NIPS 2008

NEURAL
INFORMATION
PROCESSING
SYSTEMS
CONFERENCE

Abstracts of Papers

TUTORIALS
December 8, 2008
Hyatt Regency
Vancouver, BC, Canada

CONFERENCE SESSIONS
December 8-11, 2008
Hyatt Regency
Vancouver, BC, Canada

MINI SYMPOSIA
December 11, 2008
Hyatt Regency
Vancouver, BC, Canada

WORKSHOP
December 12-13, 2008
The Westin Resort & Spa
The Hilton Whistler Resort & Spa
Whistler, BC, Canada

Neural Information
Processing Systems
Foundation

NIPS 2008

Abstracts of Papers

TUTORIALS
December 8, 2008
Hyatt Regency
Vancouver, BC, Canada

CONFERENCE SESSIONS
December 8-11, 2008
Hyatt Regency
Vancouver, BC, Canada

MINI SYMPOSIA
December 11, 2008
Hyatt Regency
Vancouver, BC, Canada

WORKSHOP
December 12-13, 2008
The Westin Resort & Spa
The Hilton Whistler Resort & Spa
Whistler, BC, Canada

Sponsored by the Neural Information Processing Systems Foundation, Inc.

There are 5 invited talks and 250 accepted papers selected from a total of 1016 submission considered by the program committee. Because the program stresses interdisciplinary interactions, there are no parallel sessions. Papers presented at the conference will appear in "Advances in Neural Information Processing Systems 21", edited by Daphne Koller, Yoshua Bengio, Dale Schuurmans, and Léon Bottou.

Neural Information Processing Systems Foundation

Contents

Authors!

Please, see the last two pages of this brochure.

Organizing Committee

General Chair	Daphne Koller, Stanford University
Program Co-Chairs	Dale Schuurmans, University of Alberta
	Yoshua Bengio, University of Montreal
Tutorial Chair	Geoff Gordon, Carnegie Mellon University
Workshop Co-Chairs	Maneesh Sahani, London College University
	Alex Smola, Yahoo! Research
Demonstration Chair	Ralf Herbrich, Microsoft Research
Publications Chair	Leon Bottou, NEC Labs America
Electronic Proceedings Co-Chairs	Aron Culotta, Southeastern Louisiana University
	Andrew McCallum, University of Massachusetts, Amherst
Publicity Chair	Antonio Torralba, MIT
Volunteers Chair	Yasemin Altun, Max Planck Institute

Program Committee

Program Co-Chairs	Dale Schuurmans, University of Alberta
	Yoshua Bengio, University of Montreal
Program Committee Area Chairs	Jean-Yves Audibert, Ecole Nationale des Ponts et Chaussées
	Francis Bach, INRIA & Ecole Normale Supérieure
	Kristin Bennett, Rensselaer Polytechnic Institute
	Michael Bowling, University of Alberta
	Aaron Courville, Université de Montréal
	Koby Crammer, University of Pennsylvania
	Sanjoy Dasgupta, University of California, San Diego
	Nathaniel Daw, New York University
	Eleazar Eskin, University of California, Los Angeles
	David Fleet, University of Toronto
	Paolo Frasconi, Università di Firenze
	Arthur Gretton, Max Planck Institute
	Tony Jebara, Columbia University
	Chris Manning, Stanford University
	Ron Meir, Technion
	Noboru Murata, Waseda University
	Erkki Oja, Helsinki University of Technology
	Doina Precup, McGill University
	Stefan Schaal, University of Southern California
	Fei Sha, University of Southern California
	Alan Stocker, New York University
	Ingo Steinwart, Los Alamos National Laboratory
	Erik Sudderth, University of California, Berkeley
	Yee-Whye Teh, University College London
	Antonio Torralba, Massachusetts Institute of Technology
	Larry Wasserman, Carnegie Mellon University
	Max Welling, University of California, Irvine

NIPS Foundation Officers and Board Members

Sponsors

NIPS gratefully acknowledges the generosity of those individuals and organizations who have provided financial support for the NIPS 2008 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
Google
Pascal
Yahoo
Intel
Boeing
IBM
Willow Garage
D.E. Shaw
Numenta
Toyota Research
Springer

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 acknowlege the exceptional work of:

Nelson Loyola, Workflow Master
Lee Campbell
Chris Hiestand
Sheri Leone
Mary Ellen Perry

Awards

Outstanding Student Paper Awards

Influence of graph construction on graph-based clustering measures
Markus Maier, Ulrike von Luxburg and Matthias Hein

Inferring rankings under constrained sensing
Srikanth Jagabathula and Devavrat Shah

Student Paper Honorable Mentions

Stress, noradrenaline, and realistic prediction of mouse behaviour using reinforcement learning
Gediminas Luksys, Carmen Sandi and Wulfram Gerstner

On Computational Power and the Order-Chaos Phase Transition in Reservoir Computing
Benjamin Schrauwen, Lars Buesing and Robert Legenstein

Program Highlights

Monday, December 8th

Tuesday, December 9th

Wednesday, December 10th

Thursday, December 11th

7:30am–9:00am	Breakfast: *Perspectives Level*

8:00am–9:00am	Breakfast: also in *Regency A — Conference Level*
8:00am–1:00pm	Registration Desk Open
8:00am–1:00pm	Internet Access Room Open
8:30am–10:10am	**Oral Session:** Nonparametric Processes, Scene Processing and Image Statistics 46
10:10am–10:40am	Break
10:40am–12:00am	**Oral Session:** Attention and Mind 46
	Invited Talk: *Theory of Mind with fMRI* Rebecca Saxe
12:00pm	Close of Main Conference
1:30pm–2:50pm	Mini Symposia
2:50pm–3:10pm	Break
3:10pm–4:30pm	Mini Symposia
2:00pm–6:00pm	Buses depart for Workshops

Schedule of Presentations

Monday, December 8th

9:30am–11:30am **Tutorial Session 1**

Compressed Sensing
Emmanuel Candes
Location: Regency E/F

Prequential Analysis
Philip Dawid
Location: Regency D

1:00pm–3:00pm **Tutorial Session 2**

Agnostic Learning: Algorithms and Theory
Adam Kalai
Location: Regency E/F

The Neurobiology of Decision Making
Michael Shadlen
Location: Regency D

3:30pm–5:30pm **Tutorial Session 3**

Statistical Models of Visual Images
Eero Simoncelli
Location: Regency E/F

The Smoothed Analysis of Algorithms
Daniel Spielman
Location: Regency D

6:30pm–7:45pm **Buffet, Opening Remarks and Awards**

Dinner buffet open until 9pm; dessert buffet until 10pm.

7:45pm–8:25pm **Oral Session:**
Clustering
Chair: Fei Sha

8:25pm–8:45pm Spotlights

Evaluating probabilities under high-dimensional latent variable models, IAIN MURRAY and RUSLAN SALAKHUTDINOV, University of Toronto 76

8:45pm–12:00am **Poster Session**

M1 ***Logistic Normal Priors for Unsupervised Probabilistic Grammar Induction,*** SHAY B. COHEN, KEVIN GIMPEL and NOAH A. SMITH, Carnegie Mellon University 58

M2 ***Regularized Learning with Networks of Features,*** TED SANDLER, University of Pennsylvania, JOHN BLITZER, University of California, Berkeley, PARTHA PRATIM TALUKDAR and LYLE H. UNGAR, University of Pennsylvania .. 58

M3 ***Look Ma, No Hands: Analyzing the Monotonic Feature Abstraction for Text Classification,*** DOUG DOWNEY, Northwestern University, and OREN ETZIONI, University of Washington 58

M4 ***Syntactic Topic Models,*** JORDAN L BOYD-GRABER and DAVID BLEI, Princeton University 59

M5 ***Semi-supervised Learning with Weakly-Related Unlabeled Data : Towards Better Text Categorization,*** LIU YANG, Carnegie Mellon University, RONG JIN, Michigan State University, and RAHUL SUKTHANKAR, Intel Research and Carnegie Mellon University 59

M6 ***Generative and Discriminative Learning with Unknown Labeling Bias,*** MIROSLAV DUDÍK, Carnegie Mellon University, and STEVEN PHILLIPS, AT&T Labs-Research 60

M7 ***Large Margin Taxonomy Embedding for Document Categorization,*** KILIAN Q WEINBERGER and OLIVIER CHAPELLE, Yahoo! Research 60

M8 ***Learning Taxonomies by Dependence Maximization,*** MATTHEW BLASCHKO and ARTHUR GRETTON, Max Planck Institute for Biological Cybernetics 61

M9 ***Structured ranking learning using cumulative distribution networks,*** JIM C. HUANG and BRENDAN J. FREY, University of Toronto 61

M10 ***Overlaying classifiers: a practical approach for optimal ranking,*** STÉPHAN J.M. CLÉMENÇON, Institut Telecom, and NICOLAS VAYATIS, CMLA - CNRS & UniverSud 62

M11 ***Predictive Indexing for Fast Search,*** SHARAD GOEL, JOHN LANGFORD and ALEXANDER L STREHL, Yahoo! Research 62

M12 ***Integrating Locally Learned Causal Structures with Overlapping Variables,*** ROBERT E. TILLMAN, Carnegie Mellon University, DAVID DANKS and CLARK GLYMOUR, Carnegie Mellon University and Institute for Human and Machine Cognition 62

M13 ***Finding Latent Causes in Causal Networks: an Efficient Approach Based on Markov Blankets,*** JEAN-PHILIPPE PELLET and ANDRÉ ELISSEEFF, IBM Zurich Research Lab 63

Tuesday, December 9th

8:30am–9:50am **Oral Session:**
Sensorimotor Control
Chair: Arthur Gretton

8:30am **Invited Talk:** ***Computations in Human Sensorimotor Control,*** DANIEL WOLPERT, University of Cambridge 97

9:30am ***Kernel-ARMA for Hand Tracking and Brain-Machine interfacing During 3D Motor Control,*** LAVI SHPIGELMAN, HAGAI LALAZAR and EILON VAADIA, Hebrew University 97

9:50am–10:00am **Spotlights**

Unifying the Sensory and Motor Components of Sensorimotor Adaptation, ADRIAN HAITH, University of Edinburgh, CARL P.T. JACKSON, R CHRIS MIALL, University of Birmingham, and SETHU VIJAYAKUMAR, University of Edinburgh 110

Effects of Stimulus Type and of Error-Correcting Code Design on BCI Speller Performance, JEREMY HILL, MPI for Biological Cybernetics, JASON FARQUHAR, Donders Institute for Brain, Cognition and Behaviour, SUZANNA MARTENS, FELIX BIESSMANN and BERNHARD SCHÖLKOPF, MPI for Biological Cybernetics 135

Bayesian Experimental Design of Magnetic Resonance Imaging Sequences, MATTHIAS W SEEGER, Saarland University, HANNES NICKISCH, ROLF POHMANN and BERNHARD SCHÖLKOPF, MPI for Biological Cybernetics 128

How memory biases affect information transmission: A rational analysis of serial reproduction, JING XU and THOMAS L GRIFFITHS, University of California, Berkeley 138

A rational model of preference learning and choice prediction by children, CHRIS LUCAS, THOMAS L GRIFFITHS, University of California, Berkeley, FEI XU, University of British Columbia, and CHRISTINE FAWCETT, University of California, Berkeley 139

Optimal Response Initiation: Why Recent Experience Matters, MATT JONES, MICHAEL MOZER and SACHIKO KINOSHITA, University of Colorado 137

Analyzing human feature learning as nonparametric Bayesian inference, JOSEPH AUSTERWEIL and THOMAS L GRIFFITHS, University of California, Berkeley 135

A Massively Parallel Digital Learning Processor, HANS PETER GRAF, SRIHARI CADAMBI, IGOR DURDANOVIC, VENKATA JAKKULA, MURUGAN SANKARADASS, ERIC COSATTO and SRIMAT CHAKRADHAR, NEC Laboratories America 134

Mind the Duality Gap: Logarithmic regret algorithms for online optimization, Shai Shalev-Shwartz and Sham M Kakade, Toyota Technological Institute . . . 123

An Online Algorithm for Maximizing Submodular Functions, Matthew Streeter, Google, Inc., and Daniel Golovin, Carnegie Mellon University . . . 112

2:00pm–3:20pm **Oral Session:**
Combinatorial Approximation
Chair: Erik Sudderth

2:00pm **Invited Talk:** ***Online Stochastic Combinatorial Optimization,*** Pascal van Hentenryck, Brown University . . . 102

3:00pm ***MAS: a multiplicative approximation scheme for probabilistic inference,*** Ydo Wexler and Christopher Meek, Microsoft Research . . . 102

3:20pm–3:30pm **Spotlights**

Clusters and Coarse Partitions in LP Relaxations, David Sontag, Massachusetts Institute of Technology, Amir Globerson, Hebrew University, and Tommi Jaakkola, Massachusetts Institute of Technology . . . 129

Improved Moves for Truncated Convex Models, M Pawan Kumar, University of Oxford, and Philip H S Torr, Oxford Brookes University . . . 131

Learning Bounded Treewidth Bayesian Networks, Gal Elidan, Hebrew University, and Stephen Gould, Stanford University . . . 117

Bayesian Network Score Approximation using a Metagraph Kernel, Benjamin Yackley, Department of Computer Science, University of New Mexico, Eduardo Corona, New York University, and Terran Lane, Department of Computer Science, University of New Mexico . . . 118

Near-minimax recursive density estimation on the binary hypercube, Maxim Raginsky, Duke University, Svetlana Lazebnik, University of North Carolina at Chapel Hill, Rebecca Willett and Jorge Silva, Duke University . . . 119

Bounds on marginal probability distributions, Joris M Mooij, MPI for Biological Cybernetics, and Hilbert J Kappen, Radboud University . . . 128

Posterior Consistency of the Silverman g-prior in Bayesian Model Choice, zhihua zhang, Michael I. Jordan, University of California, Berkeley, and Dit-Yan Yeung, Hong Kong University of Science & Technology . . . 124

Improving on Expectation Propagation, Manfred Opper, TU Berlin, Ulrich Paquet, University of Cambridge, and Ole Winther, University of Copenhagen . . . 126

The Gaussian Process Density Sampler, Ryan Adams, University of Cambridge, Iain Murray, University of Toronto, and David MacKay, University of Cambridge . . . 128

7:30pm–12:00am Poster Session

7:30pm–12:00am Demonstrations

Wednesday, December 10th

8:30am–9:50am **Oral Session:**
Neural Coding
Chair: Koby Crammer

8:30am **Invited Talk:** ***Connectome: the quest to deconstruct the brain,*** SEBASTIAN SEUNG, Howard Hughes Medical Institute and MIT 145

9:30am ***Modeling Short-term Noise Dependence of Spike Counts in Macaque Prefrontal Cortex,*** ARNO ONKEN, STEFFEN GRÜNEWÄLDER, Technische Universität Berlin, MATTHIAS MUNK, Max Planck Institute for Biological Cybernetics, and KLAUS OBERMAYER, Technische Universität Berlin 145

9:50am–10:00am **Spotlights**

Characterizing neural dependencies with Poisson copula models, PIETRO BERKES, Brandeis University, FRANK WOOD and JONATHAN PILLOW, Gatsby Computational Neuroscience Unit, UCL 157

A general framework for investigating how far the decoding process in the brain can be simplified, MASAFUMI OIZUMI, The University of Tokyo, TOSHIYUKI ISHII, RIKEN Brain Science Institute, KAZUYA ISHIBASHI, The University of Tokyo, TOSHIHIKO HOSOYA, RIKEN Brain Science Institute, and MASATO OKADA, The University of Tokyo 156

Estimating the Location and Orientation of Complex, Correlated Neural Activity using MEG, DAVID WIPF, JULIA OWEN, University of California, San Francisco, HAGAI T ATTIAS, KENSUKE SEKIHARA, Golden Metallic Inc, and SRIKANTAN S NAGARAJAN, University of California, San Francisco 159

Nonparametric sparse hierarchical models describe V1 fMRI responses to natural images, PRADEEP RAVIKUMAR, VINCENT Q VU, BIN YU, THOMAS NASELARIS, KENDRICK KAY and JACK GALLANT, University of California, Berkeley 186

Artificial Olfactory Brain for Mixture Identification, MEHMET K MUEZZINOGLU, ALEXANDER VERGARA, RAMON HUERTA, University of California San Diego, THOMAS NOWOTNY, University of Sussex, NIKOLAI RULKOV, HENRY ABARBANEL, ALLEN SELVERSTON and MIKHAIL RABINOVICH, University of California San Diego 159

Spike Feature Extraction Using Informative Samples, ZHI YANG, QI ZHAO and WENTAI LIU, University of California at Santa Cruz 156

The Conjoint Effect of Divisive Normalization and Orientation Selectivity on Redundancy Reduction, FABIAN H. SINZ and MATTHIAS BETHGE, MPI for Biological Cybernetics 158

The Infinite Factorial Hidden Markov Model, JURGEN VAN GAEL, University of Cambridge, YEE WHYE TEH, Gatsby Computational Neuroscience Unit, UCL, and ZOUBIN GHAHRAMANI, CMU & University of Cambridge 181

Mixed Membership Stochastic Blockmodels, EDOARDO M. AIROLDI, DAVID M. BLEI, Princeton University, STEPHEN E. FIENBERG and ERIC P. XING, Carnegie Mellon University 176

Supervised Bipartite Graph Inference, YOSHIHIRO YAMANISHI, Mines ParisTech 164

Stochastic Relational Models for Large-scale Dyadic Data using MCMC, SHENGHUO ZHU, KAI YU and YIHONG GONG, NEC Laboratories America, Inc. 177

Using matrices to model symbolic relationship, ILYA SUTSKEVER and GEOFFREY E HINTON, University of Toronto 181

2:00pm–3:20pm **Oral Session:**
Physics and High Order Statistics
Chair: Tony Jebara

2:00pm **Invited Talk:** ***Machine Learning in High Energy Physics,*** HARRISON PROSPER, Florida State University 151

3:00pm ***Characteristic Kernels on Groups and Semigroups,*** KENJI FUKUMIZU, Institute of Statistical Mathematics, BHARATH K SRIPERUMBUDUR, University of California, San Diego and Max Planck Institute for Biological Cybernetics, ARTHUR GRETTON and BERNHARD SCHÖLKOPF, Max Planck Institute for Biological Cybernetics 151

3:20pm–3:30pm **Spotlights**

Kernel Measures of Independence for non-iid Data, XINHUA ZHANG, NICTA and Australian National University, LE SONG, Carnegie Mellon University, ARTHUR GRETTON, Max Planck Institute for Biological Cybernetics, and ALEX J. SMOLA, Yahoo! Research 170

Multi-label Multiple Kernel Learning, SHUIWANG JI, LIANG SUN, Arizona State University, RONG JIN, Michigan State University, and JIEPING YE, Arizona State University 163

Scalable Algorithms for String Kernels with Inexact Matching, PAVEL P KUKSA, PAI-HSI HUANG and VLADIMIR PAVLOVIC, Rutgers University 164

Kernelized Sorting, NOVI QUADRIANTO, Australian National University, LE SONG, CMU, and ALEX J. SMOLA, Yahoo! Research 174

Learning with Consistency between Inductive Functions and Kernels, HAIXUAN YANG, IRWIN KING and MICHAEL LYU, The Chinese University of Hong Kong 171

4:00pm–5:20pm **Oral Session:**
Sparse Learning and Ranking
Chair: Jean-Yves Audibert

5:20pm–5:30pm **Spotlights**

7:30pm–12:00am **Poster Session**

W74 ***Shared Segmentation of Natural Scenes Using Dependent Pitman-Yor Processes,*** ERIK SUDDERTH and MICHAEL JORDAN, University of California, Berkeley 193

W75 ***Model selection and velocity estimation using novel priors for motion patterns,*** SHUANG WU, HONGJING LU and ALAN YUILLE, University of California, Los Angeles 149

W76 ***Using Bayesian Dynamical Systems for Motion Template Libraries,*** SILVIA CHIAPPA, JENS KOBER and JAN PETERS, MPI for Biological Cybernetics 185

W77 ***Dynamic visual attention: searching for coding length increments,*** XIAODI HOU, California Institute of Technology, and LIQING ZHANG, Shanghai Jiao Tong University 185

W78 ***Offline Handwriting Recognition with Multidimensional Recurrent Neural Networks,*** ALEX GRAVES, TU Munich, and JUERGEN SCHMIDHUBER, IDSIA and TU Munich 186

W79 ***Nonparametric sparse hierarchical models describe V1 fMRI responses to natural images,*** PRADEEP RAVIKUMAR, VINCENT Q VU, BIN YU, THOMAS NASELARIS, KENDRICK KAY and JACK GALLANT, University of California, Berkeley 186

W80 ***Reducing statistical dependencies in natural signals using radial Gaussianization,*** SIWEI LYU, University at Albany, SUNY, and EERO SIMONCELLI, New York University 194

7:30pm–12:00am Demonstrations

1b ***Visualizing NIPS Cooperations using Multiple Maps t-SNE,*** LAURENS VAN DER MAATEN, Maastricht University, MICC-IKAT, and GEOFFREY HINTON, University of Toronto 187

2b ***Real-time Topology Learning,*** KYNAN ENG, TOBI DELBRUCK, MATTHIAS SCHRAG and MARTIN BOERLIN, ETH Zurich 188

3b ***Infer.NET: Software for Graphical Models,*** TOM MINKA, JOHN WINN, JOHN GUIVER and ANITHA KANNAN, Microsoft Research 188

4b ***RL-Glue: From Grid Worlds to Sensor Rich Robots,*** BRIAN TANNER, ADAM WHITE and RICHARD S. SUTTON, University of Alberta 189

5b ***The Chiara: A New Robotic Platform for Education and Research,*** DAVID S. TOURETZKY, ETHAN J. TIRA-THOMPSON, GAKU SATO and DAVID P. RICE, Carnegie Mellon University 190

6b ***Machine Perception for Human Machine Interaction,*** PAUL RUVOLO, MARIAN STEWART BARTLETT, NICHOLAS BUTKO, CLAUDIA LAINSCSEK, GWEN LITTLEWORTH, JACOB WHITEHILL, TINGFAN WU and JAVIER MOVELLAN, University of California at San Diego 190

Thursday, December 11th

8:30am–10:10am **Oral Session:**
Nonparametric Processes, Scene Processing, and Image Statistics
Chair: Max Welling

10:40am–12:00pm **Oral Session:**
Attention and Mind
Chair: Aaron Courville

1:30pm–4:30pm **Mini Symposia**
Parallel sessions with one break.

