

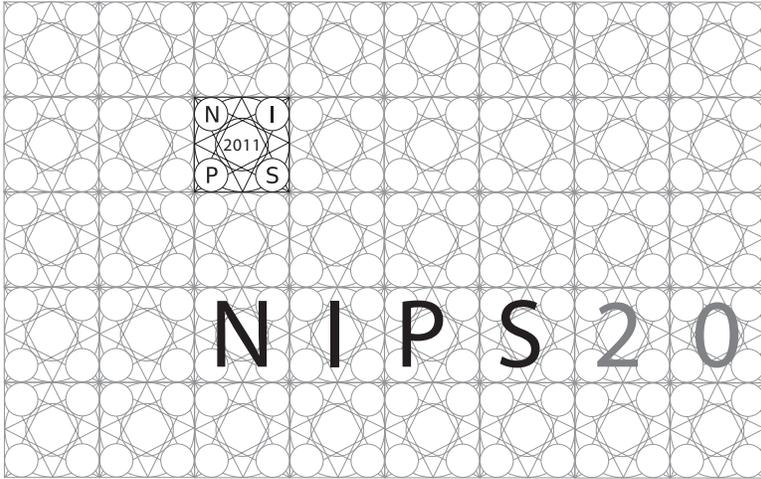


Neural Information Processing Systems

GRANADA, SPAIN DECEMBER 12TH -15TH



2011 CONFERENCE BOOK



11

Abstracts of Papers

TUTORIALS

December 12, 2011

Granada Congress and
Exhibition Centre, Granada, Spain

CONFERENCE SESSIONS

December 12-15, 2011

Granada Congress and
Exhibition Centre, Granada, Spain

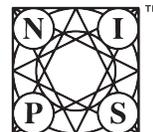
WORKSHOPS

December 16-17, 2011

Melia Sierra Nevada & Melia Sol y
Nieve, Sierra Nevada, Spain

Sponsored by the Neural Information Processing
System Foundation, Inc

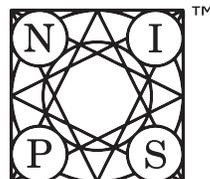
The technical program includes six invited talks and 306 accepted papers, selected from a total of 1,400 submissions considered by the program committee. Because the conference stresses interdisciplinary interactions, there are no parallel sessions. Papers presented at the conference will appear in "Advances in Neural Information Processing Systems 23," edited by Rich Zemel, John Shawe-Taylor, Peter Bartlett, Fernando Pereira and Kilian Weinberger.



Neural Information
Processing Systems
Foundation

TABLE OF CONTENTS

Organizing Committee	3	WEDNESDAY	
Program Committee	3	Oral Sessions	
NIPS Foundation Offices and Board Members	4	Sessions 6 - 10, Abstracts	73
Core Logistics Team	4	Spotlights Sessions	
Awards	4	Sessions 5 - 7, Abstracts	73
Sponsors	5	Poster Sessions	
		Sessions W1 - W101	76
PROGRAM HIGHLIGHTS	6	Abstracts	80
MONDAY		Location of Presentations	79
Tutorials		Demonstrations	103
Sessions 1 - 3, Abstracts	8		
Spotlights Session	12	THURSDAY	
Poster Sessions		Oral Sessions	
Sessions M1 - M102,	12	Sessions 11 - 14, Abstracts	105
Abstracts	16		
Location of Presentations	15	Reviewers	107
		Author Index	110
TUESDAY		Future Conference Locations	112
Oral Sessions			
Sessions 1 - 5, Abstracts	40		
Spotlights Sessions			
Sessions 2 - 4, Abstracts	40		
Poster Sessions			
Sessions T1 - T103	43		
Abstracts	48		
Location of Presentations	47		
Demonstrations	71		



Neural Information
Processing Systems
Foundation

ORGANIZING COMMITTEE

General Chairs	John Shawe-Taylor , University College London, Richard Zemel , University of Toronto
Program Chairs	Peter Bartlett , Queensland Univ. of Technology & UC Berkeley, Fernando Pereira , Google Research
Spanish Ambassador	Jesus Cortes , University of Granada, Spain
Tutorials Chair	Max Welling , University of California, Irvine
Workshop Chairs	Fernando Perez-Cruz , University Carlos III in Madrid, Spain ; Jeff Bilmes , University of Washington
Demonstration Chair	Samy Bengio , Google Research
Publications Chair & Electronic Proceedings Chair	Kilian Weinberger , Washington University in St. Louis
Program Manager	David Hall , University of California, Berkeley

PROGRAM COMMITTEE

Cedric Archambeau (Xerox Research Centre Europe)	Jan Peters (Max Planck Institute of Intelligent Systems, Tübingen)
Andreas Argyriou (Toyota Technological Institute at Chicago)	Jon Pillow (University of Texas, Austin)
Peter Auer (Montanuniversität Leoben)	Joelle Pineau (McGill University)
Mikhail Belkin (Ohio State University)	Ali Rahimi (San Francisco, CA)
Chiru Bhattacharyya (Indian Institute of Computer Science)	Sasha Rakhlin (University of Pennsylvania)
Charles Cadieu (University of California, Berkeley)	Pradeep Ravikumar (University of Texas, Austin)
Michael Collins (Columbia University)	Ruslan Salakhutdinov (MIT)
Ronan Collobert (IDIAP Research Institute)	Sunita Sarawagi (IIT Bombay)
Hal Daume III (University of Maryland)	Thomas Serre (Brown University)
Fei Fei Li (Stanford University)	Shai Shalev-Shwartz (The Hebrew University of Jerusalem)
Rob Fergus (New York University)	Ingo Steinwart (Universität Stuttgart)
Maria Florina Balcan (Georgia Tech)	Amar Subramanya (Google)
Kenji Fukumizu (Institute of Statistical Mathematics)	Masashi Sugiyama (Tokyo Institute of Technology)
Amir Globerson (The Hebrew University of Jerusalem)	Koji Tsuda (National Institute of Advanced Industrial Science and Technology)
Sally Goldman (Google)	Raquel Urtasun (Toyota Technological Institute at Chicago)
Noah Goodman (Stanford University)	Manik Varma (Microsoft)
Alexander Gray (Georgia Tech)	Nicolas Vayatis (Ecole Normale Supérieure de Cachan)
Katherine Heller (MIT)	Jean-Philippe Vert (Mines ParisTech)
Guy Lebanon (Georgia Tech)	Hanna Wallach (University of Massachusetts Amherst)
Mate Lengyel (University of Cambridge)	Frank Wood (Columbia University)
Roger Levy (University of California, San Diego)	Eric Xing (Carnegie Mellon University)
Hang Li (Microsoft)	Yuan Yao (Peking University)
Chih-Jen Lin (National Taiwan University)	Kai Yu (NEC Labs)
Phil Long (Google)	Tong Zhang (Rutgers University)
Yi Ma (University of Illinois at Urbana-Champaign)	Jerry Zhu (University of Wisconsin-Madison)
Remi Munos (INRIA, Lille)	

NIPS would like to especially thank Microsoft Research for their donation of Conference Management Toolkit (CMT) software and server space.

NIPS FOUNDATION OFFICERS & BOARD MEMBERS

President	Terrence Sejnowski , The Salk Institute
Treasurer	Marian Stewart Bartlett , University of California, San Diego
Secretary	Michael Mozer , University of Colorado, Boulder
Legal Advisor	Phil Sotel , Pasadena, CA
Executive	John Lafferty , Carnegie Mellon University; Chris Williams , University of Edinburgh; Dale Schuurmans , University of Alberta, Canada; Yoshua Bengio , University of Montreal, Canada; Daphne Koller , Stanford University; John C. Platt , Microsoft Research; Bernhard Schölkopf , Max Planck Institute for Biological Cybernetics, Tübingen
Advisory Board	Sue Becker , McMaster University, Ontario, Canada, Gary Blasdel , Harvard Medical School, Jack Cowan , University of Chicago, Thomas G. Dietterich , Oregon State University, Stephen Hanson , Rutgers University, Michael I. Jordan , UC Berkeley, Michael Kearns , University of Pennsylvania, Scott Kirkpatrick , Hebrew University, Jerusalem, Richard Lippmann , Massachusetts Institute of Technology, Todd K. Leen , Oregon Graduate Institute, Bartlett Mel , University of Southern California, John Moody , International Computer Science Institute, Berkeley and Portland, Gerald Tesauro , IBM Watson Labs, Dave Touretzky , Carnegie Mellon University, Sebastian Thrun , Stanford University, Lawrence Saul , University of California, San Diego, Sara A. Solla , Northwestern University Medical School, Yair Weiss , Hebrew University of Jerusalem
Emeritus Members	T. L. Fine , Cornell University, Eve Marder , Brandeis University

CORE LOGISTICS TEAM

The running of NIPS would not be possible without the help of many volunteers, students, researchers and administrators who donate their valuable time and energy to assist the conference in various ways. However, there is a core team at the Salk Institute whose tireless efforts make the conference run smoothly and efficiently every year. This year, NIPS would particularly like to acknowledge the exceptional work of:

Lee Campbell - IT Manager
Chris Hiestand - Webmaster
Mary Ellen Perry - Executive Director
Montse Gamez - Administrator
Ramona Marchand - Administrator

AWARDS

OUTSTANDING STUDENT PAPER AWARDS

Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials

Philipp Krähenbühl * and Vladlen Koltun

Priors over Recurrent Continuous Time Processes

Ardavan Saeedi * and Alexandre Bouchard-Côte

Fast and Accurate k-means For Large Datasets

Michael Shindler * Alex Wong, and Adam Meyerson

* Winner

STUDENT PAPER HONORABLE MENTIONS

Learning Sparse Representations of High Dimensional Data on Large Scale Dictionaries

Zhen James Xiang * Hao Xu, and Peter Ramadge

The Manifold Tangent Classifier

Salah Rifai *, Yann Dauphin *, Pascal Vincent, Yoshua Bengio, and Xavier Muller

SPONSORS

NIPS gratefully acknowledges the generosity of those individuals and organizations who have provided financial support for the NIPS 2011 conference. The financial support enabled us to sponsor student travel and participation, the outstanding student paper awards, the demonstration track and the opening buffet.



Microsoft[®]
Research

Google[™]



Artificial Intelligence

DE Shaw & Co

TWO 2σ SIGMA

IBM Research



TOYOTA



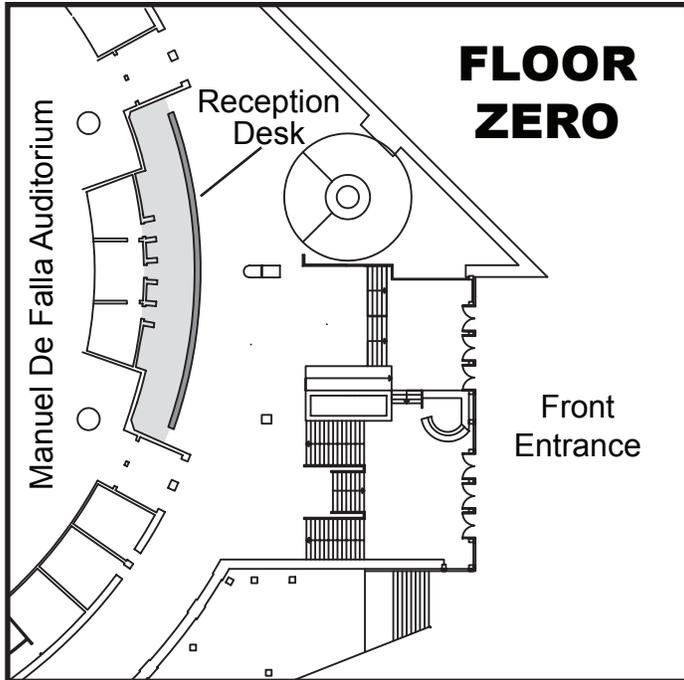
eBay[®]
Research Labs

YAHOO!
LABS

Telefonica

The Springer logo, which is a chess knight piece, followed by the text "Springer" and "Machine Learning Journal" below it.

PROGRAM HIGHLIGHTS



TUESDAY (CONTINUED)

10:40 – 11:10 am

Spotlights Session 2

11:10 – 11:30 am

Oral Session 2: **On the Completeness of First-Order Knowledge Compilation for Lifted Probabilistic Inference**

G. Van den Broeck

11:30 am – 12:00 - Coffee Break

12:00 – 12:40 pm

Oral Session 3: **Modelling Genetic Variations using Fragmentation-Coagulation Processes**

Y. Teh, C. Blundell, L. Elliott

Priors over Recurrent Continuous Time Processes

A. Saeedi, A. Bouchard-Côté

12:40 – 1:10 pm

Spotlights Session 3

1:10 – 1:30 pm

Oral Session 4: **Generalization Bounds and Consistency for Latent Structural Probit and Ramp Loss**

D. Mcallester, J. Keshet

1:30 – 2 pm

Spotlights Session 4

2 – 4 pm - Break

4:00 – 5:30 pm

Oral Session 5: **Invited Talk: Sparsity: algorithms, approximations, and analysis**

Anna Gilbert

Learning Sparse Representations of High Dimensional Data on Large Scale Dictionaries

Z. Xiang, H. Xu, P. Ramadge

High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity

P. Loh, M. Wainwright

5:30 – 5:45 pm - Coffee Break

5:45 – 11:59 pm

Demonstrations
Andalucia II

5:45 – 11:59 pm

Poster Session
Floor One

MONDAY DEC 12TH

7:30 am – 6:30 pm

Registration Desk Open
Floor Zero

8 am – 6:30 pm

Internet Access Room Open

6:30 – 6:40 pm

Opening Remarks, Awards and Reception

6:40 – 7:00 pm

Spotlight Session 1

7:00 – 11:59 pm

Poster Session
Floor One

TUESDAY DEC 13TH

8 am – 5:30 pm

Registration Desk Open
Floor Zero

9:30 – 10:40 am

Oral Session 1: **Posner Lecture: Learning About Sensorimotor Data**
Richard Sutton

Oral Session 1: **A Non-Parametric Approach to Dynamic Programming**

O. Kroemer, J. Peters

PROGRAM HIGHLIGHTS

WEDNESDAY DEC 14TH

8 am – 5:30 pm

Registration Desk Open
Floor Zero

9:30 – 10:40 am

Oral Session 6: **Posner Lecture: From kernels to causal inference**
Bernhard Schölkopf

High-Dimensional Graphical Model Selection: Tractable Graph Families and Necessary Conditions

A. Anandkumar, V. Tan, A. Willsky

10:40 – 11:10 am

Spotlights Session 5

11:10 – 11:30 am

Oral Session 7: **Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials**
P. Krähenbühl, V. Koltun

11:30 am – 12:00 - Coffee Break

12:00 – 12:40 pm

Oral Session 8: **Efficient Online Learning via Randomized Rounding**
N. Cesa-Bianchi, O. Shamir

Convergence Rates of Inexact Proximal-Gradient Methods for Convex Optimization

M. Schmidt, N. Le Roux, F. Bach

12:40 – 1:10 pm

Spotlights Session 6

1:10 – 1:30 pm

Oral Session 9: **k-NN Regression Adapts to Local Intrinsic Dimension**
S. Kpotufe

1:30 – 2:00 pm

Spotlights Session 7

2:00 – 4:00 pm - Break

4:00 – 5:30 pm

Oral Session 10: **Invited Talk: Natural Algorithms**
Bernard Chazelle

Iterative Learning for Reliable Crowdsourcing Systems

D. Karger, S. Oh, D. Shah

A Collaborative Mechanism for Crowdsourcing Prediction Problems

J. Abernethy, R. Frongillo

5:30 – 5:45 pm - Coffee Break

5:45 – 11:59 pm

Demonstrations
Andalucia II

5:45 – 11:59 pm

Poster Session
Floor One

THURSDAY DEC 15TH

8 AM – 12:00 pm

Registration Desk Open
Floor Zero

9:30 – 10:40 am

Oral Session 11: **Invited Talk: The Neuronal Replicator Hypothesis: Novel Mechanisms for Information Transfer and Search in the Brain**
Eörs Szathmáry, Chrisantha Fernando

Continuous-Time Regression Models for Longitudinal Networks

D. Vu, A. Asuncion, D. Hunter, P. Smyth

10:40 – 11:20 am

Oral Session 12: **Scalable Training of Mixture Models via Coresets**
D. Feldman, M. Faulkner, A. Krause

Fast and Accurate k-means For Large Datasets

M. Shindler, A. Wong, A. Meyerson

11:20 am – 12:00 - Coffee Break

12:00 – 1:10 pm

Oral Session 13: **Invited Talk: Diversity in an Olfactory Network**
Gilles Laurent

Empirical models of spiking in neural populations

J. Macke, L. Buesing, J. Cunningham, B. Yu, K., Shenoy, M. Sahani

1:10 – 1:50 pm

Oral Session 14: **The Manifold Tangent Classifier**
S. Rifai, Y. Dauphin, P. Vincent, Y., Bengio, X. Muller

Reconstructing Patterns of Information Diffusion from Incomplete Observations

F. Chierichetti, J. Kleinberg, D. Liben-Nowell

3:00 pm Bus to Sierra Nevada

5:00 pm Alhambra Tour Bus Boarding

5:30 pm Alhambra Tour Bus Boarding

8:00 pm Bus to Sierra Nevada

MONDAY TUTORIALS

9:30 AM – 11:30 AM - Tutorial Session 1

Flexible, Multivariate Point Process Models for Unlocking the Neural Code

Jonathan Pillow

Location: Andulucia II & III

Linear Programming Relaxations for Graphical Models

Amir Globerson, Tommi Jaakkola

Location: Manuel de Falla

12:00 – 2:00 PM - Tutorial Session 2

Lagrangian Relaxation Algorithms for Inference in Natural Language Processing

Alexander Rush, Michael Collins

Location: Andulucia II & III

Modern Bayesian Nonparametrics

Peter Orbanz, Yee Whye Teh

Location: Manuel de Falla

4:00 PM – 6:00 PM - Tutorial Session 3

Tutorial: Graphical Models for the Internet

Amr Ahmed, Alexander Smola

Location: Andulucia II & III

Tutorial: Information Theory in Learning and Control

Naftali Tishby

Location: Manuel de Falla

Tutorial Session 1, 9:30 – 11:30am

Tutorial: ***Flexible, Multivariate Point Process Models for Unlocking the Neural Code***

Jonathan Pillow

A central goal of computational neuroscience is to understand the neural code, the semantic relationship between neural spike trains and the extrinsic (sensory, motor, & cognitive) variables that they represent. One powerful approach to this problem involves “cascade” point process models, which describe the neural encoding process in terms of a cascade of three stages: (1) linear dimensionality-reduction of a high-dimensional stimulus space; (2) a nonlinear transformation from feature space to spike rate; and (3) an inhomogeneous, conditional renewal (e.g., Poisson) spiking process. These models have been shown to provide accurate descriptions of single- and multi-neuron spike responses in a wide variety of brain areas, and have shed light on the fundamental units (rates, spike times, correlations, oscillations) that neurons use to convey information. Recent innovations have focused on extending these models to incorporate richer nonlinear dependencies and dynamics, and to capture more biologically realistic features of neural spike trains.

In this tutorial, I will provide a general introduction to cascade neural encoding models and then discuss some more recent advances, including models for non-Poisson spike trains and correlated neural population responses. Topics will include: Poisson & renewal processes, reverse correlation, spike-triggered average / covariance (STA/STC) analysis, inverse regression, maximally informative dimensions (MID), generalized linear models (GLMs), Ising models, latent variable / shared-

ABSTRACTS OF TUTORIALS

noise models, functional connectivity, advanced regularization methods, and model-based (Bayesian) techniques for decoding multi-neuron spike trains.

Jonathan Pillow is an assistant professor in Psychology and Neurobiology at the University of Texas at Austin. He graduated from the University of Arizona in 1997 with a degree in mathematics and philosophy, and was a U.S. Fulbright fellow in Morocco in 1998. He received his Ph.D. in neuroscience from NYU in 2005, and was a Royal Society postdoctoral research fellow at the Gatsby Computational Neuroscience Unit, UCL from 2005 to 2008. His recent work involves statistical methods for understanding the neural code in single neurons and neural populations, and his lab conducts psychophysical experiments designed to test Bayesian models of human sensory perception.

Tutorial Session 1, 9:30 – 11:30am

Tutorial: ***Linear Programming Relaxations for Graphical Models***

Amir Globerson, Tommi Jaakkola

Many probabilistic modeling tasks rely on solving challenging inference problems. These combinatorial problems arise, e.g., in predicting likely values for variables as in selecting and orienting residues in protein design, parsing in natural language processing, or when learning the model structure itself. In many cases, the inference problems involve densely connected variables (or higher order dependences) and are provably hard. However, recent research has shown that some of these difficulties can be overcome by a careful choice of approximation schemes and learning algorithms. These have yielded very encouraging results in wide array of fields, from machine vision and natural language processing to computational biology and signal processing. In this tutorial, we will focus on linear programming (LP) relaxations which have been particularly successful in solving inference problems. Intuitively, LP relaxations decompose a complex problem into a set of simpler subproblems that are subsequently encouraged to agree. If the subproblems agree about a solution, then the solution is the optimal one, otherwise it is fractional. Geometrically, the relaxation maintains an outer bound approximation to a polytope whose vertexes correspond to valid solutions. We will introduce and explain key ideas behind recent approaches, discuss when they can and cannot be applied, how they can be integrated into supervised learning schemes and what efficient message passing algorithms exist for solving them. We will also discuss how similar ideas can be used for calculating marginals. Examples from several applications will be provided, including computational biology, natural language processing (see also a separate tutorial on Dual Decomposition for NLP), and structure learning.

Amir Globerson is senior lecturer at the School of Engineering and Computer Science at the Hebrew University. He received a PhD in computational neuroscience from the Hebrew University, and was a Rothschild postdoctoral fellow at MIT. He joined the Hebrew University in 2008. His research interests include graphical models and probabilistic inference, convex optimization, robust learning and natural language processing.

Tommi Jaakkola is a professor of Electrical Engineering and Computer Science at MIT. He received an M.Sc. degree in theoretical physics from Helsinki University of Technology, and Ph.D. from MIT in computational neuroscience. Following a Sloan postdoctoral fellowship in computational molecular biology, he joined the MIT faculty in 1998. His research interests include statistical inference, graphical models, and large scale modern estimation problems with predominantly incomplete data..

Tutorial Session 2, 12:00 - 2:00pm

Tutorial: **Lagrangian Relaxation Algorithms for Inference in Natural Language Processing**
Alexander Rush, Michael Collins

There has been a long history in combinatorial optimization of methods that exploit structure in complex problems, using methods such as dual decomposition or Lagrangian relaxation. These methods leverage the observation that complex inference problems can often be decomposed into efficiently solvable sub-problems. Recent work within the machine learning community has explored algorithms for MAP inference in graphical models using these methods, as an alternative for example to maximum-product loopy belief propagation.

In this tutorial, we give an overview of Lagrangian relaxation for inference problems in natural language processing. The goals of the tutorial will be two-fold:

1) to give an introduction to key inference problems in NLP: for example problems arising in machine translation, sequence modeling, parsing, and information extraction.

2) to give a formal and practical overview of Lagrangian relaxation as a method for deriving inference algorithms for NLP problems. In general, the algorithms we describe combine combinatorial optimization methods (for example dynamic programming, exact belief propagation, minimum spanning tree, all-pairs shortest path) with subgradient methods from the optimization community. Formal guarantees for the algorithms come from a close relationship to linear programming relaxations.

For many of the problems that we consider, the resulting algorithms produce exact solutions, with certificates of optimality, on the vast majority of examples. In practice the algorithms are efficient for problems that are either NP-hard (as is the case for non-projective parsing, or for phrase-based translation), or for problems that are solvable in polynomial time using dynamic programming, but where the traditional exact algorithms are far too expensive to be practical.

Alexander Rush is a Ph.D. candidate in computer science at MIT. His research explores novel decoding methods for problems in natural language processing with applications to parsing and statistical machine translation. His paper "Dual Decomposition for Parsing with Non-Projective Head Automata" received the best paper award at EMNLP 2010.

Michael Collins is the Vikram S. Pandit Professor of computer science at Columbia University. His research is focused on topics including statistical parsing, structured prediction problems in machine learning, and applications including machine translation, dialog systems, and speech recognition. His awards include a Sloan fellowship, an NSF career award, and best paper awards at EMNLP (2002, 2004, and 2010), UAI (2004 and 2005), and CoNLL 2008.

Tutorial Session 2, 12:00 - 2:00pm

Tutorial: **Modern Bayesian Nonparametrics**
Peter Orbanz, Yee Whye Teh

A nonparametric model is a model on an infinite dimensional parameter space. The parameter space represents the set of all possible solutions for a given learning problem -- for example, the set of smooth functions in nonlinear regression, or of all probability densities in a density estimation problem. A Bayesian nonparametric model defines a prior distribution on such an infinite dimensional space, where the traditional prior assumptions (e.g. "the parameter is likely to be small") are replaced by structural assumptions ("the function is likely to be smooth"), and learning then requires computation of the posterior distribution given data.

The tutorial will provide a high-level introduction to modern Bayesian nonparametrics. Since first attracting attention at NIPS a decade ago, the field has evolved substantially in the machine learning, statistics and probability communities. We now have a much improved understanding of how novel models can be used effectively in applications, of their theoretical properties, of techniques for posterior computation, and of how they can be combined to fit the requirements of a given problem. In the tutorial, we will survey the current state of the art with a focus on recent developments of interest in machine learning.

Peter Orbanz is a research fellow at the University of Cambridge. He holds a PhD degree from ETH Zurich and will join the Statistics Faculty at Columbia University as an Assistant Professor in 2012. He is interested in the mathematical and algorithmic aspects of Bayesian nonparametric models and of related learning technologies.

Yee Whye Teh is a reader at the Gatsby Computational Neuroscience Unit, UCL. He obtained his PhD from the University of Toronto, and did postdoctoral work at the University of California, Berkeley and National University of Singapore. He is interested in developing probabilistic and Bayesian methodologies for solving machine learning problems.

Tutorial Session 3, 4:00 – 6:00 pm

Tutorial: **Graphical Models for the Internet**
Amr Ahmed, Alexander Smola

In this tutorial we give an overview over applications and scalable inference in graphical models for the internet. Structured data analysis has become a key enabling technique to process significant amounts of data, ranging from entity extraction on webpages to sentiment and topic analysis for news articles and comments. Our tutorial covers large scale sampling and optimization methods for Nonparametric Bayesian models such as Latent Dirichlet Allocation, both from a statistics and a systems perspective. Subsequently we give an overview over a range of generative models to elicit sentiment, ideology, time dependence, hierarchical structure, and multilingual similarity from data. We conclude with an overview of recent advances in (semi)supervised information extraction methods based on conditional random fields and related undirected graphical models.

Amr Ahmed is a Research Scientist at Yahoo! Research. He got his M.Sc and PhD from the School of Computer Science at Carnegie Mellon University in 2009 and 2011 respectively. He is interested in graphical models and Bayesian non-parametric statistics with an eye towards building efficient inference algorithms for such models that scale to the size of the data on the internet. On the application side, he is interested in information retrieval over structured sources, social media (blogs, news stream, twitter), user modeling and personalization.

Alex Smola is Principal Research Scientist at Yahoo! Research and adjunct Professor at the Australian National University. Prior to joining Yahoo! in September 2008 he was group leader of the machine learning program at NICTA, a federally funded research center in Canberra, Australia. His role involved leading a team of up to 35 researchers, programmers, PhD students, visitors, and interns. Prior to that he held positions at the Australian National University, GMD FIRST in Berlin and AT&T Research. He has written and edited 4 books, published over 160 papers, and organized several Summer Schools.

Tutorial Session 3, 4:00 – 6:00 pm

Tutorial: **Information Theory in Learning and Control**
Naftali Tishby

The concept of information plays a fundamental role in modern theories of cognition, in particular as regards perception, learning and behavior. However, the formal links between information theory and cognitive neuroscience remain elusive, and information theoretic measures are all too frequently misapplied.

In this tutorial I present a principled overview of some recent links between Shannon's theory of communication and statistical learning theory, and then put these in a more general framework of information theory of perception and control. I begin with the well-known links between statistical inference and information, from simple hypothesis testing, parameter estimation and the concept of sufficient statistics. An information theoretic generalization of minimal sufficient statistics leads to a natural optimization problem; i.e., the information bottleneck principle (IB), which is directly connected to classical models of communication with side information. While the IB optimization problem is generally non-convex, it is efficiently solvable for multivariate Gaussian variables. This special case was recently generalized using the Kernel trick to a wide class of dimensionality reduction problems, similar to Kernel-CCA. This version makes the information bottleneck method completely applicable for a wide range of practical problems. I will discuss the advantages of these algorithms over K-CCA and describe the importance of the information tradeoff (information curve) for hierarchical data representation and feature extraction. I will also discuss some finite sample properties and generalization bounds for this method.

In the second part of the tutorial I begin with the (Kelly) gambling problem and show that it can be extended to a general computational theory of value-information tradeoff, for both MDP and POMDP settings. This will provide a unified theoretical framework for optimal control and information seeking algorithms, and one that has the potential to be a principled model of perception-action cycles. I will then discuss the concept of predictive information and show how it provides useful bounds on the information flow in perception and control, and how it can be applied in robotics and neuroscience.

Naftali Tishby, the director of the Interdisciplinary Center for Neural Computation (ICNC), holds the Ruth and Stan Flinkman Chair in Brain Research at the Edmond and Lily Safra Center for Brain Science (ELSC). He was the founding chair of the Hebrew University's new computer engineering program, and a director of its Leibniz Research Center in Computer Science. Dr. Tishby received his Ph.D. (1985) in theoretical physics from the Hebrew University and was a research staff member at MIT and Bell Labs from 1985 to 1991. He was also a visiting professor at Princeton NECI, the University of Pennsylvania, the University of California at Santa Barbara, and at IBM research labs. Dr. Tishby is a leader of machine learning research and computational neuroscience. He was among the first to introduce methods from statistical physics into learning theory, and dynamical systems techniques in speech processing. His current research is at the interface between computer science, statistical physics and computational neuroscience and concerns the foundations of biological information processing and the connections between dynamics and information. He has introduced, with his colleagues, new theoretical frameworks for optimal adaptation and efficient information representation in biology, such as the Information Bottleneck method, the Minimum Information principle for neural coding, and information theory of Perception-Action Cycles.



MONDAY CONFERENCE

MONDAY - CONFERENCE

MONDAY, DECEMBER 12TH

6:30 – 6:40PM - OPENING REMARKS, AWARDS & RECEPTION

Tapas Reception.

SPOTLIGHT SESSION

SESSION 1 - 6:40 – 7:00 PM

Session Chair: Rob Fergus

- **Structural equations and divisive normalization for energy-dependent component analysis**
Jun-ichiro Hirayama, Aapo Hyvarinen, Kyoto University
Subject area: ICA, PCA, CCA & Other Linear Models
See abstract, page 29 (M57)
- **Uniqueness of Belief Propagation on Signed Graphs**
Yusuke Watanabe, The Institute of Statistical Mathematics
Subject area: Approximate Inference
See abstract, page 35 (M87)
- **Probabilistic amplitude and frequency demodulation**
Richard Turner, Maneesh Sahani, Gatsby Unit, UCL
Subject area: Probabilistic Models and Methods
See abstract, page 34 (M79)
- **On the accuracy of l1-filtering of signals with block-sparse structure**, A. Iouditski, UJF; F. Kilinc Karzan, Carnegie Mellon University; A. Nemirovski, Georgia Institute of Technology; B. Polyak, Institute for Control Sciences, RAS Moscow
Subject area: Statistical Learning Theory
See abstract, page 31 (M66)
- **Active dendrites: adaptation to spike-based communication**
Balazs Ujfalussy, Mate Lengyel, University of Cambridge
Subject area: Computational Neural Models
See abstract, page 19 (M15)

POSTER SESSION

AND RECEPTION - 7:00 – 11:59 PM

- M1 **A reinterpretation of the policy oscillation phenomenon in approximate policy iteration**, P. Wagner
- M2 **MAP Inference for Bayesian Inverse Reinforcement Learning**, J. Choi, K. Kim
- M3 **Monte Carlo Value Iteration with Macro-Actions**, Z. Lim, D. Hsu, L. Sun
- M4 **Reinforcement Learning using Kernel-Based Stochastic Factorization**, A. Barreto, D. Precup, J. Pineau
- M5 **Policy Gradient Coagent Networks**, P. Thomas
- M6 **Analysis and Improvement of Policy Gradient Estimation**, T. Zhao, H. Hachiya, G. Niu, M. Sugiyama
- M7 **Efficient Offline Communication Policies for Factored Multiagent POMDPs**, J. Messias, M. Spaan, P. Lima
- M8 **Speedy Q-Learning**, M. Gheshlaghi Azar, R. Munos, M. Ghavamzadeh, H. Kappen
- M9 **Optimal Reinforcement Learning for Gaussian Systems**, P. Hennig
- M10 **Clustering via Dirichlet Process Mixture Models for Portable Skill Discovery**, S. Niekum, A. Barto
- M11 **Nonlinear Inverse Reinforcement Learning with Gaussian Processes**, S. Levine, Z. Popovic, V. Koltun
- M12 **From Stochastic Nonlinear Integrate-and-Fire to Generalized Linear Models**, S. Mensi, R. Naud, W. Gerstner
- M13 **A Brain-Machine Interface Operating with a Real-Time Spiking Neural Network Control Algorithm**, J. Dethier, P. Nuyujukian, C. Eliasmith, T. Stewart, S. Elasaad, K. Shenoy, K. Boahen
- M14 **Energetically Optimal Action Potentials**, M. Stemmler, B. Sengupta, S. Laughlin, J. Niven
- M15 **Active dendrites: adaptation to spike-based communication**, B. Ujfalussy, M. Lengyel
- M16 **Inferring spike-timing-dependent plasticity from spike train data**, I. Stevenson, K. Koerding
- M17 **Dynamical segmentation of single trials from population neural data**, B. Petreska, B. Yu, J. Cunningham, G. Santhanam, S. Ryu, K. Shenoy, M. Sahani
- M18 **Emergence of Multiplication in a Biophysical Model of a Wide-Field Visual Neuron for Computing Object Approaches: Dynamics, Peaks, & Fits**, M. Keil
- M19 **Why The Brain Separates Face Recognition From Object Recognition**, J. Leibo, J. Mutch, T. Poggio
- M20 **Estimating time-varying input signals and ion channel states from a single voltage trace of a neuron**, R. Kobayashi, Y. Tsubo, P. Lansky, S. Shinomoto
- M21 **How biased are maximum entropy models?**, J. Macke, I. Murray, P. Latham
- M22 **Joint 3D Estimation of Objects and Scene Layout**, A. Geiger, C. Wojek, R. Urtasun
- M23 **Pylon Model for Semantic Segmentation**, V. Lempitsky, A. Vedaldi, A. Zisserman
- M24 **Higher-Order Correlation Clustering for Image Segmentation**, S. Kim, S. Nowozin, P. Kohli, C. Yoo

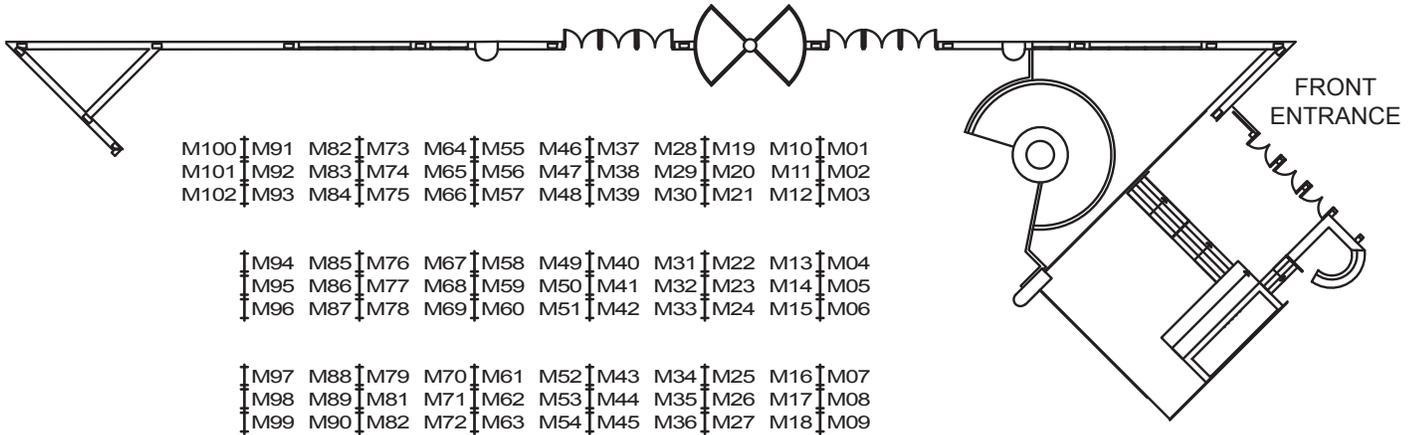
MONDAY - CONFERENCE

- M25 **Unsupervised learning models of primary cortical receptive fields and receptive field plasticity**, A. Saxe, M. Bhand, R. Mudur, B. Suresh, A. Ng
- M26 **Transfer Learning by Borrowing Examples**, J. Lim, R. Salakhutdinov, A. Torralba
- M27 **Large-Scale Category Structure Aware Image Categorization**, B. Zhao, F. Li, E. Xing
- M28 **Hierarchical Matching Pursuit for Recognition: Architecture and Fast Algorithms**, L. Bo, X. Ren, D. Fox
- M29 **Portmanteau Vocabularies for Multi-Cue Image Representation**, F. Khan, J. van de Weijer, A. Bagdanov, M. Vanrell
- M30 **PiCoDes: Learning a Compact Code for Novel-Category Recognition**, A. Bergamo, L. Torresani, A. Fitzgibbon
- M31 **Orthogonal Matching Pursuit with Replacement**, P. Jain, A. Tewari, I. Dhillon
- M32 **SpaRCS: Recovering low-rank and sparse matrices from compressive measurements** A. Waters, A. Sankaranarayanan, R. Baraniuk
- M33 **Signal Estimation Under Random Time-Warpings and Nonlinear Signal Alignment**, S. Kurtek, A. Srivastava, W. Wu
- M34 **Inverting Grice's Maxims to Learn Rules from Natural Language Extractions**, M. Sorower, T. Dietterich, J. Doppa, W. Orr, P. Tadepalli, X. Fern
- M35 **Multi-View Learning of Word Embeddings via CCA**, P. Dhillon, D. Foster, L. Ungar
- M36 **Active Ranking using Pairwise Comparisons**, K. Jamieson, R. Nowak
- M37 **Co-Training for Domain Adaptation**, M. Chen, K. Weinberger, J. Blitzer
- M38 **Efficient anomaly detection using bipartite k-NN graphs**, K. Sricharan, A. Hero
- M39 **A Maximum Margin Multi-Instance Learning Framework for Image Categorization**, H. Wang, H. Huang, F. Kamangar, F. Nie, C. Ding
- M40 **Advice Refinement in Knowledge-Based SVMs**, G. Kunapuli, R. Maclin, J. Shavlik
- M41 **Multiclass Boosting: Theory and Algorithms**, M. Saberian, N. Vasconcelos
- M42 **Boosting with Maximum Adaptive Sampling**, C. Dubout, F. Fleuret
- M43 **Kernel Embeddings of Latent Tree Graphical Models**, L. Song, A. Parikh, E. Xing
- M44 **Hierarchical Multitask Structured Output Learning for Large-scale Sequence Segmentation**, N. Goernitz, C. Widmer, G. Zeller, A. Kahles, S. Sonnenburg, G. Raetsch
- M45 **Submodular Multi-Label Learning**, J. Petterson, T. Caetano
- M46 **Algorithms for Hyper-Parameter Optimization**, J. Bergstra, R. Bardenet, Y. Bengio, B. Kégl
- M47 **Non-parametric Group Orthogonal Matching Pursuit for Sparse Learning with Multiple Kernels**, V. Sindhwani, A. Lozano
- M48 **Manifold Precs: An Annealing Technique for Diverse Sampling of Manifolds**, N. Shroff, P. Turaga, R. Chellappa
- M49 **Group Anomaly Detection using Flexible Genre Models**, L. Xiong, B. Póczos, J. Schneider
- M50 **Matrix Completion for Image Classification**, R. Cabral, F. De la Torre, J. Costeira, A. Bernardino
- M51 **Selecting Receptive Fields in Deep Networks**, A. Coates, A. Ng
- M52 **Co-regularized Multi-view Spectral Clustering**, A. Kumar, P. Rai, H. Daume III
- M53 **Learning with the weighted trace-norm under arbitrary sampling distributions**, R. Foygel, R. Salakhutdinov, O. Shamir, N. Srebro
- M54 **Active Learning with a Drifting Distribution**, L. Yang
- M55 **Bayesian Partitioning of Large-Scale Distance Data**, D. Adamez, V. Roth
- M56 **Beyond Spectral Clustering - Tight Relaxations of Balanced Graph Cuts**, M. Hein, S. Setzer
- M57 **Structural equations and divisive normalization for energy-dependent component analysis**, J. Hirayama, A. Hyvarinen
- M58 **Generalised Coupled Tensor Factorisation**, K. Yilmaz, A. Cemgil, U. Simsekli
- M59 **Similarity-based Learning via Data Driven Embeddings**, P. Kar, P. Jain
- M60 **Metric Learning with Multiple Kernels**, J. Wang, H. Do, A. Woznica, A. Kalousis
- M61 **Regularized Laplacian Estimation and Fast Eigenvector Approximation**, P. Perry, M. Mahoney
- M62 **Hierarchical Topic Modeling for Analysis of Time-Evolving Personal Choices**, X. Zhang, D. Dunson, L. Carin
- M63 **Efficient Learning of Generalized Linear and Single Index Models with Isotonic Regression**, S. Kakade, A. Kalai, V. Kanade, O. Shamir

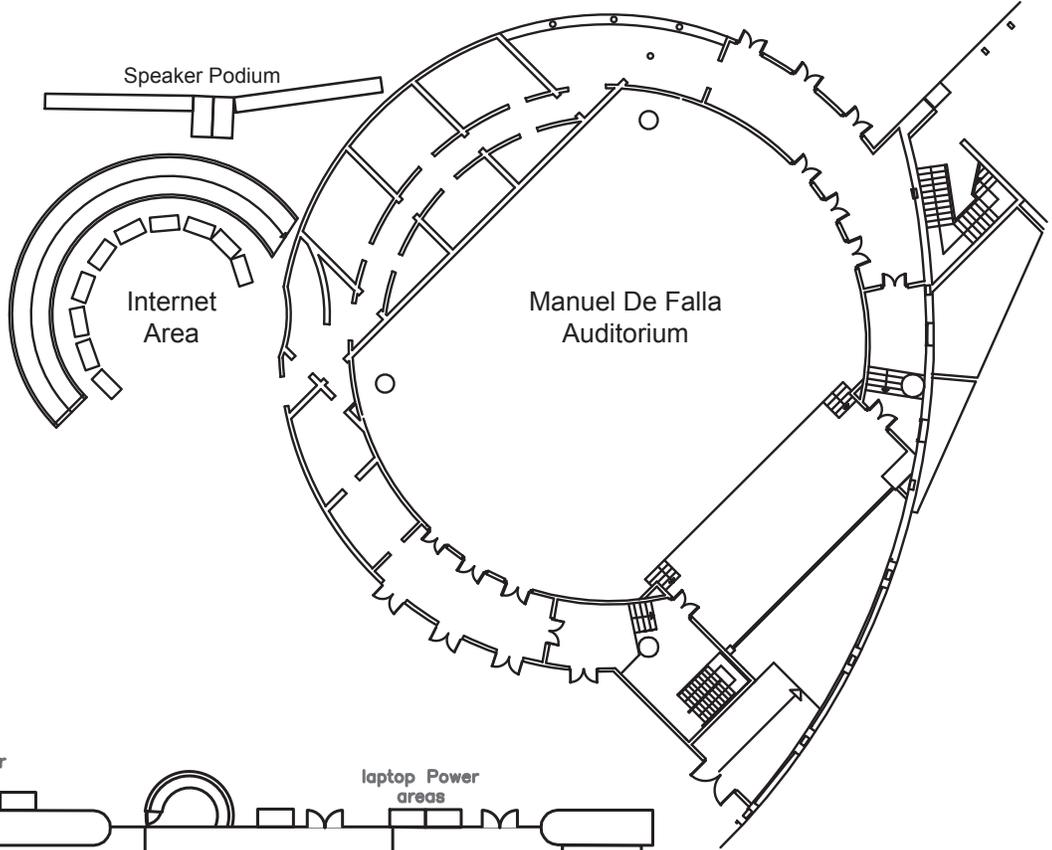
MONDAY - CONFERENCE

- M64 **Unifying Non-Maximum Likelihood Learning Objectives with Minimum KL Contraction**, S. Lyu
- M65 **Statistical Performance of Convex Tensor Decomposition**, R. Tomioka, T. Suzuki, K. Hayashi, H. Kashima
- M66 **On the accuracy of l1-filtering of signals with block-sparse structure**, A. Iouditski, F. Kilinc Karzan, A. Nemirovski, B. Polyak
- M67 **Committing Bandits**, L. Bui, R. Johari, S. Mannor
- M68 **Newtron: an Efficient Bandit algorithm for Online Multiclass Prediction**, E. Hazan, S. Kale
- M69 **Learning Eigenvectors for Free**, W. Koolen, W. Kotlowski, M. Warmuth
- M70 **Online Learning: Stochastic, Constrained, and Smoothed Adversaries**, A. Rakhlin, K. Sridharan, A. Tewari
- M71 **Optimistic Optimization of Deterministic Functions**, R. Munos
- M72 **The Impact of Unlabeled Patterns in Rademacher Complexity Theory for Kernel Classifiers**, L. Oneto, D. Anguita, A. Ghio, S. Ridella
- M73 **Unifying Framework for Fast Learning Rate of Non-Sparse Multiple Kernel Learning**, T. Suzuki
- M74 **Nearest Neighbor based Greedy Coordinate Descent**, I. Dhillon, P. Ravikumar, A. Tewari
- M75 **Agnostic Selective Classification**, Y. Wiener, R. El-Yaniv
- M76 **Greedy Model Averaging**, D. Dai, T. Zhang
- M77 **Confidence Sets for Network Structure**, D. Choi, P. Wolfe, E. Airoldi
- M78 **On Tracking The Partition Function**, G. Desjardins, A. Courville, Y. Bengio
- M79 **Probabilistic amplitude and frequency demodulation**, R. Turner, M. Sahani
- M80 **Structure Learning for Optimization**, . Yang, A. Rahimi
- M81 **Spike and Slab Variational Inference for Multi-Task and Multiple Kernel Learning**, M. Titsias, M. Lázaro-Gredilla
- M82 **Thinning Measurement Models and Questionnaire Design**, R. Silva
- M83 **An Application of Tree-Structured Expectation Propagation for Channel Decoding**, P. Olmos, L. Salamanca, J. Murillo Fuentes, F. Perez-Cruz
- M84 **Global Solution of Fully-Observed Variational Bayesian Matrix Factorization is Column-Wise Independent**, S. Nakajima, M. Sugiyama, S. Babacan
- M85 **Quasi-Newton Methods for Markov Chain**, Monte Carlo Y. Zhang, C. Sutton
- M86 **Nonstandard Interpretations of Probabilistic Programs for Efficient Inference**, D. Wingate, N. Goodman, A. Stuhlmüller, J. Siskind
- M87 **Uniqueness of Belief Propagation on Signed Graphs**, Y. Watanabe
- M88 **Non-conjugate Variational Message Passing for Multinomial and Binary Regression**, D. Knowles, T. Minka
- M89 **Ranking annotators for crowdsourced labeling tasks**, V. Raykar, S. Yu
- M90 **Gaussian process modulated renewal processes**, V. Rao, Y. Teh
- M91 **Infinite Latent SVM for Classification and Multi-task Learning**, J. Zhu, N. Chen, E. Xing
- M92 **Spatial distance dependent Chinese Restaurant Process for image segmentation**, S. Ghosh, A. Ungureanu, E. Sudderth, D. Blei
- M93 **Analytical Results for the Error in Filtering of Gaussian Processes**, A. Susemihl, R. Meir, M. Opper
- M94 **Robust Multi-Class Gaussian Process Classification**, D. Hernández-lobato, J. Hernández-Lobato, P. Dupont
- M95 **Additive Gaussian Processes**, D. Duvenaud, H. Nickisch, C. Rasmussen
- M96 **A Global Structural EM Algorithm for a Model of Cancer Progression**, A. Tofigh, E. Sjlund, M. Hglund, J. Lagergren
- M97 **Collective Graphical Models**, D. Sheldon, T. Dietterich
- M98 **Simultaneous Sampling and Multi-Structure Fitting with Adaptive Reversible Jump MCMC**, T. Pham, T. Chin, J. Yu, D. Suter
- M99 **Causal Discovery with Cyclic Additive Noise Models**, J. Mooij, D. Janzing, T. Heskes, B. Schölkopf
- M100 **A Model for Temporal Dependencies in Event Streams**, A. Gunawardana, C. Meek, P. Xu
- M101 **Facial Expression Transfer with Input-Output Temporal Restricted Boltzmann Machines**, M. Zeiler, G. Taylor, L. Sigal, I. Matthews, R. Fergus
- M102 **Learning Auto-regressive Models from Sequence and Non-sequence Data**, T. Huang, J. Schneider

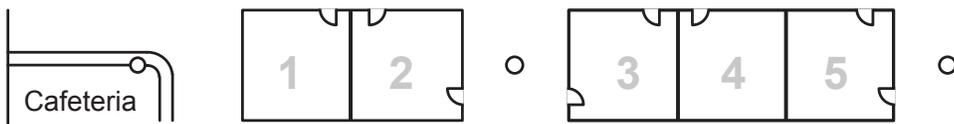
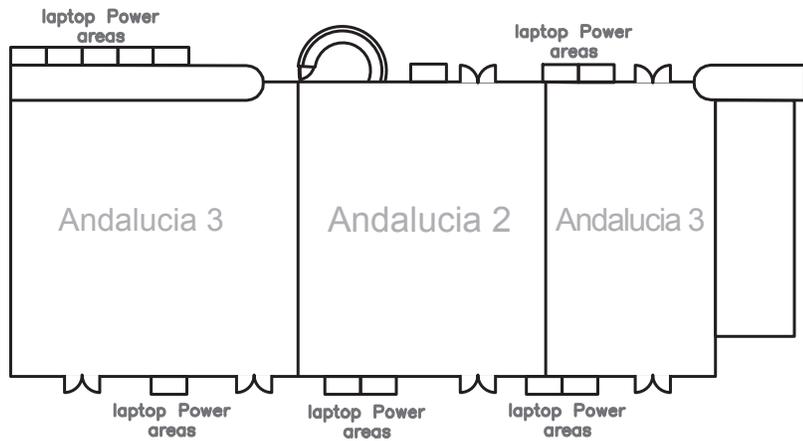
MONDAY POSTER FLOORPLAN



FLOOR ONE



To Cafeteria



M1 A reinterpretation of the policy oscillation phenomenon in approximate policy iteration

P. Wagner pwagner@cis.hut.fi
Aalto University School of Science

A majority of approximate dynamic programming approaches to the reinforcement learning problem can be categorized into greedy value function methods and value-based policy gradient methods. The former approach, although fast, is well known to be susceptible to the policy oscillation phenomenon. We take a fresh view to this phenomenon by casting a considerable subset of the former approach as a limiting special case of the latter. We explain the phenomenon in terms of this view and illustrate the underlying mechanism with artificial examples. We also use it to derive the constrained natural actor-critic algorithm that can interpolate between the aforementioned approaches. In addition, it has been suggested in the literature that the oscillation phenomenon might be subtly connected to the grossly suboptimal performance in the Tetris benchmark problem of all attempted approximate dynamic programming methods. We report empirical evidence against such a connection and in favor of an alternative explanation. Finally, we report scores in the Tetris problem that improve on existing dynamic programming based results.

Subject Area: Control and Reinforcement Learning

M2 MAP Inference for Bayesian Inverse Reinforcement Learning

J. Choi jdchoi@ai.kaist.ac.kr
K. Kim kekim@cs.kaist.ac.kr
KAIST

The difficulty in inverse reinforcement learning (IRL) arises in choosing the best reward function since there are typically an infinite number of reward functions that yield the given behaviour data as optimal. Using a Bayesian framework, we address this challenge by using the maximum a posteriori (MAP) estimation for the reward function, and show that most of the previous IRL algorithms can be modeled into our framework. We also present a gradient method for the MAP estimation based on the (sub)differentiability of the posterior distribution. We show the effectiveness of our approach by comparing the performance of the proposed method to those of the previous algorithms.

Subject Area: Control and Reinforcement Learning

M3 Monte Carlo Value Iteration with Macro-Actions

Z. Lim zhanweiz@gmail.com
D. Hsu dyhsu@comp.nus.edu.sg
L. Sun leews@comp.nus.edu.sg
National University of Singapore

POMDP planning faces two major computational challenges: large state spaces and long planning horizons. The recently introduced Monte Carlo Value Iteration (MCVI) can tackle POMDPs with very large discrete state spaces or continuous state spaces, but its performance degrades when faced with long planning horizons. This paper presents *Macro-MCVI*, which extends MCVI by exploiting macro-actions for temporal abstraction. We

provide sufficient conditions for *Macro-MCVI* to inherit the good theoretical properties of MCVI. *Macro-MCVI* does not require explicit construction of probabilistic models for macro-actions and is thus easy to apply in practice. Experiments show that *Macro-MCVI* substantially improves the performance of MCVI with suitable macro-actions.

Subject Area: Control and Reinforcement Learning

M4 Reinforcement Learning using Kernel-Based Stochastic Factorization

A. Barreto amsb@cs.mcgill.ca
D. Precup dprecup@cs.mcgill.ca
J. Pineau jpineau@cs.mcgill.ca
McGill University

Kernel-based reinforcement-learning (KBRL) is a method for learning a decision policy from a set of sample transitions which stands out for its strong theoretical guarantees. However, the size of the approximator grows with the number of transitions, which makes the approach impractical for large problems. In this paper we introduce a novel algorithm to improve the scalability of KBRL. We resort to a special decomposition of a transition matrix, called stochastic factorization, to fix the size of the approximator while at the same time incorporating all the information contained in the data. The resulting algorithm, kernel-based stochastic factorization (KBSF), is much faster but still converges to a unique solution. We derive a theoretical upper bound for the distance between the value functions computed by KBRL and KBSF. The effectiveness of our method is illustrated with computational experiments on four reinforcement-learning problems, including a difficult task in which the goal is to learn a neurostimulation policy to suppress the occurrence of seizures in epileptic rat brains. We empirically demonstrate that the proposed approach is able to compress the information contained in KBRL's model. Also, on the tasks studied, KBSF outperforms two of the most prominent reinforcement-learning algorithms, namely least-squares policy iteration and fitted Q-iteration.

Subject Area: Control and Reinforcement Learning

M5 Policy Gradient Coagent Networks

P. Thomas pthomas@cs.umass.edu
University of Massachusetts Amherst

We present a novel class of actor-critic algorithms for actors consisting of sets of interacting modules. We present, analyze theoretically, and empirically evaluate an update rule for each module, which requires only local information: the module's input, output, and the TD error broadcast by a critic. Such updates are necessary when computation of compatible features becomes prohibitively difficult and are also desirable to increase the biological plausibility of reinforcement learning methods.

Subject Area: Cognitive Science

M6 Analysis and Improvement of Policy Gradient Estimation

T. Zhao tingting@sg.cs.titech.ac.jp
 H. Hachiya hachiya@sg.cs.titech.ac.jp
 G. Niu gang@sg.cs.titech.ac.jp
 M. Sugiyama sugi@cs.titech.ac.jp
 Tokyo Institute of Technology

Policy gradient is a useful model-free reinforcement learning approach, but it tends to suffer from instability of gradient estimates. In this paper, we analyze and improve the stability of policy gradient methods. We first prove that the variance of gradient estimates in the PGPE (policy gradients with parameter-based exploration) method is smaller than that of the classical REINFORCE method under a mild assumption. We then derive the optimal baseline for PGPE, which contributes to further reducing the variance. We also theoretically show that PGPE with the optimal baseline is more preferable than REINFORCE with the optimal baseline in terms of the variance of gradient estimates. Finally, we demonstrate the usefulness of the improved PGPE method through experiments.
 Subject Area: Control and Reinforcement Learning

M7 Efficient Offline Communication Policies for Factored Multiagent POMDPs

J. Messias jmessias@isr.ist.utl.pt
 P. Lima pal@isr.ist.utl.pt
 Instituto Superior T ecnico
 M. Spaan m.t.j.spaan@tudelft.nl
 Delft University of Technology

Factored Decentralized Partially Observable Markov Decision Processes (Dec-POMDPs) form a powerful framework for multiagent planning under uncertainty, but optimal solutions require a rigid history-based policy representation. In this paper we allow inter-agent communication which turns the problem in a centralized Multiagent POMDP (MPOMDP). We map belief distributions over state factors to an agent's local actions by exploiting structure in the joint MPOMDP policy. The key point is that when sparse dependencies between the agents' decisions exist, often the belief over its local state factors is sufficient for an agent to unequivocally identify the optimal action, and communication can be avoided. We formalize these notions by casting the problem into convex optimization form, and present experimental results illustrating the savings in communication that we can obtain.
 Subject Area: Control and Reinforcement Learning

M8 Speedy Q-Learning

M. Gheshlaghi Azar m.azar@science.ru.nl
 H. Kappen b.kappen@science.ru.nl
 Radboud University of Nijmegen
 R. Munos remi.munos@inria.fr
 M. Ghavamzadeh mohammad.ghavamzadeh@inria.fr
 INRIA Lille - Nord Europe

We introduce a new convergent variant of Q-learning, called speedy Q-learning, to address the problem of

slow convergence in the standard form of the Q-learning algorithm. We prove a PAC bound on the performance of SQL, which shows that for an MDP with n state-action pairs and the discount factor γ only $T = O(\log(n)/(\epsilon^2(1-\gamma)^4))$ steps are required for the SQL algorithm to converge to an ϵ -optimal action-value function with high probability. This bound has a better dependency on $1/\epsilon$ and $1/(1-\gamma)$, and thus, is tighter than the best available result for Q-learning. Our bound is also superior to the existing results for both model-free and model-based instances of batch Q-value iteration that are considered to be more efficient than the incremental methods like Q-learning.

Subject Area: Control and Reinforcement Learning

M9 Optimal Reinforcement Learning for Gaussian Systems

P. Hennig philipp.hennig@gmail.com
 Max Planck Institute for Intelligent Systems

The exploration-exploitation trade-off is among the central challenges of reinforcement learning. The optimal Bayesian solution is intractable in general. This paper studies to what extent analytic statements about optimal learning are possible if all beliefs are Gaussian processes. A first order approximation of learning of both loss and dynamics, for nonlinear, time-varying systems in continuous time and space, subject to a relatively weak restriction on the dynamics, is described by an infinite-dimensional partial differential equation. An approximate finite-dimensional projection gives an impression for how this result may be helpful.

Subject Area: Control and Reinforcement Learning

M10 Clustering via Dirichlet Process Mixture Models for Portable Skill Discovery

S. Niekum sniekum@cs.umass.edu
 A. Barto barto@cs.umass.edu
 University of Massachusetts

Skill discovery algorithms in reinforcement learning typically identify single states or regions in state space that correspond to task-specific subgoals. However, such methods do not directly address the question of how many distinct skills are appropriate for solving the tasks that the agent faces. This can be highly inefficient when many identified subgoals correspond to the same underlying skill, but are all used individually as skill goals. Furthermore, skills created in this manner are often only transferable to tasks that share identical state spaces, since corresponding subgoals across tasks are not merged into a single skill goal. We show that these problems can be overcome by clustering subgoal data defined in an *agent-space* and using the resulting clusters as templates for skill termination conditions. Clustering via a Dirichlet process mixture model is used to discover a minimal, sufficient collection of portable skills.

