn efficient, and semantically meaningful verb representations.
****

Bayesian estimation of discrete entropy with mixtures of stick-breaking priors
Evan Archer, Il Memming Park, Jonathan Pillow
We consider the problem of estimating Shannon's entropy H in the under-sampled regime, where the number of possible symbols may be unknown or countably infinite. Pitman-Yor processes (a generalization of Dirichlet processes) provide tractable prior distributions over the space of countably infinite discrete distributions, and have found major applications in Bayesian non-parametric statistics and machine learning. Here we show that they also provide natural priors for Bayesian entropy estimation, due to the remarkable fact that the moments of the induced posterior distribution over H can be computed analytically.

We derive formulas for the posterior mean (Bayes' least squares estimate) and variance under such priors. Moreover, we show that a fixed Dirichlet or Pitman-Yor process prior implies a narrow prior on H, meaning the prior strongly determines the entropy estimate in the under-sampled regime. We derive a family of continuous mixing measures such that the resulting mixture of Pitman-Yor processes produces an approximately flat (improper) prior over H. We explore the theoretical properties of the resulting estimator, and show that it performs well on data sampled from both exponential and power-law tailed distributions.

****

Timely Object Recognition

Sergey Karayev, Tobias Baumgartner, Mario Fritz, Trevor Darrell

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

****

Efficient high dimensional maximum entropy modeling via symmetric partition functions

Paul Vernaza, Drew Bagnell

The application of the maximum entropy principle to sequence modeling has been popularized by methods such as Conditional Random Fields (CRFs). However, these approaches are generally limited to modeling paths in discrete spaces of low dimensionality. We consider the problem of modeling distributions over paths in continuous spaces of high dimensionality---a problem for which inference is generally intractable. Our main contribution is to show that maximum entropy modeling of high-dimensional, continuous paths is tractable as long as the constrained features possess a certain kind of low dimensional structure. In this case, we show that the associated {\em partition function} is symmetric and that this symmetry can be exploited to compute the partition function efficiently in a compressed form. Empirical results are given showing an application of our method to maximum entropy modeling of high dimensional human motion capture data.

****

Topic-Partitioned Multinetwork Embeddings

Peter Krafft, Juston Moore, Bruce Desmarais, Hanna Wallach

We introduce a joint model of network content and context designed for exploratory analysis of email networks via visualization of topic-specific communication patterns. Our model is an admixture model for text and network attributes which uses multinomial distributions over words as mixture components for explaining text and latent Euclidean positions of actors as mixture components for explaining network attributes. We validate the appropriateness of our model by achieving state-of-the-art performance on a link prediction task and by achieving semantic coherence equivalent to that of latent Dirichlet allocation. We demonstrate the capability of our model for descriptive, explanatory, and exploratory analysis by investigating the inferred topic-specific communication patterns of a new government email dataset, the New Hanover County email corpus.

****

Recovery of Sparse Probability Measures via Convex Programming

Mert Pilanci, Laurent Ghaoui, Venkat Chandrasekaran

We consider the problem of cardinality penalized optimization of a convex function over the probability simplex with additional convex constraints. It's well-known that the classical L1 regularizer fails to promote sparsity on the probability simplex since L1 norm on the probability simplex is trivially constant. We propose a direct relaxation of the minimum cardinality problem and show that it can be efficiently solved using convex programming. As a first application we consider recovering a sparse probability measure given moment constraints, in which our formulation becomes linear programming, hence can be solved very efficiently. A sufficient condition for exact recovery of the minimum cardinality solution is derived for arbitrary affine constraints. We then develop a penalized version

for the noisy setting which can be solved using second order cone programs. The proposed method outperforms known heuristics based on L1 norm. As a second application we consider convex clustering using a sparse Gaussian mixture and compare our results with the well known soft k-means algorithm.
****

Proximal Newton-type methods for convex optimization
Jason D. Lee, Yuekai Sun, Michael Saunders
We seek to solve convex optimization problems in composite form:
****

Learning visual motion in recurrent neural networks
Marius Pachitariu, Maneesh Sahani
We present a dynamic nonlinear generative model for visual motion based on a latent representation of binary-gated Gaussian variables. Trained on sequences of images, the model learns to represent different movement directions in different variables. We use an online approximate-inference scheme that can be mapped to the dynamics of networks of neurons. Probed with drifting grating stimuli and moving bars of light, neurons in the model show patterns of responses analogous to those of direction-selective simple cells in primary visual cortex. Most model neurons also show speed tuning and respond equally well to a range of motion directions and speeds aligned to the constraint line of their respective preferred speed. We show how these computations are enabled by a specific pattern of recurrent connections learned by the model.
****

Graphical Models via Generalized Linear Models
Eunho Yang, Genevera Allen, Zhandong Liu, Pradeep Ravikumar
Undirected graphical models, or Markov networks, such as Gaussian graphical models and Ising models enjoy popularity in a variety of applications. In many settings, however, data may not follow a Gaussian or binomial distribution assumed by these models. We introduce a new class of graphical models based on generalized linear models (GLM) by assuming that node-wise conditional distributions arise from exponential families. Our models allow one to estimate networks for a wide class of exponential distributions, such as the Poisson, negative binomial, and exponential, by fitting penalized GLMs to select the neighborhood for each node. A major contribution of this paper is the rigorous statistical analysis showing that with high probability, the neighborhood of our graphical models can be recovered exactly. We provide examples of high-throughput genomic networks learned via our GLM graphical models for multinomial and Poisson distributed data.
****

Searching for objects driven by context
Bogdan Alexe, Nicolas Heess, Yee Teh, Vittorio Ferrari
The dominant visual search paradigm for object class detection is sliding windows. Although simple and effective, it is also wasteful, unnatural and rigidly hardwired. We propose strategies to search for objects which intelligently explore the space of windows by making sequential observations at locations decided based on previous observations. Our strategies adapt to the class being searched and to the content of a particular test image. Their driving force is exploiting context as the statistical relation between the appearance of a window and its location relative to the object, as observed in the training set. In addition to being more elegant than sliding windows, we demonstrate experimentally on the PASCAL VOC 2010 dataset that our strategies evaluate two orders of magnitude fewer windows while at the same time achieving higher detection accuracy.
****

Learning Mixtures of Tree Graphical Models
Anima Anandkumar, Daniel J. Hsu, Furong Huang, Sham M. Kakade
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****

Convex Multi-view Subspace Learning

Martha White, Xinhua Zhang, Dale Schuurmans, Yao-liang Yu
Subspace learning seeks a low dimensional representation of data that enables ac curate reconstruction. However, in many applications, data is obtained from mul tiple sources rather than a single source (e.g. an object might be viewed by cam eras at different angles, or a document might consist of text and images). The conditional independence of separate sources imposes constraints on their shared latent representation, which, if respected, can improve the quality of the lear ned low dimensional representation. In this paper, we present a convex formulat ion of multi-view subspace learning that enforces conditional independence while reducing dimensionality. For this formulation, we develop an efficient algorit hm that recovers an optimal data reconstruction by exploiting an implicit convex regularizer, then recovers the corresponding latent representation and reconstr uction model, jointly and optimally. Experiments illustrate that the proposed m ethod produces high quality results.
****
Learning Invariant Representations of Molecules for Atomization Energy Predictio n
Grégoire Montavon, Katja Hansen, Siamac Fazli, Matthias Rupp, Franziska Biegler, Andreas Ziehe, Alexandre Tkatchenko, Anatole Lilienfeld, Klaus-Robert Müller
The accurate prediction of molecular energetics in chemical compound space is a crucial ingredient for rational compound design. The inherently graph-like, non -vectorial nature of molecular data gives rise to a unique and difficult machine learning problem. In this paper, we adopt a learning-from-scratch approach wher e quantum-mechanical molecular energies are predicted directly from the raw mole cular geometry. The study suggests a benefit from setting flexible priors and en forcing invariance stochastically rather than structurally. Our results improve the state-of-the-art by a factor of almost three, bringing statistical methods o ne step closer to the holy grail of ''chemical accuracy''.
****
On Multilabel Classification and Ranking with Partial Feedback
Claudio Gentile, 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.
****
A dynamic excitatory-inhibitory network in a VLSI chip for spiking information r eregistrations
Juan Huo
Requests for name changes in the electronic proceedings will be accepted with 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 Non-Stochastic Experts
Varun Kanade, Zhenming Liu, Bozidar Radunovic
We consider the online distributed non-stochastic experts problem, where the dis tributed system consists of one coordinator node that is connected to k sites, a nd the sites are required to communicate with each other via the coordinator. At each time-step t, one of the k site nodes has to pick an expert from the set {1 , . . . , n}, and the same site receives information about payoffs of all expert s for that round. The goal of the distributed system is to minimize regret at ti me horizon T, while simultaneously keeping communication to a minimum. The two e xtreme solutions to this problem are: (i) Full communication: This essentially s imulates the non-distributed setting to obtain the optimal O(\sqrt{log(n)T}) reg ret bound at the cost of T communication. (ii) No communication: Each site runs an independent copy – the regret is O(\sqrt{log(n)kT}) and the communication is 0. This paper shows the difficulty of simultaneously achieving regret asymptotic ally better than \sqrt{kT} and communication better than T. We give a novel algo rithm that for an oblivious adversary achieves a non-trivial trade-off: regret O

(\sqrt{k^{5(1+\epsilon)/6} T}) and communication O(T/k^\epsilon), for any value of \epsilon in (0, 1/5). We also consider a variant of the model, where the coordinator picks the expert. In this model, we show that the label-efficient forecaster of Cesa-Bianchi et al. (2005) already gives us strategy that is near optimal in regret vs communication trade-off.
****

Deep Learning of Invariant Features via Simulated Fixations in Video
Will Zou, Shenghuo Zhu, Kai Yu, Andrew Ng
We apply salient feature detection and tracking in videos to simulate ■xations and smooth pursuit in human vision. With tracked sequences as input, a hierarchical network of modules learns invariant features using a temporal slowness constraint. The network encodes invariance which are increasingly complex with hierarchy. Although learned from videos, our features are spatial instead of spatial-temporal, and well suited for extracting features from still images. We applied our features to four datasets (COIL-100, Caltech 101, STL-10, PubFig), and observe a consistent improvement of 4% to 5% in classi■cation accuracy. With this approach, we achieve state-of-the-art recognition accuracy 61% on STL-10 dataset.
****

Homeostatic plasticity in Bayesian spiking networks as Expectation Maximization with posterior constraints
Stefan Habenschuss, Johannes Bill, Bernhard Nessler
Recent spiking network models of Bayesian inference and unsupervised learning frequently assume either inputs to arrive in a special format or employ complex computations in neuronal activation functions and synaptic plasticity rules. Here we show in a rigorous mathematical treatment how homeostatic processes, which have previously received little attention in this context, can overcome common theoretical limitations and facilitate the neural implementation and performance of existing models. In particular, we show that homeostatic plasticity can be understood as the enforcement of a 'balancing' posterior constraint during probabilistic inference and learning with Expectation Maximization. We link homeostatic dynamics to the theory of variational inference, and show that nontrivial terms, which typically appear during probabilistic inference in a large class of models, drop out. We demonstrate the feasibility of our approach in a spiking Winner-Take-All architecture of Bayesian inference and learning. Finally, we sketch how the mathematical framework can be extended to richer recurrent network architectures. Altogether, our theory provides a novel perspective on the interplay of homeostatic processes and synaptic plasticity in cortical microcircuits, and points to an essential role of homeostasis during inference and learning in spiking networks.
****

Nonparametric Bayesian Inverse Reinforcement Learning for Multiple Reward Functions
Jaedeug Choi, Kee-eung Kim
We present a nonparametric Bayesian approach to inverse reinforcement learning (IRL) for multiple reward functions. Most previous IRL algorithms assume that the behaviour data is obtained from an agent who is optimizing a single reward function, but this assumption is hard to be met in practice. Our approach is based on integrating the Dirichlet process mixture model into Bayesian IRL. We provide an efficient Metropolis-Hastings sampling algorithm utilizing the gradient of the posterior to estimate the underlying reward functions, and demonstrate that our approach outperforms the previous ones via experiments on a number of problem domains.
****

Human memory search as a random walk in a semantic network
Joseph Austerweil, Joshua T. Abbott, Thomas Griffiths
The human mind has a remarkable ability to store a vast amount of information in memory, and an even more remarkable ability to retrieve these experiences when needed. Understanding the representations and algorithms that underlie human memory search could potentially be useful in other information retrieval settings, including internet search. Psychological studies have revealed clear regularitie

s in how people search their memory, with clusters of semantically related items tending to be retrieved together. These findings have recently been taken as evidence that human memory search is similar to animals foraging for food in patchy environments, with people making a rational decision to switch away from a cluster of related information as it becomes depleted. We demonstrate that the results that were taken as evidence for this account also emerge from a random walk on a semantic network, much like the random web surfer model used in internet search engines. This offers a simpler and more unified account of how people search their memory, postulating a single process rather than one process for exploring a cluster and one process for switching between clusters.
****

Multi-criteria Anomaly Detection using Pareto Depth Analysis
Ko-jen Hsiao, Kevin Xu, Jeff Calder, Alfred Hero
We consider the problem of identifying patterns in a data set that exhibit anomalous behavior, often referred to as anomaly detection. In most anomaly detection algorithms, the dissimilarity between data samples is calculated by a single criterion, such as Euclidean distance. However, in many cases there may not exist a single dissimilarity measure that captures all possible anomalous patterns. In such a case, multiple criteria can be defined, and one can test for anomalies by scalarizing the multiple criteria by taking some linear combination of them. If the importance of the different criteria are not known in advance, the algorithm may need to be executed multiple times with different choices of weights in the linear combination. In this paper, we introduce a novel non-parametric multi-criteria anomaly detection method using Pareto depth analysis (PDA). PDA uses the concept of Pareto optimality to detect anomalies under multiple criteria without having to run an algorithm multiple times with different choices of weights. The proposed PDA approach scales linearly in the number of criteria and is provably better than linear combinations of the criteria.
****

A Spectral Algorithm for Latent Dirichlet Allocation
Anima Anandkumar, Dean P. Foster, Daniel J. Hsu, Sham M. Kakade, Yi-kai 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.
****

The Perturbed Variation
Maayan Harel, Shie Mannor
We introduce a new discrepancy score between two distributions that gives an indication on their \emph{similarity}. While much research has been done to determine if two samples come from exactly the same distribution, much less research considered the problem of determining if two finite samples come from similar distributions. The new score gives an intuitive interpretation of similarity; it optimally perturbs the distributions so that they best fit each other. The score is defined between distributions, and can be efficiently estimated from samples. We provide convergence bounds of the estimated score, and develop hypothesis testing procedures that test if two data sets come from similar distributions. The statistical power of this procedures is presented in simulations. We also compare the score's capacity to detect similarity with that of other known measures on real data.
****

Discriminatively Trained Sparse Code Gradients for Contour Detection
Ren Xiaofeng, Liefeng Bo
Finding contours in natural images is a fundamental problem that serves as the basis of many tasks such as image segmentation and object recognition. At the core of contour detection technologies are a set of hand-designed gradient features, used by most existing approaches including the state-of-the-art Global Pb (gPb) operator. In this work, we show that contour detection accuracy can be significantly improved by computing Sparse Code Gradients (SCG), which measure contrast using patch representations automatically learned through sparse coding. We u

se K-SVD and Orthogonal Matching Pursuit for efficient dictionary learning and e
ncoding, and use multi-scale pooling and power transforms to code oriented local
 neighborhoods before computing gradients and applying linear SVM. By extracting
 rich representations from pixels and avoiding collapsing them prematurely, Spar
se Code Gradients effectively learn how to measure local contrasts and find cont
ours. We improve the F-measure metric on the BSDS500 benchmark to 0.74 (up from
0.71 of gPb contours). Moreover, our learning approach can easily adapt to nove
l sensor data such as Kinect-style RGB-D cameras: Sparse Code Gradients on depth
 images and surface normals lead to promising contour detection using depth and
depth+color, as verified on the NYU Depth Dataset. Our work combines the concep
t of oriented gradients with sparse representation and opens up future possibili
ties for learning contour detection and segmentation.
****

A Bayesian Approach for Policy Learning from Trajectory Preference Queries
Aaron Wilson, Alan Fern, Prasad Tadepalli
We consider the problem of learning control policies via trajectory preference q
ueries to an expert. In particular, the learning agent can present an expert wit
h short runs of a pair of policies originating from the same state and the exper
t then indicates the preferred trajectory. The agent's goal is to elicit a laten
t target policy from the expert with as few queries as possible. To tackle this
problem we propose a novel Bayesian model of the querying process and introduce
two methods that exploit this model to actively select expert queries. Experimen
tal results on four benchmark problems indicate that our model can effectively l
earn policies from trajectory preference queries and that active query selection
 can be substantially more efficient than random selection.
****

Kernel Hyperalignment
Alexander Lorbert, Peter J. Ramadge
We offer a regularized, kernel extension of the multi-set, orthogonal Procrustes
 problem, or hyperalignment. Our new method, called Kernel Hyperalignment, expan
ds the scope of hyperalignment to include nonlinear measures of similarity and e
nables the alignment of multiple datasets with a large number of base features.
With direct application to fMRI data analysis, kernel hyperalignment is well-sui
ted for multi-subject alignment of large ROIs, including the entire cortex. We c
onducted experiments using real-world, multi-subject fMRI data.
****

Multi-Task Averaging
Sergey Feldman, Maya Gupta, Bela Frigyik
We present a multi-task learning approach to jointly estimate the means of multi
ple independent data sets. The proposed multi-task averaging (MTA) algorithm res
ults in a convex combination of the single-task averages. We derive the optimal
amount of regularization, and show that it can be effectively estimated. Simulat
ions and real data experiments demonstrate that MTA both maximum likelihood and
 James-Stein estimators, and that our approach to estimating the amount of regul
arization rivals cross-validation in performance but is more computationally eff
icient.
****

A Unifying Perspective of Parametric Policy Search Methods for Markov Decision P
rocesses
Thomas Furmston, David Barber
Parametric policy search algorithms are one of the methods of choice for the opt
imisation of Markov Decision Processes, with Expectation Maximisation and natura
l gradient ascent being considered the current state of the art in the field. In
 this article we provide a unifying perspective of these two algorithms by showi
ng that their step-directions in the parameter space are closely related to the
search direction of an approximate Newton method. This analysis leads naturally
to the consideration of this approximate Newton method as an alternative gradien
t-based method for Markov Decision Processes. We are able show that the algorith
m has numerous desirable properties, absent in the naive application of Newton's
 method, that make it a viable alternative to either Expectation Maximisation or

natural gradient ascent. Empirical results suggest that the algorithm has excel lent convergence and robustness properties, performing strongly in comparison to both Expectation Maximisation and natural gradient ascent.
****

Convergence and Energy Landscape for Cheeger Cut Clustering
Xavier Bresson, Thomas Laurent, David Uminsky, James Brecht
Unsupervised clustering of scattered, noisy and high-dimensional data points is an important and difficult problem. Continuous relaxations of balanced cut probl ems yield excellent clustering results. This paper provides rigorous convergenc e results for two algorithms that solve the relaxed Cheeger Cut minimization. T he first algorithm is a new steepest descent algorithm and the second one is a s light modification of the Inverse Power Method algorithm \cite{pro:HeinBuhler10O neSpec}. While the steepest descent algorithm has better theoretical convergence properties, in practice both algorithm perform equally. We also completely ch aracterize the local minima of the relaxed problem in terms of the original bala nced cut problem, and relate this characterization to the convergence of the alg orithms.
****

A Divide-and-Conquer Method for Sparse Inverse Covariance Estimation
Cho-jui Hsieh, Arindam Banerjee, Inderjit Dhillon, Pradeep Ravikumar
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-auth ors prior to requesting a name change in the electronic proceedings.
****

Nonconvex Penalization Using Laplace Exponents and Concave Conjugates
Zhihua Zhang, Bojun Tu
In this paper we study sparsity-inducing nonconvex penalty functions using L´evy processes. We de■ne such a penalty as the Laplace exponent of a subordina- tor. Accordingly, we propose a novel approach for the construction of sparsity- indu cing nonconvex penalties. Particularly, we show that the nonconvex logarith- mic (LOG) and exponential (EXP) penalty functions are the Laplace exponents of Gamm a and compound Poisson subordinators, respectively. Additionally, we explore the concave conjugate of nonconvex penalties. We ■nd that the LOG and EXP penalties are the concave conjugates of negative Kullback-Leiber (KL) dis- tance function s. Furthermore, the relationship between these two penalties is due to asymmetri city of the KL distance.
****

How They Vote: Issue-Adjusted Models of Legislative Behavior
Sean Gerrish, David Blei
We develop a probabilistic model of legislative data that uses the text of the b ills to uncover lawmakers' positions on specific political issues. Our model ca n be used to explore how a lawmaker's voting patterns deviate from what is expec ted and how that deviation depends on what is being voted on. We derive approxim ate posterior inference algorithms based on variational methods. Across 12 years of legislative data, we demonstrate both improvement in heldout predictive perf ormance and the model's utility in interpreting an inherently multi-dimensional space.
****

Multiclass Learning Approaches: A Theoretical Comparison with Implications
Amit Daniely, Sivan Sabato, Shai 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.
****

Dynamical And-Or Graph Learning for Object Shape Modeling and Detection
Xiaolong Wang, Liang Lin
This paper studies a novel discriminative part-based model to represent and reco gnize object shapes with an “And-Or graph”. We define this model consisting of t

hree layers: the leaf-nodes with collaborative edges for localizing local parts, the or-nodes specifying the switch of leaf-nodes, and the root-node encoding the global verification. A discriminative learning algorithm, extended from the CCCP [23], is proposed to train the model in a dynamical manner: the model structure (e.g., the configuration of the leaf-nodes associated with the or-nodes) is automatically determined with optimizing the multi-layer parameters during the iteration. The advantages of our method are two-fold. (i) The And-Or graph model enables us to handle well large intra-class variance and background clutters for object shape detection from images. (ii) The proposed learning algorithm is able to obtain the And-Or graph representation without requiring elaborate supervision and initialization. We validate the proposed method on several challenging databases (e.g., INRIA-Horse, ETHZ-Shape, and UIUC-People), and it outperforms the state-of-the-arts approaches.
****

Augmented-SVM: Automatic space partitioning for combining multiple non-linear dynamics
Ashwini Shukla, Aude Billard
Non-linear dynamical systems (DS) have been used extensively for building generative models of human behavior. Its applications range from modeling brain dynamics to encoding motor commands. Many schemes have been proposed for encoding robot motions using dynamical systems with a single attractor placed at a predefined target in state space. Although these enable the robots to react against sudden perturbations without any re-planning, the motions are always directed towards a single target. In this work, we focus on combining several such DS with distinct attractors, resulting in a multi-stable DS. We show its applicability in reach-to-grasp tasks where the attractors represent several grasping points on the target object. While exploiting multiple attractors provides more flexibility in recovering from unseen perturbations, it also increases the complexity of the underlying learning problem. Here we present the Augmented-SVM (A-SVM) model which inherits region partitioning ability of the well known SVM classifier and is augmented with novel constraints derived from the individual DS. The new constraints modify the original SVM dual whose optimal solution then results in a new class of support vectors (SV). These new SV ensure that the resulting multi-stable DS incurs minimum deviation from the original dynamics and is stable at each of the attractors within a finite region of attraction. We show, via implementations on a simulated 10 degrees of freedom mobile robotic platform, that the model is capable of real-time motion generation and is able to adapt on-the-fly to perturbations.
****

3D Social Saliency from Head-mounted Cameras
Hyun Park, Eakta Jain, Yaser Sheikh
Yaser Sheikh
****

Coupling Nonparametric Mixtures via Latent Dirichlet Processes
Dahua Lin, John Fisher
Mixture distributions are often used to model complex data. In this paper, we develop a new method that jointly estimates mixture models over multiple data sets by exploiting the statistical dependencies between them. Specifically, we introduce a set of latent Dirichlet processes as sources of component models (atoms), and for each data set, we construct a nonparametric mixture model by combining sub-sampled versions of the latent DPs. Each mixture model may acquire atoms from different latent DPs, while each atom may be shared by multiple mixtures. This multi-to-multi association distinguishes the proposed method from prior constructions that rely on tree or chain structures, allowing mixture models to be coupled more flexibly. In addition, we derive a sampling algorithm that jointly infers the model parameters and present experiments on both document analysis and image modeling.
****

Multiclass Learning with Simplex Coding
Youssef Mroueh, Tomaso Poggio, Lorenzo Rosasco, Jean-jeacques Slotine

In this paper we dicuss a novel framework for multiclass learning, defined by a suitable coding/decoding strategy, namely the simplex coding, that allows to generalize to multiple classes a relaxation approach commonly used in binary classification. In this framework a relaxation error analysis can be developed avoiding constraints on the considered hypotheses class. Moreover, we show that in this setting it is possible to derive the first provably consistent regularized methods with training/tuning complexity which is {\em independent} to the number of classes. Tools from convex analysis are introduced that can be used beyond the scope of this paper.
****

Clustering Sparse Graphs
Yudong Chen, Sujay Sanghavi, Huan Xu
We develop a new algorithm to cluster sparse unweighted graphs -- i.e. partition the nodes into disjoint clusters so that there is higher density within clusters, and low across clusters. By sparsity we mean the setting where both the in-cluster and across cluster edge densities are very small, possibly vanishing in the size of the graph. Sparsity makes the problem noisier, and hence more difficult to solve. Any clustering involves a tradeoff between minimizing two kinds of errors: missing edges within clusters and present edges across clusters. Our insight is that in the sparse case, these must be {\em penalized differently}. We analyze our algorithm's performance on the natural, classical and widely studied ``planted partition'' model (also called the stochastic block model); we show that our algorithm can cluster sparser graphs, and with smaller clusters, than all previous methods. This is seen empirically as well.
****

On-line Reinforcement Learning Using Incremental Kernel-Based Stochastic Factorization
Andre Barreto, Doina Precup, Joelle Pineau
The ability to learn a policy for a sequential decision problem with continuous state space using on-line data is a long-standing challenge. This paper presents a new reinforcement-learning algorithm, called iKBSF, which extends the benefits of kernel-based learning to the on-line scenario. As a kernel-based method, the proposed algorithm is stable and has good convergence properties. However, unlike other similar algorithms,iKBSF's space complexity is independent of the number of sample transitions, and as a result it can process an arbitrary amount of data. We present theoretical results showing that iKBSF can approximate (to any level of accuracy) the value function that would be learned by an equivalent batch non-parametric kernel-based reinforcement learning approximator. In order to show the effectiveness of the proposed algorithm in practice, we apply iKBSF to the challenging three-pole balancing task, where the ability to process a large number of transitions is crucial for achieving a high success rate.
****

Accelerated Training for Matrix-norm Regularization: A Boosting Approach
Xinhua Zhang, Dale Schuurmans, Yao-liang Yu
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****

Supervised Learning with Similarity Functions
Purushottam Kar, Prateek Jain
We address the problem of general supervised learning when data can only be accessed through an (indefinite) similarity function between data points. Existing work on learning with indefinite kernels has concentrated solely on binary/multiclass classification problems. We propose a model that is generic enough to handle any supervised learning task and also subsumes the model previously proposed for classification. We give a ''goodness'' criterion for similarity functions w.r.t. a given supervised learning task and then adapt a well-known landmarking technique to provide efficient algorithms for supervised learning using ''good'' similarity functions. We demonstrate the effectiveness of our model on three impor

tant supervised learning problems: a) real-valued regression, b) ordinal regress ion and c) ranking where we show that our method guarantees bounded generalizati on error. Furthermore, for the case of real-valued regression, we give a natural goodness definition that, when used in conjunction with a recent result in spar se vector recovery, guarantees a sparse predictor with bounded generalization er ror. Finally, we report results of our learning algorithms on regression and ord inal regression tasks using non-PSD similarity functions and demonstrate the eff ectiveness of our algorithms, especially that of the sparse landmark selection a lgorithm that achieves significantly higher accuracies than the baseline methods while offering reduced computational costs.

****

A Simple and Practical Algorithm for Differentially Private Data Release

Moritz Hardt, Katrina Ligett, Frank Mcsherry

We present a new algorithm for differentially private data release, based on a s imple combination of the Exponential Mechanism with the Multiplicative Weights u pdate rule. Our MWEM algorithm achieves what are the best known and nearly opti mal theoretical guarantees, while at the same time being simple to implement and experimentally more accurate on actual data sets than existing techniques.

****

Persistent Homology for Learning Densities with Bounded Support

Florian Pokorny, Hedvig Kjellström, Danica Kragic, Carl Ek

We present a novel method for learning densities with bounded support which enab les us to incorporate `hard' topological constraints. In particular, we show how emerging techniques from computational algebraic topology and the notion of Per sistent Homology can be combined with kernel based methods from Machine Learnin g for the purpose of density estimation. The proposed formalism facilitates lear ning of models with bounded support in a principled way, and -- by incorporating Persistent Homology techniques in our approach -- we are able to encode algebra ic-topological constraints which are not addressed in current state-of the art p robabilistic models. We study the behaviour of our method on two synthetic examp les for various sample sizes and exemplify the benefits of the proposed approach on a real-world data-set by learning a motion model for a racecar. We show how to learn a model which respects the underlying topological structure of the race track, constraining the trajectories of the car.

****

Proper losses for learning from partial labels

Jesús Cid-sueiro

This paper discusses the problem of calibrating posterior class probabilities fr om partially labelled data. Each instance is assumed to be labelled as belonging to one of several candidate categories, at most one of them being true. We gene ralize the concept of proper loss to this scenario, establish a necessary and su fficient condition for a loss function to be proper, and we show a direct proced ure to construct a proper loss for partial labels from a conventional proper los s. The problem can be characterized by the mixing probability matrix relating th e true class of the data and the observed labels. An interesting result is that the full knowledge of this matrix is not required, and losses can be constructed that are proper in a subset of the probability simplex.

****

Predicting Action Content On-Line and in Real Time before Action Onset – an Intr acranial Human Study

Uri Maoz, Shengxuan Ye, Ian Ross, Adam Mamelak, Christof Koch

The ability to predict action content from neural signals in real time before th e ac- tion occurs has been long sought in the neuroscienti■c study of decision-m aking, agency and volition. On-line real-time (ORT) prediction is important for under- standing the relation between neural correlates of decision-making and co nscious, voluntary action as well as for brain-machine interfaces. Here, epileps y patients, implanted with intracranial depth microelectrodes or subdural grid e lectrodes for clinical purposes, participated in a “matching-pennies” game again st an opponent. In each trial, subjects were given a 5 s countdown, after which they had to raise their left or right hand immediately as the “go” signal appear

ed on a computer screen. They won a ■xed amount of money if they raised a differ ent hand than their opponent and lost that amount otherwise. The question we her e studied was the extent to which neural precursors of the subjects' decisions c an be detected in intracranial local ■eld potentials (LFP) prior to the onset of the action. We found that combined low-frequency (0.1–5 Hz) LFP signals from 10 electrodes were predictive of the intended left-/right-hand movements before th e onset of the go signal. Our ORT system predicted which hand the patient would raise 0.5 s before the go signal with 68±3% accuracy in two patients. Based on t hese results, we constructed an ORT system that tracked up to 30 electrodes simu ltaneously, and tested it on retrospective data from 7 patients. On average, we could predict the correct hand choice in 83% of the trials, which rose to 92% if we let the system drop 3/10 of the trials on which it was less con■dent. Our sy stem demonstrates— for the ■rst time—the feasibility of accurately predicting a binary action on single trials in real time for patients with intracranial recor dings, well before the action occurs.

****

Classification Calibration Dimension for General Multiclass Losses

Harish G. Ramaswamy, Shivani Agarwal

We study consistency properties of surrogate loss functions for general multicla ss classification problems, defined by a general loss matrix. We extend the noti on of classification calibration, which has been studied for binary and multicla ss 0-1 classification problems (and for certain other specific learning problems), to the general multiclass setting, and derive necessary and sufficient condit ions for a surrogate loss to be classification calibrated with respect to a loss matrix in this setting. We then introduce the notion of \emph{classification ca libration dimension} of a multiclass loss matrix, which measures the smallest si ze' of a prediction space for which it is possible to design a convex surrogate that is classification calibrated with respect to the loss matrix. We derive bot h upper and lower bounds on this quantity, and use these results to analyze vari ous loss matrices. In particular, as one application, we provide a different rou te from the recent result of Duchi et al.\ (2010) for analyzing the difficulty o f designinglow-dimensional' convex surrogates that are consistent with respect t o pairwise subset ranking losses. We anticipate the classification calibration d imension may prove to be a useful tool in the study and design of surrogate loss es for general multiclass learning problems.

****

Analog readout for optical reservoir computers

Anteo Smerieri, François Duport, Yvon Paquot, Benjamin Schrauwen, Marc Haelterma n, Serge Massar

Reservoir computing is a new, powerful and flexible machine learning technique t hat is easily implemented in hardware. Recently, by using a time-multiplexed arc hitecture, hardware reservoir computers have reached performance comparable to d igital implementations. Operating speeds allowing for real time information oper ation have been reached using optoelectronic systems. At present the main perfor mance bottleneck is the readout layer which uses slow, digital postprocessing. W e have designed an analog readout suitable for time-multiplexed optoelectronic r eservoir computers, capable of working in real time. The readout has been built and tested experimentally on a standard benchmark task. Its performance is bette r than non-reservoir methods, with ample room for further improvement. The prese nt work thereby overcomes one of the major limitations for the future developmen t of hardware reservoir computers.