Abstracts of Tutorials

Tutorial Session 1, 9:30am–11:30am

Tutorial: *Compressed Sensing*

Emmanuel Candes emmanuel@acm.caltech.edu
California Institute of Technology, Pasadena, CA, United States

One of the central tenets of signal processing and data acquisition is the Shannon/Nyquist sampling theory: the number of samples needed to capture a signal is dictated by its bandwidth. This tutorial surveys a novel sampling or sensing theory which goes somewhat against this conventional wisdom. This theory now known as “Compressed Sensing” or “Compressive Sampling” allows the faithful recovery of signals and images from what appear to be highly incomplete sets of data, i.e. from far fewer measurements or data bits than traditional methods use. We will present the key ideas underlying this new sampling or sensing theory, will survey some of the most important results, and discuss some of the most exciting current developments. We will emphasize the practicality and the broad applicability of this technique, and discuss what we believe are far reaching implications; e.g. procedures for sensing and compressing data simultaneously and much faster. Finally, there are already many ongoing efforts to build a new generation of sensing devices based on compressed sensing and we will discuss remarkable recent progress in this area as well.

Emmanuel Candes received his B. Sc. degree from the Ecole Polytechnique (France) in 1993, and the Ph.D. degree in statistics from Stanford University in 1998. He is the Ronald and Maxine Linde Professor of Applied and Computational Mathematics at the California Institute of Technology. Prior to joining Caltech, he was an Assistant Professor of Statistics at Stanford University, 1998–2000. His research interests are in computational harmonic analysis, multiscale analysis, approximation theory, statistical estimation and detection with applications to the imaging sciences, signal processing, scientific computing, inverse problems. Other topics of interest include theoretical computer science, mathematical optimization, and information theory. Dr. Candes received the Third Popov Prize in Approximation Theory in 2001, and the DOE Young Investigator Award in 2002. He was selected as an Alfred P. Sloan Research Fellow in 2001. He co-authored a paper that won the Best Paper Award of the European Association for Signal, Speech and Image Processing (EURASIP) in 2003. He was selected as the main lecturer at the NSF-sponsored 29th Annual Spring Lecture Series in the Mathematical Sciences in 2004 and as the Aziz Lecturer in 2007. He has also given plenary and keynote addresses at major international conferences including ICIAM 2007 and ICIP 2007. In 2005, he was awarded the James H. Wilkinson Prize in Numerical Analysis and Scientific Computing by SIAM. Finally, he is the recipient of the 2006 Alan T. Waterman Medal awarded by the US National Science Foundation

Tutorial Session 1, 9:30am–11:30am

Tutorial: *Prequential Analysis*

Philip Dawid A.P.Dawid@statslab.cam.ac.uk
Centre for Mathematical Sciences, University of Cambridge, Cambridge, United Kingdom

"Prequential" is a portmanteau word for predictive sequential–a broad statistical methodology founded on a view of data (like Mark Twain's view of history) as "just one darned thing after another". This tutorial will outline the basic concepts of prequential analysis, and some of its properties and applications. Suppose we have been sold a method that purports to learn projectible regularities in a data-sequence. Prequential analysis assesses how well this works by contrasting its one-step ahead forecasts with realised outcomes. Like cross-validation, by never using an observation to contribute to its own forecast this procedure avoids over-optimistic "substitution bias". But, unlike cross-validation, it also avoids indirect substitution bias, since no two observations can each contribute to the other's forecast. This leads to more reliable performance, even when there is no natural ordering in the data. It also allows the development of a fruitful and elegant general theory. There are close connexions with martingale theory, wih the theory of online learning with expert advice, and with game-theoretic probability. Many traditional statistical concepts, such as consistency and efficiency, as well as practical techniques such as model selection and inference from misspecified models, take on an interesting new aspect when reconsidered from the prequential point of view.

Philip Dawid is Professor of Statistics at the University of Cambridge, having been Pearson Professor of Statistics at University College London from 1989-2007. He is Chartered Statistician and Fellow of the Royal Statistical Society, which has awarded him the Guy Medal in Bronze and in Silver; elected Fellow of the Institute of Mathematical Statistics; elected Member of the International Statistical Institute; and a Member of the Organising Committee for the Valencia International Meetings on Bayesian Statistics. He has served as Editor of the Journal of the Royal Statistical Society (Series B) and Biometrika, and (currently) of Bayesian Analysis. He was President of the International Society for Bayesian Analysis in 2000. Much of his research has focused on the logical foundations of probability and statistics, with emphasis on the Bayesian approach. His co-authored book "Probabilistic Networks and Expert Systems" (Springer, 1999) won the 2002 DeGroot Prize. He is interested in logical problems of structuring legal evidence, and led an international research project applying Bayesian networks to complex cases of forensic identification from DNA profiles. He recently directed a multidisciplinary research programme "Evidence, Inference and Enquiry" at University College London.

Tutorial Session 2, 1:00pm–3:00pm

Tutorial: *Agnostic Learning: Algorithms and Theory*

Adam Kalai adum@microsoft.com
Microsoft Research New Engand, Cambridge, MA, United States

Agnostic Learning (Kearns, Schapire and Sellie '92; Haussler '90) is a branch of Computational Learning Theory in which little or no assumptions are made about the true function being learned. The results are powerful, noise-tolerant binary-classification algorithms. Like Valiant's seminal PAC model, the algorithms are required to be computationally efficient. Unlike the PAC model, it does not assume anything about the "true" labeling function, and hence Agnostic Learning can also be viewed as learning with arbitrary or even adversarial noise. Recent progress has shown that several classical algorithms can be improved and made agnostic. Moreover, the agnostic perspective sheds light on several key Machine Learning notions, such as Fourier learning, SVMs, Active Learning, Decision Tree Learning, and Boosting. Hence, this tutorial on Agnostic Learning also presents an opportunity to overview and reconsider many beautiful notions from Computational Learning Theory.

Adam Tauman Kalai received his BA (1996) from Harvard, and MA (1998) and PhD (2001) under the supervision of Avrim Blum from CMU. After an NSF postdoctoral fellowship at M.I.T. with Santosh Vempala, he served as an assistant professor at the Toyota Technological institute at Chicago and then at Georgia Tech. He is now a Senior Research Scientist at Microsoft Research New England. His honors include an NSF CAREER award, and an Alfred P. Sloan fellowship. His research focuses on computational learning theory, game theory, algorithms, and online optimization.

Tutorial Session 2, 1:00pm–3:00pm

Tutorial: *The Neurobiology of Decision Making*

Michael Shadlen shadlen@u.washington.edu
School of Medicine, University of Washington, Seattle, WA, United States

This tutorial lecture describes recent advances in our understanding of how simple decisions are implemented in the brain. The study of decision-making opens a window on the neural basis of many other higher cognitive capacities which also use information in a contingent fashion and in a flexible time frame — from the immediacy of sensory events or the need to control a body in real time. I will attempt to connect the findings from neural recording and microstimulation experiments to the broader computational problem of inference on data streams. The neurobiology inspires us to reformulate many inference/reasoning problems with explicit incorporation of stopping rules and an interplay between time and probability. These insights were anticipated by Alan Turing in his codebreaking work during World War II, and they were developed by Abraham Wald into the field of Sequential Analysis. Besides its mathematical elegance and strategic importance, this computational mechanism may be essential for higher brain function. If so, the principles revealed by the study of decision-making may one day lead to new treatments for neurological disorders affecting our most cherished cognitive abilities.

Michael Shadlen MD, PhD is an Investigator of the Howard Hughes Medical Institute and Professor of Physiology & Biophysics at the University of Washington, where is also an adjunct Professor of Neurology. He performed undergraduate and medical studies at Brown University and obtained a PhD from UC Berkeley in visual neuroscience under the guidance of Ralph D. Freeman. He received postgraduate clinical training in Neurology at Stanford Medical Center. He then returned to basic neuroscience as a fellow in the laboratory of William T. Newsome, where he began to work on the neurobiology of decision-making. Shadlen studies neurons in the association cortex that process information from the visual cortex to give rise to interpretations, decisions, and plans for behavior. His experiments combine electrophysiology and behavioral and computational methods to advance our knowledge of higher brain function.

Tutorial Session 3, 3:30pm–5:30pm

Tutorial: *Statistical Models of Visual Images*

Eero Simoncelli eero.simoncelli@nyu.edu
New York University, New York, NY, United States

An overview of the problem of statistical image modeling, providing a set of examples from the image processing, computer vision, and computational neuroscience literatures. I'll describe the empirical observations that led to these models, and the use of these models in various applications.

Eero P. Simoncelli received the B.S. degree in Physics in 1984 from Harvard University, studied applied mathematics at Cambridge University for a year and a half, and then received the M.S. degree in 1988 and the Ph.D. degree in 1993, both in Electrical Engineering from the Massachusetts Institute of Technology. He was an Assistant Professor in the Computer and Information Science department at the University of Pennsylvania from 1993 until 1996. He moved to New York University in September of 1996, where he is currently a Professor in Neural Science, Mathematics, and Psychology. In August 2000, he became an Associate Investigator of the Howard Hughes Medical Institute, under their new program in Computational Biology. His research interests span a wide range of topics in the representation and analysis of visual images, in both machine and biological systems.

Tutorial Session 3, 3:30pm–5:30pm

Tutorial: *The Smoothed Analysis of Algorithms*

Daniel Spielman spielman@cs.yale.edu
Department of Computer Science, Yale University, New Haven, CT, United States

Theorists have long been challenged by the existence of remarkable algorithms that are known by scientists and engineers to work well in practice, but whose theoretical analyses have been negative or unconvincing. The root of the problem is that algorithms are usually analyzed in one of two ways: by worst-case or average-case analysis. The former can improperly suggest that an algorithm will perform poorly, while the latter can be unconvincing because the random inputs it considers may fail to resemble those encountered in practice. Smoothed analysis can help explain the success of many algorithms that both worst-case and average-case analyses cannot. In smoothed analysis, we measure the performance of an algorithm under slight random perturbations of arbitrary inputs, and bound the performance as a function of the size of the input and the magnitude of the perturbation. If an algorithm has low smoothed complexity, then it should perform well if its inputs are subject to random noise. We will explain how smoothed analysis has been used to analyze the behavior of many algorithms and heuristics, including the simplex algorithm, the Perceptron algorithm Gaussian Elimination with partial pivoting, and k-means clustering.

Daniel Spielman is a Professor of Applied Mathematics and Computer Science at Yale University. He received his B.A. in Mathematics and Computer Science from Yale in 1992, and his Ph.D in Applied Mathematics from M.I.T. in 1995. He spent a year as a postdoc in Computer Science at U.C. Berkeley, and then taught in the Applied Mathematics department at M.I.T. until 2005. The awards he has received include the 1995 ACM Doctoral Dissertation Award, the 2002 IEEE Information Theory Paper Award, and most recently the 2008 Gödel Prize for his work on Smoothed Analysis. His main research interests include the design and analysis of algorithms, graph theory, and combinatorial scientific computing.

Abstracts of Presentations

Monday, December 8th

Oral Session (7:45pm–8:25pm): "Clustering"

7:45pm *Measures of Clustering Quality: A Working Set of Axioms for Clustering*

Shai Ben-David shai@cs.uwaterloo.ca
University of Waterloo, waterloo, Canada
Margareta Ackerman mackerma@uwaterloo.ca
University of Waterloo, Waterloo, Canada

Aiming towards the development of a general clustering theory, we discuss abstract axiomatization for clustering. In this respect, we follow up on the work of Kelinberg, ([Kleinberg]) that showed an impossibility result for such axiomatization. We argue that an impossibility result is not an inherent feature of clustering, but rather, to a large extent, it is an artifact of the specific formalism used in [Kleinberg].

As opposed to previous work focusing on *clustering functions*, we propose to address *clustering quality measures* as the primitive object to be axiomatized. We show that principles like those formulated in Kleinberg's axioms can be readily expressed in the latter framework without leading to inconsistency.

A clustering-quality measure is a function that, given a data set and its partition into clusters, returns a non-negative real number representing how 'strong' or 'conclusive' the clustering is. We analyze what clustering-quality measures should look like and introduce a set of requirements ('axioms') that express these requirement and extend the translation of Kleinberg's axioms to our framework.

We propose several natural clustering quality measures, all satisfying the proposed axioms. In addition, we show that the proposed clustering quality can be computed in polynomial time.

8:05pm *Influence of graph construction on graph-based clustering measures*

Markus Maier markus.maier@tuebingen.mpg.de
Ulrike von Luxburg ulrike.luxburg@tuebingen.mpg.de
MPI for Biological Cybernetics, Tuebingen, Germany
Matthias Hein hein@cs.uni-sb.de
Saarland University, Saarbruecken, Germany

Graph clustering methods such as spectral clustering are defined for general weighted graphs. In machine learning, however, data often is not given in form of a graph, but in terms of similarity (or distance) values between points. In this case, first a neighborhood graph is constructed using the similarities between the points and then a graph clustering algorithm is applied to this graph. In this paper we investigate the influence of the construction of the similarity graph on the clustering results, from a theoretical point of view. We first study the convergence of graph clustering criteria such as the normalized cut (Ncut) as the sample size tends to infinity. We find that the limit expressions are different for different types of graph, for example the r-neighborhood graph or the k-nearest neighbor graph. In plain words: Ncut on a knn graph does something systematically different than Ncut on an r-neighborhood graph! This finding shows that graph clustering criteria cannot be studied independently of the kind of graph they will be applied to. We also provide examples which show how those differences lead to big differences in clustering results in practice.

Spotlights (8:25pm–8:45pm)

- ***On the Reliability of Clustering Stability in the Large Sample Regime***
  OHAD SHAMIR and NAFTALI TISHBY, The Hebrew University.
  See abstract, page 67.

- ***Spectral Clustering with Perturbed Data***
  LING HUANG, Intel Research, DONGHUI YAN, MICHAEL JORDAN, University of California, Berkeley, and NINA TAFT, Intel Research.
  See abstract, page 68.

- ***A "Shape Aware" Model for semi-supervised Learning of Objects and its Context***
  ABHINAV GUPTA, University of Maryland, JIANBO SHI, University of Pennsylvania, and LARRY DAVIS, University of Maryland.
  See abstract, page 81.

- ***One sketch for all: Theory and Application of Conditional Random Sampling***
  PING LI, Cornell University, KENNETH W CHURCH, Microsoft, and TREVOR HASTIE, Stanford University.
  See abstract, page 70.

- ***Gates***
  TOM MINKA and JOHN WINN, Microsoft Research.
  See abstract, page 75.

- ***Fast Computation of Posterior Mode in Multi-Level Hierarchical Models***
  LIANG ZHANG, Duke University, and DEEPAK AGARWAL, Yahoo! Research.
  See abstract, page 75.

- ***Relative Performance Guarantees for Approximate Inference in Latent Dirichlet Allocation***
  Indraneel Mukherjee and David Blei, Princeton University.
  See abstract, page 74.

- ***Efficient Inference in Phylogenetic InDel Trees***
  Alexandre Bouchard-Côté, Michael Jordan and Dan Klein, University of California, Berkeley.
  See abstract, page 63.

- ***Large Margin Taxonomy Embedding for Document Categorization***
  Kilian Q Weinberger and Olivier Chapelle, Yahoo! Research.
  See abstract, page 60.

- ***Look Ma, No Hands: Analyzing the Monotonic Feature Abstraction for Text Classification***
  Doug Downey, Northwestern University, and Oren Etzioni, University of Washington.
  See abstract, page 58.

- ***Syntactic Topic Models***
  Jordan L Boyd-Graber and David Blei, Princeton University.
  See abstract, page 59.

- ***Semi-supervised Learning with Weakly-Related Unlabeled Data : Towards Better Text Categorization***
  Liu Yang, Carnegie Mellon University, Rong Jin, Michigan State University, and Rahul Sukthankar, Intel Research and Carnegie Mellon University.
  See abstract, page 59.

- ***Unsupervised Learning of Visual Sense Models for Polysemous Words***
  Kate Saenko, MIT CSAIL, and Trevor Darrell, University of California, Berkeley.
  See abstract, page 75.

- ***Deep Learning with Kernel Regularization for Visual Recognition***
  Kai Yu, Wei Xu and Yihong Gong, NEC Laboratories America, Inc..
  See abstract, page 78.

- ***Evaluating probabilities under high-dimensional latent variable models***
  Iain Murray and Ruslan Salakhutdinov, University of Toronto.
  See abstract, page 76.

Poster Session (8:45pm–12:00am)

M1 *Logistic Normal Priors for Unsupervised Probabilistic Grammar Induction*

Shay B. Cohen scohen@cs.cmu.edu
Kevin Gimpel kgimpel@cs.cmu.edu
Noah A. Smith nasmith@cs.cmu.edu
Carnegie Mellon University, Language Technologies Institute, School of Computer Science, Pittsburgh, PA, United States

We explore a novel approach for modeling probabilistic grammars, a family of distributions over discrete structures that includes hidden Markov models and probabilistic context-free grammars. Our model extends the correlated topic model to probabilistic grammars, using the logistic normal distribution as a prior over the grammar parameters. We derive a variational EM algorithm for that model, and then experiment with the task of unsupervised grammar induction for natural language dependency parsing, and show that our model achieves superior results over previous models that use priors with different properties.

M2 *Regularized Learning with Networks of Features*

Ted Sandler tsandler@cis.upenn.edu
University of Pennsylvania, Philadelphia, PA, United States
John Blitzer blitzer@cs.berkeley.edu
University of California, Berkeley, CA, United States
Partha Pratim Talukdar partha@cis.upenn.edu
University of Pennsylvania, Philadelphia, United States
Lyle H. Ungar ungar@cis.upenn.edu
University of Pennsylvania, Philadelphia, PA, United States

For many supervised learning problems, we possess prior knowledge about which features yield similar information about the target variable. In predicting the topic of a document, we might know that two words are synonyms, or when performing image recognition, we know which pixels are adjacent. Such synonymous or neighboring features are near-duplicates and should therefore be expected to have similar weights in a good model. Here we present a framework for regularized learning in settings where one has prior knowledge about which features are expected to have similar and dissimilar weights. This prior knowledge is encoded as a graph whose vertices represent features and whose edges represent similarities and dissimilarities between them. During learning, each feature's weight is penalized by the amount it differs from the average weight of its neighbors. For text classification, regularization using graphs of word co-occurrences outperforms manifold learning and compares favorably to other recently proposed semi-supervised learning methods. For sentiment analysis, feature graphs constructed from declarative human knowledge, as well as from auxiliary task learning, significantly improve prediction accuracy.

M3 *Look Ma, No Hands: Analyzing the Monotonic Feature Abstraction for Text Classification*

Doug Downey ddowney@eecs.northwestern.edu
Northwestern University, Evanston, IL, United States
Oren Etzioni etzioni@cs.washington.edu
University of Washington, Seattle, WA, United States

Is accurate classification possible in the absence of hand-labeled data? This paper introduces the Monotonic Feature (MF) abstraction—where the probability of class membership increases monotonically with the MF's value. The paper proves that when an MF is given, PAC learning is possible with no hand-labeled data under certain assumptions.

We argue that MFs arise naturally in a broad range of textual classification applications. On the classic "20 Newsgroups" data set, a learner given an MF and unlabeled data achieves classification accuracy equal to that of a state-of-the-art semi-supervised learner relying on 160 hand-labeled examples. Even when MFs are not given as input, their presence or absence can be determined from a small amount of hand-labeled data, which yields a new semi-supervised learning method that reduces error by 15% on the 20 Newsgroups data.

Spotlight presentation, Monday, 8:25pm.

M4 *Syntactic Topic Models*

Jordan L Boyd-Graber jbg@princeton.edu
Computer Science, Princeton University, Princeton, NJ
David Blei blei@cs.princeton.edu
Princeton University, Princeton, NJ, United States

We develop the syntactic topic model (STM), a nonparametric Bayesian model of parsed documents. The STM generates words that are both thematically and syntactically constrained, which combines the semantic insights of topic models with the syntactic information available from parse trees. Each word of a sentence is generated by a distribution that combines document-specific topic weights and parse-tree specific syntactic transitions. Words are assumed generated in an order that respects the parse tree. We derive an approximate posterior inference method based on variational methods for hierarchical Dirichlet processes, and we report qualitative and quantitative results on both synthetic data and hand-parsed documents.

Spotlight presentation, Monday, 8:25pm.

M5 *Semi-supervised Learning with Weakly-Related Unlabeled Data : Towards Better Text Categorization*

Liu Yang liuy@cs.cmu.edu
Machine Learning Department, Carnegie Mellon
University, Pittsburgh, PA, United States
Rong Jin rongjin@cse.msu.edu
Michigan State University, East Lansing, MI, United States
Rahul Sukthankar rahuls@cs.cmu.edu
Intel Research and Carnegie Mellon University, Pittsburgh, PA, United States

The cluster assumption is exploited by most semi-supervised learning (SSL) methods. However, if the unlabeled data is merely weakly related to the target classes, it becomes questionable whether driving the decision boundary to the low density regions of the unlabeled data will help the classification. In such case, the cluster assumption may not be valid; and consequently how to leverage this type of unlabeled data to enhance the classification accuracy becomes a challenge. We introduce "Semi-supervised Learning with Weakly-Related Unlabeled Data" (SSLW), an inductive method that builds upon the maximum-margin approach, towards a better usage of weakly-related unlabeled information. Although the SSLW could improve a wide range of classification tasks, in this paper, we focus on text categorization with a small training pool. The key assumption behind this work is that, even with different topics, the word usage patterns across different corpora tends to be consistent. To this end, SSLW estimates the optimal word-correlation matrix that is consistent with both the co-occurrence information derived from the weakly-related unlabeled documents and the labeled documents. For empirical evaluation, we present a direct comparison with a number of state-of-the-art methods for inductive semi-supervised learning and text categorization; and we show that SSLW results in a significant improvement in categorization accuracy, equipped with a small training set and an unlabeled resource that is weakly related to the test beds.

Spotlight presentation, Monday, 8:25pm.