Subject Area: Control and Reinforcement Learning

MONDAY ABSTRACTS

M11 Nonlinear Inverse Reinforcement Learning with Gaussian Processes

S. Levine svlevine@cs.stanford.edu
V. Koltun vladlen@stanford.edu
Stanford University
Z. Popovic zoran@washington.edu
University of Washington

We present a probabilistic algorithm for nonlinear inverse reinforcement learning. The goal of inverse reinforcement learning is to learn the reward function in a Markov decision process from expert demonstrations. While most prior inverse reinforcement learning algorithms represent the reward as a linear combination of a set of features, we use Gaussian processes to learn the reward as a nonlinear function, while also determining the relevance of each feature to the expert's policy. Our probabilistic algorithm allows complex behaviors to be captured from suboptimal stochastic demonstrations, while automatically balancing the simplicity of the learned reward structure against its consistency with the observed actions.

Subject Area: Control and Reinforcement Learning

M12 From Stochastic Nonlinear Integrate-and-Fire to Generalized Linear Models

S. Mensi skander.mensi@epfl.ch
Brain-Mind Institute
R. Naud richard.naud@epfl.ch
W. Gerstner wulfram.gerstner@epfl.ch
Ecole Polytechnique Federale de Lausanne

Variability in single neuron models is typically implemented either by a stochastic Leaky-Integrate-and-Fire model or by a model of the Generalized Linear Model (GLM) family. We use analytical and numerical methods to relate state-of-the-art models from both schools of thought. First we find the analytical expressions relating the subthreshold voltage from the Adaptive Exponential Integrate-and-Fire model (AdEx) to the Spike-Response Model with escape noise (SRM) as an example of a GLM). Then we calculate numerically the link-function that provides the firing probability given a deterministic membrane potential. We find a mathematical expression for this link-function and test the ability of the GLM to predict the firing probability of a neuron receiving complex stimulation. Comparing the prediction performance of various link-functions, we find that a GLM with an exponential link-function provides an excellent approximation to the Adaptive Exponential Integrate-and-Fire with colored-noise input. These results help to understand the relationship between the different approaches to stochastic neuron models.

Subject Area: Neuroscience

M13 A Brain-Machine Interface Operating with a Real-Time Spiking Neural Network Control Algorithm

J. Dethier jdethier@stanford.edu
P. Nuyujukian nips.npl.stanford@herag.com
S. Elasaad shauki@stanford.edu
K. Shenoy shenoy@stanford.edu
K. Boahen boahen@stanford.edu
Stanford University
C. Eliasmith celiasmith@uwaterloo.ca
T. Stewart tcstewar@uwaterloo.ca
University of Waterloo

Motor prostheses aim to restore function to disabled patients. Despite compelling proof of concept systems, barriers to clinical translation remain. One challenge is to develop a low-power, fully-implantable system that dissipates only minimal power so as not to damage tissue. To this end, we implemented a Kalman-filter based decoder via a spiking neural network (SNN) and tested it in brain-machine interface (BMI) experiments with a rhesus monkey. The Kalman filter was trained to predict the arm's velocity and mapped on to the SNN using the Neural Engineering Framework (NEF). A 2,000-neuron embedded Matlab SNN implementation runs in real-time and its closed-loop performance is quite comparable to that of the standard Kalman filter. The success of this closed-loop decoder holds promise for hardware SNN implementations of statistical signal processing algorithms on neuromorphic chips, which may offer power savings necessary to overcome a major obstacle to the successful clinical translation of neural motor prostheses.

Subject Area: Neuroscience\Brain-computer Interfaces

M14 Energetically Optimal Action Potentials

M. Stemmler stemmler@bio.lmu.de
LMU Munich
B. Sengupta bs393@cam.ac.uk
S. Laughlin sl104@cam.ac.uk
Cambridge University
J. Niven nivenje@googlemail.com
University of Sussex

Most action potentials in the nervous system take on the form of strong, rapid, and brief voltage deflections known as spikes, in stark contrast to other action potentials, such as in the heart, that are characterized by broad voltage plateaus. We derive the shape of the neuronal action potential from first principles, by postulating that action potential generation is strongly constrained by the brain's need to minimize energy expenditure. For a given height of an action potential, the least energy is consumed when the underlying currents obey the bang-bang principle: the currents giving rise to the spike should be intense, yet short-lived, yielding spikes with sharp onsets and offsets. Energy optimality predicts features in the biophysics that are not per se required for producing the characteristic neuronal action potential: sodium currents should be extraordinarily powerful and inactivate with voltage; both potassium and sodium currents should have kinetics that have a bell-shaped voltage-dependence; and the cooperative action of multiple 'gates' should start the flow of current.

Subject Area: Neuroscience\Computational Neural Models

M15 Active dendrites: adaptation to spike-based communication

B. Ujfalussy bbu20@cam.ac.uk
 M. Lengyel m.lengyel@eng.cam.ac.uk
 University of Cambridge

Computational analyses of dendritic computations often assume stationary inputs to neurons, ignoring the pulsatile nature of spike-based communication between neurons and the moment-to-moment fluctuations caused by such spiking inputs. Conversely, circuit computations with spiking neurons are usually formalized without regard to the rich nonlinear nature of dendritic processing. Here we address the computational challenge faced by neurons that compute and represent analogue quantities but communicate with digital spikes, and show that reliable computation of even purely linear functions of inputs can require the interplay of strongly nonlinear subunits within the postsynaptic dendritic tree. Our theory predicts a matching of dendritic nonlinearities and synaptic weight distributions to the joint statistics of presynaptic inputs. This approach suggests normative roles for some puzzling forms of nonlinear dendritic dynamics and plasticity.

Subject Area: Neuroscience\Computational Neural Models

M16 Inferring spike-timing-dependent plasticity from spike train data

I. Stevenson i-stevenson@berkeley.edu
 University of California, Berkeley
 K. Koerding kk@northwestern.edu

Synaptic plasticity underlies learning and is thus central for development, memory, and recovery from injury. However, it is often difficult to detect changes in synaptic strength in vivo, since intracellular recordings are experimentally challenging. Here we present two methods aimed at inferring changes in the coupling between pairs of neurons from extracellularly recorded spike trains. First, using a generalized bilinear model with Poisson output we estimate time-varying coupling assuming that all changes are spike-timing-dependent. This approach allows model-based estimation of STDP modification functions from pairs of spike trains. Then, using recursive point-process adaptive filtering methods we estimate more general variation in coupling strength over time. Using simulations of neurons undergoing spike-timing dependent modification, we show that the true modification function can be recovered. Using multi-electrode data from motor cortex we then illustrate the use of this technique on in vivo data.

Subject Area: Neuroscience\Computational Neural Models

M17 Dynamical segmentation of single trials from population neural data

B. Petreska biljana@gatsby.ucl.ac.uk
 M. Sahani maneesh@gatsby.ucl.ac.uk
 Gatsby Unit, UCL
 B. Yu byronyu@cmu.edu
 Carnegie Mellon University
 J. Cunningham jcunnin@stanford.edu
 S. Ryu seoulman@stanford.edu
 K. Shenoy shenoy@stanford.edu
 Stanford University
 G. Santhanam gopal@nerur.com

Simultaneous recordings of many neurons embedded within a recurrently-connected cortical network may provide concurrent views into the dynamical processes of that network, and thus its computational function. In principle, these dynamics might be identified by purely unsupervised, statistical means. Here, we show that a Hidden Switching Linear Dynamical Systems (HSLDS) model—in which multiple linear dynamical laws approximate a nonlinear and potentially non-stationary dynamical process—is able to distinguish different dynamical regimes within single-trial motor cortical activity associated with the preparation and initiation of hand movements. The regimes are identified without reference to behavioural or experimental epochs, but nonetheless transitions between them correlate strongly with external events whose timing may vary from trial to trial. The HSLDS model also performs better than recent comparable models in predicting the firing rate of an isolated neuron based on the firing rates of others, suggesting that it captures more of the “shared variance” of the data. Thus, the method is able to trace the dynamical processes underlying the coordinated evolution of network activity in a way that appears to reflect its computational role.”

Subject Area: Neuroscience\Computational Neural Models

M18 Emergence of Multiplication in a Biophysical Model of a Wide-Field Visual Neuron for Computing Object Approaches: Dynamics, Peaks, & Fits

M. Keil matskeil@ub.edu
 University of Barcelona

Many species show avoidance reactions in response to looming object approaches. In locusts, the corresponding escape behavior correlates with the activity of the lobula giant movement detector (LGMD) neuron. During an object approach, its firing rate was reported to gradually increase until a peak is reached, and then it declines quickly. The η -function predicts that the LGMD activity is a product between an exponential function of angular size $\exp(-\Theta)$ and angular velocity $\dot{\Theta}$, and that peak activity is reached before time-to-contact (ttc). The η -function has become the prevailing LGMD model because it reproduces many experimental observations, and even experimental evidence for the multiplicative operation was reported. Several inconsistencies remain unresolved, though. Here we address these issues with a new model (ψ -model), which explicitly connects Θ and $\dot{\Theta}$ to biophysical quantities. The ψ -model avoids biophysical problems associated with implementing $\exp(\cdot)$, implements the multiplicative operation of η via divisive inhibition, and explains why activity peaks could occur after ttc. It consistently predicts response

features of the LGMD, and provides excellent fits to published experimental data, with goodness of fit measures comparable to corresponding fits with the η -function.
Subject Area: Neuroscience

M19 Why The Brain Separates Face Recognition From Object Recognition

J. Leibo jzleibo@mit.edu
J. Mutch jmutch@mit.edu
T. Poggio tp@csail.mit.edu
Massachusetts Institute of Technology

Many studies have uncovered evidence that visual cortex contains specialized regions involved in processing faces but not other object classes. Recent electrophysiology studies of cells in several of these specialized regions revealed that at least some of these regions are organized in a hierarchical manner with viewpoint-specific cells projecting to downstream viewpoint-invariant identity-specific cells (Freiwald and Tsao 2010). A separate computational line of reasoning leads to the claim that some transformations of visual inputs that preserve viewed object identity are class-specific. In particular, the 2D images evoked by a face undergoing a 3D rotation are not produced by the same image transformation (2D) that would produce the images evoked by an object of another class undergoing the same 3D rotation. However, within the class of faces, knowledge of the image transformation evoked by 3D rotation can be reliably transferred from previously viewed faces to help identify a novel face at a new viewpoint. We show, through computational simulations, that an architecture which applies this method of gaining invariance to class-specific transformations is effective when restricted to faces and fails spectacularly when applied across object classes. We argue here that in order to accomplish viewpoint-invariant face identification from a single example view, visual cortex must separate the circuitry involved in discounting 3D rotations of faces from the generic circuitry involved in processing other objects. The resulting model of the ventral stream of visual cortex is consistent with the recent physiology results showing the hierarchical organization of the face processing network.
Subject Area: Neuroscience\Computational Neural Models

M20 Estimating time-varying input signals and ion channel states from a single voltage trace of a neuron

R. Kobayashi kobayashi@cns.ci.ritsumei.ac.jp
Ritsumeikan University
Y. Tsubo yasuhirotsubo@riken.jp
RIKEN
P. Lansky lansky@biomed.cas.cz
Academy of Sciences
S. Shinomoto shinomoto@scphys.kyoto-u.ac.jp
Kyoto University

State-of-the-art statistical methods in neuroscience have enabled us to fit mathematical models to experimental data and subsequently to infer the dynamics of hidden parameters underlying the observable phenomena. Here, we develop a Bayesian method for inferring the time-varying mean and variance of the synaptic input, along with the dynamics of each ion channel from a single voltage trace of a neuron. An estimation problem may be formulated on the basis of

the state-space model with prior distributions that penalize large fluctuations in these parameters. After optimizing the hyperparameters by maximizing the marginal likelihood, the state-space model provides the time-varying parameters of the input signals and the ion channel states. The proposed method is tested not only on the simulated data from the Hodgkin-Huxley type models but also on experimental data obtained from a cortical slice *in vitro*.
Subject Area: Neuroscience\Neural Coding

M21 How biased are maximum entropy models?

J. Macke jakob@gatsby.ucl.ac.uk
P. Latham pel@gatsby.ucl.ac.uk
University College London
I. Murray i.murray@ed.ac.uk
University of Edinburgh

Maximum entropy models have become popular statistical models in neuroscience and other areas in biology, and can be useful tools for obtaining estimates of mutual information in biological systems. However, maximum entropy models fit to small data sets can be subject to sampling bias; i.e. the true entropy of the data can be severely underestimated. Here we study the sampling properties of estimates of the entropy obtained from maximum entropy models. We show that if the data is generated by a distribution that lies in the model class, the bias is equal to the number of parameters divided by twice the number of observations. However, in practice, the true distribution is usually outside the model class, and we show here that this misspecification can lead to much larger bias. We provide a perturbative approximation of the maximally expected bias when the true model is out of model class, and we illustrate our results using numerical simulations of an Ising model; i.e. the second-order maximum entropy distribution on binary data.
Subject Area: Neuroscience\Neural Coding

M22 Joint 3D Estimation of Objects and Scene Layout

A. Geiger geiger@kit.edu
KIT
C. Wojek cwojek@mpi-inf.mpg.de
R. Urtasun rurtasun@ttic.edu
TTI-Chicago

We propose a novel generative model that is able to reason jointly about the 3D scene layout as well as the 3D location and orientation of objects in the scene. In particular, we infer the scene topology, geometry as well as traffic activities from a short video sequence acquired with a single camera mounted on a moving car. Our generative model takes advantage of dynamic information in the form of vehicle tracklets as well as static information coming from semantic labels and geometry (i.e., vanishing points). Experiments show that our approach outperforms a discriminative baseline based on multiple kernel learning (MKL) which has access to the same image information. Furthermore, as we reason about objects in 3D, we are able to significantly increase the performance of state-of-the-art object detectors in their ability to estimate object orientation.
Subject Area: Vision

M23 Pylon Model for Semantic Segmentation

V. Lempitsky victorlempitsky@gmail.com
 Yandex
 A. Vedaldi vedaldi@robots.ox.ac.uk
 A. Zisserman az@robots.ox.ac.uk
 University of Oxford

Graph cut optimization is one of the standard workhorses of image segmentation since for binary random field representations of the image, it gives globally optimal results and there are efficient polynomial time implementations. Often, the random field is applied over a flat partitioning of the image into non-intersecting elements, such as pixels or super-pixels. In the paper we show that if, instead of a flat partitioning, the image is represented by a hierarchical segmentation tree, then the resulting energy combining unary and boundary terms can still be optimized using graph cut (with all the corresponding benefits of global optimality and efficiency). As a result of such inference, the image gets partitioned into a set of segments that may come from different layers of the tree.

We apply this formulation, which we call the *pylon model*, to the task of semantic segmentation where the goal is to separate an image into areas belonging to different semantic classes. The experiments highlight the advantage of inference on a segmentation tree (over a flat partitioning) and demonstrate that the optimization in the pylon model is able to flexibly choose the level of segmentation across the image. Overall, the proposed system has superior segmentation accuracy on several datasets (Graz-02, Stanford background) compared to previously suggested approaches.

Subject Area: Vision\Image Segmentation

M24 Higher-Order Correlation Clustering for Image Segmentation

S. Kim sungwoong.kim01@gmail.com
 C. Yoo cdyoo@ee.kaist.ac.kr
 KAIST
 S. Nowozin senowozi@microsoft.com
 P. Kohli pkohli@microsoft.com
 Microsoft Research

For many of the state-of-the-art computer vision algorithms, image segmentation is an important preprocessing step. As such, several image segmentation algorithms have been proposed, however, with certain reservation due to high computational load and many hand-tuning parameters. Correlation clustering, a graph-partitioning algorithm often used in natural language processing and document clustering, has the potential to perform better than previously proposed image segmentation algorithms. We improve the basic correlation clustering formulation by taking into account higher-order cluster relationships. This improves clustering in the presence of local boundary ambiguities. We first apply the pairwise correlation clustering to image segmentation over a pairwise superpixel graph and then develop higher-order correlation clustering over a hypergraph that considers higher-order relations among superpixels. Fast inference is possible by linear programming relaxation, and also effective parameter learning framework by structured

support vector machine is possible. Experimental results on various datasets show that the proposed higher-order correlation clustering outperforms other state-of-the-art image segmentation algorithms.

Subject Area: Vision\Image Segmentation

M25 Unsupervised learning models of primary cortical receptive fields and receptive field plasticity

A. Saxe asaxe@stanford.edu
 M. Bhand mbhand@cs.stanford.edu
 R. Mudur rmudur@stanford.edu
 B. Suresh bipins@cs.stanford.edu
 A. Ng ang@cs.stanford.edu
 Stanford University

The efficient coding hypothesis holds that neural receptive fields are adapted to the statistics of the environment, but is agnostic to the timescale of this adaptation, which occurs on both evolutionary and developmental timescales. In this work we focus on that component of adaptation which occurs during an organism's lifetime, and show that a number of unsupervised feature learning algorithms can account for features of normal receptive field properties across multiple primary sensory cortices. Furthermore, we show that the same algorithms account for altered receptive field properties in response to experimentally altered environmental statistics. Based on these modeling results we propose these models as phenomenological models of receptive field plasticity during an organism's lifetime. Finally, due to the success of the same models in multiple sensory areas, we suggest that these algorithms may provide a constructive realization of the theory, first proposed by Mountcastle (1978), that a qualitatively similar learning algorithm acts throughout primary sensory cortices.

Subject Area: Vision\Natural Scene Statistics

M26 Transfer Learning by Borrowing Examples for Multiclass Object Detection

J. Lim lim@csail.mit.edu
 A. Torralba torralba@csail.mit.edu
 Massachusetts Institute of Technology
 R. Salakhutdinov rsalakh@utstat.toronto.edu
 University of Toronto

Despite the recent trend of increasingly large datasets for object detection, there still exist many classes with few training examples. To overcome this lack of training data for certain classes, we propose a novel way of augmenting the training data for each class by borrowing and transforming examples from other classes. Our model learns which training instances from other classes to borrow and how to transform the borrowed examples so that they become more similar to instances from the target class. Our experimental results demonstrate that our new object detector, with borrowed and transformed examples, improves upon the current state-of-the-art detector on the challenging SUN09 object detection dataset.

Subject Area: Vision\Object Recognition

M27 Large-Scale Category Structure Aware Image Categorization

B. Zhao zhaobinhere@hotmail.com
 E. Xing epxing@cs.cmu.edu
 Carnegie Mellon University
 F. Li feifeili@cs.stanford.edu
 Stanford University

Most previous research on image categorization has focused on medium-scale data sets, while large-scale image categorization with millions of images from thousands of categories remains a challenge. With the emergence of structured large-scale dataset such as the ImageNet, rich information about the conceptual relationships between images, such as a tree hierarchy among various image categories, become available. As human cognition of complex visual world benefits from underlying semantic relationships between object classes, we believe a machine learning system can and should leverage such information as well for better performance. In this paper, we employ such semantic relatedness among image categories for large-scale image categorization. Specifically, a category hierarchy is utilized to properly define loss function and select common set of features for related categories. An efficient optimization method based on proximal approximation and accelerated parallel gradient method is introduced. Experimental results on a subset of ImageNet containing 1.2 million images from 1000 categories demonstrate the effectiveness and promise of our proposed approach.

Subject Area: Vision\Object Recognition

M28 Hierarchical Matching Pursuit for Image Classification: Architecture and Fast Algorithms

L. Bo lfb@cs.washington.edu
 D. Fox fox@cs.washington.edu
 University of Washington
 X. Ren xiaofeng.ren@intel.com
 Intel Labs Seattle

Extracting good representations from images is essential for many computer vision tasks. In this paper, we propose hierarchical matching pursuit (HMP), which builds a feature hierarchy layer-by-layer using an efficient matching pursuit encoder. It includes three modules: batch (tree) orthogonal matching pursuit, spatial pyramid max pooling, and contrast normalization. We investigate the architecture of HMP, and show that all three components are critical for good performance. To speed up the orthogonal matching pursuit, we propose a batch tree orthogonal matching pursuit that is particularly suitable to encode a large number of observations that share the same large dictionary. HMP is scalable and can efficiently handle full-size images. In addition, HMP enables linear support vector machines (SVMs) to match the performance of nonlinear SVMs while being scalable to large datasets. We compare HMP with many state-of-the-art algorithms including convolutional deep belief networks, SIFT based single layer sparse coding, and kernel based feature learning. HMP consistently yields superior accuracy on three types of visual recognition problems: object recognition (Caltech-101), scene recognition (MIT-Scene), and static event recognition (UIUC-Sports).

Subject Area: Vision\Object Recognition

M29 Portmanteau Vocabularies for Multi-Cue Image Representation

F. Khan fahad@cvc.uab.es
 M. Vanrell maria@cvc.uab.es
 J. van de Weijer Joost@cvc.uab.es
 Computer Vision Center
 A. Bagdanov bagdanov@dsi.unifi.it
 University of Florence.

We describe a novel technique for feature combination in the bag-of-words model of image classification. Our approach builds discriminative compound words from primitive cues learned independently from training images. Our main observation is that modeling joint-cue distributions independently is more statistically robust for typical classification problems than attempting to empirically estimate the dependent, joint-cue distribution directly. We use Information theoretic vocabulary compression to find discriminative combinations of cues and the resulting vocabulary of *portmanteau*¹ words is compact, has the cue binding property, and supports individual weighting of cues in the final image representation. State-of-the-art results on both the Oxford Flower-102 and Caltech-UCSD Bird-200 datasets demonstrate the effectiveness of our technique compared to other, significantly more complex approaches to multi-cue image representation

Subject Area: Vision\Object Recognition

M30 PiCoDes: Learning a Compact Code for Novel-Category Recognition

A. Bergamo aleb@cs.dartmouth.edu
 L. Torresani lorenzo@cs.dartmouth.edu
 Dartmouth
 A. Fitzgibbon awf@microsoft.com
 Microsoft Research

We introduce PiCoDes: a very compact image descriptor which nevertheless allows high performance on object category recognition. In particular, we address novel-category recognition: the task of defining indexing structures and image representations which enable a large collection of images to be searched for an object category that was not known when the index was built. Instead, the training images defining the category are supplied at query time. We explicitly learn descriptors of a given length (from as small as 16 bytes per image) which have good object-recognition performance. In contrast to previous work in the domain of object recognition, we do not choose an arbitrary intermediate representation, but explicitly learn short codes. In contrast to previous approaches to learn compact codes, we optimize explicitly for (an upper bound on) classification performance. Optimization directly for binary features is difficult and nonconvex, but we present an alternation scheme and convex upper bound which demonstrate excellent performance in practice. PiCoDes of 256 bytes match the accuracy of the current best known classifier for the Caltech256 benchmark, but they decrease the database storage size by a factor of 100 and speed-up the training and testing of novel classes by orders of magnitude.

Subject Area: Vision\Visual Features

M31 Orthogonal Matching Pursuit with Replacement

P. Jain prajain@microsoft.com
 A. Tewari ambujtewari@gmail.com
 I. Dhillon inderjit@cs.utexas.edu
 University of Texas at Austin

In this paper, we consider the problem of compressed sensing where the goal is to recover almost all the sparse vectors using a small number of *fixed* linear measurements. For this problem, we propose a novel partial hard-thresholding operator leading to a general family of iterative algorithms. While one extreme of the family yields well known hard thresholding algorithms like ITI and HTP[17, 10], the other end of the spectrum leads to a novel algorithm that we call Orthogonal Matching Pursuit with Replacement (OMPR). OMPR, like the classic greedy algorithm OMP, adds exactly one coordinate to the support at each iteration, based on the correlation with the current residual. However, unlike OMP, OMPR also removes one coordinate from the support. This simple change allows us to prove the best known guarantees for OMPR in terms of the Restricted Isometry Property (a condition on the measurement matrix). In contrast, OMP is known to have very weak performance guarantees under RIP. We also extend OMPR using locality sensitive hashing to get OMPR-Hash, the first provably sub-linear (in dimensionality) algorithm for sparse recovery. Our proof techniques are novel and flexible enough to also permit the tightest known analysis of popular iterative algorithms such as CoSaMP and Subspace Pursuit. We provide experimental results on large problems providing recovery for vectors of size up to million dimensions. We demonstrate that for large-scale problems our proposed methods are more robust and faster than the existing methods.

Subject Area: Speech and Signal Processing

M32 SpaRCS: Recovering low-rank and sparse matrices from compressive measurements

A. Waters andrew.e.waters@rice.edu
 A. Sankaranarayanan saswin@rice.edu
 R. Baraniuk richb@rice.edu
 Rice University

We consider the problem of recovering a matrix M that is the sum of a low-rank matrix L and a sparse matrix S from a small set of linear measurements of the form $y = A(M) = A(L + S)$. This model subsumes three important classes of signal recovery problems: compressive sensing, affine rank minimization, and robust principal component analysis. We propose a natural optimization problem for signal recovery under this model and develop a new greedy algorithm called SpaRCS to solve it. SpaRCS inherits a number of desirable properties from the state-of-the-art CoSaMP and ADMiRA algorithms, including exponential convergence and efficient implementation. Simulation results with video compressive sensing, hyperspectral imaging, and robust matrix completion data sets demonstrate both the accuracy and efficacy of the algorithm.

Subject Area: Speech and Signal Processing

M33 Signal Estimation Under Random Time-Warpings and Nonlinear Signal Alignment

S. Kurtek skurtek@stat.fsu.edu
 A. Srivastava anuj@stat.fsu.edu
 W. Wu wwu@stat.fsu.edu
 Florida State University

While signal estimation under random amplitudes, phase shifts, and additive noise is studied frequently, the problem of estimating a deterministic signal under random time-warpings has been relatively unexplored. We present a novel framework for estimating the unknown signal that utilizes the action of the warping group to form an equivalence relation between signals. First, we derive an estimator for the equivalence class of the unknown signal using the notion of Karcher mean on the quotient space of equivalence classes. This step requires the use of Fisher-Rao Riemannian metric and a square-root representation of signals to enable computations of distances and means under this metric. Then, we define a notion of the center of a class and show that the center of the estimated class is a consistent estimator of the underlying unknown signal. This estimation algorithm has many applications: (1) registration/alignment of functional data, (2) separation of phase/amplitude components of functional data, (3) joint demodulation and carrier estimation, and (4) sparse modeling of functional data. Here we demonstrate only (1) and (2): Given signals are temporally aligned using nonlinear warpings and, thus, separated into their phase and amplitude components. The proposed method for signal alignment is shown to have state of the art performance using Berkeley growth, handwritten signatures, and neuroscience spike train data.

Subject Area: Speech and Signal Processing

M34 Inverting Grice's Maxims to Learn Rules from Natural Language Extractions

M. Sorower ssorower@gmail.com
 T. Dietterich tgdt@cs.orst.edu
 J. Doppa doppa@eecs.oregonstate.edu
 W. Orr orr@eecs.oregonstate.edu
 P. Tadepalli tadepall@eecs.oregonstate.edu
 X. Fern xfern@eecs.oregonstate.edu
 Oregon State University

We consider the problem of learning rules from natural language text sources. These sources, such as news articles and web texts, are created by a writer to communicate information to a reader, where the writer and reader share substantial domain knowledge. Consequently, the texts tend to be concise and mention the minimum information necessary for the reader to draw the correct conclusions. We study the problem of learning domain knowledge from such concise texts, which is an instance of the general problem of learning in the presence of missing data. However, unlike standard approaches to missing data, in this setting we know that facts are more likely to be missing from the text in cases where the reader can infer them from the facts that are mentioned combined with the domain knowledge. Hence, we can explicitly model this "missingness" process and invert it via probabilistic inference to learn the underlying domain knowledge. This paper introduces a *mention model* that models the probability of facts being mentioned in the text based on

what other facts have already been mentioned and domain knowledge in the form of Horn clause rules. Learning must simultaneously search the space of rules and learn the parameters of the mention model. We accomplish this via an application of Expectation Maximization within a Markov Logic framework. An experimental evaluation on synthetic and natural text data shows that the method can learn accurate rules and apply them to new texts to make correct inferences. Experiments also show that the method out-performs the standard EM approach that assumes mentions are missing at random.”

Subject Area: Applications\Natural Language Processing

M35 Multi-View Learning of Word Embeddings via CCA

P. Dhillon dhillon@cis.upenn.edu
 D. Foster foster@wharton.upenn.edu
 L. Ungar ungar@cis.upenn.edu
 University of Pennsylvania

Recently, there has been substantial interest in using large amounts of unlabeled data to learn word representations which can then be used as features in supervised classifiers for NLP tasks. However, most current approaches are slow to train, do not model context of the word, and lack theoretical grounding. In this paper, we present a new learning method, Low Rank Multi-View Learning (LR-MVL) which uses a fast spectral method to estimate low dimensional context-specific word representations from unlabeled data. These representation features can then be used with any supervised learner. LR-MVL is extremely fast, gives guaranteed convergence to a global optimum, is theoretically elegant, and achieves state-of-the-art performance on named entity recognition (NER) and chunking problems.

Subject Area: Applications\Natural Language Processing

M36 Active Ranking using Pairwise Comparisons

K. Jamieson kgjamieson@wisc.edu
 R. Nowak nowak@ece.wisc.edu
 University of Wisconsin-Madison

This paper examines the problem of ranking a collection of objects using pairwise comparisons (rankings of two objects). In general, the ranking of n objects can be identified by standard sorting methods using $n \log n$ pairwise comparisons. We are interested in natural situations in which relationships among the objects may allow for ranking using far fewer pairwise comparisons. {Specifically, we assume that the objects can be embedded into a d -dimensional Euclidean space and that the rankings reflect their relative distances from a common reference point in R^d . We show that under this assumption the number of possible rankings grows like n^{2d} and demonstrate an algorithm that can identify a randomly selected ranking using just slightly more than $d \log n$ adaptively selected pairwise comparisons, on average.} If instead the comparisons are chosen at random, then almost all pairwise comparisons must be made in order to identify any ranking. In addition, we propose a robust, error-tolerant algorithm that only requires that the pairwise comparisons are probably correct. Experimental studies with synthetic and real datasets support the conclusions of our theoretical analysis.

Subject Area: Applications\Ranking

M37 Co-Training for Domain Adaptation

M. Chen mc15@cse.wustl.edu
 K. Weinberger kilian@wustl.edu
 Washington University
 J. Blitzer blitzer@google.com
 Google Research

Domain adaptation algorithms seek to generalize a model trained in a *source* domain to a new *target* domain. In many practical cases, the source and target distributions can differ substantially, and in some cases crucial target features may not have support in the source domain. In this paper we introduce an algorithm that bridges the gap between source and target domains by slowly adding both the target features and instances in which the current algorithm is the most confident. Our algorithm is a variant of co-training, and we name it CODA (Co-training for domain adaptation). Unlike the original co-training work, we do not assume a particular feature split. Instead, for each iteration of co-training, we add target features and formulate a single optimization problem which simultaneously learns a target predictor, a split of the feature space into views, and a shared subset of source and target features to include in the predictor. CODA significantly out-performs the state-of-the-art on the 12-domain benchmark data set of Blitzer et al.. Indeed, over a wide range (65 of 84 comparisons) of target supervision, ranging from no labeled target data to a relatively large number of target labels, CODA achieves the best performance.

Subject Area: Supervised Learning

M38 Efficient anomaly detection using bipartite k -NN graphs

K. Sricharan kksreddy@umich.edu
 A. Hero hero@umich.edu
 University of Michigan

Learning minimum volume sets of an underlying nominal distribution is a very effective approach to anomaly detection. Several approaches to learning minimum volume sets have been proposed in the literature, including the K -point nearest neighbor graph (K -kNNG) algorithm based on the geometric entropy minimization (GEM) principle [4]. The K -kNNG detector, while possessing several desirable characteristics, suffers from high computation complexity, and in [4] a simpler heuristic approximation, the leave-one-out kNNG (L1O-kNNG) was proposed. In this paper, we propose a novel bipartite k -nearest neighbor graph (BP-kNNG) anomaly detection scheme for estimating minimum volume sets. Our bipartite estimator retains all the desirable theoretical properties of the K -kNNG, while being computationally simpler than the K -kNNG and the surrogate L1O-kNNG detectors. We show that BP-kNNG is asymptotically consistent in recovering the p -value of each test point. Experimental results are given that illustrate the superior performance of BP-kNNG as compared to the L1O-kNNG and other state of the art anomaly detection schemes.

Subject Area: Supervised Learning

M39 Maximum Margin Multi-Instance Learning

H. Wang	huawangcs@gmail.com
H. Huang	heng@uta.edu
F. Kamangar	kamangar@uta.edu
F. Nie	feipingnie@gmail.com
C. Ding	chqding@uta.edu
UTA	

Multi-instance learning (MIL) considers input as bags of instances, in which labels are assigned to the bags. MIL is useful in many real-world applications. For example, in image categorization semantic meanings (labels) of an image mostly arise from its regions (instances) instead of the entire image (bag). Existing MIL methods typically build their models using the *Bag-to-Bag* (B2B) distance, which are often computationally expensive and may not truly reflect the semantic similarities. To tackle this, in this paper we approach MIL problems from a new perspective using the *Class-to-Bag* (C2B) distance, which directly assesses the relationships between the classes and the bags. Taking into account the two major challenges in MIL, high heterogeneity on data and weak label association, we propose a novel Maximum Margin Multi-Instance Learning (M³I) approach to parameterize the C2B distance by introducing the class specific distance metrics and the locally adaptive significance coefficients. We apply our new approach to the automatic image categorization tasks on three (one single-label and two multi-label) benchmark data sets. Extensive experiments have demonstrated promising results that validate the proposed method.

Subject Area: Supervised Learning/Classification

M40 Advice Refinement in Knowledge-Based SVMs

G. Kunapuli	gkunapuli@gmail.com
J. Shavlik	shavlik@cs.wisc.edu
University of Wisconsin at Madison	
R. Maclin	rmaclin@d.umn.edu
University of Minnesota Duluth	

Knowledge-based support vector machines (KBSVMs) incorporate advice from domain experts, which can improve generalization significantly. A major limitation that has not been fully addressed occurs when the expert advice is imperfect, which can lead to poorer models. We propose a model that extends KBSVMs and is able to not only learn from data and advice, but also simultaneously improve the advice. The proposed approach is particularly effective for knowledge discovery in domains with few labeled examples. The proposed model contains bilinear constraints, and is solved using two iterative approaches: successive linear programming and a constrained concave-convex approach. Experimental results demonstrate that these algorithms yield useful refinements to expert advice, as well as improve the performance of the learning algorithm overall.

Subject Area: Supervised Learning/Classification

M41 Multiclass Boosting: Theory and Algorithms

M. Saberian	saberian@ucsd.edu
N. Vasconcelos	nuno@ece.ucsd.edu
UC San Diego	

The problem of multiclass boosting is considered. A new framework, based on multi-dimensional codewords and predictors is introduced. The optimal set of codewords is derived, and a margin enforcing loss proposed. The resulting risk is minimized by gradient descent on a multidimensional functional space. Two algorithms are proposed: 1) CD-MCBoost, based on coordinate descent, updates one predictor component at a time, 2) GD-MCBoost, based on gradient descent, updates all components jointly. The algorithms differ in the weak learners that they support but are both shown to be 1) Bayes consistent, 2) margin enforcing, and 3) convergent to the global minimum of the risk. They also reduce to AdaBoost when there are only two classes. Experiments show that both methods outperform previous multiclass boosting approaches on a number of datasets.

Subject Area: Supervised Learning

M42 Boosting with Maximum Adaptive Sampling

C. Dubout	charles.dubout@idiap.ch
F. Fleuret	francois.fleuret@idiap.ch
Idiap Research Institute	

Classical Boosting algorithms, such as AdaBoost, build a strong classifier without concern about the computational cost. Some applications, in particular in computer vision, may involve up to millions of training examples and features. In such contexts, the training time may become prohibitive. Several methods exist to accelerate training, typically either by sampling the features, or the examples, used to train the weak learners. Even if those methods can precisely quantify the speed improvement they deliver, they offer no guarantee of being more efficient than any other, given the same amount of time. This paper aims at shading some light on this problem, i.e. given a fixed amount of time, for a particular problem, which strategy is optimal in order to reduce the training loss the most. We apply this analysis to the design of new algorithms which estimate on the fly at every iteration the optimal trade-off between the number of samples and the number of features to look at in order to maximize the expected loss reduction. Experiments in object recognition with two standard computer vision data-sets show that the adaptive methods we propose outperform basic sampling and state-of-the-art bandit methods.

Subject Area: Supervised Learning

M43 Kernel Embeddings of Latent Tree Graphical Models

L. Song lesong@cs.cmu.edu
 A. Parikh apparikh@cs.cmu.edu
 E. Xing epxing@cs.cmu.edu
 Carnegie Mellon University

Latent tree graphical models are natural tools for expressing long range and hierarchical dependencies among many variables which are common in computer vision, bioinformatics and natural language processing problems. However, existing models are largely restricted to discrete and Gaussian variables due to computational constraints; furthermore, algorithms for estimating the latent tree structure and learning the model parameters are largely restricted to heuristic local search. We present a method based on kernel embeddings of distributions for latent tree graphical models with continuous and non-Gaussian variables. Our method can recover the latent tree structures with provable guarantees and perform local-minimum free parameter learning and efficient inference. Experiments on simulated and real data show the advantage of our proposed approach.

Subject Area: Supervised Learning\Kernel Methods

M44 Hierarchical Multitask Structured Output Learning for Large-scale Sequence Segmentation

N. Goernitz nico.goernitz@tu-berlin.de
 Technical University Berlin
 C. Widmer cwidmer@tuebingen.mpg.de
 G. Raetsch Gunnar.Raetsch@tuebingen.mpg.de
 A. Kahles andre.kahles@tuebingen.mpg.de
 Max Planck Society
 G. Zeller georg.zeller@gmail.com
 EMBL
 S. Sonnenburg soeren@sonnenburgs.de
 TomTom

We present a novel regularization-based Multitask Learning (MTL) formulation for Structured Output (SO) prediction for the case of hierarchical task relations. Structured output learning often results in difficult inference problems and requires large amounts of training data to obtain accurate models. We propose to use MTL to exploit information available for related structured output learning tasks by means of hierarchical regularization. Due to the combination of example sets, the cost of training models for structured output prediction can easily become infeasible for real world applications. We thus propose an efficient algorithm based on bundle methods to solve the optimization problems resulting from MTL structured output learning. We demonstrate the performance of our approach on gene finding problems from the application domain of computational biology. We show that 1) our proposed solver achieves much faster convergence than previous methods and 2) that the Hierarchical SO-MTL approach clearly outperforms considered non-MTL methods.

Subject Area: Supervised Learning

M45 Submodular Multi-Label Learning

J. Petterson james.petterson@nicta.com.au
 T. Caetano tiberio.caetano@nicta.com.au
 NICTA/Australian National University

In this paper we present an algorithm to learn a multi-label classifier which attempts at directly optimising the F-score. The key novelty of our formulation is that we explicitly allow for assortative (submodular) pairwise label interactions, i.e., we can leverage the co-occurrence of pairs of labels in order to improve the quality of prediction. Prediction in this model consists of minimising a particular submodular set function, what can be accomplished exactly and efficiently via graph-cuts. Learning however is substantially more involved and requires the solution of an intractable combinatorial optimisation problem. We present an approximate algorithm for this problem and prove that it is sound in the sense that it never predicts incorrect labels. We also present a nontrivial test of a sufficient condition for our algorithm to have found an optimal solution. We present experiments on benchmark multi-label datasets, which attest the value of our proposed technique. We also make available source code that enables the reproduction of our experiments.

Subject Area: Supervised Learning

M46 Algorithms for Hyper-Parameter Optimization

J. Bergstra bergstra@rowland.harvard.edu
 Harvard
 R. Bardenet bardenet@lri.fr
 B. Kégl balazs.kegl@gmail.com
 University of Paris-Sud/CNRS
 Y. Bengio bengioy@iro.umontreal.ca
 University of Montreal

Several recent advances to the state of the art in image classification benchmarks have come from better configurations of existing techniques rather than novel approaches to feature learning. Traditionally, hyper-parameter optimization has been the job of humans because they can be very efficient in regimes where only a few trials are possible. Presently, computer clusters and GPU processors make it possible to run more trials and we show that algorithmic approaches can find better results. We present hyper-parameter optimization results on tasks of training neural networks and deep belief networks (DBNs). We optimize hyper-parameters using random search and two new greedy sequential methods based on the expected improvement criterion. Random search has been shown to be sufficiently efficient for learning neural networks for several datasets, but we show it is unreliable for training DBNs. The sequential algorithms are applied to the most difficult DBN learning problems from [Larochelle et al., 2007] and find significantly better results than the best previously reported. This work contributes novel techniques for making response surface models $P(x|y)$ in which many elements of hyper-parameter assignment (x) are known to be irrelevant given particular values of other elements.

Subject Area: Algorithms and Architectures

M47 Non-parametric Group Orthogonal Matching Pursuit for Sparse Learning with Multiple Kernels

V. Sindhwani vikas.sindhwani@gmail.com
 A. Lozano aclozano@us.ibm.com
 IBM T.J. Watson Research Center

We consider regularized risk minimization in a large dictionary of Reproducing kernel Hilbert Spaces (RKHSs) over which the target function has a sparse representation. This setting, commonly referred to as Sparse Multiple Kernel Learning (MKL), may be viewed as the non-parametric extension of group sparsity in linear models. While the two dominant algorithmic strands of sparse learning, namely convex relaxations using l_1 norm (e.g., Lasso) and greedy methods (e.g., OMP), have both been rigorously extended for group sparsity, the sparse MKL literature has so far mainly adopted the former with mild empirical success. In this paper, we close this gap by proposing a Group-OMP based framework for sparse multiple kernel learning. Unlike l_1 -MKL, our approach decouples the sparsity regularizer (via a direct l_0 constraint) from the smoothness regularizer (via RKHS norms) which leads to better empirical performance as well as a simpler optimization procedure that only requires a black-box single-kernel solver. The algorithmic development and empirical studies are complemented by theoretical analyses in terms of Rademacher generalization bounds and sparse recovery conditions analogous to those for OMP [27] and Group-OMP [16].

Subject Area: Supervised Learning

M48 Manifold Precs: An Annealing Technique for Diverse Sampling of Manifolds

N. Shroff nshroff@umiacs.umd.edu
 P. Turaga pturaga@umiacs.umd.edu
 R. Chellappa Rama@umiacs.umd.edu
 University of Maryland College Park

In this paper, we consider the ‘Precis’ problem of sampling K representative yet diverse data points from a large dataset. This problem arises frequently in applications such as video and document summarization, exploratory data analysis, and pre-filtering. We formulate a general theory which encompasses not just traditional techniques devised for vector spaces, but also non-Euclidean manifolds, thereby enabling these techniques to shapes, human activities, textures and many other image and video based datasets. We propose intrinsic manifold measures for measuring the quality of a selection of points with respect to their representative power, and their diversity. We then propose efficient algorithms to optimize the cost function using a novel annealing-based iterative alternation algorithm. The proposed formulation is applicable to manifolds of known geometry as well as to manifolds whose geometry needs to be estimated from samples. Experimental results show the strength and generality of the proposed approach.

Subject Area: Unsupervised & Semi-supervised Learning

M49 Group Anomaly Detection using Flexible Genre Models

L. Xiong lxiong@cs.cmu.edu
 B. Póczos bapoczos@cs.cmu.edu
 J. Schneider schneide@cs.cmu.edu
 Carnegie Mellon University

An important task in exploring and analyzing real-world data sets is to detect unusual and interesting phenomena. In this paper, we study the group anomaly detection problem. Unlike traditional anomaly detection research that focuses on data points, our goal is to discover anomalous aggregated behaviors of groups of points. For this purpose, we propose the *Flexible Genre Model* (FGM). FGM is designed to characterize data groups at both the point level and the group level so as to detect various types of group anomalies. We evaluate the effectiveness of FGM on both synthetic and real data sets including images and turbulence data, and show that it is superior to existing approaches in detecting group anomalies.

Subject Area: Unsupervised & Semi-supervised Learning

M50 Matrix Completion for Multi-label Image Classification

R. Cabral rscabral@cmu.edu
 J. Costeira jpc@isr.ist.utl.pt
 A. Bernardino alex@isr.ist.utl.pt
 ISR/ Instituto Superior Tecnico
 F. De la Torre fforre@cs.cmu.edu
 Carnegie Mellon University

Recently, image categorization has been an active research topic due to the urgent need to retrieve and browse digital images via semantic keywords. This paper formulates image categorization as a multi-label classification problem using recent advances in matrix completion. Under this setting, classification of testing data is posed as a problem of completing unknown label entries on a data matrix that concatenates training and testing features with training labels. We propose two convex algorithms for matrix completion based on a Rank Minimization criterion specifically tailored to visual data, and prove its convergence properties. A major advantage of our approach w.r.t. standard discriminative classification methods for image categorization is its robustness to outliers, background noise and partial occlusions both in the feature and label space. Experimental validation on several datasets shows how our method outperforms state-of-the-art algorithms, while effectively capturing semantic concepts of classes.

Subject Area: Unsupervised & Semi-supervised Learning

M51 Selecting Receptive Fields in Deep Networks

A. Coates acoates@cs.stanford.edu
 A. Ng ang@cs.stanford.edu
 Stanford University

Recent deep learning and unsupervised feature learning systems that learn from unlabeled data have achieved high performance in benchmarks by using extremely large architectures with many features (hidden units) at each layer. Unfortunately, for such large architectures the number of parameters usually grows quadratically in the width of the network, thus necessitating hand-coded “local receptive fields” that limit the number of connections from lower level features to higher ones (e.g., based on spatial locality). In this paper we propose a fast method to choose these connections that may be incorporated into a wide variety of unsupervised training methods. Specifically, we choose local receptive fields that group together those low-level features that are most similar to each other according to a pairwise similarity metric. This approach allows us to harness the advantages of local receptive fields (such as improved scalability, and reduced data requirements) when we do not know how to specify such receptive fields by hand or where our unsupervised training algorithm has no obvious generalization to a topographic setting. We produce results showing how this method allows us to use even simple unsupervised training algorithms to train successful multi-layered networks that achieve state-of-the-art results on CIFAR and STL datasets: 82.0% and 60.1% accuracy, respectively.”

Subject Area: Unsupervised & Semi-supervised Learning

M52 Co-regularized Multi-view Spectral Clustering

A. Kumar abhishek@cs.umd.edu
 H. Daume III me@hal3.name
 University of Maryland
 P. Rai piyush@cs.utah.edu
 University of Utah

In many clustering problems, we have access to multiple views of the data each of which could be individually used for clustering. Exploiting information from multiple views, one can hope to find a clustering that is more accurate than the ones obtained using the individual views. Since the true clustering would assign a point to the same cluster irrespective of the view, we can approach this problem by looking for clusterings that are consistent across the views, i.e., corresponding data points in each view should have same cluster membership. We propose a spectral clustering framework that achieves this goal by co-regularizing the clustering hypotheses, and propose two co-regularization schemes to accomplish this. Experimental comparisons with a number of baselines on two synthetic and three real-world datasets establish the efficacy of our proposed approaches.

Subject Area: Unsupervised & Semi-supervised Learning

M53 Learning with the weighted trace-norm under arbitrary sampling distributions

R. Foygel rina@uchicago.edu
 University of Chicago
 R. Salakhutdinov rsalakhu@utstat.toronto.edu
 University of Toronto
 O. Shamir ohadsh@microsoft.com
 Microsoft Research
 N. Srebro nati@ttic.edu
 TTI-Chicago

We provide rigorous guarantees on learning with the weighted trace-norm under arbitrary sampling distributions. We show that the standard weighted-trace norm might fail when the sampling distribution is not a product distribution (i.e. when row and column indexes are not selected independently), present a corrected variant for which we establish strong learning guarantees, and demonstrate that it works better in practice. We provide guarantees when weighting by either the true or empirical sampling distribution, and suggest that even if the true distribution is known (or is uniform), weighting by the empirical distribution may be beneficial.

Subject Area: Unsupervised & Semi-supervised Learning

M54 Active Learning with a Drifting Distribution

L. Yang liuy@cs.cmu.edu
 Carnegie Mellon University

We study the problem of active learning in a stream-based setting, allowing the distribution of the examples to change over time. We prove upper bounds on the number of prediction mistakes and number of label requests for established disagreement-based active learning algorithms, both in the realizable case and under Tsybakov noise. We further prove minimax lower bounds for this problem.

Subject Area: Unsupervised & Semi-supervised Learning

M55 Bayesian Partitioning of Large-Scale Distance Data

D. Adametz david.adametz@unibas.ch
 V. Roth volker.roth@unibas.ch
 University of Basel

A Bayesian approach to partitioning distance matrices is presented. It is inspired by the ‘Translation-Invariant Wishart-Dirichlet’ process (TIWD) in (Vogt et al., 2010) and shares a number of advantageous properties like the fully probabilistic nature of the inference model, automatic selection of the number of clusters and applicability in semi-supervised settings. In addition, our method (which we call *fastTIWD*) overcomes the main shortcoming of the original TIWD, namely its high computational costs. The *fastTIWD* reduces the workload in each iteration of a Gibbs sampler from $O(n^3)$ in the TIWD to $O(n^2)$. Our experiments show that this cost reduction does not compromise the quality of the inferred partitions. With this new method it is now possible to ‘mine’ large relational datasets with a probabilistic model, thereby automatically detecting new and potentially interesting clusters.

Subject Area: Unsupervised & Semi-supervised Learning

M56 Beyond Spectral Clustering - Tight Relaxations of Balanced Graph Cuts

M. Hein hein@cs.uni-saarland.de
 S. Setzer setzer@mia.uni-saarland.de
 Saarland University

Spectral clustering is based on the spectral relaxation of the normalized/ratio graph cut criterion. While the spectral relaxation is known to be loose, it has been shown recently that a non-linear eigenproblem yields a tight relaxation of the Cheeger cut. In this paper, we extend this result considerably by providing a characterization of all balanced graph cuts which allow for a tight relaxation. Although the resulting optimization problems are non-convex and non-smooth, we provide an efficient first-order scheme which scales to large graphs. Moreover, our approach comes with the quality guarantee that given any partition as initialization the algorithm either outputs a better partition or it stops immediately.

Subject Area: Unsupervised & Semi-supervised Learning

M57 Structural equations and divisive normalization for energy-dependent component analysis

J. Hirayama hirayama@robot.kuass.kyoto-u.ac.jp
 Kyoto University
 A. Hyvarinen aapo.hyvarinen@helsinki.fi

Components estimated by independent component analysis and related methods are typically not independent in real data. A very common form of nonlinear dependency between the components is correlations in their variances or energies. Here, we propose a principled probabilistic model to model the energy-correlations between the latent variables. Our two-stage model includes a linear mixing of latent signals into the observed ones like in ICA. The main new feature is a model of the energy-correlations based on the structural equation model (SEM), in particular, a Linear Non-Gaussian SEM. The SEM is closely related to divisive normalization which effectively reduces energy correlation. Our new two-stage model enables estimation of both the linear mixing and the interactions related to energy-correlations, without resorting to approximations of the likelihood function or other non-principled approaches. We demonstrate the applicability of our method with synthetic dataset, natural images and brain signals.

Subject Area: Unsupervised and Semi-supervised Learning\ICA, PCA, CCA & Other Linear Models

M58 Generalised Coupled Tensor Factorisation

K. Yilmaz kenan@sibnet.com.tr
 A. Cemgil taylan.cemgil@boun.edu.tr
 U. Simsekli umut.simsekli@boun.edu.tr
 Bogazici University

We derive algorithms for generalised tensor factorisation (GTF) by building upon the well-established theory of Generalised Linear Models. Our algorithms are general in the sense that we can compute arbitrary factorisations in a message passing framework, derived for a broad class of exponential family distributions including special cases such as Tweedie's distributions corresponding to β -divergences. By bounding the step size of the Fisher

Scoring iteration of the GLM, we obtain general updates for real data and multiplicative updates for non-negative data. The GTF framework is, then extended easily to address the problems when multiple observed tensors are factorised simultaneously. We illustrate our coupled factorisation approach on synthetic data as well as on a musical audio restoration problem.

Subject Area: Unsupervised & Semi-supervised Learning

M59 Similarity-based Learning via Data Driven Embeddings

P. Kar purushot@cse.iitk.ac.in
 Indian Institute of Technology Kanpur
 P. Jain prajain@microsoft.com

We consider the problem of classification using similarity/distance functions over data. Specifically, we propose a framework for defining the *goodness* of a (dis)similarity function with respect to a given learning task and propose algorithms that have guaranteed generalization properties when working with such good functions. Our framework unifies and generalizes the frameworks proposed by (Balcan-Blum 2006) and (Wang et al 2007). An attractive feature of our framework is its adaptability to data - we do not promote a fixed notion of goodness but rather let data dictate it. We show, by giving theoretical guarantees that the goodness criterion best suited to a problem can itself be learned which makes our approach applicable to a variety of domains and problems. We propose a *landmarking*-based approach to obtaining a classifier from such learned goodness criteria. We then provide a novel diversity based heuristic to perform task-driven selection of landmark points instead of random selection. We demonstrate the effectiveness of our goodness criteria learning method as well as the landmark selection heuristic on a variety of similarity-based learning datasets and benchmark UCI datasets on which our method consistently outperforms existing approaches by a significant margin.

Subject Area: Supervised Learning

M60 Metric Learning with Multiple Kernels

J. Wang Jun.Wang@unige.ch
 H. Do Huyen.Do@unige.ch
 A. Woznica Adam.Woznica@unige.ch
 A. Kalousis Alexandros.Kalousis@unige.ch
 University of Geneva

Metric learning has become a very active research field. The most popular representative--Mahalanobis metric learning--can be seen as learning a linear transformation and then computing the Euclidean metric in the transformed space. Since a linear transformation might not always be appropriate for a given learning problem, kernelized versions of various metric learning algorithms exist. However, the problem then becomes finding the appropriate kernel function. Multiple kernel learning addresses this limitation by learning a linear combination of a number of predefined kernels; this approach can be also readily used in the context of multiple-source learning to fuse different data sources. Surprisingly, and despite the extensive work on multiple kernel learning for SVMs, there has been no work in the area of metric learning with multiple kernel learning. In this paper we fill this gap

and present a general approach for metric learning with multiple kernel learning. Our approach can be instantiated with different metric learning algorithms provided that they satisfy some constraints. Experimental evidence suggests that our approach outperforms metric learning with an unweighted kernel combination and metric learning with cross-validation based kernel selection.

Subject Area: Unsupervised & Semi-supervised Learning

M61 Regularized Laplacian Estimation and Fast Eigenvector Approximation

P. Perry pperry@stern.nyu.edu
 NYU
 M. Mahoney mahoneymw@gmail.com
 Stanford

Recently, Mahoney and Orecchia demonstrated that popular diffusion-based procedures to compute a quick approximation to the first nontrivial eigenvector of a data graph Laplacian exactly solve certain regularized Semi-Definite Programs (SDPs). In this paper, we extend that result by providing a statistical interpretation of their approximation procedure. Our interpretation will be analogous to the manner in which ℓ^2 -regularized or ℓ^1 -regularized ℓ^2 regression (often called Ridge regression and Lasso regression, respectively) can be interpreted in terms of a Gaussian prior or a Laplace prior, respectively, on the coefficient vector of the regression problem. Our framework will imply that the solutions to the Mahoney-Orecchia regularized SDP can be interpreted as regularized estimates of the pseudoinverse of the graph Laplacian. Conversely, it will imply that the solution to this regularized estimation problem can be computed very quickly by running, e.g., the fast diffusion-based PageRank procedure for computing an approximation to the first nontrivial eigenvector of the graph Laplacian. Empirical results are also provided to illustrate the manner in which approximate eigenvector computation implicitly performs statistical regularization, relative to running the corresponding exact algorithm.

Subject Area: Unsupervised & Semi-supervised Learning

M62 Hierarchical Topic Modeling for Analysis of Time-Evolving Personal Choices

X. Zhang xianxingzhang@gmail.com
 D. Dunson dunson@stat.duke.edu
 L. Carin lcarin@duke.edu
 Duke University

The nested Chinese restaurant process is extended to design a nonparametric topic-model tree for representation of human choices. Each tree branch corresponds to a type of person, and each node (topic) has a corresponding probability vector over items that may be selected. The observed data are assumed to have associated temporal covariates (corresponding to the time at which choices are made), and we wish to impose that with increasing time it is more probable that topics deeper in the tree are utilized. This structure is imposed by developing a new "change point" stick-breaking model that is coupled with a Poisson and product-of-gammas construction. To share topics across the tree nodes, topic distributions are drawn from a

Dirichlet process. As a demonstration of this concept, we analyze real data on course selections of undergraduate students at Duke University, with the goal of uncovering and concisely representing structure in the curriculum and in the characteristics of the student body."

Subject Area: Unsupervised & Semi-supervised Learning

M63 Efficient Learning of Generalized Linear and Single Index Models with Isotonic Regression

S. Kakade sham@tti-c.org
 Microsoft Research
 A. Kalai adam@microsoft.com
 O. Shamir ohadsh@microsoft.com
 Microsoft Research
 V. Kanade vkanade@fas.harvard.edu
 Harvard University

Generalized Linear Models (GLMs) and Single Index Models (SIMs) provide powerful generalizations of linear regression, where the target variable is assumed to be a (possibly unknown) 1-dimensional function of a linear predictor. In general, these problems entail non-convex estimation procedures, and, in practice, iterative local search heuristics are often used. Kalai and Sastry (2009) provided the first provably efficient method, the *Isotron* algorithm, for learning SIMs and GLMs, under the assumption that the data is in fact generated under a GLM and under certain monotonicity and Lipschitz (bounded slope) constraints. The *Isotron* algorithm interleaves steps of perceptron-like updates with isotonic regression (fitting a one-dimensional non-decreasing function). However, to obtain provable performance, the method requires a fresh sample every iteration. In this paper, we provide algorithms for learning GLMs and SIMs, which are both computationally and statistically efficient. We modify the isotonic regression step in *Isotron* to fit a Lipschitz monotonic function, and also provide an efficient $O(n \log(n))$ algorithm for this step, improving upon the previous $O(n^2)$ algorithm. We provide a brief empirical study, demonstrating the feasibility of our algorithms in practice.

Subject Area: Learning Theory

M64 Unifying Non-Maximum Likelihood Learning Objectives with Minimum KL Contraction

S. Lyu lsw@cs.albany.edu
 University at Albany SUNY

When used to learn high dimensional parametric probabilistic models, the classical maximum likelihood (ML) learning often suffers from computational intractability, which motivates the active developments of non-ML learning methods. Yet, because of their divergent motivations and forms, the objective functions of many non-ML learning methods are seemingly unrelated, and there lacks a unified framework to understand them. In this work, based on an information geometric view of parametric learning, we introduce a general non-ML learning principle termed as minimum KL contraction, where we seek optimal parameters that minimizes the contraction of the KL divergence between the two distributions after they are transformed with a KL contraction operator. We then

show that the objective functions of several important or recently developed non-ML learning methods, including contrastive divergence [12], noise-contrastive estimation [11], partial likelihood [7], non-local contrastive objectives [31], score matching [14], pseudo-likelihood [3], maximum conditional likelihood [17], maximum mutual information [2], maximum marginal likelihood [9], and conditional and marginal composite likelihood [24], can be unified under the minimum KL contraction framework with different choices of the KL contraction operators.

Subject Area: Learning Theory

M65 Statistical Performance of Convex Tensor Decomposition

R. Tomioka tomioka@mist.i.u-tokyo.ac.jp
 T. Suzuki s-tajji@stat.t.u-tokyo.ac.jp
 University of Tokyo
 K. Hayashi kohei-h@is.naist.jp
 Nara Institute of Science and Technology
 H. Kashima kashima@mist.i.u-tokyo.ac.jp

We analyze the statistical performance of a recently proposed convex tensor decomposition algorithm. Conventionally tensor decomposition has been formulated as non-convex optimization problems, which hindered the analysis of their performance. We show under some conditions that the mean squared error of the convex method scales linearly with the quantity we call the normalized rank of the true tensor. The current analysis naturally extends the analysis of convex low-rank matrix estimation to tensors. Furthermore, we show through numerical experiments that our theory can precisely predict the scaling behaviour in practice.