****

Perfect Dimensionality Recovery by Variational Bayesian PCA

Shinichi Nakajima, Ryota Tomioka, Masashi Sugiyama, S. Babacan

The variational Bayesian (VB) approach is one of the best tractable approximatio ns to the Bayesian estimation, and it was demonstrated to perform well in many a pplications. However, its good performance was not fully understood theoreticall y. For example, VB sometimes produces a sparse solution, which is regarded as a practical advantage of VB, but such sparsity is hardly observed in the rigorous Bayesian estimation. In this paper, we focus on probabilistic PCA and give more

theoretical insight into the empirical success of VB. More specifically, for the situation where the noise variance is unknown, we derive a sufficient condition for perfect recovery of the true PCA dimensionality in the large-scale limit when the size of an observed matrix goes to infinity. In our analysis, we obtain bounds for a noise variance estimator and simple closed-form solutions for other parameters, which themselves are actually very useful for better implementation of VB-PCA.

****

Compressive neural representation of sparse, high-dimensional probabilities
Zachary Pitkow

This paper shows how sparse, high-dimensional probability distributions could be represented by neurons with exponential compression. The representation is a novel application of compressive sensing to sparse probability distributions rather than to the usual sparse signals. The compressive measurements correspond to expected values of nonlinear functions of the probabilistically distributed variables. When these expected values are estimated by sampling, the quality of the compressed representation is limited only by the quality of sampling. Since the compression preserves the geometric structure of the space of sparse probability distributions, probabilistic computation can be performed in the compressed domain. Interestingly, functions satisfying the requirements of compressive sensing can be implemented as simple perceptrons. If we use perceptrons as a simple model of feedforward computation by neurons, these results show that the mean activity of a relatively small number of neurons can accurately represent a high-dimensional joint distribution implicitly, even without accounting for any noise correlations. This comprises a novel hypothesis for how neurons could encode probabilities in the brain.

****

Shifting Weights: Adapting Object Detectors from Image to Video
Kevin Tang, Vignesh Ramanathan, Li Fei-fei, Daphne Koller

Typical object detectors trained on images perform poorly on video, as there is a clear distinction in domain between the two types of data. In this paper, we tackle the problem of adapting object detectors learned from images to work well on videos. We treat the problem as one of unsupervised domain adaptation, in which we are given labeled data from the source domain (image), but only unlabeled data from the target domain (video). Our approach, self-paced domain adaptation, seeks to iteratively adapt the detector by re-training the detector with automatically discovered target domain examples, starting with the easiest first. At each iteration, the algorithm adapts by considering an increased number of target domain examples, and a decreased number of source domain examples. To discover target domain examples from the vast amount of video data, we introduce a simple, robust approach that scores trajectory tracks instead of bounding boxes. We also show how rich and expressive features specific to the target domain can be incorporated under the same framework. We show promising results on the 2011 TRECVID Multimedia Event Detection and LabelMe Video datasets that illustrate the benefit of our approach to adapt object detectors to video.

****

Minimization of Continuous Bethe Approximations: A Positive Variation
Jason Pacheco, Erik Sudderth

We develop convergent minimization algorithms for Bethe variational approximations which explicitly constrain marginal estimates to families of valid distributions. While existing message passing algorithms define fixed point iterations corresponding to stationary points of the Bethe free energy, their greedy dynamics do not distinguish between local minima and maxima, and can fail to converge. For continuous estimation problems, this instability is linked to the creation of invalid marginal estimates, such as Gaussians with negative variance. Conversely, our approach leverages multiplier methods with well-understood convergence properties, and uses bound projection methods to ensure that marginal approximations are valid at all iterations. We derive general algorithms for discrete and Gaussian pairwise Markov random fields, showing improvements over standard loopy belief propagation. We also apply our method to a hybrid model with both discrete

and continuous variables, showing improvements over expectation propagation.
****
Optimal Regularized Dual Averaging Methods for Stochastic Optimization
Xi Chen, Qihang Lin, Javier Pena
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Gradient Weights help Nonparametric Regressors
Samory Kpotufe, Abdeslam Boularias
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Regularized Off-Policy TD-Learning
Bo Liu, Sridhar Mahadevan, Ji Liu
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Locating Changes in Highly Dependent Data with Unknown Number of Change Points
Azadeh Khaleghi, Daniil Ryabko
The problem of multiple change point estimation is considered for sequences with unknown number of change points. A consistency framework is suggested that is suitable for highly dependent time-series, and an asymptotically consistent algorithm is proposed. In order for the consistency to be established the only assumption required is that the data is generated by stationary ergodic time-series distributions. No modeling, independence or parametric assumptions are made; the data are allowed to be dependent and the dependence can be of arbitrary form. The theoretical results are complemented with experimental evaluations.
****
Controlled Recognition Bounds for Visual Learning and Exploration
Vasiliy Karasev, Alessandro Chiuso, Stefano Soatto
We describe the tradeoff between the performance in a visual recognition problem and the control authority that the agent can exercise on the sensing process. We focus on the problem of “visual search” of an object in an otherwise known and static scene, propose a measure of control authority, and relate it to the expected risk and its proxy (conditional entropy of the posterior density). We show this analytically, as well as empirically by simulation using the simplest known model that captures the phenomenology of image formation, including scaling and occlusions. We show that a “passive” agent given a training set can provide no guarantees on performance beyond what is afforded by the priors, and that an “omnipotent” agent, capable of infinite control authority, can achieve arbitrarily good performance (asymptotically).
****
Multi-Stage Multi-Task Feature Learning
Pinghua Gong, Jieping Ye, Chang-shui Zhang
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Fast Resampling Weighted v-Statistics
Chunxiao Zhou, Jiseong Park, Yun Fu
In this paper, a novel, computationally fast, and alternative algorithm for computing weighted v-statistics in resampling both univariate and multivariate data is proposed. To avoid any real resampling, we have linked this problem with fi

nite group action and converted it into a problem of orbit enumeration. For furt
her computational cost reduction, an efficient method is developed to list all o
rbits by their symmetry order and calculate all index function orbit sums and da
ta function orbit sums recursively. The computational complexity analysis shows
reduction in the computational cost from n! or nn level to low-order polynomial
level.
****

Rational inference of relative preferences
Nisheeth Srivastava, Paul R. Schrater
Statistical decision theory axiomatically assumes that the relative desirability
 of different options that humans perceive is well described by assigning them o
ption-specific scalar utility functions. However, this assumption is refuted by
observed human behavior, including studies wherein preferences have been shown t
o change systematically simply through variation in the set of choice options pr
esented. In this paper, we show that interpreting desirability as a relative com
parison between available options at any particular decision instance results in
 a rational theory of value-inference that explains heretofore intractable viola
tions of rational choice behavior in human subjects. Complementarily, we also ch
aracterize the conditions under which a rational agent selecting optimal options
 indicated by dynamic value inference in our framework will behave identically t
o one whose preferences are encoded using a static ordinal utility function.
****

Latent Graphical Model Selection: Efficient Methods for Locally Tree-like Graphs
Anima Anandkumar, Ragupathyraj Valluvan
Requests for name changes in the electronic proceedings will be accepted with 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.
****

On the connections between saliency and tracking
Vijay Mahadevan, Nuno Vasconcelos
A model connecting visual tracking and saliency has recently been proposed. This
 model is based on the saliency hypothesis for tracking which postulates that tr
acking is achieved by the top-down tuning, based on target features, of discrimi
nant center-surround saliency mechanisms over time. In this work, we identify th
ree main predictions that must hold if the hypothesis were true: 1) tracking rel
iability should be larger for salient than for non-salient targets, 2) tracking
reliability should have a dependence on the defining variables of saliency, name
ly feature contrast and distractor heterogeneity, and must replicate the depende
nce of saliency on these variables, and 3) saliency and tracking can be implemen
ted with common low level neural mechanisms. We confirm that the first two predi
ctions hold by reporting results from a set of human behavior studies on the con
nection between saliency and tracking. We also show that the third prediction ho
lds by constructing a common neurophysiologically plausible architecture that ca
n computationally solve both saliency and tracking. This architecture is fully c
ompliant with the standard physiological models of V1 and MT, and with what is k
nown about attentional control in area LIP, while explaining the results of the
human behavior experiments.
****

Symbolic Dynamic Programming for Continuous State and Observation POMDPs
Zahra Zamani, Scott Sanner, Pascal Poupart, Kristian Kersting
Partially-observable Markov decision processes (POMDPs) provide a powerful model
 for real-world sequential decision-making problems. In recent years, point- bas
ed value iteration methods have proven to be extremely effective techniques for
■nding (approximately) optimal dynamic programming solutions to POMDPs when an i
nitial set of belief states is known. However, no point-based work has provided
exact point-based backups for both continuous state and observation spaces, whic
h we tackle in this paper. Our key insight is that while there may be an in■nite
 number of possible observations, there are only a ■nite number of observation p
artitionings that are relevant for optimal decision-making when a ■nite, ■xed se

t of reachable belief states is known. To this end, we make two important contri
butions: (1) we show how previous exact symbolic dynamic pro- gramming solutions
 for continuous state MDPs can be generalized to continu- ous state POMDPs with
discrete observations, and (2) we show how this solution can be further extended
 via recently developed symbolic methods to continuous state and observations to
 derive the minimal relevant observation partitioning for potentially correlated
, multivariate observation spaces. We demonstrate proof-of- concept results on u
ni- and multi-variate state and observation steam plant control.
****

Imitation Learning by Coaching
He He, Jason Eisner, Hal Daume
Imitation Learning has been shown to be successful in solving many challenging r
eal-world problems. Some recent approaches give strong performance guarantees by
 training the policy iteratively. However, it is important to note that these gu
arantees depend on how well the policy we found can imitate the oracle on the
training data. When there is a substantial difference between the oracle's abil
ity and the learner's policy space, we may fail to find a policy that has low e
rror on the training set. In such cases, we propose to use a coach that demonstr
ates easy-to-learn actions for the learner and gradually approaches the oracle.
 By a reduction of learning by demonstration to online learning, we prove that
coaching can yield a lower regret bound than using the oracle. We apply our algo
rithm to a novel cost-sensitive dynamic feature selection problem, a hard decisi
on problem that considers a user-specified accuracy-cost trade-off. Experimenta
l results on UCI datasets show that our method outperforms state-of-the-art imit
ation learning methods in dynamic features selection and two static feature sele
ction methods.
****

Learning about Canonical Views from Internet Image Collections
Elad Mezuman, Yair Weiss
Although human object recognition is supposedly robust to viewpoint, much resear
ch on human perception indicates that there is a preferred or “canonical” view o
f objects. This phenomenon was discovered more than 30 years ago but the canonic
al view of only a small number of categories has been validated experimentally.
Moreover, the explanation for why humans prefer the canonical view over other vi
ews remains elusive. In this paper we ask: Can we use Internet image collections
 to learn more about canonical views? We start by manually finding the most comm
on view in the results returned by Internet search engines when queried with the
 objects used in psychophysical experiments. Our results clearly show that the m
ost likely view in the search engine corresponds to the same view preferred by h
uman subjects in experiments. We also present a simple method to find the most l
ikely view in an image collection and apply it to hundreds of categories. Using
the new data we have collected we present strong evidence against the two most p
rominent formal theories of canonical views and provide novel constraints for ne
w theories.
****

A lattice filter model of the visual pathway
Karol Gregor, Dmitri Chklovskii
Early stages of visual processing are thought to decorrelate, or whiten, the inc
oming temporally varying signals. Because the typical correlation time of natura
l stimuli, as well as the extent of temporal receptive fields of lateral genicul
ate nucleus (LGN) neurons, is much greater than neuronal time constants, such de
correlation must be done in stages combining contributions of multiple neurons.
We propose to model temporal decorrelation in the visual pathway with the lattic
e filter, a signal processing device for stage-wise decorrelation of temporal si
gnals. The stage-wise architecture of the lattice filter maps naturally onto the
 visual pathway (photoreceptors -> bipolar cells -> retinal ganglion cells -> LG
N) and its filter weights can be learned using Hebbian rules in a stage-wise seq
uential manner. Moreover, predictions of neural activity from the lattice filter
 model are consistent with physiological measurements in LGN neurons and fruit f
ly second-order visual neurons. Therefore, the lattice filter model is a useful

abstraction that may help unravel visual system function.
****
Multi-scale Hyper-time Hardware Emulation of Human Motor Nervous System Based on Spiking Neurons using FPGA
C. Niu, Sirish Nandyala, Won Sohn, Terence Sanger
Our central goal is to quantify the long-term progression of pediatric neurological diseases, such as a typical 10-15 years progression of child dystonia. To this purpose, quantitative models are convincing only if they can provide multi-scale details ranging from neuron spikes to limb biomechanics. The models also need to be evaluated in hyper-time, i.e. significantly faster than real-time, for producing useful predictions. We designed a platform with digital VLSI hardware for multi-scale hyper-time emulations of human motor nervous systems. The platform is constructed on a scalable, distributed array of Field Programmable Gate Array (FPGA) devices. All devices operate asynchronously with 1 millisecond time granularity, and the overall system is accelerated to 365x real-time. Each physiological component is implemented using models from well documented studies and can be flexibly modified. Thus the validity of emulation can be easily advised by neurophysiologists and clinicians. For maximizing the speed of emulation, all calculations are implemented in combinational logic instead of clocked iterative circuits. This paper presents the methodology of building FPGA modules in correspondence to components of a monosynaptic spinal loop. Results of emulated activities are shown. The paper also discusses the rationale of approximating neural circuitry by organizing neurons with sparse interconnections. In conclusion, our platform allows introducing various abnormalities into the neural emulation such that the emerging motor symptoms can be analyzed. It compels us to test the origins of childhood motor disorders and predict their long-term progressions.
****
Recognizing Activities by Attribute Dynamics
Weixin Li, Nuno Vasconcelos
In this work, we consider the problem of modeling the dynamic structure of human activities in the attributes space. A video sequence is first represented in a semantic feature space, where each feature encodes the probability of occurrence of an activity attribute at a given time. A generative model, denoted the binary dynamic system (BDS), is proposed to learn both the distribution and dynamics of different activities in this space. The BDS is a non-linear dynamic system, which extends both the binary principal component analysis (PCA) and classical linear dynamic systems (LDS), by combining binary observation variables with a hidden Gauss-Markov state process. In this way, it integrates the representation power of semantic modeling with the ability of dynamic systems to capture the temporal structure of time-varying processes. An algorithm for learning BDS parameters, inspired by a popular LDS learning method from dynamic textures, is proposed. A similarity measure between BDSs, which generalizes the Binet-Cauchy kernel for LDS, is then introduced and used to design activity classifiers. The proposed method is shown to outperform similar classifiers derived from the kernel dynamic system (KDS) and state-of-the-art approaches for dynamics-based or attribute-based action recognition.
****
Statistical Consistency of Ranking Methods in A Rank-Differentiable Probability Space
Yanyan Lan, Jiafeng Guo, Xueqi Cheng, Tie-yan Liu
This paper is concerned with the statistical consistency of ranking methods. Recently, it was proven that many commonly used pairwise ranking methods are inconsistent with the weighted pairwise disagreement loss (WPDL), which can be viewed as the true loss of ranking, even in a low-noise setting. This result is interesting but also surprising, given that the pairwise ranking methods have been shown very effective in practice. In this paper, we argue that the aforementioned result might not be conclusive, depending on what kind of assumptions are used. We give a new assumption that the labels of objects to rank lie in a rank-differentiable probability space (RDPS), and prove that the pairwise ranking methods become consistent with WPDL under this assumption. What is especially inspiring is

that RDPS is actually not stronger than but similar to the low-noise setting. Our studies provide theoretical justifications of some empirical findings on pairwise ranking methods that are unexplained before, which bridge the gap between theory and applications.
****

A Scalable CUR Matrix Decomposition Algorithm: Lower Time Complexity and Tighter Bound
Shusen Wang, Zhihua Zhang
The CUR matrix decomposition is an important extension of Nyström approximation to a general matrix. It approximates any data matrix in terms of a small number of its columns and rows. In this paper we propose a novel randomized CUR algorithm with an expected relative-error bound. The proposed algorithm has the advantages over the existing relative-error CUR algorithms that it possesses tighter theoretical bound and lower time complexity, and that it can avoid maintaining the whole data matrix in main memory. Finally, experiments on several real-world datasets demonstrate significant improvement over the existing relative-error algorithms.
****

Efficient Bayes-Adaptive Reinforcement Learning using Sample-Based Search
Arthur Guez, David Silver, Peter Dayan
Bayesian model-based reinforcement learning is a formally elegant approach to learning optimal behaviour under model uncertainty, trading off exploration and exploitation in an ideal way. Unfortunately, finding the resulting Bayes-optimal policies is notoriously taxing, since the search space becomes enormous. In this paper we introduce a tractable, sample-based method for approximate Bayes-optimal planning which exploits Monte-Carlo tree search. Our approach outperformed prior Bayesian model-based RL algorithms by a significant margin on several well-known benchmark problems -- because it avoids expensive applications of Bayes rule within the search tree by lazily sampling models from the current beliefs. We illustrate the advantages of our approach by showing it working in an infinite state space domain which is qualitatively out of reach of almost all previous work in Bayesian exploration.
****

Dual-Space Analysis of the Sparse Linear Model
Yi Wu, David Wipf
Sparse linear (or generalized linear) models combine a standard likelihood function with a sparse prior on the unknown coefficients. These priors can conveniently be expressed as a maximization over zero-mean Gaussians with different variance hyperparameters. Standard MAP estimation (Type I) involves maximizing over both the hyperparameters and coefficients, while an empirical Bayesian alternative (Type II) first marginalizes the coefficients and then maximizes over the hyperparameters, leading to a tractable posterior approximation. The underlying cost functions can be related via a dual-space framework from Wipf et al. (2011), which allows both the Type I or Type II objectives to be expressed in either coefficient or hyperparmeter space. This perspective is useful because some analyses or extensions are more conducive to development in one space or the other. Herein we consider the estimation of a trade-off parameter balancing sparsity and data fit. As this parameter is effectively a variance, natural estimators exist by assessing the problem in hyperparameter (variance) space, transitioning natural ideas from Type II to solve what is much less intuitive for Type I. In contrast, for analyses of update rules and sparsity properties of local and global solutions, as well as extensions to more general likelihood models, we can leverage coefficient-space techniques developed for Type I and apply them to Type II. For example, this allows us to prove that Type II-inspired techniques can be successful recovering sparse coefficients when unfavorable restricted isometry properties (RIP) lead to failure of popular L1 reconstructions. It also facilitates the analysis of Type II when non-Gaussian likelihood models lead to intractable integrations.
****

Neuronal Spike Generation Mechanism as an Oversampling, Noise-shaping A-to-D con

verter
Dmitri Chklovskii, Daniel Soudry
We explore the hypothesis that the neuronal spike generation mechanism is an ana log-to-digital converter, which rectifies low-pass filtered summed synaptic curr ents and encodes them into spike trains linearly decodable in post-synaptic neur ons. To digitally encode an analog current waveform, the sampling rate of the sp ike generation mechanism must exceed its Nyquist rate. Such oversampling is cons istent with the experimental observation that the precision of the spike-generat ion mechanism is an order of magnitude greater than the cut-off frequency of den dritic low-pass filtering. To achieve additional reduction in the error of analo g-to-digital conversion, electrical engineers rely on noise-shaping. If noise-sh aping were used in neurons, it would introduce correlations in spike timing to r educe low-frequency (up to Nyquist) transmission error at the cost of high-frequ ency one (from Nyquist to sampling rate). Using experimental data from three dif ferent classes of neurons, we demonstrate that biological neurons utilize noise- shaping. We also argue that rectification by the spike-generation mechanism may improve energy efficiency and carry out de-noising. Finally, the zoo of ion chan nels in neurons may be viewed as a set of predictors, various subsets of which a re activated depending on the statistics of the input current.
****
Multiple Operator-valued Kernel Learning
Hachem Kadri, Alain Rakotomamonjy, Philippe Preux, Francis Bach
Positive definite operator-valued kernels generalize the well-known notion of re producing kernels, and are naturally adapted to multi-output learning situations . This paper addresses the problem of learning a finite linear combination of in finite-dimensional operator-valued kernels which are suitable for extending func tional data analysis methods to nonlinear contexts. We study this problem in the case of kernel ridge regression for functional responses with an lr-norm constr aint on the combination coefficients. The resulting optimization problem is more involved than those of multiple scalar-valued kernel learning since operator-va lued kernels pose more technical and theoretical issues. We propose a multiple o perator-valued kernel learning algorithm based on solving a system of linear ope rator equations by using a block coordinate-descent procedure. We experimentally validate our approach on a functional regression task in the context of finger movement prediction in brain-computer interfaces.
****
No-Regret Algorithms for Unconstrained Online Convex Optimization
Brendan Mcmahan, Matthew Streeter
Some of the most compelling applications of online convex optimization, includin g online prediction and classification, are unconstrained: the natural feasible set is R^n. Existing algorithms fail to achieve sub-linear regret in this setti ng unless constraints on the comparator point x* are known in advance. We prese nt an algorithm that, without such prior knowledge, offers near-optimal regret b ounds with respect to any choice of x. In particular, regret with respect to x = 0 is constant. We then prove lower bounds showing that our algorithm's guaran tees are optimal in this setting up to constant factors.
****
Learning Partially Observable Models Using Temporally Abstract Decision Trees
Erik Talvitie
This paper introduces timeline trees, which are partial models of partially obse rvable environments. Timeline trees are given some specific predictions to make and learn a decision tree over history. The main idea of timeline trees is to us e temporally abstract features to identify and split on features of key events, spread arbitrarily far apart in the past (whereas previous decision-tree-based m ethods have been limited to a finite suffix of history). Experiments demonstrate that timeline trees can learn to make high quality predictions in complex, part ially observable environments with high-dimensional observations (e.g. an arcade game).
****
Emergence of Object-Selective Features in Unsupervised Feature Learning

Adam Coates, Andrej Karpathy, Andrew Ng
Recent work in unsupervised feature learning has focused on the goal of discovering high-level features from unlabeled images. Much progress has been made in this direction, but in most cases it is still standard to use a large amount of labeled data in order to construct detectors sensitive to object classes or other complex patterns in the data. In this paper, we aim to test the hypothesis that unsupervised feature learning methods, provided with only unlabeled data, can learn high-level, invariant features that are sensitive to commonly-occurring objects. Though a handful of prior results suggest that this is possible when each object class accounts for a large fraction of the data (as in many labeled datasets), it is unclear whether something similar can be accomplished when dealing with completely unlabeled data. A major obstacle to this test, however, is scale: we cannot expect to succeed with small datasets or with small numbers of learned features. Here, we propose a large-scale feature learning system that enables us to carry out this experiment, learning 150,000 features from tens of millions of unlabeled images. Based on two scalable clustering algorithms (K-means and agglomerative clustering), we find that our simple system can discover features sensitive to a commonly occurring object class (human faces) and can also combine these into detectors invariant to significant global distortions like large translations and scale.
****

CPRL -- An Extension of Compressive Sensing to the Phase Retrieval Problem
Henrik Ohlsson, Allen Yang, Roy Dong, Shankar Sastry
While compressive sensing (CS) has been one of the most vibrant and active research fields in the past few years, most development only applies to linear models. This limits its application and excludes many areas where CS ideas could make a difference. This paper presents a novel extension of CS to the phase retrieval problem, where intensity measurements of a linear system are used to recover a complex sparse signal. We propose a novel solution using a lifting technique -- CPRL, which relaxes the NP-hard problem to a nonsmooth semidefinite program. Our analysis shows that CPRL inherits many desirable properties from CS, such as guarantees for exact recovery. We further provide scalable numerical solvers to accelerate its implementation. The source code of our algorithms will be provided to the public.
****

Learning optimal spike-based representations
Ralph Bourdoukan, David Barrett, Sophie Deneve, Christian K. Machens
How do neural networks learn to represent information? Here, we address this question by assuming that neural networks seek to generate an optimal population representation for a fixed linear decoder. We define a loss function for the quality of the population read-out and derive the dynamical equations for both neurons and synapses from the requirement to minimize this loss. The dynamical equations yield a network of integrate-and-fire neurons undergoing Hebbian plasticity. We show that, through learning, initially regular and highly correlated spike trains evolve towards Poisson-distributed and independent spike trains with much lower firing rates. The learning rule drives the network into an asynchronous, balanced regime where all inputs to the network are represented optimally for the given decoder. We show that the network dynamics and synaptic plasticity jointly balance the excitation and inhibition received by each unit as tightly as possible and, in doing so, minimize the prediction error between the inputs and the decoded outputs. In turn, spikes are only signalled whenever this prediction error exceeds a certain value, thereby implementing a predictive coding scheme. Our work suggests that several of the features reported in cortical networks, such as the high trial-to-trial variability, the balance between excitation and inhibition, and spike-timing dependent plasticity, are simply signatures of an efficient, spike-based code.
****

Collaborative Ranking With 17 Parameters
Maksims Volkovs, Richard Zemel
The primary application of collaborate filtering (CF) is to recommend a small se

t of items to a user, which entails ranking. Most approaches, however, formulate the CF problem as rating prediction, overlooking the ranking perspective. In this work we present a method for collaborative ranking that leverages the strengths of the two main CF approaches, neighborhood- and model-based. Our novel method is highly efficient, with only seventeen parameters to optimize and a single hyperparameter to tune, and beats the state-of-the-art collaborative ranking methods. We also show that parameters learned on one dataset yield excellent results on a very different dataset, without any retraining.
****

Small-Variance Asymptotics for Exponential Family Dirichlet Process Mixture Models

Ke Jiang, Brian Kulis, Michael Jordan

Links between probabilistic and non-probabilistic learning algorithms can arise by performing small-variance asymptotics, i.e., letting the variance of particular distributions in a graphical model go to zero. For instance, in the context of clustering, such an approach yields precise connections between the k-means and EM algorithms. In this paper, we explore small-variance asymptotics for exponential family Dirichlet process (DP) and hierarchical Dirichlet process (HDP) mixture models. Utilizing connections between exponential family distributions and Bregman divergences, we derive novel clustering algorithms from the asymptotic limit of the DP and HDP mixtures that feature the scalability of existing hard clustering methods as well as the flexibility of Bayesian nonparametric models. We focus on special cases of our analysis for discrete-data problems, including topic modeling, and we demonstrate the utility of our results by applying variants of our algorithms to problems arising in vision and document analysis.
****

Deep Representations and Codes for Image Auto-Annotation

Ryan Kiros, Csaba Szepesvári

The task of assigning a set of relevant tags to an image is challenging due to the size and variability of tag vocabularies. Consequently, most existing algorithms focus on tag assignment and fix an often large number of hand-crafted features to describe image characteristics. In this paper we introduce a hierarchical model for learning representations of full sized color images from the pixel level, removing the need for engineered feature representations and subsequent feature selection. We benchmark our model on the STL-10 recognition dataset, achieving state-of-the-art performance. When our features are combined with TagProp (Guillaumin et al.), we outperform or compete with existing annotation approaches that use over a dozen distinct image descriptors. Furthermore, using 256-bit codes and Hamming distance for training TagProp, we exchange only a small reduction in performance for efficient storage and fast comparisons. In our experiments, using deeper architectures always outperform shallow ones.
****

Patient Risk Stratification for Hospital-Associated C. diff as a Time-Series Classification Task

Jenna Wiens, Eric Horvitz, John Guttag

Requests for name changes in the electronic proceedings will be accepted with 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.
****

Meta-Gaussian Information Bottleneck

Melanie Rey, Volker Roth

We present a reformulation of the information bottleneck (IB) problem in terms of copula, using the equivalence between mutual information and negative copula entropy. Focusing on the Gaussian copula we extend the analytical IB solution available for the multivariate Gaussian case to distributions with a Gaussian dependence structure but arbitrary marginal densities, also called meta-Gaussian distributions. This opens new possibles applications of IB to continuous data and provides a solution more robust to outliers.
****

Efficient Monte Carlo Counterfactual Regret Minimization in Games with Many Player Actions
Neil Burch, Marc Lanctot, Duane Szafron, Richard Gibson
Counterfactual Regret Minimization (CFR) is a popular, iterative algorithm for computing strategies in extensive-form games. The Monte Carlo CFR (MCCFR) variants reduce the per iteration time cost of CFR by traversing a sampled portion of the tree. The previous most effective instances of MCCFR can still be very slow in games with many player actions since they sample every action for a given player. In this paper, we present a new MCCFR algorithm, Average Strategy Sampling (AS), that samples a subset of the player's actions according to the player's average strategy. Our new algorithm is inspired by a new, tighter bound on the number of iterations required by CFR to converge to a given solution quality. In addition, we prove a similar, tighter bound for AS and other popular MCCFR variants. Finally, we validate our work by demonstrating that AS converges faster than previous MCCFR algorithms in both no-limit poker and Bluff.
****
Fusion with Diffusion for Robust Visual Tracking
Yu Zhou, Xiang Bai, Wenyu Liu, Longin Latecki
A weighted graph is used as an underlying structure of many algorithms like semi-supervised learning and spectral clustering. The edge weights are usually deter-mined by a single similarity measure, but it often hard if not impossible to capture all relevant aspects of similarity when using a single similarity measure. In par-ticular, in the case of visual object matching it is beneficial to integrate different similarity measures that focus on different visual representations. In this paper, a novel approach to integrate multiple similarity measures is pro-posed. First pairs of similarity measures are combined with a diffusion process on their tensor product graph (TPG). Hence the diffused similarity of each pair of ob-jects becomes a function of joint diffusion of the two original similarities, which in turn depends on the neighborhood structure of the TPG. We call this process Fusion with Diffusion (FD). However, a higher order graph like the TPG usually means significant increase in time complexity. This is not the case in the proposed approach. A key feature of our approach is that the time complexity of the dif-fusion on the TPG is the same as the diffusion process on each of the original graphs, Moreover, it is not necessary to explicitly construct the TPG in our frame-work. Finally all diffused pairs of similarity measures are combined as a weighted sum. We demonstrate the advantages of the proposed approach on the task of visual tracking, where different aspects of the appearance similarity between the target object in frame t and target object candidates in frame t+1 are integrated. The obtained method is tested on several challenge video sequences and the experimental results show that it outperforms state-of-the-art tracking methods.
****
Convolutional-Recursive Deep Learning for 3D Object Classification
Richard Socher, Brody Huval, Bharath Bath, Christopher D. Manning, Andrew Ng
Recent advances in 3D sensing technologies make it possible to easily record color and depth images which together can improve object recognition. Most current methods rely on very well-designed features for this new 3D modality. We in- troduce a model based on a combination of convolutional and recursive neural networks (CNN and RNN) for learning features and classifying RGB-D images. The CNN layer learns low-level translationally invariant features which are then given as inputs to multiple, ■xed-tree RNNs in order to compose higher order fea- tures. RNNs can be seen as combining convolution and pooling into one ef■cient, hierarchical operation. Our main result is that even RNNs with random weights compose powerful features. Our model obtains state of the art performance on a standard RGB-D object dataset while being more accurate and faster during train- ing and testing than comparable architectures such as two-layer CNNs.
****
Transferring Expectations in Model-based Reinforcement Learning
Trung Nguyen, Tomi Silander, Tze Leong
We study how to automatically select and adapt multiple abstractions or represen

tations of the world to support model-based reinforcement learning. We address t
he challenges of transfer learning in heterogeneous environments with varying ta
sks. We present an efficient, online framework that, through a sequence of tasks
, learns a set of relevant representations to be used in future tasks. Without p
re-defined mapping strategies, we introduce a general approach to support transf
er learning across different state spaces. We demonstrate the potential impact o
f our system through improved jumpstart and faster convergence to near optimum p
olicy in two benchmark domains.
****

Modelling Reciprocating Relationships with Hawkes Processes
Charles Blundell, Jeff Beck, Katherine A. Heller
We present a Bayesian nonparametric model that discovers implicit social structu
re from interaction time-series data. Social groups are often formed implicitly,
 through actions among members of groups. Yet many models of social networks use
 explicitly declared relationships to infer social structure. We consider a part
icular class of Hawkes processes, a doubly stochastic point process, that is abl
e to model reciprocity between groups of individuals. We then extend the Infinit
e Relational Model by using these reciprocating Hawkes processes to parameterise
 its edges, making events associated with edges co-dependent through time. Our m
odel outperforms general, unstructured Hawkes processes as well as structured Po
isson process-based models at predicting verbal and email turn-taking, and milit
ary conflicts among nations.
****

Ancestor Sampling for Particle Gibbs
Fredrik Lindsten, Thomas Schön, Michael Jordan
We present a novel method in the family of particle MCMC methods that we refer t
o as particle Gibbs with ancestor sampling (PG-AS). Similarly to the existing PG
 with backward simulation (PG-BS) procedure, we use backward sampling to (consid
erably) improve the mixing of the PG kernel. Instead of using separate forward a
nd backward sweeps as in PG-BS, however, we achieve the same effect in a single
forward sweep. We apply the PG-AS framework to the challenging class of non-Mark
ovian state-space models. We develop a truncation strategy of these models that
is applicable in principle to any backward-simulation-based method, but which is
 particularly well suited to the PG-AS framework. In particular, as we show in a
 simulation study, PG-AS can yield an order-of-magnitude improved accuracy relat
ive to PG-BS due to its robustness to the truncation error. Several application
examples are discussed, including Rao-Blackwellized particle smoothing and infer
ence in degenerate state-space models.
****