M6 *Generative and Discriminative Learning with Unknown Labeling Bias*

Miroslav Dudík mdudik@cmu.edu
Carnegie Mellon University, Pittsburgh, PA, United States
Steven Phillips phillips@research.att.com
AT&T Labs-Research, Florham Park, NJ, United States

We apply robust Bayesian decision theory to improve both generative and discriminative learners under bias in class proportions in labeled training data, when the true class proportions are unknown. For the generative case, we derive an entropy-based weighting that maximizes expected log likelihood under the worst-case true class proportions. For the discriminative case, we derive a multinomial logistic model that minimizes worst-case conditional log loss. We apply our theory to the modeling of species geographic distributions from presence data, an extreme case of label bias since there is no absence data. On a benchmark dataset, we find that entropy-based weighting offers an improvement over constant estimates of class proportions, consistently reducing log loss on unbiased test data.

M7 *Large Margin Taxonomy Embedding for Document Categorization*

Kilian Q Weinberger kilianweinberger@yahoo.de
Yahoo! Research, Mountain View, CA, United States
Olivier Chapelle chap@yahoo-inc.com
Yahoo! Research, Santa Clara, CA, United States

Applications of multi-class classification, such as document categorization, often appear in cost-sensitive settings. Recent work has significantly improved the state of the art by moving beyond "flat" classification through incorporation of class hierarchies [Cai and Hoffman 04]. We present a novel algorithm that goes beyond hierarchical classification and estimates the latent semantic space that underlies the class hierarchy. In this space, each class is represented by a prototype and classification is done with the simple nearest neighbor rule. The optimization of the semantic space incorporates large margin constraints that ensure that for each instance the correct class prototype is closer than any other. We show that our optimization is convex and can be solved efficiently for large data sets. Experiments on the OHSUMED medical journal data base yield state-of-the-art results on topic categorization.

Spotlight presentation, Monday, 8:25pm.

M8 *Learning Taxonomies by Dependence Maximization*

Matthew Blaschko matthew.blaschko@tuebingen.mpg.de
Arthur Gretton arthur@tuebingen.mpg.de
Max Planck Institute for Biological Cybernetics, Tuebingen, Germany

We introduce a family of unsupervised algorithms, numerical taxonomy clustering, to simultaneously cluster data, and to learn a taxonomy that encodes the relationship between the clusters. The algorithms work by maximizing the dependence between the taxonomy and the original data. The resulting taxonomy is a more informative visualization of complex data than simple clustering; in addition, taking into account the relations between different clusters is shown to substantially improve the quality of the clustering, when compared with state-of-the-art algorithms in the literature (both spectral clustering and a previous dependence maximization approach). We demonstrate our algorithm on image and text data.

M9 *Structured ranking learning using cumulative distribution networks*

Jim C. Huang jim@psi.toronto.edu
Brendan J. Frey frey@psi.toronto.edu
Probabilistic and Statistical Inference Group, University of Toronto, Toronto, ON, Canada

Ranking is at the heart of many information retrieval applications. Unlike standard regression or classification in which we predict outputs independently, in ranking we are interested in predicting structured outputs so that misranking one object can significantly affect whether we correctly rank the other objects. In practice, the problem of ranking involves a large number of objects to be ranked and either approximate structured prediction methods are required, or assumptions of independence between object scores must be made in order to make the problem tractable. We present a probabilistic method for learning to rank using the graphical modelling framework of cumulative distribution networks (CDNs), where we can take into account the structure inherent to the problem of ranking by modelling the joint cumulative distribution functions (CDFs) over multiple pairwise preferences. We apply our framework to the problem of document retrieval in the case of the OHSUMED benchmark dataset. We will show that the RankNet, ListNet and

ListMLE probabilistic models can be viewed as particular instances of CDNs and that our proposed framework allows for the exploration of a broad class of flexible structured loss functionals for ranking learning.

M10 *Overlaying classifiers: a practical approach for optimal ranking*

Stéphan J.M. Clémençon stephan.clemencon@telecom-paristech.fr
LTCI, Telecom ParisTech and CNRS, Institut Telecom, PARIS, France
Nicolas Vayatis nicolas.vayatis@gmail.com
Ecole Normale Supérieure de Cachan, CMLA - CNRS & UniverSud, Cachan, France

ROC curves are one of the most widely used displays to evaluate performance of scoring functions. In the paper, we propose a statistical method for directly optimizing the ROC curve. The target is known to be the regression function up to an increasing transformation and this boils down to recovering the level sets of the latter. We propose to use classifiers obtained by empirical risk minimization of a weighted classification error and then to construct a scoring rule by overlaying these classifiers. We show the consistency and rate of convergence to the optimal ROC curve of this procedure in terms of supremum norm and also, as a byproduct of the analysis, we derive an empirical estimate of the optimal ROC curve.

M11 *Predictive Indexing for Fast Search*

Sharad Goel goel@yahoo-inc.com
John Langford jl@yahoo-inc.com
Alexander L Strehl strehl@yahoo-inc.com
Yahoo! Research, New York, NY, United States

We tackle the computational problem of query-conditioned search. Given a machine-learned scoring rule and a query distribution, we build a predictive index by precomputing lists of potential results sorted based on an expected score of the result over future queries. The predictive index datastructure supports an anytime algorithm for approximate retrieval of the top elements. The general approach is applicable to webpage ranking, internet advertisement, and approximate nearest neighbor search. It is particularly effective in settings where standard techniques (e.g., inverted indices) are intractable. We experimentally find substantial improvement over existing methods for internet advertisement and approximate nearest neighbors.

M12 *Integrating Locally Learned Causal Structures with Overlapping Variables*

Robert E. Tillman rtillman@andrew.cmu.edu
Department of Philosophy and Machine Learning Department, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, United States
David Danks ddanks@cmu.edu
Clark Glymour cg09@andrew.cmu.edu
Department of Philosophy, Carnegie Mellon University, Pittsburgh, PA, United States; and Institute for Human and Machine Cognition, Pensacola, FL, United States

In many domains, data are distributed among datasets that share only some variables; other recorded variables may occur in only one dataset. While there are multiple asymptotically correct, informative algorithms for discovering causal relations from a single dataset, even with missing values and hidden variables, until now there have been no such reliable procedures for distributed data with overlapping variables. We describe a new asymptotically correct procedure, ION, that discovers a minimal equivalence class of causal structures from such data using local independence constraints learned from the marginal datasets and evaluate its performance, using synthetic and real-world data, against causal discovery algorithms for single complete datasets and applying Structural EM, a heuristic structure learning procedure for data with missing values, to the concatenated data.

M13 *Finding Latent Causes in Causal Networks: an Efficient Approach Based on Markov Blankets*

Jean-Philippe Pellet jep@zurich.ibm.com
IBM Zurich Research Lab, Rüschlikon, and Swiss Institute of Technology Zurich, Switzerland
André Elisseeff ael@zurich.ibm.com
IBM Zurich Research Lab, Rüschlikon, Switzerland

Causal structure-discovery techniques usually assume that all causes of more than one variable are observed. This is the so-called causal sufficiency assumption. In practice, it is untestable, and often violated. In this paper, we present an efficient causal structure-learning algorithm, suited for causally insufficient data. Similar to algorithms such as IC* and FCI, the proposed approach drops the causal sufficiency assumption and learns a structure that indicates (potential) latent causes for pairs of observed variables. Assuming a constant local density of the data-generating graph, our algorithm makes a quadratic number of conditional-independence tests w.r.t. the number of variables. We show with experiments that our algorithm is comparable to the state-of-the-art FCI algorithm in accuracy, while being several orders of magnitude faster on large problems. We conclude that MBCS* makes a new range of causally insufficient problems computationally tractable.

M14 *An Empirical Analysis of Domain Adaptation Algorithms for Genomic Sequence Analysis*

Gabriele Schweikert gabriele.schweikert@tuebingen.mpg.de
Friedrich Miescher Laboratory of the Max Planck Society, MPI for Biological Cybernetics, MPI for Developmental Biology, Tübingen, Germany
Christian Widmer cwidmer@tuebingen.mpg.de
Friedrich Miescher Laboratory of the Max Planck Society, Tübingen, Germany
Bernhard Schölkopf bs@tuebingen.mpg.de
MPI for Biological Cybernetics, Tübingen, Germany
Gunnar Rätsch gunnar.raetsch@tuebingen.mpg.de
Friedrich Miescher Laboratory of the Max Planck Society, Tübingen, Germany

We study the problem of domain transfer for a supervised classification task in mRNA splicing. We consider a number of recent domain transfer methods from machine learning, including some that are novel, and evaluate them on genomic sequence data from model organisms of varying evolutionary distance. We find that in cases where the organisms are not closely related, the use of domain adaptation methods can help improve classification performance.

M15 *Efficient Inference in Phylogenetic InDel Trees*

Alexandre Bouchard-Côté bouchard@cs.berkeley.edu
Michael Jordan jordan@cs.berkeley.edu
Dan Klein klein@cs.berkeley.edu
University of California, Berkeley, CA, United States

Accurate and effcient inference in evolutionary trees is a central problem in computational biology. While classical treatments have made unrealistic site independence assumptions, ignoring insertions and deletions, realistic approaches require tracking insertions and deletions along the phylogenetic tree—a challenging and unsolved computational problem. We propose a new ancestry resampling procedure for inference in evolutionary trees. We evaluate our method in two problem domains—multiple sequence alignment and reconstruction of ancestral sequences—and show substantial improvement over the current state of the art.

Spotlight presentation, Monday, 8:25pm.

M16 *Differentiable Sparse Coding*

Drew Bagnell dbagnell+nips@ri.cmu.edu
David M Bradley davembradley@gmail.com
Carnegie Mellon University, Pittsburgh, PA, United States

Prior work has shown that features which appear to be biologically plausible as well as empirically useful can be found by sparse coding with a prior such as a laplacian (L_1) that promotes sparsity. We show how smoother priors can preserve the benefits of these sparse priors while adding stability to the Maximum A-Posteriori (MAP) estimate that makes it more useful for prediction problems. Additionally, we show how to calculate the derivative of the MAP estimate efficiently with implicit differentiation. One prior that can be differentiated this way is KL-regularization. We demonstrate its effectiveness on a wide variety of applications, and find that online optimization of the parameters of the KL-regularized model can significantly improve prediction performance.

M17 *PSDBoost: Matrix-Generation Linear Programming for Positive Semidefinite Matrices Learning*

Chunhua Shen chhshen@gmail.com
Alan Welsh alan.welsh@anu.edu.au
Lei Wang lei.wang@anu.edu.au
NICTA (National ICT Australia), Canberra, Australia

In this work, we consider the problem of learning a positive semidefinite matrix. The critical issue is how to preserve positive semidefiniteness during the course of learning. Our algorithm is mainly inspired by LPBoost [1] and the general greedy convex optimization framework of Zhang [2]. We demonstrate the essence of the algorithm, termed PSDBoost (positive semidefinite Boosting), by focusing on a few different applications in machine learning. The proposed PSDBoost algorithm extends traditional Boosting algorithms in that its parameter is a positive semidefinite matrix with trace being one instead of a classifier. PSDBoost is based on the observation that any trace-one positive semidefinitematrix can be decomposed into linear convex combinations of trace-one rank-one matrices, which serve as base learners of PSDBoost. Numerical experiments are presented.

M18 *Multi-stage Convex Relaxation for Learning with Sparse Regularization*

TONG ZHANG tzhang@stat.rutgers.edu
Rutgers University, Piscataway, NJ, United States

We study learning formulations with non-convex regularizaton that are natural for sparse linear models. There are two approaches to this problem: (1) Heuristic methods such as gradient descent that only find a local minimum. A drawback of this approach is the lack of theoretical guarantee showing that the local minimum gives a good solution. (2) Convex relaxation such as L_1-regularization that solves the problem under some conditions. However it often leads to sub-optimal sparsity in reality.

This paper tries to remedy the above gap between theory and practice. In particular, we investigate a multi-stage convex relaxation scheme for solving problems with non-convex regularization. Theoretically, we analyze the behavior of a resulting two-stage relaxation scheme for the capped-L_1 regularization. Our performance bound shows that the procedure is superior to the standard L_1 convex relaxation for learning sparse targets. Experiments confirm the effectiveness of this method on some simulation and real data.

M19 *The Infinite Hierarchical Factor Regression Model*

PIYUSH RAI piyush@cs.utah.edu
HAL DAUME III me@hal3.name
School of Computing, University of Utah, Salt Lake City, UT, United States

We propose a nonparametric Bayesian factor regression model that accounts for uncertainty in the number of factors, and the relationship between factors. To accomplish this, we propose a sparse variant of the Indian Buffet Process and couple this with a hierarchical model over factors, based on Kingman's coalescent. We apply this model to two problems (factor analysis and factor regression) in gene-expression data analysis.

M20 *Estimating vector fields using sparse basis field expansions*

STEFAN HAUFE haufe@cs.tu-berlin.de
TU Berlin, Dept. of Computer Science, Machine Learning Laboratory, Berlin, Germany
VADIM V NIKULIN vadim.nikulin@charite.de
Charité University Medicine, Dept. of Neurology, Campus Benjamin Franklin, Berlin, Germany
ANDREAS ZIEHE ziehe@cs.tu-berlin.de
KLAUS-ROBERT MÜLLER krm@cs.tu-berlin.de
TU Berlin, Dept. of Computer Science, Machine Learning Laboratory, Berlin, Germany
GUIDO NOLTE guido.nolte@first.fraunhofer.de
Fraunhofer Institute FIRST (IDA), Berlin, Germany

We introduce a novel framework for estimating vector fields using sparse basis field expansions (S-FLEX). The notion of basis fields, which are an extension of scalar basis functions, arises naturally in our framework from a rotational invariance requirement. We consider a regression setting as well as inverse problems. All variants discussed lead to second-order cone programming formulations. While our framework is generally applicable to any type of vector field, we focus in this paper on applying it to solving the EEG/MEG inverse problem. It is shown that significantly more precise and neurophysiologically more plausible location and shape estimates of cerebral current sources from EEG/MEG measurements become possible with our method when comparing to the state-of-the-art.

M21 *Sparsity of SVMs that use the epsilon-insensitive loss*

Ingo Steinwart is.nips2008@gmail.com
CCS-3, Los Alamos National Laboratory, Los Alamos, NM, United States
Andreas Christmann andreas.christmann@uni-bayreuth.de
Department of Mathematics, Chair Stochastics, University of Bayreuth, Bayreuth, Germany

In this paper lower and upper bounds for the number of support vectors are derived for support vector machines (SVMs) based on the epsilon-insensitive loss function. It turns out that these bounds are asymptotically tight under mild assumptions on the data generating distribution. Finally, we briefly discuss a trade-off in epsilon between sparsity and accuracy if the SVM is used to estimate the conditional median.

M22 *Model Selection in Gaussian Graphical Models: High-Dimensional Consistency of ℓ_1-regularized MLE*

Pradeep Ravikumar pradeepr@stat.berkeley.edu
Garvesh Raskutti garveshr@stat.berkeley.edu
University of California, Berkeley, CA, United States
Martin Wainwright wainwrig@eecs.berkeley.edu
Bin Yu binyu@stat.berkeley.edu
Department of EECS and Department of Statistics, University of California, Berkeley, CA, United States

We study the performance of the ℓ_1-regularized maximum likelihood estimator for the problem of estimating the graph structure associated with a Gaussian Markov random field (GMRF). We consider the performance of the estimator in the high-dimensional setting, where the number of nodes in the graph p, the number of edges in the graph s and the maximum node degree d, are allowed to grow as a function of the number of samples n. Our main result provides sufficient conditions on the quadruple (n, p, s, d) for the ℓ_1-regularized MLE estimator to recover all the edges of the graph with high probability. Under some conditions on the model covariance, we show that model selection can be achieved for sample sizes $n = \Omega(\max\{s, d^2\} \log(p))$, with the error decaying as $\mathcal{O}(\exp(-c \log(p)))$ for some constant c. Empirical simulations suggest that the rate is tight for graphs where the maximum node degree d scales linearly with p, but could be tightened for graphs with $d = \mathcal{O}(1)$.

M23 *Clustering via LP-based Stabilities*

Nikos Komodakis komod@csd.uoc.gr
University of Crete, Heraklion, Greece
Nikos Paragios nikos.paragios@ecp.fr
Ecole Centrale de Paris, Chatenay-Malabry, France
Georgios Tziritas tziritas@csd.uoc.gr
University of Crete, Heraklion, Greece

A novel center-based clustering algorithm is proposed in this paper. We first formulate clustering as an NP-hard linear integer program and we then use linear programming and the duality theory to derive the solution of this optimization problem. This leads to an efficient and very general algorithm, which works in the dual domain, and can cluster data based on an arbitrary set of distances. Despite its generality, it is independent of initialization (unlike EM-like methods such as K-means), has guaranteed convergence, can automatically determine the number of clusters, and can also provide online optimality bounds about the quality of the estimated clustering solutions. To deal with the most critical issue in a center-based clustering algorithm (selection of cluster centers), we also introduce the notion of stability of a cluster center, which is a well defined LP-based quantity that plays a key role to our algorithm's success. Furthermore, we also introduce, what we call, the margins (another key ingredient in our algorithm), which can be roughly thought of as dual counterparts to stabilities and allow us to obtain computationally efficient approximations to the latter. Promising experimental results demonstrate the potentials of our method.

M24 *Measures of Clustering Quality: A Working Set of Axioms for Clustering*

Shai Ben-David and Margareta Ackerman, University of Waterloo.
Oral presentation, Monday, 7:45pm. See abstract, page 55.

M25 *On the Reliability of Clustering Stability in the Large Sample Regime*

Ohad Shamir ohadsh@cs.huji.ac.il
Naftali Tishby tishby@cs.huji.ac.il
The Hebrew University, Jerusalem, Israel

Clustering stability is an increasingly popular family of methods for performing model selection in data clustering. The basic idea is that the chosen model should be stable under perturbation or resampling of the data. Despite being reasonably effective in practice, these methods are not well understood theoretically, and present some difficulties. In particular, when the data is assumed to be sampled from an underlying distribution, the solutions returned by the clustering algorithm will usually become more and more stable as the sample size increases. This raises a potentially serious practical difficulty with these methods, because it means there might be some hard-to-compute sample size, beyond which clustering stability estimators 'break down' and become unreliable in detecting the most stable model. Namely, all models will be relatively stable, with differences in their stability measures depending mostly on random and meaningless sampling artifacts.

In this paper, we provide a set of general sufficient conditions, which ensure the reliability of clustering stability estimators in the large sample regime. In contrast to previous work, which concentrated on specific toy distributions or specific idealized clustering frameworks, here we make no such assumptions. We then exemplify how these conditions apply to several important families of clustering algorithms, such as maximum likelihood clustering, certain types of kernel clustering, and centroid-based clustering with any Bregman divergence. In addition, we explicitly derive the non-trivial asymptotic behavior of these estimators, for any framework satisfying our conditions. This can help us understand what is considered a 'stable' model by these estimators, at least for large enough samples.

Spotlight presentation, Monday, 8:25pm.

M26 *Spectral Clustering with Perturbed Data*

LING HUANG ling.huang@intel.com
Intel Research Berkeley, CA, United States
DONGHUI YAN dhyan@stat.berkeley.edu
MICHAEL JORDAN jordan@cs.berkeley.edu
University of California, Berkeley, CA, United States
NINA TAFT nina.taft@intel.com
Intel Research Berkeley, CA, United States

Spectral clustering is useful for a wide-ranging set of applications in areas such as biological data analysis, image processing and data mining. However, the computational and/or communication resources required by the method in processing large-scale data sets are often prohibitively high, and practitioners are often required to perturb the original data in various ways (quantization, downsampling, etc) before invoking a spectral algorithm. In this paper, we use stochastic perturbation theory to study the effects of data perturbation on the performance of spectral clustering. We show that the error under perturbation of spectral clustering is closely related to the perturbation of the eigenvectors of the Laplacian matrix. From this result we derive approximate upper bounds on the clustering error. We show that this bound is tight empirically across a wide range of problems, suggesting that it can be used in practical settings to determine the amount of data reduction allowed in order to meet a specification of permitted loss in clustering performance.

Spotlight presentation, Monday, 8:25pm.

M27 *Influence of graph construction on graph-based clustering measures*

MARKUS MAIER, ULRIKE VON LUXBURG, MPI for Biological Cybernetics, and MATTHIAS HEIN, Saarland University.
Oral presentation, Monday, 8:05pm. *See abstract, page 55.*

M28 *Regularized Co-Clustering with Dual Supervision*

Vikas Sindhwani vsindhw@us.ibm.com
Jianying Hu jyhu@us.ibm.com
Aleksandra Mojsilovic aleksand@us.ibm.com
IBM Research, Yorktown Heights, NY, United States

By attempting to simultaneously partition both the rows (examples) and columns (features) of a data matrix, Co-clustering algorithms often demonstrate surprisingly impressive performance improvements over traditional one-sided (row) clustering techniques. A good clustering of features may be seen as a combinatorial transformation of the data matrix, effectively enforcing a form of regularization that may lead to a better clustering of examples (and vice-versa). In many applications, partial supervision in the form of a few row labels as well as column labels may be available to potentially assist co-clustering. In this paper, we develop two novel semi-supervised multi-class classification algorithms motivated respectively by spectral bipartite graph partitioning and matrix approximation (e.g., non-negative matrix factorization) formulations for co-clustering. These algorithms (i) support dual supervision in the form of labels for both examples and/or features, (ii) provide principled predictive capability on out-of-sample test data, and (iii) arise naturally from the classical Representer theorem applied to regularization problems posed on a collection of Reproducing Kernel Hilbert Spaces. Empirical results demonstrate the effectiveness and utility of our algorithms.

M29 *Cyclizing Clusters via Zeta Function of a Graph*

Deli Zhao dlzhao@ie.cuhk.edu.hk
Xiaoou Tang xtang@ie.cuhk.edu.hk
Department of Information Engineering, The Chinese University of Hong Kong, Hong Kong, China

Detecting underlying clusters from large-scale data plays a central role in machine learning research. In this paper, we tackle the problem of clustering complex data of multiple distributions and multiple scales. To this end, we develop an algorithm named Zeta l-links (Zell) which consists of two parts: Zeta merging with a similarity graph and an initial set of small clusters derived from local l-links of the graph. More specifically, we propose to structurize a cluster using cycles in the associated subgraph. A new mathematical tool, Zeta function of a graph, is introduced for the integration of all cycles, leading to a structural descriptor of a cluster in determinantal form. The popularity character of a cluster is conceptualized as the global fusion of variations of such a structural descriptor by means of the leave-one-out strategy in the cluster. Zeta merging proceeds, in the agglomerative fashion, according to the maximum incremental popularity among all pairwise clusters. Experiments on toy data clustering, imagery pattern clustering, and image segmentation show the competitive performance of Zell. The 98.1% accuracy, in the sense of the normalized mutual information (NMI), is obtained on the FRGC face data of 16028 samples and 466 facial clusters.