Subject Area: Theory

M66 On the accuracy of l1-filtering of signals with block-sparse structure

A. Louditski Anatoli.louditski@imag.fr
 UJF
 F. Kilinc Karzan fkilinc@andrew.cmu.edu
 Carnegie Mellon University
 A. Nemirovski nemirovs@isye.gatech.edu
 Georgia Institute of Technology
 B. Polyak boris@ipu.ru
 Institute for Control Sciences, RAS Moscow

We discuss new methods for the recovery of signals with block-sparse structure, based on ℓ_1 -minimization. Our emphasis is on the efficiently computable error bounds for the recovery routines. We optimize these bounds with respect to the method parameters to construct the estimators with improved statistical properties. We justify the proposed approach with an oracle inequality which links the properties of the recovery algorithms and the best estimation performance.

Subject Area: Theory

M67 Committing Bandits

L. Bui locbui@ieee.org
 S. Mannor shie@ee.technion.ac.il
 Technion
 R. Johari ramesh.johari@stanford.edu
 Stanford University

We consider a multi-armed bandit problem where there are two phases. The first phase is an experimentation phase where the decision maker is free to explore multiple options. In the second phase the decision maker has to commit to one of the arms and stick with it. Cost is incurred during both phases with a higher cost during the experimentation phase. We analyze the regret in this setup, and both propose algorithms and provide upper and lower bounds that depend on the ratio of the duration of the experimentation phase to the duration of the commitment phase. Our analysis reveals that if given the choice, it is optimal to experiment $\Theta(\ln T)$ steps and then commit, where T is the time horizon.

Subject Area: Theory\Online Learning

M68 Newtron: an Efficient Bandit algorithm for Online Multiclass Prediction

E. Hazan ehazan@ie.technion.ac.il
 Technion
 S. Kale satyen.kale@gmail.com
 IBM

We present an efficient algorithm for the problem of online multiclass prediction with bandit feedback in the fully adversarial setting. We measure its regret with respect to the log-loss defined in [AR09], which is parameterized by a scalar α . We prove that the regret of NEWTRON is $O(\log T)$ when α is a constant that does not vary with horizon T , and at most $O(T^{2/3})$ if α is allowed to increase to infinity with T . For $\alpha = O(\log T)$, the regret is bounded by $O(\sqrt{T})$, thus solving the open problem of [KSST08, AR09]. Our algorithm is based on a novel application of the online Newton method [HAK07]. We test our algorithm and show it to perform well in experiments, even when α is a small constant.

Subject Area: Theory\Online Learning

M69 Learning Eigenvectors for Free

W. Koolen wouter@cs.rhul.ac.uk
 Royal Holloway, University of London
 W. Kotlowski W.T.Kotlowski@cwi.nl
 Centrum Wiskunde en Informatica
 M. Warmuth manfred@cse.ucsc.edu
 Univ. of Calif. at Santa Cruz

We extend the classical problem of predicting a sequence of outcomes from a finite alphabet to the matrix domain. In this extension, the alphabet of n outcomes is replaced by the set of all dyads, i.e. outer products uu^T where u is a vector in R^n of unit length. Whereas in the classical case the goal is to learn (i.e. sequentially predict as well as) the best multinomial distribution, in the matrix case we desire to learn the density matrix that best explains the observed sequence of dyads. We show how popular

online algorithms for learning a multinomial distribution can be extended to learn density matrices. Intuitively, learning the n^2 parameters of a density matrix is much harder than learning the n parameters of a multinomial distribution. Completely surprisingly, we prove that the worst-case regrets of certain classical algorithms and their matrix generalizations are identical. The reason is that the worst-case sequence of dyads share a common eigensystem, i.e. the worst case regret is achieved in the classical case. So these matrix algorithms learn the eigenvectors without any regret.

Subject Area: Theory\Online Learning

M70 Online Learning: Stochastic, Constrained, and Smoothed Adversaries

A. Rakhlin rakhlin@gmail.com
University of Pennsylvania
K. Sridharan karthik@ttic.edu
Toyota Technological Institute at Chicago
A. Tewari ambujtewari@gmail.com
UT Austin

Learning theory has largely focused on two main learning scenarios: the classical statistical setting where instances are drawn i.i.d. from a fixed distribution, and the adversarial scenario whereby at every time step the worst instance is revealed to the player. It can be argued that in the real world neither of these assumptions is reasonable. We define the minimax value of a game where the adversary is restricted in his moves, capturing stochastic and non-stochastic assumptions on data. Building on the sequential symmetrization approach, we define a notion of distribution-dependent Rademacher complexity for the spectrum of problems ranging from i.i.d. to worst-case. The bounds let us immediately deduce variation-type bounds. We study a smoothed online learning scenario and show that exponentially small amount of noise can make function classes with infinite Littlestone dimension learnable.

Subject Area: Theory\Online Learning

M71 Optimistic Optimization of a Deterministic Function without the Knowledge of its Smoothness

R. Munos remi.munos@inria.fr
INRIA Lille - Nord Europe

We consider a global optimization problem of a deterministic function f in a semi-metric space, given a finite budget of n evaluations. The function f is assumed to be locally smooth (around one of its global maxima) with respect to a semi-metric ℓ . We describe two algorithms based on optimistic exploration that use a hierarchical partitioning of the space at all scales. A first contribution is an algorithm, DOO, that requires the knowledge of ℓ . We report a finite-sample performance bound in terms of a measure of the quantity of near-optimal states. We then define a second algorithm, SOO, which does not require the knowledge of the semi-metric ℓ under which f is smooth, and whose performance is almost as good as DOO optimally-fitted.

Subject Area: Theory\Online Learning

M72 The Impact of Unlabeled Patterns in Rademacher Complexity Theory for Kernel Classifiers

L. Oneto luca.oneto@unige.it
D. Anguita davide.anguita@unige.it
A. Ghio Alessandro.Ghio@unige.it
S. Ridella sandro.ridella@unige.it
University of Genoa, Italy

We derive here new generalization bounds, based on Rademacher Complexity theory, for model selection and error estimation of linear (kernel) classifiers, which exploit the availability of unlabeled samples. In particular, two results are obtained: the first one shows that, using the unlabeled samples, the confidence term of the conventional bound can be reduced by a factor of three; the second one shows that the unlabeled samples can be used to obtain much tighter bounds, by building localized versions of the hypothesis class containing the optimal classifier.

Subject Area: Theory\Statistical Learning Theory

M73 Unifying Framework for Fast Learning Rate of Non-Sparse Multiple Kernel Learning

T. Suzuki s-taiji@stat.t.u-tokyo.ac.jp
University of Tokyo

In this paper, we give a new generalization error bound of Multiple Kernel Learning (MKL) for a general class of regularizations. Our main target in this paper is dense type regularizations including ℓ_p -MKL that imposes ℓ_p -mixed-norm regularization instead of ℓ_1 -mixed-norm regularization. According to the recent numerical experiments, the sparse regularization does not necessarily show a good performance compared with dense type regularizations. Motivated by this fact, this paper gives a general theoretical tool to derive fast learning rates that is applicable to arbitrary monotone norm-type regularizations in a unifying manner. As a by-product of our general result, we show a fast learning rate of ℓ_p -MKL that is tightest among existing bounds. We also show that our general learning rate achieves the minimax lower bound. Finally, we show that, when the complexities of candidate reproducing kernel Hilbert spaces are inhomogeneous, dense type regularization shows better learning rate compared with sparse ℓ_1 regularization.

Subject Area: Theory\Statistical Learning Theory

M74 Nearest Neighbor based Greedy Coordinate Descent

I. Dhillon inderjit@cs.utexas.edu
P. Ravikumar pradeepr@cs.utexas.edu
A. Tewari ambujtewari@gmail.com
University of Texas, Austin

Increasingly, optimization problems in machine learning, especially those arising from high-dimensional statistical estimation, have a large number of variables. Modern statistical estimators developed over the past decade have statistical or sample complexity that depends only weakly on the number of parameters when there is some structure to the problem, such as sparsity. A central question is whether similar advances can be made in their computational complexity as well. In this paper, we propose strategies that indicate that such advances can indeed be made. In particular, we investigate the greedy

coordinate descent algorithm, and note that performing the greedy step efficiently weakens the costly dependence on the problem size provided the solution is sparse. We then propose a suite of methods that perform these greedy steps efficiently by a reduction to nearest neighbor search. We also devise a more amenable form of greedy descent for composite non-smooth objectives; as well as several approximate variants of such greedy descent. We develop a practical implementation of our algorithm that combines greedy coordinate descent with locality sensitive hashing. Without tuning the latter data structure, we are not only able to significantly speed up the vanilla greedy method, but also outperform cyclic descent when the problem size becomes large. Our results indicate the effectiveness of our nearest neighbor strategies, and also point to many open questions regarding the development of computational geometric techniques tailored towards first-order optimization methods. Subject Area: Theory\Statistical Learning Theory

M75 Agnostic Selective Classification

Y. Wiener yair.wiener@gmail.com
 R. El-Yaniv rani@cs.technion.ac.il
 Technion

For a learning problem whose associated excess loss class is (β, B) -Bernstein, we show that it is theoretically possible to track the same classification performance of the best (unknown) hypothesis in our class, provided that we are free to abstain from prediction in some region of our choice. The (probabilistic) volume of this rejected region of the domain is shown to be diminishing at rate $O(B\theta(\sqrt{1/m})^\beta)$, where θ is Hanneke's disagreement coefficient. The strategy achieving this performance has computational barriers because it requires empirical error minimization in an agnostic setting. Nevertheless, we heuristically approximate this strategy and develop a novel selective classification algorithm using constrained SVMs. We show empirically that the resulting algorithm consistently outperforms the traditional rejection mechanism based on distance from decision boundary. Subject Area: Theory\Statistical Learning Theory

M76 Greedy Model Averaging

D. Dai dongdai916@gmail.com
 T. Zhang tzhang@stat.rutgers.edu
 Rutgers University

This paper considers the problem of combining multiple models to achieve a prediction accuracy not much worse than that of the best single model for least squares regression. It is known that if the models are mis-specified, model averaging is superior to model selection. Specifically, let n be the sample size, then the worst case regret of the former decays at the rate of $O(1/n)$ while the worst case regret of the latter decays at the rate of $O(1/\sqrt{n})$. In the literature, the most important and widely studied model averaging method that achieves the optimal $O(1/n)$ average regret is the exponential weighted model averaging (EWMA) algorithm. However this method suffers from several limitations. The purpose of this paper is to present a new greedy model averaging procedure that improves EWMA. We prove strong theoretical guarantees for the new procedure and illustrate our theoretical results with empirical examples. Subject Area: Theory\Statistical Learning Theory

M77 Confidence Sets for Network Structure

D. Choi dchoi@seas.harvard.edu
 P. Wolfe patrick@seas.harvard.edu
 E. Airoldi airoldi@fas.harvard.edu
 Harvard University

Latent variable models are frequently used to identify structure in dichotomous network data, in part because they give rise to a Bernoulli product likelihood that is both well understood and consistent with the notion of exchangeable random graphs. In this article we propose conservative confidence sets that hold with respect to these underlying Bernoulli parameters as a function of any given partition of network nodes, enabling us to assess estimates of *residual* network structure, that is, structure that cannot be explained by known covariates and thus cannot be easily verified by manual inspection. We demonstrate the proposed methodology by analyzing student friendship networks from the National Longitudinal Survey of Adolescent Health that include race, gender, and school year as covariates. We employ a stochastic expectation-maximization algorithm to fit a logistic regression model that includes these explanatory variables as well as a latent stochastic blockmodel component and additional node-specific effects. Although maximum-likelihood estimates do not appear consistent in this context, we are able to evaluate confidence sets as a function of different blockmodel partitions, which enables us to qualitatively assess the significance of estimated residual network structure relative to a baseline, which models covariates but lacks block structure. Subject Area: Probabilistic Models and Methods

M78 On Tracking The Partition Function

G. Desjardins desjagui@iro.umontreal.ca
 A. Courville aaron.courville@umontreal.ca
 Y. Bengio bengioy@iro.umontreal.ca
 University of Montreal

Markov Random Fields (MRFs) have proven very powerful both as density estimators and feature extractors for classification. However, their use is often limited by an inability to estimate the partition function Z . In this paper, we exploit the gradient descent training procedure of restricted Boltzmann machines (a type of MRF) to track the log partition function during learning. Our method relies on two distinct sources of information: (1) estimating the change ΔZ incurred by each gradient update, (2) estimating the difference in Z over a small set of tempered distributions using bridge sampling. The two sources of information are then combined using an inference procedure similar to Kalman filtering. Learning MRFs through Tempered Stochastic Maximum Likelihood, we can estimate Z using no more temperatures than are required for learning. Comparing to both exact values and estimates using annealed importance sampling (AIS), we show on several datasets that our method is able to accurately track the log partition function. In contrast to AIS, our method provides this estimate at each time-step, at a computational cost similar to that required for training alone. Subject Area: Probabilistic Models and Methods

M79 Probabilistic Amplitude and Frequency Demodulation

R. Turner ret26@cam.ac.uk
 M. Sahani maneesh@gatsby.ucl.ac.uk
 Gatsby Unit, UCL

A number of recent scientific and engineering problems require signals to be decomposed into a product of a slowly varying positive envelope and a quickly varying carrier whose instantaneous frequency also varies slowly over time. Although signal processing provides algorithms for so-called amplitude- and frequency-demodulation (AFD), there are well known problems with all of the existing methods. Motivated by the fact that AFD is ill-posed, we approach the problem using probabilistic inference. The new approach, called probabilistic amplitude and frequency demodulation (PAFD), models instantaneous frequency using an auto-regressive generalization of the von Mises distribution, and the envelopes using Gaussian auto-regressive dynamics with a positivity constraint. A novel form of expectation propagation is used for inference. We demonstrate that although PAFD is computationally demanding, it outperforms previous approaches on synthetic and real signals in clean, noisy and missing data settings.
 Subject Area: Probabilistic Models and Methods

M80 Structure Learning for Optimization

S. Yang yang@cs.washington.edu
 University of Washington
 A. Rahimi ali@redbowlabs.com
 Red Bow Labs

We describe a family of global optimization procedures that automatically decompose optimization problems into smaller loosely coupled problems, then combine the solutions of these with message passing algorithms. We show empirically that these methods excel in avoiding local minima and produce better solutions with fewer function evaluations than existing global optimization methods. To develop these methods, we introduce a notion of coupling between variables of optimization that generalizes the notion of coupling that arises from factoring functions into terms that involve small subsets of the variables. It therefore subsumes the notion of independence between random variables in statistics, sparseness of the Hessian in nonlinear optimization, and the generalized distributive law. Despite being more general, this notion of coupling is easier to verify empirically -- making structure estimation easy -- yet it allows us to migrate well-established inference methods on graphical models to the setting of global optimization.
 Subject Area: Probabilistic Models and Methods

M81 Spike and Slab Variational Inference for Multi-Task and Multiple Kernel Learning

M. Titsias mtitsias@cs.man.ac.uk
 University of Manchester
 M. Lázaro-Gredilla lazarox@gmail.com
 Universidad Carlos III de Madrid

We introduce a variational Bayesian inference algorithm which can be widely applied to sparse linear models. The

algorithm is based on the spike and slab prior which, from a Bayesian perspective, is the golden standard for sparse inference. We apply the method to a general multi-task and multiple kernel learning model in which a common set of Gaussian process functions is linearly combined with task-specific sparse weights, thus inducing relation between tasks. This model unifies several sparse linear models, such as generalized linear models, sparse factor analysis and matrix factorization with missing values, so that the variational algorithm can be applied to all these cases. We demonstrate our approach in multi-output Gaussian process regression, multi-class classification, image processing applications and collaborative filtering.
 Subject Area: Probabilistic Models and Methods

M82 Thinning Measurement Models and Questionnaire Design

R. Silva ricardo@stats.ucl.ac.uk
 University College London

Inferring key unobservable features of individuals is an important task in the applied sciences. In particular, an important source of data in fields such as marketing, social sciences and medicine is questionnaires: answers in such questionnaires are noisy measures of target unobserved features. While comprehensive surveys help to better estimate the latent variables of interest, aiming at a high number of questions comes at a price: refusal to participate in surveys can go up, as well as the rate of missing data; quality of answers can decline; costs associated with applying such questionnaires can also increase. In this paper, we cast the problem of refining existing models for questionnaire data as follows: solve a constrained optimization problem of preserving the maximum amount of information found in a latent variable model using only a subset of existing questions. The goal is to find an optimal subset of a given size. For that, we first define an information theoretical measure for quantifying the quality of a reduced questionnaire. Three different approximate inference methods are introduced to solve this problem. Comparisons against a simple but powerful heuristic are presented.
 Subject Area: Probabilistic Models and Methods

M83 An Application of Tree-Structured Expectation Propagation for Channel Decoding

P. Olmos olmos@us.es
 L. Salamanca salamanca@us.es
 J. Murillo Fuentes murillo@us.es
 Universidad de Sevilla
 F. Perez-Cruz fernandop@ieee.org
 University Carlos III in Madrid

We show an application of a tree structure for approximate inference in graphical models using the expectation propagation algorithm. These approximations are typically used over graphs with short-range cycles. We demonstrate that these approximations also help in sparse graphs with long-range loops, as the ones used in coding theory to approach channel capacity. For asymptotically large sparse graph, the expectation propagation algorithm together with the tree structure yields a completely disconnected approximation to the graphical model but, for

for finite-length practical sparse graphs, the tree structure approximation to the code graph provides accurate estimates for the marginal of each variable.
Subject Area: Probabilistic Models and Methods

M84 Global Solution of Fully-Observed Variational Bayesian Matrix Factorization is Column-Wise Independent

S. Nakajima shinnkj23@gmail.com
Nikon Corporation
M. Sugiyama sugi@cs.titech.ac.jp
Tokyo Institute of Technology
S. Babacan dbabacan@illinois.edu
University of Illinois at Urbana-Champaign

Variational Bayesian matrix factorization (VBMF) efficiently approximates the posterior distribution of factorized matrices by assuming *matrix-wise independence* of the two factors. A recent study on fully-observed VBMF showed that, under a stronger assumption that the two factorized matrices are column-wise independent, the global optimal solution can be analytically computed. However, it was not clear how restrictive the column-wise independence assumption is. In this paper, we prove that the global solution under *matrix-wise independence* is actually column-wise independent, implying that the column-wise independence assumption is harmless. A practical consequence of our theoretical finding is that the global solution under *matrix-wise independence* (which is a standard setup) can be obtained analytically in a computationally very efficient way without any iterative algorithms. We experimentally illustrate advantages of using our analytic solution in probabilistic principal component analysis.

Subject Area: Probabilistic Models and Methods

M85 Quasi-Newton Methods for Markov Chain Monte Carlo

Y. Zhang s0956889@sms.ed.ac.uk
C. Sutton csutton@inf.ed.ac.uk
University of Edinburgh

The performance of Markov chain Monte Carlo methods is often sensitive to the scaling and correlations between the random variables of interest. An important source of information about the local correlation and scale is given by the Hessian matrix of the target distribution, but this is often either computationally expensive or infeasible. In this paper we propose MCMC samplers that make use of quasi-Newton approximations from the optimization literature, that approximate the Hessian of the target distribution from previous samples and gradients generated by the sampler. A key issue is that MCMC samplers that depend on the history of previous states are in general not valid. We address this problem by using limited memory quasi-Newton methods, which depend only on a fixed window of previous samples. On several real world datasets, we show that the quasi-Newton sampler is a more effective sampler than standard Hamiltonian Monte Carlo at a fraction of the cost of MCMC methods that require higher-order derivatives.

Subject Area: Probabilistic Models and Methods

M86 Nonstandard Interpretations of Probabilistic Programs for Efficient Inference

D. Wingate wingated@mit.edu
N. Goodman ngoodman@stanford.edu
A. Stuhlmüller andreas@stuhlmueeller.info
Massachusetts Institute of Technology
J. Siskind qobi@purdue.edu
Purdue University

Probabilistic programming languages allow modelers to specify a stochastic process using syntax that resembles modern programming languages. Because the program is in machine-readable format, a variety of techniques from compiler design and program analysis can be used to examine the structure of the distribution represented by the probabilistic program. We show how nonstandard interpretations of probabilistic programs can be used to craft efficient inference algorithms: information about the structure of a distribution (such as gradients or dependencies) is generated as a monad-like side computation while executing the program. These interpretations can be easily coded using special-purpose objects and operator overloading. We implement two examples of nonstandard interpretations in two different languages, and use them as building blocks to construct inference algorithms: automatic differentiation, which enables gradient based methods, and provenance tracking, which enables efficient construction of global proposals.

Subject Area: Probabilistic Models and Methods

M87 Uniqueness of Belief Propagation on Signed Graphs

Y. Watanabe watay@ism.ac.jp
The Institute of Statistical Mathematics

While loopy Belief Propagation (LBP) has been utilized in a wide variety of applications with empirical success, it comes with few theoretical guarantees. Especially, if the interactions of random variables in a graphical model are strong, the behaviors of the algorithm can be difficult to analyze due to underlying phase transitions. In this paper, we develop a novel approach to the uniqueness problem of the LBP fixed point; our new “necessary and sufficient” condition is stated in terms of graphs and signs, where the sign denotes the types (attractive/repulsive) of the interaction (i.e., compatibility function) on the edge. In all previous works, uniqueness is guaranteed only in the situations where the strength of the interactions are “sufficiently” small in certain senses. In contrast, our condition covers arbitrary strong interactions on the specified class of signed graphs. The result of this paper is based on the recent theoretical advance in the LBP algorithm; the connection with the graph zeta function.

Subject Area: Probabilistic Models and Methods

M88 Non-conjugate Variational Message Passing for Multinomial and Binary Regression

D. Knowles dak33@cam.ac.uk
 University of Cambridge
 T. Minka minka@microsoft.com
 Microsoft Research Ltd

Variational Message Passing (VMP) is an algorithmic implementation of the Variational Bayes (VB) method which applies only in the special case of conjugate exponential family models. We propose an extension to VMP, which we refer to as Non-conjugate Variational Message Passing (NCVMP) which aims to alleviate this restriction while maintaining modularity, allowing choice in how expectations are calculated, and integrating into an existing message-passing framework: Infer.NET. We demonstrate NCVMP on logistic binary and multinomial regression. In the multinomial case we introduce a novel variational bound for the softmax factor which is tighter than other commonly used bounds whilst maintaining computational tractability. Subject Area: Probabilistic Models and Methods

M89 Ranking annotators for crowdsourced labeling tasks

V. Raykar vikasraykar@gmail.com
 S. Yu shipeng.yu@gmail.com
 Siemens Medical Solutions

With the advent of crowdsourcing services it has become quite cheap and reasonably effective to get a dataset labeled by multiple annotators in a short amount of time. Various methods have been proposed to estimate the consensus labels by correcting for the bias of annotators with different kinds of expertise. Often we have low quality annotators or *spammers*--annotators who assign labels randomly (e.g., without actually looking at the instance). Spammers can make the cost of acquiring labels very expensive and can potentially degrade the quality of the consensus labels. In this paper we formalize the notion of a spammer and define a score which can be used to rank the annotators--with the spammers having a score close to zero and the good annotators having a high score close to one. Subject Area: Probabilistic Models and Methods

M90 Gaussian process modulated renewal processes

V. Rao vrao@gatsby.ucl.ac.uk
 Y. Teh ywteh@gatsby.ucl.ac.uk
 Gatsby Computational Neuroscience Unit, UCL

Renewal processes are generalizations of the Poisson process on the real line, whose intervals are drawn i.i.d. from some distribution. Modulated renewal processes allow these distributions to vary with time, allowing the introduction nonstationarity. In this work, we take a nonparametric Bayesian approach, modeling this nonstationarity with a Gaussian process. Our approach is based on the idea of *uniformization*, allowing us to draw exact samples from an otherwise intractable distribution. We develop a novel and efficient MCMC sampler for posterior inference. In our experiments, we test these on a number of synthetic and real datasets. Subject Area: Probabilistic Models and Methods

M91 Infinite Latent SVM for Classification and Multi-task Learning

J. Zhu junzhu@cs.cmu.edu
 E. Xing epxing@cs.cmu.edu
 Carnegie Mellon University
 N. Chen chenn07@mails.tsinghua.edu.cn
 Tsinghua University

Unlike existing nonparametric Bayesian models, which rely solely on specially conceived priors to incorporate domain knowledge for discovering improved latent representations, we study nonparametric Bayesian inference with regularization on the desired posterior distributions. While priors can indirectly affect posterior distributions through Bayes' theorem, imposing posterior regularization is arguably more direct and in some cases can be much easier. We particularly focus on developing *infinite latent support vector machines* (iLSVM) and *multi-task infinite latent support vector machines* (MT-iLSVM), which explore the large-margin idea in combination with a nonparametric Bayesian model for discovering predictive latent features for classification and multi-task learning, respectively. We present efficient inference methods and report empirical studies on several benchmark datasets. Our results appear to demonstrate the merits inherited from both large-margin learning and Bayesian nonparametrics. Subject Area: Probabilistic Models and Methods

M92 Spatial distance dependent Chinese restaurant processes for image segmentation

S. Ghosh sghosh@cs.brown.edu
 E. Sudderth sudderth@cs.brown.edu
 Brown University
 A. Ungureanu andrei.b.ungureanu@gmail.com
 Morgan Stanley
 D. Blei blei@cs.princeton.edu
 Princeton University

The distance dependent Chinese restaurant process (ddCRP) was recently introduced to accommodate random partitions of non-exchangeable data. The ddCRP clusters data in a biased way: each data point is more likely to be clustered with other data that are near it in an external sense. This paper examines the ddCRP in a spatial setting with the goal of natural image segmentation. We explore the biases of the spatial ddCRP model and propose a novel hierarchical extension better suited for producing "human-like" segmentations. We then study the sensitivity of the models to various distance and appearance hyperparameters, and provide the first rigorous comparison of nonparametric Bayesian models in the image segmentation domain. On unsupervised image segmentation, we demonstrate that similar performance to existing nonparametric Bayesian models is possible with substantially simpler models and algorithms." Subject Area: Probabilistic Models and Methods

M93 Analytical Results for the Error in Filtering of Gaussian Processes

A. Susemihl alex.susemihl@bccn-berlin.de
 M. Opper manfred.opper@tu-berlin.de
 Berlin Institute of Technology
 R. Meir rmeir@ee.technion.ac.il
 Technion

Bayesian filtering of stochastic stimuli has received a great deal of attention recently. It has been applied to describe the way in which biological systems dynamically represent and make decisions about the environment. There have been no exact results for the error in the biologically plausible setting of inference on point process, however. We present an exact analysis of the evolution of the mean-squared error in a state estimation task using Gaussian-tuned point processes as sensors. This allows us to study the dynamics of the error of an optimal Bayesian decoder, providing insights into the limits obtainable in this task. This is done for Markovian and a class of non-Markovian Gaussian processes. We find that there is an optimal tuning width for which the error is minimized. This leads to a characterization of the optimal encoding for the setting as a function of the statistics of the stimulus, providing a mathematically sound primer for an ecological theory of sensory processing.

Subject Area: Probabilistic Models and Methods

M94 Robust Multi-Class Gaussian Process Classification

D. Hernández-lobato danielhernandezlobato@gmail.com
 P. Dupont Pierre.Dupont@uclouvain.be
 Louvain University
 J. Hernández-Lobato jmh233@cam.ac.uk
 Cambridge University

Multi-class Gaussian Process Classifiers (MGPCs) are often affected by over-fitting problems when labeling errors occur far from the decision boundaries. To prevent this, we investigate a robust MGPC (RMGPC) which considers labeling errors independently of their distance to the decision boundaries. Expectation propagation is used for approximate inference. Experiments with several datasets in which noise is injected in the class labels illustrate the benefits of RMGPC. This method performs better than other Gaussian process alternatives based on considering latent Gaussian noise or heavy-tailed processes. When no noise is injected in the labels, RMGPC still performs equal or better than the other methods. Finally, we show how RMGPC can be used for successfully identifying data instances which are difficult to classify accurately in practice.

Subject Area: Probabilistic Models and Methods

M95 Additive Gaussian Processes

D. Duvenaud dkd23@cam.ac.uk
 C. Rasmussen cer54@cam.ac.uk
 University of Cambridge
 H. Nickisch hannes@nickisch.org
 Philips Research

We introduce a Gaussian process model of functions which are additive. An additive function is one which decomposes into a sum of low-dimensional functions, each depending on only a subset of the input variables. Additive GPs generalize both Generalized Additive Models, and

the standard GP models which use squared-exponential kernels. Hyperparameter learning in this model can be seen as Bayesian Hierarchical Kernel Learning (HKL). We introduce an expressive but tractable parameterization of the kernel function, which allows efficient evaluation of all input interaction terms, whose number is exponential in the input dimension. The additional structure discoverable by this model results in increased interpretability, as well as state-of-the-art predictive power in regression tasks.

Subject Area: Probabilistic Models and Methods

M96 A Global Structural EM Algorithm for a Model of Cancer Progression

A. Tofigh ali.tofigh@mcgill.ca
 McGill University
 E. Sjlund erik.sjlund@sbc.su.se
 Stockholm University
 M. Høglund mattias.hoglund@med.lu.se
 Lund University
 J. Lagergren jensl@csc.kth.se
 KTH

Cancer has complex patterns of progression that include converging as well as diverging progressional pathways. Vogelstein's path model of colon cancer was a pioneering contribution to cancer research. Since then, several attempts have been made at obtaining mathematical models of cancer progression, devising learning algorithms, and applying these to cross-sectional data. Beerewinkel et al. provided, what they coined, EM-like algorithms for Oncogenetic Trees (OTs) and mixtures of such. Given the small size of current and future data sets, it is important to minimize the number of parameters of a model. For this reason, we too focus on tree-based models and introduce Hidden-variable Oncogenetic Trees (HOTs). In contrast to OTs, HOTs allow for errors in the data and thereby provide more realistic modeling. We also design global structural EM algorithms for learning HOTs and mixtures of HOTs (HOT-mixtures). The algorithms are global in the sense that, during the M-step, they find a structure that yields a global maximum of the expected complete log-likelihood rather than merely one that improves it. The algorithm for single HOTs performs very well on reasonable-sized data sets, while that for HOT-mixtures requires data sets of sizes obtainable only with tomorrow's more cost-efficient technologies.

Subject Area: Probabilistic Models and Methods

M97 Collective Graphical Models

D. Sheldon sheldon@eecs.oregonstate.edu
 T. Dietterich tg@cs.orst.edu
 Oregon State University

There are many settings in which we wish to fit a model of the behavior of individuals but where our data consist only of aggregate information (counts or low-dimensional contingency tables). This paper introduces *Collective Graphical Models*—a framework for modeling and probabilistic inference that operates directly on the sufficient statistics of the individual model. We derive a highly-efficient Gibbs sampling algorithm for sampling from the posterior distribution of the sufficient statistics conditioned on noisy aggregate observations, prove its correctness, and demonstrate its effectiveness experimentally.

Subject Area: Probabilistic Models and Methods

M98 Simultaneous Sampling and Multi-Structure Fitting with Adaptive Reversible Jump MCMC

T. Pham trung@cs.adelaide.edu.au
 T. Chin tjchin@cs.adelaide.edu.au
 J. Yu jin.yu@adelaide.edu.au
 D. Suter dsuter@cs.adelaide.edu.au
 The University of Adelaide

Multi-structure model fitting has traditionally taken a two-stage approach: First, sample a (large) number of model hypotheses, then select the subset of hypotheses that optimise a joint fitting and model selection criterion. This disjoint two-stage approach is arguably suboptimal and inefficient - if the random sampling did not retrieve a good set of hypotheses, the optimised outcome will not represent a good fit. To overcome this weakness we propose a new multi-structure fitting approach based on Reversible Jump MCMC. Instrumental in raising the effectiveness of our method is an adaptive hypothesis generator, whose proposal distribution is learned incrementally and online. We prove that this adaptive proposal satisfies the diminishing adaptation property crucial for ensuring ergodicity in MCMC. Our method effectively conducts hypothesis sampling and optimisation simultaneously, and gives superior computational efficiency over other methods.

Subject Area: Probabilistic Models and Methods

M99 On Causal Discovery with Cyclic Additive Noise Models

J. Mooij j.mooij@cs.ru.nl
 T. Heskes t.heskes@cs.ru.nl
 Radboud University Nijmegen
 D. Janzing dominik.janzing@tuebingen.mpg.de
 B. Schölkopf bs@tuebingen.mpg.de
 Max Planck Institute for Intelligent Systems

We study a particular class of cyclic causal models, where each variable is a (possibly nonlinear) function of its parents and additive noise. We prove that the causal graph of such models is generically identifiable in the bivariate, Gaussian-noise case. We also propose a method to learn such models from observational data. In the acyclic case, the method reduces to ordinary regression, but in the more challenging cyclic case, an additional term arises in the loss function, which makes it a special case of nonlinear independent component analysis. We illustrate the proposed method on synthetic data.

Subject Area: Probabilistic Models and Methods

M100A Model for Temporal Dependencies in Event Streams

A. Gunawardana aselag@microsoft.com
 C. Meek meek@microsoft.com
 Microsoft Research
 P. Xu puyangxu@gmail.com
 Johns Hopkins University

We introduce the Piecewise-Constant Conditional Intensity Model, a model for learning temporal dependencies in event streams. We describe a closed-form Bayesian

approach to learning these models, and describe an importance sampling algorithm for forecasting future events using these models, using a proposal distribution based on Poisson superposition. We then use synthetic data, supercomputer event logs, and web search query logs to illustrate that our learning algorithm can efficiently learn nonlinear temporal dependencies, and that our importance sampling algorithm can effectively forecast future events.

Subject Area: Probabilistic Models and Methods

M101 Facial Expression Transfer with Input-Output Temporal Restricted Boltzmann Machines

M. Zeiler zeiler@cs.nyu.edu
 G. Taylor gwtaylor@cs.nyu.edu
 R. Fergus fergus@cs.nyu.edu
 New York University
 L. Sigal lsigal@disneyresearch.com
 I. Matthews iainm@disneyresearch.com
 Disney Research Pittsburgh

We present a type of Temporal Restricted Boltzmann Machine that defines a probability distribution over an output sequence conditional on an input sequence. It shares the desirable properties of RBMs: efficient exact inference, an exponentially more expressive latent state than HMMs, and the ability to model nonlinear structure and dynamics. We apply our model to a challenging real-world graphics problem: facial expression transfer. Our results demonstrate improved performance over several baselines modeling high-dimensional 2D and 3D data.

Subject Area: Probabilistic Models and Methods

M102 Learning Auto-regressive Models from Sequence and Non-sequence Data

T. Huang tzukuoh@cs.cmu.edu
 J. Schneider schneide@cs.cmu.edu
 Carnegie Mellon University

Vector Auto-regressive models (VAR) are useful tools for analyzing time series data. In quite a few modern time series modelling tasks, the collection of reliable time series turns out to be a major challenge, either due to the slow progression of the dynamic process of interest, or inaccessibility of repetitive measurements of the same dynamic process over time. In those situations, however, we observe that it is often easier to collect a large amount of non-sequence samples, or snapshots of the dynamic process of interest. In this work, we assume a small amount of time series data are available, and propose methods to incorporate non-sequence data into penalized least-square estimation of VAR models. We consider non-sequence data as samples drawn from the stationary distribution of the underlying VAR model, and devise a novel penalization scheme based on the discrete-time Lyapunov equation concerning the covariance of the stationary distribution. Experiments on synthetic and video data demonstrate the effectiveness of the proposed methods.

Subject Area: Probabilistic Models and Methods



TUESDAY CONFERENCE

ORAL SESSION

SESSION 1 - 9:30 – 10:40 AM



Session Chair: Remi Munos

POSNER LECTURE: Learning About Sensorimotor Data

Richard Sutton sutton@cs.ualberta.ca
University of Alberta

Temporal-difference (TD) learning of reward predictions underlies both reinforcement-learning algorithms and the standard dopamine model of reward-based learning in the brain. This confluence of computational and neuroscientific ideas is perhaps the most successful since the Hebb synapse. Can it be extended beyond reward? The brain certainly predicts many things other than reward---such as in a forward model of the consequences of various ways of behaving---and TD methods can be used to make these predictions. The idea and advantages of using TD methods to learn large numbers of predictions about many states and stimuli, in parallel, have been apparent since the 1990s, but technical issues have prevented this vision from being practically implemented...until now. A key breakthrough was the development of a new family of gradient-TD methods, introduced at NIPS in 2008 (by Maei, Szepesvari, and myself). Using these methods, and other ideas, we are now able to learn thousands of non-reward predictions in real-time at 10Hz from a single sensorimotor data stream from a physical robot. These predictions are temporally extended (ranging up to tens of seconds of anticipation), goal oriented, and policy contingent. The new algorithms enable learning to be off-policy and in parallel, resulting in dramatic increases in the amount that can be learned in a given amount of time. Our effective learning rate scales linearly with computational resources. On a consumer laptop we can learn thousands of predictions in real-time. On a larger computer, or on a comparable laptop in a few years, the same methods could learn millions of meaningful predictions about different alternate ways of behaving. These predictions in aggregate constitute a rich detailed model of the world that can support planning methods such as approximate dynamic programming.

Richard S. Sutton is a professor and iCORE chair in the department of computing science at the University of Alberta. He is a fellow of the Association for the Advancement of Artificial Intelligence and co-author of the textbook "Reinforcement Learning: An Introduction" from MIT Press. Before joining the University of Alberta in 2003, he worked in industry at AT&T and GTE Labs, and in academia at the University of Massachusetts. He received a PhD in computer science from the University of Massachusetts in 1984 and a BA in psychology from Stanford University in 1978. Rich's research interests center on the learning problems facing a decision-maker interacting with its environment, which he sees as central to artificial intelligence. He is also interested in animal learning psychology, in connectionist networks, and generally in systems that continually improve their representations and models of the world.

A Non-Parametric Approach to Dynamic Programming

Oliver Kroemer oliverkro@googlemail.com
Jan Peters mail@jan-peters.net
Technische Universitaet Darmstadt

In this paper, we consider the problem of policy evaluation for continuous-state systems. We present a non-parametric approach to policy evaluation, which uses kernel density estimation to represent the system. The true form of the value function for this model can be determined, and can be computed using Galerkin's method. Furthermore, we also present a unified view of several well-known policy evaluation methods. In particular, we show that the same Galerkin method can be used to derive Least-Squares Temporal Difference learning, Kernelized Temporal Difference learning, and a discrete-state Dynamic Programming solution, as well as our proposed method. In a numerical evaluation of these algorithms, the proposed approach performed better than the other methods
Subject Area: Control and Reinforcement Learning

SPOTLIGHT SESSION

SESSION 2 - 10:40 – 11:10 AM



Session Chair: Remi Munos

- **Action-Gap Phenomenon in Reinforcement Learning**
A. Farahmand, McGill University
Subject Area: Control and Reinforcement Learning
See abstract, page 48 (T5)
- **The Fixed Points of Off-Policy TD**
J. Kolter, MIT
Subject Area: Control and Reinforcement Learning
See abstract, page 48 (T6)
- **Inductive reasoning about chimeric creatures**
C. Kemp, Carnegie Mellon University
Subject Area: Cognitive Science
See abstract, page 50 (T14)
- **Evaluating computational models of preference learning**
A. Jern, C. Lucas, C. Kemp, Carnegie Mellon University
Subject Area: Cognitive Science
See abstract, page 50 (T13)
- **Identifying Alzheimer's Disease-Related Brain Regions from Multi-Modality Neuroimaging Data using Sparse Composite Linear Discrimination Analysis**
S. Huang, J. Li, J. Ye, T. Wu, Arizona State University
K. Chen, A. Fleisher, E. Reiman, Banner Alzheimer's University
Subject Area: Brain Imaging
See abstract, page 51 (T16)

- **Decoding of Finger Flexion from Electrocorticographic Signals Using Switching Non-Parametric Dynamic Systems**
Z. Wang, Q. Ji, Rensselaer Polytechnic Institute; G. Schalk, Wadsworth Center
Subject Area: Brain-computer Interfaces & Neural Prostheses
See abstract, page 51 (T17)
- **Active learning of neural response functions with Gaussian processes**
Mijung Park, Greg Horwitz, Jonathan Pillow, UT Austin
Subject Area: Neural Coding
See abstract, page 52 (T20)



ORAL SESSION

SESSION 2 - 11:10 – 11:30 AM

Session Chair: Michael Collins

On the Completeness of First-Order Knowledge Compilation for Lifted Probabilistic Inference

Guy Van den Broeck
guy.vandenbroeck@cs.kuleuven.be
Katholieke Universiteit Leuven

Probabilistic logics are receiving a lot of attention today because of their expressive power for knowledge representation and learning. However, this expressivity is detrimental to the tractability of inference, when done at the propositional level. To solve this problem, various lifted inference algorithms have been proposed that reason at the first-order level, about groups of objects as a whole. Despite the existence of various lifted inference approaches, there are currently no completeness results about these algorithms. The key contribution of this paper is that we introduce a formal definition of lifted inference that allows us to reason about the completeness of lifted inference algorithms relative to a particular class of probabilistic models. We then show how to obtain a completeness result using a first-order knowledge compilation approach for theories of formulae containing up to two logical variables.

Subject Area: Structured and Relational Data



ORAL SESSION

SESSION 3 - 12:00 – 12:40 PM

Session Chair: Amir Globerson

Modelling Genetic Variations using Fragmentation-Coagulation Processes

Yee Whye Teh ywteh@gatsby.ucl.ac.uk
Charles Blundell c.blundell@gatsby.ucl.ac.uk
Gatsby Computational Neuroscience Unit, UCL
Lloyd Elliott elliott@gatsby.ucl.ac.uk
University College London

We propose a novel class of Bayesian nonparametric models for sequential data called fragmentation-coagulation processes (FCPs). FCPs model a set of

sequences using a partition-valued Markov process which evolves by splitting and merging clusters. An FCP is exchangeable, projective, stationary and reversible, and its equilibrium distributions are given by the Chinese restaurant process. As opposed to hidden Markov models, FCPs allow for flexible modelling of the number of clusters, and they avoid label switching non-identifiability problems. We develop an efficient Gibbs sampler for FCPs which uses uniformization and the forward-backward algorithm. Our development of FCPs is motivated by applications in population genetics, and we demonstrate the utility of FCPs on problems of genotype imputation with phased and unphased SNP data.

Subject Area: Bayesian Nonparametrics

Priors over Recurrent Continuous Time Processes

Ardavan Saeedi ardavan.s@stat.ubc.ca
Alexandre Bouchard-Côté bouchard@stat.ubc.ca
University of British Columbia

We introduce the Gamma-Exponential Process (GEP), a prior over a large family of continuous time processes. A hierarchical version of this prior (HGEP; the Hierarchical GEP) yields a useful model for analyzing complex time series. Models based on HGEPs display many attractive properties: conjugacy, exchangeability and closed-form predictive distribution for the waiting times, and exact Gibbs updates for the time scale parameters. After establishing these properties, we show how posterior inference can be carried efficiently using Particle MCMC methods. This yields a MCMC algorithm that can resample entire sequences atomically while avoiding the complications of introducing slice and stick auxiliary variables. We applied our model to the problem of estimating the disease progression in Multiple Sclerosis, and to RNA evolutionary modeling. In both domains, we found that our model outperformed the standard rate matrix estimation approach.

Subject Area: Bayesian Nonparametrics



SPOTLIGHT SESSION

SESSION 3 - 12:40 – 1:10 PM

Session Chair: Amir Globerson

- **Neuronal Adaptation for Sampling-Based Probabilistic Inference in Perceptual Bistability**
D. Reichert, P. Series, A. Storkey, Univ. of Edinburgh
Subject Area: Computational Neural Models
See abstract, page 51 (T18)
- **Sequence learning with hidden units in spiking neural networks**
J. Brea, Bern University, W. Senn & J. Pfister, Cambridge University
Subject Area: Computational Neural Models
See abstract, page 52 (T19)
- **Information Rates and Optimal Decoding in Large Neural Populations**
K. Rahnama Rad, L. Paninski, Columbia University
Subject Area: Probabilistic Models and Methods
See abstract, page 66 (T85)

TUESDAY - CONFERENCE

- **A blind sparse deconvolution method for neural spike identification**
C. Ekanadham, D. Tranchina, E. Simoncelli, Courant Institute, New York University
Subject Area: Approximate Inference
See abstract, page 67 (T88)
- **Accelerated Adaptive Markov Chain for Partition Function Computation**
S. Ermon, C. Gomes, A. Sabharwal, IBM Watson Research Center; B. Selman, Cornell University
Subject Area: Approximate Inference
See abstract, page 67 (T89)
- **The Kernel Beta Process**
L. Ren, Y. Wang, D. Dunson, L. Carin, Duke University
Subject Area: Bayesian Nonparametrics
See abstract, page 69 (T95)
- **Solving Decision Problems with Limited Information**
D. Maua, C. de Campos, Dalle Molle Institute for Artificial Intelligence
Subject Area: Graphical Models
See abstract, page 70 (T99)
- **Non-Asymptotic Analysis of Stochastic Approximation Algorithms for Machine Learning**
F. Bach, INRIA - Ecole Normale Supérieure; E. Moulines, Telecom Paristech
Subject Area: Stochastic Methods
See abstract, page 62 (T64)
- **Online Submodular Set Cover, Ranking, and Repeated Active Learning**
A. Guillory, J. Bilmes, University of Washington
Subject Area: Online Learning
See abstract, page 65 (T80)
- **Sparse Estimation with Structured Dictionaries**
D. Wipf, Microsoft Research Asia
Subject Area: Sparsity and Feature Selection
See abstract, page 59 (T52)
- **Universal low-rank matrix recovery from Pauli measurements**
Y. Liu, National Institute of Standards and Technology
Subject Area: Theory
See abstract, page 63 (T71)
- **See the Tree Through the Lines: The Shazoo Algorithm**
F. Vitale, N. Cesa-Bianchi and G. Zappella, Università degli Studi di Milano; C. Gentile, Università dell'Insubria
Subject Area: Online Learning
See abstract, page 66 (T82)
- **On U-processes and clustering performance**
S. Cléménçon, Telecom Paris Tech
Subject Area: Statistical Learning Theory
See abstract, page 66 (T84)

ORAL SESSION

SESSION 4 - 1:10 – 1:30 PM

Session Chair: Shai Shalev-Shwartz

Generalization Bounds and Consistency for Latent Structural Probit and Ramp Loss

David Mcallester mcallester@ttic.edu
Joseph Keshet jkeshet@ttic.edu
TTI-Chicago

We consider latent structural versions of probit loss and ramp loss. We show that these surrogate loss functions are consistent in the strong sense that for any feature map (finite or infinite dimensional) they yield predictors approaching the infimum task loss achievable by any linear predictor over the given features. We also give finite sample generalization bounds (convergence rates) for these loss functions. These bounds suggest that probit loss converges more rapidly. However, ramp loss is more easily optimized and may ultimately be more practical..
Subject Area: Learning with Structured Data

SPOTLIGHT SESSION

SESSION 4 - 1:30 – 2:00 AM

Session Chair: Shai Shalev-Shwartz

- **Algorithms and hardness results for parallel large margin learning**
R. Servedio, Columbia University; P. Long, Google
Subject Area: Learning Theory
See abstract, page 63 (T68)

ORAL SESSION

SESSION 5 - 4:00 – 5:30 PM

Session Chair: Pradeep Ravikumar

INVITED TALK: Sparsity: algorithms, approximations, and analysis

Anna Gilbert annacg@umich.edu
University of Michigan

The last 15 years we have seen an explosion in the role of sparsity in mathematical signal and image processing, signal and image acquisition and reconstruction algorithms, and myriad applications. It is also central to machine learning. I will present an overview of the mathematical theory and several fundamental algorithmic results, including a fun application to solving Sudoku puzzles..

Anna Gilbert received an S.B. degree from the University of Chicago and a Ph.D. from Princeton University, both in mathematics. In 1997, she was a postdoctoral fellow at Yale University and AT&T Labs-Research. From 1998 to 2004, she was a member of technical staff at AT&T Labs-Research in Florham Park, NJ. Since then she has been with the Department of Mathematics at the University of Michigan, where she is now a Professor. She has received several awards, including a Sloan Research Fellowship (2006), an NSF CAREER award (2006), the National Academy of Sciences Award for Initiatives in Research (2008), the Association of Computing Machinery (ACM) Douglas Engelbart

Best Paper award (2008), and the EURASIP Signal Processing Best Paper award (2010). Her research interests include analysis, probability, networking, and algorithms. She is especially interested in randomized algorithms with applications to harmonic analysis, signal and image processing, networking, and massive datasets.

Learning Sparse Representations of High Dimensional Data on Large Scale Dictionaries

Zhen James Xiang zxiang@princeton.edu
 Hao Xu haoxu@princeton.edu
 Peter Ramadge ramadge@princeton.edu
 Princeton University

Learning sparse representations on data adaptive dictionaries is a state-of-the-art method for modeling data. But when the dictionary is large and the data dimension is high, it is a computationally challenging problem. We explore three aspects of the problem. First, we derive new, greatly improved screening tests that quickly identify codewords that are guaranteed to have zero weights. Second, we study the properties of random projections in the context of learning sparse representations. Finally, we develop a hierarchical framework that uses incremental random projections and screening to learn, in small stages, a hierarchically structured dictionary for sparse representations. Empirical results show that our framework can learn informative hierarchical sparse representations more efficiently.

Subject Area: None of the above

High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity

Po-Ling Loh ploh@berkeley.edu
 Martin Wainwright wainwrig@eecs.berkeley.edu
 UC Berkeley

Although the standard formulations of prediction problems involve fully-observed and noiseless data drawn in an i.i.d. manner, many applications involve noisy and/or missing data, possibly involving dependencies. We study these issues in the context of high-dimensional sparse linear regression, and propose novel estimators for the cases of noisy, missing, and/or dependent data. Many standard approaches to noisy or missing data, such as those using the EM algorithm, lead to optimization problems that are inherently non-convex, and it is difficult to establish theoretical guarantees on practical algorithms. While our approach also involves optimizing non-convex programs, we are able to both analyze the statistical error associated with any global optimum, and prove that a simple projected gradient descent algorithm will converge in polynomial time to a small neighborhood of the set of global minimizers. On the statistical side, we provide non-asymptotic bounds that hold with high probability for the cases of noisy, missing, and/or dependent data. On the computational side, we prove that under the same types of conditions required for statistical consistency, the projected gradient algorithm will converge at geometric rates to a near-global minimizer. We illustrate these theoretical predictions with simulations, showing agreement with the predicted scalings. Subject Area: Supervised Learning



POSTER SESSION

AND RECEPTION - 5:45 – 11:59 PM

- T1 A Non-Parametric Approach to Dynamic Programming**, O. Kroemer, J. Peters
- T2 Periodic Finite State Controllers for Efficient POMDP and DEC-POMDP Planning**, J. Pajarinen, J. Peltonen
- T3 Transfer from Multiple MDPs**, A. Lazaric, M. Restelli
- T4 Variance Reduction in Monte-Carlo Tree Search**, J. Veness, M. Lanctot, M. Bowling
- T5 Action-Gap Phenomenon in Reinforcement Learning**, A. Farahmand
- T6 The Fixed Points of Off-Policy TD**, J. Kolter
- T7 Convergent Fitted Value Iteration with Linear Function Approximation**, D. Lizotte
- T8 Blending Autonomous Exploration and Apprenticeship Learning**, T. Walsh, D. Hewlett, C. Morrison
- T9 Selecting the State-Representation in Reinforcement Learning**, O. Maillard, R. Munos, D. Ryabko
- T10 A Reinforcement Learning Theory for Homeostatic Regulation**, M. Keramati, B. Gutkin
- T11 Environmental statistics and the trade-off between model-based and TD learning in humans**, D. Simon, N. Daw
- T12 TD γ : Re-evaluating Complex Backups in Temporal Difference Learning**, G. Konidaris, S. Niekum, P. Thomas
- T13 Evaluating computational models of preference learning**, A. Jern, C. Lucas, C. Kemp
- T14 Inductive reasoning about chimeric creatures**, C. Kemp
- T15 Predicting response time and error rates in visual search**, B. Chen, V. Navalpakkam, P. Perona
- T16 Identifying Alzheimer's Disease-Related Brain Regions from Multi-Modality Neuroimaging Data using Sparse Composite Linear Discrimination Analysis**, S. Huang, J. Li, J. Ye, T. Wu, K. Chen, A. Fleisher, E. Reiman
- T17 Decoding of Finger Flexion from Electrographic Signals Using Switching Non-Parametric Dynamic Systems**, Z. Wang, G. Schalk, Q. Ji

TUESDAY - CONFERENCE

- T18** **Neuronal Adaptation for Sampling-Based Probabilistic Inference in Perceptual Bistability**, D. Reichert, P. Series, A. Storkey
- T19** **Sequence learning with hidden units in spiking neural networks**, J. Brea, W. Senn, J. Pfister
- T20** **Active learning of neural response functions with Gaussian processes**, M. Park, G. Horwitz, J. Pillow
- T21** **Recovering Intrinsic Images with a Global Sparsity Prior on Reflectance**, P. Gehler, C. Rother, M. Kiefel, L. Zhang, B. Schölkopf
- T22** **Semi-supervised Regression via Parallel Field Regularization**, B. Lin, C. Zhang, X. He
- T23** **Video Annotation and Tracking with Active Learning**, C. Vondrick, D. Ramanan
- T24** **Learning Probabilistic Non-Linear Latent Variable Models for Tracking Complex Activities**, A. Yao, J. Gall, L. Gool, R. Urtasun
- T25** **Image Parsing with Stochastic Scene Grammar**, Y. Zhao, S. Zhu
- T26** **Maximal Cliques that Satisfy Hard Constraints with Application to Deformable Object Model Learning**, X. Wang, X. Bai, X. Yang, W. Liu, L. Latecki
- T27** **Generalized Lasso based Approximation of Sparse Coding for Visual Recognition**, N. Morioka, S. Satoh
- T28** **Semantic Labeling of 3D Point Clouds for Indoor Scenes**, H. Koppula, A. Anand, T. Joachims, A. Saxena
- T29** **An Unsupervised Decontamination Procedure For Improving The Reliability Of Human Judgments**, M. Mozer, B. Link, H. Pashler
- T30** **Learning to Search Efficiently in High Dimensions**, Z. Li, H. Ning, L. Cao, T. Zhang, Y. Gong, T. Huang
- T31** **Inferring Interaction Networks using the IBP applied to microRNA Target Prediction**, H. Le, Z. Bar-Joseph
- T32** **Learning Patient-Specific Cancer Survival Distributions as a Sequence of Dependent Regressors**, C. Yu, R. Greiner, H. Lin, V. Baracos
- T33** **History distribution matching method for predicting effectiveness of HIV combination therapies**, J. Bogojeska
- T34** **An Empirical Evaluation of Thompson Sampling**, O. Chapelle, L. Li
- T35** **Hashing Algorithms for Large-Scale Learning**, P. Li, A. Shrivastava, J. Moore, A. König
- T36** **Learning Sparse Representations of High Dimensional Data on Large Scale Dictionaries**, Z. Xiang, H. Xu, P. Ramadge
- T37** **Relative Density-Ratio Estimation for Robust Distribution Comparison**, M. Yamada, T. Suzuki, T. Kanamori, H. Hachiya, M. Sugiyama
- T38** **Sparse Bayesian Multi-Task Learning**, C. Archambeau, S. Guo, O. Zoeter
- T39** **High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity**, P. Loh, M. Wainwright
- T40** **Learning Anchor Planes for Classification**, Z. Zhang, L. Ladicky, P. Torr, A. Saffari
- T41** **ShareBoost: Efficient multiclass learning with feature sharing**, S. Shalev-Shwartz, Y. Wexler, A. Shashua
- T42** **A Two-Stage Weighting Framework for Multi-Source Domain Adaptation**, Q. Sun, R. Chattopadhyay, S. Panchanathan, J. Ye
- T43** **The Local Rademacher Complexity of Lp-Norm Multiple Kernel Learning**, M. Kloft, G. Blanchard
- T44** **Maximum Margin Multi-Label Structured Prediction**, C. Lampert
- T45** **Generalization Bounds and Consistency for Latent Structural Probit and Ramp Loss**, D. Mcallester, J. Keshet
- T46** **Sparse Recovery with Brownian Sensing**, A. Carpentier, O. Maillard, R. Munos
- T47** **Sparse Features for PCA-Like Linear Regression**, C. Boutsidis, P. Drineas, M. Magdon-Ismail
- T48** **Shaping Level Sets with Submodular Functions**, F. Bach
- T49** **Greedy Algorithms for Structurally Constrained High Dimensional Problems**, A. Tewari, P. Ravikumar, I. Dhillon
- T50** **Trace Lasso: a trace norm regularization for correlated designs**, E. Grave, G. Obozinski, F. Bach
- T51** **Robust Lasso with missing and grossly corrupted observations**, N. Nguyen, N. Nasrabadi, T. Tran
- T52** **Sparse Estimation with Structured Dictionaries**, D. Wipf
- T53** **Learning a Distance Metric from a Network**, B. Shaw, B. Huang, T. Jebara
- T54** **A Denoising View of Matrix Completion**, W. Wang, M. Carreira-Perpinan, Z. Lu

TUESDAY - CONFERENCE

- T55 Crowdclustering**, R. Gomes, P. Welinder, A. Krause, P. Perona
- T56 Demixed Principal Component Analysis**, W. Brendel, R. Romo, C. Machens
- T57 Nonnegative dictionary learning in the exponential noise model for adaptive music signal representation**, O. Dikmen, C. Févotte
- T58 Target Neighbor Consistent Feature Weighting for Nearest Neighbor Classification**, I. Takeuchi, M. Sugiyama
- T59 Penalty Decomposition Methods for Rank Minimization**, Y. Zhang, Z. Lu
- T60 Statistical Tests for Optimization Efficiency**, L. Boyles, A. Korattikara, D. Ramanan, M. Welling
- T61 Prismatic Algorithm for Discrete D.C. Programming Problem**, Y. Kawahara, T. Washio
- T62 Sparse Inverse Covariance Matrix Estimation Using Quadratic Approximation**, C. Hsieh, M. Sustik, I. Dhillon, P. Ravikumar
- T63 A Convergence Analysis of Log-Linear Training**, S. Wiesler, H. Ney
- T64 Non-Asymptotic Analysis of Stochastic Approximation Algorithms for Machine Learning**, F. Bach, E. Moulines
- T65 Better Mini-Batch Algorithms via Accelerated Gradient Methods**, A. Cotter, O. Shamir, N. Srebro, K. Sridharan
- T66 PAC-Bayesian Analysis of Contextual Bandits**, Y. Seldin, P. Auer, F. Laviolette, J. Shawe-Taylor, R. Ortner
- T67 Spectral Methods for Learning Multivariate Latent Tree Structure**, A. anandkumar, K. Chaudhuri, D. Hsu, S. Kakade, L. Song, T. Zhang
- T68 Algorithms and hardness results for parallel large margin learning**, R. Servedio, P. Long
- T69 Composite Multiclass Losses**, E. Vernet, R. Williamson, M. Reid
- T70 Autonomous Learning of Action Models for Planning**, N. Mehta, P. Tadepalli, A. Fern
- T71 Universal low-rank matrix recovery from Pauli measurements**, Y. Liu
- T72 A More Powerful Two-Sample Test in High Dimensions using Random Projection**, M. Lopes, L. Jacob, M. Wainwright
- T73 Prediction strategies without loss**, M. Kapralov, R. Panigrahy
- T74 Learning in Hilbert vs. Banach Spaces: A Measure Embedding Viewpoint**, B. Sriperumbudur, K. Fukumizu, G. Lanckriet
- T75 On Strategy Stitching in Large Extensive Form Multiplayer Games**, R. Gibson, D. Szafron
- T76 Multi-Bandit Best Arm Identification**, V. Gabillon, M. Ghavamzadeh, A. Lazaric, S. Bubeck
- T77 Linear Submodular Bandits and their Application to Diversified Retrieval**, Y. Yue, C. Guestrin
- T78 Adaptive Hedge**, T. van Erven, P. Grunwald, W. Koolen, S. Rooij
- T79 On the Universality of Online Mirror Descent**, N. Srebro, K. Sridharan, A. Tewari
- T80 Online Submodular Set Cover, Ranking, and Repeated Active Learning**, A. Guillory, J. Bilmes
- T81 Finite Time Analysis of Stratified Sampling for Monte Carlo**, A. Carpentier, R. Munos
- T82 See the Tree Through the Lines: The Shazoo Algorithm**, F. Vitale, N. Cesa-Bianchi, C. Gentile, G. Zappella
- T83 Generalizing from Several Related Classification Tasks to a New Unlabeled Sample**, G. Blanchard, G. Lee, C. Scott
- T84 On U-processes and clustering performance**, S. Cléménçon
- T85 Information Rates and Optimal Decoding in Large Neural Populations**, K. Rahnama Rad, L. Paninski
- T86 EigenNet: A Bayesian hybrid of generative and conditional models for sparse learning**, Y. Qi, F. Yan
- T87 Learning unbelievable marginal probabilities**, X. Pitkow, Y. Ahmadian, K. Miller
- T88 A blind sparse deconvolution method for neural spike identification**, C. Ekanadham, D. Tranchina, E. Simoncelli
- T89 Accelerated Adaptive Markov Chain for Partition Function Computation**, S. Ermon, C. Gomes, A. Sabharwal, B. Selman
- T90 Message-Passing for Approximate MAP Inference with Latent Variables**, J. Jiang, P. Rai, H. Daume III
- T91 Priors over Recurrent Continuous Time Processes**, A. Saeedi, A. Bouchard-Côté
- T92 Modelling Genetic Variations using Fragmentation-Coagulation Processes**, Y. Teh, C. Blundell, L. Elliott
- T93 Variational Gaussian Process Dynamical Systems**, A. Damianou, M. Titsias, N. Lawrence