Interpreting prediction markets: a stochastic approach
Rafael Frongillo, Nicholás Della Penna, Mark D. Reid
We strengthen recent connections between prediction markets and learning by show
ing that a natural class of market makers can be understood as performing stocha
stic mirror descent when trader demands are sequentially drawn from a fixed dist
ribution. This provides new insights into how market prices (and price paths) ma
y be interpreted as a summary of the market's belief distribution by relating th
em to the optimization problem being solved. In particular, we show that the sta
tionary point of the stochastic process of prices generated by the market is equ
al to the market's Walrasian equilibrium of classic market analysis. Together, t
hese results suggest how traditional market making mechanisms might be replaced
with general purpose learning algorithms while still retaining guarantees about
their behaviour.
****

Nonparanormal Belief Propagation (NPNBP)
Gal Elidan, Cobi Cario
The empirical success of the belief propagation approximate inference algorithm
has inspired numerous theoretical and algorithmic advances. Yet, for continuous
non-Gaussian domains performing belief propagation remains a challenging task: r
ecent innovations such as nonparametric or kernel belief propagation, while usef
ul, come with a substantial computational cost and offer little theoretical guar

antees, even for tree structured models. In this work we present Nonparanormal BP for performing efficient inference on distributions parameterized by a Gaussian copulas network and any univariate marginals. For tree structured networks, our approach is guaranteed to be exact for this powerful class of non-Gaussian models. Importantly, the method is as efficient as standard Gaussian BP, and its convergence properties do not depend on the complexity of the univariate marginals, even when a nonparametric representation is used.
****

Adaptive Stratified Sampling for Monte-Carlo integration of Differentiable functions
Alexandra Carpentier, Rémi Munos
We consider the problem of adaptive stratified sampling for Monte Carlo integration of a differentiable function given a finite number of evaluations to the function. We construct a sampling scheme that samples more often in regions where the function oscillates more, while allocating the samples such that they are well spread on the domain (this notion shares similitude with low discrepancy). We prove that the estimate returned by the algorithm is almost as accurate as the estimate that an optimal oracle strategy (that would know the variations of the function everywhere) would return, and we provide a finite-sample analysis.
****

Bayesian Nonparametric Modeling of Suicide Attempts
Francisco Ruiz, Isabel Valera, Carlos Blanco, Fernando Pérez-Cruz
The National Epidemiologic Survey on Alcohol and Related Conditions (NESARC) database contains a large amount of information, regarding the way of life, medical conditions, depression, etc., of a representative sample of the U.S. population. In the present paper, we are interested in seeking the hidden causes behind the suicide attempts, for which we propose to model the subjects using a nonparametric latent model based on the Indian Buffet Process (IBP). Due to the nature of the data, we need to adapt the observation model for discrete random variables. We propose a generative model in which the observations are drawn from a multinomial-logit distribution given the IBP matrix. The implementation of an efficient Gibbs sampler is accomplished using the Laplace approximation, which allows us to integrate out the weighting factors of the multinomial-logit likelihood model. Finally, the experiments over the NESARC database show that our model properly captures some of the hidden causes that model suicide attempts.
****

Non-linear Metric Learning
Dor Kedem, Stephen Tyree, Fei Sha, Gert Lanckriet, Kilian Q. Weinberger
In this paper, we introduce two novel metric learning algorithms, $\chi 2$-LMNN and GB-LMNN, which are explicitly designed to be non-linear and easy-to-use. The two approaches achieve this goal in fundamentally different ways: $\chi 2$-LMNN inherits the computational benefits of a linear mapping from linear metric learning, but uses a non-linear $\chi 2$-distance to explicitly capture similarities within histogram data sets; GB-LMNN applies gradient-boosting to learn non-linear mappings directly in function space and takes advantage of this approach's robustness, speed, parallelizability and insensitivity towards the single additional hyper-parameter. On various benchmark data sets, we demonstrate these methods not only match the current state-of-the-art in terms of kNN classification error, but in the case of $\chi 2$-LMNN, obtain best results in 19 out of 20 learning settings.
****

Putting Bayes to sleep
Dmitry Adamskiy, Manfred K. K. Warmuth, Wouter M. Koolen
We consider sequential prediction algorithms that are given the predictions from a set of models as inputs. If the nature of the data is changing over time in that different models predict well on different segments of the data, then adaptivity is typically achieved by mixing into the weights in each round a bit of the initial prior (kind of like a weak restart). However, what if the favored models in each segment are from a small subset, i.e. the data is likely to be predicted well by models that predicted well before? Curiously, fitting such ''sparse composite models'' is achieved by mixing in a bit of all the past posteriors. Thi

s self-referential updating method is rather peculiar, but it is efficient and g
ives superior performance on many natural data sets. Also it is important becaus
e it introduces a long-term memory: any model that has done well in the past can
 be recovered quickly. While Bayesian interpretations can be found for mixing in
 a bit of the initial prior, no Bayesian interpretation is known for mixing in p
ast posteriors. We build atop the ''specialist'' framework from the online lear
ning literature to give the Mixing Past Posteriors update a proper Bayesian foun
dation. We apply our method to a well-studied multitask learning problem and obt
ain a new intriguing efficient update that achieves a significantly better bound
.
****

Sparse Approximate Manifolds for Differential Geometric MCMC
Ben Calderhead, Mátyás Sustik
One of the enduring challenges in Markov chain Monte Carlo methodology is the de
velopment of proposal mechanisms to make moves distant from the current point, t
hat are accepted with high probability and at low computational cost. The recent
 introduction of locally adaptive MCMC methods based on the natural underlying R
iemannian geometry of such models goes some way to alleviating these problems fo
r certain classes of models for which the metric tensor is analytically tractabl
e, however computational efficiency is not assured due to the necessity of poten
tially high-dimensional matrix operations at each iteration. In this paper we fi
rstly investigate a sampling-based approach for approximating the metric tensor
and suggest a valid MCMC algorithm that extends the applicability of Riemannian
Manifold MCMC methods to statistical models that do not admit an analytically co
mputable metric tensor. Secondly, we show how the approximation scheme we consid
er naturally motivates the use of l1 regularisation to improve estimates and obt
ain a sparse approximate inverse of the metric, which enables stable and sparse
approximations of the local geometry to be made. We demonstrate the application
of this algorithm for inferring the parameters of a realistic system of ordinary
 differential equations using a biologically motivated robust student-t error mo
del, for which the expected Fisher Information is analytically intractable.
****

Bayesian active learning with localized priors for fast receptive field characte
rization
Mijung Park, Jonathan Pillow
Active learning can substantially improve the yield of neurophysiology experimen
ts by adaptively selecting stimuli to probe a neuron's receptive field (RF) in r
eal time. Bayesian active learning methods maintain a posterior distribution ove
r the RF, and select stimuli to maximally reduce posterior entropy on each time
step. However, existing methods tend to rely on simple Gaussian priors, and do
not exploit uncertainty at the level of hyperparameters when determining an opti
mal stimulus. This uncertainty can play a substantial role in RF characterizati
on, particularly when RFs are smooth, sparse, or local in space and time. In th
is paper, we describe a novel framework for active learning under hierarchical,
conditionally Gaussian priors. Our algorithm uses sequential Markov Chain Monte
 Carlo sampling (''particle filtering'' with MCMC) over hyperparameters to const
ruct a mixture-of-Gaussians representation of the RF posterior, and selects opti
mal stimuli using an approximate infomax criterion. The core elements of this a
lgorithm are parallelizable, making it computationally efficient for real-time e
xperiments. We apply our algorithm to simulated and real neural data, and show
that it can provide highly accurate receptive field estimates from very limited
data, even with a small number of hyperparameter samples.
****

Deep Neural Networks Segment Neuronal Membranes in Electron Microscopy Images
Dan Ciresan, Alessandro Giusti, Luca Gambardella, Jürgen Schmidhuber
Requests for name changes in the electronic proceedings will be accepted with 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 from the Wisdom of Crowds by Minimax Entropy
Dengyong Zhou, Sumit Basu, Yi Mao, John Platt
An important way to make large training sets is to gather noisy labels from crowds of nonexperts. We propose a minimax entropy principle to improve the quality of these labels. Our method assumes that labels are generated by a probability distribution over workers, items, and labels. By maximizing the entropy of this distribution, the method naturally infers item confusability and worker expertise. We infer the ground truth by minimizing the entropy of this distribution, which we show minimizes the Kullback-Leibler (KL) divergence between the probability distribution and the unknown truth. We show that a simple coordinate descent scheme can optimize minimax entropy. Empirically, our results are substantially better than previously published methods for the same problem.
****
Parametric Local Metric Learning for Nearest Neighbor Classification
Jun Wang, Alexandros Kalousis, Adam Woznica
We study the problem of learning local metrics for nearest neighbor classification. Most previous works on local metric learning learn a number of local unrelated metrics. While this ''independence'' approach delivers an increased flexibility its downside is the considerable risk of overfitting. We present a new parametric local metric learning method in which we learn a smooth metric matrix function over the data manifold. Using an approximation error bound of the metric matrix function we learn local metrics as linear combinations of basis metrics defined on anchor points over different regions of the instance space. We constrain the metric matrix function by imposing on the linear combinations manifold regularization which makes the learned metric matrix function vary smoothly along the geodesics of the data manifold. Our metric learning method has excellent performance both in terms of predictive power and scalability. We experimented with several large-scale classification problems, tens of thousands of instances, and compared it with several state of the art metric learning methods, both global and local, as well as to SVM with automatic kernel selection, all of which it outperforms in a significant manner.
****
The topographic unsupervised learning of natural sounds in the auditory cortex
Hiroki Terashima, Masato Okada
The computational modelling of the primary auditory cortex (A1) has been less fruitful than that of the primary visual cortex (V1) due to the less organized properties of A1. Greater disorder has recently been demonstrated for the tonotopy of A1 that has traditionally been considered to be as ordered as the retinotopy of V1. This disorder appears to be incongruous, given the uniformity of the neocortex; however, we hypothesized that both A1 and V1 would adopt an efficient coding strategy and that the disorder in A1 reflects natural sound statistics. To provide a computational model of the tonotopic disorder in A1, we used a model that was originally proposed for the smooth V1 map. In contrast to natural images, natural sounds exhibit distant correlations, which were learned and reflected in the disordered map. The auditory model predicted harmonic relationships among neighbouring A1 cells; furthermore, the same mechanism used to model V1 complex cells reproduced nonlinear responses similar to the pitch selectivity. These results contribute to the understanding of the sensory cortices of different modalities in a novel and integrated manner.
****
Efficient Sampling for Bipartite Matching Problems
Maksims Volkovs, Richard Zemel
Bipartite matching problems characterize many situations, ranging from ranking in information retrieval to correspondence in vision. Exact inference in real-world applications of these problems is intractable, making efficient approximation methods essential for learning and inference. In this paper we propose a novel {\it sequential matching} sampler based on the generalization of the Plackett-Luce model, which can effectively make large moves in the space of matchings. This allows the sampler to match the difficult target distributions common in these problems: highly multimodal distributions with well separated modes. We present

experimental results with bipartite matching problems - ranking and image corres
pondence - which show that the sequential matching sampler efficiently approxima
tes the target distribution, significantly outperforming other sampling approach
es.
****
Volume Regularization for Binary Classification
Koby Crammer, Tal 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-auth
ors prior to requesting a name change in the electronic proceedings.
****
Causal discovery with scale-mixture model for spatiotemporal variance dependenci
es
Zhitang Chen, Kun Zhang, Laiwan Chan
In conventional causal discovery, structural equation models (SEM) are directly
 applied to the observed variables, meaning that the causal effect can be represe
nted as a function of the direct causes themselves. However, in many real world
 problems, there are significant dependencies in the variances or energies, which
 indicates that causality may possibly take place at the level of variances or e
nergies. In this paper, we propose a probabilistic causal scale-mixture model wi
th spatiotemporal variance dependencies to represent a specific type of generati
ng mechanism of the observations. In particular, the causal mechanism including
 contemporaneous and temporal causal relations in variances or energies is repres
ented by a Structural Vector AutoRegressive model (SVAR). We prove the identifia
bility of this model under the non-Gaussian assumption on the innovation process
es. We also propose algorithms to estimate the involved parameters and discover
 the contemporaneous causal structure. Experiments on synthesis and real world da
ta are conducted to show the applicability of the proposed model and algorithms.
****
Neurally Plausible Reinforcement Learning of Working Memory Tasks
Jaldert Rombouts, Pieter Roelfsema, Sander Bohte
A key function of brains is undoubtedly the abstraction and maintenance of infor
mation from the environment for later use. Neurons in association cortex play an
 important role in this process: during learning these neurons become tuned to r
elevant features and represent the information that is required later as a persi
stent elevation of their activity. It is however not well known how these neuron
s acquire their task-relevant tuning. Here we introduce a biologically plausible
 learning scheme that explains how neurons become selective for relevant informa
tion when animals learn by trial and error. We propose that the action selection
 stage feeds back attentional signals to earlier processing levels. These feedba
ck signals interact with feedforward signals to form synaptic tags at those conn
ections that are responsible for the stimulus-response mapping. A globally relea
sed neuromodulatory signal interacts with these tagged synapses to determine the
 sign and strength of plasticity. The learning scheme is generic because it can
 train networks in different tasks, simply by varying inputs and rewards. It expl
ains how neurons in association cortex learn to (1) temporarily store task-relev
ant information in non-linear stimulus-response mapping tasks and (2) learn to o
ptimally integrate probabilistic evidence for perceptual decision making.
****
Simultaneously Leveraging Output and Task Structures for Multiple-Output Regress
ion
Piyush Rai, Abhishek Kumar, Hal Daume
Multiple-output regression models require estimating multiple functions, one for
 each output. To improve parameter estimation in such models, methods based on s
tructural regularization of the model parameters are usually needed. In this pap
er, we present a multiple-output regression model that leverages the covariance
 structure of the functions (i.e., how the multiple functions are related with ea
ch other) as well as the conditional covariance structure of the outputs. This i
s in contrast with existing methods that usually take into account only one of t

hese structures. More importantly, unlike most of the other existing methods, no ne of these structures need be known a priori in our model, and are learned from the data. Several previously proposed structural regularization based multiple -output regression models turn out to be special cases of our model. Moreover, i n addition to being a rich model for multiple-output regression, our model can a lso be used in estimating the graphical model structure of a set of variables (m ultivariate outputs) conditioned on another set of variables (inputs). Experimen tal results on both synthetic and real datasets demonstrate the effectiveness of our method.
****

Feature Clustering for Accelerating Parallel Coordinate Descent
Chad Scherrer, Ambuj Tewari, Mahantesh Halappanavar, David Haglin
Requests for name changes in the electronic proceedings will be accepted with 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.
****

Phoneme Classification using Constrained Variational Gaussian Process Dynamical System
Hyunsin Park, Sungrack Yun, Sanghyuk Park, Jongmin Kim, Chang Yoo
This paper describes a new acoustic model based on variational Gaussian process dynamical system (VGPDS) for phoneme classification. The proposed model overcome s the limitations of the classical HMM in modeling the real speech data, by adop ting a nonlinear and nonparametric model. In our model, the GP prior on the dyna mics function enables representing the complex dynamic structure of speech, whil e the GP prior on the emission function successfully models the global dependenc y over the observations. Additionally, we introduce variance constraint to the o riginal VGPDS for mitigating sparse approximation error of the kernel matrix. Th e effectiveness of the proposed model is demonstrated with extensive experimenta l results including parameter estimation, classification performance on the synt hetic and benchmark datasets.
****

Fast Variational Inference in the Conjugate Exponential Family
James Hensman, Magnus Rattray, Neil Lawrence
We present a general method for deriving collapsed variational inference algorit hms for probabilistic models in the conjugate exponential family. Our method uni fies many existing approaches to collapsed variational inference. Our collapsed variational inference leads to a new lower bound on the marginal likelihood. We exploit the information geometry of the bound to derive much faster optimization methods based on conjugate gradients for these models. Our approach is very gen eral and is easily applied to any model where the mean field update equations ha ve been derived. Empirically we show significant speed-ups for probabilistic mod els optimized using our bound.
****

Identifiability and Unmixing of Latent Parse Trees
Daniel J. Hsu, Sham M. Kakade, Percy S. Liang
This paper explores unsupervised learning of parsing models along two directions . First, which models are identifiable from infinite data? We use a general te chnique for numerically checking identifiability based on the rank of a Jacobian matrix, and apply it to several standard constituency and dependency parsing mo dels. Second, for identifiable models, how do we estimate the parameters effici ently? EM suffers from local optima, while recent work using spectral methods c annot be directly applied since the topology of the parse tree varies across sen tences. We develop a strategy, unmixing, which deals with this additional compl exity for restricted classes of parsing models.
****

On the (Non-)existence of Convex, Calibrated Surrogate Losses for Ranking
Clément Calauzènes, Nicolas Usunier, Patrick Gallinari
We study surrogate losses for learning to rank, in a framework where the ranking s are induced by scores and the task is to learn the scoring function. We focus

on the calibration of surrogate losses with respect to a ranking evaluation metr
ic, where the calibration is equivalent to the guarantee that near-optimal value
s of the sur- rogate risk imply near-optimal values of the risk de■ned by the ev
aluation metric. We prove that if a surrogate loss is a convex function of the s
cores, then it is not calibrated with respect to two evaluation metrics widely u
sed for search engine evaluation, namely the Average Precision and the Expected
Reciprocal Rank. We also show that such convex surrogate losses cannot be calibr
ated with respect to the Pairwise Disagreement, an evaluation metric used when l
earning from pair- wise preferences. Our results cast lights on the intrinsic di
f■culty of some ranking problems, as well as on the limitations of learning-to-r
ank algorithms based on the minimization of a convex surrogate risk.
****

Learning with Partially Absorbing Random Walks
Xiao-ming Wu, Zhenguo Li, Anthony So, John Wright, Shih-fu Chang
Requests for name changes in the electronic proceedings will be accepted with 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.
****

Relax and Randomize : From Value to Algorithms
Sasha Rakhlin, Ohad Shamir, Karthik Sridharan
We show a principled way of deriving online learning algorithms from a minimax a
nalysis. Various upper bounds on the minimax value, previously thought to be non
-constructive, are shown to yield algorithms. This allows us to seamlessly recov
er known methods and to derive new ones, also capturing such ''unorthodox'' meth
ods as Follow the Perturbed Leader and the R^2 forecaster. Understanding the inh
erent complexity of the learning problem thus leads to the development of algori
thms. To illustrate our approach, we present several new algorithms, including a
family of randomized methods that use the idea of a ''random play out''. New ve
rsions of the Follow-the-Perturbed-Leader algorithms are presented, as well as m
ethods based on the Littlestone's dimension, efficient methods for matrix comple
tion with trace norm, and algorithms for the problems of transductive learning a
nd prediction with static experts.
****

Inverse Reinforcement Learning through Structured Classification
Edouard Klein, Matthieu Geist, Bilal Piot, Olivier Pietquin
This paper adresses the inverse reinforcement learning (IRL) problem, that is in
ferring a reward for which a demonstrated expert behavior is optimal. We introdu
ce a new algorithm, SCIRL, whose principle is to use the so-called feature expec
tation of the expert as the parameterization of the score function of a multi-cl
ass classifier. This approach produces a reward function for which the expert po
licy is provably near-optimal. Contrary to most of existing IRL algorithms, SCIR
L does not require solving the direct RL problem. Moreover, with an appropriate
heuristic, it can succeed with only trajectories sampled according to the expert
behavior. This is illustrated on a car driving simulator.
****

Efficient and direct estimation of a neural subunit model for sensory coding
Brett Vintch, Andrew Zaharia, J Movshon, Eero Simoncelli
Many visual and auditory neurons have response properties that are well explaine
d by pooling the rectified responses of a set of self-similar linear filters. Th
ese filters cannot be found using spike-triggered averaging (STA), which estimat
es only a single filter. Other methods, like spike-triggered covariance (STC), d
efine a multi-dimensional response subspace, but require substantial amounts of
data and do not produce unique estimates of the linear filters. Rather, they pro
vide a linear basis for the subspace in which the filters reside. Here, we defin
e a 'subunit' model as an LN-LN cascade, in which the first linear stage is rest
ricted to a set of shifted ("convolutional") copies of a common filter, and the
first nonlinear stage consists of rectifying nonlinearities that are identical f
or all filter outputs; we refer to these initial LN elements as the 'subunits' o
f the receptive field. The second linear stage then computes a weighted sum of t

he responses of the rectified subunits. We present a method for directly fitting this model to spike data. The method performs well for both simulated and real data (from primate V1), and the resulting model outperforms STA and STC in terms of both cross-validated accuracy and efficiency.
****
Discriminative Learning of Sum-Product Networks
Robert Gens, Pedro Domingos
Sum-product networks are a new deep architecture that can perform fast, exact in- ference on high-treewidth models. Only generative methods for training SPNs have been proposed to date. In this paper, we present the ■rst discriminative training algorithms for SPNs, combining the high accuracy of the former with the representational power and tractability of the latter. We show that the class of tractable discriminative SPNs is broader than the class of tractable generative ones, and propose an ef■cient backpropagation-style algorithm for computing the gradient of the conditional log likelihood. Standard gradient descent suffers from the diffusion problem, but networks with many layers can be learned reliably using “hard” gradient descent, where marginal inference is replaced by MPE inference (i.e., inferring the most probable state of the non-evidence variables). The resulting updates have a simple and intuitive form. We test discriminative SPNs on standard image classi■cation tasks. We obtain the best results to date on the CIFAR-10 dataset, using fewer features than prior methods with an SPN architec- ture that learns local image structure discriminatively. We also report the highest published test accuracy on STL-10 even though we only use the labeled portion of the dataset.
****
Stochastic optimization and sparse statistical recovery: Optimal algorithms for high dimensions
Alekh Agarwal, Sahand Negahban, 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.
****
Bandit Algorithms boost Brain Computer Interfaces for motor-task selection of a brain-controlled button
Joan Fruitet, Alexandra Carpentier, Maureen Clerc, Rémi Munos
A brain-computer interface (BCI) allows users to “communicate” with a computer without using their muscles. BCI based on sensori-motor rhythms use imaginary motor tasks, such as moving the right or left hand to send control signals. The performances of a BCI can vary greatly across users but also depend on the tasks used, making the problem of appropriate task selection an important issue. This study presents a new procedure to automatically select as fast as possible a discriminant motor task for a brain-controlled button. We develop for this purpose an adaptive algorithm UCB-classif based on the stochastic bandit theory. This shortens the training stage, thereby allowing the exploration of a greater variety of tasks. By not wasting time on inefficient tasks, and focusing on the most promising ones, this algorithm results in a faster task selection and a more efficient use of the BCI training session. Comparing the proposed method to the standard practice in task selection, for a fixed time budget, UCB-classif leads to an improve classification rate, and for a fix classification rate, to a reduction of the time spent in training by 50%.
****
Localizing 3D cuboids in single-view images
Jianxiong Xiao, Bryan Russell, Antonio Torralba
In this paper we seek to detect rectangular cuboids and localize their corners in uncalibrated single-view images depicting everyday scenes. In contrast to recent approaches that rely on detecting vanishing points of the scene and grouping line segments to form cuboids, we build a discriminative parts-based detector that models the appearance of the cuboid corners and internal edges while enforcing consistency to a 3D cuboid model. Our model is invariant to the different 3D v

iewpoints and aspect ratios and is able to detect cuboids across many different object categories. We introduce a database of images with cuboid annotations tha t spans a variety of indoor and outdoor scenes and show qualitative and quantita tive results on our collected database. Our model out-performs baseline detector s that use 2D constraints alone on the task of localizing cuboid corners.
****

A Convex Formulation for Learning Scale-Free Networks via Submodular Relaxation
Aaron Defazio, Tibério Caetano
A key problem in statistics and machine learning is the determination of network structure from data. We consider the case where the structure of the graph to b e reconstructed is known to be scale-free. We show that in such cases it is natu ral to formulate structured sparsity inducing priors using submodular functions, and we use their Lovasz extension to obtain a convex relaxation. For tractable classes such as Gaussian graphical models, this leads to a convex optimization p roblem that can be efficiently solved. We show that our method results in an imp rovement in the accuracy of reconstructed networks for synthetic data. We also s how how our prior encourages scale-free reconstructions on a bioinfomatics datas et.
****

Hamming Distance Metric Learning
Mohammad Norouzi, David J. Fleet, Russ R. Salakhutdinov
Motivated by large-scale multimedia applications we propose to learn mappings fr om high-dimensional data to binary codes that preserve semantic similarity. Bina ry codes are well suited to large-scale applications as they are storage efficie nt and permit exact sub-linear kNN search. The framework is applicable to broad families of mappings, and uses a flexible form of triplet ranking loss. We overc ome discontinuous optimization of the discrete mappings by minimizing a piecewis e-smooth upper bound on empirical loss, inspired by latent structural SVMs. We develop a new loss-augmented inference algorithm that is quadratic in the code l ength. We show strong retrieval performance on CIFAR-10 and MNIST, with promisi ng classification results using no more than kNN on the binary codes.
****

Co-Regularized Hashing for Multimodal Data
Yi Zhen, Dit-Yan Yeung
Hashing-based methods provide a very promising approach to large-scale similarit y search. To obtain compact hash codes, a recent trend seeks to learn the hash functions from data automatically. In this paper, we study hash function learni ng in the context of multimodal data. We propose a novel multimodal hash functi on learning method, called Co-Regularized Hashing (CRH), based on a boosted co-r egularization framework. The hash functions for each bit of the hash codes are learned by solving DC (difference of convex functions) programs, while the learn ing for multiple bits proceeds via a boosting procedure so that the bias introdu ced by the hash functions can be sequentially minimized. We empirically compare CRH with two state-of-the-art multimodal hash function learning methods on two publicly available data sets.
****

The Coloured Noise Expansion and Parameter Estimation of Diffusion Processes
Simon Lyons, Amos J. Storkey, Simo Särkkä
Stochastic differential equations (SDE) are a natural tool for modelling systems that are inherently noisy or contain uncertainties that can be modelled as stoc hastic processes. Crucial to the process of using SDE to build mathematical mode ls is the ability to estimate parameters of those models from observed data. Ove r the past few decades, significant progress has been made on this problem, but we are still far from having a definitive solution. We describe a novel method o f approximating a diffusion process that we show to be useful in Markov chain Mo nte-Carlo (MCMC) inference algorithms. We take the ‘white’ noise that drives a d iffusion process and decompose it into two terms. The first is a ‘coloured noise ’ term that can be deterministically controlled by a set of auxilliary variables . The second term is small and enables us to form a linear Gaussian ‘small noise ’ approximation. The decomposition allows us to take a diffusion process of inte

rest and cast it in a form that is amenable to sampling by MCMC methods. We expl ain why many state-of-the-art inference methods fail on highly nonlinear inferen ce problems. We demonstrate experimentally that our method performs well in such situations. Our results show that this method is a promising new tool for use i n inference and parameter estimation problems.
****

How Prior Probability Influences Decision Making: A Unifying Probabilistic Model
Yanping Huang, Timothy Hanks, Mike Shadlen, Abram L. Friesen, Rajesh PN Rao
How does the brain combine prior knowledge with sensory evidence when making dec isions under uncertainty? Two competing descriptive models have been proposed ba sed on experimental data. The first posits an additive offset to a decision var iable, implying a static effect of the prior. However, this model is inconsisten t with recent data from a motion discrimination task involving temporal integrat ion of uncertain sensory evidence. To explain this data, a second model has been proposed which assumes a time-varying influence of the prior. Here we present a normative model of decision making that incorporates prior knowledge in a princ ipled way. We show that the additive offset model and the time-varying prior mo del emerge naturally when decision making is viewed within the framework of part ially observable Markov decision processes (POMDPs). Decision making in the mod el reduces to (1) computing beliefs given observations and prior information in a Bayesian manner, and (2) selecting actions based on these beliefs to maximize the expected sum of future rewards. We show that the model can explain both da ta previously explained using the additive offset model as well as more recent data on the time-varying influence of prior knowledge on decision making.
****

Structured Learning of Gaussian Graphical Models
Karthik Mohan, Mike Chung, Seungyeop Han, Daniela Witten, Su-in Lee, Maryam Faze l
We consider estimation of multiple high-dimensional Gaussian graphical models co rresponding to a single set of nodes under several distinct conditions. We assum e that most aspects of the networks are shared, but that there are some structur ed differences between them. Specifically, the network differences are generated from node perturbations: a few nodes are perturbed across networks, and most or all edges stemming from such nodes differ between networks. This corresponds to a simple model for the mechanism underlying many cancers, in which the gene reg ulatory network is disrupted due to the aberrant activity of a few specific gen es. We propose to solve this problem using the structured joint graphical lasso, a convex optimization problem that is based upon the use of a novel symmetric o verlap norm penalty, which we solve using an alternating directions method of m ultipliers algorithm. Our proposal is illustrated on synthetic data and on an ap plication to brain cancer gene expression data.
****

Entropy Estimations Using Correlated Symmetric Stable Random Projections
Ping Li, Cun-hui Zhang
Methods for efficiently estimating the Shannon entropy of data streams have impo rtant applications in learning, data mining, and network anomaly detections (e.g ., the DDoS attacks). For nonnegative data streams, the method of Compressed Cou nting (CC) based on maximally-skewed stable random projections can provide accur ate estimates of the Shannon entropy using small storage. However, CC is no long er applicable when entries of data streams can be below zero, which is a common scenario when comparing two streams. In this paper, we propose an algorithm for entropy estimation in general data streams which allow negative entries. In our method, the Shannon entropy is approximated by the finite difference of two corr elated frequency moments estimated from correlated samples of symmetric stable r andom variables. Our experiments confirm that this method is able to substantial ly better approximate the Shannon entropy compared to the prior state-of-the-art .
****

Coding efficiency and detectability of rate fluctuations with non-Poisson neuron al firing

Shinsuke Koyama
Statistical features of neuronal spike trains are known to be non-Poisson. Here, we investigate the extent to which the non-Poissonian feature affects the efficiency of transmitting information on fluctuating firing rates. For this purpose, we introduce the Kullbuck-Leibler (KL) divergence as a measure of the efficiency of information encoding, and assume that spike trains are generated by time-rescaled renewal processes. We show that the KL divergence determines the lower bound of the degree of rate fluctuations below which the temporal variation of the firing rates is undetectable from sparse data. We also show that the KL divergence, as well as the lower bound, depends not only on the variability of spikes in terms of the coefficient of variation, but also significantly on the higher-order moments of interspike interval (ISI) distributions. We examine three specific models that are commonly used for describing the stochastic nature of spikes (the gamma, inverse Gaussian (IG) and lognormal ISI distributions), and find that the time-rescaled renewal process with the IG distribution achieves the largest KL divergence, followed by the lognormal and gamma distributions.
****
Nyström Method vs Random Fourier Features: A Theoretical and Empirical Comparison
Tianbao Yang, Yu-feng Li, Mehrdad Mahdavi, Rong Jin, Zhi-Hua Zhou
Both random Fourier features and the Nyström method have been successfully applied to efficient kernel learning. In this work, we investigate the fundamental difference between these two approaches, and how the difference could affect their generalization performances. Unlike approaches based on random Fourier features where the basis functions (i.e., cosine and sine functions) are sampled from a distribution {\it independent} from the training data, basis functions used by the Nyström method are randomly sampled from the training examples and are therefore {\it data dependent}. By exploring this difference, we show that when there is a large gap in the eigen-spectrum of the kernel matrix, approaches based the Nyström method can yield impressively better generalization error bound than random Fourier features based approach. We empirically verify our theoretical findings on a wide range of large data sets.
****
Spiking and saturating dendrites differentially expand single neuron computation capacity
Romain Cazé, Mark Humphries, Boris Gutkin
The integration of excitatory inputs in dendrites is non-linear: multiple excitatory inputs can produce a local depolarization departing from the arithmetic sum of each input's response taken separately. If this depolarization is bigger than the arithmetic sum, the dendrite is spiking; if the depolarization is smaller, the dendrite is saturating. Decomposing a dendritic tree into independent dendritic spiking units greatly extends its computational capacity, as the neuron then maps onto a two layer neural network, enabling it to compute linearly non-separable Boolean functions (lnBFs). How can these lnBFs be implemented by dendritic architectures in practise? And can saturating dendrites equally expand computational capacity? To adress these questions we use a binary neuron model and Boolean algebra. First, we confirm that spiking dendrites enable a neuron to compute lnBFs using an architecture based on the disjunctive normal form (DNF). Second, we prove that saturating dendrites as well as spiking dendrites also enable a neuron to compute lnBFs using an architecture based on the conjunctive normal form (CNF). Contrary to the DNF-based architecture, a CNF-based architecture leads to a dendritic unit tuning that does not imply the neuron tuning, as has been observed experimentally. Third, we show that one cannot use a DNF-based architecture with saturating dendrites. Consequently, we show that an important family of lnBFs implemented with a CNF-architecture can require an exponential number of saturating dendritic units, whereas the same family implemented with either a DNF-architecture or a CNF-architecture always require a linear number of spiking dendritic unit. This minimization could explain why a neuron spends energetic resources to make its dendrites spike.
****