M30 *An Efficient Sequential Monte Carlo Algorithm for Coalescent Clustering*

Dilan Gorur dilan@gatsby.ucl.ac.uk
Yee Whye Teh ywteh@gatsby.ucl.ac.uk
Gatsby Computational Neuroscience Unit, UCL, London, United Kingdom

We propose an efficient sequential Monte Carlo inference scheme for the recently proposed coalescent clustering model (Teh et al, 2008). Our algorithm has a quadratic runtime while those in (Teh et al, 2008) is cubic. In experiments, we were surprised to find that in addition to being more efficient, it is also a better sampler than the best in (Teh et al, 2008), when measured in terms of variance of estimated likelihood and effective sample size.

M31 *Spectral Hashing*

Yair Weiss yweiss@cs.huji.ac.il
Hebrew University, Jerusalem, Israel
Antonio Torralba torralba@csail.mit.edu
CSAIL MIT, Cambridge, MA, United States
Rob Fergus fergus@cs.nyu.edu
Courant Institute - New York University, New York, NY, United States

Semantic hashing seeks compact binary codes of datapoints so that the Hamming distance between codewords correlates with semantic similarity. Hinton et al. used a clever implementation of autoencoders to find such codes. In this paper, we show that the problem of finding a best code for a given dataset is closely related to the problem of graph partitioning and can be shown to be NP hard. By relaxing the original problem, we obtain a spectral method whose solutions are simply a subset of thresh- olded eigenvectors of the graph Laplacian. By utilizing recent results on convergence of graph Laplacian eigenvectors to the Laplace-Beltrami eigen- functions of manifolds, we show how to efficiently calculate the code of a novel datapoint. Taken together, both learning the code and applying it to a novel point are extremely simple. Our experiments show that our codes significantly outperform the state-of-the art.

M32 *One sketch for all: Theory and Application of Conditional Random Sampling*

Ping Li pingli@cornell.edu
Cornell University, Ithaca, NY, United States
Kenneth W Church church@microsoft.com
Microsoft, Redmond, WA, United States
Trevor Hastie hastie@stanford.edu
Stanford University, Stanford, CA, United States

Conditional Random Sampling (CRS) was originally proposed for efficiently computing pairwise (l_2, l_1) distances, in *static*, *large-scale*, and *sparse* data sets such as text and Web data. It was previously presented using a heuristic argument.

This study extends CRS to handle *dynamic* or *streaming* data, which much better reflect the real-world situation than assuming static data. Compared with other known *sketching* algorithms for dimension reductions such as *stable random projections*, CRS exhibits a significant advantage in that it is "one-sketch-for-all." In particular, we demonstrate that CRS can be applied to efficiently compute the l_p distance and the Hilbertian metrics, both are popular in machine learning.

Although a fully rigorous analysis of CRS is difficult, we prove that, with a simple modification, CRS is rigorous at least for an important application of computing Hamming norms. A generic estimator and an approximate variance formula are provided and tested on various applications, for computing Hamming norms, Hamming distances, and χ^2 distances.

Spotlight presentation, Monday, 8:25pm.

M33 *Theory of matching pursuit*

Zakria Hussain z.hussain@cs.ucl.ac.uk
John Shawe-Taylor jst@cs.ucl.ac.uk
University College London, London, United Kingdom

We analyse matching pursuit for kernel principal components analysis by proving that the sparse subspace it produces is a sample compression scheme. We show that this bound is tighter than the KPCA bound of Shawe-Taylor et al [swck-05] and highly predictive of the size of the subspace needed to capture most of the variance in the data. We analyse a second matching pursuit algorithm called kernel matching pursuit (KMP) which does not correspond to a sample compression scheme. However, we give a novel bound that views the choice of subspace of the KMP algorithm as a compression scheme and hence provide a VC bound to upper bound its future loss. Finally we describe how the same bound can be applied to other matching pursuit related algorithms.

M34 *Rademacher Complexity Bounds for Non-I.I.D. Processes*

Mehryar Mohri mohri@cs.nyu.edu
Courant Institute of Mathematical Sciences, Google Research, New York, NY, United States
Afshin Rostamizadeh rostami@cs.nyu.edu
Courant Institute, New York University, New York, NY, United States

This paper presents the first data-dependent generalization bounds for non-i.i.d. settings based on the notion of Rademacher complexity. Our bounds extend to the non-i.i.d. case existing Rademacher complexity bounds derived for the i.i.d. setting. These bounds provide a strict generalization of the ones found in the i.i.d. case, and can also be used within the standard i.i.d. scenario. They apply to the standard scenario of beta-mixing stationary sequences examined in many previous studies of non-i.i.d. settings and benefit form the crucial advantages of Rademacher complexity over other measures of the complexity of hypothesis classes. In particular, they are data-dependent and measure the complexity of a class of hypotheses based on the training sample. The empirical Rademacher complexity can be estimated from finite samples and lead to tighter bounds.

M35 *On the Complexity of Linear Prediction: Risk Bounds, Margin Bounds, and Regularization*

Sham M Kakade sham@tti-c.org
Karthik Sridharan karthik@tti-c.org
Ambuj Tewari tewari@tti-c.org
Toyota Technological Institute, Chicago, IL, United States

We provide sharp bounds for Rademacher and Gaussian complexities of (constrained) linear classes. These bounds make short work of providing a number of corollaries including: risk bounds for linear prediction (including settings where the weight vectors are constrained by either L_2 or L_1 constraints), margin bounds (including both L_2 and L_1 margins, along with more general notions based on relative entropy), a proof of the PAC-Bayes theorem, and L_2 covering numbers (with L_p norm constraints and relative entropy constraints). In addition to providing a unified analysis, the results herein provide some of the sharpest risk and margin bounds (improving upon a number of previous results). Interestingly, our results show that the uniform convergence rates of empirical risk minimization algorithms tightly match the regret bounds of online learning algorithms for linear prediction (up to a constant factor of 2).

M36 *Risk Bounds for Randomized Sample Compressed Classifiers*

Mohak Shah mohak@cim.mcgill.ca
Centre for Intelligent Machines, McGill University, Montreal, QC, Canada

We derive risk bounds for the randomized classifiers in Sample Compression settings where the classifier-specification utilizes two sources of information viz. the compression set and the message string. By extending the recently proposed Occam's Hammer principle to the data-dependent settings, we derive point-wise versions of the bounds on the stochastic sample compressed classifiers and also recover the corresponding classical PAC-Bayes bound. We further show how these compare favorably to the existing results.

M37 *Efficient Direct Density Ratio Estimation for Non-stationarity Adaptation and Outlier Detection*

Takafumi Kanamori kanamori@is.nagoya-u.ac.jp
Graduate School of Information Science, Nagoya university, Nagoya, Japan
Shohei Hido hido@jp.ibm.com
IBM Research, Tokyo Research Laboratory, Yamato-shi, Japan
Masashi Sugiyama sugi@cs.titech.ac.jp
Department of Computer Science, Tokyo Institute of Technology, Tokyo, Japan

We address the problem of estimating the ratio of two probability density functions (a.k.a.∼the importance). The importance values can be used for various succeeding tasks such as non-stationarity adaptation or outlier detection. In this paper, we propose a new importance estimation method that has a closed-form solution; the leave-one-out cross-validation score can also be computed analytically. Therefore, the proposed method is computationally very efficient and numerically stable. We also elucidate theoretical properties of the proposed method such as the convergence rate and approximation error bound. Numerical experiments show that the proposed method is comparable to the best existing method in accuracy, while it is computationally more efficient than competing approaches.

M38 *An Homotopy Algorithm for the Lasso with Online Observations*

PIERRE GARRIGUES garrigue@eecs.berkeley.edu
LAURENT EL GHAOUI elghaoui@eecs.berkeley.edu
University of California at Berkeley, Berkeley, CA, United States

It has been shown that the problem of ℓ_1-penalized least-square regression commonly referred to as the Lasso or Basis Pursuit DeNoising leads to solutions that are sparse and therefore achieves model selection. We propose in this paper an algorithm to solve the Lasso with online observations. We introduce an optimization problem that allows us to compute an homotopy from the current solution to the solution after observing a new data point. We compare our method to Lars and present an application to compressed sensing with sequential observations. Our approach can also be easily extended to compute an homotopy from the current solution to the solution after removing a data point, which leads to an efficient algorithm for leave-one-out cross-validation.

M39 *Supervised Exponential Family Principal Component Analysis via Convex Optimization*

YUHONG GUO yuhongguo.cs@gmail.com
Australian National University, Canberra, Australia

Recently, supervised dimensionality reduction has been gaining attention, owing to the realization that data labels are often available and strongly suggest important underlying structures in the data. In this paper, we present a novel convex supervised dimensionality reduction approach based on exponential family PCA, which is able to avoid the local optima of the EM learning. Moreover, by introducing a sample-based empirical approximation to exponential family models, it avoids the limitation of the prevailing Gaussian assumptions of standard PCA, and produces a kernelized formulation for nonlinear supervised dimensionality reduction. A training algorithm is then devised based on a subgradient bundle method, whose scalability can be gained through a coordinate descent procedure. The advantage of our global optimization approach is demonstrated by empirical results over both synthetic and real data.

M40 *Convergence and Rate of Convergence of a Manifold-Based Dimension Reduction Algorithm*

ANDREW SMITH andrewsmith81@gmail.com
XIAOMING HUO xiaoming@isye.gatech.edu
HONGYUAN ZHA zha@cc.gatech.edu
Georgia Institute of Technology, Atlanta, GA, United States

We study the convergence and the rate of convergence of a local manifold learning algorithm: LTSA. The main technical tool is the perturbation analysis on the linear invariant subspace that corresponds to the solution of LTSA. We derive a worst-case upper bound of errors for LTSA which naturally leads to a convergence result. We then derive the rate of convergence for LTSA in a special case.

M41 *Diffeomorphic Dimensionality Reduction*

Christian Walder christian.walder@tuebingen.mpg.de
Bernhard Schölkopf bs@tuebingen.mpg.de
Max Planck Institute for Biological Cybernetics, Tübingen, Germany

This paper introduces a new approach to constructing meaningful lower dimensional representations of sets of data points. We argue that constraining the mapping between the high and low dimensional spaces to be a diffeomorphism is a natural way of ensuring that pairwise distances are approximately preserved. Accordingly we develop an algorithm which diffeomorphically maps the data near to a lower dimensional subspace and then projects onto that subspace. The problem of solving for the mapping is transformed into one of solving for an Eulerian flow field which we compute using ideas from kernel methods. We demonstrate the efficacy of our approach on various real world data sets.

M42 *Transfer Learning by Distribution Matching for Targeted Advertising*

Steffen Bickel bickel@cs.uni-potsdam.de
Christoph Sawade sawade@cs.uni-potsdam.de
Tobias Scheffer scheffer@cs.uni-potsdam.de
University of Potsdam, Potsdam, Germany

We address the problem of learning classifiers for several related tasks that may differ in their joint distribution of input and output variables. For each task, small – possibly even empty – labeled samples and large unlabeled samples are available. While the unlabeled samples reflect the target distribution, the labeled samples may be biased. This setting is motivated by the problem of predicting sociodemographic features for users of web portals, based on the content which they have accessed. Here, questionnaires offered to a portion of each portal's users produce biased samples. We derive a transfer learning procedure that produces resampling weights which match the pool of all examples to the target distribution of any given task. Transfer learning enables us to make predictions even for new portals with few or no training data and improves the overall prediction accuracy.

M43 *Relative Performance Guarantees for Approximate Inference in Latent Dirichlet Allocation*

Indraneel Mukherjee imukherj@cs.princeton.edu
David Blei blei@cs.princeton.edu
Princeton University, Princeton, NJ, United States

Hierarchical probabilistic modeling of discrete data has emerged as a powerful tool for text analysis. Posterior inference in such models is intractable, and practitioners rely on approximate posterior inference methods such as variational inference or Gibbs sampling. There has been much research in designing better approximations, but there is yet little theoretical understanding of which of the available techniques are appropriate, and in which data analysis settings. In this paper we provide the beginnings of such understanding. We analyze the improvement that the recently proposed collapsed variational inference (CVB) provides over mean field variational inference (VB) in latent Dirichlet allocation. We prove that the difference in the tightness of the bound on the likelihood of

a document decreases as $O(k-1)+\sqrt{\log m/m}$, where k is the number of topics in the model and m is the number of words in a document. As a consequence, the advantage of CVB over VB is lost for long documents but increases with the number of topics. We demonstrate empirically that the theory holds, using simulated text data and two text corpora. We provide practical guidelines for choosing an approximation.

Spotlight presentation, Monday, 8:25pm.

M44 *Unsupervised Learning of Visual Sense Models for Polysemous Words*

Kate Saenko saenko@mit.edu
MIT CSAIL, Cambridge, MA, United States
Trevor Darrell trevor@cs.berkeley.edu
EECS and ICSI, University of California, Berkeley, CA, United States

Polysemy is a problem for methods that exploit image search engines to build object category models. Existing unsupervised approaches do not take word sense into consideration. We propose a new method that uses a dictionary to learn models of visual word sense from a large collection of unlabeled web data. The use of LDA to discover a latent sense space makes the model robust despite the very limited nature of dictionary definitions. The definitions are used to learn a distribution in the latent space that best represents a sense. The algorithm then uses the text surrounding image links to retrieve images with high probability of a particular dictionary sense. An object classifier is trained on the resulting sense-specific images. We evaluate our method on a dataset obtained by searching the web for polysemous words. Category classification experiments show that our dictionary-based approach outperforms baseline methods.

Spotlight presentation, Monday, 8:25pm.

M45 *Gates*

Tom Minka minka@microsoft.com
John Winn jwinn@microsoft.com
Microsoft Research, Cambridge, United Kingdom

Gates are a new notation for representing mixture models and context-sensitive independence in factor graphs. Factor graphs provide a natural representation for message-passing algorithms, such as expectation propagation. However, message passing in mixture models is not well captured by factor graphs unless the entire mixture is represented by one factor, because the message equations have a containment structure. Gates capture this containment structure graphically, allowing both the independences and the message-passing equations for a model to be readily visualized. Different variational approximations for mixture models can be understood as different ways of drawing the gates in a model. We present general equations for expectation propagation and variational message passing in the presence of gates.

Spotlight presentation, Monday, 8:25pm.

M46 *Fast Computation of Posterior Mode in Multi-Level Hierarchical Models*

Liang Zhang lz9@stat.duke.edu
Department of Statistical Science, Duke University, Durham, NC, United States
Deepak Agarwal dagarwal@yahoo-inc.com
Yahoo! Research, Santa Clara, CA, United States

Multi-level hierarchical models provide an attractive framework for incorporating correlations induced in a response variable organized in a hierarchy. Model fitting is challenging, especially for hierarchies with large number of nodes. We provide a novel algorithm based on a multi-scale Kalman filter that is both scalable and easy to implement. For non-Gaussian responses, quadratic approximation to the log-likelihood results in biased estimates. We suggest a bootstrap strategy to correct such biases. Our method is illustrated through simulation studies and analyses of real world data sets in health care and online advertising.

Spotlight presentation, Monday, 8:25pm.

M47 *Evaluating probabilities under high-dimensional latent variable models*

Iain Murray murray@cs.toronto.edu
University of Toronto, Toronto, ON, Canada
Ruslan Salakhutdinov rsalakhu@cs.toronto.edu
University of Toronto, Toronto, Canada

We present a simple new Monte Carlo algorithm for evaluating probabilities of observations in complex latent variable models, such as Deep Belief Networks. While the method is based on Markov chains, estimates based on short runs are formally unbiased. In expectation, the log probability of a test set will be underestimated, and this could form the basis of a probabilistic bound. The method is much cheaper than gold-standard annealing-based methods and only slightly more expensive than the cheapest Monte Carlo methods. We give examples of the new method substantially improving simple variational bounds at modest extra cost.

Spotlight presentation, Monday, 8:25pm.

M48 *Counting Solution Clusters in Graph Coloring Problems Using Belief Propagation*

Lukas Kroc kroc@cs.cornell.edu
Ashish Sabharwal sabhar@cs.cornell.edu
Bart Selman selman@cs.cornell.edu
Cornell University, Ithaca, NY, United States

We show that an important and computationally challenging solution space feature of the graph coloring problem (COL), namely the number of clusters of solutions, can be accurately estimated by a technique very similar to one for counting the number of solutions. This cluster counting approach can be naturally written in terms of a new factor graph derived from the factor graph representing the COL instance. Using a variant of the Belief Propagation inference framework, we can efficiently approximate cluster counts in random COL problems over a large range of graph densities. We illustrate the algorithm on instances with up to 100,000 vertices. Moreover, we supply a methodology for computing the number of clusters exactly using advanced techniques from the knowledge compilation literature. This methodology scales up to several hundred variables.

M49 *Fast High-dimensional Kernel Summations Using the Monte Carlo Multipole Method*

Dongryeol Lee dongryel@cc.gatech.edu
Georgia Institute of Technology, Computational Science and Engineering, Atlanta, GA, United States
Alexander G Gray agray@cc.gatech.edu
Georgia Institute of Technology, College of Computing, Atlanta, GA, United States

We propose a new fast Gaussian summation algorithm for high-dimensional datasets with high accuracy. First, we extend the original fast multipole-type methods to use approximation schemes with both hard and probabilistic error. Second, we utilize a new data structure called subspace tree which maps each data point in the node to its lower dimensional mapping as determined by any linear dimension reduction method such as PCA. This new data structure is suitable for reducing the cost of each pairwise distance computation, the most dominant cost in many kernel methods. Our algorithm guarantees probabilistic relative error on each kernel sum, and can be applied to high-dimensional Gaussian summations which are ubiquitous inside many kernel methods as the key computational bottleneck. We provide empirical speedup results on low to high-dimensional datasets up to 89 dimensions.

M50 *Kernel Change-point Analysis*

Zaïd Harchaoui zaid.harchaoui@telecom-paristech.fr
LTCI, CNRS and TELECOM ParisTech, Paris, France
Francis Bach francis.bach@mines.org
INRIA - Ecole Normale Superieure, Paris, France
Eric Moulines eric.moulines@telecom-paristech.fr
LTCI, CNRS and TELECOM ParisTech, Paris, France

We introduce a kernel-based method for change-point analysis within a sequence of temporal observations. Change-point analysis of an (unlabelled) sample of observations consists in, first, testing whether a change in the distribution occurs within the sample, and second, if a change occurs, estimating the change-point instant after which the distribution of the observations switches from one distribution to another different distribution. We propose a test statistics based upon the maximum kernel Fisher discriminant ratio as a measure of homogeneity between segments. We derive its limiting distribution under the null hypothesis (no change occurs), and establish the consistency under the alternative

hypothesis (a change occurs). This allows to build a statistical hypothesis testing procedure for testing the presence of change-point, with a prescribed false-alarm probability and detection probability tending to one in the large-sample setting. If a change actually occurs, the test statistics also yields an estimator of the change-point location. Promising experimental results in temporal segmentation of mental tasks from BCI data and pop song indexation are presented.

M51 *Deep Learning with Kernel Regularization for Visual Recognition*

KAI YU kyu@sv.nec-labs.com
WEI XU xw@sv.nec-labs.com
YIHONG GONG ygong@sv.nec-labs.com
NEC Laboratories America, Inc., Cupertino, CA, United States

In this paper we aim to train deep neural networks for rapid visual recognition. The task is highly challenging, largely due to the lack of a meaningful regularizer on the functions realized by the networks. We propose a novel regularization method that takes advantage of kernel methods, where a given kernel function represents prior knowledge about the recognition task of interest. We derive an efficient algorithm using stochastic gradient descent, and demonstrate encouraging results on a wide range of recognition tasks, in terms of both accuracy and speed.

Spotlight presentation, Monday, 8:25pm.

M52 *Implicit Mixtures of Restricted Boltzmann Machines*

VINOD NAIR vnair@cs.toronto.edu
University of Toronto, Toronto, ON, Canada
GEOFFREY E HINTON hinton@cs.toronto.edu
University of Toronto, Toronto, Canada

We present a mixture model whose components are Restricted Boltzmann Machines (RBMs). This possibility has not been considered before because computing the partition function of an RBM is intractable, which appears to make learning a mixture of RBMs intractable as well. Surprisingly, when formulated as a third-order Boltzmann machine, such a mixture model *can* be learned tractably using contrastive divergence. The energy function of the model captures three-way interactions among visible units, hidden units, and a single hidden multinomial unit that represents the cluster labels. The distinguishing feature of this model is that, unlike other mixture models, the mixing proportions are not explicitly parameterized. Instead, they are defined implicitly via the energy function and depend on all the parameters in the model. We present results for the MNIST and NORB datasets showing that the implicit mixture of RBMs learns clusters that reflect the class structure in the data.

M53 *Estimation of Information Theoretic Measures for Continuous Random Variables*

FERNANDO PEREZ-CRUZ fp@princeton.edu
Princeton University, Princeton, NJ, United States
University Carlos III, Madrid, Leganes Spain

We analyze the estimation of information theoretic measures of continuous random variables such as: differential entropy, mutual information or Kullback-Leibler divergence. The objective of this paper is two-fold. First, we prove that the information theoretic measure estimates using the k-nearest-neighbor density estimation with fixed k converge almost surely, even though the k-nearest-neighbor density estimation with fixed k does not converge to its true measure. Second, we show that the information theoretic measure estimates do not converge for k growing linearly with the number of samples. Nevertheless, these nonconvergent estimates can be used for solving the two-sample problem and assessing if two random variables are independent. We show that the two-sample and independence tests based on these nonconvergent estimates compare favorably with the maximum mean discrepancy test and the Hilbert Schmidt independence criterion, respectively.

M54 *Privacy-preserving logistic regression*

KAMALIKA CHAUDHURI kamalika@soe.ucsd.edu
CLAIRE MONTELEONI cmontel@cs.ucsd.edu
University of California San Diego, La Jolla, CA, United States

This paper addresses the important tradeoff between privacy and learnability, when designing algorithms for learning from private databases. We focus on privacy-preserving logistic regression. First we apply an idea of Dwork et al. to design a privacy-preserving logistic regression algorithm. This involves bounding the sensitivity of regularized logistic regression, and perturbing the learned classifier with noise proportional to the sensitivity. We then show, both theoretically and experimentally, that for certain data distributions, this algorithm has poor learning generalization, compared with standard regularized logistic regression.