TUESDAY - CONFERENCE



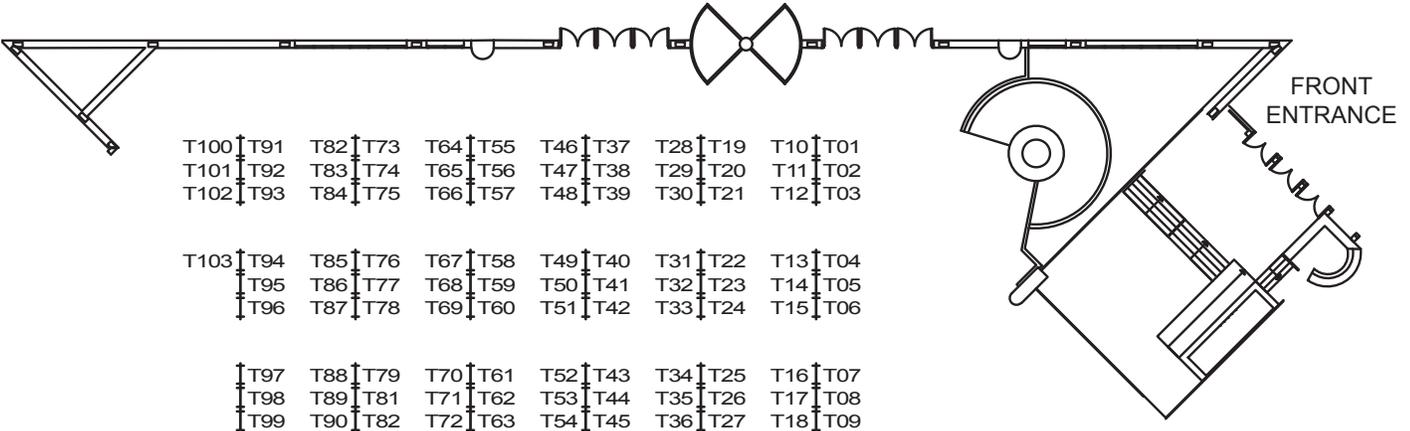
DEMONSTRATIONS

5:45 – 11:59 PM

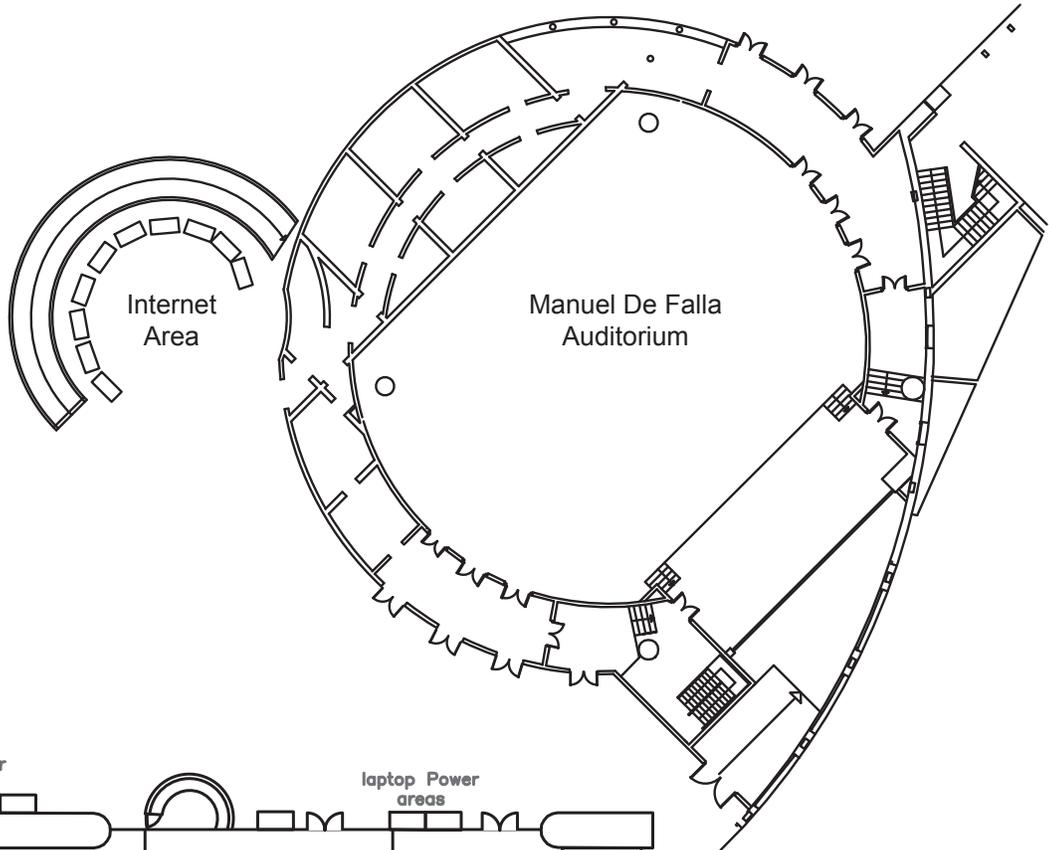
- T94 The Doubly Correlated Nonparametric Topic Model**, D. Kim, E. Sudderth
- T95 The Kernel Beta Process**, L. Ren, Y. Wang, D. Dunson, L. Carin
- T96 An Exact Algorithm for F-Measure Maximization**, K. Dembczynski, W. Waegeman, W. Cheng, E. Hullermeier
- T97 Contextual Gaussian Process Bandit Optimization**, A. Krause, C. Ong
- T98 Automated Refinement of Bayes Networks' Parameters based on Test Ordering Constraints**, O. Khan, P. Poupart, J. Agosta
- T99 Solving Decision Problems with Limited Information**, D. Maua, C. de Campos
- T100 Learning Higher-Order Graph Structure with Features by Structure Penalty**, S. Ding, G. Wahba, X. Zhu
- T101 On the Completeness of First-Order Knowledge Compilation for Lifted Probabilistic Inference**, G. Van den Broeck
- T102 Inference in continuous time changepoint point models**, F. Stimberg, M. Opper, G. Sanguinetti, A. Ruttor
- T103 Comparative Analysis of Viterbi Training and Maximum Likelihood Estimation for HMMs**, A. Allahverdyan, A. Galstyan

- 1A Reproducing biologically realistic firing patterns on a highly-accelerated neuromorphic hardware system**, M. Schwartz
- 2A A smartphone 3D functional brain scanner**, C. Stahlhut, A. Stopczynski, J. Larsen, M. Petersen, L. Hansen
- 3A SENNA Natural Language Processing Demo**, R. Collobert
- 4A Haptic Belt with Pedestrian Detection**, J. Feng, M. Rasi, A. Ng, Q. Le, M. Quigley, J. Chen, T. Low, W. Zou

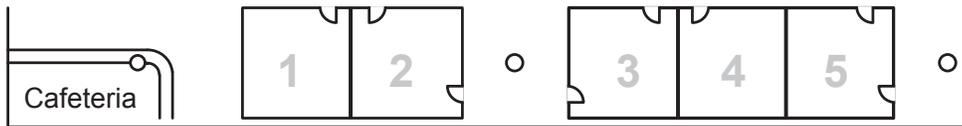
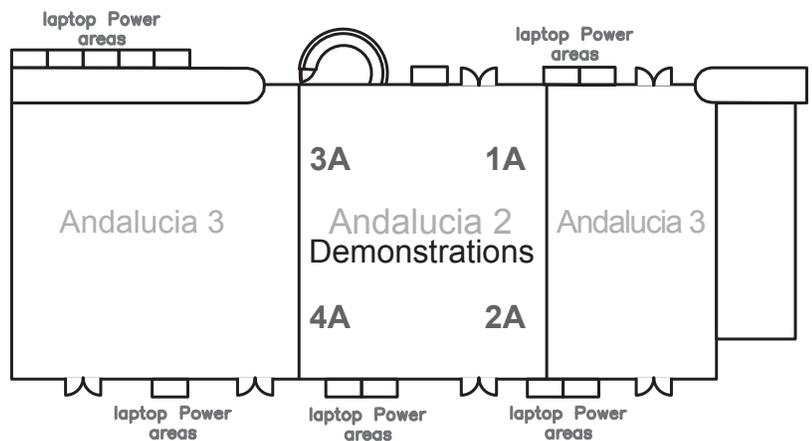
TUESDAY POSTER FLOORPLAN



FLOOR ONE



To Cafeteria
↓



T1 A Non-Parametric Approach to Dynamic Programming

O. Kroemer oliverkro@googlemail.com
 J. Peters mail@jan-peters.net
 Technische Universitaet Darmstadt

In this paper, we consider the problem of policy evaluation for continuous-state systems. We present a non-parametric approach to policy evaluation, which uses kernel density estimation to represent the system. The true form of the value function for this model can be determined, and can be computed using Galerkin's method. Furthermore, we also present a unified view of several well-known policy evaluation methods. In particular, we show that the same Galerkin method can be used to derive Least-Squares Temporal Difference learning, Kernelized Temporal Difference learning, and a discrete-state Dynamic Programming solution, as well as our proposed method. In a numerical evaluation of these algorithms, the proposed approach performed better than the other methods.
 Subject Area: Control and Reinforcement Learning

T2 Periodic Finite State Controllers for Efficient POMDP and DEC-POMDP Planning

J. Pajarinen Joni.Pajarinen@tkk.fi
 J. Peltonen jaakko.peltonen@tkk.fi
 Aalto University

Applications such as robot control and wireless communication require planning under uncertainty. Partially observable Markov decision processes (POMDPs) plan policies for single agents under uncertainty and their decentralized versions (DEC-POMDPs) find a policy for multiple agents. The policy in infinite-horizon POMDP and DEC-POMDP problems has been represented as finite state controllers (FSCs). We introduce a novel class of periodic FSCs, composed of layers connected only to the previous and next layer. Our *periodic* FSC method finds a deterministic finite-horizon policy and converts it to an initial periodic infinite-horizon policy. This policy is optimized by a new infinite-horizon algorithm to yield deterministic periodic policies, and by a new expectation maximization algorithm to yield stochastic periodic policies. Our method yields better results than earlier planning methods and can compute larger solutions than with regular FSCs.
 Subject Area: Control and Reinforcement Learning

T3 Transfer from Multiple MDPs

A. Lazaric alessandro.lazaric@inria.fr
 INRIA Lille-Nord Europe
 M. Restelli restelli@elet.polimi.it
 Politecnico di Milano

Transfer reinforcement learning (RL) methods leverage on the experience collected on a set of source tasks to speed-up RL algorithms. A simple and effective approach is to transfer samples from source tasks and include them in the training set used to solve a target task. In this paper, we investigate the theoretical properties of this transfer method and we introduce novel algorithms adapting the transfer process on the basis of the similarity between

source and target tasks. Finally, we report illustrative experimental results in a continuous chain problem.
 Subject Area: Control and Reinforcement Learning

T4 Variance Reduction in Monte-Carlo Tree Search

J. Veness jveness@gmail.com
 M. Lanctot lanctot@cs.ualberta.ca
 M. Bowling bowling@cs.ualberta.ca
 University of Alberta

Monte-Carlo Tree Search (MCTS) has proven to be a powerful, generic planning technique for decision-making in single-agent and adversarial environments. The stochastic nature of the Monte-Carlo simulations introduces errors in the value estimates, both in terms of bias and variance. Whilst reducing bias (typically through the addition of domain knowledge) has been studied in the MCTS literature, comparatively little effort has focused on reducing variance. This is somewhat surprising, since variance reduction techniques are a well-studied area in classical statistics. In this paper, we examine the application of some standard techniques for variance reduction in MCTS, including common random numbers, antithetic variates and control variates. We demonstrate how these techniques can be applied to MCTS and explore their efficacy on three different stochastic, single-agent settings: Pig, Can't Stop and Dominion.
 Subject Area: Control and Reinforcement Learning

T5 Action-Gap Phenomenon in Reinforcement Learning

A. Farahmand amirf@ualberta.ca
 McGill University

Many practitioners of reinforcement learning problems have observed that oftentimes the performance of the agent reaches very close to the optimal performance even though the estimated (action-)value function is still far from the optimal one. The goal of this paper is to explain and formalize this phenomenon by introducing the concept of the action-gap regularity. As a typical result, we prove that for an agent following the greedy policy π^A with respect to an action-value function Q^A , the performance loss $\mathbb{E}[V^*(X) - V^{\pi^A}(X)]$ is upper bounded by $O(\|Q^A - Q^*\|_{\infty}^{1+\zeta})$, in which $\zeta \geq 0$ is the parameter quantifying the action-gap regularity. For $\zeta > 0$, our results indicate smaller performance loss compared to what previous analyses had suggested. Finally, we show how this regularity affects the performance of the family of approximate value iteration algorithms.
 Subject Area: Control and Reinforcement Learning

T6 The Fixed Points of Off-Policy TD

J. Kolter kolter@csail.mit.edu
 Massachusetts Institute of Technology

Off-policy learning, the ability for an agent to learn about a policy other than the one it is following, is a key element of Reinforcement Learning, and in recent years there has been much work on developing Temporal Different (TD) algorithms that are guaranteed to converge under off-policy sampling. It has remained an open question, however, whether anything can be said a priori about the quality of the TD solution when off-policy sampling is employed with

function approximation. In general the answer is no: for arbitrary off-policy sampling the error of the TD solution can be unboundedly large, even when the approximator can represent the true value function well. In this paper we propose a novel approach to address this problem: we show that by considering a certain convex subset of off-policy distributions we can indeed provide guarantees as to the solution quality similar to the on-policy case. Furthermore, we show that we can efficiently project on to this convex set using only samples generated from the system. The end result is a novel TD algorithm that has approximation guarantees even in the case of off-policy sampling and which empirically outperforms existing TD methods.

Subject Area: Control and Reinforcement Learning

T7 Convergent Fitted Value Iteration with Linear Function Approximation

D. Lizotte dlizotte@uwaterloo.ca
University of Waterloo

Fitted value iteration (FVI) with ordinary least squares regression is known to diverge. We present a new method, "Expansion-Constrained Ordinary Least Squares" (ECOLS), that produces a linear approximation but also guarantees convergence when used with FVI. To ensure convergence, we constrain the least squares regression operator to be a non-expansion in the ∞ -norm. We show that the space of function approximators that satisfy this constraint is more rich than the space of "averagers," we prove a minimax property of the ECOLS residual error, and we give an efficient algorithm for computing the coefficients of ECOLS based on constraint generation. We illustrate the algorithmic convergence of FVI with ECOLS in a suite of experiments, and discuss its properties."

Subject Area: Control and Reinforcement Learning

T8 Blending Autonomous Exploration and Apprenticeship Learning

T. Walsh thomasjwalsh@gmail.com
University of Kansas
D. Hewlett dhewlett@cs.arizona.edu
Google
C. Morrison clayton@sista.arizona.edu
University of Arizona

We present theoretical and empirical results for a framework that combines the benefits of apprenticeship and autonomous reinforcement learning. Our approach modifies an existing apprenticeship learning framework that relies on teacher demonstrations and does not necessarily explore the environment. The first change is replacing previously used Mistake Bound model learners with a recently proposed framework that melds the KWIK and Mistake Bound supervised learning protocols. The second change is introducing a communication of expected utility from the student to the teacher. The resulting system only uses teacher traces when the agent needs to learn concepts it cannot efficiently learn on its own.

Subject Area: Cognitive Science

T9 Selecting the State-Representation in Reinforcement Learning

O. Maillard odalricambrym.maillard@gmail.com
Montan University at Leoben
R. Munos remi.munos@inria.fr
D. Ryabko daniil.ryabko@inria.fr
INRIA

The problem of selecting the right state-representation in a reinforcement learning problem is considered. Several models (functions mapping past observations to a finite set) of the observations are given, and it is known that for at least one of these models the resulting state dynamics are indeed Markovian. Without knowing neither which of the models is the correct one, nor what are the probabilistic characteristics of the resulting MDP, it is required to obtain as much reward as the optimal policy for the correct model (or for the best of the correct models, if there are several). We propose an algorithm that achieves that, with a regret of order $T^{3/4}$ where T is the horizon time.

Subject Area: Cognitive Science

T10 A Reinforcement Learning Theory for Homeostatic Regulation

M. Keramati m.mahdi.keramati@gmail.com
B. Gutkin boris.gutkin@ens.fr
Group for Neural Theory

Reinforcement learning models address animal's behavioral adaptation to its changing "external" environment, and are based on the assumption that Pavlovian, habitual and goal-directed responses seek to maximize reward acquisition. Negative-feedback models of homeostatic regulation, on the other hand, are concerned with behavioral adaptation in response to the "internal" state of the animal, and assume that animals' behavioral objective is to minimize deviations of some key physiological variables from their hypothetical setpoints. Building upon the drive-reduction theory of reward, we propose a new analytical framework that integrates learning and regulatory systems, such that the two seemingly unrelated objectives of reward maximization and physiological-stability prove to be identical. The proposed theory shows behavioral adaptation to both internal and external states in a disciplined way. We further show that the proposed framework allows for a unified explanation of some behavioral phenomenon like motivational sensitivity of different associative learning mechanism, anticipatory responses, interaction among competing motivational systems, and risk aversion."

Subject Area: Cognitive Science

T11 Environmental statistics and the trade-off between model-based and TD learning in humans

D. Simon dylex@nyu.edu
 N. Daw nathaniel.daw@nyu.edu
 New York University

There is much evidence that humans and other animals utilize a combination of model-based and model-free RL methods. Although it has been proposed that these systems may dominate according to their relative statistical efficiency in different circumstances, there is little specific evidence -- especially in humans -- as to the details of this trade-off. Accordingly, we examine the relative performance of different RL approaches under situations in which the statistics of reward are differentially noisy and volatile. Using theory and simulation, we show that model-free TD learning is relatively most disadvantaged in cases of high volatility and low noise. We present data from a decision-making experiment manipulating these parameters, showing that humans shift learning strategies in accord with these predictions. The statistical circumstances favoring model-based RL are also those that promote a high learning rate, which helps explain why, in psychology, the distinction between these strategies is traditionally conceived in terms of rule-based vs. incremental learning. Subject Area: Cognitive Science

T12 TD γ : Re-evaluating Complex Backups in Temporal Difference Learning

G. Konidaris gdk@csail.mit.edu
 Massachusetts Institute of Technology
 S. Niekum sniekum@cs.umass.edu
 P. Thomas pthomas@cs.umass.edu
 University of Massachusetts Amherst

We show that the λ -return target used in the TD(λ) family of algorithms is the maximum likelihood estimator for a specific model of how the variance of an n -step return estimate increases with n . We introduce the γ -return estimator, an alternative target based on a more accurate model of variance, which defines the TD γ family of complex-backup temporal difference learning algorithms. We derive TD γ , the γ -return equivalent of the original TD(λ) algorithm, which eliminates the λ parameter but can only perform updates at the end of an episode and requires time and space proportional to the episode length. We then derive a second algorithm, TD $\gamma(C)$, with a capacity parameter C . TD $\gamma(C)$ requires C times more time and memory than TD(λ) and is incremental and online. We show that TD γ outperforms TD(λ) for any setting of λ on 4 out of 5 benchmark domains, and that TD $\gamma(C)$ performs as well as or better than TD γ for intermediate settings of C . Subject Area: Control and Reinforcement Learning

T13 Evaluating the inverse decision-making approach to preference learning

A. Jern ajern@cmu.edu
 C. Lucas cglucas@gmail.com
 C. Kemp ckemp@cmu.edu
 Carnegie Mellon University

Psychologists have recently begun to develop computational accounts of how people infer others' preferences from their behavior. The inverse decision-making approach proposes that people infer preferences by inverting a generative model of decision-making. Existing data sets, however, do not provide sufficient resolution to thoroughly evaluate this approach. We introduce a new preference learning task that provides a benchmark for evaluating computational accounts and use it to compare the inverse decision-making approach to a feature-based approach, which relies on a discriminative combination of decision features. Our data support the inverse decision-making approach to preference learning. Subject Area: Cognitive Science

T14 Inductive reasoning about chimeric creatures

C. Kemp ckemp@cmu.edu
 Carnegie Mellon University

Given one feature of a novel animal, humans readily make inferences about other features of the animal. For example, winged creatures often fly, and creatures that eat fish often live in the water. We explore the knowledge that supports these inferences and compare two approaches. The first approach proposes that humans rely on abstract representations of dependency relationships between features, and is formalized here as a graphical model. The second approach proposes that humans rely on specific knowledge of previously encountered animals, and is formalized here as a family of exemplar models. We evaluate these models using a task where participants reason about chimeras, or animals with pairs of features that have not previously been observed to co-occur. The results support the hypothesis that humans rely on explicit representations of relationships between features. Subject Area: Cognitive Science

T15 Predicting response time and error rates in visual search

B. Chen bchen3@caltech.edu
 V. Navalpakkam vidhya@caltech.edu
 P. Perona perona@gmail.com
 Caltech

A model of human visual search is proposed. It predicts both response time (RT) and error rates (RT) as a function of image parameters such as target contrast and clutter. The model is an ideal observer, in that it optimizes the Bayes ratio of target present vs target absent. The ratio is computed on the firing pattern of V1/V2 neurons, modeled by Poisson distributions. The optimal mechanism for integrating information over time is shown to be a 'soft max' of diffusions, computed over the visual field by 'hypercolumns' of neurons that share the same receptive field and have different response properties to image features. An approximation of the optimal Bayesian observer, based on integrating local decisions, rather than diffusions, is also derived; it is shown experimentally to produce very similar predictions. A psychophysics experiment is proposed that may discriminate between which mechanism is used in the human brain. Subject Area: Neuroscience

T16 Identifying Alzheimer's Disease-Related Brain Regions from Multi-Modality Neuroimaging Data using Sparse Composite Linear Discrimination Analysis

S. Huang shuang31@asu.edu
 J. Li jing.li.8@asu.edu
 J. Ye jieping.ye@asu.edu
 T. Wu teresa.wu@asu.edu
 Arizona State University
 K. Chen kwei.chen@bannerhealth.com
 A. Fleisher adam.fleisher@bannerhealth.com
 E. Reiman eric.reiman@bannerhealth.com
 Banner Alzheimer's Institute

Diagnosis of Alzheimer's disease (AD) at the early stage of the disease development is of great clinical importance. Current clinical assessment that relies primarily on cognitive measures proves low sensitivity and specificity. The fast growing neuroimaging techniques hold great promise. Research so far has focused on single neuroimaging modalities. However, as different modalities provide complementary measures for the same disease pathology, fusion of multi-modality data may increase the statistical power in identification of disease-related brain regions. This is especially true for early AD, at which stage the disease-related regions are most likely to be weak-effect regions that are difficult to be detected from a single modality alone. We propose a sparse composite linear discriminant analysis model (SCLDA) for identification of disease-related brain regions of early AD from multi-modality data. SCLDA uses a novel formulation that decomposes each LDA parameter into a product of a common parameter shared by all the modalities and a parameter specific to each modality, which enables joint analysis of all the modalities and borrowing strength from one another. We prove that this formulation is equivalent to a penalized likelihood with non-convex regularization, which can be solved by the DC ((difference of convex functions) programming. We show that in using the DC programming, the property of the non-convex regularization in terms of preserving weak-effect features can be nicely revealed. We perform extensive simulations to show that SCLDA outperforms existing competing algorithms on feature selection, especially on the ability for identifying weak-effect features. We apply SCLDA to the Magnetic Resonance Imaging (MRI) and Positron Emission Tomography (PET) images of 49 AD patients and 67 normal controls (NC). Our study identifies disease-related brain regions consistent with findings in the AD literature.

Subject Area: Neuroscience\Brain Imaging

T17 Decoding of Finger Flexion from Electrocorticographic Signals Using Switching Non-Parametric Dynamic Systems

Z. Wang wangz6@rpi.edu
 Q. Ji qji@ecse.rpi.edu
 Rensselaer Polytechnic Institute
 G. Schalk schalk@wadsworth.org
 Wadsworth Center

Brain-computer interfaces (BCIs) use brain signals to

convey a user's intent. Some BCI approaches begin by decoding kinematic parameters of movements from brain signals, and then proceed to using these signals, in absence of movements, to allow a user to control an output. Recent results have shown that electrocorticographic (ECoG) recordings from the surface of the brain in humans can give information about kinematic parameters (e.g., hand velocity or finger flexion). The decoding approaches in these demonstrations usually employed classical classification/regression algorithms that derive a linear mapping between brain signals and outputs. However, they typically only incorporate little prior information about the target kinematic parameter. In this paper, we show that different types of anatomical constraints that govern finger flexion can be exploited in this context. Specifically, we incorporate these constraints in the construction, structure, and the probabilistic functions of a switched non-parametric dynamic system (SNDS) model. We then apply the resulting SNDS decoder to infer the flexion of individual fingers from the same ECoG dataset used in a recent study. Our results show that the application of the proposed model, which incorporates anatomical constraints, improves decoding performance compared to the results in the previous work. Thus, the results presented in this paper may ultimately lead to neurally controlled hand prostheses with full fine-grained finger articulation.

Subject Area: Neuroscience

T18 Neuronal Adaptation for Sampling-Based Probabilistic Inference in Perceptual Bistability

D. Reichert d.p.reichert@sms.ed.ac.uk
 P. Series pseries@inf.ed.ac.uk
 A. Storkey a.storkey@ed.ac.uk
 University of Edinburgh

It has been argued that perceptual multistability reflects probabilistic inference performed by the brain when sensory input is ambiguous. Alternatively, more traditional explanations of multistability refer to low-level mechanisms such as neuronal adaptation. We employ a Deep Boltzmann Machine (DBM) model of cortical processing to demonstrate that these two different approaches can be combined in the same framework. Based on recent developments in machine learning, we show how neuronal adaptation can be understood as a mechanism that improves probabilistic, sampling-based inference. Using the ambiguous Necker cube image, we analyze the perceptual switching exhibited by the model. We also examine the influence of spatial attention, and explore how binocular rivalry can be modeled with the same approach. Our work joins earlier studies in demonstrating how the principles underlying DBMs relate to cortical processing, and offers novel perspectives on the neural implementation of approximate probabilistic inference in the brain.

Subject Area: Neuroscience

T19 Sequence learning with hidden units in spiking neural networks

J. Brea brea@pyl.unibe.ch
 Universit{a}t Bern
 W. Senn senn@pyl.unibe.ch
 J. Pfister jean-pascal.pfister@eng.cam.ac.uk
 Cambridge University

We consider a statistical framework in which recurrent networks of spiking neurons learn to generate spatio-temporal spike patterns. Given biologically realistic stochastic neuronal dynamics we derive a tractable learning rule for the synaptic weights towards hidden and visible neurons that leads to optimal recall of the training sequences. We show that learning synaptic weights towards hidden neurons significantly improves the storing capacity of the network. Furthermore, we derive an approximate online learning rule and show that our learning rule is consistent with Spike-Timing Dependent Plasticity in that if a presynaptic spike shortly precedes a postsynaptic spike, potentiation is induced and otherwise depression is elicited.

Subject Area: Neuroscience

T20 Active learning of neural response functions with Gaussian processes

M. Park mjpark@mail.utexas.edu
 G. Horwitz ghorwitz@uw.edu
 J. Pillow pillow@mail.utexas.edu
 University of Texas at Austin

A sizable literature has focused on the problem of estimating a low-dimensional feature space capturing a neuron's stimulus sensitivity. However, comparatively little work has addressed the problem of estimating the nonlinear function from feature space to a neuron's output spike rate. Here, we use a Gaussian process (GP) prior over the infinite-dimensional space of nonlinear functions to obtain Bayesian estimates of the "nonlinearity" in the linear-nonlinear-Poisson (LNP) encoding model. This offers flexibility, robustness, and computational tractability compared to traditional methods (e.g., parametric forms, histograms, cubic splines). Most importantly, we develop a framework for optimal experimental design based on uncertainty sampling. This involves adaptively selecting stimuli to characterize the nonlinearity with as little experimental data as possible, and relies on a method for rapidly updating hyperparameters using the Laplace approximation. We apply these methods to data from color-tuned neurons in macaque V1. We estimate nonlinearities in the 3D space of cone contrasts, which reveal that V1 combines cone inputs in a highly nonlinear manner. With simulated experiments, we show that optimal design substantially reduces the amount of data required to estimate this nonlinear combination rule."

Subject Area: Neuroscience

T21 Recovering Intrinsic Images with a Global Sparsity Prior on Reflectance

P. Gehler pgehler@mpi-inf.mpg.de
 C. Rother carrot@microsoft.com
 Microsoft
 M. Kiefel martin.kiefel@tuebingen.mpg.de
 L. Zhang lumin.zhang@gmail.com
 B. Schölkopf bs@tuebingen.mpg.de
 Max Planck Institute for Intelligent Systems

We address the challenging task of decoupling material properties from lighting properties given a single image. In the last two decades virtually all works have concentrated on exploiting edge information to address this problem. We take a different route by introducing a new prior on reflectance, that models reflectance values as being drawn from a sparse set of basis colors. This results in a Random Field model with global, latent variables (basis colors) and pixel-accurate output reflectance values. We show that without edge information high-quality results can be achieved, that are on par with methods exploiting this source of information. Finally, we present competitive results by integrating an additional edge model. We believe that our approach is a solid starting point for future development in this domain.

Subject Area: Vision

T22 Semi-supervised Regression via Parallel Field Regularization

B. Lin binbinlin@zju.edu.cn
 C. Zhang pluskid@gmail.com
 X. He xiaofeihe@cad.zju.edu.cn
 Zhejiang University

This paper studies the problem of semi-supervised learning from the vector field perspective. Many of the existing work use the graph Laplacian to ensure the smoothness of the prediction function on the data manifold. However, beyond smoothness, it is suggested by recent theoretical work that we should ensure second order smoothness for achieving faster rates of convergence for semi-supervised regression problems. To achieve this goal, we show that the second order smoothness measures the linearity of the function, and the gradient field of a linear function has to be a parallel vector field. Consequently, we propose to find a function which minimizes the empirical error, and simultaneously requires its gradient field to be as parallel as possible. We give a continuous objective function on the manifold and discuss how to discretize it by using random points. The discretized optimization problem turns out to be a sparse linear system which can be solved very efficiently. The experimental results have demonstrated the effectiveness of our proposed approach.

Subject Area: Vision

T23 Video Annotation and Tracking with Active Learning

C. Vondrick cvondric@ics.uci.edu
D. Ramanan dramanan@ics.uci.edu

We introduce a novel active learning framework for video annotation. By judiciously choosing which frames a user should annotate, we can obtain highly accurate tracks with minimal user effort. We cast this problem as one of active learning, and show that we can obtain excellent performance by querying frames that, if annotated, would produce a large expected change in the estimated object track. We implement a constrained tracker and compute the expected change for putative annotations with efficient dynamic programming algorithms. We demonstrate our framework on four datasets, including two benchmark datasets constructed with key frame annotations obtained by Amazon Mechanical Turk. Our results indicate that we could obtain equivalent labels for a small fraction of the original cost.

Subject Area: Vision\Motion and Tracking

T24 Learning Probabilistic Non-Linear Latent Variable Models for Tracking Complex Activities

A. Yao yaoa@vision.ee.ethz.ch
J. Gall gall@vision.ee.ethz.ch
ETH Zurich
L. Gool vangool@vision.ee.ethz.ch
R. Urtasun rurtasun@ttic.edu
TTI-Chicago

A common approach for handling the complexity and inherent ambiguities of 3D human pose estimation is to use pose priors learned from training data. Existing approaches however, are either too simplistic (linear), too complex to learn, or can only learn latent spaces from “simple data”, i.e., single activities such as walking or running. In this paper, we present an efficient stochastic gradient descent algorithm that is able to learn probabilistic non-linear latent spaces composed of multiple activities. Furthermore, we derive an incremental algorithm for the online setting which can update the latent space without extensive relearning. We demonstrate the effectiveness of our approach on the task of monocular and multi-view tracking and show that our approach outperforms the state-of-the-art.”

Subject Area: Vision\Motion and Tracking

T25 Image Parsing with Stochastic Scene Grammar

Y. Zhao ybzhao@ucla.edu
S. Zhu sczhu@stat.ucla.edu
University of California, Los Angeles

This paper proposes a parsing algorithm for scene understanding which includes four aspects: computing 3D scene layout, detecting 3D objects (e.g. furniture), detecting 2D faces (windows, doors etc.), and segmenting background. In contrast to previous scene labeling work that applied discriminative classifiers to pixels (or super-pixels), we use a generative Stochastic Scene Grammar (SSG). This grammar represents the compositional structures of visual entities from scene categories, 3D

foreground/background, 2D faces, to 1D lines. The grammar includes three types of production rules and two types of contextual relations. Production rules: (i) AND rules represent the decomposition of an entity into sub-parts; (ii) OR rules represent the switching among sub-types of an entity; (iii) SET rules represent an ensemble of visual entities. Contextual relations: (i) Cooperative “+” relations represent positive links between binding entities, such as hinged faces of a object or aligned boxes; (ii) Competitive “-” relations represents negative links between competing entities, such as mutually exclusive boxes. We design an efficient MCMC inference algorithm, namely Hierarchical cluster sampling, to search in the large solution space of scene configurations. The algorithm has two stages: (i) Clustering: It forms all possible higher-level structures (clusters) from lower-level entities by production rules and contextual relations. (ii) Sampling: It jumps between alternative structures (clusters) in each layer of the hierarchy to find the most probable configuration (represented by a parse tree). In our experiment, we demonstrate the superiority of our algorithm over existing methods on public dataset. In addition, our approach achieves richer structures in the parse tree.

Subject Area: Vision\Natural Scene Statistics

T26 Maximal Cliques that Satisfy Hard Constraints with Application to Deformable Object Model Learning

X. Wang wxghust@gmail.com
X. Bai xiang.bai@gmail.com
X. Yang xingwei@temple.edu
W. Liu liuwy@hust.edu.cn
Huazhong University of Science and Technology
L. Latecki latecki@temple.edu
Temple University

We propose a novel inference framework for finding maximal cliques in a weighted graph that satisfy hard constraints. The constraints specify the graph nodes that must belong to the solution as well as mutual exclusions of graph nodes, i.e., sets of nodes that cannot belong to the same solution. The proposed inference is based on a novel particle filter algorithm with state permutations. We apply the inference framework to a challenging problem of learning part-based, deformable object models. Two core problems in the learning framework, matching of image patches and finding salient parts, are formulated as two instances of the problem of finding maximal cliques with hard constraints. Our learning framework yields discriminative part based object models that achieve very good detection rate, and outperform other methods on object classes with large deformation.

Subject Area: Vision\Object Recognition

T27 Generalized Lasso based Approximation of Sparse Coding for Visual Recognition

N. Morioka nmorioka@gmail.com
University of New South Wales
S. Satoh satoh@nii.ac.jp
NII

Sparse coding, a method of explaining sensory data with as few dictionary bases as possible, has attracted much attention in computer vision. For visual object category recognition, ℓ_1 regularized sparse coding is combined with spatial pyramid representation to obtain state-of-the-art performance. However, because of its iterative optimization, applying sparse coding onto every local feature descriptor extracted from an image database can become a major bottleneck. To overcome this computational challenge, this paper presents "Generalized Lasso based Approximation of Sparse coding" (GLAS). By representing the distribution of sparse coefficients with slice transform, we fit a piece-wise linear mapping function with generalized lasso. We also propose an efficient post-refinement procedure to perform mutual inhibition between bases which is essential for an overcomplete setting. The experiments show that GLAS obtains comparable performance to ℓ_1 regularized sparse coding, yet achieves significant speed up demonstrating its effectiveness for large-scale visual recognition problems."

Subject Area: Vision\Object Recognition

T28 Semantic Labeling of 3D Point Clouds for Indoor Scenes

H. Koppula hema@cs.cornell.edu
A. Anand aa755@cs.cornell.edu
T. Joachims tj@cs.cornell.edu
A. Saxena asaxena@cs.cornell.edu
Cornell University

Inexpensive RGB-D cameras that give an RGB image together with depth data have become widely available. In this paper, we use this data to build 3D point clouds of full indoor scenes such as an office and address the task of semantic labeling of these 3D point clouds. We propose a graphical model that captures various features and contextual relations, including the local visual appearance and shape cues, object co-occurrence relationships and geometric relationships. With a large number of object classes and relations, the model's parsimony becomes important and we address that by using multiple types of edge potentials. The model admits efficient approximate inference, and we train it using a maximum-margin learning approach. In our experiments over a total of 52 3D scenes of homes and offices (composed from about 550 views, having 2495 segments labeled with 27 object classes), we get a performance of 84.06% in labeling 17 object classes for offices, and 73.38% in labeling 17 object classes for home scenes. Finally, we applied these algorithms successfully on a mobile robot for the task of finding objects in large cluttered rooms.

Subject Area: Vision\Visual Perception

T29 An Unsupervised Decontamination Procedure For Improving The Reliability Of Human Judgments

M. Mozer mozer@colorado.edu
B. Link link@colorado.edu
University of Colorado
H. Pashler hpashler@ucsd.edu
University of California, San Diego

Psychologists have long been struck by individuals' limitations in expressing their internal sensations, impressions, and evaluations via rating scales. Instead of using an absolute scale, individuals rely on reference points from recent experience. This *relativity of judgment* limits the informativeness of responses on surveys, questionnaires, and evaluation forms. Fortunately, the cognitive processes that map stimuli to responses are not simply noisy, but rather are influenced by recent experience in a lawful manner. We explore techniques to remove sequential dependencies, and thereby *decontaminate* a series of ratings to obtain more meaningful human judgments. In our formulation, the problem is to infer latent (subjective) impressions from a sequence of stimulus labels (e.g., movie names) and responses. We describe an unsupervised approach that simultaneously recovers the impressions and parameters of a contamination model that predicts how recent judgments affect the current response. We test our *iterated impression inference*, or I^3 , algorithm in three domains: rating the gap between dots, the desirability of a movie based on an advertisement, and the morality of an action. We demonstrate significant objective improvements in the quality of the recovered impressions.

Subject Area: Applications

T30 Learning to Search Efficiently in High Dimensions

Z. Li zhenli3@illinois.edu
Y. Gong ygongca@gmail.com
T. Huang huang@ifp.illinois.edu
UIUC
H. Ning huazhong@google.com
Google Inc
L. Cao liangliang.cao@gmail.com
IBM T. J. Watson Research Center
T. Zhang tzhang@stat.rutgers.edu
Rutgers University

High dimensional similarity search in large scale databases becomes an important challenge due to the advent of Internet. For such applications, specialized data structures are required to achieve computational efficiency. Traditional approaches relied on algorithmic constructions that are often data independent (such as Locality Sensitive Hashing) or weakly dependent (such as kd-trees, k -means trees). While supervised learning algorithms have been applied to related problems, those proposed in the literature mainly focused on learning hash codes optimized for compact embedding of the data rather than search efficiency. Consequently such an embedding has to be used with linear scan or another search algorithm. Hence learning to hash does not directly address the search efficiency issue. This paper considers a new framework that applies supervised learning to

directly optimize a data structure that supports efficient large scale search. Our approach takes both search quality and computational cost into consideration. Specifically, we learn a boosted search forest that is optimized using pairwise similarity labeled examples. The output of this search forest can be efficiently converted into an inverted indexing data structure, which can leverage modern text search infrastructure to achieve both scalability and efficiency. Experimental results show that our approach significantly outperforms the start-of-the-art learning to hash methods (such as spectral hashing), as well as state-of-the-art high dimensional search algorithms (such as LSH and k -means trees).

Subject Area: Applications

T31 Inferring Interaction Networks using the IBP applied to microRNA Target Prediction

H. Le hple@cs.cmu.edu
 Z. Bar-Joseph zivbj@cs.cmu.edu
 Carnegie Mellon University

Determining interactions between entities and the overall organization and clustering of nodes in networks is a major challenge when analyzing biological and social network data. Here we extend the Indian Buffet Process (IBP), a nonparametric Bayesian model, to integrate noisy interaction scores with properties of individual entities for inferring interaction networks and clustering nodes within these networks. We present an application of this method to study how microRNAs regulate mRNAs in cells. Analysis of synthetic and real data indicates that the method improves upon prior methods, correctly recovers interactions and clusters, and provides accurate biological predictions.

Subject Area: Applications

T32 Learning Patient-Specific Cancer Survival Distributions as a Sequence of Dependent Regressors

C. Yu chunnam@cs.ualberta.ca
 R. Greiner rgreiner@ualberta.ca
 V. Baracos Vickie.Baracos@albertahealthservices.ca
 University of Alberta
 H. Lin H.Lin-21@sms.ed.ac.uk

An accurate model of patient survival time can help in the treatment and care of cancer patients. The common practice of providing survival time estimates based only on population averages for the site and stage of cancer ignores many important individual differences among patients. In this paper, we propose a local regression method for learning patient-specific survival time distribution based on patient attributes such as blood tests and clinical assessments. When tested on a cohort of more than 2000 cancer patients, our method gives survival time predictions that are much more accurate than popular survival analysis models such as the Cox and Aalen regression models. Our results also show that using patient-specific attributes can reduce the prediction error on survival time by as much as 20% when compared to using cancer site and stage only.

Subject Area: Applications

T33 History distribution matching method for predicting effectiveness of HIV combination therapies

J. Bogojeska jasmına@mpi-inf.mpg.de
 Max-Planck Institute for Informatics

This paper presents an approach that predicts the effectiveness of HIV combination therapies by simultaneously addressing several problems affecting the available HIV clinical data sets: the different treatment backgrounds of the samples, the uneven representation of the levels of therapy experience, the missing treatment history information, the uneven therapy representation and the unbalanced therapy outcome representation. The computational validation on clinical data shows that, compared to the most commonly used approach that does not account for the issues mentioned above, our model has significantly higher predictive power. This is especially true for samples stemming from patients with longer treatment history and samples associated with rare therapies. Furthermore, our approach is at least as powerful for the remaining samples.

Subject Area: Applications

T34 An Empirical Evaluation of Thompson Sampling

O. Chapelle chap@yahoo-inc.com
 L. Li lihongli.cs@gmail.com
 Yahoo! Research

Thompson sampling is one of oldest heuristic to address the exploration / exploitation trade-off, but it is surprisingly not very popular in the literature. We present here some empirical results using Thompson sampling on simulated and real data, and show that it is highly competitive. And since this heuristic is very easy to implement, we argue that it should be part of the standard baselines to compare against.

Subject Area: Applications

T35 Hashing Algorithms for Large-Scale Learning

P. Li pingli@cornell.edu
 A. Shrivastava anshu@cs.cornell.edu
 J. Moore jlmo@cs.cornell.edu
 Cornell University
 A. König chrisko@microsoft.com
 Microsoft Research

Minwise hashing is a standard technique in the context of search for efficiently computing set similarities. The recent development of b -bit minwise hashing provides a substantial improvement by storing only the lowest b bits of each hashed value. In this paper, we demonstrate that b -bit minwise hashing can be naturally integrated with linear learning algorithms such as linear SVM and logistic regression, to solve large-scale and high-dimensional statistical learning tasks, especially when the data do not fit in memory. We compare b -bit minwise hashing with the Count-Min (CM) and Vowpal Wabbit (VW) algorithms, which have essentially the same variances as random projections. Our theoretical and empirical comparisons illustrate that b -bit minwise hashing is significantly more accurate (at the same storage cost) than VW (and random projections) for binary data.

Subject Area: Dimensionality Reduction

T36 Learning Sparse Representations of High Dimensional Data on Large Scale Dictionaries

Z. Xiang zxiang@princeton.edu
 H. Xu haoxu@princeton.edu
 P. Ramadge ramadge@princeton.edu
 Princeton University

Learning sparse representations on data adaptive dictionaries is a state-of-the-art method for modeling data. But when the dictionary is large and the data dimension is high, it is a computationally challenging problem. We explore three aspects of the problem. First, we derive new, greatly improved screening tests that quickly identify codewords that are guaranteed to have zero weights. Second, we study the properties of random projections in the context of learning sparse representations. Finally, we develop a hierarchical framework that uses incremental random projections and screening to learn, in small stages, a hierarchically structured dictionary for sparse representations. Empirical results show that our framework can learn informative hierarchical sparse representations more efficiently.

Subject Area: None of the above

T37 Relative Density-Ratio Estimation for Robust Distribution Comparison

M. Yamada yamada@sg.cs.titech.ac.jp
 H. Hachiya hachiya@sg.cs.titech.ac.jp
 M. Sugiyama sugi@cs.titech.ac.jp
 Tokyo Institute of Technology
 T. Suzuki s-taiji@stat.t.u-tokyo.ac.jp
 University of Tokyo
 T. Kanamori kanamori@is.nagoya-u.ac.jp
 Nagoya university

Divergence estimators based on direct approximation of density-ratios without going through separate approximation of numerator and denominator densities have been successfully applied to machine learning tasks that involve distribution comparison such as outlier detection, transfer learning, and two-sample homogeneity test. However, since density-ratio functions often possess high fluctuation, divergence estimation is still a challenging task in practice. In this paper, we propose to use relative divergences for distribution comparison, which involves approximation of relative density-ratios. Since relative density-ratios are always smoother than corresponding ordinary density-ratios, our proposed method is favorable in terms of the non-parametric convergence speed. Furthermore, we show that the proposed divergence estimator has asymptotic variance independent of the model complexity under a parametric setup, implying that the proposed estimator hardly overfits even with complex models. Through experiments, we demonstrate the usefulness of the proposed approach.

Subject Area: Algorithms and Architectures

T38 Sparse Bayesian Multi-Task Learning

C. Archambeau cedric.p.archambeau@gmail.com
 S. Guo shengboguo@gmail.com
 O. Zoeter Onno.Zoeter@xrce.xerox.com
 Xerox Research Centre Europe

We propose a new sparse Bayesian model for multi-task regression and classification. The model is able to capture correlations between tasks, or more specifically a low-rank approximation of the covariance matrix, while being sparse in the features. We introduce a general family of group sparsity inducing priors based on matrix-variate Gaussian scale mixtures. We show the amount of sparsity can be learnt from the data by combining an approximate inference approach with type II maximum likelihood estimation of the hyperparameters. Empirical evaluations on data sets from biology and vision demonstrate the applicability of the model, where on both regression and classification tasks it achieves competitive predictive performance compared to previously proposed methods.

Subject Area: Supervised Learning

T39 High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity

P. Loh ploh@berkeley.edu
 M. Wainwright wainwrig@eecs.berkeley.edu
 UC Berkeley

Although the standard formulations of prediction problems involve fully-observed and noiseless data drawn in an i.i.d. manner, many applications involve noisy and/or missing data, possibly involving dependencies. We study these issues in the context of high-dimensional sparse linear regression, and propose novel estimators for the cases of noisy, missing, and/or dependent data. Many standard approaches to noisy or missing data, such as those using the EM algorithm, lead to optimization problems that are inherently non-convex, and it is difficult to establish theoretical guarantees on practical algorithms. While our approach also involves optimizing non-convex programs, we are able to both analyze the statistical error associated with any global optimum, and prove that a simple projected gradient descent algorithm will converge in polynomial time to a small neighborhood of the set of global minimizers. On the statistical side, we provide non-asymptotic bounds that hold with high probability for the cases of noisy, missing, and/or dependent data. On the computational side, we prove that under the same types of conditions required for statistical consistency, the projected gradient descent algorithm will converge at geometric rates to a near-global minimizer. We illustrate these theoretical predictions with simulations, showing agreement with the predicted scalings.

Subject Area: Supervised Learning

T40 Learning Anchor Planes for Classification

Z. Zhang ziming.zhang@brookes.ac.uk
 P. Torr philiptorr@brookes.ac.uk
 Oxford Brookes
 L. Ladicky lubor@robots.ox.ac.uk
 University of Oxford
 A. Saffari amir@ymer.org
 Sony Computer Entertainment Europe

Local Coordinate Coding (LCC) [18] is a method for modeling functions of data lying on non-linear manifolds. It provides a set of anchor points which form a local coordinate system, such that each data point on the manifold can be approximated by a linear combination of its anchor points, and the linear weights become the local coordinate coding. In this paper we propose encoding data using orthogonal anchor planes, rather than anchor points. Our method needs only a few orthogonal anchor planes for coding, and it can linearize any (α, β, p) -Lipschitz smooth nonlinear function with a fixed expected value of the upper-bound approximation error on any high dimensional data. In practice, the orthogonal coordinate system can be easily learned by minimizing this upper bound using singular value decomposition (SVD). We apply our method to model the coordinates locally in linear SVMs for classification tasks, and our experiment on MNIST shows that using only 50 anchor planes our method achieves 1.72% error rate, while LCC achieves 1.90% error rate using 4096 anchor points.

Subject Area: Supervised Learning\Classification

T41 ShareBoost: Efficient multiclass learning with feature sharing

S. Shalev-Shwartz shai.shwartz@gmail.com
 Y. Wexler Yonatan.Wexler@orcam.com
 A. Shashua shashua@cs.huji.ac.il
 Hebrew University of Jerusalem

Multiclass prediction is the problem of classifying an object into a relevant target class. We consider the problem of learning a multiclass predictor that uses only few features, and in particular, the number of used features should increase sub-linearly with the number of possible classes. This implies that features should be shared by several classes. We describe and analyze the ShareBoost algorithm for learning a multiclass predictor that uses few shared features. We prove that ShareBoost efficiently finds a predictor that uses few shared features (if such a predictor exists) and that it has a small generalization error. We also describe how to use ShareBoost for learning a non-linear predictor that has a fast evaluation time. In a series of experiments with natural data sets we demonstrate the benefits of ShareBoost and evaluate its success relatively to other state-of-the-art approaches.

Subject Area: Supervised Learning\Classification

T42 A Two-Stage Weighting Framework for Multi-Source Domain Adaptation

Q. Sun qian_sun@asu.edu
 R. Chattopadhyay rchattop@asu.edu
 S. Panchanathan panch@asu.edu
 J. Ye jiejing.ye@asu.edu
 Arizona State University

Discriminative learning when training and test data belong to different distributions is a challenging and complex task. Often times we have very few or no labeled data from the test or target distribution but may have plenty of labeled data from multiple related sources with different distributions. The difference in distributions may be in both marginal and conditional probabilities. Most of the existing domain adaptation work focuses on the marginal probability distribution difference between the domains, assuming that the conditional probabilities are similar. However in many real world applications, conditional probability distribution differences are as commonplace as marginal probability differences. In this paper we propose a two-stage domain adaptation methodology which combines weighted data from multiple sources based on marginal probability differences (first stage) as well as conditional probability differences (second stage), with the target domain data. The weights for minimizing the marginal probability differences are estimated independently, while the weights for minimizing conditional probability differences are computed simultaneously by exploiting the potential interaction among multiple sources. We also provide a theoretical analysis on the generalization performance of the proposed multi-source domain adaptation formulation using the weighted Rademacher complexity measure. Empirical comparisons with existing state-of-the-art domain adaptation methods using three real-world datasets demonstrate the effectiveness of the proposed approach.

Subject Area: Supervised Learning\Classification

T43 The Local Rademacher Complexity of L_p -Norm Multiple Kernel Learning

M. Kloft mkloft@mail.tu-berlin.de
 TU Berlin
 G. Blanchard gilles.blanchard@math.uni-potsdam.de
 Universit"t Potsdam (DE)

We derive an upper bound on the local Rademacher complexity of ℓ_p -norm multiple kernel learning, which yields a tighter excess risk bound than global approaches. Previous local approaches analyzed the case $p = 1$ only while our analysis covers all cases $1 \leq p \leq \infty$, assuming the different feature mappings corresponding to the different kernels to be uncorrelated. We also show a lower bound that shows that the bound is tight, and derive consequences regarding excess loss, namely fast convergence rates of the order $O(n^{-\frac{\alpha}{\alpha+1}})$, where α is the minimum eigenvalue decay rate of the individual kernels.

Subject Area: Supervised Learning\Kernel Methods

T44 Maximum Margin Multi-Label Structured Prediction

C. Lampert chl@ist.ac.at
IST Austria

We study multi-label prediction for structured output spaces, a problem that occurs, for example, in object detection in images, secondary structure prediction in computational biology, and graph matching with symmetries. Conventional multi-label classification techniques are typically not applicable in this situation, because they require explicit enumeration of the label space, which is infeasible in case of structured outputs. Relying on techniques originally designed for single-label structured prediction, in particular structured support vector machines, results in reduced prediction accuracy, or leads to infeasible optimization problems. In this work we derive a maximum-margin training formulation for multi-label structured prediction that remains computationally tractable while achieving high prediction accuracy. It also shares most beneficial properties with single-label maximum-margin approaches, in particular a formulation as a convex optimization problem, efficient working set training, and PAC-Bayesian generalization bounds.
Subject Area: Supervised Learning

T45 Generalization Bounds and Consistency for Latent Structural Probit and Ramp Loss

D. Mcallester mcallester@ttic.edu
J. Keshet jkeshet@ttic.edu
TTI-Chicago

We consider latent structural versions of probit loss and ramp loss. We show that these surrogate loss functions are consistent in the strong sense that for any feature map (finite or infinite dimensional) they yield predictors approaching the infimum task loss achievable by any linear predictor over the given features. We also give finite sample generalization bounds (convergence rates) for these loss functions. These bounds suggest that probit loss converges more rapidly. However, ramp loss is more easily optimized and may ultimately be more practical.
Subject Area: Supervised Learning

T46 Sparse Recovery with Brownian Sensing

A. Carpentier alexandra.carpentier@inria.fr
R. Munos remi.munos@inria.fr
INRIA Lille - Nord Europe
O. Maillard odalricambrym.maillard@gmail.com
Montanuniversit{a}t Leoben

We consider the problem of recovering the parameter $\alpha \in \mathbb{R}^k$ of a sparse function f , i.e. the number of non-zero entries of α is small compared to the number K of features, given noisy evaluations of f at a set of well-chosen sampling points. We introduce an additional randomisation process, called Brownian sensing, based on the computation of stochastic integrals, which produces a Gaussian sensing matrix, for which good recovery properties are proven independently on the number of sampling points N , even

when the features are arbitrarily non-orthogonal. Under the assumption that f is Hölder continuous with exponent at least $1/2$, we provide an estimate $\hat{\alpha}$ of the parameter such that $\|\hat{\alpha} - \alpha\|_2 = O(\|\eta\|_2/\sqrt{N})$, where η is the observation noise. The method uses a set of sampling points uniformly distributed along a one-dimensional curve selected according to the features. We report numerical experiments illustrating our method.

Subject Area: Supervised Learning

T47 Sparse Features for PCA-Like Linear Regression

C. Boutsidis boutsc@cs.rpi.edu
P. Drineas drinep@cs.rpi.edu
M. Magdon-Ismail magdon@cs.rpi.edu
RPI

Principal Components Analysis (PCA) is often used as a feature extraction procedure. Given a matrix $X \in \mathbb{R}^{n \times d}$, whose rows represent n data points with respect to d features, the top k right singular vectors of X (the so-called *eigenfeatures*), are arbitrary linear combinations of all available features. The eigenfeatures are very useful in data analysis, including the regularization of linear regression. Enforcing sparsity on the eigenfeatures, i.e., forcing them to be linear combinations of only a *small* number of actual features (as opposed to all available features), can promote better generalization error and improve the interpretability of the eigenfeatures. We present deterministic and randomized algorithms that construct such sparse eigenfeatures while *probably* achieving in-sample performance comparable to regularized linear regression. Our algorithms are relatively simple and practically efficient, and we demonstrate their performance on several data sets.

Subject Area: Supervised Learning

T48 Shaping Level Sets with Submodular Functions

F. Bach francis.bach@ens.fr
INRIA - Ecole Normale Supérieure

We consider a class of sparsity-inducing regularization terms based on submodular functions. While previous work has focused on non-decreasing functions, we explore symmetric submodular functions and their Lov{a}z extensions. We show that the Lov{a}z extension may be seen as the convex envelope of a function that depends on level sets (i.e., the set of indices whose corresponding components of the underlying predictor are greater than a given constant): this leads to a class of convex structured regularization terms that impose prior knowledge on the level sets, and not only on the supports of the underlying predictors. We provide a unified set of optimization algorithms, such as proximal operators, and theoretical guarantees (allowed level sets and recovery conditions). By selecting specific submodular functions, we give a new interpretation to known norms, such as the total variation; we also define new norms, in particular ones that are based on order statistics with application to clustering and outlier detection, and on noisy cuts in graphs with application to change point detection in the presence of outliers.

Subject Area: Supervised Learning

T49 Greedy Algorithms for Structurally Constrained High Dimensional Problems

A. Tewari ambujtewari@gmail.com
 P. Ravikumar pradeepr@cs.utexas.edu
 I. Dhillon inderjit@cs.utexas.edu
 University of Texas at Austin

A hallmark of modern machine learning is its ability to deal with high dimensional problems by exploiting structural assumptions that limit the degrees of freedom in the underlying model. A deep understanding of the capabilities and limits of high dimensional learning methods under specific assumptions such as sparsity, group sparsity, and low rank has been attained. Efforts (Negahban et al., 2010, Chandrasekaran et al., 2010) are now underway to distill this valuable experience by proposing general unified frameworks that can achieve the twin goals of summarizing previous analyses and enabling their application to notions of structure hitherto unexplored. Inspired by these developments, we propose and analyze a general computational scheme based on a greedy strategy to solve convex optimization problems that arise when dealing with structurally constrained high-dimensional problems. Our framework not only unifies existing greedy algorithms by recovering them as special cases but also yields novel ones. Finally, we extend our results to infinite dimensional problems by using interesting connections between smoothness of norms and behavior of martingales in Banach spaces.

Subject Area: Supervised Learning

T50 Trace Lasso: a trace norm regularization for correlated designs

E. Grave edouard.grave@inria.fr
 G. Obozinski Guillaume.Obozinski@ens.fr
 F. Bach francis.bach@ens.fr
 INRIA - Ecole Normale Supérieure

Using the ℓ_1 -norm to regularize the estimation of the parameter vector of a linear model leads to an unstable estimator when covariates are highly correlated. In this paper, we introduce a new penalty function which takes into account the correlation of the design matrix to stabilize the estimation. This norm, called the trace Lasso, uses the trace norm of the selected covariates, which is a convex surrogate of their rank, as the criterion of model complexity. We analyze the properties of our norm, describe an optimization algorithm based on reweighted least-squares, and illustrate the behavior of this norm on synthetic data, showing that it is more adapted to strong correlations than competing methods such as the elastic net.