Active Learning of Model Evidence Using Bayesian Quadrature
Michael Osborne, Roman Garnett, Zoubin Ghahramani, David K. Duvenaud, Stephen J. Roberts, Carl Rasmussen
Numerical integration is an key component of many problems in scientific computing, statistical modelling, and machine learning. Bayesian Quadrature is a model-based method for numerical integration which, relative to standard Monte Carlo methods, offers increased sample efficiency and a more robust estimate of the uncertainty in the estimated integral. We propose a novel Bayesian Quadrature approach for numerical integration when the integrand is non-negative, such as the case of computing the marginal likelihood, predictive distribution, or normalising constant of a probabilistic model. Our approach approximately marginalises the quadrature model's hyperparameters in closed form, and introduces an active learning scheme to optimally select function evaluations, as opposed to using Monte Carlo samples. We demonstrate our method on both a number of synthetic benchmarks and a real scientific problem from astronomy.
****
Diffusion Decision Making for Adaptive k-Nearest Neighbor Classification
Yung-kyun Noh, Frank Park, Daniel Lee
This paper sheds light on some fundamental connections of the diffusion decision making model of neuroscience and cognitive psychology with k-nearest neighbor classification. We show that conventional k-nearest neighbor classification can be viewed as a special problem of the diffusion decision model in the asymptotic situation. Applying the optimal strategy associated with the diffusion decision model, an adaptive rule is developed for determining appropriate values of k in k-nearest neighbor classification. Making use of the sequential probability ratio test (SPRT) and Bayesian analysis, we propose five different criteria for adaptively acquiring nearest neighbors. Experiments with both synthetic and real datasets demonstrate the effectivness of our classification criteria.
****
Bayesian Hierarchical Reinforcement Learning
Feng Cao, Soumya Ray
We describe an approach to incorporating Bayesian priors in the maxq framework for hierarchical reinforcement learning (HRL). We define priors on the primitive environment model and on task pseudo-rewards. Since models for composite tasks can be complex, we use a mixed model-based/model-free learning approach to find an optimal hierarchical policy. We show empirically that (i) our approach results in improved convergence over non-Bayesian baselines, given sensible priors, (ii) task hierarchies and Bayesian priors can be complementary sources of information, and using both sources is better than either alone, (iii) taking advantage of the structural decomposition induced by the task hierarchy significantly reduces the computational cost of Bayesian reinforcement learning and (iv) in this framework, task pseudo-rewards can be learned instead of being manually specified, leading to automatic learning of hierarchically optimal rather than recursively optimal policies.
****
Compressive Sensing MRI with Wavelet Tree Sparsity
Chen Chen, Junzhou Huang
Requests for name changes in the electronic proceedings will be accepted with 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.
****
Mixability in Statistical Learning
Tim Erven, Peter Grünwald, Mark D. Reid, Robert C. Williamson
Statistical learning and sequential prediction are two different but related formalisms to study the quality of predictions. Mapping out their relations and transferring ideas is an active area of investigation. We provide another piece of the puzzle by showing that an important concept in sequential prediction, the mixability of a loss, has a natural counterpart in the statistical setting, which we call stochastic mixability. Just as ordinary mixability characterizes fast ra

tes for the worst-case regret in sequential prediction, stochastic mixability ch
aracterizes fast rates in statistical learning. We show that, in the special cas
e of log-loss, stochastic mixability reduces to a well-known (but usually unname
d) martingale condition, which is used in existing convergence theorems for mini
mum description length and Bayesian inference. In the case of 0/1-loss, it reduc
es to the margin condition of Mammen and Tsybakov, and in the case that the mode
l under consideration contains all possible predictors, it is equivalent to ordi
nary mixability.
****

Symmetric Correspondence Topic Models for Multilingual Text Analysis
Kosuke Fukumasu, Koji Eguchi, Eric Xing
Topic modeling is a widely used approach to analyzing large text collections. A
small number of multilingual topic models have recently been explored to discove
r latent topics among parallel or comparable documents, such as in Wikipedia. Ot
her topic models that were originally proposed for structured data are also appl
icable to multilingual documents. Correspondence Latent Dirichlet Allocation (Co
rrLDA) is one such model; however, it requires a pivot language to be specified
in advance. We propose a new topic model, Symmetric Correspondence LDA (SymCorrL
DA), that incorporates a hidden variable to control a pivot language, in an exte
nsion of CorrLDA. We experimented with two multilingual comparable datasets extr
acted from Wikipedia and demonstrate that SymCorrLDA is more effective than some
 other existing multilingual topic models.
****

Learning Halfspaces with the Zero-One Loss: Time-Accuracy Tradeoffs
Aharon Birnbaum, Shai 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.
****

Learning High-Density Regions for a Generalized Kolmogorov-Smirnov Test in High-
Dimensional Data
Assaf Glazer, Michael Lindenbaum, Shaul Markovitch
We propose an efficient, generalized, nonparametric, statistical Kolmogorov-Smir
nov test for detecting distributional change in high-dimensional data. To implem
ent the test, we introduce a novel, hierarchical, minimum-volume sets estimator
to represent the distributions to be tested. Our work is motivated by the need t
o detect changes in data streams, and the test is especially efficient in this c
ontext. We provide the theoretical foundations of our test and show its superior
ity over existing methods.
****

Factorial LDA: Sparse Multi-Dimensional Text Models
Michael Paul, Mark Dredze
Multi-dimensional latent variable models can capture the many latent factors in
a text corpus, such as topic, author perspective and sentiment. We introduce fac
torial LDA, a multi-dimensional latent variable model in which a document is inf
luenced by K different factors, and each word token depends on a K-dimensional v
ector of latent variables. Our model incorporates structured word priors and lea
rns a sparse product of factors. Experiments on research abstracts show that our
 model can learn latent factors such as research topic, scientific discipline, a
nd focus (e.g. methods vs. applications.) Our modeling improvements reduce test
perplexity and improve human interpretability of the discovered factors.
****

Augment-and-Conquer Negative Binomial Processes
Mingyuan Zhou, Lawrence Carin
By developing data augmentation methods unique to the negative binomial (NB) dis
tribution, we unite seemingly disjoint count and mixture models under the NB p
rocess framework. We develop fundamental properties of the models and derive eff
icient Gibbs sampling inference. We show that the gamma-NB process can be reduc
ed to the hierarchical Dirichlet process with normalization, highlighting its un

ique theoretical, structural and computational advantages. A variety of NB proce sses with distinct sharing mechanisms are constructed and applied to topic model ing, with connections to existing algorithms, showing the importance of inferrin g both the NB dispersion and probability parameters.
****

Large Scale Distributed Deep Networks
Jeffrey Dean, Greg Corrado, Rajat Monga, Kai Chen, Matthieu Devin, Mark Mao, Mar c'aurelio Ranzato, Andrew Senior, Paul Tucker, Ke Yang, Quoc Le, Andrew Ng
Recent work in unsupervised feature learning and deep learning has shown that b eing able to train large models can dramatically improve performance. In this paper, we consider the problem of training a deep network with billions of para meters using tens of thousands of CPU cores. We have developed a software fram ework called DistBelief that can utilize computing clusters with thousands of m achines to train large models. Within this framework, we have developed two al gorithms for large-scale distributed training: (i) Downpour SGD, an asynchronou s stochastic gradient descent procedure supporting a large number of model repl icas, and (ii) Sandblaster, a framework that supports for a variety of distribu ted batch optimization procedures, including a distributed implementation of L- BFGS. Downpour SGD and Sandblaster L-BFGS both increase the scale and speed of deep network training. We have successfully used our system to train a deep n etwork 100x larger than previously reported in the literature, and achieves sta te-of-the-art performance on ImageNet, a visual object recognition task with 16 million images and 21k categories. We show that these same techniques dramati cally accelerate the training of a more modestly sized deep network for a comme rcial speech recognition service. Although we focus on and report performance o f these methods as applied to training large neural networks, the underlying al gorithms are applicable to any gradient-based machine learning algorithm.
****

Structure estimation for discrete graphical models: Generalized covariance matri ces and their inverses
Po-ling Loh, Martin J. Wainwright
We investigate a curious relationship between the structure of a discrete graphi cal model and the support of the inverse of a generalized covariance matrix. We show that for certain graph structures, the support of the inverse covariance ma trix of indicator variables on the vertices of a graph re■ects the conditional i ndependence structure of the graph. Our work extends results that have previousl y been es- tablished only in the context of multivariate Gaussian graphical mode ls, thereby addressing an open question about the signi■cance of the inverse cov ariance ma- trix of a non-Gaussian distribution. Based on our population-level r esults, we show how the graphical Lasso may be used to recover the edge structur e of cer- tain classes of discrete graphical models, and present simulations to verify our theoretical results.
****

Joint Modeling of a Matrix with Associated Text via Latent Binary Features
Xianxing Zhang, Lawrence Carin
A new methodology is developed for joint analysis of a matrix and accompanying d ocuments, with the documents associated with the matrix rows/columns. The docume nts are modeled with a focused topic model, inferring latent binary features (to pics) for each document. A new matrix decomposition is developed, with latent bi nary features associated with the rows/columns, and with imposition of a low-ran k constraint. The matrix decomposition and topic model are coupled by sharing th e latent binary feature vectors associated with each. The model is applied to ro ll-call data, with the associated documents defined by the legislation. State-of -the-art results are manifested for prediction of votes on a new piece of legisl ation, based only on the observed text legislation. The coupling of the text and legislation is also demonstrated to yield insight into the properties of the ma trix decomposition for roll-call data.
****

Near-Optimal MAP Inference for Determinantal Point Processes
Jennifer Gillenwater, Alex Kulesza, Ben Taskar

Determinantal point processes (DPPs) have recently been proposed as computationally efficient probabilistic models of diverse sets for a variety of applications, including document summarization, image search, and pose estimation. Many DPP inference operations, including normalization and sampling, are tractable; however, finding the most likely configuration (MAP), which is often required in practice for decoding, is NP-hard, so we must resort to approximate inference. Because DPP probabilities are log-submodular, greedy algorithms have been used in the past with some empirical success; however, these methods only give approximation guarantees in the special case of DPPs with monotone kernels. In this paper we propose a new algorithm for approximating the MAP problem based on continuous techniques for submodular function maximization. Our method involves a novel continuous relaxation of the log-probability function, which, in contrast to the multilinear extension used for general submodular functions, can be evaluated and differentiated exactly and efficiently. We obtain a practical algorithm with a 1/4-approximation guarantee for a general class of non-monotone DPPs. Our algorithm also extends to MAP inference under complex polytope constraints, making it possible to combine DPPs with Markov random fields, weighted matchings, and other models. We demonstrate that our approach outperforms greedy methods on both synthetic and real-world data.

****

Image Denoising and Inpainting with Deep Neural Networks

Junyuan Xie, Linli Xu, Enhong Chen

We present a novel approach to low-level vision problems that combines sparse coding and deep networks pre-trained with denoising auto-encoder (DA). We propose an alternative training scheme that successfully adapts DA, originally designed for unsupervised feature learning, to the tasks of image denoising and blind inpainting. Our method achieves state-of-the-art performance in the image denoising task. More importantly, in blind image inpainting task, the proposed method provides solutions to some complex problems that have not been tackled before. Specifically, we can automatically remove complex patterns like superimposed text from an image, rather than simple patterns like pixels missing at random. Moreover, the proposed method does not need the information regarding the region that requires inpainting to be given a priori. Experimental results demonstrate the effectiveness of the proposed method in the tasks of image denoising and blind inpainting. We also show that our new training scheme for DA is more effective and can improve the performance of unsupervised feature learning.

****

Strategic Impatience in Go/NoGo versus Forced-Choice Decision-Making

Pradeep Shenoy, Angela J. Yu

Two-alternative forced choice (2AFC) and Go/NoGo (GNG) tasks are behavioral choice paradigms commonly used to study sensory and cognitive processing in choice behavior. While GNG is thought to isolate the sensory/decisional component by removing the need for response selection, a consistent bias towards the Go response (higher hits and false alarm rates) in the GNG task suggests possible fundamental differences in the sensory or cognitive processes engaged in the two tasks. Existing mechanistic models of these choice tasks, mostly variants of the drift-diffusion model (DDM; [1,2]) and the related leaky competing accumulator models [3,4] capture various aspects of behavior but do not address the provenance of the Go bias. We postulate that this ``impatience'' to go is a strategic adjustment in response to the implicit asymmetry in the cost structure of GNG: the NoGo response requires waiting until the response deadline, while a Go response immediately terminates the current trial. We show that a Bayes-risk minimizing decision policy that minimizes both error rate and average decision delay naturally exhibits the experimentally observed bias. The optimal decision policy is formally equivalent to a DDM with a time-varying threshold that initially rises after stimulus onset, and collapses again near the response deadline. The initial rise is due to the fading temporal advantage of choosing the Go response over the fixed-delay NoGo response. We show that fitting a simpler, fixed-threshold DDM to the optimal model reproduces the counterintuitive result of a higher threshold in GNG than 2AFC decision-making, previously observed in direct DDM fit to behavior

al data [2], although such approximations cannot reproduce the Go bias. Thus, ob served discrepancies between GNG and 2AFC decision-making may arise from rationa l strategic adjustments to the cost structure, and need not imply additional dif ferences in the underlying sensory and cognitive processes.

****

Cost-Sensitive Exploration in Bayesian Reinforcement Learning

Dongho Kim, Kee-eung Kim, Pascal Poupart

In this paper, we consider Bayesian reinforcement learning (BRL) where actions i ncur costs in addition to rewards, and thus exploration has to be constrained in terms of the expected total cost while learning to maximize the expected long-t erm total reward. In order to formalize cost-sensitive exploration, we use the c onstrained Markov decision process (CMDP) as the model of the environment, in wh ich we can naturally encode exploration requirements using the cost function. We extend BEETLE, a model-based BRL method, for learning in the environment with c ost constraints. We demonstrate the cost-sensitive exploration behaviour in a nu mber of simulated problems.

****

MCMC for continuous-time discrete-state systems

Vinayak Rao, Yee Teh

We propose a simple and novel framework for MCMC inference in continuous-time di screte-state systems with pure jump trajectories. We construct an exact MCMC sam pler for such systems by alternately sampling a random discretization of time gi ven a trajectory of the system, and then a new trajectory given the discretizati on. The first step can be performed efficiently using properties of the Poisson process, while the second step can avail of discrete-time MCMC techniques based on the forward-backward algorithm. We compare our approach to particle MCMC and a uniformization-based sampler, and show its advantages.

****

Spectral Learning of General Weighted Automata via Constrained Matrix Completion

Borja Balle, Mehryar Mohri

Many tasks in text and speech processing and computational biology require es- t imating functions mapping strings to real numbers. A broad class of such func- t ions can be de■ned by weighted automata. Spectral methods based on the sin- gula r value decomposition of a Hankel matrix have been recently proposed for learnin g a probability distribution represented by a weighted automaton from a training sample drawn according to this same target distribution. In this paper, we show how spectral methods can be extended to the problem of learning a general weigh ted automaton from a sample generated by an arbitrary distribution. The main obs truction to this approach is that, in general, some entries of the Hankel matrix may be missing. We present a solution to this problem based on solving a constr ained matrix completion problem. Combining these two ingredients, matrix complet ion and spectral method, a whole new family of algorithms for learning general w eighted automata is obtained. We present generalization bounds for a particular algorithm in this family. The proofs rely on a joint stability analysis of matri x completion and spectral learning.

****

Learning with Recursive Perceptual Representations

Oriol Vinyals, Yangqing Jia, Li Deng, Trevor Darrell

Linear Support Vector Machines (SVMs) have become very popular in vision as part of state-of-the-art object recognition and other classification tasks but requi re high dimensional feature spaces for good performance. Deep learning methods c an find more compact representations but current methods employ multilayer perce ptrons that require solving a difficult, non-convex optimization problem. We pro pose a deep non-linear classifier whose layers are SVMs and which incorporates r andom projection as its core stacking element. Our method learns layers of linea r SVMs recursively transforming the original data manifold through a random proj ection of the weak prediction computed from each layer. Our method scales as lin ear SVMs, does not rely on any kernel computations or nonconvex optimization, an d exhibits better generalization ability than kernel-based SVMs. This is especia lly true when the number of training samples is smaller than the dimensionality

of data, a common scenario in many real-world applications. The use of random pr
ojections is key to our method, as we show in the experiments section, in which
we observe a consistent improvement over previous --often more complicated-- met
hods on several vision and speech benchmarks.
****

Scaled Gradients on Grassmann Manifolds for Matrix Completion
Thanh Ngo, Yousef Saad
This paper describes gradient methods based on a scaled metric on the Grassmann
manifold for low-rank matrix completion. The proposed methods significantly impr
ove canonical gradient methods especially on ill-conditioned matrices, while mai
ntaining established global convegence and exact recovery guarantees. A connecti
on between a form of subspace iteration for matrix completion and the scaled gra
dient descent procedure is also established. The proposed conjugate gradient met
hod based on the scaled gradient outperforms several existing algorithms for mat
rix completion and is competitive with recently proposed methods.
****

Link Prediction in Graphs with Autoregressive Features
Emile Richard, Stephane Gaiffas, Nicolas Vayatis
In the paper, we consider the problem of link prediction in time-evolving graphs
. We assume that certain graph features, such as the node degree, follow a vecto
r autoregressive (VAR) model and we propose to use this information to improve t
he accuracy of prediction. Our strategy involves a joint optimization procedure
over the space of adjacency matrices and VAR matrices which takes into account b
oth sparsity and low rank properties of the matrices. Oracle inequalities are de
rived and illustrate the trade-offs in the choice of smoothing parameters when m
odeling the joint effect of sparsity and low rank property. The estimate is comp
uted efficiently using proximal methods through a generalized forward-backward a
gorithm.
****

A Generative Model for Parts-based Object Segmentation
S. Eslami, Christopher Williams
The Shape Boltzmann Machine (SBM) has recently been introduced as a state-of-the
-art model of foreground/background object shape. We extend the SBM to account f
or the foreground object's parts. Our model, the Multinomial SBM (MSBM), can cap
ture both local and global statistics of part shapes accurately. We combine the
MSBM with an appearance model to form a fully generative model of images of obje
cts. Parts-based image segmentations are obtained simply by performing probabili
stic inference in the model. We apply the model to two challenging datasets whic
h exhibit significant shape and appearance variability, and find that it obtains
 results that are comparable to the state-of-the-art.
****

Dimensionality Dependent PAC-Bayes Margin Bound
Chi Jin, Liwei Wang
Margin is one of the most important concepts in machine learning. Previous margi
n bounds, both for SVM and for boosting, are dimensionality independent. A major
 advantage of this dimensionality independency is that it can explain the excell
ent performance of SVM whose feature spaces are often of high or infinite dimens
ion. In this paper we address the problem whether such dimensionality independen
cy is intrinsic for the margin bounds. We prove a dimensionality dependent PAC-B
ayes margin bound. The bound is monotone increasing with respect to the dimensio
n when keeping all other factors fixed. We show that our bound is strictly sharp
er than a previously well-known PAC-Bayes margin bound if the feature space is o
f finite dimension; and the two bounds tend to be equivalent as the dimension go
es to infinity. In addition, we show that the VC bound for linear classifiers ca
n be recovered from our bound under mild conditions. We conduct extensive experi
ments on benchmark datasets and find that the new bound is useful for model sele
ction and is significantly sharper than the dimensionality independent PAC-Bayes
 margin bound as well as the VC bound for linear classifiers.
****

MAP Inference in Chains using Column Generation

David Belanger, Alexandre Passos, Sebastian Riedel, Andrew McCallum
Linear chains and trees are basic building blocks in many applications of graphi cal models. Although exact inference in these models can be performed by dynami c programming, this computation can still be prohibitively expensive with non-tr ivial target variable domain sizes due to the quadratic dependence on this size. Standard message-passing algorithms for these problems are inefficient because they compute scores on hypotheses for which there is strong negative local evid ence. For this reason there has been significant previous interest in beam sear ch and its variants; however, these methods provide only approximate inference. This paper presents new efficient exact inference algorithms based on the combi nation of it column generation and pre-computed bounds on the model's cost struc ture. Improving worst-case performance is impossible. However, our method subst antially speeds real-world, typical-case inference in chains and trees. Experim ents show our method to be twice as fast as exact Viterbi for Wall Street Journa l part-of-speech tagging and over thirteen times faster for a joint part-of-spee d and named-entity-recognition task. Our algorithm is also extendable to new te chniques for approximate inference, to faster two-best inference, and new opport unities for connections between inference and learning.
****

Cocktail Party Processing via Structured Prediction
Yuxuan Wang, Deliang Wang
While human listeners excel at selectively attending to a conversation in a cock tail party, machine performance is still far inferior by comparison. We show tha t the cocktail party problem, or the speech separation problem, can be effective ly approached via structured prediction. To account for temporal dynamics in spe ech, we employ conditional random fields (CRFs) to classify speech dominance wit hin each time-frequency unit for a sound mixture. To capture complex, nonlinear relationship between input and output, both state and transition feature functio ns in CRFs are learned by deep neural networks. The formulation of the problem a s classification allows us to directly optimize a measure that is well correlate d with human speech intelligibility. The proposed system substantially outperfor ms existing ones in a variety of noises.
****

Fused sparsity and robust estimation for linear models with unknown variance
Arnak Dalalyan, Yin Chen
In this paper, we develop a novel approach to the problem of learning sparse rep resentations in the context of fused sparsity and unknown noise level. We propos e an algorithm, termed Scaled Fused Dantzig Selector (SFDS), that accomplishes t he aforementioned learning task by means of a second-order cone program. A speci al emphasize is put on the particular instance of fused sparsity corresponding to the learning in presence of outliers. We establish finite sample risk bounds and carry out an experimental evaluation on both synthetic and real data.
****

On the Sample Complexity of Robust PCA
Matthew Coudron, Gilad Lerman
Requests for name changes in the electronic proceedings will be accepted with 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 to Discover Social Circles in Ego Networks
Jure Leskovec, Julian Mcauley
Our personal social networks are big and cluttered, and currently there is no go od way to organize them. Social networking sites allow users to manually categor ize their friends into social circles (e.g. circles' on Google+, andlists' on Fa cebook and Twitter), however they are laborious to construct and must be updated whenever a user's network grows. We define a novel machine learning task of ide ntifying users' social circles. We pose the problem as a node clustering problem on a user's ego-network, a network of connections between her friends. We devel op a model for detecting circles that combines network structure as well as user

profile information. For each circle we learn its members and the circle-specif
ic user profile similarity metric. Modeling node membership to multiple circles
allows us to detect overlapping as well as hierarchically nested circles. Experi
ments show that our model accurately identifies circles on a diverse set of data
from Facebook, Google+, and Twitter for all of which we obtain hand-labeled gro
und-truth data.
****

Exact and Stable Recovery of Sequences of Signals with Sparse Increments via Dif
ferential _1-Minimization
Demba Ba, Behtash Babadi, Patrick Purdon, Emery Brown
Requests for name changes in the electronic proceedings will be accepted with 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.
****

Tight Bounds on Profile Redundancy and Distinguishability
Jayadev Acharya, Hirakendu Das, Alon Orlitsky
The minimax KL-divergence of any distribution from all distributions in a collec
tion P has several practical implications. In compression, it is called redundan
cy and represents the least additional number of bits over the entropy needed to
encode the output of any distribution in P. In online es- timation and learning
, it is the lowest expected log-loss regret when guessing a sequence of random v
alues generated by a distribution in P. In hypothesis testing, it upper bounds t
he largest number of distinguishable distributions in P. Motivated by problems r
anging from population estimation to text classi■cation and speech recognition,
several machine-learning and information-theory researchers have recently consid
ered label-invariant observations and properties induced by i.i.d. distributions
. A suf■cient statistic for all these properties is the data’s pro■le, the multi
set of the number of times each data element appears. Improving on a sequence of
previous works, we show that the redun- dancy of the collection of distribution
s induced over pro■les by length-n i.i.d. sequences is between 0.3 · n1/3 and n1
/3 log2 n, in particular, establishing its exact growth power.
****

On Triangular versus Edge Representations --- Towards Scalable Modeling of Netwo
rks
Qirong Ho, Junming Yin, Eric Xing
Requests for name changes in the electronic proceedings will be accepted with no
questions asked. However name changes may cause bibliographic tracking issues.
Authors are asked to consider this carefully and discuss it with their co-auth
ors prior to requesting a name change in the electronic proceedings.
****

A Better Way to Pretrain Deep Boltzmann Machines
Geoffrey E. Hinton, Russ R. Salakhutdinov
We describe how the pre-training algorithm for Deep Boltzmann Machines (DBMs) is
related to the pre-training algorithm for Deep Belief Networks and we show that
under certain conditions, the pre-training procedure improves the variational l
ower bound of a two-hidden-layer DBM. Based on this analysis, we develop a diffe
rent method of pre-training DBMs that distributes the modelling work more evenly
over the hidden layers. Our results on the MNIST and NORB datasets demonstrate
that the new pre-training algorithm allows us to learn better generative models.
****

Semi-supervised Eigenvectors for Locally-biased Learning
Toke Hansen, Michael W. Mahoney
In many applications, one has information, e.g., labels that are provided in a
semi-supervised manner, about a specific target region of a large data set, and
one wants to perform machine learning and data analysis tasks nearby that pre-
specified target region. Locally-biased problems of this sort are particularly
challenging for popular eigenvector-based machine learning and data analysis t
ools. At root, the reason is that eigenvectors are inherently global quantities.
In this paper, we address this issue by providing a methodology to construct s

emi-supervised eigenvectors of a graph Laplacian, and we illustrate how these l
ocally-biased eigenvectors can be used to perform locally-biased machine learni
ng. These semi-supervised eigenvectors capture successively-orthogonalized dire
ctions of maximum variance, conditioned on being well-correlated with an input
seed set of nodes that is assumed to be provided in a semi-supervised manner. W
e also provide several empirical examples demonstrating how these semi-supervis
ed eigenvectors can be used to perform locally-biased learning.
****

The variational hierarchical EM algorithm for clustering hidden Markov models
Emanuele Coviello, Gert Lanckriet, Antoni Chan
In this paper, we derive a novel algorithm to cluster hidden Markov models (HMM
s) according to their probability distributions. We propose a variational hierar
chical EM algorithm that i) clusters a given collection of HMMs into groups of H
MMs that are similar, in terms of the distributions they represent, and ii) char
acterizes each group by a ``cluster center'', i.e., a novel HMM that is represen
tative for the group. We illustrate the benefits of the proposed algorithm on hi
erarchical clustering of motion capture sequences as well as on automatic music
tagging.
****

Scalable nonconvex inexact proximal splitting
Suvrit Sra
We study large-scale, nonsmooth, nonconconvex optimization problems. In particul
ar, we focus on nonconvex problems with \emph{composite} objectives. This class
of problems includes the extensively studied convex, composite objective problem
s as a special case. To tackle composite nonconvex problems, we introduce a powe
rful new framework based on asymptotically \emph{nonvanishing} errors, avoiding
the common convenient assumption of eventually vanishing errors. Within our fram
ework we derive both batch and incremental nonconvex proximal splitting algorith
ms. To our knowledge, our framework is first to develop and analyze incremental
\emph{nonconvex} proximal-splitting algorithms, even if we disregard the ability
to handle nonvanishing errors. We illustrate our theoretical framework by showi
ng how it applies to difficult large-scale, nonsmooth, and nonconvex problems.
****

Bayesian nonparametric models for ranked data
Francois Caron, Yee Teh
We develop a Bayesian nonparametric extension of the popular Plackett-Luce choic
e model that can handle an infinite number of choice items. Our framework is b
ased on the theory of random atomic measures, with the prior specified by a gamm
a process. We derive a posterior characterization and a simple and effective Gi
bbs sampler for posterior simulation. We then develop a time-varying extension
of our model, and apply our model to the New York Times lists of weekly bestsell
ing books.
****

GenDeR: A Generic Diversified Ranking Algorithm
Jingrui He, Hanghang Tong, Qiaozhu Mei, Boleslaw Szymanski
Diversified ranking is a fundamental task in machine learning. It is broadly app
licable in many real world problems, e.g., information retrieval, team assemblin
g, product search, etc. In this paper, we consider a generic setting where we ai
m to diversify the top-k ranking list based on an arbitrary relevance function a
nd an arbitrary similarity function among all the examples. We formulate it as a
n optimization problem and show that in general it is NP-hard. Then, we show tha
t for a large volume of the parameter space, the proposed objective function enj
oys the diminishing returns property, which enables us to design a scalable, gre
edy algorithm to find the near-optimal solution. Experimental results on real da
ta sets demonstrate the effectiveness of the proposed algorithm.
****

Accuracy at the Top
Stephen Boyd, Corinna Cortes, Mehryar Mohri, Ana Radovanovic
Requests for name changes in the electronic proceedings will be accepted with 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.
****

Approximating Equilibria in Sequential Auctions with Incomplete Information and
Multi-Unit Demand

Amy Greenwald, Jiacui Li, Eric Sodomka

In many large economic markets, goods are sold through sequential auctions. Such
domains include eBay, online ad auctions, wireless spectrum auctions, and the D
utch flower auctions. Bidders in these domains face highly complex decision-maki
ng problems, as their preferences for outcomes in one auction often depend on th
e outcomes of other auctions, and bidders have limited information about factors
that drive outcomes, such as other bidders' preferences and past actions. In
this work, we formulate the bidder's problem as one of price prediction (i.e., l
earning) and optimization. We define the concept of stable price predictions and
show that (approximate) equilibrium in sequential auctions can be characterized
as a profile of strategies that (approximately) optimize with respect to such (
approximately) stable price predictions. We show how equilibria found with our f
ormulation compare to known theoretical equilibria for simpler auction domains,
and we find new approximate equilibria for a more complex auction domain where a
nalytical solutions were heretofore unknown.
****

Multiresolution Gaussian Processes

Emily Fox, David Dunson

We propose a multiresolution Gaussian process to capture long-range, non-Markovi
an dependencies while allowing for abrupt changes. The multiresolution GP hiera
rchically couples a collection of smooth GPs, each defined over an element of a
random nested partition. Long-range dependencies are captured by the top-level
GP while the partition points define the abrupt changes. Due to the inherent co
njugacy of the GPs, one can analytically marginalize the GPs and compute the con
ditional likelihood of the observations given the partition tree. This allows f
or efficient inference of the partition itself, for which we employ graph-theore
tic techniques. We apply the multiresolution GP to the analysis of Magnetoencep
halography (MEG) recordings of brain activity.
****

Burn-in, bias, and the rationality of anchoring

Falk Lieder, Tom Griffiths, Noah Goodman

Bayesian inference provides a unifying framework for addressing problems in mach
ine learning, artificial intelligence, and robotics, as well as the problems fac
ing the human mind. Unfortunately, exact Bayesian inference is intractable in al
l but the simplest models. Therefore minds and machines have to approximate Baye
sian inference. Approximate inference algorithms can achieve a wide range of tim
e-accuracy tradeoffs, but what is the optimal tradeoff? We investigate time-accu
racy tradeoffs using the Metropolis-Hastings algorithm as a metaphor for the min
d's inference algorithm(s). We find that reasonably accurate decisions are possi
ble long before the Markov chain has converged to the posterior distribution, i.
e. during the period known as burn-in. Therefore the strategy that is optimal su
bject to the mind's bounded processing speed and opportunity costs may perform s
o few iterations that the resulting samples are biased towards the initial value
. The resulting cognitive process model provides a rational basis for the anchor
ing-and-adjustment heuristic. The model's quantitative predictions are tested ag
ainst published data on anchoring in numerical estimation tasks. Our theoretical
and empirical results suggest that the anchoring bias is consistent with approx
imate Bayesian inference.
****

Truly Nonparametric Online Variational Inference for Hierarchical Dirichlet Proc
esses

Michael Bryant, Erik Sudderth

Variational methods provide a computationally scalable alternative to Monte Carl
o methods for large-scale, Bayesian nonparametric learning. In practice, howeve
r, conventional batch and online variational methods quickly become trapped in l

ocal optima. In this paper, we consider a nonparametric topic model based on the hierarchical Dirichlet process (HDP), and develop a novel online variational inference algorithm based on split-merge topic updates. We derive a simpler and faster variational approximation of the HDP, and show that by intelligently splitting and merging components of the variational posterior, we can achieve substantially better predictions of test data than conventional online and batch variational algorithms. For streaming analysis of large datasets where batch analysis is infeasible, we show that our split-merge updates better capture the nonparametric properties of the underlying model, allowing continual learning of new topics.

****

3D Object Detection and Viewpoint Estimation with a Deformable 3D Cuboid Model

Sanja Fidler, Sven Dickinson, Raquel Urtasun

This paper addresses the problem of category-level 3D object detection. Given a monocular image, our aim is to localize the objects in 3D by enclosing them with tight oriented 3D bounding boxes. We propose a novel approach that extends the well-acclaimed deformable part-based model[Felz.] to reason in 3D. Our model represents an object class as a deformable 3D cuboid composed of faces and parts, which are both allowed to deform with respect to their anchors on the 3D box. We model the appearance of each face in fronto-parallel coordinates, thus effectively factoring out the appearance variation induced by viewpoint. Our model reasons about face visibility patters called aspects. We train the cuboid model jointly and discriminatively and share weights across all aspects to attain efficiency. Inference then entails sliding and rotating the box in 3D and scoring object hypotheses. While for inference we discretize the search space, the variables are continuous in our model. We demonstrate the effectiveness of our approach in indoor and outdoor scenarios, and show that our approach outperforms the state-of-the-art in both 2D[Felz09] and 3D object detection[Hedau12].

****

Risk-Aversion in Multi-armed Bandits

Amir Sani, Alessandro Lazaric, Rémi Munos

In stochastic multi--armed bandits the objective is to solve the exploration--exploitation dilemma and ultimately maximize the expected reward. Nonetheless, in many practical problems, maximizing the expected reward is not the most desirable objective. In this paper, we introduce a novel setting based on the principle of risk--aversion where the objective is to compete against the arm with the best risk--return trade--off. This setting proves to be intrinsically more difficult than the standard multi-arm bandit setting due in part to an exploration risk which introduces a regret associated to the variability of an algorithm. Using variance as a measure of risk, we introduce two new algorithms, we investigate their theoretical guarantees, and we report preliminary empirical results.