We then provide a second privacy-preserving regularized logistic regression algorithm. Our algorithm is based on a new privacy-preserving technique: solving a perturbed optimization problem. We prove that our algorithm preserves privacy in the model due to Dwork et al., and we provide learning guarantees. We show that our algorithm performs almost as well as standard regularized logistic regression, in terms of generalization error. Our privacy-preserving technique may also be of more general interest, as it can be extended to work with any convex loss function that is differentiable at all points and Lipschitz, and it does not depend on the sensitivity of the function. Our work also reveals an interesting connection between regularization and privacy.

M55 *Robust Near-Isometric Matching via Structured Learning of Graphical Models*

Julian McAuley julian.mcauley@nicta.com.au
Tiberio Caetano tiberio.caetano@nicta.com.au
NICTA, Canberra, Australia
Alex J. Smola alex@smola.org
Yahoo! Research, Santa-Clara, CA, United States

Models for near-rigid shape matching are typically based on distance-related features, in order to infer matches that are consistent with the isometric assumption. However, real shapes from image datasets, even when expected to be related by "almost isometric" transformations, are actually subject not only to noise but also, to some limited degree, to variations in appearance and scale. In this paper, we introduce a graphical model that parameterises appearance, distance, and angle features and we learn all of the involved parameters via structured prediction. The outcome is a model for near-rigid shape matching which is robust in the sense that it is able to capture the possibly limited but still important scale and appearance variations. Our experimental results reveal substantial improvements upon recent successful models, while maintaining similar running times.

M56 *Shape-Based Object Localization for Descriptive Classification*

Geremy Heitz gaheitz@stanford.edu
Gal Elidan galel@cs.stanford.edu
Benjamin Packer bpacker@cs.stanford.edu
Daphne Koller koller@cs.stanford.edu
Stanford University, Stanford, CA, United States

Discriminative tasks, including object categorization and detection, are central components of high-level computer vision. Sometimes, however, we are interested in more refined aspects of the object in an image, such as pose or particular regions. In this paper we develop a method (LOOPS) for learning a shape and image feature model that can be trained on a particular object class, and used to outline instances of the class in novel images. Furthermore, while the training data consists of uncorresponded outlines, the resulting LOOPS model contains a set of landmark points that appear consistently across instances, and can be accurately localized in an image. The resulting localization can then be used to address a range of tasks, including descriptive classification, search, and clustering.

M57 *Supervised Dictionary Learning*

Julien Mairal julien.mairal@inria.fr
Francis Bach francis.bach@mines.org
INRIA - Ecole Normale Superieure, Paris, France
Jean Ponce jean.ponce@ens.fr
Ecole Normale Superieure, Paris, France
Guillermo Sapiro guille@ece.umn.edu
University of Minnesota, Minneapolis, USA
Andrew Zisserman az@robots.ox.ac.uk
University of Oxford, Oxford, United Kingdom

It is now well established that sparse signal models are well suited to restoration tasks and can effectively be learned from audio, image, and video data. Recent research has been aimed at learning discriminative sparse models instead of purely reconstructive ones. This paper proposes a new step in that direction with a novel sparse representation for signals belonging to different classes in terms of a shared dictionary and multiple decision functions. It is shown that the linear variant of the model admits a simple probabilistic interpretation, and that its most general variant also admits a simple interpretation in terms of kernels. An optimization framework for learning all the components of the proposed model is presented, along with experiments on standard handwritten digit and texture classification tasks.

M58 *A "Shape Aware" Model for semi-supervised Learning of Objects and its Context*

Abhinav Gupta agupta@cs.umd.edu
University of Maryland, College Park, College Park, MD, United States
Jianbo Shi jshi@cis.upenn.edu
University of Pennsylvania, Philadelphia, PA, United States
Larry Davis lsd@cs.umd.edu
University of Maryland, College Park, College Park, MD, United States

Integrating semantic and syntactic analysis is essential for document analysis. Using an analogous reasoning, we present an approach that combines bag-of-words and spatial models to perform semantic and syntactic analysis for recognition of an object based on its internal appearance and its context. We argue that while object recognition requires modeling relative spatial locations of image features within the object, a bag-of-word is sufficient for representing context.

Learning such a model from weakly labeled data involves labeling of features into two classes: foreground(object) or "informative" background(context). labeling. We present a "shape-aware" model which utilizes contour information for efficient and accurate labeling of features in the image. Our approach iterates between an MCMC-based labeling and contour based labeling of features to integrate co-occurrence of features and shape similarity.

Spotlight presentation, Monday, 8:25pm.

M59 *Multiscale Random Fields with Application to Contour Grouping*

Longin Jan Latecki latecki@temple.edu
Department of CIS, Temple University, Philadelphia, PA, United States
ChengEn Lu luchengen@gmail.com
Department of Electronics and Information Engineering, Huazhong University of Science and Technology, Wuhan National Laboratory for Optoelectronics, Wuhan, China
Marc Sobel marc.sobel@temple.edu
Temple University, Philadelphia, PA, United States
Xiang Bai xiang.bai@gmail.com
Department of Electronics and Information Engineering, Huazhong University of Science and Technology, Wuhan National Laboratory for Optoelectronics, Wuhan, China

We introduce a new interpretation of multiscale random fields (MSRFs) that admits efficient optimization in the framework of regular (single level) random fields (RFs). It is based on a new operator, called append, that combines sets of random variables (RVs) to single RVs. We assume that a MSRF can be decomposed into disjoint trees that link RVs at different pyramid levels. The append operator is then applied to map RVs in each tree structure to a single RV. We demonstrate the usefulness of the proposed approach on a challenging task involving grouping contours of target shapes in images. MSRFs provide a natural representation of multiscale contour models, which are needed in order to cope with unstable contour decompositions. The append operator allows us to find optimal image labels using the classical framework of relaxation labeling, Alternative methods like Markov Chain Monte Carlo (MCMC) could also be used.

M60 *Adaptive Template Matching with Shift-Invariant Semi-NMF*

Jonathan Le Roux leroux@hil.t.u-tokyo.ac.jp
The University of Tokyo, Tokyo, Japan
Alain de Cheveigné alain.de.cheveigne@ens.fr
CNRS, Université Paris Descartes, ENS, Paris, France
Lucas C. Parra parra@ccny.cuny.edu
Department of Biomedical Engineering, City University of New York, New York, NY, United States

How does one extract unknown but stereotypical events that are linearly superimposed within a signal with variable latencies and variable amplitudes? One could think of using template matching or matching pursuit to find the arbitrarily shifted linear components. However, traditional matching approaches require that the templates be known *a priori*. To overcome this restriction we use instead semi Non-Negative Matrix Factorization (semi-NMF) that we extend to allow for time shifts when matching the templates to the signal. The algorithm estimates templates directly from the data along with their non-negative amplitudes. The resulting method can be thought of as an adaptive template matching procedure. We demonstrate the procedure on the task of extracting spikes from single channel extracellular recordings. On these data the algorithm essentially performs spike detection and unsupervised spike clustering. Results on simulated data and extracellular recordings indicate that the method performs well for signal-to-noise ratios of 6dB or higher and that spike templates are recovered accurately provided they are sufficiently different.

M61 *Hierarchical Fisher Kernels for Longitudinal Data*

Lu Zhengdong zhengdon@csee.ogi.edu
Todd Leen tleen@csee.ogi.edu
Dept. of Computer Science & Engineering, Oregon Health & Science University, Beaverton, OR, United States
Jeffrey Kaye kaye@ohsu.edu
OHSU, Oregon Center for Aging & Technology, Portland, OR, United States

We develop new techniques for time series classification based on hierarchical Bayesian generative models (called mixed-effect models) and the Fisher kernel derived from them. A key advantage of the new formulation is that one can compute the Fisher information matrix despite varying sequence lengths and sampling times. We therefore can avoid the ad hoc replacement of Fisher information matrix with the identity matrix commonly used in literature, which destroys the geometrical grounding of the kernel construction. In contrast, our construction retains the proper geometric structure resulting in a kernel that is properly invariant under change of coordinates in the model parameter space. Experiments on detecting cognitive decline show that classifiers based on the proposed kernel out-perform those based on generative models and other feature extraction routines.

M62 *Accelerating Bayesian Inference over Nonlinear Differential Equations with Gaussian Processes*

BEN CALDERHEAD bc@dcs.gla.ac.uk
MARK GIROLAMI girolami@dcs.gla.ac.uk
Department of Computing Science, University of Glasgow, United Kingdom
NEIL D LAWRENCE neill@cs.man.ac.uk
School of Computer Science, University of Manchester, United Kingdom

Identification and comparison of nonlinear dynamical systems using noisy and sparse experimental data is a vital task in many fields, however current methods are computationally expensive and prone to error due in part to the nonlinear nature of the likelihood surfaces induced. We present an accelerated sampling procedure which enables Bayesian inference of parameters in nonlinear ordinary and delay differential equations via the novel use of Gaussian processes (GP). Our method involves GP regression over time-series data, and the resulting derivative and time delay estimates make parameter inference possible without solving the dynamical system explicitly, resulting in dramatic savings of computational time. We demonstrate the speed and statistical accuracy of our approach using examples of both ordinary and delay differential equations, and provide a comprehensive comparison with current state of the art methods.

M63 *Hierarchical Semi-Markov Conditional Random Fields for Recursive Sequential Data*

THE TRUYEN TRAN thetruyen.tran@postgrad.curtin.edu.au
DINH PHUNG d.phung@curtin.edu.au
Dept of Computing, Curtin University of Technology, Bentley, WA, Australia
HUNG BUI bui@ai.sri.com
SRI International, Menlo Park, CA, United States
SVETHA VENKATESH s.venkatesh@curtin.edu.au
Dept of Computing, Curtin University of Technology, Bentley, WA, Australia

Inspired by the hierarchical hidden Markov models (HHMM), we present the hierarchical semi-Markov conditional random field (HSCRF), a generalisation of embedded undirected Markov chains to model complex hierarchical, nested Markov processes. It is parameterised in a discriminative framework and has polynomial time algorithms for learning and inference. Importantly, we consider partially-supervised learning and propose algorithms for generalised partially-supervised learning and constrained inference. We demonstrate the HSCRF in two applications: (i) recognising human activities of daily living (ADLs) from indoor surveillance cameras, and (ii) noun-phrase chunking. We show that the HSCRF is capable of learning rich hierarchical models with reasonable accuracy in both fully and partially observed data cases.

M64 *Non-stationary dynamic Bayesian networks*

Joshua W Robinson josh@cs.duke.edu
Alexander J Hartemink amink@cs.duke.edu
Department of Computer Science, Duke University, Durham, NC, United States

A principled mechanism for identifying conditional dependencies in time-series data is provided through structure learning of dynamic Bayesian networks (DBNs). An important assumption of DBN structure learning is that the data are generated by a stationary process—an assumption that is not true in many important settings. In this paper, we introduce a new class of graphical models called non-stationary dynamic Bayesian networks, in which the conditional dependence structure of the underlying data-generation process is permitted to change over time. Non-stationary dynamic Bayesian networks represent a new framework for studying problems in which the structure of a network is evolving over time. We define the non-stationary DBN model, present an MCMC sampling algorithm for learning the structure of the model from time-series data under different assumptions, and demonstrate the effectiveness of the algorithm on both simulated and biological data.

M65 *Local Gaussian Process Regression for Real Time Online Model Learning*

Duy Nguyen-Tuong duy.nguyen-tuong@tuebingen.mpg.de
Max Planck Institute for Biological Cybernetics, Tuebingen, Germany
Matthias Seeger matthias.seeger@tuebingen.mpg.de
Jan Peters mail@jan-peters.net
MPI for Biological Cybernetics, Tuebingen, Germany

Learning in real-time applications, e.g., online approximation of the inverse dynamics model for model-based robot control, requires fast online regression techniques. Inspired by local learning, we propose a method to speed up standard Gaussian Process regression (GPR) with local GP models (LGP). The training data is partitioned in local regions, for each an individual GP model is trained. The prediction for a query point is performed by weighted estimation using nearby local models. Unlike other GP approximations, such as mixtures of experts, we use a distance based measure for partitioning of the data and weighted prediction. The proposed method achieves online learning and prediction in real-time. Comparisons with other nonparametric regression methods show that LGP has higher accuracy than LWPR and close to the performance of standard GPR and nu-SVR.

M66 *Online Prediction on Large Diameter Graphs*

Mark Herbster m.herbster@cs.ucl.ac.uk
Guy Lever g.lever@cs.ucl.ac.uk
Massimiliano Pontil m.pontil@cs.ucl.ac.uk
Department of Computer Science, University College London, London, United Kingdom

Current online learning algorithms for predicting the labelling of a graph have an important limitation in the case of large diameter graphs: the number of mistakes made by such algorithms may be proportional to the square root of the number of vertices, even when tackling simple problems. We overcome this problem with an efficient algorithm which achieves a logarithmic mistake bound. Furthermore, current algorithms are optimised for data which exhibits cluster-structure; we give an additional algorithm which performs well both in the presence of cluster structure and on large diameter graphs.

M67 *Mortal Multi-Armed Bandits*

Deepayan Chakrabarti deepay@yahoo-inc.com
Yahoo! Research, Mountain View, CA, United States
Ravi Kumar ravikumar@yahoo-inc.com
Yahoo! Research, Sunnyvale, CA, United States
Filip Radlinski filiprad@microsoft.com
Microsoft Research, Cambridge, United Kingdom
Eli Upfal eli@cs.brown.edu
Brown University, Providence, RI, United States

We formulate and study a new variant of the k-armed bandit problem, motivated by e-commerce applications. In our model, arms have (stochastic) lifetime after which they expire. In this setting an algorithm needs to continuously explore new arms, in contrast to the standard k-armed bandit model in which arms are available indefinitely and exploration is reduced once an optimal arm is identified with near-certainty. The main motivation for our setting is online-advertising, where ads have limited lifetime due to, for example, the nature of their content and their campaign budget. An algorithm needs to choose among a large collection of ads, more than can be fully explored within the ads' lifetime. We present an optimal algorithm for the state-aware (deterministic reward function) case, and build on this technique to obtain an algorithm for the state-oblivious (stochastic reward function) case. Empirical studies on various reward distributions, including one derived from a real-world ad serving application, show that the proposed algorithms significantly outperform the standard multi-armed bandit approaches applied to these settings.

M68 *Biasing Approximate Dynamic Programming with a Lower Discount Factor*

Marek Petrik petrik@cs.umass.edu
University of Massachusetts Amherst, Amherst, MA, United States
Bruno Scherrer bruno.scherrer@loria.fr
INRIA Lorraine, Vandoeuvre les Nancy, France

Most algorithms for solving Markov decision processes rely on a discount factor, which ensures their convergence. It is generally assumed that using an artificially low discount factor will improve the convergence rate, while sacrificing the solution quality. We however demonstrate that using an artificially low discount factor may significantly improve the solution quality, when used in approximate dynamic programming. We propose two explanations of this phenomenon. The first justification follows directly from the standard approximation error bounds: using a lower discount factor may decrease the approximation error bounds. However, we also show that these bounds are loose, a thus their decrease does not entirely justify the improved solution quality. We thus propose another justification: when the rewards are received only sporadically (as in the case of Tetris), we can derive tighter bounds, which support a significant improvement in the solution quality with a decreased discount factor.

M69 *Particle Filter-based Policy Gradient in POMDPs*

Pierre-Arnaud Coquelin pierre-arnaud.coquelin@predictandcontrol.com
INRIA Lille - Nord Europe, SequeL project, France
Romain Deguest rd2304@columbia.edu
Ecole Polytechnique, CMAP, France, and Columbia University, New York, United States
Remi Munos remi.munos@inria.fr
INRIA Lille - Nord Europe, SequeL project, France

Our setting is a Partially Observable Markov Decision Process with continuous state, observation and action spaces. Decisions are based on a Particle Filter for estimating the belief state given past observations. We consider a policy gradient approach for parameterized policy optimization. For that purpose, we investigate sensitivity analysis of the performance measure with respect to the parameters of the policy, focusing on Finite Difference (FD) techniques. We show that the naive FD is subject to variance explosion because of the non-smoothness of the resampling procedure. We propose a more sophisticated FD method which overcomes this problem and establish its consistency.

M70 *MDPs with Non-Deterministic Policies*

Mahdi Milani Fard mahdi.milanifard@mail.mcgill.ca
Joelle Pineau jpineau@cs.mcgill.ca
McGill University, Montreal, QC, Canada

Markov Decision Processes (MDPs) have been extensively studied and used in the context of planning and decision-making, and many methods exist to find the optimal policy for problems modelled as MDPs. Although finding the optimal policy is sufficient in many domains, in certain applications such as decision support systems where the policy is executed by a human (rather than a machine), finding all possible near-optimal policies might be useful as it provides more flexibility to the person executing the policy. In this paper we introduce the new concept of non-deterministic MDP policies, and address the question of finding near-optimal non-deterministic policies. We propose two solutions to this problem, one based on a Mixed Integer Program and the other one based on a search algorithm. We include experimental results obtained from applying this framework to optimize treatment choices in the context of a medical decision support system.

M71 *A Convergent $O(n)$ Temporal-difference Algorithm for Off-policy Learning with Linear Function Approximation*

Richard S Sutton rich@richsutton.com
University of Alberta, Edmonton, Canada
Csaba Szepesvari szepesva@cs.ualberta.ca
University of Alberta, Edmonton, AB, Canada
Hamid Reza Maei maei@cs.ualberta.ca
Department of Computing Science, University of Alberta, Edmonton, AB, Canada

We introduce the first temporal-difference learning algorithm that is stable with linear function approximation and off-policy training, for any finite Markov decision process, target policy, and exciting behavior policy, and whose complexity scales linearly in the number of parameters. We consider an i.i.d. policy-evaluation setting in which the data need not come from on-policy experience. The gradient temporal-difference (GTD) algorithm estimates the expected update vector of the TD(0) algorithm and performs stochastic gradient descent on its L_2 norm. Our analysis proves that its expected update is in the direction of the gradient, assuring convergence under the usual stochastic approximation conditions to the same least-squares solution as found by the LSTD, but without its quadratic computational complexity. GTD is online and incremental, and does not involve multiplying by products of likelihood ratios as in importance-sampling methods.

M72 *Bounding Performance Loss in Approximate MDP Homomorphisms*

Jonathan Taylor jonathan.taylor@utoronto.ca
Department of Computer Science, University of Toronto, Toronto, ON, Canada
Prakash Panangaden prakash@cs.mcgill.ca
Doina Precup dprecup@cs.mcgill.ca
School of Computer Science, McGill University, Montreal, QC, Canada

We define a metric for measuring behavior similarity between states in a Markov decision process (MDP), in which action similarity is taken into account. We show that the kernel of our metric corresponds exactly to the classes of states defined by MDP homomorphisms (Ravindran & Barto, 2003). We prove that the difference in the optimal value function of different states can be upper-bounded by the value of this metric, and that the bound is tighter than that provided by bisimulation metrics (Ferns et al. 2004, 2005). Our results hold both for discrete and for continuous actions. We provide an algorithm for constructing approximate homomorphisms, by using this metric to identify states that can be grouped together, as well as actions that can be matched. Previous research on this topic is based mainly on heuristics.

M73 *Skill Characterization Based on Betweenness*

Ozgur Simsek ozgur@cs.umass.edu
Andrew Barto barto@cs.umass.edu
University of Massachusetts Amherst, Amherst, MA, United States

We present a characterization of a useful class of skills based on a graphical representation of an agent's interaction with its environment. Our characterization uses betweenness, a measure of centrality on graphs. It may be used directly to form a set of skills suitable for a given environment. More importantly, it serves as a useful guide for developing online, incremental skill discovery algorithms that do not rely on knowing or representing the environment graph in its entirety.

M74 *Bayesian Model of Behaviour in Economic Games*

Debajyoti Ray dray@caltech.edu
Caltech, Pasadena, CA, United States
Brooks King-Casas bkcasas@cpu.bcm.edu
P. Read Montague rmontague@hnl.bcm.edu
Baylor College of Medicine, Houston. Texas.
Peter Dayan dayan@gatsby.ucl.ac.uk
Gatsby Computational Neuroscience Unit, London, United Kingdom

Classical Game Theoretic approaches that make strong rationality assumptions have difficulty modeling observed behaviour in Economic games of human subjects. We investigate the role of finite levels of iterated reasoning and non-selfish utility functions in a Partially Observable Markov Decision Process model that incorporates Game Theoretic notions of interactivity. Our generative model captures a broad class of characteristic behaviours in a multi-round Investment game. We invert the generative process for a recognition model that is used to classify 200 subjects playing an Investor-Trustee game against randomly matched opponents.

M75 *Temporal Difference Based Actor Critic Learning - Convergence and Neural Implementation*

Dotan Di Castro dot@tx.technion.ac.il
Dmitry Volkinshtein dmitryv@tx.technion.ac.il
Ron Meir rmeir@ee.technion.ac.il
Dept. of Electrical Engineering, Technion, Haifa, Israel

Actor-critic algorithms for reinforcement learning are achieving renewed popularity due to their good convergence properties in situations where other approaches often fail (e.g., when function approximation is involved). Interestingly, there is growing evidence that actor-critic approaches based on transient dopamine signals play a key role in biological learning through the cortical and basal ganglia loops. We derive a temporal difference based actor critic learning algorithm, for which convergence can be proved without assuming separate time scales for the actor and the critic. The approach is demonstrated by applying it to networks of spiking neurons. The established relation between the transient dopamine signal and the temporal difference signal lends support to the biological relevance of such algorithms.

M76 *On the asymptotic equivalence between differential Hebbian and temporal difference learning using a local third factor*

Christoph Kolodziejski kolo@bccn-goettingen.de
Bernstein Center for Computational Neuroscience, Göttingen, Germany,

and Georg-August-University Göttingen, Göttingen, Germany
Bernd Porr b.porr@elec.gla.ac.uk
University of Glasgow, Glasgow, United Kingdom
Minija Tamosiunaite m.tamosiunaite@if.vdu.lt
Bernstein Center for Computational Neuroscience, Göttingen, Germany,
Georg-August-University Göttingen, Göttingen, Germany,
and Vytautas Magnus University, Kaunas, Lithuania
Florentin Wörgötter worgott@bccn-goettingen.de
Bernstein Center for Computational Neuroscience, Göttingen, Germany,
and Georg-August-University Göttingen, Göttingen, Germany

In this theoretical contribution we provide mathematical proof that two of the most important classes of network learning - correlation-based differential Hebbian learning and reward-based temporal difference learning - are asymptotically equivalent when timing the learning with a local modulatory signal. This opens the opportunity to consistently reformulate most of the abstract reinforcement learning framework from a correlation based perspective that is more closely related to the biophysics of neurons.