Subject Area: Supervised Learning

T51 Robust Lasso with missing and grossly corrupted observations

N. Nguyen namctin@gmail.com
 T. Tran tract@jhu.edu
 Johns Hopkins University
 N. Nasrabadi nasser.m.nasrabadi@us.army.mil
 Army Research Lab

This paper studies the problem of accurately recovering a sparse vector β^* from highly corrupted linear measurements $y = \beta^* + e^* + w$ where e^* is a sparse error vector whose nonzero entries may be unbounded and w is a bounded noise. We propose a so-called extended Lasso optimization which takes into consideration sparse prior information of both β^* and e^* . Our first result shows that the extended Lasso can faithfully recover both the regression and the corruption vectors. Our analysis is relied on a notion of extended restricted eigenvalue for the design matrix X . Our second set of results applies to a general class of Gaussian design matrix X with i.i.d rows $N(0, \Sigma)$, for which we provide a surprising phenomenon: the extended Lasso can recover exact signed supports of both β^* and e^* from only $\Omega(k \log p \log n)$ observations, even the fraction of corruption is arbitrarily close to one. Our analysis also shows that this amount of observations required to achieve exact signed support is optimal.

Subject Area: Supervised Learning

T52 Sparse Estimation with Structured Dictionaries

D. Wipf davidwipf@gmail.com
 Microsoft Research Asia

In the vast majority of recent work on sparse estimation algorithms, performance has been evaluated using ideal or quasi-ideal dictionaries (e.g., random Gaussian or Fourier) characterized by unit ℓ_2 norm, incoherent columns or features. But in reality, these types of dictionaries represent only a subset of the dictionaries that are actually used in practice (largely restricted to idealized compressive sensing applications). In contrast, herein sparse estimation is considered in the context of structured dictionaries possibly exhibiting high coherence between arbitrary groups of columns and/or rows. Sparse penalized regression models are analyzed with the purpose of finding, to the extent possible, regimes of dictionary invariant performance. In particular, a Type II Bayesian estimator with a dictionary-dependent sparsity penalty is shown to have a number of desirable invariance properties leading to provable advantages over more conventional penalties such as the ℓ_1 norm, especially in areas where existing theoretical recovery guarantees no longer hold. This can translate into improved performance in applications such as model selection with correlated features, source localization, and compressive sensing with constrained measurement directions.

Subject Area: Supervised Learning

T53 Learning a Distance Metric from a Network

B. Shaw blake@cs.columbia.edu
 B. Huang bert@cs.columbia.edu
 T. Jebara jebara@cs.columbia.edu
 Columbia University

Many real-world networks are described by both connectivity information and features for every node. To better model and understand these networks, we present structure preserving metric learning (SPML), an algorithm for learning a Mahalanobis distance metric from a network such that the learned distances are tied to the inherent connectivity structure of the network. Like

the graph embedding algorithm structure preserving embedding, SPML learns a metric which is structure preserving, meaning a connectivity algorithm such as k-nearest neighbors will yield the correct connectivity when applied using the distances from the learned metric. We show a variety of synthetic and real-world experiments where SPML predicts link patterns from node features more accurately than standard techniques. We further demonstrate a method for optimizing SPML based on stochastic gradient descent which removes the running-time dependency on the size of the network and allows the method to easily scale to networks of thousands of nodes and millions of edges.

Subject Area: Unsupervised & Semi-supervised Learning

T54 A Denoising View of Matrix Completion

W. Wang wwang5@ucmerced.edu
 M. Carreira-Perpinan mcarreira-perpinan@ucmerced.edu
 University of California, Merced
 Z. Lu zhengdol@microsoft.com
 MSRA

In matrix completion, we are given a matrix where the values of only some of the entries are present, and we want to reconstruct the missing ones. Much work has focused on the assumption that the data matrix has low rank. We propose a more general assumption based on denoising, so that we expect that the value of a missing entry can be predicted from the values of neighboring points. We propose a nonparametric version of denoising based on local, iterated averaging with mean-shift, possibly constrained to preserve local low-rank manifold structure. The few user parameters required (the denoising scale, number of neighbors and local dimensionality) and the number of iterations can be estimated by cross-validating the reconstruction error. Using our algorithms as a postprocessing step on an initial reconstruction (provided by e.g. a low-rank method), we show consistent improvements with synthetic, image and motion-capture data.

Subject Area: Unsupervised & Semi-supervised Learning

T55 Crowdclustering

R. Gomes gomes@caltech.edu
 P. Welinder welinder@caltech.edu
 Caltech
 A. Krause krausea@ethz.ch
 ETH Zurich
 P. Perona andrea@vision.caltech.edu
 Caltech

Is it possible to crowdsource categorization? Amongst the challenges: (a) each annotator has only a partial view of the data, (b) different annotators may have different clustering criteria and may produce different numbers of categories, (c) the underlying category structure may be hierarchical. We propose a Bayesian model of how annotators may approach clustering and show how one may infer clusters/categories, as well as annotator parameters, using this model. Our experiments, carried out on large collections of images, suggest that Bayesian crowdclustering works well and may be superior to single-expert annotations.

Subject Area: Unsupervised & Semi-supervised Learning

T56 Demixed Principal Component Analysis

W. Brendel wieland.brendel@ens.fr
 Ecole Normale Supérieure
 R. Romo rromo@ifc.unam.mx
 Universidad Autónoma de México
 C. Machens christian.machens@neuro.fchampalimaud.org

In many experiments, the data points collected live in high-dimensional observation spaces, yet can be assigned a set of labels or parameters. In electrophysiological recordings, for instance, the responses of populations of neurons generally depend on mixtures of experimentally controlled parameters. The heterogeneity and diversity of these parameter dependencies can make visualization and interpretation of such data extremely difficult. Standard dimensionality reduction techniques such as principal component analysis (PCA) can provide a succinct and complete description of the data, but the description is constructed independent of the relevant task variables and is often hard to interpret. Here, we start with the assumption that a particularly informative description is one that reveals the dependency of the high-dimensional data on the individual parameters. We show how to modify the loss function of PCA so that the principal components seek to capture both the maximum amount of variance about the data, while also depending on a minimum number of parameters. We call this method demixed principal component analysis (dPCA) as the principal components here segregate the parameter dependencies. We phrase the problem as a probabilistic graphical model, and present a fast Expectation-Maximization (EM) algorithm. We demonstrate the use of this algorithm for electrophysiological data and show that it serves to demix the parameter-dependence of a neural population response.

Subject Area: Unsupervised & Semi-supervised Learning

T57 Nonnegative dictionary learning in the exponential noise model for adaptive music signal representation

O. Dikmen dikmen@telecom-paristech.fr
 C. Févotte fevotte@telecom-paristech.fr
 Telecom ParisTech

In this paper we describe a maximum likelihood likelihood approach for dictionary learning in the multiplicative exponential noise model. This model is prevalent in audio signal processing where it underlies a generative composite model of the power spectrogram. Maximum joint likelihood estimation of the dictionary and expansion coefficients leads to a nonnegative matrix factorization problem where the Itakura-Saito divergence is used. The optimality of this approach is in question because the number of parameters (which include the expansion coefficients) grows with the number of observations. In this paper we describe a variational procedure for optimization of the marginal likelihood, i.e., the likelihood of the dictionary where the activation coefficients have been integrated out (given a specific prior). We compare the output of both maximum joint likelihood estimation (i.e., standard Itakura-Saito NMF) and maximum marginal likelihood estimation (MMLE) on real and synthetic datasets. The MMLE

approach is shown to embed automatic model order selection, akin to automatic relevance determination.

Subject Area: Unsupervised & Semi-supervised Learning

T58 Target Neighbor Consistent Feature Weighting for Nearest Neighbor Classification

I. Takeuchi takeuchi.ichiro@nitech.ac.jp
Nagoya Institute of Technology
M. Sugiyama sugi@cs.titech.ac.jp
Tokyo Institute of Technology

We consider feature selection and weighting for nearest neighbor classifiers. A technical challenge in this scenario is how to cope with the discrete update of nearest neighbors when the feature space metric is changed during the learning process. This issue, called the target neighbor change, was not properly addressed in the existing feature weighting and metric learning literature. In this paper, we propose a novel feature weighting algorithm that can exactly and efficiently keep track of the correct target neighbors via sequential quadratic programming. To the best of our knowledge, this is the first algorithm that guarantees the consistency between target neighbors and the feature space metric. We further show that the proposed algorithm can be naturally combined with regularization path tracking, allowing computationally efficient selection of the regularization parameter. We demonstrate the effectiveness of the proposed algorithm through experiments.

Subject Area: Unsupervised & Semi-supervised Learning

T59 Penalty Decomposition Methods for Rank Minimization

Y. Zhang yza30@sfu.ca
Z. Lu zhaosong@sfu.ca
Simon Fraser University

In this paper we consider general rank minimization problems with rank appearing in either objective function or constraint. We first show that a class of matrix optimization problems can be solved as lower dimensional vector optimization problems. As a consequence, we establish that a class of rank minimization problems have closed form solutions. Using this result, we then propose penalty decomposition methods for general rank minimization problems. The convergence results of the PD methods have been shown in the longer version of the paper. Finally, we test the performance of our methods by applying them to matrix completion and nearest low-rank correlation matrix problems. The computational results demonstrate that our methods generally outperform the existing methods in terms of solution quality and/or speed.

Subject Area: Optimization

T60 Statistical Tests for Optimization Efficiency

L. Boyles lboyles@uci.edu
A. Korattikara akoratti@uci.edu
D. Ramanan dramanan@ics.uci.edu
M. Welling welling@ics.uci.edu
University of California Irvine

Learning problems such as logistic regression are typically formulated as pure optimization problems defined on some loss function. We argue that this view ignores the fact that the loss function depends on stochastically generated data which in turn determines an intrinsic scale of precision for statistical estimation. By considering the statistical properties of the update variables used during the optimization (e.g. gradients), we can construct frequentist hypothesis tests to determine the reliability of these updates. We utilize subsets of the data for computing updates, and use the hypothesis tests for determining when the batch-size needs to be increased. This provides computational benefits and avoids overfitting by stopping when the batch-size has become equal to size of the full dataset. Moreover, the proposed algorithms depend on a single interpretable parameter --- the probability for an update to be in the wrong direction --- which is set to a single value across all algorithms and datasets. In this paper, we illustrate these ideas on three L_1 regularized coordinate algorithms: L_1 -regularized L_2 -loss SVMs, L_1 -regularized logistic regression, and the Lasso, but we emphasize that the underlying methods are much more generally applicable.

Subject Area: Optimization

T61 Prismatic Algorithm for Discrete D.C. Programming Problem

Y. Kawahara kawahara@ar.sanken.osaka-u.ac.jp
T. Washio washio@ar.sanken.osaka-u.ac.jp
Osaka University

In this paper, we propose the first exact algorithm for minimizing the difference of two submodular functions (D.S.), i.e., the discrete version of the D.C. programming problem. The developed algorithm is a branch-and-bound-based algorithm which responds to the structure of this problem through the relationship between submodularity and convexity. The D.S. programming problem covers a broad range of applications in machine learning because this generalizes the optimization of a wide class of set functions. We empirically investigate the performance of our algorithm, and illustrate the difference between exact and approximate solutions respectively obtained by the proposed and existing algorithms in feature selection and discriminative structure learning.

Subject Area: Optimization

T62 Sparse Inverse Covariance Matrix Estimation Using Quadratic Approximation

C. Hsieh cjhsieh@cs.utexas.edu
M. Sustik sustik@cs.utexas.edu
I. Dhillon inderjit@cs.utexas.edu
P. Ravikumar pradeepr@cs.utexas.edu
University of Texas, Austin

The L_1 regularized Gaussian maximum likelihood estimator has been shown to have strong statistical guarantees in recovering a sparse inverse covariance matrix, or alternatively the underlying graph structure of a Gaussian Markov Random Field, from very limited samples. We propose a novel algorithm for solving the resulting optimization problem which is a regularized log-

determinant program. In contrast to other state-of-the-art methods that largely use first order gradient information, our algorithm is based on Newton's method and employs a quadratic approximation, but with some modifications that leverage the structure of the sparse Gaussian MLE problem. We show that our method is superlinearly convergent, and also present experimental results using synthetic and real application data that demonstrate the considerable improvements in performance of our method when compared to other state-of-the-art methods.
Subject Area: Optimization

T63 A Convergence Analysis of Log-Linear Training

S. Wiesler wiesler@cs.rwth-aachen.de
H. Ney ney@cs.rwth-aachen.de
RWTH Aachen

Log-linear models are widely used probability models for statistical pattern recognition. Typically, log-linear models are trained according to a convex criterion. In recent years, the interest in log-linear models has greatly increased. The optimization of log-linear model parameters is costly and therefore an important topic, in particular for large-scale applications. Different optimization algorithms have been evaluated empirically in many papers. In this work, we analyze the optimization problem analytically and show that the training of log-linear models can be highly ill-conditioned. We verify our findings on two handwriting tasks. By making use of our convergence analysis, we obtain good results on a large-scale continuous handwriting recognition task with a simple and generic approach.
Subject Area: Optimization

T64 Non-Asymptotic Analysis of Stochastic Approximation Algorithms for Machine Learning

F. Bach francis.bach@ens.fr
INRIA - Ecole Normale Supérieure
E. Moulines eric.moulines@telecom-paristech.fr
Telecom Paristech

We consider the minimization of a convex objective function defined on a Hilbert space, which is only available through unbiased estimates of its gradients. This problem includes standard machine learning algorithms such as kernel logistic regression and least-squares regression, and is commonly referred to as a stochastic approximation problem in the operations research community. We provide a non-asymptotic analysis of the convergence of two well-known algorithms, stochastic gradient descent (a.k.a. ~Robbins-Monro algorithm) as well as a simple modification where iterates are averaged (a.k.a. ~Polyak-Ruppert averaging). Our analysis suggests that a learning rate proportional to the inverse of the number of iterations, while leading to the optimal convergence rate in the strongly convex case, is not robust to the lack of strong convexity or the setting of the proportionality constant. This situation is remedied when using slower decays together with averaging, robustly leading to the optimal rate of convergence. We illustrate our theoretical results with simulations on synthetic and standard datasets.
Subject Area: Optimization

T65 Better Mini-Batch Algorithms via Accelerated Gradient Methods

A. Cotter cotter@ttic.edu
N. Srebro nati@ttic.edu
K. Sridharan karthik@ttic.edu
Toyota Technological Institute at Chicago
O. Shamir ohadsh@microsoft.com
Microsoft Research

Mini-batch algorithms have recently received significant attention as a way to speed-up stochastic convex optimization problems. In this paper, we study how such algorithms can be improved using accelerated gradient methods. We provide a novel analysis, which shows how standard gradient methods may sometimes be insufficient to obtain a significant speed-up. We propose a novel accelerated gradient algorithm, which deals with this deficiency, and enjoys a uniformly superior guarantee. We conclude our paper with experiments on real-world datasets, which validates our algorithm and substantiates our theoretical insights.
Subject Area: Optimization\Stochastic Methods

T66 PAC-Bayesian Analysis of Contextual Bandits

Y. Seldin seldin@tuebingen.mpg.de
Max Planck Institute for Intelligent Systems
P. Auer auer@unileoben.ac.at
R. Ortner ronald.ortner@unileoben.ac.at
University of Leoben
F. Lavolette francois.lavolette@ift.ulaval.ca
Université Laval
J. Shawe-Taylor jst@cs.ucl.ac.uk
University College London

We derive an instantaneous (per-round) data-dependent regret bound for stochastic multiarmed bandits with side information (also known as contextual bandits). The scaling of our regret bound with the number of states (contexts) N goes as $\sqrt{N I_{pt}(S;A)}$, where $I_{pt}(S;A)$ is the mutual information between states and actions (the side information) used by the algorithm at round t . If the algorithm uses all the side information, the regret bound scales as $\sqrt{N \ln K}$, where K is the number of actions (arms). However, if the side information $I_{pt}(S;A)$ is not fully used, the regret bound is significantly tighter. In the extreme case, when $I_{pt}(S;A) = 0$, the dependence on the number of states reduces from linear to logarithmic. Our analysis allows to provide the algorithm large amount of side information, let the algorithm to decide which side information is relevant for the task, and penalize the algorithm only for the side information that it is using de facto. We also present an algorithm for multiarmed bandits with side information with computational complexity that is a linear in the number of actions.
Subject Area: Learning Theory

T67 Spectral Methods for Learning Multivariate Latent Tree Structure

a. anandkumar a.anandkumar@uci.edu
 U.C.Irvine
 K. Chaudhuri kamalika@cs.ucsd.edu
 UC San Diego
 D. Hsu danielhsu@gmail.com
 S. Kakade sham@tti-c.org
 Microsoft Research
 L. Song lesong@cs.cmu.edu
 Carnegie Mellon University
 T. Zhang tzhang@stat.rutgers.edu
 Rutgers University

This work considers the problem of learning the structure of multivariate linear tree models, which include a variety of directed tree graphical models with continuous, discrete, and mixed latent variables such as linear-Gaussian models, hidden Markov models, Gaussian mixture models, and Markov evolutionary trees. The setting is one where we only have samples from certain observed variables in the tree, and our goal is to estimate the tree structure (i.e., the graph of how the underlying hidden variables are connected to each other and to the observed variables). We propose the Spectral Recursive Grouping algorithm, an efficient and simple bottom-up procedure for recovering the tree structure from independent samples of the observed variables. Our finite sample size bounds for exact recovery of the tree structure reveal certain natural dependencies on underlying statistical and structural properties of the underlying joint distribution. Furthermore, our sample complexity guarantees have no explicit dependence on the dimensionality of the observed variables, making the algorithm applicable to many high-dimensional settings. At the heart of our algorithm is a spectral quartet test for determining the relative topology of a quartet of variables from second-order statistics.
 Subject Area: Learning Theory

T68 Algorithms and hardness results for parallel large margin learning

R. Servedio rocco@cs.columbia.edu
 Columbia University
 P. Long plong@google.com
 Google

We study the fundamental problem of learning an unknown large-margin halfspace in the context of parallel computation. Our main positive result is a parallel algorithm for learning a large-margin halfspace that is based on interior point methods from convex optimization and fast parallel algorithms for matrix computations. We show that this algorithm learns an unknown γ -margin halfspace over n dimensions using $\text{poly}(n, 1/\gamma)$ processors and runs in time $\tilde{O}(1/\gamma) + O(\log n)$. In contrast, naive parallel algorithms that learn a γ -margin halfspace in time that depends polylogarithmically on n have $\Omega(1/\gamma^2)$ runtime dependence on γ . Our main negative result deals with boosting, which is a standard approach to learning large-margin halfspaces. We give an information-theoretic proof that in the original PAC framework, in which a weak learning algorithm is provided as an oracle that is called by the booster, boosting cannot be parallelized: the ability

to call the weak learner multiple times in parallel within a single boosting stage does not reduce the overall number of successive stages of boosting that are required.
 Subject Area: Learning Theory

T69 Composite Multiclass Losses

E. Vernet evernet@ens-cachan.fr
 ENS Cachan
 R. Williamson Bob.Williamson@anu.edu.au
 M. Reid mark.reid@anu.edu.au
 The Australian National University

We consider loss functions for multiclass prediction problems. We show when a multiclass loss can be expressed as a “proper composite loss”, which is the composition of a proper loss and a link function. We extend existing results for binary losses to multiclass losses. We determine the stationarity condition, Bregman representation, order-sensitivity, existence and uniqueness of the composite representation for multiclass losses. We also show that the integral representation for binary proper losses can not be extended to multiclass losses. We subsume existing results on “classification calibration” by relating it to properness. We draw conclusions concerning the design of multiclass losses.
 Subject Area: Learning Theory

T70 Autonomous Learning of Action Models for Planning

N. Mehta mehtane@eecs.oregonstate.edu
 P. Tadepalli tadepall@eecs.oregonstate.edu
 A. Fern afern@eecs.oregonstate.edu

This paper introduces two new frameworks for learning action models for planning. In the mistake-bounded planning framework, the learner has access to a planner for the given model representation, a simulator, and a planning problem generator, and aims to learn a model with at most a polynomial number of faulty plans. In the planned exploration framework, the learner does not have access to a problem generator and must instead design its own problems, plan for them, and converge with at most a polynomial number of planning attempts. The paper reduces learning in these frameworks to concept learning with one-sided error and provides algorithms for successful learning in both frameworks. A specific family of hypothesis spaces is shown to be efficiently learnable in both the frameworks.
 Subject Area: Learning Theory

T71 Universal low-rank matrix recovery from Pauli measurements

Y. Liu yikailiu00@gmail.com
 National Institute of Standards and Technology

We study the problem of reconstructing an unknown matrix M of rank r and dimension d using $O(rd \text{ poly } \log d)$ Pauli measurements. This has applications in quantum state tomography, and is a non-commutative analogue of a well-known problem in compressed sensing: recovering a sparse vector from a few of its Fourier coefficients. We show that almost all sets of $O(rd \log^6 d)$ Pauli measurements satisfy the rank- r restricted isometry

property (RIP). This implies that M can be recovered from a fixed (“universal”) set of Pauli measurements, using nuclear-norm minimization (e.g., the matrix Lasso), with nearly-optimal bounds on the error. A similar result holds for any class of measurements that use an orthonormal operator basis whose elements have small operator norm. Our proof uses Dudley’s inequality for Gaussian processes, together with bounds on covering numbers obtained via entropy duality.”

Subject Area: Theory

T72 A More Powerful Two-Sample Test in High Dimensions using Random Projection

M. Lopes mlopes@stat.berkeley.edu
 L. Jacob laurent@stat.berkeley.edu
 M. Wainwright wainwrig@eecs.berkeley.edu
 UC Berkeley

We consider the hypothesis testing problem of detecting a shift between the means of two multivariate normal distributions in the high-dimensional setting, allowing for the data dimension p to exceed the sample size n . Our contribution is a new test statistic for the two-sample test of means that integrates a random projection with the classical Hotelling T^2 statistic. Working within a high-dimensional framework that allows $(p, n) \rightarrow \infty$, we first derive an asymptotic power function for our test, and then provide sufficient conditions for it to achieve greater power than other state-of-the-art tests. Using ROC curves generated from simulated data, we demonstrate superior performance against competing tests in the parameter regimes anticipated by our theoretical results. Lastly, we illustrate an advantage of our procedure with comparisons on a high-dimensional gene expression dataset involving the discrimination of different types of cancer.

Subject Area: Theory

T73 Prediction strategies without loss

M. Kapralov kapralov@stanford.edu
 Stanford University
 R. Panigrahy rina@microsoft.com
 Microsoft Research

Consider a sequence of bits where we are trying to predict the next bit from the previous bits. Assume we are allowed to say ‘predict 0’ or ‘predict 1’, and our payoff is +1 if the prediction is correct and -1 otherwise. We will say that at each point in time the loss of an algorithm is the number of wrong predictions minus the number of right predictions so far. In this paper we are interested in algorithms that have essentially zero (expected) loss over any string at any point in time and yet have small regret with respect to always predicting 0 or always predicting 1. For a sequence of length T our algorithm has regret $14\epsilon T$ and loss $2\sqrt{T}\epsilon^{-c}$ in expectation for all strings. We show that the tradeoff between loss and regret is optimal up to constant factors. Our techniques extend to the general setting of N experts, where the related problem of trading off regret to the best expert for regret to the ‘special’ expert has been studied by Even-Dar et al. (COLT’07). We obtain essentially zero loss with respect to the special expert and optimal loss/regret tradeoff, improving upon the results of Even-Dar et al

(COLT’07) and settling the main question left open in their paper. The strong loss bounds of the algorithm have some surprising consequences. First, we obtain a parameter free algorithm for the experts problem that has optimal regret bounds with respect to k -shifting optima, i.e. bounds with respect to the optimum that is allowed to change arms multiple times. Moreover, for any window of size N the regret of our algorithm to any expert never exceeds $O(\sqrt{N}(\log N + \log T))$, where N is the number of experts and T is the time horizon, while maintaining the essentially zero loss property.

Subject Area: Theory

T74 Learning in Hilbert vs. Banach Spaces: A Measure Embedding Viewpoint

B. Sriperumbudur bharathsv.ucsd@gmail.com
 G. Lanckriet gert@ece.ucsd.edu
 U.C. San Diego
 K. Fukumizu fukumizu@ism.ac.jp
 Institute of Statistical Mathematics

The goal of this paper is to investigate the advantages and disadvantages of learning in Banach spaces over Hilbert spaces. While many works have been carried out in generalizing Hilbert methods to Banach spaces, in this paper, we consider the simple problem of learning a Parzen window classifier in a reproducing kernel Banach space (RKBS)—which is closely related to the notion of embedding probability measures into an RKBS—in order to carefully understand its pros and cons over the Hilbert space classifier. We show that while this generalization yields richer distance measures on probabilities compared to its Hilbert space counterpart, it however suffers from serious computational drawback limiting its practical applicability, which therefore demonstrates the need for developing efficient learning algorithms in Banach spaces.

Subject Area: Theory

T75 On Strategy Stitching in Large Extensive Form Multiplayer Games

R. Gibson rggibson@cs.ualberta.ca
 D. Szafron dszafron@ualberta.ca
 University of Alberta

Computing a good strategy in a large extensive form game often demands an extraordinary amount of computer memory, necessitating the use of abstraction to reduce the game size. Typically, strategies from abstract games perform better in the real game as the granularity of abstraction is increased. This paper investigates two techniques for stitching a base strategy in a coarse abstraction of the full game tree, to expert strategies in fine abstractions of smaller subtrees. We provide a general framework for creating static experts, an approach that generalizes some previous strategy stitching efforts. In addition, we show that static experts can create strong agents for both 2-player and 3-player Leduc and Limit Texas Hold’em poker, and that a specific class of static experts can be preferred among a number of alternatives. Furthermore, we describe a poker agent that used static experts and won the 3-player events of the 2010 Annual Computer Poker Competition.

Subject Area: Theory

T76 Multi-Bandit Best Arm Identification

V. Gabillon victor.gabillon@inria.fr
 M. Ghavamzadeh mohammad.ghavamzadeh@inria.fr
 A. Lazaric alessandro.lazaric@inria.fr
 INRIA Lille-Nord Europe
 S. Bubeck sbubeck@princeton.edu
 Princeton University

We study the problem of identifying the best arm in each of the bandits in a multi-bandit multi-armed setting. We first propose an algorithm called Gap-based Exploration (GapE) that focuses on the arms whose mean is close to the mean of the best arm in the same bandit (i.e., small gap). We then introduce an algorithm, called GapE-V, which takes into account the variance of the arms in addition to their gap. We prove an upper-bound on the probability of error for both algorithms. Since GapE and GapE-V need to tune an exploration parameter that depends on the complexity of the problem, which is often unknown in advance, we also introduce variations of these algorithms that estimate this complexity online. Finally, we evaluate the performance of these algorithms and compare them to other allocation strategies on a number of synthetic problems.

Subject Area: Theory\Online Learning

T77 Linear Submodular Bandits and their Application to Diversified Retrieval

Y. Yue yisongyue@cmu.edu
 C. Guestrin guestrin@cs.cmu.edu
 Carnegie Mellon University

Diversified retrieval and online learning are two core research areas in the design of modern information retrieval systems. In this paper, we propose the linear submodular bandits problem, which is an online learning setting for optimizing a general class of feature-rich submodular utility models for diversified retrieval. We present an algorithm, called LSBGREEDY, and prove that it efficiently converges to a near-optimal model. As a case study, we applied our approach to the setting of personalized news recommendation, where the system must recommend small sets of news articles selected from tens of thousands of available articles each day. In a live user study, we found that LSBGREEDY significantly outperforms existing online learning approaches.

Subject Area: Theory\Online Learning

T78 Adaptive Hedge

T. van Erven tim@timvanerven.nl
 VU University
 P. Grunwald pdg@cwi.nl
 S. Rooij S.de.Rooij@cwi.nl
 CWI
 W. Koolen wouter@cs.rhul.ac.uk
 Royal Holloway, University of London

Most methods for decision-theoretic online learning are based on the Hedge algorithm, which takes a parameter called the learning rate. In most previous analyses the learning rate was carefully tuned to obtain optimal worst-

case performance, leading to suboptimal performance on easy instances, for example when there exists an action that is significantly better than all others. We propose a new way of setting the learning rate, which adapts to the difficulty of the learning problem: in the worst case our procedure still guarantees optimal performance, but on easy instances it achieves much smaller regret. In particular, our adaptive method achieves constant regret in a probabilistic setting, when there exists an action that on average obtains strictly smaller loss than all other actions. We also provide a simulation study comparing our approach to existing methods.

Subject Area: Theory

T79 On the Universality of Online Mirror Descent

N. Srebro nati@ttic.edu
 K. Sridharan karthik@ttic.edu
 Toyota Technological Institute at Chicago
 A. Tewari ambujtewari@gmail.com
 UT Austin

We show that for a general class of convex online learning problems, Mirror Descent can always achieve a (nearly) optimal regret guarantee.

Subject Area: Theory\Online Learning

T80 Online Submodular Set Cover, Ranking, and Repeated Active Learning

A. Guillory guillory@cs.washington.edu
 J. Bilmes bilmes@ee.washington.edu
 University of Washington

We propose an online prediction version of submodular set cover with connections to ranking and repeated active learning. In each round, the learning algorithm chooses a sequence of items. The algorithm then receives a monotone submodular function and suffers loss equal to the cover time of the function: the number of items needed, when items are selected in order of the chosen sequence, to achieve a coverage constraint. We develop an online learning algorithm whose loss converges to approximately that of the best sequence in hindsight. Our proposed algorithm is readily extended to a setting where multiple functions are revealed at each round and to bandit and contextual bandit settings.

Subject Area: Theory

T81 Finite Time Analysis of Stratified Sampling for Monte Carlo

A. Carpentier alexandra.carpentier@inria.fr
 R. Munos remi.munos@inria.fr
 INRIA Lille - Nord Europe

We consider the problem of stratified sampling for Monte-Carlo integration. We model this problem in a multi-armed bandit setting, where the arms represent the strata, and the goal is to estimate a weighted average of the mean values of the arms. We propose a strategy that samples the arms according to an upper bound on their standard deviations and compare its estimation quality to an ideal allocation that would know the standard deviations of

the arms. We provide two regret analyses: a distribution-dependent bound $\tilde{O}(n^{-3/2})$ that depends on a measure of the disparity of the arms, and a distribution-free bound $\tilde{O}(n^{-4/3})$ that does not. To the best of our knowledge, such a finite-time analysis is new for this problem.

Subject Area: Theory

T82 See the Tree Through the Lines: The Shazoo Algorithm

F. Vitale fabio.vitale@unimi.it
 N. Cesa-Bianchi nicolo.cesa-bianchi@unimi.it
 G. Zappella giovanni.zappella@unimi.it
 Universit  degli studi di Milano
 C. Gentile claudio.gentile@uninsubria.it
 Universita' dell'Insubria

Predicting the nodes of a given graph is a fascinating theoretical problem with applications in several domains. Since graph sparsification via spanning trees retains enough information while making the task much easier, trees are an important special case of this problem. Although it is known how to predict the nodes of an unweighted tree in a nearly optimal way, in the weighted case a fully satisfactory algorithm is not available yet. We fill this hole and introduce an efficient node predictor, Shazoo, which is nearly optimal on any weighted tree. Moreover, we show that Shazoo can be viewed as a common nontrivial generalization of both previous approaches for unweighted trees and weighted lines. Experiments on real-world datasets confirm that Shazoo performs well in that it fully exploits the structure of the input tree, and gets very close to (and sometimes better than) less scalable energy minimization methods.

Subject Area: Theory\Online Learning

T83 Generalizing from Several Related Classification Tasks to a New Unlabeled Sample

G. Blanchard gilles.blanchard@math.uni-potsdam.de
 Universit  Potsdam (DE)
 G. Lee gyemin@umich.edu
 C. Scott cscott@eecs.umich.edu
 University of Michigan

We consider the problem of assigning class labels to an unlabeled test data set, given several labeled training data sets drawn from similar distributions. This problem arises in several applications where data distributions fluctuate because of biological, technical, or other sources of variation. We develop a distribution-free, kernel-based approach to the problem. This approach involves identifying an appropriate reproducing kernel Hilbert space and optimizing a regularized empirical risk over the space. We present generalization error analysis, describe universal kernels, and establish universal consistency of the proposed methodology. Experimental results on flow cytometry data are presented.

Subject Area: Theory

T84 On U-processes and clustering performance

S. Cl men on stephan.clemencon@telecom-paristech.fr
 Telecom ParisTech

Many clustering techniques aim at optimizing empirical criteria that are of the form of a U -statistic of degree two. Given a measure of dissimilarity between pairs of observations, the goal is to minimize the within cluster point scatter over a class of partitions of the feature space. It is the purpose of this paper to define a general statistical framework, relying on the theory of U -processes, for studying the performance of such clustering methods. In this setup, under adequate assumptions on the complexity of the subsets forming the partition candidates, the excess of clustering risk is proved to be of the order $O_{\mathbb{P}}(1/\sqrt{n})$. Based on recent results related to the tail behavior of degenerate U -processes, it is also shown how to establish tighter rate bounds. Model selection issues, related to the number of clusters forming the data partition in particular, are also considered.

Subject Area: Theory\Statistical Learning Theory

T85 Information Rates and Optimal Decoding in Large Neural Populations

K. Rahnama Rad kamiar@stat.columbia.edu
 L. Paninski liam@stat.columbia.edu

Many fundamental questions in theoretical neuroscience involve optimal decoding and the computation of Shannon information rates in populations of spiking neurons. In this paper, we apply methods from the asymptotic theory of statistical inference to obtain a clearer analytical understanding of these quantities. We find that for large neural populations carrying a finite total amount of information, the full spiking population response is asymptotically as informative as a single observation from a Gaussian process whose mean and covariance can be characterized explicitly in terms of network and single neuron properties. The Gaussian form of this asymptotic sufficient statistic allows us in certain cases to perform optimal Bayesian decoding by simple linear transformations, and to obtain closed-form expressions of the Shannon information carried by the network. One technical advantage of the theory is that it may be applied easily even to non-Poisson point process network models; for example, we find that under some conditions, neural populations with strong history-dependent (non-Poisson) effects carry exactly the same information as do simpler equivalent populations of non-interacting Poisson neurons with matched firing rates. We argue that our findings help to clarify some results from the recent literature on neural decoding and neuroprosthetic design.

Subject Area: Probabilistic Models and Methods

T86 EigenNet: A Bayesian hybrid of generative and conditional models for sparse learning

Y. Qi alanqi@cs.purdue.edu
 F. Yan fengyan0@gmail.com
 Purdue University

For many real-world applications, we often need to select correlated variables—such as genetic variations and imaging features associated with Alzheimer’s disease—in a high dimensional space. The correlation between variables presents a challenge to classical variable selection methods. To address this challenge, the elastic net has been developed and successfully applied to many applications. Despite its great success, the elastic net does not exploit the correlation information embedded in the data to select correlated variables. To overcome this limitation, we present a novel hybrid model, EigenNet, that uses the eigenstructures of data to guide variable selection. Specifically, it integrates a sparse conditional classification model with a generative model capturing variable correlations in a principled Bayesian framework. We develop an efficient active-set algorithm to estimate the model via evidence maximization. Experiments on synthetic data and imaging genetics data demonstrated the superior predictive performance of the EigenNet over the lasso, the elastic net, and the automatic relevance determination.

Subject Area: Probabilistic Models and Methods

T87 Learning unbelievable probabilities

X. Pitkow xaq@neurotheory.columbia.edu
 University of Rochester
 Y. Ahmadian ya2005@columbia.edu
 K. Miller ken@neurotheory.columbia.edu
 Columbia University

Loopy belief propagation performs approximate inference on graphical models with loops. One might hope to compensate for the approximation by adjusting model parameters. Learning algorithms for this purpose have been explored previously, and the claim has been made that every set of locally consistent marginals can arise from belief propagation run on a graphical model. On the contrary, here we show that many probability distributions have marginals that cannot be reached by belief propagation using any set of model parameters or any learning algorithm. We call such marginals ‘unbelievable.’ This problem occurs whenever the Hessian of the Bethe free energy is not positive-definite at the target marginals. All learning algorithms for belief propagation necessarily fail in these cases, producing beliefs or sets of beliefs that may even be worse than the pre-learning approximation. We then show that averaging inaccurate beliefs, each obtained from belief propagation using model parameters perturbed about some learned mean values, can achieve the unbelievable marginals.

Subject Area: Probabilistic Models and Methods

T88 A blind sparse deconvolution method for neural spike identification

C. Ekanadham chaitu@math.nyu.edu
 D. Tranchina tranchin@courant.nyu.edu
 E. Simoncelli eero.simoncelli@nyu.edu
 New York University

We consider the problem of estimating neural spikes from extracellular voltage recordings. Most current methods are based on clustering, which requires substantial human supervision and produces systematic errors by failing to properly handle temporally overlapping spikes. We formulate the problem as one of statistical inference, in which the recorded voltage is a noisy sum of the spike trains of each neuron convolved with its associated spike waveform. Joint maximum-a-posteriori (MAP) estimation of the waveforms and spikes is then a blind deconvolution problem in which the coefficients are sparse. We develop a block-coordinate descent method for approximating the MAP solution. We validate our method on data simulated according to the generative model, as well as on real data for which ground truth is available via simultaneous intracellular recordings. In both cases, our method substantially reduces the number of missed spikes and false positives when compared to a standard clustering algorithm, primarily by recovering temporally overlapping spikes. The method offers a fully automated alternative to clustering methods that is less susceptible to systematic errors.

Subject Area: Probabilistic Models and Methods

T89 Accelerated Adaptive Markov Chain for Partition Function Computation

S. Ermon ermonste@cs.cornell.edu
 C. Gomes gomes@cs.cornell.edu
 B. Selman selman@cs.cornell.edu
 Cornell University
 A. Sabharwal ashish.sabharwal@us.ibm.com
 IBM Watson Research Center

We propose a novel Adaptive Markov Chain Monte Carlo algorithm to compute the partition function. In particular, we show how to accelerate a flat histogram sampling technique by significantly reducing the number of ‘‘null moves’’ in the chain, while maintaining asymptotic convergence properties. Our experiments show that our method converges quickly to highly accurate solutions on a range of benchmark instances, outperforming other state-of-the-art methods such as IJGP, TRW, and Gibbs sampling both in run-time and accuracy. We also show how obtaining a so-called density of states distribution allows for efficient weight learning in Markov Logic theories.

Subject Area: Probabilistic Models and Methods

T90 Message-Passing for Approximate MAP Inference with Latent Variables

J. Jiang jiarong@umiacs.umd.edu
 H. Daume III me@hal3.name
 University of Maryland
 P. Rai piyush@cs.utah.edu
 University of Utah

We consider a general inference setting for discrete probabilistic graphical models where we seek maximum a posteriori (MAP) estimates for a subset of the random variables (max nodes), marginalizing over the rest (sum nodes). We present a hybrid message-passing algorithm to accomplish this. The hybrid algorithm passes a mix of sum and max messages depending on the type of source node (sum or max). We derive our algorithm by showing that it falls out as the solution of a particular relaxation of a variational framework. We further show that the Expectation Maximization algorithm can be seen as an approximation to our algorithm. Experimental results on synthetic and real-world datasets, against several baselines, demonstrate the efficacy of our proposed algorithm.

Subject Area: Probabilistic Models and Methods

T91 Priors over Recurrent Continuous Time Processes

A. Saeedi ardavan.s@stat.ubc.ca
 A. Bouchard-Côté bouchard@stat.ubc.ca
 University of British Columbia

We introduce the Gamma-Exponential Process (GEP), a prior over a large family of continuous time processes. A hierarchical version of this prior (HGEP; the Hierarchical GEP) yields a useful model for analyzing complex time series. Models based on HGEPs display many attractive properties: conjugacy, exchangeability and closed-form predictive distribution for the waiting times, and exact Gibbs updates for the time scale parameters. After establishing these properties, we show how posterior inference can be carried efficiently using Particle MCMC methods. This yields a MCMC algorithm that can resample entire sequences atomically while avoiding the complications of introducing slice and stick auxiliary variables. We applied our model to the problem of estimating the disease progression in Multiple Sclerosis, and to RNA evolutionary modeling. In both domains, we found that our model outperformed the standard rate matrix estimation approach.

Subject Area: Probabilistic Models and Methods

T92 Modelling Genetic Variations using Fragmentation-Coagulation Processes

Y. Teh ywteh@gatsby.ucl.ac.uk
 C. Blundell c.blundell@gatsby.ucl.ac.uk
 Gatsby Computational Neuroscience Unit, UCL
 L. Elliott elliott@gatsby.ucl.ac.uk
 University College London

We propose a novel class of Bayesian nonparametric models for sequential data called fragmentation-coagulation processes (FCPs). FCPs model a set of sequences using a partition-valued Markov process

which evolves by splitting and merging clusters. An FCP is exchangeable, projective, stationary and reversible, and its equilibrium distributions are given by the Chinese restaurant process. As opposed to hidden Markov models, FCPs allow for flexible modelling of the number of clusters, and they avoid label switching non-identifiability problems. We develop an efficient Gibbs sampler for FCPs which uses uniformization and the forward-backward algorithm. Our development of FCPs is motivated by applications in population genetics, and we demonstrate the utility of FCPs on problems of genotype imputation with phased and unphased SNP data.

Subject Area: Probabilistic Models and Methods

T93 Variational Gaussian Process Dynamical Systems

A. Damianou andreas.damianou@sheffield.ac.uk
 N. Lawrence N.Lawrence@shef.ac.uk
 University of Sheffield
 M. Titsias mtitsias@cs.man.ac.uk
 University of Manchester

High dimensional time series are endemic in applications of machine learning such as robotics (sensor data), computational biology (gene expression data), vision (video sequences) and graphics (motion capture data). Practical nonlinear probabilistic approaches to this data are required. In this paper we introduce the variational Gaussian process dynamical system. Our work builds on recent variational approximations for Gaussian process latent variable models to allow for nonlinear dimensionality reduction simultaneously with learning a dynamical prior in the latent space. The approach also allows for the appropriate dimensionality of the latent space to be automatically determined. We demonstrate the model on a human motion capture data set and a series of high resolution video sequences.

Subject Area: Probabilistic Models and Methods

T94 The Doubly Correlated Nonparametric Topic Model

D. Kim daeil@cs.brown.edu
 E. Sudderth sudderth@cs.brown.edu
 Brown University

Topic models are learned via a statistical model of variation within document collections, but designed to extract meaningful semantic structure. Desirable traits include the ability to incorporate annotations or metadata associated with documents; the discovery of correlated patterns of topic usage; and the avoidance of parametric assumptions, such as manual specification of the number of topics. We propose a doubly correlated nonparametric topic (DCNT) model, the first model to simultaneously capture all three of these properties. The DCNT models metadata via a flexible, Gaussian regression on arbitrary input features; correlations via a scalable square-root covariance representation; and nonparametric selection from an unbounded series of potential topics via a stick-breaking construction. We validate the semantic structure and predictive performance of the DCNT using a corpus of NIPS documents annotated by various metadata.

Subject Area: Probabilistic Models and Methods

T95 The Kernel Beta Process

L. Ren lren@yahoo-inc.com
 Y. Wang yw65@duke.edu
 D. Dunson dunson@stat.duke.edu
 L. Carin lcarin@duke.edu
 Duke University

A new Levy process prior is proposed for an uncountable collection of covariate-dependent feature-learning measures; the model is called the kernel beta process (KBP). Available covariates are handled efficiently via the kernel construction, with covariates assumed observed with each data sample ("customer"), and latent covariates learned for each feature ("dish"). Each customer selects dishes from an infinite buffet, in a manner analogous to the beta process, with the added constraint that a customer first decides probabilistically whether to "consider" a dish, based on the distance in covariate space between the customer and dish. If a customer does consider a particular dish, that dish is then selected probabilistically as in the beta process. The beta process is recovered as a limiting case of the KBP. An efficient Gibbs sampler is developed for computations, and state-of-the-art results are presented for image processing and music analysis tasks.

Subject Area: Probabilistic Models and Methods

T96 An Exact Algorithm for F-Measure Maximization

K. Dembczynski dembczynski@informatik.uni-marburg.de
 Poznan University of Technology
 W. Waegeman willem.waegeman@ugent.be
 Ghent University
 W. Cheng cheng@mathematik.uni-marburg.de
 E. Hullermeier eyke@informatik.uni-marburg.de
 Philipps-University at Marburg

The F-measure, originally introduced in information retrieval, is nowadays routinely used as a performance metric for problems such as binary classification, multi-label classification, and structured output prediction. Optimizing this measure remains a statistically and computationally challenging problem, since no closed-form maximizer exists. Current algorithms are approximate and typically rely on additional assumptions regarding the statistical distribution of the binary response variables. In this paper, we present an algorithm which is not only computationally efficient but also exact, regardless of the underlying distribution. The algorithm requires only a quadratic number of parameters of the joint distribution (with respect to the number of binary responses). We illustrate its practical performance by means of experimental results for multi-label classification.

Subject Area: Probabilistic Models and Methods

T97 Contextual Gaussian Process Bandit Optimization

A. Krause krausea@ethz.ch
 C. Ong chengsoon.ong@inf.ethz.ch
 ETH Zurich

How should we design experiments to maximize performance of a complex system, taking into account uncontrollable environmental conditions? How should we select relevant documents (ads) to display, given information about the user? These tasks can be formalized as contextual bandit problems, where at each round, we receive context (about the experimental conditions, the query), and have to choose an action (parameters, documents). The key challenge is to trade off exploration by gathering data for estimating the mean payoff function over the context-action space, and to exploit by choosing an action deemed optimal based on the gathered data. We model the payoff function as a sample from a Gaussian process defined over the joint context-action space, and develop CGP-UCB, an intuitive upper-confidence style algorithm. We show that by mixing and matching kernels for contexts and actions, CGP-UCB can handle a variety of practical applications. We further provide generic tools for deriving regret bounds when using such composite kernel functions. Lastly, we evaluate our algorithm on two case studies, in the context of automated vaccine design and sensor management. We show that context-sensitive optimization outperforms no or naive use of context.

Subject Area: Probabilistic Models and Methods

T98 Automated Refinement of Bayes Networks' Parameters based on Test Ordering Constraints

O. Khan ozkhan@cs.uwaterloo.ca
 P. Poupart ppoupart@cs.uwaterloo.ca
 University of Waterloo
 J. Agosta johnmark.agosta@gmail.com
 Intel Corporation

In this paper, we derive a method to refine a Bayes network diagnostic model by exploiting constraints implied by expert decisions on test ordering. At each step, the expert executes an evidence gathering test, which suggests the test's relative diagnostic value. We demonstrate that consistency with an expert's test selection leads to non-convex constraints on the model parameters. We incorporate these constraints by augmenting the network with nodes that represent the constraint likelihoods. Gibbs sampling, stochastic hill climbing and greedy search algorithms are proposed to find a MAP estimate that takes into account test ordering constraints and any data available. We demonstrate our approach on diagnostic sessions from a manufacturing scenario.

Subject Area: Probabilistic Models and Methods

T99 Solving Decision Problems with Limited Information

D. Maua denis@idsia.ch
 C. de Campos cassio@idsia.ch
 Dalle Molle Institute for Artificial Intelligence

We present a new algorithm for exactly solving decision-making problems represented as an influence diagram. We do not require the usual assumptions of no forgetting and regularity, which allows us to solve problems with limited information. The algorithm, which implements a sophisticated variable elimination procedure, is empirically shown to outperform a state-of-the-art algorithm in randomly generated problems of up to 150 variables and 10^{64} strategies.

Subject Area: Probabilistic Models and Methods

T100 Learning Higher-Order Graph Structure with Features by Structure Penalty

S. Ding sding@stat.wisc.edu
 G. Wahba wahba@stat.wisc.edu
 X. Zhu jerryzhu@cs.wisc.edu
 U. Wisconsin-Madison

In discrete undirected graphical models, the conditional independence of node labels Y is specified by the graph structure. We study the case where there is another input random vector X (e.g. observed features) such that the distribution $P(Y | X)$ is determined by functions of X that characterize the (higher-order) interactions among the Y 's. The main contribution of this paper is to learn the graph structure and the functions conditioned on X at the same time. We prove that discrete undirected graphical models with feature X are equivalent to multivariate discrete models. The reparameterization of the potential functions in graphical models by conditional log odds ratios of the latter offers advantages in representation of the conditional independence structure. The functional spaces can be flexibly determined by kernels. Additionally, we impose a Structure Lasso (SLasso) penalty on groups of functions to learn the graph structure. These groups with overlaps are designed to enforce hierarchical function selection. In this way, we are able to shrink higher order interactions to obtain a sparse graph structure.

Subject Area: Probabilistic Models and Methods

T101 On the Completeness of First-Order Knowledge Compilation for Lifted Probabilistic Inference

G. Van den Broeck guy.vandenbroeck@cs.kuleuven.be
 Katholieke Universiteit Leuven

Probabilistic logics are receiving a lot of attention today because of their expressive power for knowledge representation and learning. However, this expressivity is detrimental to the tractability of inference, when done at the propositional level. To solve this problem, various lifted inference algorithms have been proposed that reason at the first-order level, about groups of objects as a whole. Despite the existence of various lifted inference approaches, there are currently no completeness results

about these algorithms. The key contribution of this paper is that we introduce a formal definition of lifted inference that allows us to reason about the completeness of lifted inference algorithms relative to a particular class of probabilistic models. We then show how to obtain a completeness result using a first-order knowledge compilation approach for theories of formulae containing up to two logical variables.

Subject Area: Probabilistic Models and Methods

T102 Inference in continuous-time change-point models

F. Stimberg flostim@cs.tu-berlin.de
 M. Opper mo@ecs.soton.ac.uk
 A. Ruttor andreas.ruttor@tu-berlin.de
 Technische Universitaet Berlin
 G. Sanguinetti gsanguin@inf.ed.ac.uk
 University of Edinburgh

We consider the problem of Bayesian inference for continuous time multi-stable stochastic systems which can change both their diffusion and drift parameters at discrete times. We propose exact inference and sampling methodologies for two specific cases where the discontinuous dynamics is given by a Poisson process and a two-state Markovian switch. We test the methodology on simulated data, and apply it to two real data sets in finance and systems biology. Our experimental results show that the approach leads to valid inferences and non-trivial insights.

Subject Area: Probabilistic Models and Methods

T103 Comparative Analysis of Viterbi Training and Maximum Likelihood Estimation for HMMs

A. Allahverdyan armen.allahverdyan@gmail.com
 Yerevan Physics Institute
 A. Galstyan galstyan@isi.edu
 University of Southern California

We present an asymptotic analysis of Viterbi Training (VT) and contrast it with a more conventional Maximum Likelihood (ML) approach to parameter estimation in Hidden Markov Models. While ML estimator works by (locally) maximizing the likelihood of the observed data, VT seeks to maximize the probability of the most likely hidden state sequence. We develop an analytical framework based on a generating function formalism and illustrate it on an exactly solvable model of HMM with one unambiguous symbol. For this particular model the ML objective function is continuously degenerate. VT objective, in contrast, is shown to have only finite degeneracy. Furthermore, VT converges faster and results in sparser (simpler) models, thus realizing an automatic Occam's razor for HMM learning. For more general scenario VT can be worse compared to ML but still capable of correctly recovering most of the parameters.

Subject Area: Probabilistic Models and Methods

5:45 – 11:59 PM

1A Reproducing biologically realistic firing patterns on a highly-accelerated neuromorphic hardware system

Marc-Olivier Schwartz
Kirchhoff-Institut für Physik

This demonstration will feature a neuromorphic chip in operation. The chip that will be showed during the demonstration features 512 neuron circuits and 115.000 plastic synapses. The users will be able to see live membrane voltage recording from neurons on the neuromorphic chip, which will reproduce typical firing patterns as seen in biology. Users will be able to modify the parameters of the neuron circuits with a Graphical User Interface (GUI) and observe the result on a screen.

2A A smartphone 3D functional brain scanner

Carsten Stahlhut
Arkadiusz Stopczynski
Jakob Eg Larsen
Michael Petersen
Lars Hansen
Technical University of Denmark

We demonstrate a fully portable 3D real-time functional brain scanner consisting of a wireless 14-channel 'Neuroheadset' (Emotiv EPOC) and a Nokia N900 smartphone. The novelty of our system is the ability to perform real-time functional brain imaging on a smartphone device, including stimulus delivery, data acquisition, logging, brain state decoding, and 3D visualization of the cortical EEG sources. Custom-made software realized in Qt has been implemented on the phone, which allow for either the phone to process the EEG data locally or transmit it to a server when more advanced machine learning tools are preferred. Source localization is implemented locally on the phone with a 3D brain model consisting of 1,028 vertices and 2,048 triangles stored in the mobile application. Our system design benefits from the possibility of being able to integrate with multiple hardware platforms (smartphones, tablet computers, and netbooks) that are based on Linux operating systems.

3A SENNA Natural Language Processing Demo

Ronan Collobert
Idiap Research Institute

We present an interactive demo of SENNA, our end-to-end Natural Language Processing (NLP) system. The demo leverages Torch7, which provides an efficient Matlab-like environment for machine-learning algorithms. As the user types in characters, the demo displays (in a smooth and real time fashion) the outcome of a wide variety of NLP tasks (Part-Of-Speech tags, Chunking, Name Entity Recognition, Semantic Role Labeling, and Parsing). Torch7 GUI features are used to embellish the output results, as well as the interactive experience.

4A Haptic Belt with Pedestrian Detection

Jean Feng
Marc Rasi
Andrew Ng
Quoc V. Le
Morgan Quigley
Justin Chen
Tiffany Low
Will Zou
Stanford University

We built a haptic belt that sends vibrating signals to the user that informs the user of the positions of surrounding pedestrians. When the camera detects a person, a motor in that direction will vibrate. By wearing the belt, the user can develop a "sixth-sense" to identify where people are standing, even if a person stands behind the user. Our belt improves upon earlier work by transmitting high level processed data to users.

To achieve this goal, we improved and integrated both software and hardware components. The current hardware system consists of eight cameras arranged in a circle to attain a 360-degree view, a belt with twelve corresponding vibration motors, and a computer. The cameras are mounted at the top of the backpack while the computer is attached to the center of the backpack. The software component is built from a state-of-the-art pedestrian detection algorithm (from NYU [1,2]) and further improved by adding parallelism with multithreaded CPU and GPU. The whole system is optimized to operate at 1.5 frames per second with high accuracy, which is six times faster than the original algorithm.

As part of the demo, the user will wear the belt and the mobile computing backpack. They will feel the belt vibrate in the direction where the algorithm detects a pedestrian.



WEDNESDAY CONFERENCE



ORAL SESSION

SESSION 6 - 9:30 – 10:40 AM

Session Chair: Raquel Urtasun

POSNER LECTURE: From Kernels to Causal Inference,

Bernhard Schölkopf, Max Planck Institute for Intelligent Systems

Kernel methods in machine learning have expanded from tricks to construct nonlinear algorithms to general tools to assay higher order statistics and properties of distributions. They find applications also in causal inference, an intriguing field that examines causal structures by testing their probabilistic footprints. However, the links between causal inference and modern machine learning go beyond this and the talk will outline some initial thoughts how problems like covariate shift adaptation and semi-supervised learning can benefit from the causal methodology.

Bernhard Scholkopf received degrees in mathematics (London) and physics (Tubingen), and a doctorate in computer science from the Technical University Berlin. He has researched at AT&T Bell Labs, at GMD FIRST, Berlin, at the Australian National University, Canberra, and at Microsoft Research Cambridge (UK). In 2001, he was appointed scientific member of the Max Planck Society and director at the MPI for Biological Cybernetics; in 2010 he founded the Max Planck Institute for Intelligent Systems. For further information, see www.kyb.tuebingen.mpg.de/~bs.

- **High-Dimensional Graphical Model Selection: Tractable Graph Families and Necessary Conditions,** Animashree Anandkumar, UC Irvine; Vincent Tan, University of Wisconsin-Madison; Alan Willsky, MIT

We consider the problem of Ising and Gaussian graphical model selection given n i.i.d. samples from the model. We propose an efficient threshold-based algorithm for structure estimation based known as conditional mutual information test. This simple local algorithm requires only low-order statistics of the data and decides whether two nodes are neighbors in the unknown graph. Under some transparent assumptions, we establish that the proposed algorithm is structurally consistent (or sparsistent) when the number of samples scales as $n = \Omega(J_{\min}^{-4} \log p)$, where p is the number of nodes and J_{\min} is the minimum edge potential. We also prove novel non-asymptotic necessary conditions for graphical model selection.

Subject Area: Probabilistic Models and Methods



SPOTLIGHT SESSION

SESSION 5 - 10:40 – 11:10 AM

Session Chair: Raquel Urtasun

- **Minimax Localization of Structural Information in Large Noisy Matrices**
Mladen Kolar, Sivaraman Balakrishnan, Alessandro Rinaldo, Aarti Singh, Carnegie Mellon University
Subject Area: Clustering
See abstract, page 93 (W55)
- **On Learning Discrete Graphical Models using Greedy Methods**
Ali Jalali, Christopher Johnson, Pradeep Ravikumar, University of Texas at Austin
Subject Area: Model Selection & Structure Learning
See abstract, page 102 (W100)
- **Learning to Learn with Compound HD Models**
Ruslan Salakhutdinov, University of Toronto; Josh Tenenbaum & Antonio Torralba, MIT
Subject Area: Probabilistic Models and Methods
See abstract, page 100 (W89)
- **Probabilistic Joint Image Segmentation and Labeling**
Adrian Ion, TU Wien & IST Austria; Joao Carreira & Cristian Sminchisescu, University of Bonn; University of Bonn
Subject Area: Vision
See abstract, page 83 (W16)
- **Object Detection with Grammar Models**
Ross Girshick, University of Chicago, Pedro Felzenszwalb, Brown University, and David Mcallester, TTI-Chicago.
Subject Area: Object Recognition
See abstract, page 85 (W24)
- **Rapid Deformable Object Detection using Dual-Tree Branch-and-Bound**
Iasonas Kokkinos, Ecole Centrale Paris / INRIA Saclay
Subject Area: Object Recognition
See abstract, page 85 (W23)
- **Im2Text: Describing Images Using 1 Million Captioned Photographs**
Vicente Ordonez, Girish Kulkarni and Tamara Berg, Stony Brook University
Subject Area: Object Recognition
See abstract, page 85 (W21)

ORAL SESSION

SESSION 7 - 11:10 – 11:30 AM

Session Chair: Katherine Heller

- **Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials**

Philipp Krähenbühl and Vladlen Koltun, Stanford University

Most state-of-the-art techniques for multi-class image segmentation and labeling use conditional random fields defined over pixels or image regions. While region-level models often feature dense pairwise connectivity, pixel-level models are considerably larger and have only permitted sparse graph structures. In this paper, we consider fully connected CRF models defined on the complete set of pixels in an image. The resulting graphs have billions of edges, making traditional inference algorithms impractical. Our main contribution is a highly efficient approximate inference algorithm for fully connected CRF models in which the pairwise edge potentials are defined by linear combinations of Gaussian kernels. Our algorithm can approximately minimize fully connected models on tens of thousands of variables in a fraction of a second. Quantitative and qualitative results on the MSRC-21 and PASCAL VOC 2010 datasets demonstrate that full pairwise connectivity at the pixel level produces significantly more accurate segmentations and pixel-level label assignments.

Subject Area: Vision

the calculation of the gradient of the smooth term or in the proximity operator with respect to the second term. We show that the basic proximal-gradient method, the basic proximal-gradient method with a strong convexity assumption, and the accelerated proximal-gradient method achieve the same convergence rates as in the error-free case, provided the errors decrease at an appropriate rate. Our experimental results on a structured sparsity problem indicate that sequences of errors with these appealing theoretical properties can lead to practical performance improvements.