****

Approximate Message Passing with Consistent Parameter Estimation and Applications to Sparse Learning

Ulugbek Kamilov, Sundeep Rangan, Michael Unser, Alyson K. Fletcher

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

****

Gradient-based kernel method for feature extraction and variable selection

Kenji Fukumizu, Chenlei Leng

We propose a novel kernel approach to dimension reduction for supervised learning: feature extraction and variable selection; the former constructs a small number of features from predictors, and the latter finds a subset of predictors. First, a method of linear feature extraction is proposed using the gradient of regression function, based on the recent development of the kernel method. In comparison with other existing methods, the proposed one has wide applicability without strong assumptions on the regressor or type of variables, and uses computationally simple eigendecomposition, thus applicable to large data sets. Second, in

combination of a sparse penalty, the method is extended to variable selection, following the approach by Chen et al. (2010). Experimental results show that the proposed methods successfully find effective features and variables without parametric models.

****

Scalable imputation of genetic data with a discrete fragmentation-coagulation process

Lloyd Elliott, Yee Teh

We present a Bayesian nonparametric model for genetic sequence data in which a set of genetic sequences is modelled using a Markov model of partitions. The partitions at consecutive locations in the genome are related by their clusters first splitting and then merging. Our model can be thought of as a discrete time analogue of continuous time fragmentation-coagulation processes [Teh et al 2011], preserving the important properties of projectivity, exchangeability and reversibility, while being more scalable. We apply this model to the problem of genotype imputation, showing improved computational efficiency while maintaining the same accuracies as in [Teh et al 2011].

****

Non-parametric Approximate Dynamic Programming via the Kernel Method

Nikhil Bhat, Vivek Farias, Ciamac C. Moallemi

This paper presents a novel non-parametric approximate dynamic programming (ADP) algorithm that enjoys graceful, dimension-independent approximation and sample complexity guarantees. In particular, we establish both theoretically and computationally that our proposal can serve as a viable alternative to state-of-the-art parametric ADP algorithms, freeing the designer from carefully specifying an approximation architecture. We accomplish this by developing a kernel-based mathematical program for ADP. Via a computational study on a controlled queueing network, we show that our non-parametric procedure is competitive with parametric ADP approaches.

****

Probabilistic n-Choose-k Models for Classification and Ranking

Kevin Swersky, Brendan J. Frey, Daniel Tarlow, Richard Zemel, Ryan P. Adams

In categorical data there is often structure in the number of variables that take on each label. For example, the total number of objects in an image and the number of highly relevant documents per query in web search both tend to follow a structured distribution. In this paper, we study a probabilistic model that explicitly includes a prior distribution over such counts, along with a count-conditional likelihood that de■nes probabilities over all subsets of a given size. When labels are binary and the prior over counts is a Poisson-Binomial distribution, a standard logistic regression model is recovered, but for other count distributions, such priors induce global dependencies and combinatorics that appear to complicate learning and inference. However, we demonstrate that simple, ef■cient learning procedures can be derived for more general forms of this model. We illustrate the utility of the formulation by exploring applications to multi-object classi■cation, learning to rank, and top-K classi■cation.

****

Delay Compensation with Dynamical Synapses

Chi Fung, K. Wong, Si Wu

Time delay is pervasive in neural information processing. To achieve real-time tracking, it is critical to compensate the transmission and processing delays in a neural system. In the present study we show that dynamical synapses with short-term depression can enhance the mobility of a continuous attractor network to the extent that the system tracks time-varying stimuli in a timely manner. The state of the network can either track the instantaneous position of a moving stimulus perfectly (with zero-lag) or lead it with an effectively constant time, in agreement with experiments on the head-direction systems in rodents. The parameter regions for delayed, perfect and anticipative tracking correspond to network states that are static, ready-to-move and spontaneously moving, respectively, demonstrating the strong correlation between tracking performance and the intrinsic dynamics of the network. We also find that when the speed of the stimulus coinc

ides with the natural speed of the network state, the delay becomes effectively independent of the stimulus amplitude.
****

Visual Recognition using Embedded Feature Selection for Curvature Self-Similarit y

Angela Eigenstetter, Bjorn Ommer

Category-level object detection has a crucial need for informative object repres entations. This demand has led to feature descriptors of ever increasing dimensi onality like co-occurrence statistics and self-similarity. In this paper we pro pose a new object representation based on curvature self-similarity that goes be yond the currently popular approximation of objects using straight lines. Howev er, like all descriptors using second order statistics, ours also exhibits a hig h dimensionality. Although improving discriminability, the high dimensionality becomes a critical issue due to lack of generalization ability and curse of dime nsionality. Given only a limited amount of training data, even sophisticated le arning algorithms such as the popular kernel methods are not able to suppress no isy or superfluous dimensions of such high-dimensional data. Consequently, there is a natural need for feature selection when using present-day informative fea tures and, particularly, curvature self-similarity. We therefore suggest an embe dded feature selection method for SVMs that reduces complexity and improves gen eralization capability of object models. By successfully integrating the propose d curvature self-similarity representation together with the embedded feature s election in a widely used state-of-the-art object detection framework we show th e general pertinence of the approach.
****

High-Order Multi-Task Feature Learning to Identify Longitudinal Phenotypic Marke rs for Alzheimer's Disease Progression Prediction

Hua Wang, Feiping Nie, Heng Huang, Jingwen Yan, Sungeun Kim, Shannon Risacher, A ndrew Saykin, Li Shen

Alzheimer disease (AD) is a neurodegenerative disorder characterized by progress ive impairment of memory and other cognitive functions. Regression analysis has been studied to relate neuroimaging measures to cognitive status. However, wheth er these measures have further predictive power to infer a trajectory of cogniti ve performance over time is still an under-explored but important topic in AD re search. We propose a novel high-order multi-task learning model to address this issue. The proposed model explores the temporal correlations existing in data fe atures and regression tasks by the structured sparsity-inducing norms. In additi on, the sparsity of the model enables the selection of a small number of MRI mea sures while maintaining high prediction accuracy. The empirical studies, using t he baseline MRI and serial cognitive data of the ADNI cohort, have yielded promi sing results.
****

Multiresolution analysis on the symmetric group

Risi Kondor, Walter Dempsey

There is no generally accepted way to define wavelets on permutations. We addres s this issue by introducing the notion of coset based multiresolution analysis (CMRA) on the symmetric group; find the corresponding wavelet functions; and desc ribe a fast wavelet transform of O(n^p) complexity with small p for sparse signa ls (in contrast to the O(n^q n!) complexity typical of FFTs). We discuss potenti al applications in ranking, sparse approximation, and multi-object tracking.
****

Learned Prioritization for Trading Off Accuracy and Speed

Jiarong Jiang, Adam Teichert, Jason Eisner, Hal Daume

Users want natural language processing (NLP) systems to be both fast and accurat e, but quality often comes at the cost of speed. The field has been manually exp loring various speed-accuracy tradeoffs (for particular problems and datasets). We aim to explore this space automatically, focusing here on the case of agenda -based syntactic parsing \cite{kay-1986}. Unfortunately, off-the-shelf reinforce ment learning techniques fail to learn good policies: the state space is simply too large to explore naively. An attempt to counteract this by applying imitatio

n learning algorithms also fails: the ``teacher'' is far too good to successfully imitate with our inexpensive features. Moreover, it is not specifically tuned for the known reward function. We propose a hybrid reinforcement/apprenticeship learning algorithm that, even with only a few inexpensive features, can automatically learn weights that achieve competitive accuracies at significant improvements in speed over state-of-the-art baselines.
****

Learning as MAP Inference in Discrete Graphical Models
Xianghang Liu, James Petterson, Tibério Caetano
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****

Hierarchical Optimistic Region Selection driven by Curiosity
Odalric-ambrym Maillard
This paper aims to take a step forwards making the term ``intrinsic motivation'' from reinforcement learning theoretically well founded, focusing on curiosity-driven learning. To that end, we consider the setting where, a fixed partition P of a continuous space X being given, and a process \nu defined on X being unknown, we are asked to sequentially decide which cell of the partition to select as well as where to sample \nu in that cell, in order to minimize a loss function that is inspired from previous work on curiosity-driven learning. The loss on each cell consists of one term measuring a simple worst case quadratic sampling error, and a penalty term proportional to the range of the variance in that cell. The corresponding problem formulation extends the setting known as active learning for multi-armed bandits to the case when each arm is a continuous region, and we show how an adaptation of recent algorithms for that problem and of hierarchical optimistic sampling algorithms for optimization can be used in order to solve this problem. The resulting procedure, called Hierarchical Optimistic Region SElection driven by Curiosity (HORSE.C) is provided together with a finite-time regret analysis.
****

Trajectory-Based Short-Sighted Probabilistic Planning
Felipe Trevizan, Manuela Veloso
Requests for name changes in the electronic proceedings will be accepted with 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.
****

Best Arm Identification: A Unified Approach to Fixed Budget and Fixed Confidence
Victor Gabillon, Mohammad Ghavamzadeh, Alessandro Lazaric
We study the problem of identifying the best arm(s) in the stochastic multi-armed bandit setting. This problem has been studied in the literature from two different perspectives: fixed budget and fixed confidence. We propose a unifying approach that leads to a meta-algorithm called unified gap-based exploration (UGapE), with a common structure and similar theoretical analysis for these two settings. We prove a performance bound for the two versions of the algorithm showing that the two problems are characterized by the same notion of complexity. We also show how the UGapE algorithm as well as its theoretical analysis can be extended to take into account the variance of the arms and to multiple bandits. Finally, we evaluate the performance of UGapE and compare it with a number of existing fixed budget and fixed confidence algorithms.
****

On the Use of Non-Stationary Policies for Stationary Infinite-Horizon Markov Decision Processes
Bruno Scherrer, Boris Lesner
Requests for name changes in the electronic proceedings will be accepted with 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 Spatio-Temporal Architectures and Learning for Protein Structure Prediction
Pietro Lena, Ken Nagata, Pierre Baldi
Residue-residue contact prediction is a fundamental problem in protein structure prediction. Hower, despite considerable research efforts, contact prediction methods are still largely unreliable. Here we introduce a novel deep machine-learning architecture which consists of a multidimensional stack of learning modules. For contact prediction, the idea is implemented as a three-dimensional stack of Neural Networks NN^k_{ij}, where i and j index the spatial coordinates of the contact map and k indexes ''time''. The temporal dimension is introduced to capture the fact that protein folding is not an instantaneous process, but rather a progressive refinement. Networks at level k in the stack can be trained in supervised fashion to refine the predictions produced by the previous level, hence addressing the problem of vanishing gradients, typical of deep architectures. Increased accuracy and generalization capabilities of this approach are established by rigorous comparison with other classical machine learning approaches for contact prediction. The deep approach leads to an accuracy for difficult long-range contacts of about 30%, roughly 10% above the state-of-the-art. Many variations in the architectures and the training algorithms are possible, leaving room for further improvements. Furthermore, the approach is applicable to other problems with strong underlying spatial and temporal components.
****

Isotropic Hashing
Weihao Kong, Wu-jun Li
Most existing hashing methods adopt some projection functions to project the original data into several dimensions of real values, and then each of these projected dimensions is quantized into one bit (zero or one) by thresholding. Typically, the variances of different projected dimensions are different for existing projection functions such as principal component analysis (PCA). Using the same number of bits for different projected dimensions is unreasonable because larger-variance dimensions will carry more information. Although this viewpoint has been widely accepted by many researchers, it is still not verified by either theory or experiment because no methods have been proposed to find a projection with equal variances for different dimensions. In this paper, we propose a novel method, called isotropic hashing (IsoHash), to learn projection functions which can produce projected dimensions with isotropic variances (equal variances). Experimental results on real data sets show that IsoHash can outperform its counterpart with different variances for different dimensions, which verifies the viewpoint that projections with isotropic variances will be better than those with anisotropic variances.
****

Repulsive Mixtures
Francesca Petralia, Vinayak Rao, David Dunson
Discrete mixtures are used routinely in broad sweeping applications ranging from unsupervised settings to fully supervised multi-task learning. Indeed, finite mixtures and infinite mixtures, relying on Dirichlet processes and modifications, have become a standard tool. One important issue that arises in using discrete mixtures is low separation in the components; in particular, different components can be introduced that are very similar and hence redundant. Such redundancy leads to too many clusters that are too similar, degrading performance in unsupervised learning and leading to computational problems and an unnecessarily complex model in supervised settings. Redundancy can arise in the absence of a penalty on components placed close together even when a Bayesian approach is used to learn the number of components. To solve this problem, we propose a novel prior that generates components from a repulsive process, automatically penalizing redundant components. We characterize this repulsive prior theoretically and propose a Markov chain Monte Carlo sampling algorithm for posterior computation. The methods are illustrated using synthetic examples and an iris data set.
****

Forward-Backward Activation Algorithm for Hierarchical Hidden Markov Models
Kei Wakabayashi, Takao Miura
Hierarchical Hidden Markov Models (HHMMs) are sophisticated stochastic models th
at enable us to capture a hierarchical context characterization of sequence data
. However, existing HHMM parameter estimation methods require large computations
 of time complexity O(TN^{2D}) at least for model inference, where D is the dept
h of the hierarchy, N is the number of states in each level, and T is the sequen
ce length. In this paper, we propose a new inference method of HHMMs for which t
he time complexity is O(TN^{D+1}). A key idea of our algorithm is application of
 the forward-backward algorithm to ''state activation probabilities''. The notio
n of a state activation, which offers a simple formalization of the hierarchical
 transition behavior of HHMMs, enables us to conduct model inference efficiently
. We present some experiments to demonstrate that our proposed method works more
 efficiently to estimate HHMM parameters than do some existing methods such as t
he flattening method and Gibbs sampling method.
****

Finding Exemplars from Pairwise Dissimilarities via Simultaneous Sparse Recovery
Ehsan Elhamifar, Guillermo Sapiro, René Vidal
Given pairwise dissimilarities between data points, we consider the problem of f
inding a subset of data points called representatives or exemplars that can effi
ciently describe the data collection. We formulate the problem as a row-sparsity
 regularized trace minimization problem which can be solved efficiently using co
nvex programming. The solution of the proposed optimization program finds the re
presentatives and the probability that each data point is associated to each one
 of the representatives. We obtain the range of the regularization parameter for
 which the solution of the proposed optimization program changes from selecting
one representative to selecting all data points as the representatives. When dat
a points are distributed around multiple clusters according to the dissimilariti
es, we show that the data in each cluster select only representatives from that
cluster. Unlike metric-based methods, our algorithm does not require that the pa
irwise dissimilarities be metrics and can be applied to dissimilarities that are
 asymmetric or violate the triangle inequality. We demonstrate the effectiveness
 of the proposed algorithm on synthetic data as well as real-world datasets of i
mages and text.
****

Mirror Descent Meets Fixed Share (and feels no regret)
Nicolò Cesa-bianchi, Pierre Gaillard, Gabor Lugosi, Gilles Stoltz
Mirror descent with an entropic regularizer is known to achieve shifting regret
bounds that are logarithmic in the dimension. This is done using either a carefu
lly designed projection or by a weight sharing technique. Via a novel unified an
alysis, we show that these two approaches deliver essentially equivalent bounds
on a notion of regret generalizing shifting, adaptive, discounted, and other rel
ated regrets. Our analysis also captures and extends the generalized weight shar
ing technique of Bousquet and Warmuth, and can be refined in several ways, inclu
ding improvements for small losses and adaptive tuning of parameters.
****

Semi-Supervised Domain Adaptation with Non-Parametric Copulas
David Lopez-paz, Jose Hernández-lobato, Bernhard Schölkopf
A new framework based on the theory of copulas is proposed to address semi-super
vised domain adaptation problems. The presented method factorizes any multivaria
te density into a product of marginal distributions and bivariate copula functio
ns. Therefore, changes in each of these factors can be detected and corrected to
 adapt a density model across different learning domains. Importantly, we introd
uce a novel vine copula model, which allows for this factorization in a non-para
metric manner. Experimental results on regression problems with real-world data
illustrate the efficacy of the proposed approach when compared to state-of-the-a
rt techniques.
****

The Lovász ϑ function, SVMs and finding large dense subgraphs
Vinay Jethava, Anders Martinsson, Chiranjib Bhattacharyya, Devdatt Dubhashi

Requests for name changes in the electronic proceedings will be accepted with 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.
****

Slice sampling normalized kernel-weighted completely random measure mixture models

Nick Foti, Sinead Williamson

A number of dependent nonparametric processes have been proposed to model non-stationary data with unknown latent dimensionality. However, the inference algorithms are often slow and unwieldy, and are in general highly specific to a given model formulation. In this paper, we describe a wide class of nonparametric processes, including several existing models, and present a slice sampler that allows efficient inference across this class of models.
****

Automatic Feature Induction for Stagewise Collaborative Filtering

Joonseok Lee, Mingxuan Sun, Seungyeon Kim, Guy Lebanon

Recent approaches to collaborative filtering have concentrated on estimating an algebraic or statistical model, and using the model for predicting missing ratings. In this paper we observe that different models have relative advantages in different regions of the input space. This motivates our approach of using stagewise linear combinations of collaborative filtering algorithms, with non-constant combination coefficients based on kernel smoothing. The resulting stagewise model is computationally scalable and outperforms a wide selection of state-of-the-art collaborative filtering algorithms.
****

A Stochastic Gradient Method with an Exponential Convergence _Rate for Finite Training Sets

Nicolas Roux, Mark Schmidt, Francis Bach

We propose a new stochastic gradient method for optimizing the sum of■ a finite set of smooth functions, where the sum is strongly convex.■ While standard stochastic gradient methods■ converge at sublinear rates for this problem, the proposed method incorporates a memory of previous gradient values in order to achieve a linear convergence ■rate. In a machine learning context, numerical experiments indicate that the new algorithm can dramatically outperform standard■ algorithms, both in terms of optimizing the training error and reducing the test error quickly.
****

Monte Carlo Methods for Maximum Margin Supervised Topic Models

Qixia Jiang, Jun Zhu, Maosong Sun, Eric Xing

An effective strategy to exploit the supervising side information for discovering predictive topic representations is to impose discriminative constraints induced by such information on the posterior distributions under a topic model. This strategy has been adopted by a number of supervised topic models, such as MedLDA, which employs max-margin posterior constraints. However, unlike the likelihood-based supervised topic models, of which posterior inference can be carried out using the Bayes' rule, the max-margin posterior constraints have made Monte Carlo methods infeasible or at least not directly applicable, thereby limited the choice of inference algorithms to be based on variational approximation with strict mean field assumptions. In this paper, we develop two efficient Monte Carlo methods under much weaker assumptions for max-margin supervised topic models based on an importance sampler and a collapsed Gibbs sampler, respectively, in a convex dual formulation. We report thorough experimental results that compare our approach favorably against existing alternatives in both accuracy and efficiency.
****

Analyzing 3D Objects in Cluttered Images

Mohsen Hejrati, Deva Ramanan

We present an approach to detecting and analyzing the 3D configuration of objects in real-world images with heavy occlusion and clutter. We focus on the application of finding and analyzing cars. We do so with a two-stage model; the first s

tage reasons about 2D shape and appearance variation due to within-class variati on(station wagons look different than sedans) and changes in viewpoint. Rather t han using a view-based model, we describe a compositional representation that mo dels a large number of effective views and shapes using a small number of local view-based templates. We use this model to propose candidate detections and 2D e stimates of shape. These estimates are then refined by our second stage, using a n explicit 3D model of shape and viewpoint. We use a morphable model to capture 3D within-class variation, and use a weak-perspective camera model to capture vi ewpoint. We learn all model parameters from 2D annotations. We demonstrate state -of-the-art accuracy for detection, viewpoint estimation, and 3D shape reconstru ction on challenging images from the PASCAL VOC 2011 dataset.
****

A mechanistic model of early sensory processing based on subtracting sparse repr esentations
Shaul Druckmann, Tao Hu, Dmitri Chklovskii
Early stages of sensory systems face the challenge of compressing information fr om numerous receptors onto a much smaller number of projection neurons, a so cal led communication bottleneck. To make more efficient use of limited bandwidth, c ompression may be achieved using predictive coding, whereby predictable, or redu ndant, components of the stimulus are removed. In the case of the retina, Sriniv asan et al. (1982) suggested that feedforward inhibitory connections subtracting a linear prediction generated from nearby receptors implement such compression, resulting in biphasic center-surround receptive fields. However, feedback inhib itory circuits are common in early sensory circuits and furthermore their dynami cs may be nonlinear. Can such circuits implement predictive coding as well? Here , solving the transient dynamics of nonlinear reciprocal feedback circuits throu gh analogy to a signal-processing algorithm called linearized Bregman iteration we show that nonlinear predictive coding can be implemented in an inhibitory fee dback circuit. In response to a step stimulus, interneuron activity in time cons tructs progressively less sparse but more accurate representations of the stimul us, a temporally evolving prediction. This analysis provides a powerful theoreti cal framework to interpret and understand the dynamics of early sensory processi ng in a variety of physiological experiments and yields novel predictions regard ing the relation between activity and stimulus statistics.
****

Ensemble weighted kernel estimators for multivariate entropy estimation
Kumar Sricharan, Alfred Hero
Requests for name changes in the electronic proceedings will be accepted with 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.
****

Active Comparison of Prediction Models
Christoph Sawade, Niels Landwehr, Tobias Scheffer
We address the problem of comparing the risks of two given predictive models - f or instance, a baseline model and a challenger - as confidently as possible on a fixed labeling budget. This problem occurs whenever models cannot be compared o n held-out training data, possibly because the training data are unavailable or do not reflect the desired test distribution. In this case, new test instances h ave to be drawn and labeled at a cost. We devise an active comparison method tha t selects instances according to an instrumental sampling distribution. We deriv e the sampling distribution that maximizes the power of a statistical test appli ed to the observed empirical risks, and thereby minimizes the likelihood of choo sing the inferior model. Empirically, we investigate model selection problems on several classification and regression tasks and study the accuracy of the resul ting p-values.
****

Reducing statistical time-series problems to binary classification
Daniil Ryabko, Jeremie Mary
We show how binary classification methods developed to work on i.i.d. data can

be used for solving statistical problems that are seemingly unrelated to classification and concern highly-dependent time series. Specifically, the problems of time-series clustering, homogeneity testing and the three-sample problem are addressed. The algorithms that we construct for solving these problems are based on a new metric between time-series distributions, which can be evaluated using binary classification methods. Universal consistency of the proposed algorithms is proven under most general assumptions. The theoretical results are illustrated with experiments on synthetic and real-world data.
****

Max-Margin Structured Output Regression for Spatio-Temporal Action Localization
Du Tran, Junsong Yuan
Structured output learning has been successfully applied to object localization, where the mapping between an image and an object bounding box can be well captured. Its extension to action localization in videos, however, is much more challenging, because one needs to predict the locations of the action patterns both spatially and temporally, i.e., identifying a sequence of bounding boxes that track the action in video. The problem becomes intractable due to the exponentially large size of the structured video space where actions could occur. We propose a novel structured learning approach for spatio-temporal action localization. The mapping between a video and a spatio-temporal action trajectory is learned. The intractable inference and learning problems are addressed by leveraging an efficient Max-Path search method, thus makes it feasible to optimize the model over the whole structured space. Experiments on two challenging benchmark datasets show that our proposed method outperforms the state-of-the-art methods.
****

Minimizing Sparse High-Order Energies by Submodular Vertex-Cover
Andrew Delong, Olga Veksler, Anton Osokin, Yuri Boykov
Inference on high-order graphical models has become increasingly important in recent years. We consider energies with simple 'sparse' high-order potentials. Previous work in this area uses either specialized message-passing or transforms each high-order potential to the pairwise case. We take a fundamentally different approach, transforming the entire original problem into a comparatively small instance of a submodular vertex-cover problem. These vertex-cover instances can then be attacked by standard pairwise methods, where they run much faster (4--15 times) and are often more effective than on the original problem. We evaluate our approach on synthetic data, and we show that our algorithm can be useful in a fast hierarchical clustering and model estimation framework.
****

A new metric on the manifold of kernel matrices with application to matrix geometric means
Suvrit Sra
Requests for name changes in the electronic proceedings will be accepted with 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.
****

Wavelet based multi-scale shape features on arbitrary surfaces for cortical thickness discrimination
Won Kim, Deepti Pachauri, Charles Hatt, Moo. Chung, Sterling Johnson, Vikas Singh
Hypothesis testing on signals de■ned on surfaces (such as the cortical surface) is a fundamental component of a variety of studies in Neuroscience. The goal here is to identify regions that exhibit changes as a function of the clinical condition under study. As the clinical questions of interest move towards identifying very early signs of diseases, the corresponding statistical differences at the group level invariably become weaker and increasingly hard to identify. Indeed, after a multiple comparisons correction is adopted (to account for correlated statistical tests over all surface points), very few regions may survive. In contrast to hypothesis tests on point-wise measurements, in this paper, we make the case for performing statistical analysis on multi-scale shape descriptors that c

haracterize the local topological context of the signal around each surface vert
ex. Our descriptors are based on recent results from harmonic analysis, that sho
w how wavelet theory extends to non-Euclidean settings (i.e., irregular weighted
 graphs). We provide strong evidence that these descriptors successfully pick up
 group-wise differences, where traditional methods either fail or yield unsatisf
actory results. Other than this primary application, we show how the framework a
llows performing cortical surface smoothing in the native space without mappint
to a unit sphere.
****

Cardinality Restricted Boltzmann Machines
Kevin Swersky, Ilya Sutskever, Daniel Tarlow, Richard Zemel, Russ R. Salakhutdin
ov, Ryan P. Adams
The Restricted Boltzmann Machine (RBM) is a popular density model that is also g
ood for extracting features. A main source of tractability in RBM models is the
model's assumption that given an input, hidden units activate independently from
 one another. Sparsity and competition in the hidden representation is believed
to be beneficial, and while an RBM with competition among its hidden units would
 acquire some of the attractive properties of sparse coding, such constraints ar
e not added due to the widespread belief that the resulting model would become
intractable. In this work, we show how a dynamic programming algorithm develope
d in 1981 can be used to implement exact sparsity in the RBM's hidden units. We
then expand on this and show how to pass derivatives through a layer of exact sp
arsity, which makes it possible to fine-tune a deep belief network (DBN) consist
ing of RBMs with sparse hidden layers. We show that sparsity in the RBM's hidde
n layer improves the performance of both the pre-trained representations and of
the fine-tuned model.
****

Sparse Prediction with the k-Support Norm
Andreas Argyriou, Rina Foygel, Nathan Srebro
Requests for name changes in the electronic proceedings will be accepted with 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 Marginalized Particle Gaussian Process Regression
Yali Wang, Brahim Chaib-draa
We present a novel marginalized particle Gaussian process (MPGP) regression, whi
ch provides a fast, accurate online Bayesian filtering framework to model the la
tent function. Using a state space model established by the data construction pr
ocedure, our MPGP recursively filters out the estimation of hidden function valu
es by a Gaussian mixture. Meanwhile, it provides a new online method for trainin
g hyperparameters with a number of weighted particles. We demonstrate the estima
ted performance of our MPGP on both simulated and real large data sets. The resu
lts show that our MPGP is a robust estimation algorithm with high computational
efficiency, which outperforms other state-of-art sparse GP methods.
****

Iterative ranking from pair-wise comparisons
Sahand Negahban, Sewoong Oh, Devavrat Shah
The question of aggregating pairwise comparisons to obtain a global ranking over
 a collection of objects has been of interest for a very long time: be it rankin
g of online gamers (e.g. MSR’s TrueSkill system) and chess players, aggregating
social opinions, or deciding which product to sell based on transactions. In mos
t settings, in addition to obtaining ranking, finding ‘scores’ for each object (
e.g. player’s rating) is of interest to understanding the intensity of the prefe
rences.   In this paper, we propose a novel iterative rank aggregation algorithm
 for discovering scores for objects from pairwise comparisons. The algorithm has
 a natural random walk interpretation over the graph of objects with edges prese
nt between two objects if they are compared; the scores turn out to be the stati
onary probability of this random walk. The algorithm is model independent. To es
tablish the efficacy of our method, however, we consider the popular Bradley-Ter

ry-Luce (BTL) model in which each object has an associated score which determine s the probabilistic outcomes of pairwise comparisons between objects. We bound t he finite sample error rates between the scores assumed by the BTL model and tho se estimated by our algorithm. This, in essence, leads to order-optimal dependen ce on the number of samples required to learn the scores well by our algorithm. Indeed, the experimental evaluation shows that our (model independent) algorithm performs as well as the Maximum Likelihood Estimator of the BTL model and outpe rforms a recently proposed algorithm by Ammar and Shah [1].
****

Training sparse natural image models with a fast Gibbs sampler of an extended st ate space
Lucas Theis, Jascha Sohl-dickstein, Matthias Bethge
We present a new learning strategy based on an efficient blocked Gibbs sampler f or sparse overcomplete linear models. Particular emphasis is placed on statistic al image modeling, where overcomplete models have played an important role in di scovering sparse representations. Our Gibbs sampler is faster than general purpo se sampling schemes while also requiring no tuning as it is free of parameters. Using the Gibbs sampler and a persistent variant of expectation maximization, we are able to extract highly sparse distributions over latent sources from data. When applied to natural images, our algorithm learns source distributions which resemble spike-and-slab distributions. We evaluate the likelihood and quantitati vely compare the performance of the overcomplete linear model to its complete co unterpart as well as a product of experts model, which represents another overco mplete generalization of the complete linear model. In contrast to previous clai ms, we find that overcomplete representations lead to significant improvements, but that the overcomplete linear model still underperforms other models.
****

Learning from Distributions via Support Measure Machines
Krikamol Muandet, Kenji Fukumizu, Francesco Dinuzzo, Bernhard Schölkopf
This paper presents a kernel-based discriminative learning framework on probabil ity measures. Rather than relying on large collections of vectorial training exa mples, our framework learns using a collection of probability distributions that have been constructed to meaningfully represent training data. By representing these probability distributions as mean embeddings in the reproducing kernel Hil bert space (RKHS), we are able to apply many standard kernel-based learning tech niques in straightforward fashion. To accomplish this, we construct a generaliza tion of the support vector machine (SVM) called a support measure machine (SMM). Our analyses of SMMs provides several insights into their relationship to tradi tional SVMs. Based on such insights, we propose a flexible SVM (Flex-SVM) that p laces different kernel functions on each training example. Experimental results on both synthetic and real-world data demonstrate the effectiveness of our propo sed framework.
****

Learning Multiple Tasks using Shared Hypotheses
Koby Crammer, Yishay Mansour
In this work we consider a setting where we have a very large number of relate d tasks with few examples from each individual task. Rather than either learni ng each task individually (and having a large generalization error) or learnin g all the tasks together using a single hypothesis (and suffering a potentiall y large inherent error), we consider learning a small pool of {\em shared hypotheses}. Each task is then mapped to a single hypothesis in the pool (hard association). We derive VC dimension generalization bounds for our model, bas ed on the number of tasks, shared hypothesis and the VC dimension of the hypot heses class. We conducted experiments with both synthetic problems and senti ment of reviews, which strongly support our approach.
****

Minimizing Uncertainty in Pipelines
Nilesh Dalvi, Aditya Parameswaran, Vibhor Rastogi
In this paper, we consider the problem of debugging large pipelines by human lab eling. We represent the execution of a pipeline using a directed acyclic graph o

f AND and OR nodes, where each node represents a data item produced by some oper
ator in the pipeline. We assume that each operator assigns a confidence to each
of its output data. We want to reduce the uncertainty in the output by issuing q
ueries to a human expert, where a query consists of checking if a given data ite
m is correct. In this paper, we consider the problem of asking the optimal set o
f queries to minimize the resulting output uncertainty. We perform a detailed ev
aluation of the complexity of the problem for various classes of graphs. We give
 efficient algorithms for the problem for trees, and show that, for a general da
g, the problem is intractable.
****

An Integer Optimization Approach to Associative Classification
Dimitris Bertsimas, Allison Chang, Cynthia Rudin
Requests for name changes in the electronic proceedings will be accepted with 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.
****

Algorithms for Learning Markov Field Policies
Abdeslam Boularias, Jan Peters, Oliver Kroemer
We present a new graph-based approach for incorporating domain knowledge in rein
forcement learning applications. The domain knowledge is given as a weighted gra
ph, or a kernel matrix, that loosely indicates which states should have similar
optimal actions. We first introduce a bias into the policy search process by der
iving a distribution on policies such that policies that disagree with the provi
ded graph have low probabilities. This distribution corresponds to a Markov Rand
om Field. We then present a reinforcement and an apprenticeship learning algorit
hms for finding such policy distributions. We also illustrate the advantage of t
he proposed approach on three problems: swing-up cart-balancing with nonuniform
and smooth frictions, gridworlds, and teaching a robot to grasp new objects.
****

Bayesian Probabilistic Co-Subspace Addition
Lei Shi
For modeling data matrices, this paper introduces Probabilistic Co-Subspace Addi
tion (PCSA) model by simultaneously capturing the dependent structures among bot
h rows and columns. Briefly, PCSA assumes that each entry of a matrix is generat
ed by the additive combination of the linear mappings of two features, which dis
tribute in the row-wise and column-wise latent subspaces. Consequently, it captu
res the dependencies among entries intricately, and is able to model the non-Gau
ssian and heteroscedastic density. Variational inference is proposed on PCSA for
  approximate Bayesian learning, where the updating for posteriors is formulated
 into the problem of solving Sylvester equations. Furthermore, PCSA is extended
to tackling and filling missing values, to adapting its sparseness, and to model
ling tensor data. In comparison with several state-of-art approaches, experiment
s demonstrate the effectiveness and efficiency of Bayesian (sparse) PCSA on mode
ling matrix (tensor) data and filling missing values.
****

Exploration in Model-based Reinforcement Learning by Empirically Estimating Lear
ning Progress
Manuel Lopes, Tobias Lang, Marc Toussaint, Pierre-yves Oudeyer
Formal exploration approaches in model-based reinforcement learning estimate the
 accuracy of the currently learned model without consideration of the empirical
prediction error. For example, PAC-MDP approaches such as Rmax base their model
certainty on the amount of collected data, while Bayesian approaches assume a pr
ior over the transition dynamics. We propose extensions to such approaches which
 drive exploration solely based on empirical estimates of the learner's accuracy
 and learning progress. We provide a ``sanity check'' theoretical analysis, disc
ussing the behavior of our extensions in the standard stationary finite state-ac
tion case. We then provide experimental studies demonstrating the robustness of
these exploration measures in cases of non-stationary environments or where orig
inal approaches are misled by wrong domain assumptions.