M77 *An ideal observer model of infant object perception*

Charles Kemp ckemp@cmu.edu
Carnegie Mellon University, Pittsburgh, PA, United States
Fei Xu fei@psych.ubc.ca
University of British Columbia, Vancouver, BC, Canada

Before the age of 4 months, infants make inductive inferences about the motions of physical objects. Developmental psychologists have provided verbal accounts of the knowledge that supports these inferences, but often these accounts focus on categorical rather than probabilistic principles. We propose that infant object perception is guided in part by probabilistic principles like persistence: things tend to remain the same, and when they change they do so gradually. To illustrate this idea, we develop an ideal observer model that includes probabilistic formulations of rigidity and inertia. Like previous researchers, we suggest that rigid motions are expected from an early age, but we challenge the previous claim that expectations consistent with inertia are relatively slow to develop (Spelke et al., 1992). We support these arguments by modeling four experiments from the developmental literature.

M78 *Temporal Dynamics of Cognitive Control*

Jeremy Reynolds jeremy.reynolds@colorado.edu
University of Colorado, Denver, United States
Michael Mozer mozer@colorado.edu
University of Colorado, Boulder, CO, United States

Cognitive control refers to the flexible deployment of memory and attention in response to task demands and current goals. Control is often studied experimentally by presenting sequences of stimuli, some demanding a response, and others modulating the stimulus-response mapping. In these tasks, participants must maintain information about the current stimulus-response mapping in working memory. Prominent theories of cognitive control use recurrent neural nets to implement working memory, and optimize memory utilization via reinforcement learning. We present a novel perspective on cognitive control in which working memory representations are intrinsically probabilistic, and control operations that maintain and update working memory are dynamically determined via probabilistic inference. We show that our model provides a parsimonious account of behavioral and neuroimaging data, and suggest that it offers an elegant conceptualization of control in which behavior can be cast as optimal, subject to limitations on learning and the rate of information processing. Moreover, our model provides insight into how task instructions can be directly translated into appropriate behavior and then efficiently refined with subsequent task experience.

M79 *Modeling human function learning with Gaussian processes*

Thomas L Griffiths tom_griffiths@berkeley.edu
Chris Lucas clucas@berkeley.edu
Joseph Williams joseph_williams@berkeley.edu
University of California, Berkeley, Berkeley, CA, United States
Michael L Kalish kalish@louisiana.edu
Institute of Cognitive Science, Lafayette, LA, United States

Accounts of how people learn functional relationships between continuous variables have tended to focus on two possibilities: that people are estimating explicit functions, or that they are simply performing associative learning supported by similarity. We provide a rational analysis of function learning, drawing on work on regression in machine learning and statistics. Using the equivalence of Bayesian linear regression and Gaussian processes, we show that learning explicit rules and using similarity can be seen as two views of one solution to this problem. We use this insight to define a Gaussian process model of human function learning that combines the strengths of both approaches.

M80 *Generative versus discriminative training of RBMs for classification of fMRI images*

Tanya Schmah schmah@cs.toronto.edu
Geoffrey E Hinton hinton@cs.toronto.edu
Richard Zemel zemel@cs.toronto.edu
University of Toronto, Toronto, ON, Canada
Steven L Small small@uchicago.edu
Department of Neurology, The University of Chicago, Chicago, IL, United States
Stephen Strother sstrother@rotman-baycrest.on.ca
Rotman Research Institute & University of Toronto, Toronto, ON, Canada

Neuroimaging datasets often have a very large number of voxels and a very small number of training cases, which means that overfitting of models for this data can become a very serious problem. Working with a set of fMRI images from a study on stroke recovery, we consider a classification task for which logistic regression performs poorly, even when L1- or L2- regularised. We show that much better discrimination can be achieved by fitting a generative model to each separate condition and then seeing which model is most likely to have generated the data. We use discriminative fitting of exactly the same set of models to demonstrate that the superior discrimination performance is caused by the generative fitting rather than the type of model. We used restricted Boltzmann machines as our generative models, but our results suggest that many other generative models should be tried for discriminating different conditions in neuroimaging data.

M81 *Short-Term Depression in VLSI Stochastic Synapse*

Peng Xu pxu@umd.edu
Timothy K Horiuchi timmerino@gmail.com
Pamela A Abshire pabshire@umd.edu
Department of Electrical & Computer Engineering, Institute for Systems Research, University of Maryland, College Park, MD, United States

We report a compact realization of short-term depression (STD) in a VLSI stochastic synapse. The behavior of the circuit is based on a subtractive single release model of STD. Experimental results agree well with simulation and exhibit expected STD behavior: the transmitted spike train has negative autocorrelation and lower power spectral density at low frequencies which can remove redundancy in the input spike train, and the mean transmission probability is inversely proportional to the input spike rate which has been suggested as an automatic gain control mechanism in neural systems. The dynamic stochastic synapse could potentially be a powerful addition to existing deterministic VLSI spiking neural systems.

M82 *Understanding Brain Connectivity Patterns during Motor Imagery for Brain-Computer Interfacing*

Moritz Grosse-Wentrup moritz.grosse-wentrup@tuebingen.mpg.de
MPI for Biological Cybernetics, Tuebingen, Germany

EEG connectivity measures could provide a new type of feature space for inferring a subject's intention in Brain-Computer Interfaces (BCIs). However, very little is known on EEG connectivity patterns for BCIs. In this study, EEG connectivity during motor imagery (MI) of the left and right is investigated in a broad frequency range across the whole scalp by combining Beamforming with Transfer Entropy and taking into account possible volume conduction effects. Observed connectivity patterns indicate that modulation intentionally induced by MI is strongest in the gamma-band, i.e., above 35 Hz. Furthermore, modulation between MI and rest is found to be more pronounced than between MI of different hands. This is in contrast to results on MI obtained with bandpower features, and might provide an explanation for the so far only moderate success of connectivity features in BCIs. It is concluded that future studies on connectivity based BCIs should focus on high frequency bands and consider experimental paradigms that maximally vary cognitive demands between conditions.

M83 *Tracking Changing Stimuli in Continuous Attractor Neural Networks*

C. C. Alan Fung alanfung@ust.hk
K. Y. Michael Wong phkywong@ust.hk
Department of Physics, Hong Kong University of Science and Technology, Hong Kong, China
Si Wu siwu@ion.ac.cn
Institute of Neuroscience, Shanghai, China

Continuous attractor neural networks (CANNs) are emerging as promising models for describing the encoding of continuous stimuli in neural systems. Due to the translational invariance of their neuronal interactions, CANNs can hold a continuous family of neutrally stable states. In this study, we systematically explore how neutral stability of a CANN facilitates its tracking performance, a capacity believed to have wide applications in brain functions. We develop a perturbative approach that utilizes the dominant movement of the network stationary states in the state space. We quantify the distortions of the bump shape during tracking, and study their effects on the tracking performance. Results are obtained on the maximum speed for a moving stimulus to be trackable, and the reaction time to catch up an abrupt change in stimulus.

M84 *Designing neurophysiology experiments to optimally constrain receptive field models along parametric submanifolds.*

Jeremy Lewi jeremy@lewi.us
Robert Butera rbutera@ece.gatech.edu
Georgia Tech, Atlanta, GA, United States
Liam Paninski liam@stat.columbia.edu
Columbia University, NY, NY, United States

Sequential optimal design methods hold great promise for improving the efficiency of neurophysiology experiments. However, previous methods for optimal experimental design have incorporated only weak prior information about the underlying neural system (e.g., the sparseness or smoothness of the receptive field). Here we describe how to use stronger prior information, in the form of parametric models of the receptive field, in order to construct optimal stimuli and further improve the efficiency of our experiments. For example, if we believe that the receptive field is well-approximated by a Gabor function, then our method constructs stimuli that optimally constrain the Gabor parameters (orientation, spatial frequency, etc.) using as few experimental trials as possible. More generally, we may believe a priori that the receptive field lies near a known sub-manifold of the full parameter space; in this case, our method chooses stimuli in order to reduce the uncertainty along the tangent space of this sub-manifold as rapidly as possible. Applications to simulated and real data indicate that these methods may in many cases improve the experimental efficiency by an order of magnitude.

M85 *Hebbian Learning of Bayes Optimal Decisions*

Bernhard Nessler nessler@igi.tugraz.at
Michael Pfeiffer pfeiffer@igi.tugraz.at
Wolfgang Maass maass@igi.tugraz.at
TU Graz, Graz, Austria

Uncertainty is omnipresent when we perceive or interact with our environment, and the Bayesian framework provides computational methods for dealing with it. Mathematical models for Bayesian decision making typically require datastructures that are hard to implement in neural networks. This article shows that even the simplest and experimentally best supported type of synaptic plasticity, Hebbian learning, in combination with a sparse, redundant neural code, can in principle learn to infer optimal Bayesian decisions. We present a concrete Hebbian learning rule operating on log-probability ratios. Modulated by reward-signals, this Hebbian plasticity rule also provides a new perspective for understanding how Bayesian inference could support fast reinforcement learning in the brain. In particular we show that recent experimental results by Yang and Shadlen [1] on reinforcement learning of probabilistic inference in primates can be modeled in this way.

M86 *Goal-directed decision making in prefrontal cortex: a computational framework*

Matthew Botvinick matthewb@princeton.edu
James An an@princeton.edu
Princeton University, Princeton, NJ, United States

Research in animal learning and behavioral neuroscience has distinguished between two forms of action control: a habit-based form, which relies on stored action values, and a goal-directed form, which forecasts and compares action outcomes based on a model of the environment. While habit-based control has been the subject of extensive computational research, the computational principles underlying goal-directed control in animals have so far received less attention. In the present paper, we advance a computational framework for goal-directed control in animals and humans. We take three empirically motivated points as founding premises: (1) Neurons in dorsolateral prefrontal cortex represent action policies, (2) Neurons in orbitofrontal cortex represent rewards, and (3) Neural computation, across domains, can be appropriately understood as performing structured probabilistic inference. On a purely computational level, the resulting account relates closely to previous work using Bayesian inference to solve Markov decision problems, but extends this work by introducing a new algorithm, which provably converges on optimal plans. On a cognitive and neuroscientific level, the theory provides a unifying framework for several different forms of goal-directed action selection, placing emphasis on a novel form, within which orbitofrontal reward representations directly drive policy selection.

M87 *Gaussian-process factor analysis for low-dimensional single-trial analysis of neural population activity*

Byron Yu byronyu@stanford.edu
Stanford University & Gatsby Computational Neuroscience Unit, UCL
John P Cunningham jcunnin@stanford.edu
Gopal Santhanam gopals@stanford.edu
Stephen Ryu seoulman@stanford.edu
Krishna V Shenoy shenoy@stanford.edu
Department of Electrical Engineering, Stanford University, Stanford, CA, United States
Maneesh Sahani maneesh@gatsby.ucl.ac.uk
Gatsby Computational Neuroscience Unit, UCL, London, United Kingdom

We consider the problem of extracting smooth low-dimensional "neural trajectories" that summarize the activity recorded simultaneously from tens to hundreds of neurons on individual experimental trials. Beyond the benefit of visualizing the high-dimensional noisy spiking activity in a compact denoised form, such trajectories can offer insight into the dynamics of the neural circuitry underlying the recorded activity. Current methods for extracting neural trajectories involve a two-stage process: the data are first "denoised" by smoothing over time, then a static dimensionality reduction technique is applied. We first describe extensions of the two-stage methods that allow the degree of smoothing to be chosen in a principled way, and account for spiking variability that may vary both across neurons and across time. We then present a novel method for extracting neural trajectories, Gaussian-process factor analysis (GPFA), which unifies the smoothing and dimensionality reduction operations in a common probabilistic framework. We applied these methods to the activity of 61 neurons recorded simultaneously in macaque premotor and motor cortices during reach planning and execution. By adopting a goodness-of-fit metric that measures how well the activity of each neuron can be predicted by all other recorded neurons, we found that GPFA provided a better characterization of the population activity than the two-stage methods. From the extracted single-trial neural trajectories, we directly observed a convergence in neural state during motor planning, an effect suggestive of attractor dynamics that was shown indirectly by previous studies.

M88 *Self-organization using synaptic plasticity*

Vicenç Gómez vgomez@iua.upf.edu
SNN, Radboud University Nijmegen, Nijmegen, Netherlands
Andreas Kaltenbrunner andreas.kaltenbrunner@upf.edu
Vicente López vicente.lopez@barcelonamedia.org
Barcelona Media - Innovation Centre, Barcelona, Spain
Hilbert J Kappen b.kappen@science.ru.nl
Radboud University Nijmegen, Nijmegen, Netherlands

Large networks of spiking neurons show abrupt changes in their collective dynamics resembling phase transitions studied in statistical physics. An example of this phenomenon is the transition from irregular, noise-driven dynamics to regular, self-sustained behavior observed in networks of integrate-and-fire neurons as the interaction strength between the neurons increases. In this work we show how a network of spiking neurons is able to self-organize toward a critical state for which the number of possible robust periods (dynamic range) is maximized. Self-organization occurs via synaptic dynamics. The resulting plasticity rule is defined locally so that global homeostasis near the critical state is achieved by local regulation of individual synapses.

M89 *Cell Assemblies in Large Sparse Inhibitory Networks of Biologically Realistic Spiking Neurons*

Adam Ponzi adamp@oist.jp
Jeff Wickens wickens@oist.jp
Okinawa Institute of Science and Technology(OIST), Okinawa, Japan

Cell assemblies exhibiting episodes of recurrent coherent activity have been observed in several brain regions including the striatum and hippocampus CA3. Here we address the question of how coherent dynamically switching assemblies appear in large networks of biologically realistic spiking neurons interacting deterministically. We show by numerical simulations of large asymmetric inhibitory networks with fixed external excitatory drive that if the network has intermediate to sparse connectivity, the individual cells are in the vicinity of a bifurcation between a quiescent and firing state and the network inhibition varies slowly on the spiking timescale, then cells form assemblies whose members show strong positive correlation, while members of different assemblies show strong negative correlation. We show that cells and assemblies switch between firing and quiescent states with time durations consistent with a power-law. Our results are in good qualitative agreement with the experimental studies. The deterministic dynamical behaviour is related to winner-less competition shown in small closed loop inhibitory networks with heteroclinic cycles connecting saddle-points.

M90 *Bio-inspired Real Time Sensory Map Realignment in a Robotic Barn Owl*

Juan Huo j.huo@ed.ac.uk
Zhijun Yang zhijun.yang@ed.ac.uk
Alan Murray alan.murray@ed.ac.uk
The University of Edinburgh, Edinburgh

The visual and auditory map alignment in the Superior Colliculus (SC) of barn owl is important for its accurate localization for prey behavior. Prism learning or Blindness may interfere this alignment and cause loss of the capability of accurate prey. However, juvenile barn owl could recover its sensory map alignment by shifting its auditory map. The adaptation of this map alignment is believed based on activity dependent axon developing in Inferior Colliculus (IC). A model is built to explore this mechanism. In this model, axon growing process is instructed by an inhibitory network in SC while the strength of the inhibition adjusted by Spike Timing Dependent Plasticity (STDP). We test and analyze this mechanism by application of the neural structures involved in spatial localization in a robotic system.

Tuesday, December 9th

Oral Session (8:30am–9:50am): "Sensorimotor Control"

8:30am Invited Talk: *Computations in Human Sensorimotor Control*

Daniel Wolpert wolpert@eng.cam.ac.uk
Department of Engineering, University of Cambridge, Cambridge, United Kingdom

The effortless ease with which humans move our arms, our eyes, even our lips when we speak masks the true complexity of the control processes involved. This is evident when we try to build machines to perform human control tasks. While computers can now beat grandmasters at chess, no computer can yet control a robot to manipulate a chess piece with the dexterity of a six-year-old child. I will review our recent work on how the humans learn to make skilled movements covering structural learning and generalization, how we learn the dynamics of tools and how we make decisions in the face of uncertainty.

Daniel Wolpert read medical sciences at Cambridge and clinical medicine at Oxford. After working as a medical doctor for a year he completed a PhD in the Physiology Department at Oxford. He then worked as a postdoctoral fellow at MIT, before moving to the Institute of Neurology, UCL. In 2005 he took up the post of Professor of Engineering for the Life Sciences at the University of Cambridge and is a Fellow of Trinity College. His research interests are computational and experimental approaches to human sensorimotor control (www.wolpertlab.com).

9:30am *Kernel-ARMA for Hand Tracking and Brain-Machine interfacing During 3D Motor Control*

Lavi Shpigelman shpigi@gmail.com
Hagai Lalazar hagai@alice.nc.huji.ac.il
Eilon Vaadia eilonv@ekmd.huji.ac.il
The Hebrew University, Jerusalem, Israel

Using machine learning algorithms to decode intended behavior from neural activity serves a dual purpose. First, these tools can be used to allow patients to interact with their environment through a Brain-Machine Interface (BMI). Second, analysis of the characteristics of such methods can reveal the significance of various features of neural activity, stimuli and responses to the encoding-decoding task. In this study we adapted, implemented and tested a machine learning method, called Kernel Auto-Regressive Moving Average (KARMA), for the task of inferring movements from neural activity in primary motor cortex. Our version of this algorithm is used in an on-line learning setting and is updated when feedback from the last inferred sequence become available. We first used it to track real hand movements executed by a monkey in a standard 3D motor control task. We then applied it in a closed-loop BMI setting to infer intended movement, while arms were restrained, allowing a monkey to perform the task using the BMI alone. KARMA is a recurrent method that learns a nonlinear model of output dynamics. It uses similarity functions (termed kernels) to compare between inputs. These kernels can be structured to incorporate domain knowledge into the method. We compare KARMA to various state-of-the-art methods by evaluating tracking performance and present results from the KARMA based BMI experiments.

Spotlights (9:50am–10:00am)

- ***Unifying the Sensory and Motor Components of Sensorimotor Adaptation***
  Adrian Haith, University of Edinburgh, Carl P.T. Jackson, R Chris Miall, University of Birmingham, and Sethu Vijayakumar, University of Edinburgh.
  See abstract, page 110.

- ***Effects of Stimulus Type and of Error-Correcting Code Design on BCI Speller Performance***
  Jeremy Hill, MPI for Biological Cybernetics, Jason Farquhar, Donders Institute for Brain, Cognition and Behaviour, Suzanna Martens, Felix Biessmann and Bernhard Schölkopf, MPI for Biological Cybernetics.
  See abstract, page 135.

- ***Bayesian Experimental Design of Magnetic Resonance Imaging Sequences***
  Matthias W Seeger, Saarland University, Hannes Nickisch, Rolf Pohmann and Bernhard Schölkopf, MPI for Biological Cybernetics.
  See abstract, page 128.

- ***How memory biases affect information transmission: A rational analysis of serial reproduction***
  Jing Xu and Thomas L Griffiths, University of California, Berkeley.
  See abstract, page 138.

- ***A rational model of preference learning and choice prediction by children***
  Chris Lucas, Thomas L Griffiths, University of California, Berkeley, Fei Xu, University of British Columbia, and Christine Fawcett, University of California, Berkeley.
  See abstract, page 139.

- ***Optimal Response Initiation: Why Recent Experience Matters***
  Matt Jones, Michael Mozer and Sachiko Kinoshita, University of Colorado.
  See abstract, page 137.

- ***Analyzing human feature learning as nonparametric Bayesian inference***
  Joseph Austerweil and Thomas L Griffiths, University of California, Berkeley.
  See abstract, page 135.

- ***A Massively Parallel Digital Learning Processor***
  Hans Peter Graf, Srihari Cadambi, Igor Durdanovic, Venkata Jakkula, Murugan Sankaradass, Eric Cosatto and Srimat Chakradhar, NEC Laboratories America.
  See abstract, page 134.

- ***Predicting the Geometry of Metal Binding Sites from Protein Sequence***
  Paolo Frasconi, Università di Firenze, and Andrea Passerini, Università di Trento.
  See abstract, page 132.

Oral Session (10:30am–11:50am): "Learning from Reinforcement: Modeling and Control"

10:30am *Stress, noradrenaline, and realistic prediction of mouse behaviour using reinforcement learning*

GEDIMINAS LUKSYS gediminas.luksys@epfl.ch
CARMEN SANDI carmen.sandi@epfl.ch
WULFRAM GERSTNER wulfram.gerstner@epfl.ch
Ecole Polytechnique Federale de Lausanne (EPFL), Lausanne, Switzerland

Suppose we train an animal in a conditioning experiment. Can one predict how a given animal, under given experimental conditions, would perform the task? Since various factors such as stress, motivation, genetic background, and previous errors in task performance can influence animal behaviour, this appears to be a very challenging aim. Reinforcement learning (RL) models have been successful in modeling animal (and human) behaviour, but their success has been limited because of uncertainty as to how to set meta-parameters (such as learning rate, exploitation-exploration balance and future reward discount factor) that strongly influence model performance. We show that a simple RL model whose meta-parameters are controlled by an artificial neural network, fed with inputs such as stress, affective phenotype, previous task performance, and even neuromodulatory manipulations, can successfully predict mouse behaviour in the "hole-box" - a simple conditioning task. Our results also provide important insights on how stress and anxiety affect animal learning, performance accuracy, and discounting of future rewards, and on how noradrenergic systems can interact with these processes.

10:50am *Learning to Use Working Memory in Partially Observable Environments through Dopaminergic Reinforcement*

MICHAEL T TODD mttodd@princeton.edu
YAEL NIV yael@princeton.edu
Princeton University, Princeton, NJ, United States
JONATHAN D. COHEN jdc@princeton.edu
Department of Psychology and Princeton Neursocience Institute,
Princeton University, Princeton, NJ, United States

Working memory is a central topic of cognitive neuroscience because it is critical for solving real world problems in which information from multiple temporally distant sources must be combined to generate appropriate behavior. However, an often neglected fact is that learning to use working memory effectively is itself a difficult problem. The "Gating" framework is a collection of psychological models that show how dopamine can train the basal ganglia and prefrontal cortex to form useful working memory representations in certain types of problems. We bring together gating with ideas from machine learning about using finite memory systems in more general problems. Thus we present a normative Gating model that learns, by online temporal difference methods, to use working memory to maximize discounted future rewards in general partially observable settings. The model successfully solves a benchmark working memory problem, and exhibits limitations similar to those observed in human experiments. Moreover, the model introduces a concise, normative definition of high level cognitive concepts such as working memory and cognitive control in terms of maximizing discounted future rewards.