Subject Area: Convex Optimization

SPOTLIGHT SESSION

SESSION 6 - 12:40 – 1:10 PM

Session Chair: Phil Long

- **Improved Algorithms for Linear Stochastic Bandits**

Y. Abbasi-yadkori and C. Szepesvari, University of Alberta; D. Pal, Google
Subject Area: Online Learning
See abstract, page 98 (W82)

- **From Bandits to Experts: On the Value of Side-Observations**

S. Mannor, O. Shamir, Microsoft Research
Subject Area: Online Learning
See abstract, page 98 (W81)

- **Lower Bounds for Passive and Active Learning**

M. Raginsky, University of Illinois at Urbana-Champaign; A. Rakhlin, University of Pennsylvania
Subject Area: Statistical Learning Theory
See abstract, page 99 (W85)

- **Active Classification based on Value of Classifier**

T. Gao, D. Koller, Stanford University
Subject Area: Classification
See abstract, page 88 (W33)

- **Budgeted Optimization with Concurrent Stochastic-Duration Experiments**

J. Azimi, A. Fern, X. Fern, Oregon State University
Subject Area: Active Learning
See abstract, page 92 (W52)

- **Projection onto A Nonnegative Max-Heap**

J. Liu, Siemens Corporate Research; L. Sun and J. Ye, Arizona State University
Subject Area: Learning with Structured Data
See abstract, page dd

- **Phase transition in the family of p -resistances**

M. Alamgir, U. von Luxburg, Max Planck Institute for Intelligent Systems
Subject Area: Statistical Learning Theory
See abstract, page 99 (W87)

ORAL SESSION

SESSION 8 - 12:00 – 12:40 PM

Session Chair: Phil Long

- **Efficient Online Learning via Randomized Rounding**

Nicolò Cesa-Bianchi, Università degli Studi di Milano; Ohad Shamir, Microsoft Research

Most online algorithms used in machine learning today are based on variants of mirror descent or follow-the-leader. In this paper, we present an online algorithm based on a completely different approach, which combines “random playout” and randomized rounding of loss subgradients. As an application of our approach, we provide the first computationally efficient online algorithm for collaborative filtering with trace-norm constrained matrices. As a second application, we solve an open question linking batch learning and transductive online learning.

Subject Area: Online Learning

- **Convergence Rates of Inexact Proximal-Gradient Methods for Convex Optimization**

Mark Schmidt, INRIA ; Nicolas Le Roux, INRIA ; Francis Bach, INRIA - Ecole Normale Supérieure

We consider the problem of optimizing the sum of a smooth convex function and a non-smooth convex function using proximal-gradient methods, where an error is present in

ORAL SESSION



SESSION 9 - 1:10 – 1:30 PM

Session Chair: Cédric Archambeau

- ***k*-NN Regression Adapts to Local Intrinsic Dimension**
Samory Kpotufe, Max Planck Institute

Many nonparametric regressors were recently shown to converge at rates that depend only on the intrinsic dimension of data. These regressors thus escape the curse of dimension when high-dimensional data has low intrinsic dimension (e.g. a manifold). We show that k -NN regression is also adaptive to intrinsic dimension. In particular our rates are local to a query x and depend only on the way masses of balls centered at x vary with radius. Furthermore, we show a simple way to choose $k = k(x)$ locally at any x so as to nearly achieve the minimax rate at x in terms of the unknown intrinsic dimension in the vicinity of x . We also establish that the minimax rate does not depend on a particular choice of metric space or distribution, but rather that this minimax rate holds for any metric space and doubling measure.

Subject Area: Learning Theory

SPOTLIGHT SESSION



SESSION 7 - 1:30 – 2:00 PM

Session Chair: Cédric Archambeau

- ***Complexity of Inference in Latent Dirichlet Allocation***
D. Sontag, New York Univ.; D. Roy, Univ. of Cambridge
Subject Area: Topic Models
See abstract, page 95 (W66)
- ***Practical Variational Inference for Neural Networks***
A. Graves, University of Toronto
Subject Area: Neural Networks
See abstract, page 90 (W42)
- ***A Multilinear Subspace Regression Method Using Orthogonal Tensors Decompositions***
Q. Zhao and A. Cichocki, RIKEN Brain Science Institute;
C. Caiafa, D. Mandic, L. Zhang, Shanghai Jiao Tong Univ.;
T. Ball and A. Schulze-bonhage, Albert-Ludwigs-Univ.
Subject Area: Regression
See abstract, page 90 (W43)
- ***Sparse Filtering***
J. Ngiam, P. Koh, Z. Chen, S. Bhaskar, A. Ng, Stanford
Subject Area: Unsupervised & Semi-supervised Learning
See abstract, page 91 (W50)
- ***Directed Graph Embedding: an Algorithm based on Continuous Limits of Laplacian-type Operators***
D. Perrault-Joncas, M. Meila, University of Washington
Subject Area: Spectral Methods
See abstract, page 94 (W63)

- ***Noise Thresholds for Spectral Clustering***
S. Balakrishnan, M. Xu, A. Krishnamurthy, A. Singh,
Carnegie Mellon University
Subject Area: Clustering
See abstract, page 93 (W56)

ORAL SESSION



SESSION 10 - 4:00 – 5:30 PM

Session Chair: Joelle Pineau

- ***INVITED TALK: Natural Algorithms***
Bernard Chazelle, Princeton University

I will discuss the merits of an algorithmic approach to the analysis of complex self-organizing systems. I will argue that computer science, and algorithms in particular, offer a fruitful perspective on the complex dynamics of multiagent systems: for example, opinion dynamics, bird flocking, and firefly synchronization. I will give many examples and try to touch on some of the theory behind them, with an emphasis on their algorithmic nature and the particular challenges to machine learning that an algorithmic approach to dynamical systems raises.

Bernard Chazelle is Eugene Higgins professor of computer science at Princeton University, where he has been on the faculty since 1986. He has held research and faculty positions at Carnegie-Mellon University, Brown University, Ecole Polytechnique, Ecole normale supérieure, the University of Paris, and INRIA. He did extensive consulting for Xerox PARC, DEC SRC, and NEC Research, where he was President of the Board of Fellows for several years. He received his Ph.D in computer science from Yale University in 1980. He is the author of the book "The Discrepancy Method." He is a fellow of the American Academy of Arts and Sciences, the European Academy of Sciences, and the World Innovation Foundation. He is an ACM Fellow and a former Guggenheim fellow.

- ***Iterative Learning for Reliable Crowdsourcing Systems***
David Karger, Sewoong Oh, Devavrat Shah, MIT

Crowdsourcing systems, in which tasks are electronically distributed to numerous "information piece-workers", have emerged as an effective paradigm for human-powered solving of large scale problems in domains such as image classification, data entry, optical character recognition, recommendation, and proofreading. Because these low-paid workers can be unreliable, nearly all crowdsourcers must devise schemes to increase confidence in their answers, typically by assigning each task multiple times and combining the answers in some way such as majority voting. In this paper, we consider a general model of such crowdsourcing tasks, and pose the problem of minimizing the total price (i.e., number of task assignments) that must be paid to achieve a target overall reliability. We give new algorithms for deciding which tasks to assign to which workers and for inferring correct answers from the workers' answers. We show that our algorithm significantly outperforms majority voting and, in fact, are asymptotically optimal through comparison to an oracle that knows the reliability of every worker.

Subject Area: Graphical Models

- **A Collaborative Mechanism for Crowdsourcing Prediction Problems**

Jacob Abernethy, Rafael Frongillo, UC Berkeley

Machine Learning competitions such as the Netflix Prize have proven reasonably successful as a method of “crowdsourcing” prediction tasks. But these competitions have a number of weaknesses, particularly in the incentive structure they create for the participants. We propose a new approach, called a Crowdsourced Learning Mechanism, in which participants collaboratively “learn” a hypothesis for a given prediction task. The approach draws heavily from the concept of a prediction market, where traders bet on the likelihood of a future event. In our framework, the mechanism continues to publish the current hypothesis, and participants can modify this hypothesis by wagering on an update. The critical incentive property is that a participant will profit an amount that scales according to how much her update improves performance on a released test set.

Subject Area: Game Theory & Computational Economics

- W12 **Variational Learning for Recurrent Spiking Networks**, D. Rezende, D. Wierstra, W. Gerstner
- W13 **Empirical models of spiking in neural populations**, J. Macke, L. Buesing, J. Cunningham, B. Yu, K. Shenoy, M. Sahani
- W14 **Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials**, P. Krähenbühl, V. Koltun
- W15 **Fast and Balanced: Efficient Label Tree Learning for Large Scale Object Recognition**, J. Deng, S. Satheesh, A. Berg, F. Li
- W16 **Probabilistic Joint Image Segmentation and Labeling**, A. Ion, J. Carreira, C. Sminchisescu
- W17 **Heavy-tailed Distances for Gradient Based Image Descriptors**, Y. Jia, T. Darrell
- W18 **θ -MRF: Capturing Spatial and Semantic Structure in the Parameters for Scene Understanding**, C. Li, A. Saxena, T. Chen
- W19 **Learning to Agglomerate Superpixel Hierarchies**, V. Jain, S. Turaga, K. Briggman, M. Helmstaedter, W. Denk, H. Seung
- W20 **Multiple Instance Filtering**, K. Wnuk, S. Soatto
- W21 **Automatic Captioning using Billions of Photographs**, V. Ordonez, G. Kulkarni, T. Berg
- W22 **Exploiting spatial overlap to efficiently compute appearance distances between image windows**, B. Alexe, V. Petrescu, V. Ferrari
- W23 **Rapid Deformable Object Detection using Dual-Tree Branch-and-Bound**, I. Kokkinos
- W24 **Object Detection with Grammar Models**, R. Girshick, P. Felzenszwalb, D. Mcallester
- W25 **Learning a Tree of Metrics with Disjoint Visual Features**, S. Hwang, K. Grauman, F. Sha
- W26 **Learning person-object interactions for action recognition in still images**, V. Delaitre, J. Sivic, I. Laptev
- W27 **Extracting Speaker-Specific Information with a Regularized Siamese Deep Network**, K. Chen, A. Salman
- W28 **A Machine Learning Approach to Predict Chemical Reactions**, M. Kayala, P. Baldi
- W29 **RTRMC: A Riemannian trust-region method for low-rank matrix completion**, N. Boumal, P. Absil
- W30 **Randomized Algorithms for Comparison-based Search**, D. Tschoop, S. Diggavi, P. Delgosha, S. Mohajer



POSTER SESSION

AND RECEPTION - 5:45 – 11:59 PM

- W1 **A rational model of causal inference with continuous causes**, M. Pacer, T. Griffiths
- W2 **Testing a Bayesian Measure of Representativeness Using a Large Image Database**, J. Abbott, K. Heller, Z. Ghahramani, T. Griffiths
- W3 **How Do Humans Teach: On Curriculum Learning and Teaching Dimension**, F. Khan, X. Zhu, B. Mutlu
- W4 **Probabilistic Modeling of Dependencies Among Visual Short-Term Memory Representations**, E. Orhan, R. Jacobs
- W5 **Understanding the Intrinsic Memorability of Images**, P. Isola, D. Parikh, A. Torralba, A. Oliva
- W6 **An ideal observer model for identifying the reference frame of object**, s J. Austerweil, A. Friesen, T. Griffiths
- W7 **On the Analysis of Multi-Channel Neural Spike Data**, a B. Chen, D. Carlson, L. Carin
- W8 **Select and Sample: A Model of Efficient Neural Inference and Learning**, J. Shelton, J. Bornschein, A. Sheikh, P. Berkes, J. Lucke
- W9 **Neural Reconstruction with Approximate Message Passing (NeuRAMP)**, A. Fletcher, S. Rangan, L. Varshney, A. Bhargava
- W10 **Two is better than one: distinct roles for familiarity and recollection in retrieving palimpsest memories**, C. Savin, P. Dayan, M. Lengyel
- W11 **Efficient coding with a population of Linear-Nonlinear neurons**, y. karklin, E. Simoncelli

WEDNESDAY - CONFERENCE

- W31 **Active Learning Ranking from Pairwise Preferences with Almost Optimal Query Complexity**, N. Ailon
- W32 **Selective Prediction of Financial Trends with Hidden Markov Models**, D. Pidan, R. El-Yaniv
- W33 **Active Classification based on Value of Classifier**, T. Gao, D. Koller
- W34 **The Fast Convergence of Boosting**, M. Telgarsky
- W35 **Variance Penalizing AdaBoost**, P. Shivaswamy, T. Jebara
- W36 **Kernel Bayes' Rule**, K. Fukumizu, L. Song, A. Gretton
- W37 **Multiple Instance Learning on Structured Data**, D. Zhang, Y. Liu, L. Si, J. Zhang, R. Lawrence
- W38 **Structured Learning for Cell Tracking**, X. Lou, F. Hamprecht
- W39 **Projection onto A Nonnegative Max-Heap**, J. Liu, L. Sun, J. Ye
- W40 **Shallow vs. Deep Sum-Product Networks**, O. Delalleau, Y. Bengio
- W41 **Unfolding Recursive Autoencoders for Paraphrase Detection**, R. Socher, E. Huang, J. Pennin, A. Ng, C. Manning
- W42 **Practical Variational Inference for Neural Networks**, A. Graves
- W43 **A Multilinear Subspace Regression Method Using Orthogonal Tensors Decompositions**, Q. Zhao, C. Caiafa, D. Mandic, L. Zhang, T. Ball, A. Schulze-bonhage, A. Cichocki
- W44 **Sparse recovery by thresholded non-negative least squares**, M. Slawski, M. Hein
- W45 **Structured sparse coding via lateral inhibition**, a. szlam, K. Gregor, Y. LeCun
- W46 **Efficient Methods for Overlapping Group Lasso**, L. Yuan, J. Liu, J. Ye
- W47 **Generalized Beta Mixtures of Gaussians**, A. Armagan, D. Dunson, M. Clyde
- W48 **Sparse Manifold Clustering and Embedding**, E. Elhamifar, R. Vidal
- W49 **Data Skeletonization via Reeb Graphs**, X. Ge, I. Safa, M. Belkin, Y. Wang
- W50 **Sparse Filtering**, J. Ngiam, P. Koh, Z. Chen, S. Bhaskar, A. Ng
- W51 **ICA with Reconstruction Cost for Efficient Overcomplete Feature Learning**, Q. Le, A. Karpenko, J. Ngiam, A. Ng
- W52 **Budgeted Optimization with Concurrent Stochastic-Duration Experiments**, J. Azimi, A. Fern, X. Fern
- W53 **Fast and Accurate k-means For Large Datasets**, M. Shindler, A. Wong, A. Meyerson
- W54 **Scalable Training of Mixture Models via Coresets**, D. Feldman, M. Faulkner, A. Krause
- W55 **Minimax Localization of Structural Information in Large Noisy Matrices**, M. Kolar, S. Balakrishnan, A. Rinaldo, A. Singh
- W56 **Noise Thresholds for Spectral Clustering**, S. Balakrishnan, M. Xu, A. Krishnamurthy, A. Singh
- W57 **Maximum Covariance Unfolding : Manifold Learning for Bimodal Data**, V. Mahadevan, C. Wong, J. Costa Pereira, T. Liu, N. Vasconcelos, L. Saul
- W58 **The Manifold Tangent Classifier**, S. Rifai, Y. Dauphin, P. Vincent, Y. Bengio, X. Muller
- W59 **Dimensionality Reduction Using the Sparse Linear Mode**, I. I. Gkioulekas, T. Zickler
- W60 **Large-Scale Sparse Principal Component Analysis with Application to Text Data**, Y. Zhang, L. Ghaoui
- W61 **Divide-and-Conquer Matrix Factorization**, L. Mackey, A. Talwalkar, M. Jordan
- W62 **Convergent Bounds on the Euclidean Distance**, Y. Hwang, H. Ahn
- W63 **Directed Graph Embedding: an Algorithm based on Continuous Limits of Laplacian-type Operator**, s D. Perrault-Joncas, M. Meila
- W64 **Improving Topic Coherence with Regularized Topic Models**, D. Newman, E. Bonilla, W. Buntine
- W65 **Expressive Power and Approximation Errors of Restricted Boltzmann Machines**, G. Montufar, J. Rauh, N. Ay
- W66 **Complexity of Inference in Latent Dirichlet Allocation**, D. Sontag, D. Roy
- W67 **Hierarchically Supervised Latent Dirichlet Allocation**, A. Perotte, F. Wood, N. Elhadad, N. Bartlett
- W68 **Distributed Delayed Stochastic Optimization**, A. Agarwal, J. Duchi
- W69 **A concave regularization technique for sparse mixture models**, M. Larsson, J. Ugander
- W70 **Fast approximate submodular minimization**, S. Jegelka, H. Lin, J. Bilmes

WEDNESDAY - CONFERENCE

- W71 **Convergence Rates of Inexact Proximal-Gradient Methods for Convex Optimization**, M. Schmidt, N. Le Roux, F. Bach
- W72 **Linearized Alternating Direction Method with Adaptive Penalty for Low-Rank Representation**, Z. Lin, R. Liu, Z. Su
- W73 **Approximating Semidefinite Programs in Sublinear Time**, D. Garber, E. Hazan
- W74 **Hogwild: A Lock-Free Approach to Parallelizing Stochastic Gradient Descent**, B. Recht, C. Re, S. Wright, F. Niu
- W75 **Beating SGD: Learning SVMs in Sublinear Time**, E. Hazan, T. Koren, N. Srebro
- W76 **Learning large-margin halfspaces with more malicious noise**, P. Long, R. Servedio
- W77 **k-NN Regression Adapts to Local Intrinsic Dimension**, S. Kpotufe
- W78 **A Collaborative Mechanism for Crowdsourcing Prediction Problems**, J. Abernethy, R. Frongillo
- W79 **Multi-armed bandits on implicit metric spaces**, A. Slivkins
- W80 **Efficient Online Learning via Randomized Rounding**, N. Cesa-Bianchi, O. Shamir
- W81 **From Bandits to Experts: On the Value of Side-Observations**, S. Mannor, O. Shamir
- W82 **Improved Algorithms for Linear Stochastic Bandits**, Y. Abbasi-yadkori, D. Pal, C. Szepesvari
- W83 **Stochastic convex optimization with bandit feedback**, A. Agarwal, D. Foster, D. Hsu, S. Kakade, A. Rakhlin
- W84 **Predicting Dynamic Difficulty**, O. Missura, T. Gaertner
- W85 **Lower Bounds for Passive and Active Learning**, M. Raginsky, A. Rakhlin
- W86 **Optimal learning rates for least squares SVMs using Gaussian kernels**, M. Eberts, I. Steinwart
- W87 **Phase transition in the family of p-resistances**, M. Alamgir, U. von Luxburg
- W88 **Bayesian Spike-Triggered Covariance Analysis**, I. Park, J. Pillow
- W89 **Learning to Learn with Compound HD Models**, R. Salakhutdinov, J. Tenenbaum, A. Torralba
- W90 **Reconstructing Patterns of Information Diffusion from Incomplete Observations**, F. Chierichetti, J. Kleinberg, D. Liben-Nowell
- W91 **Continuous-Time Regression Models for Longitudinal Networks**, D. Vu, A. Asuncion, D. Hunter, P. Smyth
- W92 **Differentially Private M-Estimators**, J. Lei
- W93 **Bayesian Bias Mitigation for Crowdsourcing**, F. Wauthier, M. Jordan
- W94 **t-divergence Based Approximate Inference**, N. Ding, S. Vishwanathan, Y. Qi
- W95 **Query-Aware MCMC**, M. Wick, A. McCallum
- W96 **Learning sparse inverse covariance matrices in the presence of confounders**, O. Stegle, C. Lippert, J. Mooij, N. Lawrence, K. Borgwardt
- W97 **Gaussian Process Training with Input Noise**, A. McHutchon, C. Rasmussen
- W98 **Iterative Learning for Reliable Crowdsourcing Systems**, D. Karger, S. Oh, D. Shah
- W99 **High-Dimensional Graphical Model Selection: Tractable Graph Families and Necessary Conditions**, A. Anandkumar, V. Tan, A. Willsky
- W100 **On Learning Discrete Graphical Models using Greedy Methods**, A. Jalali, C. Johnson, P. Ravikumar
- W101 **Clustered Multi-Task Learning Via Alternating Structure Optimization**, J. Zhou, J. Chen, J. Ye

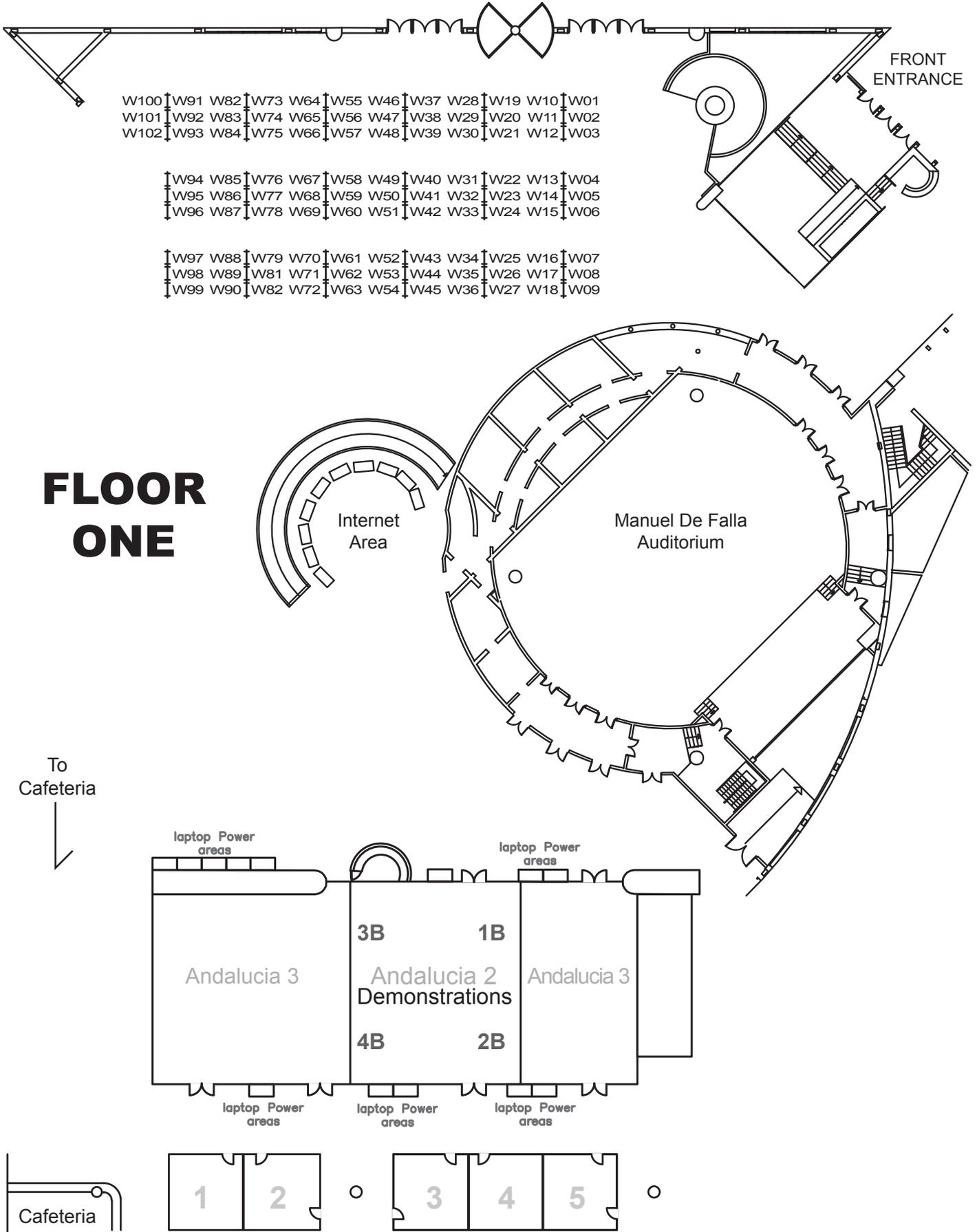


DEMONSTRATIONS

5:45 – 11:59 PM

- 1B **Real-time social media analysis with TWIMPACT**
Mikio Braun, Matthias Jugel, Klaus-Robert Müller, TU Berlin
- 2B **AIsoy1, a Robot that Perceives, Feels and Makes Decisions**
Diego García Sánchez, AIsoy Robotics S.L.; David Rios Insua, Rey Juan Carlos University
- 3B **Real-time Multi-class Segmentation using Depth Cues**
Clement Farabet, Nathan Silberman, New York University
- 4B **Contour-Based Large Scale Image Retrieval Platform**
Rong Zhou, Liqing Zhang, Shanghai Jiao Tong University

WEDNESDAY POSTER FLOORPLAN



W1 A rational model of causal inference with continuous causes

M. Pacer mpacer@berkeley.edu
 T. Griffiths tom_griffiths@berkeley.edu
 University of California, Berkeley

Rational models of causal induction have been successful in accounting for people's judgments about the existence of causal relationships. However, these models have focused on explaining inferences from discrete data of the kind that can be summarized in a 2 \times 2 contingency table. This severely limits the scope of these models, since the world often provides non-binary data. We develop a new rational model of causal induction using continuous dimensions, which aims to diminish the gap between empirical and theoretical approaches and real-world causal induction. This model successfully predicts human judgments from previous studies better than models of discrete causal inference, and outperforms several other plausible models of causal induction with continuous causes in accounting for people's inferences in a new experiment.

Subject Area: Cognitive Science

W2 Testing a Bayesian Measure of Representativeness Using a Large Image Database

J. Abbott joshua.abbott@berkeley.edu
 T. Griffiths tom_griffiths@berkeley.edu
 University of California, Berkeley
 K. Heller kheller@gmail.com
 Massachusetts Institute of Technology
 Z. Ghahramani zoubin@eng.cam.ac.uk
 University of Cambridge & CMU

How do people determine which elements of a set are most representative of that set? We extend an existing Bayesian measure of representativeness, which indicates the representativeness of a sample from a distribution, to define a measure of the representativeness of an item to a set. We show that this measure is formally related to a machine learning method known as Bayesian Sets. Building on this connection, we derive an analytic expression for the representativeness of objects described by a sparse vector of binary features. We then apply this measure to a large database of images, using it to determine which images are the most representative members of different sets. Comparing the resulting predictions to human judgments of representativeness provides a test of this measure with naturalistic stimuli, and illustrates how databases that are more commonly used in computer vision and machine learning can be used to evaluate psychological theories.

Subject Area: Cognitive Science

W3 How Do Humans Teach: On Curriculum Learning and Teaching Dimension

F. Khan faisal@cs.wisc.edu
 X. Zhu jerryzhu@cs.wisc.edu
 B. Mutlu bilge@cs.wisc.edu
 University of Wisconsin-Madison

We study the empirical strategies that humans follow as they teach a target concept with a simple 1D threshold to a robot. Previous studies of computational teaching, particularly the teaching dimension model and the curriculum learning principle, offer contradictory predictions on what optimal strategy the teacher should follow in this teaching task. We show through behavioral studies that humans employ three distinct teaching strategies, one of which is consistent with the curriculum learning principle, and propose a novel theoretical framework as a potential explanation for this strategy. This framework, which assumes a teaching goal of minimizing the learner's expected generalization error at each iteration, extends the standard teaching dimension model and offers a theoretical justification for curriculum learning.

Subject Area: Cognitive Science

W4 Probabilistic Modeling of Dependencies Among Visual Short-Term Memory Representations

E. Orhan eorhan@bcs.rochester.edu
 R. Jacobs robbie@bcs.rochester.edu
 University of Rochester

Extensive evidence suggests that items are not encoded independently in visual short-term memory (VSTM). However, previous research has not quantitatively considered how the encoding of an item influences the encoding of other items. Here, we model the dependencies among VSTM representations using a multivariate Gaussian distribution with a stimulus-dependent mean and covariance matrix. We report the results of an experiment designed to determine the specific form of the stimulus-dependence of the mean and the covariance matrix. We find that the magnitude of the covariance between the representations of two items is a monotonically decreasing function of the difference between the items' feature values, similar to a Gaussian process with a distance-dependent, stationary kernel function. We further show that this type of covariance function can be explained as a natural consequence of encoding multiple stimuli in a population of neurons with correlated responses.

Subject Area: Cognitive Science

W5 Understanding the Intrinsic Memorability of Images

P. Isola phillipi@mit.edu
 A. Torralba torralba@csail.mit.edu
 A. Oliva oliva@mit.edu
 Massachusetts Institute of Technology
 D. Parikh dparikh@ttic.edu
 TTIC

Artists, advertisers, and photographers are routinely presented with the task of creating an image that a viewer will remember. While it may seem like image memorability is purely subjective, recent work shows that it is not an inexplicable phenomenon: variation in memorability of images is consistent across subjects, suggesting that some images are intrinsically more memorable than others, independent of a subjects' contexts and biases. In this paper, we used the publicly available memorability dataset of Isola et al., and augmented the object and scene annotations with interpretable spatial, content, and aesthetic image properties. We used a feature-selection scheme with desirable explaining-away properties to determine a compact set of attributes that characterizes the memorability of any individual image. We find that images of enclosed spaces containing people with visible faces are memorable, while images of vistas and peaceful scenes are not. Contrary to popular belief, unusual or aesthetically pleasing scenes do not tend to be highly memorable. This work represents one of the first attempts at understanding intrinsic image memorability, and opens a new domain of investigation at the interface between human cognition and computer vision.

Subject Area: Cognitive Science

W6 An ideal observer model for identifying the reference frame of objects

J. Austerweil joseph.austerweil@gmail.com
 T. Griffiths tom_griffiths@berkeley.edu
 University of California, Berkeley
 A. Friesen afriesen@cs.washington.edu
 University of Washington

The object people perceive in an image can depend on its orientation relative to the scene it is in (its reference frame). For example, the images of the symbols \times and \times differ by a 45 degree rotation. Although real scenes have multiple images and reference frames, psychologists have focused on scenes with only one reference frame. We propose an ideal observer model based on nonparametric Bayesian statistics for inferring the number of reference frames in a scene and their parameters. When an ambiguous image could be assigned to two conflicting reference frames, the model predicts two factors should influence the reference frame inferred for the image: The image should be more likely to share the reference frame of the closer object (proximity) and it should be more likely to share the reference frame containing the most objects (alignment). We confirm people use both cues using a novel methodology that allows for easy testing of human reference frame inference.

Subject Area: Cognitive Science

W7 On the Analysis of Multi-Channel Neural Spike Data

B. Chen bc69@duke.edu
 D. Carlson david.carlson@duke.edu
 L. Carin lcarin@duke.edu
 Duke University

Nonparametric Bayesian methods are developed for analysis of multi-channel spike-train data, with the feature learning and spike sorting performed jointly. The feature learning and sorting are performed simultaneously across all channels. Dictionary learning is implemented via the beta-Bernoulli process, with spike sorting performed via the dynamic hierarchical Dirichlet process (dHDP), with these two models coupled. The dHDP is augmented to eliminate refractoryperiod violations, it allows the "appearance" and "disappearance" of neurons over time, and it models smooth variation in the spike statistics.

Subject Area: Neuroscience

W8 Select and Sample - A Model of Efficient Neural Inference and Learning

J. Shelton shelton@fias.uni-frankfurt.de
 A. Sheikh sheikh@fias.uni-frankfurt.de
 J. Bornschein bornschein@fias.uni-frankfurt.de
 J. Lucke luecke@fias.uni-frankfurt.de
 Goethe-University Frankfurt
 P. Berkes berkes@brandeis.edu
 Brandeis University

An increasing number of experimental studies indicate that perception encodes a posterior probability distribution over possible causes of sensory stimuli, which is used to act close to optimally in the environment. One outstanding difficulty with this hypothesis is that the exact posterior will in general be too complex to be represented directly, and thus neurons will have to represent an approximation of this distribution. Two influential proposals of efficient posterior representation by neural populations are: 1) neural activity represents samples of the underlying distribution, or 2) they represent a parametric representation of a variational approximation of the posterior. We show that these approaches can be combined for an inference scheme that retains the advantages of both: it is able to represent multiple modes and arbitrary correlations, a feature of sampling methods, and it reduces the represented space to regions of high probability mass, a strength of variational approximations. Neurally, the combined method can be interpreted as a feed-forward preselection of the relevant state space, followed by a neural dynamics implementation of Markov Chain Monte Carlo (MCMC) to approximate the posterior over the relevant states. We demonstrate the effectiveness and efficiency of this approach on a sparse coding model. In numerical experiments on artificial data and image patches, we compare the performance of the algorithms to that of exact EM, variational state space selection alone, MCMC alone, and the combined select and sample approach. The select and sample approach integrates the advantages of the sampling and variational approximations, and forms a robust, neurally plausible, and very efficient model of processing and learning in cortical networks. For sparse coding we show applications easily exceeding a thousand observed and a thousand hidden dimensions.

Subject Area: Neuroscience

W9 Neural Reconstruction with Approximate Message Passing (NeuRAMP)

A. Fletcher alyson@eecs.berkeley.edu
University of California, Berkeley
S. Rangan srangan@poly.edu
Polytechnic Institute of New York University
L. Varshney varshney@alum.mit
IBM
A. Bhargava aniruddha@wisc.edu
University of Wisconsin-Madison

Many functional descriptions of spiking neurons assume a cascade structure where inputs are passed through an initial linear filtering stage that produces a low-dimensional signal that drives subsequent nonlinear stages. This paper presents a novel and systematic parameter estimation procedure for such models and applies the method to two neural estimation problems: (i) compressed-sensing based neural mapping from multi-neuron excitation, and (ii) estimation of neural receptive yields in sensory neurons. The proposed estimation algorithm models the neurons via a graphical model and then estimates the parameters in the model using a recently-developed generalized approximate message passing (GAMP) method. The GAMP method is based on Gaussian approximations of loopy belief propagation. In the neural connectivity problem, the GAMP-based method is shown to be computational efficient, provides a more exact modeling of the sparsity, can incorporate nonlinearities in the output and significantly outperforms previous compressed-sensing methods. For the receptive field estimation, the GAMP method can also exploit inherent structured sparsity in the linear weights. The method is validated on estimation of linear nonlinear Poisson (LNP) cascade models for receptive fields of salamander retinal ganglion cells.

Subject Area: Neuroscience

W10 Two is better than one: distinct roles for familiarity and recollection in retrieving palimpsest memories

C. Savin cs664@cam.ac.uk
M. Lengyel m.lengyel@eng.cam.ac.uk
University of Cambridge
P. Dayan dayan@gatsby.ucl.ac.uk
Gatsby Computational Neuroscience Unit

Storing a new pattern in a palimpsest memory system comes at the cost of interfering with the memory traces of previously stored items. Knowing the age of a pattern thus becomes critical for recalling it faithfully. This implies that there should be a tight coupling between estimates of age, as a form of familiarity, and the neural dynamics of recollection, something which current theories omit. Using a normative model of autoassociative memory, we show that a dual memory system, consisting of two interacting modules for familiarity and recollection, has best performance for both recollection and recognition. This finding provides a new window onto actively contentious psychological and neural aspects of recognition memory.

Subject Area: Neuroscience

W11 Efficient coding of natural images with a population of noisy Linear-Nonlinear neurons

Y. karklin yan.karklin@nyu.edu
E. Simoncelli eero.simoncelli@nyu.edu
HHMI / New York University

Efficient coding provides a powerful principle for explaining early sensory coding. Most attempts to test this principle have been limited to linear, noiseless models, and when applied to natural images, have yielded oriented filters consistent with responses in primary visual cortex. Here we show that an efficient coding model that incorporates biologically realistic ingredients -- input and output noise, nonlinear response functions, and a metabolic cost on the firing rate -- predicts receptive fields and response nonlinearities similar to those observed in the retina. Specifically, we develop numerical methods for simultaneously learning the linear filters and response nonlinearities of a population of model neurons, so as to maximize information transmission subject to metabolic costs. When applied to an ensemble of natural images, the method yields filters that are center-surround and nonlinearities that are rectifying. The filters are organized into two populations, with On- and Off-centers, which independently tile the visual space. As observed in the primate retina, the Off-center neurons are more numerous and have filters with smaller spatial extent. In the absence of noise, our method reduces to a generalized version of independent components analysis, with an adapted nonlinear "contrast" function; in this case, the optimal filters are localized and oriented.

Subject Area: Neuroscience

W12 Variational Learning for Recurrent Spiking Networks

D. Rezende danilo.rezende@epfl.ch
D. Wierstra daan.wierstra@epfl.ch
W. Gerstner wulfram.gerstner@epfl.ch
Ecole Polytechnique Federale de Lausanne

We derive a plausible learning rule updating the synaptic efficacies for feedforward, feedback and lateral connections between observed and latent neurons. Operating in the context of a generative model for distributions of spike sequences, the learning mechanism is derived from variational inference principles. The synaptic plasticity rules found are interesting in that they are strongly reminiscent of experimentally found results on Spike Time Dependent Plasticity, and in that they differ for excitatory and inhibitory neurons. A simulation confirms the method's applicability to learning both stationary and temporal spike patterns.

Subject Area: Neuroscience

W13 Empirical models of spiking in neural populations

J. Macke jakob@gatsby.ucl.ac.uk
 L. Buesing lars@gatsby.ucl.ac.uk
 M. Sahani maneesh@gatsby.ucl.ac.uk
 University College London
 J. Cunningham jpc74@cam.ac.uk
 University of Cambridge
 B. Yu byronyu@cmu.edu
 Carnegie Mellon University
 K. Shenoy shenoy@stanford.edu
 Stanford University

Neurons in the neocortex code and compute as part of a locally interconnected population. Large-scale multi-electrode recording makes it possible to access these population processes empirically by fitting statistical models to unaveraged data. What statistical structure best describes the concurrent spiking of cells within a local network? We argue that in the cortex, where firing exhibits extensive correlations in both time and space and where a typical sample of neurons still reflects only a very small fraction of the local population, the most appropriate model captures shared variability by a low-dimensional latent process evolving with smooth dynamics, rather than by putative direct coupling. We test this claim by comparing a latent dynamical model with realistic spiking observations to coupled generalised linear spike-response models (GLMs) using cortical recordings. We find that the latent dynamical approach outperforms the GLM in terms of goodness-of-fit, and reproduces the temporal correlations in the data more accurately. We also compare models whose observations models are either derived from a Gaussian or point-process models, finding that the non-Gaussian model provides slightly better goodness-of-fit and more realistic population spike counts.

Subject Area: Neuroscience

W14 Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials

P. Krähenbühl philkra@gmail.com
 V. Koltun vladlen@stanford.edu
 Stanford University

Most state-of-the-art techniques for multi-class image segmentation and labeling use conditional random fields defined over pixels or image regions. While region-level models often feature dense pairwise connectivity, pixel-level models are considerably larger and have only permitted sparse graph structures. In this paper, we consider fully connected CRF models defined on the complete set of pixels in an image. The resulting graphs have billions of edges, making traditional inference algorithms impractical. Our main contribution is a highly efficient approximate inference algorithm for fully connected CRF models in which the pairwise edge potentials are defined by linear combinations of Gaussian kernels. Our algorithm can approximately minimize fully connected models on tens of thousands of variables in a fraction of a second. Quantitative and qualitative results on the MSRC-21 and PASCAL VOC 2010 datasets

demonstrate that full pairwise connectivity at the pixel level produces significantly more accurate segmentations and pixel-level label assignments.

Subject Area: Vision

W15 Fast and Balanced: Efficient Label Tree Learning for Large Scale Object Recognition

J. Deng jiadeng@cs.princeton.edu
 Princeton University
 S. Satheesh ssanjeew@stanford.edu
 F. Li feifeili@cs.stanford.edu
 Stanford University
 A. Berg aberg@cs.stonybrook.edu
 Stony Brook

We present a novel approach to efficiently learn a label tree for large scale classification with many classes. The key contribution of the approach is a technique to simultaneously determine the structure of the tree and learn the classifiers for each node in the tree. This approach also allows fine grained control over the efficiency vs accuracy trade-off in designing a label tree, leading to more balanced trees. Experiments are performed on large scale image classification with 10184 classes and 9 million images. We demonstrate significant improvements in test accuracy and efficiency with less training time and more balanced trees compared to the previous state of the art by Bengio et al.

Subject Area: Vision

W16 Probabilistic Joint Image Segmentation and Labeling

A. Ion ion@prip.tuwien.ac.at
 TU Wien & IST Austria
 J. Carreira carreira@ins.uni-bonn.de
 C. Sminchisescu cristian.sminchisescu@ins.uni-bonn.de
 University of Bonn

We present a joint image segmentation and labeling model (JSL) which, given a bag of figure-ground segment hypotheses extracted at multiple image locations and scales, constructs a joint probability distribution over both the compatible image interpretations (tilings or image segmentations) composed from those segments, and over their labeling into categories. The process of drawing samples from the joint distribution can be interpreted as first sampling tilings, modeled as maximal cliques, from a graph connecting spatially non-overlapping segments in the bag, followed by sampling labels for those segments, conditioned on the choice of a particular tiling. We learn the segmentation and labeling parameters jointly, based on Maximum Likelihood with a novel Incremental Saddle Point estimation procedure. The partition function over tilings and labelings is increasingly more accurately approximated by including incorrect configurations that a not-yet-competent model rates probable during learning. We show that the proposed methodology matches the current state of the art in the Stanford dataset, as well as in VOC2010, where 41.7% accuracy on the test set is achieved.

Subject Area: Vision

W17 Heavy-tailed Distances for Gradient Based Image Descriptors

Y. Jia jjiayq@eecs.berkeley.edu
 T. Darrell trevor@eecs.berkeley.edu
 UC Berkeley

Many applications in computer vision measure the similarity between images or image patches based on some statistics such as oriented gradients. These are often modeled implicitly or explicitly with a Gaussian noise assumption, leading to the use of the Euclidean distance when comparing image descriptors. In this paper, we show that the statistics of gradient based image descriptors often follow a heavy-tailed distribution, which undermines any principled motivation for the use of Euclidean distances. We advocate for the use of a distance measure based on the likelihood ratio test with appropriate probabilistic models that fit the empirical data distribution. We instantiate this similarity measure with the Gamma-compound-Laplace distribution, and show significant improvement over existing distance measures in the application of SIFT feature matching, at relatively low computational cost.

Subject Area: Vision

W18 θ -MRF: Capturing Spatial and Semantic Structure in the Parameters for Scene Understanding

C. Li cl758@cornell.edu
 A. Saxena asaxena@cs.cornell.edu
 T. Chen tsuhan@ece.cornell.edu
 Cornell University

For most scene understanding tasks (such as object detection or depth estimation), the classifiers need to consider contextual information in addition to the local features. We can capture such contextual information by taking as input the features/attributes from all the regions in the image. However, this contextual dependence also varies with the spatial location of the region of interest, and we therefore need a different set of parameters for each spatial location. This results in a very large number of parameters. In this work, we model the independence properties between the parameters for each location and for each task, by defining a Markov Random Field (MRF) over the parameters. In particular, two sets of parameters are encouraged to have similar values if they are spatially close or semantically close. Our method is, in principle, complementary to other ways of capturing context such as the ones that use a graphical model over the labels instead. In extensive evaluation over two different settings, of multi-class object detection and of multiple scene understanding tasks (scene categorization, depth estimation, geometric labeling), our method beats the state-of-the-art methods in all the four tasks.

Subject Area: Vision

W19 Learning to Agglomerate Superpixel Hierarchies

V. Jain jainv@janelia.hhmi.org
 Howard Hughes Medical Institute
 S. Turaga sturaga@mit.edu
 University College London
 K. Briggman Kevin.Briggman@mpimf-heidelberg.mpg.de
 M. Helmstaedter Moritz.Helmstaedter@mpimf-heidelberg.mpg.de
 W. Denk Winfried.Denk@mpimf-heidelberg.mpg.de
 Max Planck Institute for Medical Research
 H. Seung seung@mit.edu
 Massachusetts Institute of Technology

An agglomerative clustering algorithm merges the most similar pair of clusters at every iteration. The function that evaluates similarity is traditionally hand- designed, but there has been recent interest in supervised or semisupervised settings in which ground-truth clustered data is available for training. Here we show how to train a similarity function by regarding it as the action-value function of a reinforcement learning problem. We apply this general method to segment images by clustering superpixels, an application that we call Learning to Agglomerate Superpixel Hierarchies (LASH). When applied to a challenging dataset of brain images from serial electron microscopy, LASH dramatically improved segmentation accuracy when clustering supervoxels generated by state of the boundary detection algorithms. The naive strategy of directly training only supervoxel similarities and applying single linkage clustering produced less improvement.

Subject Area: Vision\Image Segmentation

W20 Multiple Instance Filtering

K. Wnuk kwnuk@cs.ucla.edu
 S. Soatto soatto@cs.ucla.edu
 UCLA

We propose a robust filtering approach based on semi-supervised and multiple instance learning (MIL). We assume that the posterior density would be unimodal if not for the effect of outliers that we do not wish to explicitly model. Therefore, we seek for a point estimate at the outset, rather than a generic approximation of the entire posterior. Our approach can be thought of as a combination of standard finite-dimensional filtering (Extended Kalman Filter, or Unscented Filter) with multiple instance learning, whereby the initial condition comes with a putative set of inlier measurements. We show how both the state (regression) and the inlier set (classification) can be estimated iteratively and causally by processing only the current measurement. We illustrate our approach on visual tracking problems whereby the object of interest (target) moves and evolves as a result of occlusions and deformations, and partial knowledge of the target is given in the form of a bounding box (training set).

Subject Area: Vision\Motion and Tracking

W21 Im2Text: Describing Images Using 1 Million Captioned Photographs

V. Ordonez vicente.ordonez@gmail.com
 G. Kulkarni girish86@gmail.com
 T. Berg tlberg@cs.sunysb.edu
 Stony Brook University

We develop and demonstrate automatic image description methods using a large captioned photo collection. One contribution is our technique for the automatic collection of this new dataset -- performing a huge number of Flickr queries and then filtering the noisy results down to 1 million images with associated visually relevant captions. Such a collection allows us to approach the extremely challenging problem of description generation using relatively simple non-parametric methods and produces surprisingly effective results. We also develop methods incorporating many state of the art, but fairly noisy, estimates of image content to produce even more pleasing results. Finally we introduce a new objective performance measure for image captioning.

Subject Area: Vision\Object Recognition

W22 Exploiting spatial overlap to efficiently compute appearance distances between image windows

B. Alexe bogdan@vision.ee.ethz.ch
 ETH ZURICH
 V. Petrescu petrescu.viviana@gmail.com
 V. Ferrari ferrari@vision.ee.ethz.ch
 University of Edinburgh

We present a computationally efficient technique to compute the distance of high-dimensional appearance descriptor vectors between image windows. The method exploits the relation between appearance distance and spatial overlap. We derive an upper bound on appearance distance given the spatial overlap of two windows in an image, and use it to bound the distances of many pairs between two images. We propose algorithms that build on these basic operations to efficiently solve tasks relevant to many computer vision applications, such as finding all pairs of windows between two images with distance smaller than a threshold, or finding the single pair with the smallest distance. In experiments on the PASCAL VOC 07 dataset, our algorithms accurately solve these problems while greatly reducing the number of appearance distances computed, and achieve larger speedups than approximate nearest neighbour algorithms based on trees [18] and on hashing [21]. For example, our algorithm finds the most similar pair of windows between two images while computing only 1% of all distances on average.

Subject Area: Vision\Object Recognition

W23 Rapid Deformable Object Detection using Dual-Tree Branch-and-Bound

I. Kokkinos iasonas.kokkinos@ecp.fr
 Ecole Centrale Paris / INRIA Saclay

In this work we use Branch-and-Bound (BB) to efficiently detect objects with deformable part models. Instead of evaluating the classifier score exhaustively over image locations and scales, we use BB to focus on promising image locations. The core problem is to compute bounds that accommodate part deformations; for this we adapt the Dual Trees data structure to our problem. We evaluate our approach using Mixture-of-Deformable Part Models. We obtain exactly the same results but are 10-20 times faster on average. We also develop a multiple-object detection variation of the system, where hypotheses for 20 categories are inserted in a common priority queue. For the problem of finding the strongest category in an image this results in up to a 100-fold speedup.

Subject Area: Vision\Object Recognition

W24 Object Detection with Grammar Models

R. Girshick rbg@cs.uchicago.edu
 University of Chicago
 P. Felzenszwalb pff@brown.edu
 Brown University
 D. Mcallester mcallester@ttic.edu
 TTI-Chicago

Compositional models provide an elegant formalism for representing the visual appearance of highly variable objects. While such models are appealing from a theoretical point of view, it has been difficult to demonstrate that they lead to performance advantages on challenging datasets. Here we develop a grammar model for person detection and show that it outperforms previous high-performance systems on the PASCAL benchmark. Our model represents people using a hierarchy of deformable parts, variable structure and an explicit model of occlusion for partially visible objects. To train the model, we introduce a new discriminative framework for learning structured prediction models from weakly-labeled data.

Subject Area: Vision\Object Recognition

W25 Learning a Tree of Metrics with Disjoint Visual Features

S. Hwang sjhwang@cs.utexas.edu
 K. Grauman grauman@cs.utexas.edu
 University of Texas at Austin
 F. Sha feisha@usc.edu
 University of Southern California

We introduce an approach to learn discriminative visual representations while exploiting external semantic knowledge about object category relationships. Given a hierarchical taxonomy that captures semantic similarity between the objects, we learn a corresponding tree of metrics (ToM). In this tree, we have one metric for each non-leaf node of the object hierarchy, and each metric is responsible for discriminating among its immediate subcategory children. Specifically, a Mahalanobis metric learned for a given node must satisfy the appropriate (dis)similarity constraints generated only among its subtree members' training instances. To further exploit the semantics, we introduce a novel regularizer coupling the metrics that prefers a sparse disjoint set of features to be selected for each metric relative to its ancestor supercategory nodes' metrics. Intuitively, this reflects that visual cues most useful to distinguish the generic classes (e.g., feline vs. canine) should be different than those cues most useful to distinguish their component fine-grained classes (e.g., Persian cat vs. Siamese cat). We validate our approach with multiple image datasets using the WordNet taxonomy, show its advantages over alternative metric learning approaches, and analyze the meaning of attribute features selected by our algorithm.
 Subject Area: Vision\Object Recognition

W26 Learning person-object interactions for action recognition in still images

V. Delaitre vincent.delaitre@ens.fr
 J. Sivic Josef.Sivic@ens.fr
 I. Laptev ivan.laptev@inria.fr
 INRIA / Ecole Normale Supérieure

We investigate a discriminatively trained model of person-object interactions for recognizing common human actions in still images. We build on the locally order-less spatial pyramid bag-of-features model, which was shown to perform extremely well on a range of object, scene and human action recognition tasks. We introduce three principal contributions. First, we replace the standard quantized local HOG/SIFT features with stronger discriminatively trained body part and object detectors. Second, we introduce new person-object interaction features based on spatial co-occurrences of individual body parts and objects. Third, we address the combinatorial problem of a large number of possible interaction pairs and propose a discriminative selection procedure using a linear support vector machine (SVM) with a sparsity inducing regularizer. Learning of action-specific body part and object interactions bypasses the difficult problem of estimating the complete human body pose configuration. Benefits of the proposed model are shown on human action recognition in consumer photographs, outperforming the strong bag-of-features baseline.
 Subject Area: Vision\Object Recognition

W27 Extracting Speaker-Specific Information with a Regularized Siamese Deep Network

K. Chen chen@cs.manchester.ac.uk
 A. Salman salmanaa@cs.man.ac.uk
 University of Manchester

Speech conveys different yet mixed information ranging from linguistic to speaker-specific components, and each of them should be exclusively used in a specific task. However, it is extremely difficult to extract a specific information component given the fact that nearly all existing acoustic representations carry all types of speech information. Thus, the use of the same representation in both speech and speaker recognition hinders a system from producing better performance due to interference of irrelevant information. In this paper, we present a deep neural architecture to extract speaker-specific information from MFCCs. As a result, a multi-objective loss function is proposed for learning speaker-specific characteristics and regularization via normalizing interference of non-speaker related information and avoiding information loss. With LDC benchmark corpora and a Chinese speech corpus, we demonstrate that a resultant speaker-specific representation is insensitive to text/languages spoken and environmental mismatches and hence outperforms MFCCs and other state-of-the-art techniques in speaker recognition. We discuss relevant issues and relate our approach to previous work.
 Subject Area: Speech and Signal Processing

W28 A Machine Learning Approach to Predict Chemical Reactions

M. Kayala mkayala@ics.uci.edu
 P. Baldi pfbaldi@ics.uci.edu
 University of California, Irvine

Being able to predict the course of arbitrary chemical reactions is essential to the theory and applications of organic chemistry. Previous approaches are not high-throughput, are not generalizable or scalable, or lack sufficient data to be effective. We describe single mechanistic reactions as concerted electron movements from an electron orbital source to an electron orbital sink. We use an existing rule-based expert system to derive a dataset consisting of 2,989 productive mechanistic steps and 6.14 million non-productive mechanistic steps. We then pose identifying productive mechanistic steps as a ranking problem: rank potential orbital interactions such that the top ranked interactions yield the major products. The machine learning implementation follows a two-stage approach, in which we first train atom level reactivity filters to prune 94.0% of non-productive reactions with less than a 0.1% false negative rate. Then, we train an ensemble of ranking models on pairs of interacting orbitals to learn a relative productivity function over single mechanistic reactions in a given system. Without the use of explicit transformation patterns, the ensemble perfectly ranks the productive mechanisms at the top 89.1% of the time, rising to 99.9% of the time when top ranked lists with at most four non-productive reactions are considered. The final system allows multi-step reaction prediction. Furthermore, it is generalizable, making reasonable predictions over reactants and conditions which the rule-based expert system does not handle.
 Subject Area: Applications

W29 RTRMC: A Riemannian trust-region method for low-rank matrix completion

N. Boumal nicolas.boumal@uclouvain.be
 P. Absil absil@inma.ucl.ac.be
 U.C.Louvain

We consider large matrices of low rank. We address the problem of recovering such matrices when most of the entries are unknown. Matrix completion finds applications in recommender systems. In this setting, the rows of the matrix may correspond to items and the columns may correspond to users. The known entries are the ratings given by users to some items. The aim is to predict the unobserved ratings. This problem is commonly stated in a constrained optimization framework. We follow an approach that exploits the geometry of the low-rank constraint to recast the problem as an unconstrained optimization problem on the Grassmann manifold. We then apply first- and second-order Riemannian trust-region methods to solve it. The cost of each iteration is linear in the number of known entries. Our methods, RTRMC 1 and 2, outperform state-of-the-art algorithms on a wide range of problem instances.

Subject Area: Applications

W30 Randomized Algorithms for Comparison-based Search

D. Tschoop dominique.tschoop@epfl.ch
 S. Diggavi suhas@ee.ucla.edu
 UCLA
 P. Delgosha pdelgosha@ee.sharif.edu
 Sharif University of Technology
 S. Mohajer smohajer@princeton.edu
 Princeton

This paper addresses the problem of finding the nearest neighbor (or one of the R -nearest neighbors) of a query object q in a database of n objects, when we can only use a comparison oracle. The comparison oracle, given two reference objects and a query object, returns the reference object most similar to the query object. The main problem we study is how to search the database for the nearest neighbor (NN) of a query, while minimizing the questions. The difficulty of this problem depends on properties of the underlying database. We show the importance of a characterization: *combinatorial disorder* D which defines approximate triangle inequalities on ranks. We present a lower bound of $\Omega(D \log^{n/D} + D^2)$ average number of questions in the search phase for any randomized algorithm, which demonstrates the fundamental role of D for worst case behavior. We develop a randomized scheme for NN retrieval in $O(D^3 \log^2 n + D \log^2 n \log \log n^{D^3})$ questions. The learning requires asking $O(D^3 \log^2 n + D \log^2 n \log \log n^{D^3})$ questions and $O(n \log^2 n / \log(2D))$ bits to store.

Subject Area: Applications

W31 Active Learning Ranking from Pairwise Preferences with Almost Optimal Query Complexity

N. Ailon nailon@cs.technion.ac.il
 Technion

Given a set V of n elements we wish to linearly order them using pairwise preference labels which may be non-transitive (due to irrationality or arbitrary noise). The goal is to linearly order the elements while disagreeing with as few pairwise preference labels as possible. Our performance is measured by two parameters: The number of disagreements (loss) and the query complexity (number of pairwise preference labels). Our algorithm adaptively queries at most $O(n \text{ poly}(\log n, \varepsilon^{-1}))$ preference labels for a regret of ε times the optimal loss. This is strictly better, and often significantly better than what non-adaptive sampling could achieve. Our main result helps settle an open problem posed by learning-to-rank (from pairwise information) theoreticians and practitioners: What is a provably correct way to sample preference labels?

Subject Area: Applications

W32 Selective Prediction of Financial Trends with Hidden Markov Models

D. Pidan dmitry77@gmail.com
 R. El-Yaniv rani@cs.technion.ac.il
 Technion

Focusing on short term trend prediction in a financial context, we consider the problem of selective prediction whereby the predictor can abstain from prediction in order to improve performance. We examine two types of selective mechanisms for HMM predictors. The first is a rejection in the spirit of Chow's well-known ambiguity principle. The second is a specialized mechanism for HMMs that identifies low quality HMM states and abstain from prediction in those states. We call this model selective HMM (sHMM). In both approaches we can trade-off prediction coverage to gain better accuracy in a controlled manner. We compare performance of the ambiguity-based rejection technique with that of the sHMM approach. Our results indicate that both methods are effective, and that the sHMM model is superior.

Subject Area: Applications

W33 Active Classification based on Value of Classifier

T. Gao tianshig@stanford.edu
 D. Koller koller@cs.stanford.edu
 Stanford University

Modern classification tasks usually involve many class labels and can be informed by a broad range of features. Many of these tasks are tackled by constructing a set of classifiers, which are then applied at test time and then pieced together in a fixed procedure determined in advance or at training time. We present an active classification process at the test time, where each classifier in a large ensemble is viewed as a potential observation that might inform our classification process. Observations are then selected dynamically based on previous observations, using a value-theoretic computation that balances an estimate of the expected classification gain from each observation as well as its computational cost. The expected classification gain is computed using a probabilistic model that uses the outcome from previous observations. This active classification process is applied at test time for each individual test instance, resulting in an efficient instance-specific decision path. We demonstrate the benefit of the active scheme on various real-world datasets, and show that it can achieve comparable or even higher classification accuracy at a fraction of the computational costs of traditional methods.

Subject Area: Supervised Learning\Classification

W34 The Fast Convergence of Boosting

M. Telgarsky mtelgars@cs.ucsd.edu
 University of California, San Diego

This manuscript considers the convergence rate of boosting under a large class of losses, including the exponential and logistic losses, where the best previous rate of convergence was $O(\exp(1/\epsilon^2))$. First, it is established that the setting of weak learnability aids the entire class, granting a rate $O(\ln(1/\epsilon))$. Next, the (disjoint) conditions under which the infimal empirical risk is attainable are characterized in terms of the sample and weak learning class, and a new proof is given for the known rate $O(\ln(1/\epsilon))$. Finally, it is established that any instance can be decomposed into two smaller instances resembling the two preceding special cases, yielding a rate $O(1/\epsilon)$, with a matching lower bound for the logistic loss. The principal technical hurdle throughout this work is the potential unattainability of the infimal empirical risk; the technique for overcoming this barrier may be of general interest.

Subject Area: Supervised Learning

W35 Variance Penalizing AdaBoost

P. Shivaswamy pannaga@cs.cornell.edu
 Cornell University
 T. Jebara jebara@cs.columbia.edu
 Columbia University

This paper proposes a novel boosting algorithm called VadaBoost which is motivated by recent empirical Bernstein bounds. VadaBoost iteratively minimizes a cost function that balances the sample mean and the sample variance of the exponential loss. Each step of the proposed algorithm minimizes the cost efficiently by providing weighted data to a weak learner rather than requiring a brute force evaluation of all possible weak learners. Thus, the proposed algorithm solves a key limitation of previous empirical Bernstein boosting methods which required brute force enumeration of all possible weak learners. Experimental results confirm that the new algorithm achieves the performance improvements of EBoost yet goes beyond decision stumps to handle any weak learner. Significant performance gains are obtained over AdaBoost for arbitrary weak learners including decision trees (CART).