****
From Deformations to Parts: Motion-based Segmentation of 3D Objects
Soumya Ghosh, Matthew Loper, Erik Sudderth, Michael Black
We develop a method for discovering the parts of an articulated object from aligned meshes capturing various three-dimensional (3D) poses. We adapt the distance dependent Chinese restaurant process (ddCRP) to allow nonparametric discovery of a potentially unbounded number of parts, while simultaneously guaranteeing a spatially connected segmentation. To allow analysis of datasets in which object instances have varying shapes, we model part variability across poses via affine transformations. By placing a matrix normal-inverse-Wishart prior on these affine transformations, we develop a ddCRP Gibbs sampler which tractably marginalizes over transformation uncertainty. Analyzing a dataset of humans captured in dozens of poses, we infer parts which provide quantitatively better motion predictions than conventional clustering methods.
****
Bayesian models for Large-scale Hierarchical Classification
Siddharth Gopal, Yiming Yang, Bing Bai, Alexandru Niculescu-mizil
A challenging problem in hierarchical classification is to leverage the hierarchical relations among classes for improving classification performance. An even greater challenge is to do so in a manner that is computationally feasible for the large scale problems usually encountered in practice. This paper proposes a set of Bayesian methods to model hierarchical dependencies among class labels using multivari- ate logistic regression. Specifically, the parent-child relationships are modeled by placing a hierarchical prior over the children nodes centered around the parame- ters of their parents; thereby encouraging classes nearby in the hierarchy to share similar model parameters. We present new, efficient variational algorithms for tractable posterior inference in these models, and provide a parallel implementa- tion that can comfortably handle large-scale problems with hundreds of thousands of dimensions and tens of thousands of classes. We run a comparative evaluation on multiple large-scale benchmark datasets that highlights the scalability of our approach, and shows a significant performance advantage over the other state-of- the-art hierarchical methods.
****
Efficient Spike-Coding with Multiplicative Adaptation in a Spike Response Model
Sander Bohte
Neural adaptation underlies the ability of neurons to maximize encoded information over a wide dynamic range of input stimuli. While adaptation is an intrinsic feature of neuronal models like the Hodgkin-Huxley model, the challenge is to integrate adaptation in models of neural computation. Recent computational models like the Adaptive Spike Response Model implement adaptation as spike-based addition of fixed-size fast spike-triggered threshold dynamics and slow spike-triggered currents. Such adaptation has been shown to accurately model neural spiking behavior over a limited dynamic range. Taking a cue from kinetic models of adaptation, we propose a multiplicative Adaptive Spike Response Model where the spike-triggered adaptation dynamics are scaled multiplicatively by the adaptation state at the time of spiking. We show that unlike the additive adaptation model, the firing rate in the multiplicative adaptation model saturates to a maximum spike-rate. When simulating variance switching experiments, the model also quantitatively fits the experimental data over a wide dynamic range. Furthermore, dynamic threshold models of adaptation suggest a straightforward interpretation of neural activity in terms of dynamic signal encoding with shifted and weighted exponential kernels. We show that when thus encoding rectified filtered stimulus signals, the multiplicative Adaptive Spike Response Model achieves a high coding efficiency and maintains this efficiency over changes in the dynamic signal range of several orders of magnitude, without changing model parameters.
****
Forging The Graphs: A Low Rank and Positive Semidefinite Graph Learning Approach
Dijun Luo, Heng Huang, Feiping Nie, Chris Ding
In many graph-based machine learning and data mining approaches, the quality of the graph is critical. However, in real-world applications, especially in semi-s

upervised learning and unsupervised learning, the evaluation of the quality of a graph is often expensive and sometimes even impossible, due the cost or the unavailability of ground truth. In this paper, we proposed a robust approach with convex optimization to ``forge'' a graph: with an input of a graph, to learn a graph with higher quality. Our major concern is that an ideal graph shall satisfy all the following constraints: non-negative, symmetric, low rank, and positive semidefinite. We develop a graph learning algorithm by solving a convex optimization problem and further develop an efficient optimization to obtain global optimal solutions with theoretical guarantees. With only one non-sensitive parameter, our method is shown by experimental results to be robust and achieve higher accuracy in semi-supervised learning and clustering under various settings. As a preprocessing of graphs, our method has a wide range of potential applications machine learning and data mining.
****

Random Utility Theory for Social Choice
Hossein Azari, David Parks, Lirong Xia
Random utility theory models an agents preferences on alternatives by drawing a real-valued score on each alternative (typically independently) from a parameterized distribution, and then ranking the alternatives according to scores. A special case that has received signicant attention is the Plackett-Luce model, for which fast inference methods for maximum likelihood estimators are available. This paper develops conditions on general random utility models that enable fast inference within a Bayesian framework through MC-EM, providing concave loglikelihood functions and bounded sets of global maxima solutions. Results on both real-world and simulated data provide support for the scalability of the approach and capability for model selection among general random utility models including Plackett-Luce.
****

Tensor Decomposition for Fast Parsing with Latent-Variable PCFGs
Michael Collins, Shay Cohen
We describe an approach to speed-up inference with latent variable PCFGs, which have been shown to be highly effective for natural language parsing. Our approach is based on a tensor formulation recently introduced for spectral estimation of latent-variable PCFGs coupled with a tensor decomposition algorithm well-known in the multilinear algebra literature. We also describe an error bound for this approximation, which bounds the difference between the probabilities calculated by the algorithm and the true probabilities that the approximated model gives. Empirical evaluation on real-world natural language parsing data demonstrates a significant speed-up at minimal cost for parsing performance.
****

Action-Model Based Multi-agent Plan Recognition
Hankz Zhuo, Qiang Yang, Subbarao Kambhampati
Multi-Agent Plan Recognition (MAPR) aims to recognize dynamic team structures and team behaviors from the observed team traces (activity sequences) of a set of intelligent agents. Previous MAPR approaches required a library of team activity sequences (team plans) be given as input. However, collecting a library of team plans to ensure adequate coverage is often difficult and costly. In this paper, we relax this constraint, so that team plans are not required to be provided beforehand. We assume instead that a set of action models are available. Such models are often already created to describe domain physics; i.e., the preconditions and effects of effects actions. We propose a novel approach for recognizing multi-agent team plans based on such action models rather than libraries of team plans. We encode the resulting MAPR problem as a \emph{satisfiability problem} and solve the problem using a state-of-the-art weighted MAX-SAT solver. Our approach also allows for incompleteness in the observed plan traces. Our empirical studies demonstrate that our algorithm is both effective and efficient in comparison to state-of-the-art MAPR methods based on plan libraries.
****

Semiparametric Principal Component Analysis
Fang Han, Han Liu

We propose two new principal component analysis methods in this paper utilizing a semiparametric model. The according methods are named Copula Component Analysis (COCA) and Copula PCA. The semiparametric model assumes that, af- ter unspeci■ed marginally monotone transformations, the distributions are multi- variate Gaussian. The COCA and Copula PCA accordingly estimate the leading eigenvectors of the correlation and covariance matrices of the latent Gaussian dis- tribution. The robust nonparametric rank-based correlation coef■cient estimator, Spearman's rho, is exploited in estimation. We prove that, under suitable condi- tions, although the marginal distributions can be arbitrarily continuous, the COCA and Copula PCA estimators obtain fast estimation rates and are feature selection consistent in the setting where the dimension is nearly exponentially large relative to the sample size. Careful numerical experiments on the synthetic and real data are conducted to back up the theoretical results. We also discuss the relationship with the transelliptical component analysis proposed by Han and Liu (2012).
****

Multi-task Vector Field Learning
Binbin Lin, Sen Yang, Chiyuan Zhang, Jieping Ye, Xiaofei He
Multi-task learning (MTL) aims to improve generalization performance by learning multiple related tasks simultaneously and identifying the shared information among tasks. Most of existing MTL methods focus on learning linear models under the supervised setting. We propose a novel semi-supervised and nonlinear approach for MTL using vector fields. A vector field is a smooth mapping from the manifold to the tangent spaces which can be viewed as a directional derivative of functions on the manifold. We argue that vector fields provide a natural way to exploit the geometric structure of data as well as the shared differential structure of tasks, both are crucial for semi-supervised multi-task learning. In this paper, we develop multi-task vector field learning (MTVFL) which learns the prediction functions and the vector fields simultaneously. MTVFL has the following key properties: (1) the vector fields we learned are close to the gradient fields of the prediction functions; (2) within each task, the vector field is required to be as parallel as possible which is expected to span a low dimensional subspace; (3) the vector fields from all tasks share a low dimensional subspace. We formalize our idea in a regularization framework and also provide a convex relaxation method to solve the original non-convex problem. The experimental results on synthetic and real data demonstrate the effectiveness of our proposed approach.
****

Local Supervised Learning through Space Partitioning
Joseph Wang, Venkatesh Saligrama
We develop a novel approach for supervised learning based on adaptively partitioning the feature space into different regions and learning local region-specific classifiers. We formulate an empirical risk minimization problem that incorporates both partitioning and classification in to a single global objective. We show that space partitioning can be equivalently reformulated as a supervised learning problem and consequently any discriminative learning method can be utilized in conjunction with our approach. Nevertheless, we consider locally linear schemes by learning linear partitions and linear region classifiers. Locally linear schemes can not only approximate complex decision boundaries and ensure low training error but also provide tight control on over-fitting and generalization error. We train locally linear classifiers by using LDA, logistic regression and perceptrons, and so our scheme is scalable to large data sizes and high-dimensions. We present experimental results demonstrating improved performance over state of the art classification techniques on benchmark datasets. We also show improved robustness to label noise.
****

Dip-means: an incremental clustering method for estimating the number of clusters
Argyris Kalogeratos, Aristidis Likas
Learning the number of clusters is a key problem in data clustering. We present dip-means, a novel robust incremental method to learn the number of data clusters that may be used as a wrapper around any iterative clustering algorithm of the

k-means family. In contrast to many popular methods which make assumptions about the underlying cluster distributions, dip-means only assumes a fundamental cluster property: each cluster to admit a unimodal distribution. The proposed algorithm considers each cluster member as a ''viewer'' and applies a univariate statistic hypothesis test for unimodality (dip-test) on the distribution of the distances between the viewer and the cluster members. Two important advantages are: i) the unimodality test is applied on univariate distance vectors, ii) it can be directly applied with kernel-based methods, since only the pairwise distances are involved in the computations. Experimental results on artificial and real datasets indicate the effectiveness of our method and its superiority over analogous approaches.

****

Unsupervised Template Learning for Fine-Grained Object Recognition

Shulin Yang, Liefeng Bo, Jue Wang, Linda Shapiro

Fine-grained recognition refers to a subordinate level of recognition, such as recognizing different species of birds, animals or plants. It differs from recognition of basic categories, such as humans, tables, and computers, in that there are global similarities in shape or structure shared within a category, and the differences are in the details of the object parts. We suggest that the key to identifying the fine-grained differences lies in finding the right alignment of image regions that contain the same object parts. We propose a template model for the purpose, which captures common shape patterns of object parts, as well as the co-occurence relation of the shape patterns. Once the image regions are aligned, extracted features are used for classification. Learning of the template model is efficient, and the recognition results we achieve significantly outperform the state-of-the-art algorithms.

****

A Nonparametric Conjugate Prior Distribution for the Maximizing Argument of a Noisy Function

Pedro Ortega, Jordi Grau-moya, Tim Genewein, David Balduzzi, Daniel Braun

We propose a novel Bayesian approach to solve stochastic optimization problems that involve ■nding extrema of noisy, nonlinear functions. Previous work has focused on representing possible functions explicitly, which leads to a two-step procedure of ■rst, doing inference over the function space and second, ■nding the extrema of these functions. Here we skip the representation step and directly model the distribution over extrema. To this end, we devise a non-parametric conjugate prior where the natural parameter corresponds to a given kernel function and the suf■cient statistic is composed of the observed function values. The resulting posterior distribution directly captures the uncertainty over the maximum of the unknown function.

****

Efficient Reinforcement Learning for High Dimensional Linear Quadratic Systems

Morteza Ibrahimi, Adel Javanmard, Benjamin 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-authors prior to requesting a name change in the electronic proceedings.

****

A Conditional Multinomial Mixture Model for Superset Label Learning

Liping Liu, Thomas Dietterich

In the superset label learning problem (SLL), each training instance provides a set of candidate labels of which one is the true label of the instance. As in ordinary regression, the candidate label set is a noisy version of the true label. In this work, we solve the problem by maximizing the likelihood of the candidate label sets of training instances. We propose a probabilistic model, the Logistic Stick- Breaking Conditional Multinomial Model (LSB-CMM), to do the job. The LSB- CMM is derived from the logistic stick-breaking process. It ■rst maps data points to mixture components and then assigns to each mixture component a label drawn from a component-speci■c multinomial distribution. The mixture components can capture underlying structure in the data, which is very useful when the model

is weakly supervised. This advantage comes at little cost, since the model intr oduces few additional parameters. Experimental tests on several real-world probl ems with superset labels show results that are competitive or superior to the st ate of the art. The discovered underlying structures also provide improved expla nations of the classi■cation predictions.
****

Value Pursuit Iteration
Amir Farahmand, Doina Precup
Value Pursuit Iteration (VPI) is an approximate value iteration algorithm that f inds a close to optimal policy for reinforcement learning and planning problems with large state spaces. VPI has two main features: First, it is a nonparametri c algorithm that finds a good sparse approximation of the optimal value function given a dictionary of features. The algorithm is almost insensitive to the numb er of irrelevant features. Second, after each iteration of VPI, the algorithm ad ds a set of functions based on the currently learned value function to the dicti onary. This increases the representation power of the dictionary in a way that i s directly relevant to the goal of having a good approximation of the optimal va lue function. We theoretically study VPI and provide a finite-sample error upper bound for it.
****

Latent Coincidence Analysis: A Hidden Variable Model for Distance Metric Learnin g
Matthew Der, Lawrence Saul
We describe a latent variable model for supervised dimensionality reduction and distance metric learning. The model discovers linear projections of high dimensi onal data that shrink the distance between similarly labeled inputs and expand t he distance between differently labeled ones. The model's continuous latent vari ables locate pairs of examples in a latent space of lower dimensionality. The mo del differs significantly from classical factor analysis in that the posterior d istribution over these latent variables is not always multivariate Gaussian. Nev ertheless we show that inference is completely tractable and derive an Expectati on-Maximization (EM) algorithm for parameter estimation. We also compare the mod el to other approaches in distance metric learning. The model's main advantage i s its simplicity: at each iteration of the EM algorithm, the distance metric is re-estimated by solving an unconstrained least-squares problem. Experiments show that these simple updates are highly effective.
****

Identification of Recurrent Patterns in the Activation of Brain Networks
Firdaus Janoos, Weichang Li, Niranjan Subrahmanya, Istvan Morocz, William Wells
Identifying patterns from the neuroimaging recordings of brain activity related to the unobservable psychological or mental state of an individual can be treat ed as a unsupervised pattern recognition problem. The main challenges, however, for such an analysis of fMRI data are: a) defining a physiologically meaningful feature-space for representing the spatial patterns across time; b) dealing with the high-dimensionality of the data; and c) robustness to the various artifacts and confounds in the fMRI time-series. In this paper, we present a network-awa re feature-space to represent the states of a general network, that enables com paring and clustering such states in a manner that is a) meaningful in terms of the network connectivity structure; b)computationally efficient; c) low-dimensio nal; and d) relatively robust to structured and random noise artifacts. This fea ture-space is obtained from a spherical relaxation of the transportation distanc e metric which measures the cost of transporting ``mass'' over the network to tr ansform one function into another. Through theoretical and empirical assessments , we demonstrate the accuracy and efficiency of the approximation, especially fo r large problems. While the application presented here is for identifying disti nct brain activity patterns from fMRI, this feature-space can be applied to the problem of identifying recurring patterns and detecting outliers in measurements on many different types of networks, including sensor, control and social netwo rks.
****

Hierarchical spike coding of sound
Yan Karklin, Chaitanya Ekanadham, Eero Simoncelli
We develop a probabilistic generative model for representing acoustic event stru cture at multiple scales via a two-stage hierarchy. The first stage consists of a spiking representation which encodes a sound with a sparse set of kernels at d ifferent frequencies positioned precisely in time. The coarse time and frequency statistical structure of the first-stage spikes is encoded by a second stage sp iking representation, while fine-scale statistical regularities are encoded by r ecurrent interactions within the first-stage. When fitted to speech data, the m odel encodes acoustic features such as harmonic stacks, sweeps, and frequency mo dulations, that can be composed to represent complex acoustic events. The model is also able to synthesize sounds from the higher-level representation and provi des significant improvement over wavelet thresholding techniques on a denoising task.
****

A Polynomial-time Form of Robust Regression
Yao-liang Yu, Özlem Aslan, Dale Schuurmans
Despite the variety of robust regression methods that have been developed, curre nt regression formulations are either NP-hard, or allow unbounded response to ev en a single leverage point. We present a general formulation for robust regressi on --Variational M-estimation--that unifies a number of robust regression method s while allowing a tractable approximation strategy. We develop an estimator tha t requires only polynomial-time, while achieving certain robustness and consiste ncy guarantees. An experimental evaluation demonstrates the effectiveness of the new estimation approach compared to standard methods.
****

Multimodal Learning with Deep Boltzmann Machines
Nitish Srivastava, Russ R. Salakhutdinov
We propose a Deep Boltzmann Machine for learning a generative model of multimoda l data. We show how to use the model to extract a meaningful representation of m ultimodal data. We find that the learned representation is useful for classifica tion and information retreival tasks, and hence conforms to some notion of seman tic similarity. The model defines a probability density over the space of multim odal inputs. By sampling from the conditional distributions over each data modal ity, it possible to create the representation even when some data modalities are missing. Our experimental results on bi-modal data consisting of images and tex t show that the Multimodal DBM can learn a good generative model of the joint sp ace of image and text inputs that is useful for information retrieval from both unimodal and multimodal queries. We further demonstrate that our model can signi ficantly outperform SVMs and LDA on discriminative tasks. Finally, we compare ou r model to other deep learning methods, including autoencoders and deep belief n etworks, and show that it achieves significant gains.
****

A nonparametric variable clustering model
Konstantina Palla, Zoubin Ghahramani, David Knowles
Factor analysis models effectively summarise the covariance structure of high di mensional data, but the solutions are typically hard to interpret. This motivate s attempting to find a disjoint partition, i.e. a clustering, of observed variab les so that variables in a cluster are highly correlated. We introduce a Bayesia n non-parametric approach to this problem, and demonstrate advantages over heuri stic methods proposed to date.
****

Transelliptical Graphical Models
Han Liu, Fang Han, Cun-hui Zhang
We advocate the use of a new distribution family—the transelliptical—for robust inference of high dimensional graphical models. The transelliptical family is an extension of the nonparanormal family proposed by Liu et al. (2009). Just as th e nonparanormal extends the normal by transforming the variables using univariat e functions, the transelliptical extends the elliptical family in the same way. We propose a nonparametric rank-based regularization estimator which achieves th

e parametric rates of convergence for both graph recovery and parameter estimation. Such a result suggests that the extra robustness and ■exibility obtained by the semiparametric transelliptical modeling incurs almost no ef■ciency loss. We also discuss the relationship between this work with the transelliptical component analysis proposed by Han and Liu (2012).
****

Collaborative Gaussian Processes for Preference Learning
Neil Houlsby, Ferenc Huszar, Zoubin Ghahramani, Jose Hernández-lobato
We present a new model based on Gaussian processes (GPs) for learning pairwise preferences expressed by multiple users. Inference is simplified by using a \emph{preference kernel} for GPs which allows us to combine supervised GP learning of user preferences with unsupervised dimensionality reduction for multi-user systems. The model not only exploits collaborative information from the shared structure in user behavior, but may also incorporate user features if they are available. Approximate inference is implemented using a combination of expectation propagation and variational Bayes. Finally, we present an efficient active learning strategy for querying preferences. The proposed technique performs favorably on real-world data against state-of-the-art multi-user preference learning algorithms.
****

Smooth-projected Neighborhood Pursuit for High-dimensional Nonparanormal Graph Estimation
Tuo Zhao, Kathryn Roeder, Han Liu
Many statistical methods gain robustness and exibility by sacricing convenient computational structure. In this paper, we illustrate this fundamental tradeoff by studying a semiparametric graphical model estimation problem. We explain how new computational techniques help to solve this type of problem. In particularly, we propose a smooth-projected neighborhood pursuit method for efciently estimating high dimensional nonparanormal graphs with theoretical guarantees. Besides new computational and theoretical analysis, we also provide an alternative view to analyze the tradeoff between computational efciency and statistical error under a smoothing optimization framework. We also report experimental results on text and stock datasets.
****

Learning Manifolds with K-Means and K-Flats
Guillermo Canas, Tomaso Poggio, Lorenzo Rosasco
We study the problem of estimating a manifold from random samples. In particular, we consider piecewise constant and piecewise linear estimators induced by k-means and k-■ats, and analyze their performance. We extend previous results for k-means in two separate directions. First, we provide new results for k-means reconstruction on manifolds and, secondly, we prove reconstruction bounds for higher-order approximation (k-■ats), for which no known results were previously available. While the results for k-means are novel, some of the technical tools are well-established in the literature. In the case of k-■ats, both the results and the mathematical tools are new.
****

Newton-Like Methods for Sparse Inverse Covariance Estimation
Figen Oztoprak, Jorge Nocedal, Steven Rennie, Peder A. Olsen
We propose two classes of second-order optimization methods for solving the sparse inverse covariance estimation problem. The first approach, which we call the Newton-LASSO method, minimizes a piecewise quadratic model of the objective function at every iteration to generate a step. We employ the fast iterative shrinkage thresholding method (FISTA) to solve this subproblem. The second approach, which we call the Orthant-Based Newton method, is a two-phase algorithm that first identifies an orthant face and then minimizes a smooth quadratic approximation of the objective function using the conjugate gradient method. These methods exploit the structure of the Hessian to efficiently compute the search direction and to avoid explicitly storing the Hessian. We show that quasi-Newton methods are also effective in this context, and describe a limited memory BFGS variant of the orthant-based Newton method. We present numerical results that suggest tha

t all the techniques described in this paper have attractive properties and cons
titute useful tools for solving the sparse inverse covariance estimation problem
. Comparisons with the method implemented in the QUIC software package are prese
nted.
****

A Neural Autoregressive Topic Model
Hugo Larochelle, Stanislas Lauly
We describe a new model for learning meaningful representations of text document
s from an unlabeled collection of documents. This model is inspired by the recen
tly proposed Replicated Softmax, an undirected graphical model of word counts th
at was shown to learn a better generative model and more meaningful document rep
resentations. Specifically, we take inspiration from the conditional mean-field
recursive equations of the Replicated Softmax in order to define a neural networ
k architecture that estimates the probability of observing a new word in a given
 document given the previously observed words. This paradigm also allows us to r
eplace the expensive softmax distribution over words with a hierarchical distrib
ution over paths in a binary tree of words. The end result is a model whose trai
ning complexity scales logarithmically with the vocabulary size instead of linea
rly as in the Replicated Softmax. Our experiments show that our model is competi
tive both as a generative model of documents and as a document representation le
arning algorithm.
****

Active Learning of Multi-Index Function Models
Tyagi Hemant, Volkan Cevher
Requests for name changes in the electronic proceedings will be accepted with 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.
****

Probabilistic Low-Rank Subspace Clustering
S. Babacan, Shinichi Nakajima, Minh Do
In this paper, we consider the problem of clustering data points into low-dimens
ional subspaces in the presence of outliers. We pose the problem using a density
 estimation formulation with an associated generative model. Based on this proba
bility model, we first develop an iterative expectation-maximization (EM) algori
thm and then derive its global solution. In addition, we develop two Bayesian me
thods based on variational Bayesian (VB) approximation, which are capable of aut
omatic dimensionality selection. While the first method is based on an alternati
ng optimization scheme for all unknowns, the second method makes use of recent r
esults in VB matrix factorization leading to fast and effective estimation. Both
 methods are extended to handle sparse outliers for robustness and can handle mi
ssing values. Experimental results suggest that proposed methods are very effect
ive in clustering and identifying outliers.
****

Fully Bayesian inference for neural models with negative-binomial spiking
Jonathan Pillow, James Scott
Characterizing the information carried by neural populations in the brain requir
es accurate statistical models of neural spike responses. The negative-binomial
 distribution provides a convenient model for over-dispersed spike counts, that
is, responses with greater-than-Poisson variability. Here we describe a powerfu
l data-augmentation framework for fully Bayesian inference in neural models with
 negative-binomial spiking. Our approach relies on a recently described latent-v
ariable representation of the negative-binomial distribution, which equates it t
o a Polya-gamma mixture of normals. This framework provides a tractable, condit
ionally Gaussian representation of the posterior that can be used to design effi
cient EM and Gibbs sampling based algorithms for inference in regression and dyn
amic factor models. We apply the model to neural data from primate retina and s
how that it substantially outperforms Poisson regression on held-out data, and r
eveals latent structure underlying spike count correlations in simultaneously re
corded spike trains.

****
Adaptive Learning of Smoothing Functions: Application to Electricity Load Forecasting
Amadou Ba, Mathieu Sinn, Yannig Goude, Pascal Pompey
This paper proposes an efficient online learning algorithm to track the smoothing functions of Additive Models. The key idea is to combine the linear representation of Additive Models with a Recursive Least Squares (RLS) filter. In order to quickly track changes in the model and put more weight on recent data, the RLS filter uses a forgetting factor which exponentially weights down observations by the order of their arrival. The tracking behaviour is further enhanced by using an adaptive forgetting factor which is updated based on the gradient of the a priori errors. Using results from Lyapunov stability theory, upper bounds for the learning rate are analyzed. The proposed algorithm is applied to 5 years of electricity load data provided by the French utility company Electricite de France (EDF). Compared to state-of-the-art methods, it achieves a superior performance in terms of model tracking and prediction accuracy.
****
Complex Inference in Neural Circuits with Probabilistic Population Codes and Topic Models
Jeff Beck, Alexandre Pouget, Katherine A. Heller
Recent experiments have demonstrated that humans and animals typically reason probabilistically about their environment. This ability requires a neural code that represents probability distributions and neural circuits that are capable of implementing the operations of probabilistic inference. The proposed probabilistic population coding (PPC) framework provides a statistically efficient neural representation of probability distributions that is both broadly consistent with physiological measurements and capable of implementing some of the basic operations of probabilistic inference in a biologically plausible way. However, these experiments and the corresponding neural models have largely focused on simple (tractable) probabilistic computations such as cue combination, coordinate transformations, and decision making. As a result it remains unclear how to generalize this framework to more complex probabilistic computations. Here we address this short coming by showing that a very general approximate inference algorithm known as Variational Bayesian Expectation Maximization can be implemented within the linear PPC framework. We apply this approach to a generic problem faced by any given layer of cortex, namely the identification of latent causes of complex mixtures of spikes. We identify a formal equivalent between this spike pattern demixing problem and topic models used for document classification, in particular Latent Dirichlet Allocation (LDA). We then construct a neural network implementation of variational inference and learning for LDA that utilizes a linear PPC. This network relies critically on two non-linear operations: divisive normalization and super-linear facilitation, both of which are ubiquitously observed in neural circuits. We also demonstrate how online learning can be achieved using a variation of Hebb’s rule and describe an extesion of this work which allows us to deal with time varying and correlated latent causes.
****
Clustering by Nonnegative Matrix Factorization Using Graph Random Walk
Zhirong Yang, Tele Hao, Onur Dikmen, Xi Chen, Erkki Oja
Nonnegative Matrix Factorization (NMF) is a promising relaxation technique for clustering analysis. However, conventional NMF methods that directly approximate the pairwise similarities using the least square error often yield mediocre performance for data in curved manifolds because they can capture only the immediate similarities between data samples. Here we propose a new NMF clustering method which replaces the approximated matrix with its smoothed version using random walk. Our method can thus accommodate farther relationships between data samples. Furthermore, we introduce a novel regularization in the proposed objective function in order to improve over spectral clustering. The new learning objective is optimized by a multiplicative Majorization-Minimization algorithm with a scalable implementation for learning the factorizing matrix. Extensive experimental results on real-world datasets show that our method has strong performance in

terms of cluster purity.
****

Graphical Gaussian Vector for Image Categorization
Tatsuya Harada, Yasuo Kuniyoshi
This paper proposes a novel image representation called a Graphical Gaussian Vector, which is a counterpart of the codebook and local feature matching approaches. In our method, we model the distribution of local features as a Gaussian Markov Random Field (GMRF) which can efficiently represent the spatial relationship among local features. We consider the parameter of GMRF as a feature vector of the image. Using concepts of information geometry, proper parameters and a metric from the GMRF can be obtained. Finally we define a new image feature by embedding the metric into the parameters, which can be directly applied to scalable linear classifiers. Our method obtains superior performance over the state-of-the-art methods in the standard object recognition datasets and comparable performance in the scene dataset. As the proposed method simply calculates the local auto-correlations of local features, it is able to achieve both high classification accuracy and high efficiency.
****

Convergence Rate Analysis of MAP Coordinate Minimization Algorithms
Ofer Meshi, Amir Globerson, Tommi Jaakkola
Finding maximum aposteriori (MAP) assignments in graphical models is an important task in many applications. Since the problem is generally hard, linear programming (LP) relaxations are often used. Solving these relaxations efficiently is thus an important practical problem. In recent years, several authors have proposed message passing updates corresponding to coordinate descent in the dual LP. However,these are generally not guaranteed to converge to a global optimum. One approach to remedy this is to smooth the LP, and perform coordinate descent on the smoothed dual. However, little is known about the convergence rate of this procedure. Here we perform a thorough rate analysis of such schemes and derive primal and dual convergence rates. We also provide a simple dual to primal mapping that yields feasible primal solutions with a guaranteed rate of convergence. Empirical evaluation supports our theoretical claims and shows that the method is highly competitive with state of the art approaches that yield global optima.
****

Distributed Probabilistic Learning for Camera Networks with Missing Data
Sejong Yoon, Vladimir Pavlovic
Probabilistic approaches to computer vision typically assume a centralized setting, with the algorithm granted access to all observed data points. However, many problems in wide-area surveillance can benefit from distributed modeling, either because of physical or computational constraints. Most distributed models to date use algebraic approaches (such as distributed SVD) and as a result cannot explicitly deal with missing data. In this work we present an approach to estimation and learning of generative probabilistic models in a distributed context where certain sensor data can be missing. In particular, we show how traditional centralized models, such as probabilistic PCA and missing-data PPCA, can be learned when the data is distributed across a network of sensors. We demonstrate the utility of this approach on the problem of distributed affine structure from motion. Our experiments suggest that the accuracy of the learned probabilistic structure and motion models rivals that of traditional centralized factorization methods while being able to handle challenging situations such as missing or noisy observations.
****

Confusion-Based Online Learning and a Passive-Aggressive Scheme
Liva Ralaivola
This paper provides the first ---to the best of our knowledge--- analysis of online learning algorithms for multiclass problems when the {\em confusion} matrix is taken as a performance measure. The work builds upon recent and elegant results on noncommutative concentration inequalities, i.e. concentration inequalities that apply to matrices, and more precisely to matrix martingales. We do establish generalization bounds for online learning algorithm and show how the theoret

ical study motivate the proposition of a new confusion-friendly learning procedu
re. This learning algorithm, called \copa (for COnfusion Passive-Aggressive) is
a passive-aggressive learning algorithm; it is shown that the update equations f
or \copa can be computed analytically, thus allowing the user from having to rec
ours to any optimization package to implement it.
****

Context-Sensitive Decision Forests for Object Detection
Peter Kontschieder, Samuel Bulò, Antonio Criminisi, Pushmeet Kohli, Marcello Pel
illo, Horst Bischof
In this paper we introduce Context-Sensitive Decision Forests - A new perspectiv
e to exploit contextual information in the popular decision forest framework for
 the object detection problem. They are tree-structured classifiers with the abi
lity to access intermediate prediction (here: classification and regression) inf
ormation during training and inference time. This intermediate prediction is ava
ilable to each sample, which allows us to develop context-based decision criteri
a, used for refining the prediction process. In addition, we introduce a novel s
plit criterion which in combination with a priority based way of constructing th
e trees, allows more accurate regression mode selection and hence improves the c
urrent context information. In our experiments, we demonstrate improved results
for the task of pedestrian detection on the challenging TUD data set when compar
ed to state-of-the-art methods.
****

Approximating Concavely Parameterized Optimization Problems
Joachim Giesen, Jens Mueller, Soeren Laue, Sascha Swiercy
Requests for name changes in the electronic proceedings will be accepted with 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.
****