11:10am *Simple Local Models for Complex Dynamical Systems*

Erik Talvitie etalviti@umich.edu
Satinder Singh baveja@umich.edu
University of Michigan, Computer Science and Engineering,
Ann Arbor, Michigan, United States

We present a novel mathematical formalism for the idea of a “local model” of an uncontrolled dynamical system, a model that makes only certain predictions in only certain situations. As a result of its restricted responsibilities, a local model may be far simpler than a complete model of the system. We then show how one might combine several local models to produce a more detailed model. We demonstrate our ability to learn a collection of local models on a large-scale example and do a preliminary empirical comparison of learning a collection of local models and some other model learning methods.

11:30am *Policy Search for Motor Primitives in Robotics*

Jens Kober jens.kober@tuebingen.mpg.de
Jan Peters mail@jan-peters.net
MPI for Biological Cybernetics, Tübingen, Germany

Many motor skills in humanoid robotics can be learned using parametrized motor primitives as done in imitation learning. However, most interesting motor learning problems are high-dimensional reinforcement learning problems often beyond the reach of current methods. In this paper, we extend previous work on policy learning from the immediate reward case to episodic reinforcement learning. We show that this results in a general, common framework also connected to policy gradient methods and yielding a novel algorithm for policy learning that is particularly well-suited for dynamic motor primitives.

The resulting algorithm is an EM-inspired algorithm applicable to complex motor learning tasks. We compare this algorithm to several well-known parametrized policy search methods and show that it outperforms them. We apply it in the context of motor learning and show that it can learn a complex Ball-in-a-Cup task using a real Barrett WAM™ robot arm.

Spotlights (11:50am–12:00pm)

- ***Multi-task Gaussian Process Learning of Robot Inverse Dynamics***
  Kian Ming A. Chai, Christopher K. I. Williams, Stefan Klanke and Sethu Vijayakumar, University of Edinburgh.
  See abstract, page 109.

- ***Multi-resolution Exploration in Continuous Spaces***
  Ali Nouri and Michael L. Littman, Rutgers University.
  See abstract, page 110.

- ***Multi-Agent Filtering with Infinitely Nested Beliefs***
  Luke Zettlemoyer, MIT, Brian Milch, Google Inc., and Leslie Pack Kaelbling, MIT.
  See abstract, page 110.

- ***Fitted Q-iteration by Advantage Weighted Regression***
  Gerhard Neumann, Graz University of Technology, and
  Jan Peters, MPI for Biological Cybernetics.
  See abstract, page 108.

- ***Signal-to-Noise Ratio Analysis of Policy Gradient Algorithms***
  John W. Roberts and Russ Tedrake, MIT CSAIL.
  See abstract, page 108.

- ***Optimization on a Budget: A Reinforcement Learning Approach***
  Paul L Ruvolo, Ian Fasel and Javier Movellan, University of California San Diego.
  See abstract, page 111.

- ***Near-optimal Regret Bounds for Reinforcement Learning***
  Peter Auer, Thomas Jaksch and Ronald Ortner, University of Leoben.
  See abstract, page 107.

- ***Algorithms for Infinitely Many-Armed Bandits***
  Yizao Wang, University of Michigan, Jean-Yves Audibert,
  Ecole Nationale des Ponts et Chaussées, and Remi Munos, INRIA.
  See abstract, page 109.

- ***Mind the Duality Gap: Logarithmic regret algorithms for online optimization***
  Shai Shalev-Shwartz and Sham M Kakade, Toyota Technological Institute.
  See abstract, page 123.

- ***An Online Algorithm for Maximizing Submodular Functions***
  Matthew Streeter, Google, Inc., and Daniel Golovin, Carnegie Mellon University.
  See abstract, page 112.

Oral Session (2:00pm–3:20pm): "Combinatorial Approximation"

2:00pm **Invited Talk:** ***Online Stochastic Combinatorial Optimization***

Pascal van Hentenryck pvh@cs.brown.edu
Brown University, Providence, RI, United States

Advances in telecommunication technologies, combined with the increasingly integrated nature of optimization applications, create a wealth of online optimization problems in scheduling, routing, and resource allocation. Moreover, in many applications, stochastic and simulation models, or massive amount of historical data, are typically available to the decision-maker. This talk presents an overview of online anticipatory algorithms for addressing this new class of applications and reports on their performance in a variety of settings. Anticipatory algorithms make decisions online, by conditionally sampling a distribution and solving the resulting optimization problems. Interestingly, many of these algorithms features innovative integration of artificial intelligence, discrete optimization, and stochastic programming techniques.

Pascal Van Hentenryck is a professor of computer science at Brown University and the director of the optimization laboratory. During the past 20 years, he developed a number of influential systems, including the pioneering CHIP system which is the foundation of all modern constraint programming systems, the Numerica system for global optimization, the optimization programming language OPL, and the programming language Comet which supports constraint-based local search, constraint programming, and mathematical programming. Most of these systems, and their foundations, are described in books published by the MIT Press and have been licensed to industry. His current research in online stochastic optimization integrates techniques from artificial intelligence, stochastic optimization, and combinatorial optimization to tackle complex decision-making applications under uncertainty. Van Hentenryck is the recipient of an 1993 NSF National Young Investigator (NYI) award, the 2002 INFORMS ICS Award for research excellence at the interface between computer science and operations research, the 2006 ACP Award for Research Excellence in Constraint Programming, best paper awards at CP'03, CP'04, and IJCAI'07, and an IBM Faculty Award in 2004. Pascal has given invited talks at many premier conferences in artificial intelligence, operations research, and programming languages, including IJCAI'97, CP'97, UAI'06, CP'AI'OR'08, SIOP'08, and ECAI'08.

3:00pm ***MAS: a multiplicative approximation scheme for probabilistic inference***

Ydo Wexler ydow@microsoft.com
Christopher Meek meek@microsoft.com
Microsoft Research, Redmond, WA, United States

We propose a multiplicative approximation scheme (MAS) for inference problems in graphical models, which can be applied to various inference algorithms. The method uses ϵ-decompositions which decompose functions used throughout the inference procedure into functions over smaller sets of variables with a known error ϵ. MAS translates these local approximations into bounds on the accuracy of the results. We show how to optimize ϵ-decompositions and provide a fast closed-form solution for an L_2 approximation. Applying MAS to the Variable Elimination inference algorithm, we introduce an algorithm we call DynaDecomp which is extremely fast in practice and provides guaranteed error bounds on the result. The superior accuracy and efficiency of DynaDecomp is demonstrated.

Spotlights (3:20pm–3:30pm)

- ***Clusters and Coarse Partitions in LP Relaxations***
  David Sontag, Massachusetts Institute of Technology, Amir Globerson, Hebrew University, and Tommi Jaakkola, Massachusetts Institute of Technology.
  See abstract, page 129.

- ***Improved Moves for Truncated Convex Models***
  M Pawan Kumar, University of Oxford, and Philip H S Torr, Oxford Brookes University.
  See abstract, page 131.

- ***Learning Bounded Treewidth Bayesian Networks***
  Gal Elidan, Hebrew University, and Stephen Gould, Stanford University.
  See abstract, page 117.

- ***Bayesian Network Score Approximation using a Metagraph Kernel***
  Benjamin Yackley, Department of Computer Science, University of New Mexico, Eduardo Corona, New York University, and Terran Lane, Department of Computer Science, University of New Mexico.
  See abstract, page 118.

- ***Near-minimax recursive density estimation on the binary hypercube***
  Maxim Raginsky, Duke University, Svetlana Lazebnik, University of North Carolina at Chapel Hill, Rebecca Willett and Jorge Silva, Duke University.
  See abstract, page 119.

- ***Bounds on marginal probability distributions***
  Joris M Mooij, MPI for Biological Cybernetics, and Hilbert J Kappen, Radboud University.
  See abstract, page 128.

- ***Posterior Consistency of the Silverman g-prior in Bayesian Model Choice***
  zhihua zhang, Michael I. Jordan, University of California, Berkeley, and Dit-Yan Yeung, Hong Kong University of Science & Technology.
  See abstract, page 124.

- ***Improving on Expectation Propagation***
  Manfred Opper, TU Berlin, Ulrich Paquet, University of Cambridge, and Ole Winther, University of Copenhagen.
  See abstract, page 126.

- ***The Gaussian Process Density Sampler***
  Ryan Adams, University of Cambridge, Iain Murray, University of Toronto, and David MacKay, University of Cambridge.
  See abstract, page 128.

- ***Efficient Sampling for Gaussian Process Inference using Control Variables***
  Michalis Titsias, Neil D Lawrence and Magnus Rattray, University of Manchester.
  See abstract, page 127.

Oral Session (4:00pm–5:20pm): "Fast Output Processing, Novelty Detection and Unlabeled Data"

4:00pm *Online Metric Learning and Fast Similarity Search*

Prateek Jain pjain@cs.utexas.edu
Brian Kulis kulis@cs.utexas.edu
Inderjit Dhillon inderjit@cs.utexa.edu
Kristen Grauman grauman@cs.utexas.edu
University of Texas at Austin, Austin, TX, United States

Metric learning algorithms can provide useful distance functions for a variety of domains, and recent work has shown good accuracy for problems where the learner can access all distance constraints at once. However, in many real applications, constraints are only available incrementally, thus necessitating methods that can perform online updates to the learned metric. Existing online algorithms offer bounds on worst-case performance, but typically do not perform well in practice as compared to their offline counterparts. We present a new online metric learning algorithm that updates a learned Mahalanobis metric based on LogDet regularization and gradient descent. We prove theoretical worst-case performance bounds, and empirically compare the proposed method against existing online metric learning algorithms. To further boost the practicality of our approach, we develop an online locality-sensitive hashing scheme which leads to efficient updates for approximate similarity search data structures. We demonstrate our algorithm on multiple datasets and show that it outperforms relevant baselines.

4:20pm *Fast Prediction on a Tree*

Mark Herbster m.herbster@cs.ucl.ac.uk
Massimiliano Pontil m.pontil@cs.ucl.ac.uk
Sergio Rojas Galeano s.rojas@cs.ucl.ac.uk
Department of Computer Science, University College London, London, United Kingdom

Given an n-vertex weighted tree with structural diameter S and a subset of m vertices, we present a technique to compute a corresponding $m \times m$ Gram matrix of the pseudoinverse of the graph Laplacian in $O(n+m^2+mS)$ time. We discuss the application of this technique to fast label prediction on a generic graph. We approximate the graph with a spanning tree and then we predict with the kernel perceptron. We address the approximation of the graph with either a minimum spanning tree or a shortest path tree. The fast computation of the pseudoinverse enables us to address prediction problems on large graphs. To this end we present experiments on two web-spam classification tasks, one of which includes a graph with 400,000 nodes and more than 10,000,000 edges. The results indicate that the accuracy of our technique is competitive with previous methods using the full graph information.

4:40pm *Beyond Novelty Detection: Incongruent Events, when General and Specific Classifiers Disagree*

Daphna Weinshall daphna@cs.huji.ac.il
School of Computer Science and Engineering, The Hebrew University of Jerusalem, Israel
Hynek Hermansky hermansky@ieee.org
John Hopkins, Baltimore, USA
Alon Zweig alon.zweig@mail.huji.ac.il
School of Computer Science and Engineering, The Hebrew University of Jerusalem, Israel
Jie Luo jluo@idiap.ch
IDIAP Research Institute, Martigny, and EPFL, Lausanne, Switzerland
Holly Jimison jimisonh@ohsu.edu
Oregon Health & Science University, Portland, OR, United States
Frank Ohl frank.ohl@ifn-magdeburg.de
Leibniz-Institut für Neurobiologie, Magdeburg, Germany
Misha Pavel pavel@bme.ogi.edu
Oregon Health & Science University, Portland, OR, United States

Unexpected stimuli are a challenge to any machine learning algorithm. Here we identify distinct types of unexpected events, focusing on 'incongruent events' - when 'general level' and 'specific level' classifiers give conflicting predictions. We define a formal framework for the representation and processing of incongruent events: starting from the notion of label hierarchy, we show how partial order on labels can be deduced from such hierarchies. For each event, we compute its probability in different ways, based on adjacent levels (according to the partial order) in the label hierarchy. An incongruent event is an event where the probability computed based on some more specific level (in accordance with the partial order) is much smaller than the probability computed based on some more general level, leading to conflicting predictions. We derive algorithms to detect incongruent events from different types of hierarchies, corresponding to class membership or part membership. Respectively, we show promising results with real data on two specific problems: Out Of Vocabulary words in speech recognition, and the identification of a new sub-class (e.g., the face of a new individual) in audio-visual facial object recognition.

5:00pm *Unlabeled data: Now it helps, now it doesn't*

Aarti Singh singh@cae.wisc.edu
Robert Nowak nowak@engr.wisc.edu
Xiaojin Zhu jerryzhu@cs.wisc.edu
University of Wisconsin-Madison, Madison, WI, United States

Empirical evidence shows that in favorable situations semi-supervised learning (SSL) algorithms can capitalize on the abundance of unlabeled training data to improve the performance of a learning task, in the sense that fewer labeled training data are needed to achieve a target error bound. However, in other situations unlabeled data do not seem to help. Recent attempts at theoretically characterizing SSL gains only provide a partial and sometimes apparently conflicting explanations of whether, and to what extent, unlabeled data can help. In this paper, we attempt to bridge the gap between the practice and theory of semi-supervised learning. We develop a finite sample analysis that characterizes the value of unlabeled data and quantifies the performance improvement of SSL compared to supervised learning. We show that there are large classes of problems for which SSL can significantly outperform supervised learning, in finite sample regimes and sometimes also in terms of error convergence rates.

Spotlights (5:20pm–5:30pm)

- ***A Transductive Bound for the Voted Classifier with an Application to Semi-supervised Learning***
  Massih Amini, Laboratoire d'Informatique de Paris 6, Francois Laviolette, Université Laval, and Nicolas Usunier, Laboratoire d'Informatique de Paris 6.
  See abstract, page 115.

- ***On the Design of Loss Functions for Classification: theory, robustness to outliers, and SavageBoost***
  Hamed Masnadi-Shirazi, UCSD, and Nuno Vasconcelos, University of California San Diego.
  See abstract, page 121.

- ***Adaptive Martingale Boosting***
  Phil Long, Google, and Rocco A. Servedio, Columbia University.
  See abstract, page 122.

- ***On the Generalization Ability of Online Strongly Convex Programming Algorithms***
  Sham M Kakade and Ambuj Tewari, Toyota Technological Institute, Chicago.
  See abstract, page 120.

- ***Weighted Sums of Random Kitchen Sinks: Replacing minimization with randomization in learning***
  Ali Rahimi, Intel Research Berkeley, and Benjamin Recht, California Institute of Technology.
  See abstract, page 120.

- ***Sparse Online Learning via Truncated Gradient***
  John Langford, Yahoo, Lihong Li and Tong Zhang, Rutgers University.
  See abstract, page 120.

- ***On the Efficient Minimization of Classification Calibrated Surrogates***
  Richard Nock, CEREGMIA - Université Antilles-Guyane, and Frank Nielsen, Ecole Polytechnique.
  See abstract, page 116.

- ***Exact Convex Confidence-Weighted Learning***
  Koby Crammer, Mark Dredze, University of Pennsylvania, and Fernando Pereira, Google, Inc..
  See abstract, page 116.

- ***Breaking Audio CAPTCHAs***
  Jennifer Tam, Jiri Simsa, Sean Hyde and Luis von Ahn, Carnegie Mellon University.
  See abstract, page 115.

- ***Nonlinear causal discovery with additive noise models***
  PATRIK O. HOYER, Helsinki Institute for Information Technology,
  DOMINIK JANZING, JORIS M MOOIJ, JONAS PETERS and BERNHARD SCHÖLKOPF, Max-Planck Institute for Biological Cybernetics.
  See abstract, page 131.

- ***Sequential effects: Superstition or rational behavior?***
  ANGELA J YU, University of California, San Diego, and
  JONATHAN D. COHEN, Princeton University.
  See abstract, page 138.

Poster Session (7:30pm–12:00am)

T1 *Regularized Policy Iteration*

AMIR M. FARAHMAND amir@cs.ualberta.ca
Department of Computing Science, University of Alberta, Edmonton, AB, Canada
MOHAMMAD GHAVAMZADEH mgh@cs.ualberta.ca
INRIA Lille - Team SequeL, Villeneuve d'Ascq, France
CSABA SZEPESVARI szepesva@cs.ualberta.ca
University of Alberta, Edmonton, AB, Canada
SHIE MANNOR shie.mannor@mcgill.ca
McGill University, Montreal, Canada

In this paper we consider approximate policy-iteration-based reinforcement learning algorithms. In order to implement a flexible function approximation scheme we propose the use of non-parametric methods with regularization, providing a convenient way to control the complexity of the function approximator. We propose two novel regularized policy iteration algorithms by adding L2-regularization to two widely-used policy evaluation methods: Bellman residual minimization (BRM) and least-squares temporal difference learning (LSTD). We derive efficient implementation for our algorithms when the approximate value-functions belong to a reproducing kernel Hilbert space. We also provide finite-sample performance bounds for our algorithms and show that they are able to achieve optimal rates of convergence under the studied conditions.

T2 *Near-optimal Regret Bounds for Reinforcement Learning*

PETER AUER auer@unileoben.ac.at
THOMAS JAKSCH thomas.jaksch@unileoben.ac.at
RONALD ORTNER rortner@unileoben.ac.at
University of Leoben, Leoben, Austria

For undiscounted reinforcement learning in Markov decision processes (MDPs) we consider the total regret of a learning algorithm with respect to an optimal policy. In order to describe the transition structure of an MDP we propose a new parameter: An MDP has diameter D if for any pair of states s, s' there is a policy which moves from s to s' in at most D steps (on average). We present a reinforcement learning algorithm with total regret $O(DS\sqrt{AT})$ after T steps for any unknown MDP with S states, A actions per state, and diameter D. This bound holds with high probability. We also present a corresponding lower bound of $\Omega(\sqrt{DSAT})$ on the total regret of any learning algorithm. Both bounds demonstrate the utility of the diameter as structural parameter of the MDP.

Spotlight presentation, Tuesday, 11:50am.

T3 *Fitted Q-iteration by Advantage Weighted Regression*

Gerhard Neumann gneumann@sbox.tugraz.at
Graz University of Technology, Graz, Austria
Jan Peters mail@jan-peters.net
MPI for Biological Cybernetics, Tuebingen, Germany

Recently, fitted Q-iteration (FQI) based methods have become more popular due to their increased sample efficiency, a more stable learning process and the higher quality of the resulting policy. However, these methods remain hard to use for continuous action spaces which frequently occur in real-world tasks, e.g., in robotics and other technical applications. The greedy action selection commonly used for the policy improvement step is particularly problematic as it is expensive for continuous actions, can cause an unstable learning process, introduces an optimization bias and results in highly non-smooth policies unsuitable for real-world systems. In this paper, we show that by using a soft-greedy action selection the policy improvement step used in FQI can be simplified to an inexpensive advantage-weighted regression. With this result, we are able to derive a new, computationally efficient FQI algorithm which can even deal with high dimensional action spaces.

Spotlight presentation, Tuesday, 11:50am.

T4 *Signal-to-Noise Ratio Analysis of Policy Gradient Algorithms*

John W. Roberts jwr@mit.edu
Russ Tedrake russt@mit.edu
MIT CSAIL, Cambridge, MA, United States

Policy gradient (PG) reinforcement learning algorithms have strong (local) convergence guarantees, but their learning performance is typically limited by a large variance in the estimate of the gradient. In this paper, we formulate the variance reduction problem by describing a signal-to-noise ratio (SNR) for policy gradient algorithms, and evaluate this SNR carefully for the popular Weight Perturbation (WP) algorithm. We confirm that SNR is a good predictor of long-term learning performance, and that in our episodic formulation, the cost-to-go function is indeed the optimal baseline. We then propose two modifications to traditional model-free policy gradient algorithms in order to optimize the SNR. First, we examine WP using anisotropic sampling distributions, which introduces a bias into the update but increases the SNR; this bias can be interpretted as following the natural gradient of the cost function. Second, we show that non-Gaussian distributions

can also increase the SNR, and argue that the optimal isotropic distribution is a 'shell' distribution with a constant magnitude and uniform distribution in direction. We demonstrate that both modifications produce substantial improvements in learning performance in challenging policy gradient experiments.

Spotlight presentation, Tuesday, 11:50am.

T5 *Policy Search for Motor Primitives in Robotics*

JENS KOBER and JAN PETERS, MPI for Biological Cybernetics.
Oral presentation, Tuesday, 11:30am. *See abstract, page 100.*

T6 *Simple Local Models for Complex Dynamical Systems*

ERIK TALVITIE and SATINDER SINGH, University of Michigan.
Oral presentation, Tuesday, 11:10am. *See abstract, page 99.*

T7 *Multi-task Gaussian Process Learning of Robot Inverse Dynamics*

KIAN MING A. CHAI k.m.a.chai@ed.ac.uk
CHRISTOPHER K. I. WILLIAMS c.k.i.williams@ed.ac.uk
STEFAN KLANKE s.klanke@ed.ac.uk
SETHU VIJAYAKUMAR sethu.vijayakumar@ed.ac.uk
University of Edinburgh, Edinburgh, United Kingdom

The inverse dynamics problem for a robotic manipulator is to compute the torques needed at the joints to drive it along a given trajectory; it is beneficial to be able to learn this function for adaptive control. A robotic manipulator will often need to be controlled while holding different loads in its end effector, giving rise to a multi-task learning problem. By placing independent Gaussian process priors over the latent functions of the inverse dynamics, we obtain a multi-task Gaussian process prior for handling multiple loads, where the inter-task similarity depends on the underlying inertial parameters. Experiments demonstrate that this multi-task formulation is effective in sharing information among the various loads, and generally improves performance over either learning only on single tasks or pooling the data over all tasks.

Spotlight presentation, Tuesday, 11:50am.