Subject Area: Supervised Learning

W36 Kernel Bayes' Rule

K. Fukumizu fukumizu@ism.ac.jp
 Institute of Statistical Mathematics
 L. Song lesong@cs.cmu.edu
 Carnegie Mellon University
 A. Gretton arthur.gretton@gmail.com
 Gatsby Unit, UCL

A nonparametric kernel-based method for realizing Bayes' rule is proposed, based on kernel representations of probabilities in reproducing kernel Hilbert spaces. The prior and conditional probabilities are expressed as empirical kernel mean and covariance operators, respectively, and the kernel mean of the posterior distribution is computed in the form of a weighted sample. The kernel Bayes' rule can be applied to a wide variety of Bayesian inference problems: we demonstrate Bayesian computation without likelihood, and filtering with a nonparametric state-space model. A consistency rate for the posterior estimate is established.

Subject Area: Supervised Learning

W37 Multiple Instance Learning on Structured Data

D. Zhang danzhang2008@gmail.com
 J. Zhang jian.zhang@gmail.com
 Purdue University
 Y. Liu yanliu.cmu@gmail.com
 University of Southern California
 L. Si lsi@cs.purdue.edu
 R. Lawrence ricklawr@us.ibm.com
 IBM TJ Watson Research Center

Most existing Multiple-Instance Learning (MIL) algorithms assume data instances and/or data bags are independently and identically distributed. But there often exists rich additional dependency/structure information between

instances/bags within many applications of MIL. Ignoring this structure information limits the performance of existing MIL algorithms. This paper explores the research problem as multiple instance learning on structured data (MILSD) and formulates a novel framework that considers additional structure information. In particular, an effective and efficient optimization algorithm has been proposed to solve the original non-convex optimization problem by using a combination of Concave-Convex Constraint Programming (CCCP) method and an adapted Cutting Plane method, which deals with two sets of constraints caused by learning on instances within individual bags and learning on structured data. Our method has the nice convergence property, with specified precision on each set of constraints. Experimental results on three different applications, i.e., webpage classification, market targeting, and protein fold identification, clearly demonstrate the advantages of the proposed method over state-of-the-art methods.
Subject Area: Supervised Learning

W38 Structured Learning for Cell Tracking

X. Lou xinghua.lou@iwr.uni-heidelberg.de
F. Hamprecht fred.hamprecht@iwr.uni-heidelberg.de
University of Heidelberg

We study the problem of learning to track a large quantity of homogeneous objects such as cell tracking in cell culture study and developmental biology. Reliable cell tracking in time-lapse microscopic image sequences is important for modern biomedical research. Existing cell tracking methods are usually kept simple and use only a small number of features to allow for manual parameter tweaking or grid search. We propose a structured learning approach that allows to learn optimum parameters automatically from a training set. This allows for the use of a richer set of features which in turn affords improved tracking compared to recently reported methods on two public benchmark sequences.
Subject Area: Supervised Learning

W39 Projection onto A Nonnegative Max-Heap

J. Liu j.liu@asu.edu
L. Sun liang.sun@asu.edu
J. Ye jieping.ye@asu.edu
Arizona State University

We consider the problem of computing the Euclidean projection of a vector of length p onto a non-negative max-heap—an ordered tree where the values of the nodes are all nonnegative and the value of any parent node is no less than the value(s) of its child node(s). This Euclidean projection plays a building block role in the optimization problem with a non-negative max-heap constraint. Such a constraint is desirable when the features follow an ordered tree structure, that is, a given feature is selected for the given regression/classification task only if its parent node is selected. In this paper, we show that such Euclidean projection problem admits an analytical solution and we develop a top-down algorithm where the key operation is to find the so-called *maximal root-tree* of the subtree rooted at each node. A naive approach for finding the maximal root-tree is to enumerate all the possible root-trees, which,

however, does not scale well. We reveal several important properties of the maximal root-tree, based on which we design a bottom-up algorithm with merge for efficiently finding the maximal root-tree. The proposed algorithm has a (worst-case) linear time complexity for a sequential list, and $O(p^2)$ for a general tree. We report simulation results showing the effectiveness of the max-heap for regression with an ordered tree structure. Empirical results show that the proposed algorithm has an expected linear time complexity for many special cases including a sequential list, a full binary tree, and a tree with depth 1.
Subject Area: Supervised Learning

W40 Shallow vs. Deep Sum-Product Networks

O. Delalleau delallea@iro.umontreal.ca
Y. Bengio bengioy@iro.umontreal.ca
University of Montreal

We investigate the representational power of sum-product networks (computation networks analogous to neural networks, but whose individual units compute either products or weighted sums), through a theoretical analysis that compares deep (multiple hidden layers) vs. shallow (one hidden layer) architectures. We prove there exist families of functions that can be represented much more efficiently with a deep network than with a shallow one, i.e. with substantially fewer hidden units. Such results were not available until now, and contribute to motivate recent research involving learning of deep sum-product networks, and more generally motivate research in Deep Learning.
Subject Area: Supervised Learning

W41 Dynamic Pooling and Unfolding Recursive Autoencoders for Paraphrase Detection

R. Socher richard@socher.org
E. Huang eh Huang@stanford.edu
J. Pennin jpennin@stanford.edu
A. Ng ang@cs.stanford.edu
C. Manning manning@stanford.edu
Stanford University

Paraphrase detection is the task of examining two sentences and determining whether they have the same meaning. In order to obtain high accuracy on this task, thorough syntactic and semantic analysis of the two statements is needed. We introduce a method for paraphrase detection based on recursive autoencoders (RAE). Our unsupervised RAEs are based on a novel unfolding objective and learn feature vectors for phrases in syntactic trees. These features are used to measure the word- and phrase-wise similarity between two sentences. Since sentences may be of arbitrary length, the resulting matrix of similarity measures is of variable size. We introduce a novel dynamic pooling layer which computes a fixed-sized representation from the variable-sized matrices. The pooled representation is then used as input to a classifier. Our method outperforms other state-of-the-art approaches on the challenging MSRP paraphrase corpus.
Subject Area: Supervised Learning

W42 Practical Variational Inference for Neural Networks

A. Graves alex.graves@gmail.com
University of Toronto

Variational methods have been previously explored as a tractable approximation to Bayesian inference for neural networks. However the approaches proposed so far have only been applicable to a few simple network architectures. This paper introduces an easy-to-implement stochastic variational method (or equivalently, minimum description length loss function) that can be applied to most neural networks. Along the way it revisits several common regularisers from a variational perspective. It also provides a simple pruning heuristic that can both drastically reduce the number of network weights and lead to improved generalisation. Experimental results are provided for a hierarchical multidimensional recurrent neural network applied to the TIMIT speech corpus.

Subject Area: Supervised Learning

W43 Multilinear Subspace Regression: An Orthogonal Tensor Decomposition Approach

Q. Zhao qbzha@brain.riken.jp
A. CICHOCKI a.cichocki@riken.jp
RIKEB Brain Science Institute
C. Caiafa ccaiafa@gmail.com
D. Mandic d.mandic@imperial.ac.uk
L. Zhang zhang-lq@cs.sjtu.edu.cn
Shanghai Jiao Tong University
T. Ball tonio.ball@uniklinik-freiburg.de
A. Schulze-bonhage andreas.schulze-bonhage@uniklinik-freiburg.de
Albert-Ludwigs-University

A multilinear subspace regression model based on so called latent variable decomposition is introduced. Unlike standard regression methods which typically employ matrix (2D) data representations followed by vector subspace transformations, the proposed approach uses tensor subspace transformations to model common latent variables across both the independent and dependent data. The proposed approach aims to maximize the correlation between the so derived latent variables and is shown to be suitable for the prediction of multidimensional dependent data from multidimensional independent data, where for the estimation of the latent variables we introduce an algorithm based on Multilinear Singular Value Decomposition (MSVD) on a specially defined cross-covariance tensor. It is next shown that in this way we are also able to unify the existing Partial Least Squares (PLS) and N-way PLS regression algorithms within the same framework. Simulations on benchmark synthetic data confirm the advantages of the proposed approach, in terms of its predictive ability and robustness, especially for small sample sizes. The potential of the proposed technique is further illustrated on a real world task of the decoding of human intracranial electrocorticogram (ECoG) from a simultaneously recorded scalp electroencephalograph (EEG).

Subject Area: Supervised Learning

W44 Sparse recovery by thresholded non-negative least squares

M. Slawski ms@cs.uni-sb.de
M. Hein hein@cs.uni-saarland.de
Saarland University

Non-negative data are commonly encountered in numerous fields, making non-negative least squares regression (NNLS) a frequently used tool. At least relative to its simplicity, it often performs rather well in practice. Serious doubts about its usefulness arise for modern high-dimensional linear models. Even in this setting - unlike first intuition may suggest - we show that for a broad class of designs, NNLS is resistant to overfitting and works excellently for sparse recovery when combined with thresholding, experimentally even outperforming L1-regularization. Since NNLS also circumvents the delicate choice of a regularization parameter, our findings suggest that NNLS may be the method of choice.

Subject Area: Supervised Learning

W45 Structured sparse coding via lateral inhibition

a. szlam aszlam@courant.nyu.edu
K. Gregor karol.gregor@gmail.com
Y. LeCun yann@cs.nyu.edu
New York University

This work describes a conceptually simple method for structured sparse coding and dictionary design. Supposing a dictionary with K atoms, we introduce a structure as a set of penalties or interactions between every pair of atoms. We describe modifications of standard sparse coding algorithms for inference in this setting, and describe experiments showing that these algorithms are efficient. We show that interesting dictionaries can be learned for interactions that encode tree structures or locally connected structures. Finally, we show that our framework allows us to learn the values of the interactions from the data, rather than having them pre-specified.

Subject Area: Supervised Learning

W46 Efficient Methods for Overlapping Group Lasso

L. Yuan lei.yuan@asu.edu
J. Liu j.liu@asu.edu
J. Ye jieping.ye@asu.edu
Arizona State University

The group Lasso is an extension of the Lasso for feature selection on (predefined) non-overlapping groups of features. The non-overlapping group structure limits its applicability in practice. There have been several recent attempts to study a more general formulation, where groups of features are given, potentially with overlaps between the groups. The resulting optimization is, however, much more challenging to solve due to the group overlaps. In this paper, we consider the efficient optimization of the overlapping group Lasso penalized problem. We reveal several key properties of the proximal operator associated with the overlapping group Lasso, and compute the proximal operator by solving the smooth and

convex dual problem, which allows the use of the gradient descent type of algorithms for the optimization. We have performed empirical evaluations using both synthetic and the breast cancer gene expression data set, which consists of 8,141 genes organized into (overlapping) gene sets. Experimental results show that the proposed algorithm is more efficient than existing state-of-the-art algorithms.

Subject Area: Supervised Learning

W47 Generalized Beta Mixtures of Gaussians

A. Armagan artin@stat.duke.edu
 D. Dunson dunson@stat.duke.edu
 M. Clyde clyde@stat.duke.edu
 Duke University

In recent years, a rich variety of shrinkage priors have been proposed that have great promise in addressing massive regression problems. In general, these new priors can be expressed as scale mixtures of normals, but have more complex forms and better properties than traditional Cauchy and double exponential priors. We first propose a new class of normal scale mixtures through a novel generalized beta distribution that encompasses many interesting priors as special cases. This encompassing framework should prove useful in comparing competing priors, considering properties and revealing close connections. We then develop a class of variational Bayes approximations through the new hierarchy presented that will scale more efficiently to the types of truly massive data sets that are now encountered routinely.

Subject Area: Supervised Learning

W48 Sparse Manifold Clustering and Embedding

E. Elhamifar ehsan@cis.jhu.edu
 R. Vidal rvidal@cis.jhu.edu

We propose an algorithm called Sparse Manifold Clustering and Embedding (SMCE) for simultaneous clustering and dimensionality reduction of data lying in multiple nonlinear manifolds. Similar to most dimensionality reduction methods, SMCE finds a small neighborhood around each data point and connects each point to its neighbors with appropriate weights. The key difference is that SMCE finds both the neighbors and the weights automatically. This is done by solving a sparse optimization problem, which encourages selecting nearby points that lie in the same manifold and approximately span a low-dimensional affine subspace. The optimal solution encodes information that can be used for clustering and dimensionality reduction using spectral clustering and embedding. Moreover, the size of the optimal neighborhood of a data point, which can be different for different points, provides an estimate of the dimension of the manifold to which the point belongs. Experiments demonstrate that our method can effectively handle multiple manifolds that are very close to each other, manifolds with non-uniform sampling and holes, as well as estimate the intrinsic dimensions of the manifolds.

Subject Area: Manifold Learning

W49 Data Skeletonization via Reeb Graphs

X. Ge gex@cse.ohio-state.edu
 I. Safa issam.safa@gmail.com
 M. Belkin mbelkin@cse.ohio-state.edu
 Y. Wang yusu@cse.ohio-state.edu
 The Ohio State University

Recovering hidden structure from complex and noisy non-linear data is one of the most fundamental problems in machine learning and statistical inference. While such data is often high-dimensional, it is of interest to approximate it with a low-dimensional or even one-dimensional space, since many important aspects of data are often intrinsically low-dimensional. Furthermore, there are many scenarios where the underlying structure is graph-like, e.g. river/road networks or various trajectories. In this paper, we develop a framework to extract, as well as to simplify, a one-dimensional “skeleton” from unorganized data using the Reeb graph. Our algorithm is very simple, does not require complex optimizations and can be easily applied to unorganized high-dimensional data such as point clouds or proximity graphs. It can also represent arbitrary graph structures in the data. We also give theoretical results to justify our method. We provide a number of experiments to demonstrate the effectiveness and generality of our algorithm, including comparisons to existing methods, such as principal curves. We believe that the simplicity and practicality of our algorithm will help to promote skeleton graphs as a data analysis tool for a broad range of applications.”

Subject Area: Unsupervised & Semi-supervised Learning

W50 Sparse Filtering

J. Ngiam jngiam@cs.stanford.edu
 P. Koh pangwei@cs.stanford.edu
 Z. Chen zhenghao@cs.stanford.edu
 S. Bhaskar sbhaskar@stanford.edu
 A. Ng ang@cs.stanford.edu
 Stanford University

Unsupervised feature learning has been shown to be effective at learning representations that perform well on image, video and audio classification. However, many existing feature learning algorithms are hard to use and require extensive hyperparameter tuning. In this work, we present sparse filtering, a simple new algorithm which is efficient and only has one hyperparameter, the number of features to learn. In contrast to most other feature learning methods, sparse filtering does not explicitly attempt to construct a model of the data distribution. Instead, it optimizes a simple cost function -- the sparsity of ℓ_2 -normalized features -- which can easily be implemented in a few lines of MATLAB code. Sparse filtering scales gracefully to handle high-dimensional inputs, and can also be used to learn meaningful features in additional layers with greedy layer-wise stacking. We evaluate sparse filtering on natural images, object classification (STL-10), and phone classification (TIMIT), and show that our method works well on a range of different modalities.

Subject Area: Unsupervised & Semi-supervised Learning

W51 ICA with Reconstruction Cost for Efficient Overcomplete Feature Learning

Q. Le quocle@stanford.edu
 A. Karpenko akarpenko@stanford.edu
 J. Ngiam jngiam@cs.stanford.edu
 A. Ng ang@cs.stanford.edu
 Stanford University

Independent Components Analysis (ICA) and its variants have been successfully used for unsupervised feature learning. However, standard ICA requires an orthonormality constraint to be enforced, which makes it difficult to learn overcomplete features. In addition, ICA is sensitive to whitening. These properties make it challenging to scale ICA to high dimensional data. In this paper, we propose a robust soft reconstruction cost for ICA that allows us to learn highly overcomplete sparse features even on unwhitened data. Our formulation reveals formal connections between ICA and sparse autoencoders, which have previously been observed only empirically. Our algorithm can be used in conjunction with off-the-shelf fast unconstrained optimizers. We show that the soft reconstruction cost can also be used to prevent replicated features in tiled convolutional neural networks. Using our method to learn highly overcomplete sparse features and tiled convolutional neural networks, we obtain competitive performances on a wide variety of object recognition tasks. We achieve state-of-the-art test accuracies on the STL-10 and Hollywood2 datasets.

Subject Area: Unsupervised & Semi-supervised Learning

W52 Budgeted Optimization with Concurrent Stochastic-Duration Experiments

J. Azimi azimi@eecs.oregonstate.edu
 A. Fern afern@eecs.oregonstate.edu
 X. Fern xfern@eecs.oregonstate.edu
 Oregon State University

Budgeted optimization involves optimizing an unknown function that is costly to evaluate by requesting a limited number of function evaluations at intelligently selected inputs. Typical problem formulations assume that experiments are selected one at a time with a limited total number of experiments, which fail to capture important aspects of many real-world problems. This paper defines a novel problem formulation with the following important extensions: 1) allowing for concurrent experiments; 2) allowing for stochastic experiment durations; and 3) placing constraints on both the total number of experiments and the total experimental time. We develop both offline and online algorithms for selecting concurrent experiments in this new setting and provide experimental results on a number of optimization benchmarks. The results show that our algorithms produce highly effective schedules compared to natural baselines.

Subject Area: Unsupervised & Semi-supervised Learning

W53 Fast and Accurate k-means For Large Datasets

M. Shindler shindler@eecs.oregonstate.edu
 Oregon State University
 A. Wong alexw@seas.ucla.edu
 University of California, Los Angeles
 A. Meyerson awm@cs.ucla.edu
 Google, Inc.

Clustering is a popular problem with many applications. We consider the k -means problem in the situation where the data is too large to be stored in main memory and must be accessed sequentially, such as from a disk, and where we must use as little memory as possible. Our algorithm is based on recent theoretical results, with significant improvements to make it practical. Our approach greatly simplifies a recently developed algorithm, both in design and in analysis, and eliminates large constant factors in the approximation guarantee, the memory requirements, and the running time. We then incorporate approximate nearest neighbor search to compute k -means in $o(nk)$ (where n is the number of data points; note that computing the cost, given a solution, takes $\Theta(nk)$ time). We show that our algorithm compares favorably to existing algorithms - both theoretically and experimentally, thus providing state-of-the-art performance in both theory and practice.

Subject Area: Unsupervised & Semi-supervised Learning

W54 Scalable Training of Mixture Models via Coresets

D. Feldman dannyf@caltech.edu
 M. Faulkner mfaulk@caltech.edu
 Caltech
 A. Krause krausea@ethz.ch
 ETH Zurich

How can we train a statistical mixture model on a massive data set? In this paper, we show how to construct coresets for mixtures of Gaussians and natural generalizations. A coreset is a weighted subset of the data, which guarantees that models fitting the coreset will also provide a good fit for the original data set. We show that, perhaps surprisingly, Gaussian mixtures admit coresets of size *independent* of the size of the data set. More precisely, we prove that a weighted set of $O(dk^3/\epsilon^2)$ data points suffices for computing a $(1 + \epsilon)$ -approximation for the optimal model on the original n data points. Moreover, such coresets can be efficiently constructed in a map-reduce style computation, as well as in a streaming setting. Our results rely on a novel reduction of statistical estimation to problems in computational geometry, as well as new complexity results about mixtures of Gaussians. We empirically evaluate our algorithms on several real data sets, including a density estimation problem in the context of earthquake detection using accelerometers in mobile phones.

Subject Area: Unsupervised & Semi-supervised Learning

W55 Minimax Localization of Structural Information in Large Noisy Matrices

M. Kolar mladenk@cs.cmu.edu
 S. Balakrishnan sbalakri@cs.cmu.edu
 A. Rinaldo arinaldo@stat.cmu.edu
 A. Singh aartisingh@cmu.edu
 Carnegie Mellon University

We consider the problem of identifying a sparse set of relevant columns and rows in a large data matrix with highly corrupted entries. This problem of identifying groups from a collection of bipartite variables such as proteins and drugs, biological species and gene sequences, malware and signatures, etc is commonly referred to as biclustering or co-clustering. Despite its great practical relevance, and although several ad-hoc methods are available for biclustering, theoretical analysis of the problem is largely non-existent. The problem we consider is also closely related to structured multiple hypothesis testing, an area of statistics that has recently witnessed a flurry of activity. We make the following contributions: i) We prove lower bounds on the minimum signal strength needed for successful recovery of a bicluster as a function of the noise variance, size of the matrix and bicluster of interest. ii) We show that a combinatorial procedure based on the scan statistic achieves this optimal limit. iii) We characterize the SNR required by several computationally tractable procedures for biclustering including element-wise thresholding, column/row average thresholding and a convex relaxation approach to sparse singular vector decomposition.

Subject Area: Unsupervised & Semi-supervised Learning

W56 Noise Thresholds for Spectral Clustering

S. Balakrishnan sbalakri@cs.cmu.edu
 M. Xu minx@cs.cmu.edu
 A. Krishnamurthy akshaykr@cs.cmu.edu
 A. Singh aartisingh@cmu.edu
 Carnegie Mellon University

Although spectral clustering has enjoyed considerable empirical success in machine learning, its theoretical properties are not yet fully developed. We analyze the performance of a spectral algorithm for hierarchical clustering and show that on a class of hierarchically structured similarity matrices, this algorithm can tolerate noise that grows with the number of data points while still perfectly recovering the hierarchical clusters with high probability. We additionally improve upon previous results for k-way spectral clustering to derive conditions under which spectral clustering makes no mistakes. Further, using minimax analysis, we derive tight upper and lower bounds for the clustering problem and compare the performance of spectral clustering to these information theoretic limits. We also present experiments on simulated and real world data illustrating our results.

Subject Area: Unsupervised & Semi-supervised Learning

W57 Maximum Covariance Unfolding : Manifold Learning for Bimodal Data

V. Mahadevan vmahadev@ucsd.edu
 C. Wong cwwong@ieee.org
 J. Costa Pereira josecp@ucsd.edu
 T. Liu tliu@ucsd.edu
 N. Vasconcelos nuno@ece.ucsd.edu
 L. Saul saul@cs.ucsd.edu
 University of California, San Diego

We propose maximum covariance unfolding (MCU), a manifold learning algorithm for simultaneous dimensionality reduction of data from different input modalities. Given high dimensional inputs from two different but naturally aligned sources, MCU computes a common low dimensional embedding that maximizes the cross-modal (inter-source) correlations while preserving the local (intra-source) distances. In this paper, we explore two applications of MCU. First we use MCU to analyze EEG-fMRI data, where an important goal is to visualize the fMRI voxels that are most strongly correlated with changes in EEG traces. To perform this visualization, we augment MCU with an additional step for metric learning in the high dimensional voxel space. Second, we use MCU to perform cross-modal retrieval of matched image and text samples from Wikipedia. To manage large applications of MCU, we develop a fast implementation based on ideas from spectral graph theory. These ideas transform the original problem for MCU, one of semidefinite programming, into a simpler problem in semidefinite quadratic linear programming.

Subject Area: Unsupervised & Semi-supervised Learning

W58 The Manifold Tangent Classifier

S. Rifai salahmeister@gmail.com
 Y. Dauphin dauphiya@iro.umontreal.ca
 P. Vincent vincentp@iro.umontreal.ca
 Y. Bengio bengioy@iro.umontreal.ca
 X. Muller xav.muller@gmail.com
 University of Montreal

We combine three important ideas present in previous work for building classifiers: the semi-supervised hypothesis (the input distribution contains information about the classifier), the unsupervised manifold hypothesis (data density concentrates near low-dimensional manifolds), and the manifold hypothesis for classification (different classes correspond to disjoint manifolds separated by low density). We exploit a new algorithm for capturing manifold structure (high-order contractive autoencoders) and we show how it builds a topological atlas of charts, each chart being characterized by the principal singular vectors of the Jacobian of a representation mapping. This representation learning algorithm can be stacked to yield a deep architecture, and we combine it with a domain knowledge-free version of the TangentProp algorithm to encourage the classifier to be insensitive to local directions changes along the manifold. Record-breaking results are obtained and we find that the learned tangent directions are very meaningful.

Subject Area: Unsupervised & Semi-supervised Learning

W59 Dimensionality Reduction Using the Sparse Linear Model

I. Gkioulekas igkiou@seas.harvard.edu
 T. Zickler zickler@eecs.harvard.edu
 Harvard University

We propose an approach for linear unsupervised dimensionality reduction, based on the sparse linear model that has been used to probabilistically interpret sparse coding. We formulate an optimization problem for learning a linear projection from the original signal domain to a lower-dimensional one in a way that approximately preserves, in expectation, pairwise inner products in the sparse domain. We derive solutions to the problem, present nonlinear extensions, and discuss relations to compressed sensing. Our experiments using facial images, texture patches, and images of object categories suggest that the approach can improve our ability to recover meaningful structure in many classes of signals.

Subject Area: Unsupervised and Semi-supervised Learning/ICA, PCA, CCA & Other Linear Models

W60 Large-Scale Sparse Principal Component Analysis with Application to Text Data

Y. Zhang zyw@eecs.berkeley.edu
 L. Ghaoui elghaoui@berkeley.edu
 UC Berkeley

Sparse PCA provides a linear combination of small number of features that maximizes variance across data. Although Sparse PCA has apparent advantages compared to PCA, such as better interpretability, it is generally thought to be computationally much more expensive. In this paper, we demonstrate the surprising fact that sparse PCA can be easier than PCA in practice, and that it can be reliably applied to very large data sets. This comes from a rigorous feature elimination pre-processing result, coupled with the favorable fact that features in real-life data typically have exponentially decreasing variances, which allows for many features to be eliminated. We introduce a fast block coordinate ascent algorithm with much better computational complexity than the existing first-order ones. We provide experimental results obtained on text corpora involving millions of documents and hundreds of thousands of features. These results illustrate how Sparse PCA can help organize a large corpus of text data in a user-interpretable way, providing an attractive alternative approach to topic models.

Subject Area: Unsupervised & Semi-supervised Learning

W61 Divide-and-Conquer Matrix Factorization

L. Mackey lmackey@cs.berkeley.edu
 A. Talwalkar ameeet@eecs.berkeley.edu
 M. Jordan jordan@cs.berkeley.edu
 University of California, Berkeley

This work introduces Divide-Factor-Combine (DFC), a parallel divide-and-conquer framework for noisy matrix factorization. DFC divides a large-scale matrix factorization task into smaller subproblems, solves each subproblem in parallel using an arbitrary base matrix factorization

algorithm, and combines the subproblem solutions using techniques from randomized matrix approximation. Our experiments with collaborative filtering, video background modeling, and simulated data demonstrate the near-linear to super-linear speed-ups attainable with this approach. Moreover, our analysis shows that DFC enjoys high-probability recovery guarantees comparable to those of its base algorithm.

Subject Area: Unsupervised & Semi-supervised Learning

W62 Convergent Bounds on the Euclidean Distance

Y. Hwang cypher@postech.ac.kr
 H. Ahn heekap@postech.ac.kr
 POSTECH

Given a set V of n vectors in d -dimensional space, we provide an efficient method for computing quality upper and lower bounds of the Euclidean distances between a pair of the vectors in V . For this purpose, we define a distance measure, called the MS-distance, by using the mean and the standard deviation values of vectors in V . Once we compute the mean and the standard deviation values of vectors in V in $O(dn)$ time, the MS-distance between them provides upper and lower bounds of Euclidean distance between a pair of vectors in V in constant time. Furthermore, these bounds can be refined further such that they converge monotonically to the exact Euclidean distance within d refinement steps. We also provide an analysis on a random sequence of refinement steps which can justify why MS-distance should be refined to provide very tight bounds in a few steps of a typical sequence. The MS-distance can be used to various problems where the Euclidean distance is used to measure the proximity or similarity between objects. We provide experimental results on the nearest and the farthest neighbor searches.

Subject Area: Unsupervised & Semi-supervised Learning

W63 Directed Graph Embedding: an Algorithm based on Continuous Limits of Laplacian-type Operators

D. Perrault-Joncas dcpj@stat.washington.edu
 M. Meila mmp@stat.washington.edu
 University of Washington

This paper considers the problem of embedding directed graphs in Euclidean space while retaining directional information. We model the observed graph as a sample from a manifold endowed with a vector field, and we design an algorithm that separates and recovers the features of this process: the geometry of the manifold, the data density and the vector field. The algorithm is motivated by our analysis of Laplacian-type operators and their continuous limit as generators of diffusions on a manifold. We illustrate the recovery algorithm on both artificially constructed and real data.

Subject Area: Unsupervised & Semi-supervised Learning

W64 Improving Topic Coherence with Regularized Topic Models

D. Newman newman@uci.edu
 University of California, Irvine
 E. Bonilla edwin.bonilla@nicta.com.au
 W. Buntine wray.buntine@nicta.com.au
 NICTA

Topic models have the potential to improve search and browsing by extracting useful semantic themes from web pages and other text documents. When learned topics are coherent and interpretable, they can be valuable for faceted browsing, results set diversity analysis, and document retrieval. However, when dealing with small collections or noisy text (e.g. web search result snippets or blog posts), learned topics can be less coherent, less interpretable, and less useful. To overcome this, we propose two methods to regularize the learning of topic models. Our regularizers work by creating a structured prior over words that reflect broad patterns in the external data. Using thirteen datasets we show that both regularizers improve topic coherence and interpretability while learning a faithful representation of the collection of interest. Overall, this work makes topic models more useful across a broader range of text data.
 Subject Area: Unsupervised & Semi-supervised Learning

W65 Expressive Power and Approximation Errors of Restricted Boltzmann Machines

G. Montufar montufar@mis.mpg.de
 J. Rauh jrauh@mis.mpg.de
 N. Ay nay@mis.mpg.de
 Max Planck Institute for Mathematics in the Sciences

We present explicit classes of probability distributions that can be learned by Restricted Boltzmann Machines (RBMs) depending on the number of units that they contain, and which are representative for the expressive power of the model. We use this to show that the maximal Kullback-Leibler divergence to the RBM model with n visible and m hidden units is bounded from above by $(n - 1) - \log(m + 1)$. In this way we can specify the number of hidden units that guarantees a sufficiently rich model containing different classes of distributions and respecting a given error tolerance.
 Subject Area: Unsupervised & Semi-supervised Learning

W66 Complexity of Inference in Latent Dirichlet Allocation

D. Sontag dsontag@cs.nyu.edu
 New York University
 D. Roy d.roy@eng.cam.ac.uk
 University of Cambridge

We consider the computational complexity of probabilistic inference in Latent Dirichlet Allocation (LDA). First, we study the problem of finding the maximum a posteriori (MAP) assignment of topics to words, where the document's topic distribution is integrated out. We show that, when the effective number of topics per document is small, exact inference takes polynomial time. In contrast,

we show that, when a document has a large number of topics, finding the MAP assignment of topics to words in LDA is NP-hard. Next, we consider the problem of finding the MAP topic distribution for a document, where the topic-word assignments are integrated out. We show that this problem is also NP-hard. Finally, we briefly discuss the problem of sampling from the posterior, showing that this is NP-hard in one restricted setting, but leaving open the general question.

Subject Area: Unsupervised & Semi-supervised Learning

W67 Hierarchically Supervised Latent Dirichlet Allocation

A. Perotte ajp2120@columbia.edu
 F. Wood fwood@stat.columbia.edu
 N. Elhadad noemie@dbmi.columbia.edu
 N. Bartlett bartlett@stat.columbia.edu
 Columbia University

We introduce hierarchically supervised latent Dirichlet allocation (HSLDA), a model for hierarchically and multiply labeled bag-of-word data. Examples of such data include web pages and their placement in directories, product descriptions and associated categories from product hierarchies, and free-text clinical records and their assigned diagnosis codes. Out-of-sample label prediction is the primary goal of this work, but improved lower-dimensional representations of the bag-of-word data are also of interest. We demonstrate HSLDA on large-scale data from clinical document labeling and retail product categorization tasks. We show that leveraging the structure from hierarchical labels improves out-of-sample label prediction substantially when compared to models that do not.

Subject Area: Unsupervised & Semi-supervised Learning

W68 Distributed Delayed Stochastic Optimization

A. Agarwal alekhagarwal@gmail.com
 J. Duchi jduchi@cs.berkeley.edu
 University of California Berkeley

We analyze the convergence of gradient-based optimization algorithms whose updates depend on delayed stochastic gradient information. The main application of our results is to the development of distributed minimization algorithms where a master node performs parameter updates while worker nodes compute stochastic gradients based on local information in parallel, which may give rise to delays due to asynchrony. Our main contribution is to show that for smooth stochastic problems, the delays are asymptotically negligible. In application to distributed optimization, we show n -node architectures whose optimization error in stochastic problems---in spite of asynchronous delays---scales asymptotically as $O(1/\sqrt{nT})$, which is known to be optimal even in the absence of delays.

Subject Area: Optimization

W69 A concave regularization technique for sparse mixture models

M. Larsson mol23@cornell.edu
 J. Ugander jhu5@cornell.edu
 Cornell University

Latent variable mixture models are a powerful tool for exploring the structure in large datasets. A common challenge for interpreting such models is a desire to impose sparsity, the natural assumption that each data point only contains few latent features. Since mixture distributions are constrained in their L_1 norm, typical sparsity techniques based on L_1 regularization become toothless, and concave regularization becomes necessary. Unfortunately concave regularization typically results in EM algorithms that must perform problematic non-concave M-step maximizations. In this work, we introduce a technique for circumventing this difficulty, using the so-called Mountain Pass Theorem to provide easily verifiable conditions under which the M-step is well-behaved despite the lacking concavity. We also develop a correspondence between logarithmic regularization and what we term the pseudo-Dirichlet distribution, a generalization of the ordinary Dirichlet distribution well-suited for inducing sparsity. We demonstrate our approach on a text corpus, inferring a sparse topic mixture model for 2,406 weblogs.
 Subject Area: Optimization

W70 On fast approximate submodular minimization

S. Jegelka jegelka@tuebingen.mpg.de
 Max Planck Institute for Intelligent Systems
 H. Lin linhui@u.washington.edu
 J. Bilmes bilmes@ee.washington.edu
 University of Washington

We are motivated by an application to extract a representative subset of machine learning training data and by the poor empirical performance we observe of the popular minimum norm algorithm. In fact, for our application, minimum norm can have a running time of about $O(n^7)$ ($O(n^5)$ oracle calls). We therefore propose a fast approximate method to minimize arbitrary submodular functions. For a large sub-class of submodular functions, the algorithm is exact. Other submodular functions are iteratively approximated by tight submodular upper bounds, and then repeatedly optimized. We show theoretical properties, and empirical results suggest significant speedups over minimum norm while retaining higher accuracies.
 Subject Area: Optimization

W71 Convergence Rates of Inexact Proximal-Gradient Methods for Convex Optimization

M. Schmidt mark.schmidt@inria.fr
 N. Le Roux nicolas@le-roux.name
 F. Bach francis.bach@ens.fr
 INRIA - Ecole Normale Supérieure

We consider the problem of optimizing the sum of a smooth convex function and a non-smooth convex function using proximal-gradient methods, where an error is present in the calculation of the gradient of the smooth term or in the proximity operator with respect to the

second term. We show that the basic proximal-gradient method, the basic proximal-gradient method with a strong convexity assumption, and the accelerated proximal-gradient method achieve the same convergence rates as in the error-free case, provided the errors decrease at an appropriate rate. Our experimental results on a structured sparsity problem indicate that sequences of errors with these appealing theoretical properties can lead to practical performance improvements.
 Subject Area: Optimization

W72 Linearized Alternating Direction Method with Adaptive Penalty for Low-Rank Representation

Z. Lin zhoulin@microsoft.com
 Microsoft Research Asia
 R. Liu rslu0705@gmail.com
 Z. Su zxsu@dlut.edu.cn
 Dalian University of Technology

Many machine learning and signal processing problems can be formulated as linearly constrained convex programs, which could be efficiently solved by the alternating direction method (ADM). However, usually the subproblems in ADM are easily solvable only when the linear mappings in the constraints are identities. To address this issue, we propose a linearized ADM (LADM) method by linearizing the quadratic penalty term and adding a proximal term when solving the subproblems. For fast convergence, we also allow the penalty to change adaptively according to a novel update rule. We prove the global convergence of LADM with adaptive penalty (LADMAP). As an example, we apply LADMAP to solve low-rank representation (LRR), which is an important subspace clustering technique yet suffers from high computation cost. By combining LADMAP with a skinny SVD representation technique, we are able to reduce the complexity $O(n^3)$ of the original ADM based method to $O(rn^2)$, where r and n are the rank and size of the representation matrix, respectively, hence making LRR possible for large scale applications. Numerical experiments verify that for LRR our LADMAP based methods are much faster than state-of-the-art algorithms.
 Subject Area: Optimization

W73 Approximating Semidefinite Programs in Sublinear Time

D. Garber dangar@cs.technion.ac.il
 E. Hazan ehazan@ie.technion.ac.il
 Technion

In recent years semidefinite optimization has become a tool of major importance in various optimization and machine learning problems. In many of these problems the amount of data in practice is so large that there is a constant need for faster algorithms. In this work we present the first sublinear time approximation algorithm for semidefinite programs which we believe may be useful for such problems in which the size of data may cause even linear time algorithms to have prohibitive running times in practice. We present the algorithm and its analysis alongside with some theoretical lower bounds and an improved algorithm for the special problem of supervised learning of a distance metric.
 Subject Area: Optimization\Convex Optimization

W74 Hogwild: A Lock-Free Approach to Parallelizing Stochastic Gradient Descent

B. Recht brecht@cs.wisc.edu
 C. Re chrisre@cs.wisc.edu
 S. Wright swright@cs.wisc.edu
 F. Niu leonn@cs.wisc.edu
 University of Wisconsin-Madison

Stochastic Gradient Descent (SGD) is a popular algorithm that can achieve state-of-the-art performance on a variety of machine learning tasks. Several researchers have recently proposed schemes to parallelize SGD, but all require performance-destroying memory locking and synchronization. This work aims to show using novel theoretical analysis, algorithms, and implementation that SGD can be implemented *without any locking*. We present an update scheme called Hogwild which allows processors access to shared memory with the possibility of overwriting each other's work. We show that when the associated optimization problem is sparse, meaning most gradient updates only modify small parts of the decision variable, then Hogwild achieves a nearly optimal rate of convergence. We demonstrate experimentally that Hogwild outperforms alternative schemes that use locking by an order of magnitude.

Subject Area: Optimization\Stochastic Methods

W75 Beating SGD: Learning SVMs in Sublinear Time

E. Hazan ehazan@ie.technion.ac.il
 T. Koren tomerk@cs.technion.ac.il
 Technion
 N. Srebro nati@ttic.edu
 TTI-Chicago

We present an optimization approach for linear SVMs based on a stochastic primal-dual approach, where the primal step is akin to an importance-weighted SGD, and the dual step is a stochastic update on the importance weights. This yields an optimization method with a sublinear dependence on the training set size, and the first method for learning linear SVMs with runtime less than the size of the training set required for learning!

Subject Area: Optimization\Stochastic Methods

W76 Learning large-margin halfspaces with more malicious noise

P. Long plong@google.com
 Google
 R. Servedio rocco@cs.columbia.edu
 Columbia University

We describe a simple algorithm that runs in time $\text{poly}(n, 1/\gamma, 1/\epsilon)$ and learns an unknown n -dimensional gamma-margin halfspace to accuracy $1 - \epsilon$ in the presence of malicious noise, when the noise rate is allowed to be as high as $\Theta(\epsilon\gamma\sqrt{\log(1/\gamma)})$. Previous efficient algorithms could only learn to accuracy ϵ in the presence of malicious noise of rate at most $\Theta(\epsilon\gamma)$. Our algorithm does not work by optimizing a convex loss function. We show that

no algorithm for learning gamma-margin halfspaces that minimizes a convex proxy for misclassification error can tolerate malicious noise at a rate greater than $\Theta(\epsilon\gamma)$; this may partially explain why previous algorithms could not achieve the higher noise tolerance of our new algorithm.

Subject Area: Learning Theory

W77 k-NN Regression Adapts to Local Intrinsic Dimension

S. Kpotufe samory@tuebingen.mpg.de
 Max Planck Institute

Many nonparametric regressors were recently shown to converge at rates that depend only on the intrinsic dimension of data. These regressors thus escape the curse of dimension when high-dimensional data has low intrinsic dimension (e.g. a manifold). We show that k -NN regression is also adaptive to intrinsic dimension. In particular our rates are local to a query x and depend only on the way masses of balls centered at x vary with radius. Furthermore, we show a simple way to choose $k = k(x)$ locally at any x so as to nearly achieve the minimax rate at x in terms of the unknown intrinsic dimension in the vicinity of x . We also establish that the minimax rate does not depend on a particular choice of metric space or distribution, but rather that this minimax rate holds for any metric space and doubling measure.

Subject Area: Learning Theory

W78 A Collaborative Mechanism for Crowdsourcing Prediction Problems

J. Abernethy jake@cs.berkeley.edu
 R. Frongillo raf@cs.berkeley.edu
 UC Berkeley

Machine Learning competitions such as the Netflix Prize have proven reasonably successful as a method of "crowdsourcing" prediction tasks. But these competitions have a number of weaknesses, particularly in the incentive structure they create for the participants. We propose a new approach, called a Crowdsourced Learning Mechanism, in which participants collaboratively "learn" a hypothesis for a given prediction task. The approach draws heavily from the concept of a prediction market, where traders bet on the likelihood of a future event. In our framework, the mechanism continues to publish the current hypothesis, and participants can modify this hypothesis by wagering on an update. The critical incentive property is that a participant will profit an amount that scales according to how much her update improves performance on a released test set.

Subject Area: Theory

W79 Multi-armed bandits on implicit metric spaces

A. Slivkins slivkins@microsoft.com
Microsoft Research

The multi-armed bandit (MAB) setting is a useful abstraction of many online learning tasks which focuses on the trade-off between exploration and exploitation. In this setting, an online algorithm has a fixed set of alternatives (“arms”), and in each round it selects one arm and then observes the corresponding reward. While the case of small number of arms is by now well-understood, a lot of recent work has focused on multi-armed bandits with (infinitely) many arms, where one needs to assume extra structure in order to make the problem tractable. In particular, in the Lipschitz MAB problem there is an underlying similarity metric space, known to the algorithm, such that any two arms that are close in this metric space have similar payoffs. In this paper we consider the more realistic scenario in which the metric space is *implicit* -- it is defined by the available structure but not revealed to the algorithm directly. Specifically, we assume that an algorithm is given a tree-based classification of arms. For any given problem instance such a classification implicitly defines a similarity metric space, but the numerical similarity information is not available to the algorithm. We provide an algorithm for this setting, whose performance guarantees (almost) match the best known guarantees for the corresponding instance of the Lipschitz MAB problem.”

Subject Area: Theory\Online Learning

W80 Efficient Online Learning via Randomized Rounding

N. Cesa-Bianchi nicolo.cesa-bianchi@unimi.it
University degli Studi di Milano
O. Shamir ohadsh@microsoft.com
Microsoft Research

Most online algorithms used in machine learning today are based on variants of mirror descent or follow-the-leader. In this paper, we present an online algorithm based on a completely different approach, which combines “random playout” and randomized rounding of loss subgradients. As an application of our approach, we provide the first computationally efficient online algorithm for collaborative filtering with trace-norm constrained matrices. As a second application, we solve an open question linking batch learning and transductive online learning.

Subject Area: Theory\Online Learning

W81 From Bandits to Experts: On the Value of Side-Observations

S. Mannor shie@ee.technion.ac.il
Technion
O. Shamir ohadsh@microsoft.com
Microsoft Research

We consider an adversarial online learning setting where a decision maker can choose an action in every stage of the game. In addition to observing the reward of the chosen

action, the decision maker gets side observations on the reward he would have obtained had he chosen some of the other actions. The observation structure is encoded as a graph, where node i is linked to node j if sampling i provides information on the reward of j . This setting naturally interpolates between the well-known “experts” setting, where the decision maker can view all rewards, and the multi-armed bandits setting, where the decision maker can only view the reward of the chosen action. We develop practical algorithms with provable regret guarantees, which depend on non-trivial graph-theoretic properties of the information feedback structure. We also provide partially-matching lower bounds.

Subject Area: Theory\Online Learning

W82 Improved Algorithms for Linear Stochastic Bandits

Y. Abbasi-yadkori abbasiya@ualberta.ca
C. Szepesvari szepesva@cs.ualberta.ca
University of Alberta
D. Pal dpal@cs.ualberta.ca
Google

We improve the theoretical analysis and empirical performance of algorithms for the stochastic multi-armed bandit problem and the linear stochastic multi-armed bandit problem. In particular, we show that a simple modification of Auer’s UCB algorithm (Auer, 2002) achieves with high probability constant regret. More importantly, we modify and, consequently, improve the analysis of the algorithm for the linear stochastic bandit problem studied by Auer (2002), Dani et al. (2008), Rusmevichientong and Tsitsiklis (2010), Li et al. (2010). Our modification improves the regret bound by a logarithmic factor, though experiments show a vast improvement. In both cases, the improvement stems from the construction of smaller confidence sets. For their construction we use a novel tail inequality for vector-valued martingales.

Subject Area: Theory\Online Learning

W83 Stochastic convex optimization with bandit feedback

A. Agarwal alekhagarwal@gmail.com
University of California Berkeley
D. Foster foster@wharton.upenn.edu
A. Rakhlin rakhlin@gmail.com
University of Pennsylvania
D. Hsu danielhsu@gmail.com
S. Kakade sham@tti-c.org
Microsoft Research

This paper addresses the problem of minimizing a convex, Lipschitz function f over a convex, compact set \mathcal{X} under a stochastic bandit feedback model. In this model, the algorithm is allowed to observe noisy realizations of the function value $f(x)$ at any query point $x \in \mathcal{X}$. We demonstrate a generalization of the ellipsoid algorithm that incurs $O(\text{poly}(d)\sqrt{T})$ regret. Since any algorithm has regret at least $\Omega(\sqrt{T})$ on this problem, our algorithm is optimal in terms of the scaling with T .

Subject Area: Theory\Online Learning

W84 Predicting Dynamic Difficulty

O. Missura olanochka@gmail.com
 T. Gaertner thomas.gaertner@iais.fraunhofer.de
 University of Bonn

Motivated by applications in electronic games as well as teaching systems, we investigate the problem of dynamic difficulty adjustment. The task here is to repeatedly find a game difficulty setting that is neither 'too easy' and bores the player, nor 'too difficult' and overburdens the player. The contributions of this paper are (i) formulation of difficulty adjustment as an online learning problem on partially ordered sets, (ii) an exponential update algorithm for dynamic difficulty adjustment, (iii) a bound on the number of wrong difficulty settings relative to the best static setting chosen in hindsight, and (iv) an empirical investigation of the algorithm when playing against adversaries.

Subject Area: Theory\Online Learning

W85 Lower Bounds for Passive and Active Learning

M. Raginsky mraginsky@gmail.com
 University of Illinois at Urbana-Champaign
 A. Rakhlin rakhlin@gmail.com
 University of Pennsylvania

We develop unified information-theoretic machinery for deriving lower bounds for passive and active learning schemes. Our bounds involve the so-called Alexander's capacity function. The supremum of this function has been recently rediscovered by Hanneke in the context of active learning under the name of "disagreement coefficient." For passive learning, our lower bounds match the upper bounds of Gine and Koltchinskii up to constants and generalize analogous results of Massart and Nédélec. For active learning, we provide first known lower bounds based on the capacity function rather than the disagreement coefficient."

Subject Area: Theory\Statistical Learning Theory

W86 Optimal learning rates for least squares SVMs using Gaussian kernels

M. Eberts Mona.Eberts@mathematik.uni-stuttgart.de
 I. Steinwart ingo.steinwart@mathematik.uni-stuttgart.de
 University of Stuttgart

We prove a new oracle inequality for support vector machines with Gaussian RBF kernels solving the regularized least squares regression problem. To this end, we apply the modulus of smoothness. With the help of the new oracle inequality we then derive learning rates that can also be achieved by a simple data-dependent parameter selection method. Finally, it turns out that our learning rates are asymptotically optimal for regression functions satisfying certain standard smoothness conditions.

Subject Area: Theory\Statistical Learning Theory

W87 Phase transition in the family of p-resistances

M. Alamgir morteza@tuebingen.mpg.de
 U. von Luxburg ulrike.luxburg@tuebingen.mpg.de
 Max Plank Institute for Intelligent Systems

We study the family of p-resistances on graphs for $p \geq 1$. This family generalizes the standard resistance distance. We prove that for any fixed graph, for $p = 1$, the p-resistance coincides with the shortest path distance, for $p = 2$ it coincides with the standard resistance distance, and for $p \rightarrow \infty$ it converges to the inverse of the minimal s-t-cut in the graph. Secondly, we consider the special case of random geometric graphs (such as k-nearest neighbor graphs) when the number n of vertices in the graph tends to infinity. We prove that an interesting phase-transition takes place. There exist two critical thresholds p^* and p^{**} such that if $p < p^*$, then the p-resistance depends on meaningful global properties of the graph, whereas if $p > p^{**}$, it only depends on trivial local quantities and does not convey any useful information. We can explicitly compute the critical values: $p^* = 1 + 1/(d - 1)$ and $p^{**} = 1 + 1/(d - 2)$ where d is the dimension of the underlying space (we believe that the fact that there is a small gap between p^* and p^{**} is an artifact of our proofs. We also relate our findings to Laplacian regularization and suggest to use q-Laplacians as regularizers, where q satisfies $1/p^* + 1/q = 1$.

Subject Area: Theory\Statistical Learning Theory

W88 Bayesian Spike-Triggered Covariance Analysis

I. Park memming@austin.utexas.edu
 J. Pillow pillow@mail.utexas.edu
 University of Texas at Austin

Neurons typically respond to a restricted number of stimulus features within the high-dimensional space of natural stimuli. Here we describe an explicit model-based interpretation of traditional estimators for a neuron's multi-dimensional feature space, which allows for several important generalizations and extensions. First, we show that traditional estimators based on the spike-triggered average (STA) and spike-triggered covariance (STC) can be formalized in terms of the "expected log-likelihood" of a Linear-Nonlinear-Poisson (LNP) model with Gaussian stimuli. This model-based formulation allows us to define maximum-likelihood and Bayesian estimators that are statistically consistent and efficient in a wider variety of settings, such as with naturalistic (non-Gaussian) stimuli. It also allows us to employ Bayesian methods for regularization, smoothing, sparsification, and model comparison, and provides Bayesian confidence intervals on model parameters. We describe an empirical Bayes method for selecting the number of features, and extend the model to accommodate an arbitrary elliptical nonlinear response function, which results in a more powerful and more flexible model for feature space inference. We validate these methods using neural data recorded extracellularly from macaque primary visual cortex."

Subject Area: Probabilistic Models and Methods

W89 Learning to Learn with Compound HD Models

R. Salakhutdinov rsalakhu@utstat.toronto.edu
 University of Toronto
 J. Tenenbaum jbt@mit.edu
 A. Torralba torralba@csail.mit.edu
 Massachusetts Institute of Technology

We introduce HD (or “Hierarchical-Deep”) models, a new compositional learning architecture that integrates deep learning models with structured hierarchical Bayesian models. Specifically we show how we can learn a hierarchical Dirichlet process (HDP) prior over the activities of the top-level features in a Deep Boltzmann Machine (DBM). This compound HDP-DBM model learns to learn novel concepts from very few training examples, by learning low-level generic features, high-level features that capture correlations among low-level features, and a category hierarchy for sharing priors over the high-level features that are typical of different kinds of concepts. We present efficient learning and inference algorithms for the HDP-DBM model and show that it is able to learn new concepts from very few examples on CIFAR-100 object recognition, handwritten character recognition, and human motion capture datasets.

Subject Area: Probabilistic Models and Methods

W90 Reconstructing Patterns of Information Diffusion from Incomplete Observations

F. Chierichetti flavio@chierichetti.name
 J. Kleinberg kleinber@cs.cornell.edu
 Cornell University
 D. Liben-Nowell dlibenno@carleton.edu
 Carleton College

Motivated by the spread of on-line information in general and on-line petitions in particular, recent research has raised the following combinatorial estimation problem. There is a tree T that we cannot observe directly (representing the structure along which the information has spread), and certain nodes randomly decide to make their copy of the information public. In the case of a petition, the list of names on each public copy of the petition also reveals a path leading back to the root of the tree. What can we conclude about the properties of the tree we observe from these revealed paths, and can we use the structure of the observed tree to estimate the size of the full unobserved tree T ? Here we provide the first algorithm for this size estimation task, together with provable guarantees on its performance. We also establish structural properties of the observed tree, providing the first rigorous explanation for some of the unusual structural phenomena present in the spread of real chain-letter petitions on the Internet.

Subject Area: Probabilistic Models and Methods

W91 Continuous-Time Regression Models for Longitudinal Networks

D. Vu dqv100@psu.edu
 D. Hunter dhunter@psu.edu
 Pennsylvania State University
 A. Asuncion asuncion@uci.edu
 P. Smyth smyth@ics.uci.edu
 University of California, Irvine

The development of statistical models for continuous-time longitudinal network data is of increasing interest in machine learning and social science. Leveraging ideas from survival and event history analysis, we introduce a continuous-time regression modeling framework for network event data that can incorporate both time-dependent network statistics and time-varying regression coefficients. We also develop an efficient inference scheme that allows our approach to scale to large networks. On synthetic and real-world data, empirical results demonstrate that the proposed inference approach can accurately estimate the coefficients of the regression model, which is useful for interpreting the evolution of the network; furthermore, the learned model has systematically better predictive performance compared to standard baseline methods.

Subject Area: Probabilistic Models and Methods

W92 Differentially Private M-Estimators

J. Lei lei09@gmail.com
 Carnegie Mellon University

This paper studies privacy preserving M-estimators using perturbed histograms. The proposed approach allows the release of a wide class of M-estimators with both differential privacy and statistical utility without knowing a priori the particular inference procedure. The performance of the proposed method is demonstrated through a careful study of the convergence rates. A practical algorithm is given and applied on a real world data set containing both continuous and categorical variables.

Subject Area: Probabilistic Models and Methods

W93 Bayesian Bias Mitigation for Crowdsourcing

F. Wauthier flw@berkeley.edu
 M. Jordan jordan@cs.berkeley.edu
 UC Berkeley

Biased labelers are a systemic problem in crowdsourcing, and a comprehensive toolbox for handling their responses is still being developed. A typical crowdsourcing application can be divided into three steps: data collection, data curation, and learning. At present these steps are often treated separately. We present Bayesian Bias Mitigation for Crowdsourcing (BBMC), a Bayesian model to unify all three. Most data curation methods account for the $\{it\}$ effects of labeler bias by modeling all labels as coming from a single latent truth. Our model captures the $\{it\}$ sources of bias by describing labelers as influenced by shared random effects. This approach can account for more complex bias patterns that arise in ambiguous or hard labeling tasks and allows us to merge data curation and learning into a single computation. Active learning

integrates data collection with learning, but is commonly considered infeasible with Gibbs sampling inference. We propose a general approximation strategy for Markov chains to efficiently quantify the effect of a perturbation on the stationary distribution and specialize this approach to active learning. Experiments show BBMC to outperform many common heuristics.

Subject Area: Probabilistic Models and Methods

W94 t-divergence Based Approximate Inference

N. Ding ding10@purdue.edu
 S. Vishwanathan vishy@stat.purdue.edu
 Y. Qi alanqi@cs.purdue.edu
 Purdue university

Approximate inference is an important technique for dealing with large, intractable graphical models based on the exponential family of distributions. We extend the idea of approximate inference to the t-exponential family by defining a new t-divergence. This divergence measure is obtained via convex duality between the log-partition function of the t-exponential family and a new t-entropy. We illustrate our approach on the Bayes Point Machine with a Student's t-prior.

Subject Area: Probabilistic Models and Methods

W95 Query-Aware MCMC

M. Wick mwick@cs.umass.edu
 A. McCallum mccallum@cs.umass.edu
 UMass Amherst

Traditional approaches to probabilistic inference such as loopy belief propagation and Gibbs sampling typically compute marginals for all the unobserved variables in a graphical model. However, in many real-world applications the user's interests are focused on a subset of the variables, specified by a query. In this case it would be wasteful to uniformly sample, say, one million variables when the query concerns only ten. In this paper we propose a query-specific approach to MCMC that accounts for the query variables and their generalized mutual information with neighboring variables in order to achieve higher computational efficiency. Surprisingly there has been almost no previous work on query-aware MCMC. We demonstrate the success of our approach with positive experimental results on a wide range of graphical models.

Subject Area: Probabilistic Models and Methods

W96 Efficient inference in matrix-variate Gaussian models with iid observation noise

O. Stegle oliver.stegle@tuebingen.mpg.de
 K. Borgwardt karsten.borgwardt@tuebingen.mpg.de
 C. Lippert christoph.lippert@tuebingen.mpg.de
 Max Planck Institute Biological Cybernetics
 J. Mooij j.mooij@cs.ru.nl
 Radboud University Nijmegen
 N. Lawrence N.Lawrence@shef.ac.uk
 University of Sheffield

Inference in matrix-variate Gaussian models has major applications for multi-output prediction and joint learning of row and column covariances from matrix-variate data. Here, we discuss an approach for efficient inference in such models that explicitly account for iid observation noise. Computational tractability can be retained by exploiting the Kronecker product between row and column covariance matrices. Using this framework, we show how to generalize the Graphical Lasso in order to learn a sparse inverse covariance between features while accounting for a low-rank confounding covariance between samples. We show practical utility on applications to biology, where we model covariances with more than 100,000 dimensions. We find greater accuracy in recovering biological network structures and are able to better reconstruct the confounders.

Subject Area: Probabilistic Models and Methods

W97 Gaussian Process Training with Input Noise

A. McHutchon ajm257@cam.ac.uk
 C. Rasmussen cer54@cam.ac.uk
 Cambridge University

In standard Gaussian Process regression input locations are assumed to be noise free. We present a simple yet effective GP model for training on input points corrupted by i.i.d. Gaussian noise. To make computations tractable we use a local linear expansion about each input point. This allows the input noise to be recast as output noise proportional to the squared gradient of the GP posterior mean. The input noise variances are inferred from the data as extra hyperparameters. They are trained alongside other hyperparameters by the usual method of maximisation of the marginal likelihood. Training uses an iterative scheme, which alternates between optimising the hyperparameters and calculating the posterior gradient. Analytic predictive moments can then be found for Gaussian distributed test points. We compare our model to others over a range of different regression problems and show that it improves over current methods.

Subject Area: Probabilistic Models and Methods

W98 Iterative Learning for Reliable Crowdsourcing Systems

D. Karger karger@mit.edu
 S. Oh sewoong79@gmail.com
 D. Shah devavrat@mit.edu
 Massachusetts Institute of Technology

Crowdsourcing systems, in which tasks are electronically distributed to numerous “information piece-workers”, have emerged as an effective paradigm for human-powered solving of large scale problems in domains such as image classification, data entry, optical character recognition, recommendation, and proofreading. Because these low-paid workers can be unreliable, nearly all crowdsourcers must devise schemes to increase confidence in their answers, typically by assigning each task multiple times and combining the answers in some way such as majority voting. In this paper, we consider a general model of such crowdsourcing tasks, and pose the problem of minimizing the total price (i.e., number of task assignments) that must be paid to achieve a target overall reliability. We give new algorithms for deciding which tasks to assign to which workers and for inferring correct answers from the workers’ answers. We show that our algorithm significantly outperforms majority voting and, in fact, are asymptotically optimal through comparison to an oracle that knows the reliability of every worker.