Kernel Latent SVM for Visual Recognition
Weilong Yang, Yang Wang, Arash Vahdat, Greg Mori
Latent SVMs (LSVMs) are a class of powerful tools that have been successfully ap
plied to many applications in computer vision. However, a limitation of LSVMs is
 that they rely on linear models. For many computer vision tasks, linear models
are suboptimal and nonlinear models learned with kernels typically perform much
better. Therefore it is desirable to develop the kernel version of LSVM. In this
 paper, we propose kernel latent SVM (KLSVM) -- a new learning framework that co
mbines latent SVMs and kernel methods. We develop an iterative training algorith
m to learn the model parameters. We demonstrate the effectiveness of KLSVM using
 three different applications in visual recognition. Our KLSVM formulation is ve
ry general and can be applied to solve a wide range of applications in computer
vision and machine learning.
****

A Linear Time Active Learning Algorithm for Link Classification
Nicolò Cesa-bianchi, Claudio Gentile, Fabio Vitale, Giovanni Zappella
Requests for name changes in the electronic proceedings will be accepted with 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 P300 BCI for the Masses: Prior Information Enables Instant Unsupervised Spelli
ng
Pieter-jan Kindermans, Hannes Verschore, David Verstraeten, Benjamin Schrauwen
The usability of Brain Computer Interfaces (BCI) based on the P300 speller is se
verely hindered by the need for long training times and many repetitions of the
same stimulus. In this contribution we introduce a set of unsupervised hierarchi
cal probabilistic models that tackle both problems simultaneously by incorporati
ng prior knowledge from two sources: information from other training subjects (t
hrough transfer learning) and information about the words being spelled (through
 language models). We show, that due to this prior knowledge, the performance of

the unsupervised models parallels and in some cases even surpasses that of supe
rvised models, while eliminating the tedious training session.
****
Dynamic Pruning of Factor Graphs for Maximum Marginal Prediction
Christoph H. Lampert
We study the problem of maximum marginal prediction (MMP) in probabilistic graph
ical models, a task that occurs, for example, as the Bayes optimal decision rule
under a Hamming loss. MMP is typically performed as a two-stage procedure: one
estimates each variable's marginal probability and then forms a prediction from
the states of maximal probability. In this work we propose a simple yet effectiv
e technique for accelerating MMP when inference is sampling-based: instead of th
e above two-stage procedure we directly estimate the posterior probability of ea
ch decision variable. This allows us to identify the point of time when we are s
ufficiently certain about any individual decision. Whenever this is the case, we
dynamically prune the variable we are confident about from the underlying facto
r graph. Consequently, at any time only samples of variable whose decision is st
ill uncertain need to be created. Experiments in two prototypical scenarios, mul
ti-label classification and image inpainting, shows that adaptive sampling can d
rastically accelerate MMP without sacrificing prediction accuracy.
****
Locally Uniform Comparison Image Descriptor
Andrew Ziegler, Eric Christiansen, David Kriegman, Serge Belongie
Keypoint matching between pairs of images using popular descriptors like SIFT or
a faster variant called SURF is at the heart of many computer vision algorithms
including recognition, mosaicing, and structure from motion. For real-time mobi
le applications, very fast but less accurate descriptors like BRIEF and related
methods use a random sampling of pairwise comparisons of pixel intensities in an
image patch. Here, we introduce Locally Uniform Comparison Image Descriptor (LU
CID), a simple description method based on permutation distances between the ord
ering of intensities of RGB values between two patches. LUCID is computable in l
inear time with respect to patch size and does not require floating point comput
ation. An analysis reveals an underlying issue that limits the potential of BRIE
F and related approaches compared to LUCID. Experiments demonstrate that LUCID i
s faster than BRIEF, and its accuracy is directly comparable to SURF while being
more than an order of magnitude faster.
****
Priors for Diversity in Generative Latent Variable Models
James Kwok, Ryan P. Adams
Probabilistic latent variable models are one of the cornerstones of machine lear
ning. They offer a convenient and coherent way to specify prior distributions o
ver unobserved structure in data, so that these unknown properties can be inferr
ed via posterior inference. Such models are useful for exploratory analysis and
visualization, for building density models of data, and for providing feature
s that can be used for later discriminative tasks. A significant limitation of t
hese models, however, is that draws from the prior are often highly redundant du
e to i.i.d. assumptions on internal parameters. For example, there is no pref
erence in the prior of a mixture model to make components non-overlapping, or in
topic model to ensure that co-ocurring words only appear in a small number of t
opics. In this work, we revisit these independence assumptions for probabilisti
c latent variable models, replacing the underlying i.i.d.\ prior with a determ
inantal point process (DPP). The DPP allows us to specify a preference for diver
sity in our latent variables using a positive definite kernel function. Using a
kernel between probability distributions, we are able to define a DPP on probab
ility measures. We show how to perform MAP inference with DPP priors in laten
t Dirichlet allocation and in mixture models, leading to better intuition for th
e latent variable representation and quantitatively improved unsupervised featur
e extraction, without compromising the generative aspects of the model.
****
Mixing Properties of Conditional Markov Chains with Unbounded Feature Functions
Mathieu Sinn, Bei Chen

Conditional Markov Chains (also known as Linear-Chain Conditional Random Fields
 in the literature) are a versatile class of discriminative models for the distr
ibution of a sequence of hidden states conditional on a sequence of observable v
ariables. Large-sample properties of Conditional Markov Chains have been first s
tudied by Sinn and Poupart [1]. The paper extends this work in two directions: f
irst, mixing properties of models with unbounded feature functions are being est
ablished; second, necessary conditions for model identifiability and the uniquen
ess of maximum likelihood estimates are being given.
****

ImageNet Classification with Deep Convolutional Neural Networks
Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton
We trained a large, deep convolutional neural network to classify the 1.3 millio
n high-resolution images in the LSVRC-2010 ImageNet training set into the 1000 d
ifferent classes. On the test data, we achieved top-1 and top-5 error rates of 3
9.7\% and 18.9\% which is considerably better than the previous state-of-the-art
 results. The neural network, which has 60 million parameters and 500,000 neuron
s, consists of five convolutional layers, some of which are followed by max-pool
ing layers, and two globally connected layers with a final 1000-way softmax. To
make training faster, we used non-saturating neurons and a very efficient GPU im
plementation of convolutional nets. To reduce overfitting in the globally connec
ted layers we employed a new regularization method that proved to be very effect
ive.
****

Stochastic Gradient Descent with Only One Projection
Mehrdad Mahdavi, Tianbao Yang, Rong Jin, Shenghuo Zhu, Jinfeng Yi
Requests for name changes in the electronic proceedings will be accepted with 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 Probability Measures with respect to Optimal Transport Metrics
Guillermo Canas, Lorenzo Rosasco
We study the problem of estimating, in the sense of optimal transport metrics, a
 measure which is assumed supported on a manifold embedded in a Hilbert space. B
y establishing a precise connection between optimal transport metrics, optimal q
uantization, and learning theory, we derive new probabilistic bounds for the per
formance of a classic algorithm in unsupervised learning (k-means), when used to
 produce a probability measure derived from the data. In the course of the analy
sis, we arrive at new lower bounds, as well as probabilistic bounds on the conve
rgence rate of the empirical law of large numbers, which, unlike existing bounds
, are applicable to a wide class of measures.
****

Effective Split-Merge Monte Carlo Methods for Nonparametric Models of Sequential
 Data
Michael C. Hughes, Emily Fox, Erik Sudderth
Applications of Bayesian nonparametric methods require learning and inference al
gorithms which efficiently explore models of unbounded complexity. We develop ne
w Markov chain Monte Carlo methods for the beta process hidden Markov model (BP-
HMM), enabling discovery of shared activity patterns in large video and motion c
apture databases. By introducing split-merge moves based on sequential allocatio
n, we allow large global changes in the shared feature structure. We also develo
p data-driven reversible jump moves which more reliably discover rare or unique
behaviors. Our proposals apply to any choice of conjugate likelihood for observe
d data, and we show success with multinomial, Gaussian, and autoregressive emiss
ion models. Together, these innovations allow tractable analysis of hundreds of
time series, where previous inference required clever initialization and at leas
t ten thousand burn-in iterations for just six sequences.
****

Scaling MPE Inference for Constrained Continuous Markov Random Fields with Conse
nsus Optimization

Stephen Bach, Matthias Broecheler, Lise Getoor, Dianne O'leary
Probabilistic graphical models are powerful tools for analyzing constrained, continuous domains. However, finding most-probable explanations (MPEs) in these models can be computationally expensive. In this paper, we improve the scalability of MPE inference in a class of graphical models with piecewise-linear and piecewise-quadratic dependencies and linear constraints over continuous domains. We derive algorithms based on a consensus-optimization framework and demonstrate their superior performance over state of the art. We show empirically that in a large-scale voter-preference modeling problem our algorithms scale linearly in the number of dependencies and constraints.
****

A systematic approach to extracting semantic information from functional MRI data
Francisco Pereira, Matthew Botvinick
This paper introduces a novel classification method for functional magnetic resonance imaging datasets with tens of classes. The method is designed to make predictions using information from as many brain locations as possible, instead of resorting to feature selection, and does this by decomposing the pattern of brain activation into differently informative sub-regions. We provide results over a complex semantic processing dataset that show that the method is competitive with state-of-the-art feature selection and also suggest how the method may be used to perform group or exploratory analyses of complex class structure.
****

Sketch-Based Linear Value Function Approximation
Marc Bellemare, Joel Veness, Michael Bowling
Hashing is a common method to reduce large, potentially infinite feature vectors to a fixed-size table. In reinforcement learning, hashing is often used in conjunction with tile coding to represent states in continuous spaces. Hashing is also a promising approach to value function approximation in large discrete domains such as Go and Hearts, where feature vectors can be constructed by exhaustively combining a set of atomic features. Unfortunately, the typical use of hashing in value function approximation results in biased value estimates due to the possibility of collisions. Recent work in data stream summaries has led to the development of the tug-of-war sketch, an unbiased estimator for approximating inner products. Our work investigates the application of this new data structure to linear value function approximation. Although in the reinforcement learning setting the use of the tug-of-war sketch leads to biased value estimates, we show that this bias can be orders of magnitude less than that of standard hashing. We provide empirical results on two RL benchmark domains and fifty-five Atari 2600 games to highlight the superior learning performance of tug-of-war hashing.
****

Unsupervised Structure Discovery for Semantic Analysis of Audio
Sourish Chaudhuri, Bhiksha Raj
Approaches to audio classification and retrieval tasks largely rely on detection-based discriminative models. We submit that such models make a simplistic assumption in mapping acoustics directly to semantics, whereas the actual process is likely more complex. We present a generative model that maps acoustics in a hierarchical manner to increasingly higher-level semantics. Our model has 2 layers with the first being generic sound units with no clear semantic associations, while the second layer attempts to find patterns over the generic sound units. We evaluate our model on a large-scale retrieval task from TRECVID 2011, and report significant improvements over standard baselines.
****

The Time-Marginalized Coalescent Prior for Hierarchical Clustering
Levi Boyles, Max Welling
We introduce a new prior for use in Nonparametric Bayesian Hierarchical Clustering. The prior is constructed by marginalizing out the time information of Kingman’s coalescent, providing a prior over tree structures which we call the Time-Marginalized Coalescent (TMC). This allows for models which factorize the tree structure and times, providing two benefits: more flexible priors may be constructe

d and more efficient Gibbs type inference can be used. We demonstrate this on an example model for density estimation and show the TMC achieves competitive experimental results.

****

Nonparametric Max-Margin Matrix Factorization for Collaborative Prediction

Minjie Xu, Jun Zhu, Bo Zhang

We present a probabilistic formulation of max-margin matrix factorization and build accordingly a nonparametric Bayesian model which automatically resolves the unknown number of latent factors. Our work demonstrates a successful example that integrates Bayesian nonparametrics and max-margin learning, which are conventionally two separate paradigms and enjoy complementary advantages. We develop an efcient variational algorithm for posterior inference, and our extensive empirical studies on large-scale MovieLens and EachMovie data sets appear to justify the aforementioned dual advantages.

****

Exponential Concentration for Mutual Information Estimation with Application to Forests

Han Liu, Larry Wasserman, John Lafferty

We prove a new exponential concentration inequality for a plug-in estimator of the Shannon mutual information. Previous results on mutual information estimation only bounded expected error. The advantage of having the exponential inequality is that, combined with the union bound, we can guarantee accurate estimators of the mutual information for many pairs of random variables simultaneously. As an application, we show how to use such a result to optimally estimate the density function and graph of a distribution which is Markov to a forest graph.

****

Slice Normalized Dynamic Markov Logic Networks

Tivadar Papai, Henry Kautz, Daniel Stefankovic

Markov logic is a widely used tool in statistical relational learning, which uses a weighted first-order logic knowledge base to specify a Markov random field (MRF) or a conditional random field (CRF). In many applications, a Markov logic network (MLN) is trained in one domain, but used in a different one. This paper focuses on dynamic Markov logic networks, where the domain of time points typically varies between training and testing. It has been previously pointed out that the marginal probabilities of truth assignments to ground atoms can change if one extends or reduces the domains of predicates in an MLN. We show that in addition to this problem, the standard way of unrolling a Markov logic theory into a MRF may result in time-inhomogeneity of the underlying Markov chain. Furthermore, even if these representational problems are not significant for a given domain, we show that the more practical problem of generating samples in a sequential conditional random field for the next slice relying on the samples from the previous slice has high computational cost in the general case, due to the need to estimate a normalization factor for each sample. We propose a new discriminative model, slice normalized dynamic Markov logic networks (SN-DMLN), that suffers from none of these issues. It supports efficient online inference, and can directly model influences between variables within a time slice that do not have a causal direction, in contrast with fully directed models (e.g., DBNs). Experimental results show an improvement in accuracy over previous approaches to online inference in dynamic Markov logic networks.

****

Continuous Relaxations for Discrete Hamiltonian Monte Carlo

Yichuan Zhang, Zoubin Ghahramani, Amos J. Storkey, Charles Sutton

Continuous relaxations play an important role in discrete optimization, but have not seen much use in approximate probabilistic inference. Here we show that a general form of the Gaussian Integral Trick makes it possible to transform a wide class of discrete variable undirected models into fully continuous systems. The continuous representation allows the use of gradient-based Hamiltonian Monte Carlo for inference, results in new ways of estimating normalization constants (partition functions), and in general opens up a number of new avenues for inference in difficult discrete systems. We demonstrate some of these continuous relaxa

tion inference algorithms on a number of illustrative problems.
****

Learning with Target Prior
Zuoguan Wang, Siwei Lyu, Gerwin Schalk, Qiang Ji
Requests for name changes in the electronic proceedings will be accepted with 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.
****

Generalization Bounds for Domain Adaptation
Chao Zhang, Lei Zhang, Jieping Ye
In this paper, we provide a new framework to study the generalization bound of the learning process for domain adaptation. Without loss of generality, we consider two kinds of representative domain adaptation settings: one is domain adaptation with multiple sources and the other is domain adaptation combining source and target data. In particular, we introduce two quantities that capture the inherent characteristics of domains. For either kind of domain adaptation, based on the two quantities, we then develop the specific Hoeffding-type deviation inequality and symmetrization inequality to achieve the corresponding generalization bound based on the uniform entropy number. By using the resultant generalization bound, we analyze the asymptotic convergence and the rate of convergence of the learning process for such kind of domain adaptation. Meanwhile, we discuss the factors that affect the asymptotic behavior of the learning process. The numerical experiments support our results.
****

Multiplicative Forests for Continuous-Time Processes
Jeremy Weiss, Sriraam Natarajan, David Page
Learning temporal dependencies between variables over continuous time is an important and challenging task. Continuous-time Bayesian networks effectively model such processes but are limited by the number of conditional intensity matrices, which grows exponentially in the number of parents per variable. We develop a partition-based representation using regression trees and forests whose parameter spaces grow linearly in the number of node splits. Using a multiplicative assumption we show how to update the forest likelihood in closed form, producing efficient model updates. Our results show multiplicative forests can be learned from few temporal trajectories with large gains in performance and scalability.
****

Variational Inference for Crowdsourcing
Qiang Liu, Jian Peng, Alexander T. Ihler
Crowdsourcing has become a popular paradigm for labeling large datasets. However, it has given rise to the computational task of aggregating the crowdsourced labels provided by a collection of unreliable annotators. We approach this problem by transforming it into a standard inference problem in graphical models, and applying approximate variational methods, including belief propagation (BP) and mean field (MF). We show that our BP algorithm generalizes both majority voting and a recent algorithm by Karger et al, while our MF method is closely related to a commonly used EM algorithm. In both cases, we find that the performance of the algorithms critically depends on the choice of a prior distribution on the workers' reliability; by choosing the prior properly, both BP and MF (and EM) perform surprisingly well on both simulated and real-world datasets, competitive with state-of-the-art algorithms based on more complicated modeling assumptions.
****

Why MCA? Nonlinear sparse coding with spike-and-slab prior for neurally plausible image encoding
Philip Sterne, Joerg Bornschein, Abdul-saboor Sheikh, Jörg Lücke, Jacquelyn Shelton
Modelling natural images with sparse coding (SC) has faced two main challenges: ■exibly representing varying pixel intensities and realistically representing low- level image components. This paper proposes a novel multiple-cause generative model of low-level image statistics that generalizes the standard SC model in t

wo crucial points: (1) it uses a spike-and-slab prior distribution for a more re alistic representation of component absence/intensity, and (2) the model uses th e highly nonlinear combination rule of maximal causes analysis (MCA) instead of a lin- ear combination. The major challenge is parameter optimization because a model with either (1) or (2) results in strongly multimodal posteriors. We show for the ■rst time that a model combining both improvements can be trained ef■cie ntly while retaining the rich structure of the posteriors. We design an exact pi ece- wise Gibbs sampling method and combine this with a variational method based on preselection of latent dimensions. This combined training scheme tackles bot h analytical and computational intractability and enables application of the mod el to a large number of observed and hidden dimensions. Applying the model to im age patches we study the optimal encoding of images by simple cells in V1 and co mpare the model’s predictions with in vivo neural recordings. In contrast to sta ndard SC, we ■nd that the optimal prior favors asymmetric and bimodal ac- tivity of simple cells. Testing our model for consistency we ■nd that the average post erior is approximately equal to the prior. Furthermore, we ■nd that the model pr edicts a high percentage of globular receptive ■elds alongside Gabor-like ■elds. Similarly high percentages are observed in vivo. Our results thus argue in favo r of improvements of the standard sparse coding model for simple cells by using ■exible priors and nonlinear combinations.

****

Tractable Objectives for Robust Policy Optimization

Katherine Chen, Michael Bowling

Robust policy optimization acknowledges that risk-aversion plays a vital role in real-world decision-making. When faced with uncertainty about the effects of ac tions, the policy that maximizes expected utility over the unknown parameters of the system may also carry with it a risk of intolerably poor performance. One might prefer to accept lower utility in expectation in order to avoid, or reduce the likelihood of, unacceptable levels of utility under harmful parameter reali zations. In this paper, we take a Bayesian approach to parameter uncertainty, but unlike other methods avoid making any distributional assumptions about the f orm of this uncertainty. Instead we focus on identifying optimization objective s for which solutions can be efficiently approximated. We introduce percentile measures: a very general class of objectives for robust policy optimization, whi ch encompasses most existing approaches, including ones known to be intractable. We then introduce a broad subclass of this family for which robust policies can be approximated efficiently. Finally, we frame these objectives in the context of a two-player, zero-sum, extensive-form game and employ a no-regret algorithm to approximate an optimal policy, with computation only polynomial in the number of states and actions of the MDP.

****

Multiple Choice Learning: Learning to Produce Multiple Structured Outputs

Abner Guzmán-rivera, Dhruv Batra, Pushmeet Kohli

The paper addresses the problem of generating multiple hypotheses for prediction tasks that involve interaction with users or successive components in a cascade . Given a set of multiple hypotheses, such components/users have the ability to automatically rank the results and thus retrieve the best one. The standard appr oach for handling this scenario is to learn a single model and then produce M-be st Maximum a Posteriori (MAP) hypotheses from this model. In contrast, we formul ate this multiple {\em choice} learning task as a multiple-output structured-out put prediction problem with a loss function that captures the natural setup of t he problem. We present a max-margin formulation that minimizes an upper-bound o n this loss-function. Experimental results on the problems of image co-segmentat ion and protein side-chain prediction show that our method outperforms conventio nal approaches used for this scenario and leads to substantial improvements in prediction accuracy.

****

Robustness and risk-sensitivity in Markov decision processes

Takayuki Osogami

We uncover relations between robust MDPs and risk-sensitive MDPs. The objective

of a robust MDP is to minimize a function, such as the expectation of cumulativ
e cost, for the worst case when the parameters have uncertainties. The objectiv
e of a risk-sensitive MDP is to minimize a risk measure of the cumulative cost w
hen the parameters are known. We show that a risk-sensitive MDP of minimizing t
he expected exponential utility is equivalent to a robust MDP of minimizing the
worst-case expectation with a penalty for the deviation of the uncertain paramet
ers from their nominal values, which is measured with the Kullback-Leibler diver
gence. We also show that a risk-sensitive MDP of minimizing an iterated risk me
asure that is composed of certain coherent risk measures is equivalent to a robu
st MDP of minimizing the worst-case expectation when the possible deviations of
uncertain parameters from their nominal values are characterized with a concave
function.
****

Communication/Computation Tradeoffs in Consensus-Based Distributed Optimization
Konstantinos Tsianos, Sean Lawlor, Michael Rabbat
Requests for name changes in the electronic proceedings will be accepted with 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 Polylog Pivot Steps Simplex Algorithm for Classification
Elad Hazan, Zohar Karnin
We present a simplex algorithm for linear programming in a linear classification
formulation. The paramount complexity parameter in linear classification proble
ms is called the margin. We prove that for margin values of practical interest o
ur simplex variant performs a polylogarithmic number of pivot steps in the worst
case, and its overall running time is near linear. This is in contrast to gener
al linear programming, for which no sub-polynomial pivot rule is known.
****

Fiedler Random Fields: A Large-Scale Spectral Approach to Statistical Network Mo
deling
Antonino Freno, Mikaela Keller, Marc Tommasi
Statistical models for networks have been typically committed to strong prior as
sumptions concerning the form of the modeled distributions. Moreover, the vast m
ajority of currently available models are explicitly designed for capturing some
specific graph properties (such as power-law degree distributions), which makes
them unsuitable for application to domains where the behavior of the target qua
ntities is not known a priori. The key contribution of this paper is twofold. Fi
rst, we introduce the Fiedler delta statistic, based on the Laplacian spectrum o
f graphs, which allows to dispense with any parametric assumption concerning the
modeled network properties. Second, we use the defined statistic to develop the
Fiedler random field model, which allows for efficient estimation of edge distr
ibutions over large-scale random networks. After analyzing the dependence struct
ure involved in Fiedler random fields, we estimate them over several real-world
networks, showing that they achieve a much higher modeling accuracy than other w
ell-known statistical approaches.
****

Majorization for CRFs and Latent Likelihoods
Tony Jebara, Anna Choromanska
The partition function plays a key role in probabilistic modeling including cond
itional random fields, graphical models, and maximum likelihood estimation. To o
ptimize partition functions, this article introduces a quadratic variational upp
er bound. This inequality facilitates majorization methods: optimization of comp
licated functions through the iterative solution of simpler sub-problems. Such b
ounds remain efficient to compute even when the partition function involves a gr
aphical model (with small tree-width) or in latent likelihood settings. For larg
e-scale problems, low-rank versions of the bound are provided and outperform LBF
GS as well as first-order methods. Several learning applications are shown and r
educe to fast and convergent update rules. Experimental results show advantages
over state-of-the-art optimization methods.

****
Feature-aware Label Space Dimension Reduction for Multi-label Classification
Yao-nan Chen, Hsuan-tien Lin
Label space dimension reduction (LSDR) is an efficient and effective paradigm for multi-label classification with many classes. Existing approaches to LSDR, such as compressive sensing and principal label space transformation, exploit only the label part of the dataset, but not the feature part. In this paper, we propose a novel approach to LSDR that considers both the label and the feature parts. The approach, called conditional principal label space transformation, is based on minimizing an upper bound of the popular Hamming loss. The minimization step of the approach can be carried out efficiently by a simple use of singular value decomposition. In addition, the approach can be extended to a kernelized version that allows the use of sophisticated feature combinations to assist LSDR. The experimental results verify that the proposed approach is more effective than existing ones to LSDR across many real-world datasets.
****
Semantic Kernel Forests from Multiple Taxonomies
Sung Hwang, Kristen Grauman, Fei Sha
When learning features for complex visual recognition problems, labeled image exemplars alone can be insufficient. While an \emph{object taxonomy} specifying the categories' semantic relationships could bolster the learning process, not all relationships are relevant to a given visual classification task, nor does a single taxonomy capture all ties that \emph{are} relevant. In light of these issues, we propose a discriminative feature learning approach that leverages \emph{multiple} hierarchical taxonomies representing different semantic views of the object categories (e.g., for animal classes, one taxonomy could reflect their phylogenic ties, while another could reflect their habitats). For each taxonomy, we first learn a tree of semantic kernels, where each node has a Mahalanobis kernel optimized to distinguish between the classes in its children nodes. Then, using the resulting \emph{semantic kernel forest}, we learn class-specific kernel combinations to select only those relationships relevant to recognize each object class. To learn the weights, we introduce a novel hierarchical regularization term that further exploits the taxonomies' structure. We demonstrate our method on challenging object recognition datasets, and show that interleaving multiple taxonomic views yields significant accuracy improvements.
****
Spectral learning of linear dynamics from generalised-linear observations with application to neural population data
Lars Buesing, Jakob H. Macke, Maneesh Sahani
Latent linear dynamical systems with generalised-linear observation models arise in a variety of applications, for example when modelling the spiking activity of populations of neurons. Here, we show how spectral learning methods for linear systems with Gaussian observations (usually called subspace identification in this context) can be extended to estimate the parameters of dynamical system models observed through non-Gaussian noise models. We use this approach to obtain estimates of parameters for a dynamical model of neural population data, where the observed spike-counts are Poisson-distributed with log-rates determined by the latent dynamical process, possibly driven by external inputs. We show that the extended system identification algorithm is consistent and accurately recovers the correct parameters on large simulated data sets with much smaller computational cost than approximate expectation-maximisation (EM) due to the non-iterative nature of subspace identification. Even on smaller data sets, it provides an effective initialization for EM, leading to more robust performance and faster convergence. These benefits are shown to extend to real neural data.
****
Assessing Blinding in Clinical Trials
Ognjen Arandjelovic
The interaction between the patient's expected outcome of an intervention and the inherent effects of that intervention can have extraordinary effects. Thus in clinical trials an effort is made to conceal the nature of the administered inte

rvention from the participants in the trial i.e. to blind it. Yet, in practice p
erfect blinding is impossible to ensure or even verify. The current standard is
follow up the trial with an auxiliary questionnaire, which allows trial particip
ants to express their belief concerning the assigned intervention and which is u
sed to compute a measure of the extent of blinding in the trial. If the estimate
d extent of blinding exceeds a threshold the trial is deemed sufficiently blinde
d; otherwise, the trial is deemed to have failed. In this paper we make several
important contributions. Firstly, we identify a series of fundamental problems o
f the aforesaid practice and discuss them in context of the most commonly used b
linding measures. Secondly, motivated by the highlighted problems, we formulate
a novel method for handling imperfectly blinded trials. We too adopt a post-tria
l feedback questionnaire but interpret the collected data using an original appr
oach, fundamentally different from those previously proposed. Unlike previous ap
proaches, ours is void of any ad hoc free parameters, is robust to small changes
in auxiliary data and is not predicated on any strong assumptions used to inter
pret participants' feedback.
****

Scalable Inference of Overlapping Communities
Prem K. Gopalan, Sean Gerrish, Michael Freedman, David Blei, David Mimno
We develop a scalable algorithm for posterior inference of overlapping communiti
es in large networks. Our algorithm is based on stochastic variational inferenc
e in the mixed-membership stochastic blockmodel. It naturally interleaves subsam
pling the network with estimating its community structure. We apply our algorit
hm on ten large, real-world networks with up to 60,000 nodes. It converges sever
al orders of magnitude faster than the state-of-the-art algorithm for MMSB, find
s hundreds of communities in large real-world networks, and detects the true com
munities in 280 benchmark networks with equal or better accuracy compared to oth
er scalable algorithms.
****

Learning Networks of Heterogeneous Influence
Nan Du, Le Song, Ming Yuan, Alex Smola
Information, disease, and influence diffuse over networks of entities in both na
tural systems and human society. Analyzing these transmission networks plays an
important role in understanding the diffusion processes and predicting events in
the future. However, the underlying transmission networks are often hidden and
incomplete, and we observe only the time stamps when cascades of events happen.
In this paper, we attempt to address the challenging problem of uncovering the h
idden network only from the cascades. The structure discovery problem is compli
cated by the fact that the influence among different entities in a network are h
eterogeneous, which can not be described by a simple parametric model. Therefore
, we propose a kernel-based method which can capture a diverse range of differen
t types of influence without any prior assumption. In both synthetic and real ca
scade data, we show that our model can better recover the underlying diffusion n
etwork and drastically improve the estimation of the influence functions between
networked entities.
****

Learning to Align from Scratch
Gary Huang, Marwan Mattar, Honglak Lee, Erik Learned-miller
Unsupervised joint alignment of images has been demonstrated to improve perfor
mance on recognition tasks such as face verification. Such alignment reduces u
ndesired variability due to factors such as pose, while only requiring weak su
pervision in the form of poorly aligned examples. However, prior work on unsu
pervised alignment of complex, real world images has required the careful sele
ction of feature representation based on hand-crafted image descriptors, in
order to achieve an appropriate, smooth optimization landscape. In this paper
, we instead propose a novel combination of unsupervised joint alignment with
unsupervised feature learning. Specifically, we incorporate deep learning into
the {\em congealing} alignment framework. Through deep learning, we obtain f
eatures that can represent the image at differing resolutions based on network
depth, and that are tuned to the statistics of the specific data being alig

ned. In addition, we modify the learning algorithm for the restricted Boltzma nn machine by incorporating a group sparsity penalty, leading to a topographic organization on the learned filters and improving subsequent alignment result s. We apply our method to the Labeled Faces in the Wild database (LFW). Usi ng the aligned images produced by our proposed unsupervised algorithm, we achi eve a significantly higher accuracy in face verification than obtained using t he original face images, prior work in unsupervised alignment, and prior work in supervised alignment. We also match the accuracy for the best available, b ut unpublished method.

****

Bayesian Warped Gaussian Processes

Miguel Lázaro-Gredilla

Warped Gaussian processes (WGP) [1] model output observations in regression task s as a parametric nonlinear transformation of a Gaussian process (GP). The use o f this nonlinear transformation, which is included as part of the probabilistic model, was shown to enhance performance by providing a better prior model on sev eral data sets. In order to learn its parameters, maximum likelihood was used. I n this work we show that it is possible to use a non-parametric nonlinear transf ormation in WGP and variationally integrate it out. The resulting Bayesian WGP i s then able to work in scenarios in which the maximum likelihood WGP failed: Low data regime, data with censored values, classification, etc. We demonstrate the superior performance of Bayesian warped GPs on several real data sets.

****

Affine Independent Variational Inference

Edward Challis, David Barber

We present a method for approximate inference for a broad class of non-conjugate probabilistic models. In particular, for the family of generalized linear model target densities we describe a rich class of variational approximating densitie s which can be best fit to the target by minimizing the Kullback-Leibler diverge nce. Our approach is based on using the Fourier representation which we show re sults in efficient and scalable inference.

****

Submodular-Bregman and the Lovász-Bregman Divergences with Applications

Rishabh Iyer, Jeff A. Bilmes

We introduce a class of discrete divergences on sets (equivalently binary vector s) that we call the submodular-Bregman divergences. We consider two kinds, de■ne d either from tight modular upper or tight modular lower bounds of a submodular function. We show that the properties of these divergences are analogous to the (standard continuous) Bregman divergence. We demonstrate how they generalize man y useful divergences, including the weighted Hamming distance, squared weighted Hamming, weighted precision, recall, conditional mutual information, and a gener alized KL-divergence on sets. We also show that the generalized Bregman divergen ce on the Lov´asz extension of a submodular function, which we call the Lov´asz-Bregman divergence, is a continuous extension of a submodular Bregman divergence . We point out a number of applications, and in particular show that a proximal algorithm de■ned through the submodular Bregman divergence pro- vides a framewor k for many mirror-descent style algorithms related to submodular function optimi zation. We also show that a generalization of the k-means algorithm using the Lo v´asz Bregman divergence is natural in clustering scenarios where ordering is im portant. A unique property of this algorithm is that computing the mean ordering is extremely ef■cient unlike other order based distance measures.

****

Optimal kernel choice for large-scale two-sample tests

Arthur Gretton, Dino Sejdinovic, Heiko Strathmann, Sivaraman Balakrishnan, Massi miliano Pontil, Kenji Fukumizu, Bharath K. Sriperumbudur

Requests for name changes in the electronic proceedings will be accepted with 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.