Subject Area: Probabilistic Models and Methods

W100 On Learning Discrete Graphical Models using Greedy Methods

A. Jalali alij@mail.utexas.edu
 C. Johnson cjohnson@cs.utexas.edu
 P. Ravikumar pradeepr@cs.utexas.edu
 University of Texas, Austin

In this paper, we address the problem of learning the structure of a pairwise graphical model from samples in a high-dimensional setting. Our first main result studies the sparsistency, or consistency in sparsity pattern recovery, properties of a forward-backward greedy algorithm as applied to general statistical models. As a special case, we then apply this algorithm to learn the structure of a discrete graphical model via neighborhood estimation. As a corollary of our general result, we derive sufficient conditions on the number of samples n , the maximum node-degree d and the problem size p , as well as other conditions on the model parameters, so that the algorithm recovers all the edges with high probability. Our result guarantees graph selection for samples scaling as $n = \Omega(d^2 \log(p))$, in contrast to existing convex-optimization based algorithms that require a sample complexity of $n = \Omega(d^3 \log(p))$. Further, the greedy algorithm only requires a restricted strong convexity condition which is typically milder than irrepresentability assumptions. We corroborate these results using numerical simulations at the end.

Subject Area: Probabilistic Models and Methods

W99 High-Dimensional Graphical Model Selection: Tractable Graph Families and Necessary Conditions

A. Anandkumar animakumar@gmail.com
 UC Irvine
 V. Tan vtan@wisc.edu
 University of Wisconsin-Madison
 A. Willsky willsky@mit.edu
 Massachusetts Institute of Technology

We consider the problem of Ising and Gaussian graphical model selection given n i.i.d. samples from the model. We propose an efficient threshold-based algorithm for structure estimation based known as conditional mutual information test. This simple local algorithm requires only low-order statistics of the data and decides whether two nodes are neighbors in the unknown graph. Under some transparent assumptions, we establish that the proposed algorithm is structurally consistent (or sparsistent) when the number of samples scales as $n = \Omega(J_{\min}^{-3} \log p)$, where p is the number of nodes and J_{\min} is the minimum edge potential. We also prove novel non-asymptotic necessary conditions for graphical model selection.

Subject Area: Probabilistic Models and Methods

W101 Clustered Multi-Task Learning Via Alternating Structure Optimization

J. Zhou jiyay.zhou@asu.edu
 J. Chen Jianhui.Chen@asu.edu
 J. Ye jieping.ye@asu.edu
 Arizona State University

Multi-task learning (MTL) learns multiple related tasks simultaneously to improve generalization performance. Alternating structure optimization (ASO) is a popular MTL method that learns a shared low-dimensional predictive structure on hypothesis spaces from multiple related tasks. It has been applied successfully in many real world applications. As an alternative MTL approach, clustered multi-task learning (CMTL) assumes that multiple tasks follow a clustered structure, i.e., tasks are partitioned into a set of groups where tasks in the same group are similar to each other, and that such a clustered structure is unknown a priori. The objectives in ASO and CMTL differ in how multiple tasks are related. Interestingly, we show in this paper the equivalence relationship between ASO and CMTL, providing significant new insights into ASO and CMTL as well as their inherent relationship. The CMTL formulation is non-convex, and we adopt a convex relaxation to the CMTL formulation. We further establish the equivalence relationship between the proposed convex relaxation of CMTL and an existing convex relaxation of ASO, and show that the proposed convex CMTL formulation is significantly more efficient especially for high-dimensional data. In addition, we present three algorithms for solving the convex CMTL formulation. We report experimental results on benchmark datasets to demonstrate the efficiency of the proposed algorithms.

Subject Area: Probabilistic Models and Methods

1B Real-time social media analysis with TWIMPACT

Mikio Braun
Matthias Jugel
Klaus-Robert Müller
TU Berlin

Social media analysis has attracted quite some interest recently. Typical questions are identifying current trends, rating the impact or influence of users, summarizing what people are saying on some topic or brand, and so on. One challenge is the enormous amount of data you need to process. Current approaches often need to filter the data first and then do the actual analysis after the event. However, ideally you would want to be able to look into the stream of social media updates in real-time as the event unfolds, and not a few days or weeks later when you've processed your data. We will showcase the real-time social media analysis engine developed by TWIMPACT in cooperation with the TU Berlin. You will be able to look at the current trends on Twitter in real-time, in addition to being able to have a look at the last year which we will bring in a pre-analyzed form, such that you can zoom in on the historic events of this year like the revolutions in Egypt, or Libya, and so on

2B AISoy1, a Robot that Perceives, Feels and Makes Decisions

Diego García Sánchez
AISoy Robotics S.L
David Ríos Insua
Rey Juan Carlos University

Guided by the principles of Affective Computing (Picard 1997) and affective decision making (Loewenstein and Werenr, 2003) we have developed AISoy1, a social emotional robot capable of interacting intelligently and emotionally with a user. At its core, there is a decision analytic model, see French and Ríos Insua (2000), which guides the robot's decision making and generates an emotional state as a consequence of evaluating how the environment surrounding it and the actions performed by the users is affecting the achievement of its main goals.

We'll show AISoy1's architecture, placing emphasis on how decision analytic and Bayesian methods support its activities when processing information received through its sensors, forecasting the evolution of its environment, adapting its emotional state and choosing the actions to be made.

3B Real-time Multi-class Segmentation using Depth Cues

Clement Farabet
Nathan Silberman
New York University

We demonstrate a real-time multi-class segmentation system. While significant progress has been made in multi-class segmentation over the last few years, per-pixel label prediction for a given image typically takes on the order of minutes. This renders use of these systems impractical for real-time applications such as robotics, navigation and human-computer interaction. Concurrent with these advances, there has been a renewed interest in the use of depth sensors following the release of the Microsoft Kinect to aid various tasks in computer vision. This work demonstrates a real-time system that provides dense label predictions for a scene given both intensity and depth images. A convolutional network is trained from a newly released depth dataset* of aligned RGB and depth frames which have been annotated with dense pixel-wise labels. Once trained, the convolutional network can be efficiently computed on the neuflow processor**, reducing its computation from a few seconds in software to about 100ms.

4B Contour-Based Large Scale Image Retrieval Platform

Rong Zhou
Liqing Zhang
Shanghai Jiao Tong University

The demonstration application is a contour-based image retrieval prototype system for the large scale database. By simulating hierarchical image processing of human visual system, it is able to retrieve relative images from large scale database according to the contour of the object in an image. In most of situation, users usually want to find images with certain contours without caring its size or location in the image. But most existing algorithms are not able to deal with such invariance problems. The existing retrieval systems usually require that the objects in the query image and in the retrieved image should have the same position and almost the same size. Different from existing algorithms, the system is developed to address these invariance issues, such as the shift-invariance and scale-invariance problem. Furthermore, the system allows two kinds of query image: a hand-drawn sketch or a natural image, which brings the better interactive user experiment and the convenience for those who don't well in drawing. The system could perform an image search from 4 million images quickly, and the response time for each query is about 3~5 seconds.



THURSDAY CONFERENCE

ORAL SESSION

SESSION 11 - 9:30 - 10:40 AM

Session Chair: Mate Lengyel

INVITED TALK: The Neuronal Replicator Hypothesis: Novel Mechanisms for Information Transfer and Search in the Brain

Eörs Szathmáry szathmary.eors@gmail.com
 Chrisantha Fernando ctf20@sussex.ac.uk
 Parmenides Foundation

Francis Crick called Gerald Edelman's "Neural Darwinism" Neural Edelmanism because he could not identify any units of evolution, i.e. entities that multiply and have hereditary variation. Whilst a sufficient condition for the production of adaptation is to satisfy George Price's equation, a condition that most hill-climbers, competitive learning, and reinforcement learning algorithms, and Edelman's neuronal groups do satisfy, a full Darwinian dynamics of populations in which there is information transfer between individuals, i.e. true units of evolution, has additional algorithmic properties, notably the capacity for the evolution of evolvability to structure exploration (proposal) distributions. This capacity for Darwinian populations to learn has inspired us to search for true units of evolution in the brain. To this end we have identified several candidate units with varying degrees of biological plausibility, and have showed how these units can replicate within neuronal tissue at timescales from milliseconds to minutes. Thus we present a theory that is legitimately called Darwinian Neurodynamics. This is joint work with Chrisantha Fernando.

Continuous-Time Regression Models for Longitudinal Networks

Duy Vu dqv100@psu.edu
 David Hunter dhunter@psu.edu
 Penn State University
 Arthur Asuncion asuncion@uci.edu
 Padhraic Smyth
 University of California, Irvine

The development of statistical models for continuous-time longitudinal network data is of increasing interest in machine learning and social science. Leveraging ideas from survival and event history analysis, we introduce a continuous-time regression modeling framework for network event data that can incorporate both time-dependent network statistics and time-varying regression coefficients. We also develop an efficient inference scheme that allows our approach to scale to large networks. On synthetic and real-world data, empirical results demonstrate that the proposed inference approach can accurately estimate the coefficients of the regression model, which is useful for interpreting the evolution of the network; furthermore, the learned model has systematically better predictive performance compared to standard baseline methods.

Subject Area: Probabilistic Models and Methods

ORAL SESSION

SESSION 12 - 10:40 - 11:20 AM

Session Chair: Sasha Rakhlin

Scalable Training of Mixture Models via Coresets

Dan Feldman dannyf@caltech.edu
 Matthew Faulkner mfaulk@caltech.edu
 California Institute of Technology
 Andreas Krause krausea@ethz.ch
 ETH Zurich

How can we train a statistical mixture model on a massive data set? In this paper, we show how to construct coresets for mixtures of Gaussians and natural generalizations. A coreset is a weighted subset of the data, which guarantees that models fitting the coreset will also provide a good fit for the original data set. We show that, perhaps surprisingly, Gaussian mixtures admit coresets of size independent of the size of the data set. More precisely, we prove that a weighted set of $O(dk^3/\epsilon^2)$ data points suffices for computing a $(1 + \epsilon)$ -approximation for the optimal model on the original n data points. Moreover, such coresets can be efficiently constructed in a map-reduce style computation, as well as in a streaming setting. Our results rely on a novel reduction of statistical estimation to problems in computational geometry, as well as new complexity results about mixtures of Gaussians. We empirically evaluate our algorithms on several real data sets, including a density estimation problem in the context of earthquake detection using accelerometers in mobile phones.

Subject Area: Clustering

Fast and Accurate k -means For Large Datasets

Michael Shindler shindler@eecs.oregonstate.edu
 Oregon State University
 Alex Wong alexw@seas.ucla.edu
 University of California, Los Angeles
 Adam Meyerson awm@cs.ucla.edu
 Google, Inc.

Clustering is a popular problem with many applications. We consider the k -means problem in the situation where the data is too large to be stored in main memory and must be accessed sequentially, such as from a disk, and where we must use as little memory as possible. Our algorithm is based on recent theoretical results, with significant improvements to make it practical. Our approach greatly simplifies a recently developed algorithm, both in design and in analysis, and eliminates large constant factors in the approximation guarantee, the memory requirements, and the running time. We then incorporate approximate nearest neighbor search to compute k -means in $O(nk)$ (where n is the number of data points; note that computing the cost, given a solution, takes $\Theta(nk)$ -time). We show that our algorithm compares favorably to existing algorithms - both theoretically and experimentally, thus providing state-of-the-art performance in both theory and practice.

Subject Area: Clustering

ORAL SESSION

SESSION 13 - 12:00 – 1:10 AM



Session Chair: Jonathan Pillow

INVITED TALK: Diversity in an Olfactory Network

Gilles Laurent laurentg@brain.mpg.de
Max-Planck-Institute for Brain Research

Olfaction is a fascinating sensory modality, for it deals with complex stimuli (usually complex mixtures of chemical analytes) and generates singular percepts (coffee, jasmin, bread etc), that remain stable over wide ranges of intensities, noise and input composition. Furthermore, olfactory systems accomplish these pattern recognition feats in only very few layers of processing. I will summarize experimental results on computation in a small olfactory system, focusing on the representation of such stimuli, on circuit dynamics, synchronization, learning rules and mechanisms of homeostatic regulation of activity.

Gilles Laurent 1979-85: Studied Veterinary Medicine and PhD in Neuroethology in Toulouse, France 1985-89: Postdoc (Neuroscience), University of Cambridge, UK 1990-2011: Professor (Biology), California Institute of Technology, Pasadena CA, USA 2009-present: Director, Max Planck Institute for Brain Research, Frankfurt/Main, Germany

Empirical models of spiking in neural populations

Jakob Macke jakob@gatsby.ucl.ac.uk
Lars Buesing lars@gatsby.ucl.ac.uk
Maneesh Sahani maneesh@gatsby.ucl.ac.uk
University College London
John Cunningham jpc74@cam.ac.uk
University of Cambridge
Byron Yu byronyu@cmu.edu
Carnegie Mellon University
Krishna Shenoy shenoy@stanford.edu
Stanford University

Neurons in the neocortex code and compute as part of a locally interconnected population. Large-scale multi-electrode recording makes it possible to access these population processes empirically by fitting statistical models to unaveraged data. What statistical structure best describes the concurrent spiking of cells within a local network? We argue that in the cortex, where firing exhibits extensive correlations in both time and space and where a typical sample of neurons still reflects only a very small fraction of the local population, the most appropriate model captures shared variability by a low-dimensional latent process evolving with smooth dynamics, rather than by putative direct coupling. We test this claim by comparing a latent dynamical model with realistic spiking observations to coupled generalised linear spike-response models (GLMs) using cortical recordings. We find that the latent dynamical approach outperforms the GLM in terms of goodness-of-fit, and reproduces the temporal correlations in the data more accurately. We also compare models whose observations models are either derived from a Gaussian or point-process models, finding that the non-Gaussian model provides slightly better goodness-of-fit and more realistic population spike counts.

ORAL SESSION

SESSION 14 - 1:10 - 1:50 PM



Session Chair: Ronan Collobert

The Manifold Tangent Classifier

Salah Rifai salahmeister@gmail.com
Yann Dauphin dauphiya@iro.umontreal.ca
Pascal Vincent vincentp@iro.umontreal.ca
Yoshua Bengio bengioy@iro.umontreal.ca
Xavier Muller xav.muller@gmail.com
University of Montreal

We combine three important ideas present in previous work for building classifiers: the semi-supervised hypothesis (the input distribution contains information about the classifier), the unsupervised manifold hypothesis (data density concentrates near low-dimensional manifolds), and the manifold hypothesis for classification (different classes correspond to disjoint manifolds separated by low density). We exploit a new algorithm for capturing manifold structure (high-order contractive autoencoders) and we show how it builds a topological atlas of charts, each chart being characterized by the principal singular vectors of the Jacobian of a representation mapping. This representation learning algorithm can be stacked to yield a deep architecture, and we combine it with a domain knowledge-free version of the TangentProp algorithm to encourage the classifier to be insensitive to local directions changes along the manifold. Record-breaking results are obtained and we find that the learned tangent directions are very meaningful.
Subject Area:

Reconstructing Patterns of Information Diffusion from Incomplete Observations

Flavio Chierichetti flavio@chierichetti.name
Jon Kleinberg kleinber@cs.cornell.edu
Cornell University
David Liben-Nowell dlibenno@carleton.edu
Carleton College

Motivated by the spread of on-line information in general and on-line petitions in particular, recent research has raised the following combinatorial estimation problem. There is a tree T that we cannot observe directly (representing the structure along which the information has spread), and certain nodes randomly decide to make their copy of the information public. In the case of a petition, the list of names on each public copy of the petition also reveals a path leading back to the root of the tree. What can we conclude about the properties of the tree we observe from these revealed paths, and can we use the structure of the observed tree to estimate the size of the full unobserved tree T ? Here we provide the first algorithm for this size estimation task, together with provable guarantees on its performance. We also establish structural properties of the observed tree, providing the first rigorous explanation for some of the unusual structural phenomena present in the spread of real chain-letter petitions on the Internet.
Subject Area:

REVIEWERS

Ryan Adams	Charles Blundell	Aaron Courville	David Fleet
Alekh Agarwal	Phil Blunsom	David Cox	Francois Fleuret
Arvind Agarwal	Liefeng Bo	Koby Crammer	David Forsyth
Shivani Agarwal	Botond Bocs	Daniel Cremers	Dean Foster
Deepak Agrawal	Danushka Bollegala	John Cunningham	Emily Fox
Amr Ahmed	Byron Boots	Marco Cuturi	Vojtech Franc
Misha Ahrens	Antoine Bordes	Florence D'Alche-Buc	Jordan Frank
Nir Ailon	Ali Borji	Arnak Dalalyan	Michael Frank
Edoardo Airoidi	Joerg Bornschein	Amit Daniely	Alexander Fraser
Yasemin Altun	Leon Bottou	Dipanjan Das	Bill Freeman
Mauricio Alvarez	Alexandre Bouchard-Ct	Sanjoy Dasgupta	Yoav Freund
David Andrzejewski	Guillaume Bouchard	Hal Daume	Hironori Fujisawa
Andras Antos	Stephane Boucheron	Nathaniel Daw	Kenji Fukumizu
Alfred Anwander	Abdeslam Boularias	Peter Dayan	Jack Gallant
Pablo Arbelaez	Michael Bowling	Nando De Freitas	Kuzman Ganchev
Cedric Archambeau	Jordan Boyd-Graber	Fernando De la Torre	Arvind Ganesh
Andreas Argyriou	Ulf Brefeld	Marc Deisenroth	Surya Ganguli
Raman Arora	Marcus Brubaker	Ofer Dekel	Ravi Ganti
Hiroki Asari	Emma Brunskill	Olivier Delalleau	Phil Garner
Hideki Asoh	Sebastien Bubeck	Jia Deng	Gilles Gasso
Arthur Asuncion	Daphna Buchsbaum	Li Deng	Jan Gasthaus
Chris Atkeson	Michael Buice	Misha Denil	Michael Gastpar
Peter Auer	Razvan Bunescu	Inderjit Dhillon	Eric Gaussier
Joseph Austerweil	Wray Buntine	Tom Diethe	Peter Gehler
Pranjal Awasthi	Chris J. C. Burges	Thomas Dietterich	Andreas Geiger
Francis Bach	David Burkett	Laura Dietz	Andrew Gelman
Drew Bagnell	Keith Bush	Chris Ding	Claudio Gentile
Bing Bai	Lucian Busoni	Francesco Dinuzzo	Sean Gerrish
Doru Balcan	Charles Cadieu	Chuong Do	Sam Gershman
Nina Balcan	Kevin Canini	Eizaburo Doi	Mohammad Ghavamzadeh
Christopher Baldassano	Stephane Canu	Vincent J. Dorie	Martin A. Giese
Luca Baldassarre	Andrea Caponnetto	Finale Doshi	Ran Gilad-Bachrach
Pierre Baldi	Barbara Caputo	Arnaud Doucet	Jennifer Gillenwater
David Balduzzi	Constantine Caramanis	Kenji Doya	Kevin Gimpel
Dana Ballard	Lawrence Carin	Mark Dredze	Inmar Givoni
Aharon Bar Hillel	Francois Caron	Gideon Dror	Amir Globerson
Nick Barnes	Gert Cauwenberghs	Mathias Drton	Vibhav Gogate
Jonathan T. Barron	Gavin Cawley	Jan Drugowitsch	Andrew Goldberg
Peter Bartlett	Lawrence Cayton	John Duchi	Anna Goldenberg
Curzio Basso	Asli Celikyilmaz	Miroslav Dudik	Mark Goldman
Sugato Basu	Alain Celisse	Delbert Dueck	Sally Goldman
Dhruv Batra	Soumen Chakrabarti	Kevin Duh	Pollina Gollant
Peter Battaglia	Tanmoy Chakraborty	Jennifer Dy	Ben Golub
Tim Behrens	America Chambers	Michael Elad	Alon Gonen
Mikhail Belkin	Manmohan Chandraker	Jacob Eisenstein	Noah Goodman
Shai Ben-David	Jonathan Chang	Jason Eisner	Geoff Gordon
Yoshua Bengio	Kai-Wei Chang	Tal El Hay	Dilan Gorur
Philipp Berens	Olivier Chapelle	Tina Eliassi-rad	Stephen Gould
Alexander C. Berg	Gal Chechik	Gal Elidan	Joao V. Graca
Tamara Berg	Jianhui Chen	Charles Elkan	Thore Graepel
Matthias Bethge	Xi Chen	Lloyd Elliott	David Grangier
Alina Beygelzimer	Silvia Chiappa	Dominik M. Endres	Alexander Gray
Shalabh Bhatnagar	Sharat Chikkerur	Barbara Engelhardt	Russ Greiner
Indrajit Bhattacharya	Arthur Choi	Damien Ernst	Arthur Gretton
Sourangshu Bhattacharya	Seungjin Choi	Fang Fang	Remi Gribonval
Chiru Bhattacharyya	Wongun Choi	Clement Farabet	Moritz Grosse-Wentrup
Jinbo Bi	Mario Christoudias	Naomi Feldman	Liu Guangcan
Misha Bilenko	Stephan Clemencon	Gerhard J. Felipe	Shengbo Guo
Aharon Birnbaum	Adam Coates	Rob Fergus	Yuhong Guo
Gilles Blanchard	Shay Cohen	Vittorio Ferrari	Rahul Gupta
Matthew Blaschko	Michael Collins	Sanja Fidler	Michael Gutmann
David Blei	Ronan Collobert	Jenny Finkel	Steven C. Hoi
John Blitzer	Greg Corrado	Jozsef Fiser	Hiroataka Hachiya
Avrim Blum	Timothee Cour	Andrew Fitzgibbon	Ralf Haefner

REVIEWERS

Gholamreza Haffari	Ata Kaban	Victor Lempitsky	Andriy Mnih
Patrick Haffner	Hachem Kadri	Mate Lengyel	Shakir Mohamed
Aria Haghighi	Adam Kalai	Christina Leslie	Claire Monteleoni
Ulrike Hahn	Satyen Kale	Guy Lever	Joris M. Mooij
David Hall	Hetunandan Kamisetty	Roger Levy	Taesup Moon
Keith Hall	Takafumi Kanamori	Fei Fei Li	Tetsuro Morimura
Lauren Hannah	Anitha Kannan	Fuxin Li	Quaid Morris
Zaid Harchaoui	Ashish Kapoor	Hang Li	Indraneel Mukherjee
Bharath Hariharan	Bert Kappen	Lihong Li	Sayan Mukherjee
Stefan Harmeling	Yan Karklin	Ping Li	Remi Munos
Kohei Hatano	Matthias Kaschube	Xiaodong Li	Noburu Murata
James Hays	Hisashi Kashima	Percy Liang	Iain Murray
Elad Hazan	Samuel Kaski	Xuejun Liao	Hiroshi Nakagawa
Tamir Hazan	Koray Kavukcuoglu	Lek-Heng Lim	Hiroyuki Nakahara
Martial Hebert	Yoshinobu Kawahara	Chih-Jen Lin	Shinichi Nakajima
Mohamed Hebiri	Motoaki Kawanabe	Hsuan-Tien Lin	Sahand Negahban
Matthias Hein	Sathiya Keerthi	Yuanqing Lin	Blaine Nelson
Katherine A. Heller	Balazs Kegl	Zhouchen Lin	Bernhard Nessler
Philipp Hennig	Charles Kemp	Jenn Listgarten	Gerhard Neumann
Ralf Herbrich	Kristian Kersting	Ce Liu	Hani Neuvirth-Telem
Mark Herbster	Paul Kidwell	Han Liu	Andrew Ng
Tom Heskes	Seyoung Kim	Tie-Yan Liu	Yizhao Ni
Shohei Hido	Akisato Kimura	Bo Long	Hannes Nickisch
Matt Hoffman	Arto Klami	Phil Long	Alex Niculescu-Mizil
Thomas Hofmann	Robert Kleinberg	Manuel C. Lopes	Juan Carlos Niebles
Derek Hoiem	Marius Kloft	Yonatan Lowenstein	Scott Niekum
Antti Honkela	David A. Knowles	Christopher G. Lucas	Mahesan Niranjan
Cho-Jui Hsieh	Kei Kobayashi	Jorg Lucke	William Stafford Noble
Daniel J. Hsu	Jens Kober	Yi Ma	Sebastian Nowozin
Gang Hua	Kilian Koepsell	Christian Machens	Timothy J. O'Donnell
Chang Huang	Pushmeet Kohli	Hamid R. Maei	Shigeyuki Oba
Junzhou Huang	Mladen Kolar	Sridhar Mahadevan	Guillaume Obozinski
Jonathan Huggins	Vladimir Kolmogorov	Odalric-Ambrym Maillard	Alice Oh
Zakria Hussain	J. Zico Kolter	Julien Mairal	Daisuke Okanohara
Ferenc Huszar	George Konidaris	Subhansu Maji	Aude Oliva
Seth Hutchinson	Terry Koo	Hiroshi Mamitsuka	Bruno Olshausen
Tuyen N. Huynh	Anna Koop	Gideon Mann	Sylvie Ong
Tsuyoshi Ide	Samory Kpotufe	Shie Mannor	Hans Op de Beeck
Christian Igel	Andreas Krause	Oded Margalit	Manfred Opper
Alex Ihler	Balaji Krishnapuram	Ben Marlin	Gergo Orban
Shiro Ikeda	Oliver Kroemer	Andre Martins	Noga Oron
Charles Isbell	Rui Kuang	Winter Mason	Luis Ortiz
Vladimir Itskov	Brian Kulis	Jiri Matas	Sarah Osentoski
Laurent Itti	M. Pawan Kumar	Luke Maurits	Hua Ouyang
Laurent Jacob	Sanjiv Kumar	David McAllester	Jean-Francois Paiement
Nathan Jacobs	Takio Kurita	Jon McAuliffe	John Paisley
Florian Jaeger	Branislav Kveton	Andrew McCallum	Liam Paninski
Jagadeesh Jagarlamudi	James Kwok	John McCoy	Bernardino R. Paredes
SakethaNath Jagarlapudi	Simon Lacoste-Julien	Brian McFee	Il M. Park
Prateek Jain	John Lafferty	Chris Meek	Ronald Parr
Shaili Jain	Brenden Lake	Nishant A. Mehta	Adam Pauls
Vidit Jain	Christoph H. Lampert	Marina Meila	Kristiaan Pelckmans
Viren Jain	Ni Lao	Bartlett Mel	Vianney Perchet
Ali Jalali	Hugo Larochelle	Francisco S. Melo	Florent Perronnin
Dominik Janzing	Edith Law	Roland Memisevic	Jan Peters
Rodolphe Jenatton	Alessandro Lazaric	Aditya Menon	Jonas Peters
Hueihan Jhuang	Svetlana Lazebnik	Ethan Meyers	Slav Petrov
Shuiwang Ji	Nevena Lazic	Tomer Michaeli	David Pfau
Jinzhu Jia	Yann LeCun	Lily Mihalkova	Jean-Pascal Pfister
Xiaoye Jiang	Quoc Le	Krystian Mikolajczyk	Jonathan Pillow
Rong Jin	Guy Lebanon	Brian Milch	Joelle Pineau
Mark Johnson	Guillaume Lecué	Kurt Miller	Nicolas Pinto
Kresimir Josic	Honglak Lee	David Mimno	Gordon Pipa
Anatoli Juditsky	Yuh-Jye Lee	Einat Minkov	John Platt

REVIEWERS

Robert Pless	Katya Scheinberg	Akiko Takeda	Ross Williamson
Russ Poldrack	Warren Schudy	Takashi Takenouchi	Sinead Williamson
Massi Pontil	Dale Schuurmans	Ichiro Takeuchi	Andrew Wilson
Hoifung Poon	Odelia Schwartz	Eiji Takimoto	David Wingate
Doina Precup	Alexander G. Schwing	Partha Talukdar	Ole Winther
Philippe Preux	Clayton Scott	Toshiyuki Tanaka	Frank Wood
Yanjun Qi	Matthias Seeger	Matthew Tayler	John Wright
Yuan Qi	Nicola Segata	Graham Taylor	Steve Wright
Ariadna Quattoni	Yevgeny Seldin	Yee Whye Teh	Qiang Wu
Maxim Raginsky	Thomas Serre	Josh Tenebaum	Wei Wu
Ali Rahimi	Jun Sese	Timothy Teravainen	Lin Xiao
Piyush Rai	Fei Sha	Ambuj Tewari	Eric Xing
Sasha Rakhlin	Amir H. Shabani	Olivier Teytaud	Huan Xu
Alain Rakotomamonjy	Patrick Shafto	Daniel Ting	Makoto Yamada
Liva Ralaivola	Greg Shakhnarovich	Jo-Anne Ting	Yoshihiro Yamanishi
Parikshit Ram	Shai Shalev-Shwartz	Ivan Titov	Shuicheng Yan
Deva Ramanan	Tatyana Sharpee	Michalis Titsias	Keiji Yanai
Marc'Aurelio Ranzato	Xiaotong Shen	Gasper Tkacik	Allen Yang
Garvesh Raskutti	Shirish Shevade	Sinisa Todorovic	Jianchao Yang
Frederic Rattle	Tao Shi	Ryota Tomioka	Ming Yang
Magnus Rattray	Shohei Shimizu	Antonio Torralba	Ming-Hsuan Yang
Pradeep Ravikumar	Lavi Shpigelman	Alexander Toshev	Bangpeng Yao
Mark Reid	Ilya Shpitser	Long Tran	Yuan Yao
Joseph Reisinger	Marco Signoretto	Volker Tresp	Jieping Ye
Xiaofeng Ren	Michael Silver	Bill Triggs	Sun Yi
Lev Reyzin	Ozgur Simsek	Ivor Tsang	Yiming Ying
Sebastian Riedel	Vikas Sindhwani	Ioannis Tsochantaridis	Byron Yu
Philippe Rigollet	Yoram Singer	Koji Tsuda	Kai Yu
Abel Rodriguez	Sameer Singh	Srinivas Turaga	Shipeng Yu
Karl Rohe	Kaushik Sinha	Joseph Turian	Ming Yuan
Lorenzo Rosasco	Noam Slonim	Richard Turner	Yisong Yue
Saharon Rosset	Cristian Sminchisescu	Naonori Ueda	Lihi Zelnik-Manor
Afshin Rostamizadeh	David Smith	Tomer D. Ullman	Haizhang Zhang
Dan Roth	Alexander J. Smola	Lyle Ungar	Kun Zhang
Stefan Roth	Ben Snyder	Raquel Urtasun	Tong Zhang
Volker Roth	Richard Socher	Nicolas Usunier	Dengyong Zhou
Constantin Rothkopf	Imri Sofer	Benjamin Van Roy	Xueyuan Zhou
Juho Rousu	Peter Sollich	Manik Varma	Yan Zhou
Daniel M. Roy	Fritz Sommer	Nuno Vasconcelos	Zhi-Hua Zhou
Ran Rubin	Le Song	Nicolas Vayatis	Jerry Zhu
Benjamin Rubinstein	Stam Sotiropoulos	Andrea Vedaldi	Jun Zhu
Cynthia Rudin	Matthijs Spaan	Jean-Philippe Vert	Shenghuo Zhu
Sasha Rush	Peter Spirtes	Rene Vidal	Andrew Zisserman
Daniil Ryabko	Karthik Sridharan	Sethu Vijayakumar	Larry Zitnick
Sivan Sabato	Bharath Sriperumbudur	S.V.N. Vishwanathan	Onno Zoeter
Ankan Saha	Isabelle Stanton	Ed Vul	Daniel Zoran
Maneesh Sahani	Oliver Stegle	Hanna Wallach	Alon Zweig
Hiroto Saigo	Ingo Steinwart	Chong Wang	
Jun Sakuma	Ian H. Stevenson	Gang Wang	
Ruslan Salakhutdinov	Alan A. Stocker	Liwei Wang	
Mathieu Salzmann	Peter Stone	Yizhou Wang	
Sujay Sanghavi	Hao Su	Zhikun Wang	
Guido Sanguinetti	Amar Subramanya	Ziyu Wang	
Guillermo Sapiro	Masashi Sugiyama	Larry Wasserman	
Ben Sapp	Jian Sun	Kazuho Watanabe	
Sunita Sarawagi	Min Sun	Yusuke Watanabe	
Suchi Saria	Siddharth Suri	Chu Wei	
Simo Sarkka	Ilya Sutskever	Kilian Q. Weinberger	
Issei Sato	Charles Sutton	Yair Weiss	
Kengo Sato	Taiji Suzuki	Tomas Werner	
Cristina Savin	Kevin Swersky	Jason Weston	
Ashutosh Saxena	Umar Syed	Daan Wierstra	
Stefan Schaal	Csaba Szepesvari	Chris Williams	
Tom Schaul	Prasad Tadepalli	Robert C. Williamson	

AUTHOR INDEX

- Abbasi-yadkori, Yasin, 98
 Abbott, Joshua, 80
 Abernethy, Jacob, 76,97
 Absil, Pierre-Antoine, 87
 Adametz, David, 28
 Agarwal, Alekh, 95,98
 Agosta, John-Mark, 69
 Ahmadian, Yashar, 67
 Ahmed, Amr, 10
 Ahn, Hee-Kap, 94
 Ailon, Nir, 87
 Airoldi, Edoardo, 33
 Alamgir, Morteza, 99
 Alexe, Bogdan, 85
 Allahverdyan, Armen, 70
 Anand, Abhishek, 54
 Anandkumar, Animashree, 63,73,102
 Anguita, Davide, 32
 Archambeau, Cedric, 56
 Armagan, Artin, 91
 Asuncion, Arthur, 100
 Auer, Peter, 62
 Austerweil, Joseph, 81
 Ay, Nihat, 95
 Azimi, Javad, 92
 Babacan, S. Derin, 35
 Bach, Francis, 58,59,62,96
 Bagdanov, Andrew, 22
 Bai, Xiang, 53
 Balakrishnan, Sivaraman, 93
 Baldi, Pierre, 86
 Ball, Tonio, 90
 Bar-Joseph, Ziv, 55
 Baracos, Vickie, 55
 Baraniuk, Richard, 23
 Bardenet, Rémi, 26
 Barreto, Andre, 16
 Bartlett, Nicholas, 95
 Barto, Andrew, 17
 Belkin, Mikhail, 91
 Bengio, Yoshua, 26,33,89,93,106
 Berg, Alexander,
 Berg, Tamara, 85
 Bergamo, Alessandro, 22
 Bergstra, James, 26
 Berkes, Pietro, 81
 Bernardino, Alexandre, 27
 Bhand, Maneesh, 21
 Bhargava, Aniruddha, 82
 Bhaskar, Sonia, 91
 Bilmes, Jeff, 65,96
 Blanchard, Gilles, 57,66
 Blei, David, 36
 Blitzer, John, 24
 Blundell, Charles, 41,68
 Bo, Liefeng, 22
 Boahen, Kwabena, 18
 Bogojeska, Jasmina, 55
 Bonilla, Edwin, 95
 Borgwardt, Karsten, 101
 Bornschein, Jorg, 81
 Bouchard-Côté, Alexandre, 41,68
 Boumal, Nicolas, 87
 Boutsidis, Christos, 58
 Bowling, Michael, 48
 Boyles, Levi, 61
 Braun, Mikio, 103
 Brea, Johanni, 52
 Brendel, Wieland, 60
 Briggman, K, 84
 Bubeck, Sebastien, 65
 Buesing, Lars, 83,106
 Bui, Loc, 31
 Buntine, Wray, 95
 Cabral, Ricardo, 27
 Caetano, Tiberio, 26
 Caiafa, Cesar, 90
 Cao, Liangliang, 54
 Carin, Lawrence, 30,69,81
 Carlson, David, 81
 Carpentier, Alexandra, 58,65
 Carreira-Perpinan, Miguel, 60
 Carreira, Joao, 83
 Cemgil, Ali Taylan, 29
 Cesa-Bianchi, Nicolò, 66,74,98
 Chapelle, Olivier, 55
 Chattopadhyay, Rita, 57
 Chaudhuri, Kamalika, 63
 Chellappa, Rama, 27
 Chen, Bo, 50,81
 Chen, Jianhui, 102
 Chen, Justin, 71
 Chen, Ke, 86
 Chen, Kewei, 51
 Chen, Minmin, 24
 Chen, Ning, 36
 Chen, Tsuhan, 84
 Chen, Zhenghao, 91
 Cheng, Weiwei, 69
 Chierichetti, Flavio, 100,106
 Chin, Tat-jun, 38
 Choi, David, 33
 Choi, Jaedeug, 16
 Cichocki, Andrzej,90
 Clyde, Merlise, 91
 Cléménçon, Stéphane, 66
 Coates, Adam, 28
 Collins, Michael, 9
 Collobert, Ronan, 71
 Costa Pereira, Jose, 93
 Costeira, Joao, 27
 Cotter, Andrew, 62
 Courville, Aaron, 33
 Cunningham, John, 19,83,106
 Dai, Dong, 33
 Damianou, Andreas, 68
 Darrell, Trevor, 84
 Daume III, Hal, 28,68
 Dauphin, Yann, 93,106
 Daw, Nathaniel, 50
 Dayan, Peter, 82
 De Campos, Cassio,
 De la Torre, Fernando, 27
 Delaitre, Vincent, 86
 Delalleau, Olivier, 89
 Delgosha, Payam, 87
 Dembczynski, Krzysztof, 69
 Deng, Jia, 83
 Denk, Winfried, 84
 Desjardins, Guillaume, 33
 Dethier, Julie, 18
 Dhillon, Inderjit, 23,32,59,61
 Dhillon, Paramveer, 24
 Dietterich, Thomas, 23,37
 Diggavi, Suhas, 87
 Dikmen, Onur, 60
 Ding, Chris, 25
 Ding, Nan, 101
 Ding, Shilin, 70
 Do, Huyen, 29
 Doppa, Janardhan Rao, 23
 Drineas, Petros, 58
 Dubout, Charles, 25
 Duchi, John, 95
 Dunson, David, 30,69,91
 Dupont, Pierre, 37
 Duvenaud, David, 37
 Eberts, Mona, 99
 Ekanadham, Chaitanya, 67
 El-Yaniv, Ran, 33,87
 Elasaad, Shauki, 18
 Elhadad, Noemie, 95
 Elhamifar, Ehsan, 91
 Eiasmith, Chris, 18
 Elliott, Lloyd, 41,68
 Ermon, Stefano, 67
 Farabet, Clement, 103
 Farahmand, Amir-massoud, 48
 Faulkner, Matthew, 92,105
 Feldman, Dan, 92,105
 Felzenszwalb, Pedro, 85
 Feng, Jean, 71
 Fergus, Rob, 38
 Fern, Alan, 63,92
 Fern, Xiaoli, 23,92
 Fernando, Chrisantha,105
 Ferrari, Vittorio, 85
 Févotte, Cédric, 60
 Fitzgibbon, Andrew, 22
 Fleisher, Adam, 51
 Fletcher, Alyson, 82
 Fleuret, Francois, 25
 Foster, Dean, 24,98
 Fox, Dieter, 22
 Foygel, Rina, 28
 Friesen, Abram, 81
 Frongillo, Rafael, 75,97
 Fukumizu, Kenji, 64,88
 Gabillon, Victor, 65
 Gaertner, Thomas, 82,99
 Gall, Juergen, 53
 Galstyan, Aram, 70
 Gao, Tianshi, 88
 Garber, Dan, 96
 Ge, Xiaoyin, 91
 Gehler, Peter, 52
 Geiger, Andreas, 20
 Gentile, Claudio, 66
 Gerstner, Wulfram, 18
 Ghahramani, Zoubin, 65,80
 Ghaoui, Laurent, 94
 Ghavamzadeh, Mohammad, 17
 Gheshlaghi Azar, Mohammad,
 17
 Ghio, Alessandro, 32
 Ghosh, Soumya, 36
 Gibson, Richard, 64
 Gilbert, Anna, 42
 Girshick, Ross, 85
 Gkioulekas, Ioannis, 94
 Globerson, Amir, 8
 Goernitz, Nico, 26
 Gomes, Carla P., 67
 Gomes, Ryan, 60
 Gong, Yihong, 54
 Goodman, Noah, 35
 Gool, Luc, 53
 Grauman, Kristen, 86
 Grave, Edouard, 59
 Graves, Alex, 90
 Gregor, Karol, 90
 Greiner, Russ, 55
 Gretton, Arthur, 88
 Griffiths, Tom, 80,81
 Grunwald, Peter, 65
 Guestrin, Carlos, 65
 Guillory, Andrew, 65
 Gunawardana, Asela, 38
 Guo, Shengbo, 56,88
 Gutkin, Boris, 49
 Hachiya, Hirotaka, 17,56
 Hamprecht, Fred, 89
 Hansen, Lars,71
 Hayashi, Kohei, 31
 Hazan, Elad, 31,96,97
 He, Xiaofoei, 52
 Hein, Matthias, 29,90
 Heller, Katherine, 80
 Helmstaedter, Moritz, 84
 Hennig, Philipp, 17
 Hernández-Lobato, Jose
 Miguel, 37
 Hernández-lobato, Daniel, 37
 Hero, Alfred, 24
 Heskes, Tom, 38
 Hewlett, Daniel, 49
 Hglund, Mattias,37
 Hirayama, Jun-ichiro, 29
 Horwitz, Greg, 52
 Hsieh, Cho-Jui, 61
 Hsu, Daniel, 63,98
 Hsu, David, 16
 Huang, Bert, 59
 Huang, Eric, 89
 Huang, Heng, 25
 Huang, Shuai, 51
 Huang, Thomas, 38,54
 Huang, Tzu-Kuo, 38
 Hullermeier, Eyke, 69
 Hunter, David, 100,105
 Hwang, Sung Ju, 86
 Hwang, Yoonho, 94
 Hyvarinen, Aapo, 29
 Insua, David Rios,103
 Ion, Adrian, 83
 Iouditski, Anatoli, 31
 Isola, Phillip, 81
 Jaakkola, Tommi,8
 Jacob, Laurent, 64
 Jacobs, Robert, 80
 Jain, Prateek, 23
 Jain, Viren, 29,84
 Jalali, Ali, 102
 Jamieson, Kevin, 24
 Janzing, Dominik, 38
 Jebara, Tony, 59,88
 Jegelka, Stefanie, 96
 Jern, Alan, 50
 Ji, Qiang, 51
 Jia, Yangqing, 84
 Jiang, Jiarong, 68
 Joachims, Thorsten, 54
 Johari, Ramesh, 31
 Johnson, Christopher, 102
 Jordan, Michael, 94,100
 Jugel, Matthias, 103
 Kahles, Andre, 26
 Kakade, Sham, 30,63,98
 Kalai, Adam, 30
 Kale, Satyen, 31
 Kalousis, Alexandros, 29
 Kamangar, Farhad, 25
 Kanade, Varun, 30
 Kanamori, Takafumi, 56
 Kappen, Hilbert, 17
 Kapralov, Michael, 64
 Kar, Purushottam, 29
 Karger, David, 75,102
 Karklin, Yan, 82

AUTHOR INDEX

- Karpenko, Alexandre, 92
 Kashima, Hisashi, 31
 Kawahara, Yoshinobu, 61
 Kayala, Matthew, 86
 Kégl, Balázs, 26
 Keil, Matthias, 19
 Kemp, Charles, 50
 Keramati, Mehdi, 49
 Keshet, Joseph, 42,58
 Khan, Fahad, 22
 Khan, Faisal, 80
 Khan, Omar, 69
 Kiefel, Martin, 52
 Kilinc Karzan, Fatma, 31
 Kim, Dae Il, 68
 Kim, Kee-Eung, 16
 Kim, Sungwoong, 21
 Kleinberg, Jon, 100,106
 Kloft, Marius, 57
 Knowles, David, 36
 Kobayashi, Ryota, 20
 Koerding, Konrad, 19
 Koh, Pang Wei, 91
 Kohli, Pushmeet, 21
 Kokkinos, Iasonas, 85
 Kolar, Mladen, 93
 Koller, Daphne, 88
 Kolter, J. Zico, 48
 Koltun, Vladlen, 18,74,83
 Konidaris, George, 50
 König, Arnd, 55
 Koolen, Wouter, 31,65
 Koppula, Hema, 54
 Korattikara, Anoop, 61
 Koren, Tomer, 97
 Kotlowski, Wojciech, 31
 Kpotufe, Samory, 75,97
 Krause, Andreas, 60, 69,92,105
 Krishnamurthy, Akshay, 93
 Kroemer, Oliver, 40,48
 Krähenbühl, Philipp, 74,83
 Kulkarni, Girish, 85
 Kumar, Abhishek, 28
 Kunapuli, Gautam, 25
 Kurtek, Sebastian, 23
 Ladicky, Lubor, 57
 Lagergren, Jens, 37
 Lampert, Christoph, 58
 Lanckriet, Gert, 64
 Lanctot, Marc, 48
 Lansky, Petr, 20
 Laptev, Ivan, 53,86
 Larsen, Jakob Eg, 71
 Larsson, Martin, 96
 Latecki, Longin Jan, 53
 Latham, Peter, 20
 Laughlin, Simon, 18
 Laurent, Gilles, 106
 Laviolette, Francois, 62
 Lawrence, Neil, 68,101
 Lawrence, Richard, 88
 Lazaric, Alessandro, 48,65
 Lázaro-Gredilla, Miguel, 34
 Le Roux, Nicolas, 96
 LeCun, Yann, 90
 Le, Hai-Son, 55
 Le, Quoc V., 71,92
 Lee, Gyemin, 66
 Lei, Jing, 100
 Leibo, Joel, 20
 Lepitsky, Victor, 21
 Lengyel, Mate, 19,20,82
 Levine, Sergey, 18
 Li, Congcong, 84
 Li, Fei Fei, 22,83
 Li, Jing, 51
 Li, Lihong, 55
 Li, Ping, 55
 Li, Zhen, 54
 Liben-Nowell, David, 100,106
 Lim, Joseph, 21
 Lim, Zhan Wei, 16
 Lima, Pedro, 17
 Lin, Binbin, 52
 Lin, Hsiu-Chin, 55
 Lin, Hui, 55,96
 Lin, Zhouchen, 96
 Link, Benjamin, 54
 Lippert, Christoph, 101
 Liu, Jun, 89,90
 Liu, Risheng, 96
 Liu, Tom, 93
 Liu, Wenyu, 53
 Liu, Yan, 88
 Liu, Yi-Kai, 63
 Lizotte, Dan, 49
 Loh, Po-Ling, 43,65
 Long, Phil, 63,97
 Lopes, Miles, 64
 Lou, Xinghua, 89
 Low, Tiffany, 71
 Lozano, Aurelie, 27
 Lu, Zhaosong, 61
 Lu, Zhengdong, 60
 Lucas, Christopher, 50
 Lucke, Jorg, 81
 Lyu, Siwei, 30
 Machens, Christian, 60
 Macke, Jakob, 20,83,106
 Mackey, Lester, 94
 Maclin, Richard, 25
 Magdon-Ismael, Malik, 58
 Mahadevan, Vijay, 93
 Mahoney, Michael, 30
 Maillard, Odalric-Ambrym, 49,58
 Mandic, Danilo, 90
 Manning, Christopher, 89
 Mannor, Shie, 31,98
 Matthews, Iain, 38
 Maua, Denis, 70
 McCallum, Andrew, 101
 McHutchon, Andrew, 101
 Mcallester, David, 42,58,85
 Meek, Christopher, 38
 Mehta, Neville, 63
 Meila, Marina, 94
 Meir, Ron, 37
 Mensi, Skander, 18
 Messias, João, 17,99
 Meyerson, Adam, 92,105
 Miller, Ken, 67
 Minka, Tom, 36
 Missura, Olana, 99
 Mohajer, Soheil, 87
 Montufar, Guido, 95
 Mooij, Joris, 38,101
 Moore, Joshua, 55
 Morioka, Nobuyuki, 54
 Morrison, Clayton, 49
 Moulines, Eric, 62
 Mozer, Michael, 54
 Mudur, Ritvik, 21
 Muller, Klaus-Robert, 103
 Muller, Xavier, 93
 Munos, Remi, 17,32,40,49,58,65
 Murillo Fuentes, Juan José, 34
 Murray, Iain, 20
 Mutch, Jim, 20
 Mutlu, Bilge, 80
 Nakajima, Shinichi, 35
 Nasrabadi, Nasser, 59
 Naud, Richard, 18
 Navalpakkam, Vidhya, 50
 Nemirovski, Arkadi, 31
 Newman, David, 95
 Ney, Hermann, 62
 Ng, Andrew, 21,28,71,89,91,92
 Ngiam, Jiquan, 91,92
 Nguyen, Nam, 59
 Nickisch, Hannes, 37
 Nie, Feiping, 25
 Niekum, Scott, 17,50
 Ning, Huazhong, 54
 Niu, Feng, 97
 Niu, Gang, 17
 Niven, Jeremy, 18
 Nowak, Rob, 24
 Nowozin, Sebastian, 21
 Nuyujukian, Paul, 18
 Obozinski, Guillaume, 59
 Oh, Sewoong, 102
 Oliva, Aude, 81
 Olmos, Pablo, 34
 Oneto, Luca, 32
 Ong, Cheng Soon, 69
 Opper, Manfred, 37,70
 Orbanz, Peter, 9
 Ordonez, Vicente, 85
 Orhan, Emin, 80
 Orr, Walker, 23
 Ortner, Ronald, 62
 Pacer, Michael, 80
 Pajarinen, Joni, 48
 Pal, David, 98
 Panchanathan, Sethuraman, 57
 Panigrahy, Rina, 64
 Paninski, Liam, 66
 Parikh, Ankur, 26
 Parikh, Devi, 81
 Park, Il Memming, 99
 Park, Mijung, 52
 Pashler, Harold, 54
 Peltonen, Jaakko, 48
 Pennin, Jeffrey, 89
 Perez-Cruz, Fernando, 34
 Perona, Pietro, 50,60
 Perotte, Adler, 95
 Perrault-Joncas, Dominique, 94
 Perry, Patrick, 30
 Peters, Jan, 48
 Petersen, Michael, 71
 Petrescu, Viviana, 85
 Petreska, Biljana, 19
 Petterson, James, 26
 Pfister, Jean-Pascal, 52
 Pham, Trung, 38
 Pidan, Dmitry, 87
 Pillow, Jonathan, 8,52,99
 Pineau, Joelle, 16
 Pitkow, Xaq, 67
 Poczcos, Barnabas, 27
 Poggio, Tomaso, 20
 Polyak, Boris, 31
 Popovic, Zoran, 18
 Poupart, Pascal, 69
 Precup, Doina, 16
 Qi, Yuan (Alan), 67,100
 Quigley, Morgan, 71
 Raetsch, Gunnar, 26
 Raginsky, Maxim, 99
 Rahimi, Ali, 34
 Rahnema Rad, Kamiar, 66
 Rai, Piyush, 28,68
 Rakhlin, Alexander, 32,98,99
 Ramadge, Peter, 56
 Ramanan, Deva, 53,61
 Rangan, Sundeep, 82
 Rao, Vinayak, 36
 Rasi, Marc, 71
 Rasmussen, Carl Edward, 37,101
 Rauh, Johannes, 95
 Ravikumar, Pradeep, 32,59,61,102
 Raykar, Vikas, 36
 Re, Christopher, 97
 Recht, Benjamin, 97
 Reichert, David, 51
 Reid, Mark, 63
 Reiman, Eric, 51
 Ren, Lu, 69
 Ren, Xiaofeng, 22
 Restelli, Marcello, 48
 Rezende, Danilo, 82
 Ridella, Sandro, 32
 Rifai, Salah, 93,106
 Rinaldo, Alessandro, 93
 Romo, Ranulfo, 60
 Rooij, Steven, 65
 Roth, Volker, 28
 Rother, Carsten, 52
 Roy, Dan, 95
 Rutter, Andreas, 9
 Rush, Alexander, 9
 Ryabko, Daniil, 49
 Ryu, Stephen, 19
 Saberian, Mohammad, 25
 Sabharwal, Ashish, 67
 Saeedi, Ardavan, 68
 Safa, Issam, 91
 Saffari, Amir, 57
 Sahani, Maneesh, 19,34,83,106
 Salakhutdinov, Ruslan, 21,28,100
 Salamanca, Luis, 34
 Salman, Ahmad, 86
 Sanchez, Diego Garcia, 103
 Sanguinetti, Guido, 70
 Sankaranarayanan, Aswin, 23
 Santhanam, Gopal, 19
 Satheesh, Sanjeev, 83
 Satoh, Shin'ichi, 54
 Saul, Lawrence, 93
 Savin, Cristina, 82
 Saxe, Andrew, 21
 Saxena, Ashutosh, 54,84
 Chalk, Gerwin, 51
 Schmidt, Mark, 96
 Schneider, Jeff, 27,38
 Schulze-bonhage, Andreas, 90
 Schölkopf, Bernhard, 38,52,73
 Schwartz, Marc-Olivier, 71
 Scott, Clay, 66
 Seldin, Yevgeny, 62
 Selman, Bart, 67
 Sengupta, Biswa, 18
 Senn, Walter, 52
 Series, Peggy, 51
 Servedio, Rocco, 63,97

AUTHOR INDEX

- Setzer, Simon, 29
 Seung, H. Sebastian, 84
 Sha, Fei, 86
 Shah, Devavrat, 102
 Shalev-Shwartz, Shai, 57
 Shamir, Ohad, 28,30,62,98
 Shashua, Amnon, 57
 Shavlik, Jude, 25
 Shaw, Blake, 59
 Shawe-Taylor, John, 62
 Sheikh, Abdul Saboor, 81
 Sheldon, Daniel, 37
 Shelton, Jacquelyn, 81
 Shenoy, Krishna, 18,19,83,106
 Shindler, Michael, 92,105
 Shinomoto, Shigeru, 20
 Shivaswamy, Pannagadatta, 88
 Shrivastava, Anshumali, 23,55
 Shroff, Nitesh, 27
 Si, Luo, 88
 Silbert, Nathan, 103
 Sigal, Leonid, 38
 Silva, Ricardo, 34
 Simon, Dylan, 50
 Simoncelli, Eero, 67,82
 Simsekli, Umut, 29
 Sindhvani, Vikas, 17
 Singh, Aarti, 93
 Siskind, Jeffrey, 35
 Sivic, Josef, 86
 Sjlund, Erik37
 Slawski, Martin, 90
 Slivkins, Aleksandrs, 98
 Sminchisescu, Cristian, 83
 Smola, Alexander, 10
 Smyth, Padhraic, 100,105
 Soatto, Stefano, 84
 Socher, Richard, 89
 Song, Le, 26,63,88
 Sonnenburg, Soeren, 26
 Sontag, David, 95
 Sorower, Mohammad Shahed,
 23
 Spaan, Matthijs, 17
 Srebro, Nati, 28,62,65,97
 Sricharan, Kumar, 24
 Sridharan, Karthik, 32,62,65
 Sriperumbudur, Bharath, 64
 Srivastava, Anuj, 23
 Stegle, Oliver, 101
 Steinwart, Ingo, 99
 Stemmler, Martin, 18
 Stevenson, Ian, 19
 Stewart, Terrence, 18
 Stimberg, Florian, 70
 Stopczynski, Arkadiusz,71
 Storkey, Amos, 51
 Stahlhut, Carsten,71
 Stuhlmüller, Andreas, 35
 Su, Zhixun, 96
 Sudderth, Erik, 36,68
 Sugiyama, Masashi,
 17,35,56,61
 Sun, Lee, 16
 Sun, Liang, 89
 Sun, Qian, 57
 Suresh, Bipin, 21
 Susemihl, Alex, 37
 Sustik, Matyas, 61
 Suter, David, 38
 Sutton, Charles, 35
 Sutton, Richard,40
 Suzuki, Taiji, 31,32,56
 Szafron, Duane, 64
 Szathmary, Eors, 105
 Szepesvari, Csaba, 98
 Szlam, Arthur, 90
 Tadepalli, Prasad, 23,63
 Takeuchi, Ichiro, 61
 Talwalkar, Ameet, 94
 Tan, Vincent, 102
 Taylor, Graham, 38
 Teh, Yee Whye, 9,36,68
 Telgarsky, Matus, 88
 Tenenbaum, Josh, 100
 Tewari, Ambuj, 23,32,59,65
 Thomas, Philip, 16,50
 Tishby, Naftali, 10
 Titsias, Michalis, 34,68
 Tofigh, Ali, 37
 Tomioka, Ryota, 31
 Torr, Philip, 57
 Torralba, Antonio, 21,81,100
 Torresani, Lorenzo, 22
 Tran, Trac, 59
 Tranchina, Daniel, 67
 Tschopp, Dominique, 87
 Tsubo, Yasuhiro, 20
 Turaga, Pavan, 27
 Turaga, Srinivas, 84
 Turner, Richard, 34
 Ugander, Johan, 96
 Ujfalussy, Balazs, 19
 Ungar, Lyle, 24
 Ungureanu, Andrei, 36
 Urtasun, Raquel, 20,53
 Van den Broeck, Guy, 70
 Van de Weijer, Joost,22
 Van Erven, Tim, 65
 Vanrell, Maria, 22
 Varshney, Lav, 82
 Vasconcelos, Nuno, 25,93
 Vedaldi, Andrea, 21
 Veness, Joel, 48
 Vernet, Elodie, 63
 Vidal, Rene, 91
 Vincent, Pascal, 93,106
 Vishwanathan, S.V.N., 101
 Vitale, Fabio, 66
 Von Luxburg, Ulrike, 99
 Vondrick, Carl, 53
 Vu, Duy, 100,105
 Waegeman, Willem, 69
 Wagner, Paul, 16
 Wahba, Grace, 70
 Wainwright, Martin, 56,64
 Walsh, Thomas, 49
 Wang, Hua, 25
 Wang, Jun, 29
 Wang, Weiran, 60
 Wang, Xinggang, 53
 Wang, Yingjian, 69
 Wang, Yusu, 91
 Wang, Zuoguan, 51
 Warmuth, Manfred, 31
 Washio, Takashi, 61
 Watanabe, Yusuke, 35
 Waters, Andrew, 23
 Wauthier, Fabian, 100
 Weinberger, Kilian, 24
 Welinder, Peter, 60
 Welling, Max, 61
 Wexler, Yonatan, 57
 Wick, Michael, 101
 Widmer, Christian, 26
 Wiener, Yair, 33
 Wierstra, Daan, 82
 Wiesler, Simon, 62
 Williamson, Robert, 63
 Willsky, Alan, 102
 Wingate, David, 35
 Wipf, David, 59
 Wnuk, Kamil, 84
 Wojek, Christian, 20
 Wolfe, Patrick, 33
 Wong, Alex, 92,102
 Wong, Chi Wah, 93
 Wood, Frank, 95
 Woznica, Adam, 29
 Wright, Stephen, 97
 Wu, Teresa, 51
 Wu, Wei, 23
 Xiang, Zhen James, 56
 Xing, Eric, 22,26,36
 Xiong, Liang, 27
 Xu, Hao, 56
 Xu, Min, 93
 Xu, Puyang, 38
 Yamada, Makoto, 56
 Yan, Feng, 67
 Yang, Liu, 28
 Yang, Shulin, 34
 Yang, Xingwei, 53
 Yao, Angela, 53
 Ye, Jieping, 51,57,89,90,102
 Yilmaz, Kenan, 29
 Yoo, Chang D., 21
 Yu, Byron, 19,83,106
 Yu, Chun-Nam, 55
 Yu, Jin, 38
 Yu, Shipeng, 36
 Yuan, Lei, 90
 Yue, Yisong, 65
 Zappella, Giovanni, 66
 Zeiler, Matthew, 38
 Zeller, Georg, 26
 Zhang, Chiyuan, 52
 Zhang, Dan, 88
 Zhang, Jian, 88
 Zhang, Liqing, 90
 Zhang, Lumin, 52
 Zhang, Tong, 33,54,63
 Zhang, Xianxing, 30
 Zhang, Yichuan, 35
 Zhang, Yong, 61
 Zhang, Youwei, 94
 Zhang, Ziming, 57
 Zhao, Bin, 22
 Zhao, Qibin, 90
 Zhao, Tingting, 17
 Zhao, Yibiao, 53
 Zhou, Jiayu, 102
 Zhou, Rong, 103
 Zhu, Jun, 36,53
 Zhu, Song-chun, 53
 Zhu, Xiaojin (Jerry), 70,80
 Zickler, Todd, 94
 Zisserman, Andrew, 21
 Zoeter, Onno, 56
 Zou, Will, 71

NEXT CONFERENCE



2012 - 2014
LAKE TAHOE
NEVADA