****

Entangled Monte Carlo
Seong-hwan Jun, Liangliang Wang, Alexandre Bouchard-côté
We propose a novel method for scalable parallelization of SMC algorithms, Entang
led Monte Carlo simulation (EMC). EMC avoids the transmission of particles betw
een nodes, and instead reconstructs them from the particle genealogy. In partic
ular, we show that we can reduce the communication to the particle weights for e
ach machine while efficiently maintaining implicit global coherence of the paral
lel simulation. We explain methods to efficiently maintain a genealogy of partic
les from which any particle can be reconstructed. We demonstrate using examples
from Bayesian phylogenetic that the computational gain from parallelization usin
g EMC significantly outweighs the cost of particle reconstruction. The timing ex
periments show that reconstruction of particles is indeed much more efficient as
compared to transmission of particles.
****
Minimax Multi-Task Learning and a Generalized Loss-Compositional Paradigm for MT
L
Nishant Mehta, Dongryeol Lee, Alexander Gray
Since its inception, the modus operandi of multi-task learning (MTL) has been to
minimize the task-wise mean of the empirical risks. We introduce a generalize
d loss-compositional paradigm for MTL that includes a spectrum of formulations a
s a subfamily. One endpoint of this spectrum is minimax MTL: a new MTL formulati
on that minimizes the maximum of the tasks' empirical risks. Via a certain relax
ation of minimax MTL, we obtain a continuum of MTL formulations spanning minimax
MTL and classical MTL. The full paradigm itself is loss-compositional, operatin
g on the vector of empirical risks. It incorporates minimax MTL, its relaxations
, and many new MTL formulations as special cases. We show theoretically that min
imax MTL tends to avoid worst case outcomes on newly drawn test tasks in the lea
rning to learn (LTL) test setting. The results of several MTL formulations on sy
nthetic and real problems in the MTL and LTL test settings are encouraging.
****
Semi-Crowdsourced Clustering: Generalizing Crowd Labeling by Robust Distance Met
ric Learning
Jinfeng Yi, Rong Jin, Shaili Jain, Tianbao Yang, Anil Jain
One of the main challenges in data clustering is to define an appropriate simila
rity measure between two objects. Crowdclustering addresses this challenge by de
fining the pairwise similarity based on the manual annotations obtained through
crowdsourcing. Despite its encouraging results, a key limitation of crowdcluster
ing is that it can only cluster objects when their manual annotations are availa
ble. To address this limitation, we propose a new approach for clustering, calle
d \textit{semi-crowdsourced clustering} that effectively combines the low-level
features of objects with the manual annotations of a subset of the objects obtai
ned via crowdsourcing. The key idea is to learn an appropriate similarity measur
e, based on the low-level features of objects, from the manual annotations of on
ly a small portion of the data to be clustered. One difficulty in learning the p
airwise similarity measure is that there is a significant amount of noise and in
ter-worker variations in the manual annotations obtained via crowdsourcing. We a
ddress this difficulty by developing a metric learning algorithm based on the ma
trix completion method. Our empirical study with two real-world image data sets
shows that the proposed algorithm outperforms state-of-the-art distance metric l
earning algorithms in both clustering accuracy and computational efficiency.
****
Online L1-Dictionary Learning with Application to Novel Document Detection
Shiva Kasiviswanathan, Huahua Wang, Arindam Banerjee, Prem Melville
Given their pervasive use, social media, such as Twitter, have become a leading
source of breaking news. A key task in the automated identification of such news
is the detection of novel documents from a voluminous stream of text documents
in a scalable manner. Motivated by this challenge, we introduce the problem of o
nline L1-dictionary learning where unlike traditional dictionary learning, which
uses squared loss, the L1-penalty is used for measuring the reconstruction erro
r. We present an efficient online algorithm for this problem based on alternatin

g directions method of multipliers, and establish a sublinear regret bound for t
his algorithm. Empirical results on news-stream and Twitter data, shows that thi
s online L1-dictionary learning algorithm for novel document detection gives mor
e than an order of magnitude speedup over the previously known batch algorithm,
without any significant loss in quality of results. Our algorithm for online L1-
dictionary learning could be of independent interest.
****

Learning curves for multi-task Gaussian process regression
Peter Sollich, Simon Ashton
Requests for name changes in the electronic proceedings will be accepted with 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.
****

Transelliptical Component Analysis
Fang Han, Han Liu
We propose a high dimensional semiparametric scale-invariant principle compo- ne
nt analysis, named TCA, by utilize the natural connection between the ellipti- c
al distribution family and the principal component analysis. Elliptical distribu
- tion family includes many well-known multivariate distributions like multivari
- ate Gaussian, t and logistic and it is extended to the meta-elliptical by Fang
 et.al (2002) using the copula techniques. In this paper we extend the meta-elli
ptical distribution family to a even larger family, called transelliptical. We p
rove that
****

Learning the Dependency Structure of Latent Factors
Yunlong He, Yanjun Qi, Koray Kavukcuoglu, Haesun Park
In this paper, we study latent factor models with the dependency structure in th
e latent space. We propose a general learning framework which induces sparsity
on the undirected graphical model imposed on the vector of latent factors. A nov
el latent factor model SLFA is then proposed as a matrix factorization problem w
ith a special regularization term that encourages collaborative reconstruction.
The main benefit (novelty) of the model is that we can simultaneously learn the
lower-dimensional representation for data and model the pairwise relationships b
etween latent factors explicitly. An on-line learning algorithm is devised to ma
ke the model feasible for large-scale learning problems. Experimental results on
 two synthetic data and two real-world data sets demonstrate that pairwise relat
ionships and latent factors learned by our model provide a more structured way o
f exploring high-dimensional data, and the learned representations achieve the
state-of-the-art classification performance.
****

Random function priors for exchangeable arrays with applications to graphs and r
elational data
James Lloyd, Peter Orbanz, Zoubin Ghahramani, Daniel M. Roy
A fundamental problem in the analysis of structured relational data like graphs,
 networks, databases, and matrices is to extract a summary of the common struc-
ture underlying relations between individual entities. Relational data are typic
ally encoded in the form of arrays; invariance to the ordering of rows and colum
ns corresponds to exchangeable arrays. Results in probability theory due to Aldo
us, Hoover and Kallenberg show that exchangeable arrays can be represented in te
rms of a random measurable function which constitutes the natural model paramete
r in a Bayesian model. We obtain a ■exible yet simple Bayesian nonparametric mod
el by placing a Gaussian process prior on the parameter function. Ef■cient infer
ence utilises elliptical slice sampling combined with a random sparse approximat
ion to the Gaussian process. We demonstrate applications of the model to network
 data and clarify its relation to models in the literature, several of which eme
rge as special cases.
****

Bayesian Pedigree Analysis using Measure Factorization
Bonnie Kirkpatrick, Alexandre Bouchard-côté

Pedigrees, or family trees, are directed graphs used to identify sites of the ge
nome that are correlated with the presence or absence of a disease. With the ad
vent of genotyping and sequencing technologies, there has been an explosion in t
he amount of data available, both in the number of individuals and in the number
 of sites. Some pedigrees number in the thousands of individuals. Meanwhile, a
nalysis methods have remained limited to pedigrees of <100 individuals which lim
its analyses to many small independent pedigrees. Disease models, such those us
ed for the linkage analysis log-odds (LOD) estimator, have similarly been limite
d. This is because linkage anlysis was originally designed with a different tas
k in mind, that of ordering the sites in the genome, before there were technolog
ies that could reveal the order. LODs are difficult to interpret and nontrivial
 to extend to consider interactions among sites. These developments and difficu
lties call for the creation of modern methods of pedigree analysis. Drawing fro
m recent advances in graphical model inference and transducer theory, we introdu
ce a simple yet powerful formalism for expressing genetic disease models. We s
how that these disease models can be turned into accurate and efficient estimato
rs. The technique we use for constructing the variational approximation has pot
ential applications to inference in other large-scale graphical models. This me
thod allows inference on larger pedigrees than previously analyzed in the litera
ture, which improves disease site prediction.
****

Density Propagation and Improved Bounds on the Partition Function
Stefano Ermon, Ashish Sabharwal, Bart Selman, Carla P. Gomes
Given a probabilistic graphical model, its density of states is a function that,
 for any likelihood value, gives the number of configurations with that probabil
ity. We introduce a novel message-passing algorithm called Density Propagation (
DP) for estimating this function. We show that DP is exact for tree-structured g
raphical models and is, in general, a strict generalization of both sum-product
and max-product algorithms. Further, we use density of states and tree decomposi
tion to introduce a new family of upper and lower bounds on the partition functi
on. For any tree decompostion, the new upper bound based on finer-grained densit
y of state information is provably at least as tight as previously known bounds
based on convexity of the log-partition function, and strictly stronger if a gen
eral condition holds. We conclude with empirical evidence of improvement over co
nvex relaxations and mean-field based bounds.
****

A quasi-Newton proximal splitting method
Stephen Becker, Jalal Fadili
We describe efficient implementations of the proximity calculation for a useful
class of functions; the implementations exploit the piece-wise linear nature of
the dual problem. The second part of the paper applies the previous result to ac
celeration of convex minimization problems, and leads to an elegant quasi-Newton
 method. The optimization method compares favorably against state-of-the-art alt
ernatives. The algorithm has extensive applications including signal processing,
 sparse regression and recovery, and machine learning and classification.
****

Waveform Driven Plasticity in BiFeO3 Memristive Devices: Model and Implementatio
n
Christian Mayr, Paul Stärke, Johannes Partzsch, Love Cederstroem, Rene Schüffny,
 Yao Shuai, Nan Du, Heidemarie Schmidt
Requests for name changes in the electronic proceedings will be accepted with 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.
****

Multilabel Classification using Bayesian Compressed Sensing
Ashish Kapoor, Raajay Viswanathan, Prateek Jain
In this paper, we present a Bayesian framework for multilabel classification usi
ng compressed sensing. The key idea in compressed sensing for multilabel classif
ication is to first project the label vector to a lower dimensional space using

a random transformation and then learn regression functions over these projectio ns. Our approach considers both of these components in a single probabilistic mo del, thereby jointly optimizing over compression as well as learning tasks. We t hen derive an efficient variational inference scheme that provides joint posteri or distribution over all the unobserved labels. The two key benefits of the mode l are that a) it can naturally handle datasets that have missing labels and b) it can also measure uncertainty in prediction. The uncertainty estimate provided by the model naturally allows for active learning paradigms where an oracle pro vides information about labels that promise to be maximally informative for the prediction task. Our experiments show significant boost over prior methods in te rms of prediction performance over benchmark datasets, both in the fully labeled and the missing labels case. Finally, we also highlight various useful active learning scenarios that are enabled by the probabilistic model.
****

Online Sum-Product Computation Over Trees
Mark Herbster, Stephen Pasteris, Fabio Vitale
Requests for name changes in the electronic proceedings will be accepted with 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.
****

Risk Aversion in Markov Decision Processes via Near Optimal Chernoff Bounds
Teodor Moldovan, Pieter Abbeel
The expected return is a widely used objective in decision making under uncer- t ainty. Many algorithms, such as value iteration, have been proposed to optimize it. In risk-aware settings, however, the expected return is often not an appropr iate objective to optimize. We propose a new optimization objective for risk-awa re planning and show that it has desirable theoretical properties. We also draw con- nections to previously proposed objectives for risk-aware planing: minmax, ex- ponential utility, percentile and mean minus variance. Our method applies to an extended class of Markov decision processes: we allow costs to be stochastic as long as they are bounded. Additionally, we present an ef■cient algorithm for op- timizing the proposed objective. Synthetic and real-world experiments illus trate the effectiveness of our method, at scale.
****

Calibrated Elastic Regularization in Matrix Completion
Tingni Sun, Cun-hui Zhang
This paper concerns the problem of matrix completion, which is to estimate a mat rix from observations in a small subset of indices. We propose a calibrated spec trum elastic net method with a sum of the nuclear and Frobenius penalties and de velop an iterative algorithm to solve the convex minimization problem. The itera tive algorithm alternates between imputing the missing entries in the incomplete matrix by the current guess and estimating the matrix by a scaled soft-threshol ding singular value decomposition of the imputed matrix until the resulting matr ix converges. A calibration step follows to correct the bias caused by the Frobe nius penalty. Under proper coherence conditions and for suitable penalties level s, we prove that the proposed estimator achieves an error bound of nearly optima l order and in proportion to the noise level. This provides a unified analysis o f the noisy and noiseless matrix completion problems. Simulation results are pre sented to compare our proposal with previous ones.
****

Expectation Propagation in Gaussian Process Dynamical Systems
Marc Deisenroth, Shakir Mohamed
Rich and complex time-series data, such as those generated from engineering sys- tems, financial markets, videos or neural recordings are now a common feature o f modern data analysis. Explaining the phenomena underlying these diverse data s ets requires flexible and accurate models. In this paper, we promote Gaussian pr ocess dynamical systems as a rich model class appropriate for such analysis. In particular, we present a message passing algorithm for approximate inference in GPDSs based on expectation propagation. By phrasing inference as a general mes-

sage passing problem, we iterate forward-backward smoothing. We obtain more accu
rate posterior distributions over latent structures, resulting in improved pre-
dictive performance compared to state-of-the-art GPDS smoothers, which are spe-
cial cases of our general iterative message passing algorithm. Hence, we provide
 a unifying approach within which to contextualize message passing in GPDSs.
****

Finite Sample Convergence Rates of Zero-Order Stochastic Optimization Methods
Andre Wibisono, Martin J. Wainwright, Michael Jordan, John C. Duchi
Requests for name changes in the electronic proceedings will be accepted with 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.
****

Query Complexity of Derivative-Free Optimization
Kevin G. Jamieson, Robert Nowak, Ben Recht
Derivative Free Optimization (DFO) is attractive when the objective function's d
erivatives are not available and evaluations are costly. Moreover, if the func
tion evaluations are noisy, then approximating gradients by finite differences i
s difficult. This paper gives quantitative lower bounds on the performance of D
FO with noisy function evaluations, exposing a fundamental and unavoidable gap b
etween optimization performance based on noisy evaluations versus noisy gradient
s. This challenges the conventional wisdom that the method of finite differences
 is comparable to a stochastic gradient. However, there are situations in which
 DFO is unavoidable, and for such situations we propose a new DFO algorithm that
 is proved to be near optimal for the class of strongly convex objective functio
ns. A distinctive feature of the algorithm is that it only uses Boolean-valued
function comparisons, rather than evaluations. This makes the algorithm useful
in an even wider range of applications, including optimization based on paired c
omparisons from human subjects, for example. Remarkably, we show that regardles
s of whether DFO is based on noisy function evaluations or Boolean-valued functi
on comparisons, the convergence rate is the same.
****

Communication-Efficient Algorithms for Statistical Optimization
Yuchen Zhang, Martin J. Wainwright, John C. Duchi
Requests for name changes in the electronic proceedings will be accepted with 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.
****

Selecting Diverse Features via Spectral Regularization
Abhimanyu Das, Anirban Dasgupta, Ravi Kumar
Requests for name changes in the electronic proceedings will be accepted with no
 questions asked. However name changes may cause bibliographic tracking issues.
  Authors are asked to consider this carefully and discuss it with their co-auth
ors prior to requesting a name change in the electronic proceedings.
****

Fast Bayesian Inference for Non-Conjugate Gaussian Process Regression
Emtiyaz Khan, Shakir Mohamed, Kevin P. Murphy
We present a new variational inference algorithm for Gaussian processes with non
-conjugate likelihood functions. This includes binary and multi-class classifica
tion, as well as ordinal regression. Our method constructs a convex lower bound,
 which can be optimized by using an efficient fixed point update method. We then
 show empirically that our new approach is much faster than existing methods wit
hout any degradation in performance.
****

Natural Images, Gaussian Mixtures and Dead Leaves
Daniel Zoran, Yair Weiss
Simple Gaussian Mixture Models (GMMs) learned from pixels of natural image patch
es have been recently shown to be surprisingly strong performers in modeling the
 statistics of natural images. Here we provide an in depth analysis of this simp

le yet rich model. We show that such a GMM model is able to compete with even th e most successful models of natural images in log likelihood scores, denoising p erformance and sample quality. We provide an analysis of what such a model learn s from natural images as a function of number of mixture components --- includin g covariance structure, contrast variation and intricate structures such as text ures, boundaries and more. Finally, we show that the salient properties of the G MM learned from natural images can be derived from a simplified Dead Leaves mode l which explicitly models occlusion, explaining its surprising success relative to other models.

****

Memorability of Image Regions

Aditya Khosla, Jianxiong Xiao, Antonio Torralba, Aude Oliva

While long term human visual memory can store a remarkable amount of visual info rmation, it tends to degrade over time. Recent works have shown that image memor ability is an intrinsic property of an image that can be reliably estimated usin g state-of-the-art image features and machine learning algorithms. However, the class of features and image information that is forgotten has not been explored yet. In this work, we propose a probabilistic framework that models how and whic h local regions from an image may be forgotten using a data-driven approach that combines local and global images features. The model automatically discov- ers memorability maps of individual images without any human annotation. We incorpor ate multiple image region attributes in our algorithm, leading to improved memor ability prediction of images as compared to previous works.

****

Projection Retrieval for Classification

Madalina Fiterau, Artur Dubrawski

In many applications classification systems often require in the loop human inte rvention. In such cases the decision process must be transparent and comprehensi ble simultaneously requiring minimal assumptions on the underlying data distribu tion. To tackle this problem, we formulate it as an axis-alligned subspacefindin g task under the assumption that query specific information dictates the complem entary use of the subspaces. We develop a regression-based approach called RECIP that efficiently solves this problem by finding projections that minimize a non parametric conditional entropy estimator. Experiments show that the method is ac curate in identifying the informative projections of the dataset, picking the co rrect ones to classify query points and facilitates visual evaluation by users.

****

One Permutation Hashing

Ping Li, Art Owen, Cun-hui Zhang

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

****

The representer theorem for Hilbert spaces: a necessary and sufficient condition

Francesco Dinuzzo, Bernhard Schölkopf

The representer theorem is a property that lies at the foundation of regularizat ion theory and kernel methods. A class of regularization functionals is said to admit a linear representer theorem if every member of the class admits minimizer s that lie in the finite dimensional subspace spanned by the representers of the data. A recent characterization states that certain classes of regularization f unctionals with differentiable regularization term admit a linear representer th eorem for any choice of the data if and only if the regularization term is a rad ial nondecreasing function. In this paper, we extend such result by weakening th e assumptions on the regularization term. In particular, the main result of this paper implies that, for a sufficiently large family of regularization functiona ls, radial nondecreasing functions are the only lower semicontinuous regularizat ion terms that guarantee existence of a representer theorem for any choice of th e data.

****

A Geometric take on Metric Learning
Søren Hauberg, Oren Freifeld, Michael Black
Multi-metric learning techniques learn local metric tensors in different parts o f a feature space. With such an approach, even simple classifiers can be competi tive with the state-of-the-art because the distance measure locally adapts to th e structure of the data. The learned distance measure is, however, non-metric, w hich has prevented multi-metric learning from generalizing to tasks such as dime nsionality reduction and regression in a principled way. We prove that, with app ropriate changes, multi-metric learning corresponds to learning the structure of a Riemannian manifold. We then show that this structure gives us a principled w ay to perform dimensionality reduction and regression according to the learned m etrics. Algorithmically, we provide the first practical algorithm for computing geodesics according to the learned metrics, as well as algorithms for computing exponential and logarithmic maps on the Riemannian manifold. Together, these too ls let many Euclidean algorithms take advantage of multi-metric learning. We ill ustrate the approach on regression and dimensionality reduction tasks that invol ve predicting measurements of the human body from shape data.
****

Iterative Thresholding Algorithm for Sparse Inverse Covariance Estimation
Benjamin Rolfs, Bala Rajaratnam, Dominique Guillot, Ian Wong, Arian Maleki
Requests for name changes in the electronic proceedings will be accepted with 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 allocation and homogeneous partitioning for piecewise constant mean-appro ximation
Alexandra Carpentier, Odalric-ambrym Maillard
In the setting of active learning for the multi-armed bandit, where the goal of a learner is to estimate with equal precision the mean of a finite number of arm s, recent results show that it is possible to derive strategies based on finite-time confidence bounds that are competitive with the best possible strategy. We here consider an extension of this problem to the case when the arms are the cel ls of a finite partition P of a continuous sampling space X \subset \Real^d. Our goal is now to build a piecewise constant approximation of a noisy function (wh ere each piece is one region of P and P is fixed beforehand) in order to maintai n the local quadratic error of approximation on each cell equally low. Although this extension is not trivial, we show that a simple algorithm based on upper co nfidence bounds can be proved to be adaptive to the function itself in a near-op timal way, when |P| is chosen to be of minimax-optimal order on the class of \a lpha-Hölder functions.
****

Efficient coding provides a direct link between prior and likelihood in perceptu al Bayesian inference
Xue-xin Wei, Alan A. Stocker
A common challenge for Bayesian models of perception is the fact that the two fu ndamental Bayesian components, the prior distribution and the likelihood func- t ion, are formally unconstrained. Here we argue that a neural system that emulate s Bayesian inference is naturally constrained by the way it represents sensory i nfor- mation in populations of neurons. More speci■cally, we show that an ef■cie nt coding principle creates a direct link between prior and likelihood based on the underlying stimulus distribution. The resulting Bayesian estimates can show bi- ases away from the peaks of the prior distribution, a behavior seemingly at odds with the traditional view of Bayesian estimation, yet one that has been rep orted in human perception. We demonstrate that our framework correctly accounts for the repulsive biases previously reported for the perception of visual orient ation, and show that the predicted tuning characteristics of the model neurons m atch the reported orientation tuning properties of neurons in primary visual cor tex. Our results suggest that ef■cient coding is a promising hypothesis in const rain- ing Bayesian models of perceptual inference.

****
Pointwise Tracking the Optimal Regression Function
Yair Wiener, Ran El-Yaniv
Requests for name changes in the electronic proceedings will be accepted with 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.
****
Globally Convergent Dual MAP LP Relaxation Solvers using Fenchel-Young Margins
Alex Schwing, Tamir Hazan, Marc Pollefeys, Raquel Urtasun
Requests for name changes in the electronic proceedings will be accepted with 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.
****
Nonparametric Reduced Rank Regression
Rina Foygel, Michael Horrell, Mathias Drton, John Lafferty
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Density-Difference Estimation
Masashi Sugiyama, Takafumi Kanamori, Taiji Suzuki, Marthinus Plessis, Song Liu, Ichiro Takeuchi
We address the problem of estimating the difference between two probability densities. A naive approach is a two-step procedure of first estimating two densities separately and then computing their difference. However, such a two-step procedure does not necessarily work well because the first step is performed without regard to the second step and thus a small estimation error incurred in the first stage can cause a big error in the second stage. In this paper, we propose a single-shot procedure for directly estimating the density difference without separately estimating two densities. We derive a non-parametric finite-sample error bound for the proposed single-shot density-difference estimator and show that it achieves the optimal convergence rate. We then show how the proposed density-difference estimator can be utilized in L2-distance approximation. Finally, we experimentally demonstrate the usefulness of the proposed method in robust distribution comparison such as class-prior estimation and change-point detection.
****
Learning the Architecture of Sum-Product Networks Using Clustering on Variables
Aaron Dennis, Dan Ventura
The sum-product network (SPN) is a recently-proposed deep model consisting of a network of sum and product nodes, and has been shown to be competitive with state-of-the-art deep models on certain difficult tasks such as image completion. Designing an SPN network architecture that is suitable for the task at hand is an open question. We propose an algorithm for learning the SPN architecture from data. The idea is to cluster variables (as opposed to data instances) in order to identify variable subsets that strongly interact with one another. Nodes in the SPN network are then allocated towards explaining these interactions. Experimental evidence shows that learning the SPN architecture significantly improves its performance compared to using a previously-proposed static architecture.
****
Towards a learning-theoretic analysis of spike-timing dependent plasticity
David Balduzzi, Michel Besserve
This paper suggests a learning-theoretic perspective on how synaptic plasticity benefits global brain functioning. We introduce a model, the selectron, that (i) arises as the fast time constant limit of leaky integrate-and-fire neurons equipped with spiking timing dependent plasticity (STDP) and (ii) is amenable to theoretical analysis. We show that the selectron encodes reward estimates into spikes and that an error bound on spikes is controlled by a spiking margin and the s

um of synaptic weights. Moreover, the efficacy of spikes (their usefulness to ot
her reward maximizing selectrons) also depends on total synaptic strength. Final
ly, based on our analysis, we propose a regularized version of STDP, and show th
e regularization improves the robustness of neuronal learning when faced with mu
ltiple stimuli.
****

Angular Quantization-based Binary Codes for Fast Similarity Search
Yunchao Gong, Sanjiv Kumar, Vishal Verma, Svetlana Lazebnik
This paper focuses on the problem of learning binary embeddings for efficient re
trieval of high-dimensional non-negative data. Such data typically arises in a l
arge number of vision and text applications where counts or frequencies are used
 as features. Also, cosine distance is commonly used as a measure of dissimilar
ity between such vectors. In this work, we introduce a novel spherical quantizat
ion scheme to generate binary embedding of such data and analyze its properties.
 The number of quantization landmarks in this scheme grows exponentially with da
ta dimensionality resulting in low-distortion quantization. We propose a very e
fficient method for computing the binary embedding using such large number of la
ndmarks. Further, a linear transformation is learned to minimize the quantizatio
n error by adapting the method to the input data resulting in improved embedding
. Experiments on image and text retrieval applications show superior performanc
e of the proposed method over other existing state-of-the-art methods.
****

Matrix reconstruction with the local max norm
Rina Foygel, Nathan Srebro, Russ R. Salakhutdinov
We introduce a new family of matrix norms, the ''local max'' norms, generalizing
 existing methods such as the max norm, the trace norm (nuclear norm), and the w
eighted or smoothed weighted trace norms, which have been extensively used in th
e literature as regularizers for matrix reconstruction problems. We show that th
is new family can be used to interpolate between the (weighted or unweighted) tr
ace norm and the more conservative max norm. We test this interpolation on simul
ated data and on the large-scale Netflix and MovieLens ratings data, and find im
proved accuracy relative to the existing matrix norms. We also provide theoretic
al results showing learning guarantees for some of the new norms.
****

Near-optimal Differentially Private Principal Components
Kamalika Chaudhuri, Anand Sarwate, Kaushik Sinha
Principal components analysis (PCA) is a standard tool for identifying good low-
dimensional approximations to data sets in high dimension. Many current data se
ts of interest contain private or sensitive information about individuals. Algo
rithms which operate on such data should be sensitive to the privacy risks in pu
blishing their outputs. Differential privacy is a framework for developing trad
eoffs between privacy and the utility of these outputs. In this paper we invest
igate the theory and empirical performance of differentially private approximati
ons to PCA and propose a new method which explicitly optimizes the utility of th
e output. We demonstrate that on real data, there this a large performance gap
between the existing methods and our method. We show that the sample complexity
 for the two procedures differs in the scaling with the data dimension, and that
 our method is nearly optimal in terms of this scaling.
****

Mandatory Leaf Node Prediction in Hierarchical Multilabel Classification
Wei Bi, James Kwok
In hierarchical classification, the prediction paths may be required to always e
nd at leaf nodes. This is called mandatory leaf node prediction (MLNP) and is pa
rticularly useful when the leaf nodes have much stronger semantic meaning than t
he internal nodes. However, while there have been a lot of MLNP methods in hiera
rchical multiclass classification, performing MLNP in hierarchical multilabel cl
assification is much more difficult. In this paper, we propose a novel MLNP algo
rithm that (i) considers the global hierarchy structure; and (ii) can be used o
n hierarchies of both trees and DAGs. We show that one can efficiently maximize
the joint posterior probability of all the node labels by a simple greedy algori

thm. Moreover, this can be further extended to the minimization of the expected symmetric loss. Experiments are performed on a number of real-world data sets with tree- and DAG-structured label hierarchies. The proposed method consistently outperforms other hierarchical and flat multilabel classification methods.
****
Synchronization can Control Regularization in Neural Systems via Correlated Noise Processes
Jake Bouvrie, Jean-jeacques Slotine
To learn reliable rules that can generalize to novel situations, the brain must be capable of imposing some form of regularization. Here we suggest, through theoretical and computational arguments, that the combination of noise with synchronization provides a plausible mechanism for regularization in the nervous system. The functional role of regularization is considered in a general context in which coupled computational systems receive inputs corrupted by correlated noise. Noise on the inputs is shown to impose regularization, and when synchronization upstream induces time-varying correlations across noise variables, the degree of regularization can be calibrated over time. The resulting qualitative behavior matches experimental data from visual cortex.
****
Online Regret Bounds for Undiscounted Continuous Reinforcement Learning
Ronald Ortner, Daniil Ryabko
We derive sublinear regret bounds for undiscounted reinforcement learning in continuous state space. The proposed algorithm combines state aggregation with the use of upper confidence bounds for implementing optimism in the face of uncertainty. Beside the existence of an optimal policy which satisfies the Poisson equation, the only assumptions made are Hoelder continuity of rewards and transition probabilities.
****
Perceptron Learning of SAT
Alex Flint, Matthew Blaschko
Boolean satisfiability (SAT) as a canonical NP-complete decision problem is one of the most important problems in computer science. In practice, real-world SAT sentences are drawn from a distribution that may result in efficient algorithms for their solution. Such SAT instances are likely to have shared characteristics and substructures. This work approaches the exploration of a family of SAT solvers as a learning problem. In particular, we relate polynomial time solvability of a SAT subset to a notion of margin between sentences mapped by a feature function into a Hilbert space. Provided this mapping is based on polynomial time computable statistics of a sentence, we show that the existance of a margin between these data points implies the existance of a polynomial time solver for that SAT subset based on the Davis-Putnam-Logemann-Loveland algorithm. Furthermore, we show that a simple perceptron-style learning rule will find an optimal SAT solver with a bounded number of training updates. We derive a linear time computable set of features and show analytically that margins exist for important polynomial special cases of SAT. Empirical results show an order of magnitude improvement over a state-of-the-art SAT solver on a hardware verification task.
****
On Lifting the Gibbs Sampling Algorithm
Deepak Venugopal, Vibhav Gogate
Statistical relational learning models combine the power of first-order logic, the de facto tool for handling relational structure, with that of probabilistic graphical models, the de facto tool for handling uncertainty. Lifted probabilistic inference algorithms for them have been the subject of much recent research. The main idea in these algorithms is to improve the speed, accuracy and scalability of existing graphical models' inference algorithms by exploiting symmetry in the first-order representation. In this paper, we consider blocked Gibbs sampling, an advanced variation of the classic Gibbs sampling algorithm and lift it to the first-order level. We propose to achieve this by partitioning the first-order atoms in the relational model into a set of disjoint clusters such that exact lifted inference is polynomial in each cluster given an assignment to all other

atoms not in the cluster. We propose an approach for constructing such clusters and determining their complexity and show how it can be used to trade accuracy with computational complexity in a principled manner. Our experimental evaluation shows that lifted Gibbs sampling is superior to the propositional algorithm in terms of accuracy and convergence.
****

Q-MKL: Matrix-induced Regularization in Multi-Kernel Learning with Applications to Neuroimaging
Chris Hinrichs, Vikas Singh, Jiming Peng, Sterling Johnson
Multiple Kernel Learning (MKL) generalizes SVMs to the setting where one simultaneously trains a linear classifier and chooses an optimal combination of given base kernels. Model complexity is typically controlled using various norm regularizations on the vector of base kernel mixing coefficients. Existing methods, however, neither regularize nor exploit potentially useful information pertaining to how kernels in the input set 'interact'; that is, higher order kernel-pair relationships that can be easily obtained via unsupervised (similarity, geodesics), supervised (correlation in errors), or domain knowledge driven mechanisms (which features were used to construct the kernel?). We show that by substituting the norm penalty with an arbitrary quadratic function Q \succeq 0, one can impose a desired covariance structure on mixing coefficient selection, and use this as an inductive bias when learning the concept. This formulation significantly generalizes the widely used 1- and 2-norm MKL objectives. We explore the model’s utility via experiments on a challenging Neuroimaging problem, where the goal is to predict a subject’s conversion to Alzheimer’s Disease (AD) by exploiting aggregate information from several distinct imaging modalities. Here, our new model outperforms the state of the art (p-values << 10−3). We briefly discuss ramifications in terms of learning bounds (Rademacher complexity).
****

Label Ranking with Partial Abstention based on Thresholded Probabilistic Models
Weiwei Cheng, Eyke Hüllermeier, Willem Waegeman, Volkmar Welker
Several machine learning methods allow for abstaining from uncertain predictions. While being common for settings like conventional classification, abstention has been studied much less in learning to rank. We address abstention for the label ranking setting, allowing the learner to declare certain pairs of labels as being incomparable and, thus, to predict partial instead of total orders. In our method, such predictions are produced via thresholding the probabilities of pairwise preferences between labels, as induced by a predicted probability distribution on the set of all rankings. We formally analyze this approach for the Mallows and the Plackett-Luce model, showing that it produces proper partial orders as predictions and characterizing the expressiveness of the induced class of partial orders. These theoretical results are complemented by experiments demonstrating the practical usefulness of the approach.
****

Weighted Likelihood Policy Search with Model Selection
Tsuyoshi Ueno, Kohei Hayashi, Takashi Washio, Yoshinobu Kawahara
Reinforcement learning (RL) methods based on direct policy search (DPS) have been actively discussed to achieve an efficient approach to complicated Markov decision processes (MDPs). Although they have brought much progress in practical applications of RL, there still remains an unsolved problem in DPS related to model selection for the policy. In this paper, we propose a novel DPS method, {\it weighted likelihood policy search (WLPS)}, where a policy is efficiently learned through the weighted likelihood estimation. WLPS naturally connects DPS to the statistical inference problem and thus various sophisticated techniques in statistics can be applied to DPS problems directly. Hence, by following the idea of the {\it information criterion}, we develop a new measurement for model comparison in DPS based on the weighted log-likelihood.
****