T8 *Algorithms for Infinitely Many-Armed Bandits*

YIZAO WANG yizwang@umich.edu
University of Michigan, Ann Arbor, MI, United States
JEAN-YVES AUDIBERT audibert@certis.enpc.fr
Ecole Nationale des Ponts et Chaussées, Champs-sur-Marne, France
REMI MUNOS remi.munos@inria.fr
INRIA, Villeneuve d'Ascq, France

We consider multi-armed bandit problems where the number of arms is larger than the possible number of experiments. We make a stochastic assumption on the mean-reward of a new selected arm which characterizes its probability of being a near-optimal arm. Our assumption is weaker than in previous works. We describe algorithms based on upper-confidence-bounds applied to a restricted set of randomly selected arms and provide upper-bounds on the resulting expected regret. We also derive a lower-bound which matchs (up to logarithmic factors) the upper-bound in some cases.

Spotlight presentation, Tuesday, 11:50am.

T9 *Multi-resolution Exploration in Continuous Spaces*

Ali Nouri nouri@cs.rutgers.edu
Michael L. Littman mlittman@cs.rutgers.edu
Department of Computer Science, Rutgers University, Piscataway, NJ, United States

The essence of exploration is acting to try to decrease uncertainty. We propose a new methodology for representing uncertainty in continuous-state control problems. Our approach, multi-resolution exploration (MRE), uses a hierarchical mapping to identify regions of the state space that would benefit from additional samples. We demonstrate MRE's broad utility by using it to speed up learning in a prototypical model-based and value-based reinforcement-learning method. Empirical results show that MRE improves upon state-of-the-art exploration approaches.

Spotlight presentation, Tuesday, 11:50am.

T10 *Multi-Agent Filtering with Infinitely Nested Beliefs*

Luke Zettlemoyer lsz@csail.mit.edu
MIT CSAIL, Cambridge, MA, United States
Brian Milch brian@google.com
Google Inc., Mountain View, CA, United States
Leslie Pack Kaelbling lpk@csail.mit.edu
MIT CSAIL, Cambridge, MA, United States

In partially observable worlds with many agents, nested beliefs are formed when agents simultaneously reason about the unknown state of the world and the beliefs of the other agents. The multi-agent filtering problem is to efficiently represent and update these beliefs through time as the agents act in the world. In this paper, we formally define an infinite sequence of nested beliefs about the state of the world at the current time t and present a filtering algorithm that maintains a finite representation which can be used to generate these beliefs. In some cases, this representation can be updated exactly in constant time; we also present a simple approximation scheme to compact beliefs if they become too complex. In experiments, we demonstrate efficient filtering in a range of multi-agent domains.

Spotlight presentation, Tuesday, 11:50am.

T11 *Unifying the Sensory and Motor Components of Sensorimotor Adaptation*

Adrian Haith adrian.haith@ed.ac.uk
University of Edinburgh, Edinburgh, United Kingdom
Carl P.T. Jackson c.p.jackson.1@bham.ac.uk
R Chris Miall r.c.miall@bham.ac.uk
University of Birmingham, Birmingham, United Kingdom
Sethu Vijayakumar sethu.vijayakumar@ed.ac.uk
University of Edinburgh, Edinburgh, United Kingdom

Adaptation of visually guided reaching movements in novel visuomotor environments (e.g. wearing prism goggles) comprises not only motor adaptation but also substantial sensory adaptation, corresponding to shifts in the perceived spatial location of visual and proprioceptive cues. Previous computational models of the sensory component of visuomotor adaptation have assumed that it is driven purely by the discrepancy introduced between visual and proprioceptive estimates of hand position and is independent of any motor component of adaptation. We instead propose a unified model in which sensory and motor adaptation are jointly driven by optimal Bayesian estimation of the sensory and motor contributions to perceived errors. Our model is able to account for patterns of performance errors during visuomotor adaptation as well as the subsequent perceptual aftereffects. This unified model also makes the surprising prediction that force field adaptation will elicit similar perceptual shifts, even though there is never any discrepancy between visual and proprioceptive observations. We confirm this prediction with an experiment.

Spotlight presentation, Tuesday, 9:50am.

T12 *The Recurrent Temporal Restricted Boltzmann Machine*

Ilya Sutskever ilya@cs.utoronto.ca
Geoffrey E Hinton hinton@cs.toronto.edu
Graham Taylor gwtaylor@cs.toronto.edu
University of Toronto, Toronto, ON, Canada

The Temporal Restricted Boltzmann Machine (TRBM) is a probabilistic model for sequences that is able to successfully model (i.e., generate nice-looking samples of) several very high dimensional sequences, such as motion capture data and the pixels of low resolution videos of balls bouncing in a box. The major disadvantage of the TRBM is that exact inference is extremely hard, since even computing a Gibbs update for a single variable of the posterior is exponentially expensive. This difficulty has necessitated the use of a heuristic inference procedure, that nonetheless was accurate enough for successful learning. In this paper we introduce the Recurrent TRBM, which is a very slight modification of the TRBM for which exact inference is very easy and exact gradient learning is almost tractable. We demonstrate that the RTRBM is better than an analogous TRBM at generating motion capture and videos of bouncing balls.

T13 *Optimization on a Budget: A Reinforcement Learning Approach*

Paul L Ruvolo pruvolo@cs.ucsd.edu
Ian Fasel ianfasel@cs.utexas.edu
Javier Movellan movellan@mplab.ucsd.edu
University of California San Diego, La Jolla, United States

Many popular optimization algorithms, like the Levenberg-Marquardt algorithm (LMA), use heuristic-based "controllers" that modulate the behavior of the optimizer during the optimization process. For example, in the LMA a damping parameter is dynamically modified based on a set rules that were developed using various heuristic arguments. Reinforcement learning (RL) is a machine learning approach to learn optimal controllers by examples and thus is an obvious candidate to improve the heuristic-based controllers implicit in the most popular and heavily used optimization algorithms.

Improving the performance of off-the-shelf optimizers is particularly important for time-constrained optimization problems. For example the LMA algorithm has become popular for many real-time computer vision problems, including object tracking from video, where only a small amount of time can be allocated to the optimizer on each incoming video frame.

Here we show that a popular modern reinforcement learning technique using a very simply state space can dramatically improve the performance of general purpose optimizers, like the LMA. Most surprisingly the controllers learned for a particular domain appear to work very well also on very different optimization domains. For example we used RL methods to train a new controller for the damping parameter of the LMA. This controller was trained on a collection of classic, relatively small, non-linear regression problems. The modified LMA performed better than the standard LMA on these problems. Most surprisingly, it also dramatically outperformed the standard LMA on a difficult large scale computer vision problem for which it had not been trained before. Thus the controller appeared to have extracted control rules that were not just domain specific but generalized across a wide range of optimization domains.

Spotlight presentation, Tuesday, 11:50am.

T14 *An Online Algorithm for Maximizing Submodular Functions*

Matthew Streeter mstreeter@google.com
Google, Inc., Pittsburgh, PA, United States
Daniel Golovin dgolovin@cs.cmu.edu
Carnegie Mellon University, Pittsburgh, PA, United States

We present an algorithm for solving a broad class of online resource allocation problems. Our online algorithm can be applied in environments where abstract jobs arrive one at a time, and one can complete the jobs by investing time in a number of abstract activities, according to some schedule. We assume that the fraction of jobs completed by a schedule is a monotone, submodular function of a set of pairs (v,t), where t is the time invested in activity v. Under this assumption, our online algorithm performs near-optimally according to two natural metrics: (i) the fraction of jobs completed within time T, for some fixed deadline $T > 0$, and (ii) the average time required to complete each job. We evaluate our algorithm experimentally by using it to learn, online, a schedule for allocating CPU time among solvers entered in the 2007 SAT solver competition.

Spotlight presentation, Tuesday, 11:50am.

T15 *Online Metric Learning and Fast Similarity Search*

Prateek Jain, Brian Kulis, Inderjit Dhillon and
Kristen Grauman, University of Texas at Austin.
Oral presentation, Tuesday, 4:00pm. See abstract, page 104.

T16 *Fast Prediction on a Tree*

Mark Herbster, Massimiliano Pontil and Sergio
Rojas Galeano, University College London.
Oral presentation, Tuesday, 4:20pm. See abstract, page 104.

T17 *Correlated Bigram LSA for Unsupervised Language Model Adaptation*

Yik-Cheung (Wilson) Tam yct@cs.cmu.edu
Tanja Schultz tanja@cs.cmu.edu
Carnegie Mellon University, Pittsburgh, PA, United States

We present a correlated bigram LSA approach for unsupervised LM adaptation for automatic speech recognition. The model is trained using efficient variational EM and smoothed using the proposed fractional Kneser-Ney smoothing which handles fractional counts. We address the scalability issue to large training corpora via bootstrapping of bigram LSA from unigram LSA. For LM adaptation, unigram and bigram LSA are integrated into the background N-gram LM via marginal adaptation and linear interpolation respectively. Experimental results on the Mandarin RT04 test set show that applying unigram and bigram LSA together yields 6%-8% relative perplexity reduction and 2.5% relative character error rate reduction which is statistically significant compared to applying only unigram LSA. On the large-scale evaluation on Arabic, word error rate reduction from bigram LSA is statistically significant compared to the unadapted baseline.

T18 *Learning the Semantic Correlation: An Alternative Way to Gain from Unlabeled Text*

Yi Zhang yizhang1@cs.cmu.edu
Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA, United States
Jeff Schneider schneide@cs.cmu.edu
Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, United States
Artur Dubrawski awd@cs.cmu.edu
Auton Lab, Carnegie Mellon University, Pittsburgh, PA, United States

In this paper, we address the question of what kind of knowledge is generally transferable from unlabeled text. We suggest and analyze the semantic correlation of words as a generally transferable structure of the language and propose a new method to learn this structure using an appropriately chosen latent variable model. This semantic correlation contains structural information of the language space and can be used to control the joint shrinkage of model parameters for any specific task in the same space through regularization. In an empirical study, we construct 190 different text classification tasks from a real-world benchmark, and the unlabeled documents are a mixture from all these tasks. We test the ability of various algorithms to use the mixed unlabeled text to enhance all classification tasks. Empirical results show that the proposed approach is a reliable and scalable method for semi-supervised learning, regardless of the source of unlabeled data, the specific task to be enhanced, and the prediction model used.

T19 *A Scalable Hierarchical Distributed Language Model*

Andriy Mnih amnih@cs.toronto.edu
Geoffrey E Hinton hinton@cs.toronto.edu
Department of Computer Science, University of Toronto, Canada

Neural probabilistic language models (NPLMs) have been shown to be competitive with and occasionally superior to the widely-used n-gram language models. The main drawback of NPLMs is their extremely long training and testing times. Morin and Bengio have proposed a hierarchical language model built around a binary tree of words which was two orders of magnitude faster than the non-hierarchical language model it was based on. However, it performed considerably worse than its non-hierarchical counterpart in spite of using a word tree created using expert knowledge. We introduce a fast hierarchical language model along with a simple feature-based algorithm for automatic construction of word trees from data. We then show that the resulting models can outperform non-hierarchical models and achieve state-of-the-art performance.

T20 *Tighter Bounds for Structured Estimation*

Olivier Chapelle chap@yahoo-inc.com
Yahoo! Research, Santa Clara, CA, United States
Chuong B. Do chuongdo@cs.stanford.edu
Quoc Le quocle@stanford.edu
Stanford University, Stanford, CA, United States
Alex J. Smola alex@smola.org
Yahoo! Research, Santa-Clara, CA, United States
Choon Hui Teo choonhui.teo@gmail.com
Statistical Machine Learning Program, NICTA; Australian National University, Canberra, Australia

Large-margin structured estimation methods work by minimizing a convex upper bound of loss functions. While they allow for efficient optimization algorithms, these convex formulations are not tight and sacrifice the ability to accurately model the true loss. We present tighter non-convex bounds based on generalizing the notion of a ramp loss from binary classification to structured estimation. We show that a small modification of existing optimization algorithms suffices to solve this modified problem. On structured prediction tasks such as protein sequence alignment and web page ranking, our algorithm leads to improved accuracy.

T21 *Beyond Novelty Detection: Incongruent Events, when General and Specific Classifiers Disagree*

Daphna Weinshall, The Hebrew University, Hynek Hermansky, John Hopkins, Alon Zweig, The Hebrew University, Jie Luo, IDIAP, Holly Jimison, Oregon Health & Science University, Frank Ohl, Leibniz-Institut für Neurobiologie, and Misha Pavel, Oregon Health & Science University.
Oral presentation, Tuesday, 4:40pm. See abstract, page 104.

T22 *Unlabeled data: Now it helps, now it doesn't*

Aarti Singh, Robert Nowak and Xiaojin Zhu, University of Wisconsin.
Oral presentation, Tuesday, 5:00pm. See abstract, page 105.

T23 *Breaking Audio CAPTCHAs*

Jennifer Tam jdtam@cs.cmu.edu
Jiri Simsa jsimsa@cs.cmu.edu
Sean Hyde seanh@andrew.cmu.edu
Luis von Ahn biglou@cs.cmu.edu
Carnegie Mellon University, Pittsburgh, PA, United States

CAPTCHAs are computer-generated tests that humans can pass but current computer systems cannot. CAPTCHAs provide a method for automatically distinguishing a human from a computer program, and therefore can protect Web services from abuse by so-called "bots." Most CAPTCHAs consist of distorted images, usually text, for which a user must provide some description. Unfortunately, visual CAPTCHAs limit access to the millions of visually impaired people using the Web. Audio CAPTCHAs were created to solve this accessibility issue; however, the security of audio CAPTCHAs was never formally tested. Some visual CAPTCHAs have been broken using machine learning techniques, and we propose using similar ideas to test the security of audio CAPTCHAs. Audio CAPTCHAs are generally composed of a set of words to be identified, layered on top of noise. We analyzed the security of current audio CAPTCHAs from popular Web sites by using AdaBoost, SVM, and k-NN, and achieved correct solutions for test samples with accuracy up to 71%. Such accuracy is enough to consider these CAPTCHAs broken. Training several different machine learning algorithms on different types of audio CAPTCHAs allowed us to analyze the strengths and weaknesses of the algorithms so that we could suggest a design for a more robust audio CAPTCHA.

Spotlight presentation, Tuesday, 5:20am.

T24 *A Transductive Bound for the Voted Classifier with an Application to Semi-supervised Learning*

Massih Amini amini@poleia.lip6.fr
Laboratoire d'Informatique de Paris 6, Université Paris 6, Paris, France,and National Research Council, Gatineau, Canada
Francois Laviolette francois.laviolette@ift.ulaval.ca
Université Laval, Québec, Canada
Nicolas Usunier usunier@poleia.lip6.fr
Laboratoire d'Informatique de Paris 6, Université Paris 6, Paris, France

In this paper we present two transductive bounds on the risk of the majority vote estimated over partially labeled training sets. Our first bound involves the margin distribution of the classifier and a risk bound on its associate Gibbs classifier. The bound is tight when so is the Gibbs's bound and when the errors of the majority vote classifier is concentrated on a zone of low margin. In semi-supervised learning, considering the margin as an indicator of confidence constitutes the working hypothesis of algorithms which search the decision boundary on low density regions. Following this assumption, we propose to bound the error probability of the voted classifier on the examples for whose margins are above a fixed threshold. As an application, we propose a self-learning algorithm which iteratively assigns pseudo-labels to the set of unlabeled training examples that have their margin above a threshold obtained from this bound. Empirical results on different datasets show the effectiveness of our approach compared to the same algorithm and the TSVM in which the threshold is fixed manually.

Spotlight presentation, Tuesday, 5:20pm.

T25 *On the Efficient Minimization of Classification Calibrated Surrogates*

Richard Nock rnock@martinique.univ-ag.fr
CEREGMIA - Université Antilles-Guyane, Schoelcher, Martinique
Frank Nielsen nielsen@lix.polytechnique.fr
Ecole Polytechnique, Palaiseau, France

Bartlett *et al* (2006) recently proved that a ground condition for convex surrogates, classification calibration, ties up the minimization of the surrogates and classification risks, and left as an important problem the algorithmic questions about the minimization of these surrogates. In this paper, we propose an algorithm which provably minimizes any classification calibrated surrogate strictly convex and differentiable — a set whose losses span the exponential, logistic and squared losses —, with boosting-type guaranteed convergence rates under a weak learning assumption. A particular subclass of these surrogates, that we call balanced convex surrogates, has a key rationale that ties it to maximum likelihood estimation, zero-sum games and the set of losses that satisfy some of the most common requirements for losses in supervised learning. We report experiments on more than 50 readily available domains of 11 flavors of the algorithm, that shed light on new surrogates, and the potential of data dependent strategies to *tune* surrogates.

Spotlight presentation, Tuesday, 5:20pm.

T26 *Exact Convex Confidence-Weighted Learning*

Koby Crammer crammer@cis.upenn.edu
Mark Dredze mdredze@cis.upenn.edu
University of Pennsylvania, Philadelphia, PA, United States
Fernando Pereira pereira@google.com
Google, Inc., Mountain View, CA, United States

Confidence-weighted (CW) learning [6], an online learning method for linear classifiers, maintains a Gaussian distributions over weight vectors, with a covariance matrix that represents uncertainty about weights and correlations. Confidence constraints ensure that a weight vector drawn from the hypothesis distribution correctly classifies examples with a specified probability. Within this framework, we derive a new convex form of the constraint and analyze it in the mistake bound model. Empirical evaluation with both synthetic and text data shows our version of CW learning achieves lower cumulative and out-of-sample errors than commonly used first-order and second-order online methods.

Spotlight presentation, Tuesday, 5:20pm.

T27 *An Extended Level Method for Efficient Multiple Kernel Learning*

Zenglin Xu zlxu@cse.cuhk.edu.hk
The Chinese University of Hong Kong, Hong Kong
Rong Jin rongjin@cse.msu.edu
Michigan State University, East Lansing, MI, United States
Irwin King king@cse.cuhk.edu.hk
Michael Lyu lyu@cse.cuhk.edu.hk
The Chinese University of Hong Kong, Hong Kong

We consider the problem of multiple kernel learning (MKL), which can be formulated as a convex-concave problem. In the past, two efficient methods, i.e., Semi-Infinite Linear Programming (SILP) and Subgradient Descent (SD), have been proposed for large-scale multiple kernel learning. Despite their success, both methods have their own shortcomings: (a) the SD method utilizes the gradient of only the current solution, and (b) the SILP method does not regularize the approximate solution obtained from the cutting plane model. In this work, we extend the level method, which was originally designed for optimizing non-smooth objective functions, to convex-concave optimization, and apply it to multiple kernel learning. The extended level method overcomes the drawbacks of SILP and SD by exploiting all the gradients computed in past iterations and by regularizing the solution via a projection to a level set. Empirical study with eight UCI datasets shows that the extended level method can significantly improve efficiency by saving on average 91.9% of computational time over the SILP method and 70.3% over the SD method.

T28 *Learning Bounded Treewidth Bayesian Networks*

Gal Elidan galel@huji.ac.il
Hebrew University, Jerusalem, Israel
Stephen Gould sgould@cs.stanford.edu
Stanford University, Stanford, CA, United States

With the increased availability of data for complex domains, it is desirable to learn Bayesian network structures that are sufficiently expressive for generalization while also allowing for tractable inference. While the method of thin junction trees can, in principle, be used for this purpose, its fully greedy nature makes it prone to overfitting, particularly when data is scarce. In this work we present a novel method for learning Bayesian networks of bounded treewidth that employs global structure modifications and that is polynomial in the size of the graph and the treewidth bound. At the heart of our method is a triangulated graph that we dynamically update in a way that facilitates the addition of chain structures that increase the bound on the model's treewidth by at most one. We demonstrate the effectiveness of our "treewidth-friendly" method on several real-life datasets. Importantly, we also show that by using global operators, we are able to achieve better generalization even when learning Bayesian networks of unbounded treewidth.

Spotlight presentation, Tuesday, 3:20pm.

T29 *Bayesian Network Score Approximation using a Metagraph Kernel*

Benjamin Yackley benj@cs.unm.edu
Department of Computer Science, University of New Mexico, Albuquerque, NM, United States
Eduardo Corona corona@cims.nyu.edu
Courant Institute of Mathematical Sciences, New York University, New York, NY, United States
Terran Lane terran@cs.unm.edu
Department of Computer Science, University of New Mexico, Albuquerque, NM, United States

Many interesting problems, including Bayesian network structure-search, can be cast in terms of finding the optimum value of a function over the space of graphs. However, this function is often expensive to compute exactly. We here present a method derived from the study of reproducing-kernel Hilbert spaces which takes advantage of the regular structure of the space of all graphs on a fixed number of nodes to obtain approximations to the desired function quickly and with reasonable accuracy. We then test this method on both a small testing set and a real-world Bayesian network; the results suggest that not only is this method reasonably accurate, but that the BDe score itself varies quadratically over the space of all graphs.

Spotlight presentation, Tuesday, 3:20pm.

T30 *Partially Observed Maximum Entropy Discrimination Markov Networks*

Jun Zhu jun-zhu@mails.tsinghua.edu.cn
Department of Computer Science and Technology, Tsinghua University, Beijing, China
Eric P. Xing epxing@cs.cmu.edu
Bo Zhang dcszb@tsinghua.edu.cn
School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, United States

Learning graphical models with hidden variables can offer semantic insights to complex data and lead to salient structured predictors without relying on expensive, sometime unattainable fully annotated training data. While likelihood-based methods have been extensively explored, to our knowledge, learning structured prediction models with latent variables based on the max-margin principle remains largely an open problem. In this paper, we present a partially observed Maximum Entropy Discrimination Markov Network (PoMEN) model that attempts to combine the advantages of Bayesian and margin based paradigms for learning Markov networks from partially labeled data. PoMEN leads to an averaging prediction rule that resembles a Bayes predictor that is more robust to overfitting, but is also built on the desirable discriminative laws resemble those of the M^3N. We develop an EM-style algorithm utilizing existing convex optimization algorithms for M^3N as a subroutine. We demonstrate competent performance of PoMEN over existing methods on a real-world web data extraction task.

T31 *Near-minimax recursive density estimation on the binary hypercube*

Maxim Raginsky m.raginsky@duke.edu
Duke University, Durham, NC, United States
Svetlana Lazebnik lazebnik@cs.unc.edu
University of North Carolina at Chapel Hill, Chapel Hill, NC, United States
Rebecca Willett willett@duke.edu
Jorge Silva jg.silva@duke.edu
Duke University, Durham, NC, United States

This paper describes a recursive estimation procedure for multivariate binary densities using orthogonal expansions. For d covariates, there are 2^d basis coefficients to estimate, which renders conventional approaches computationally prohibitive when d is large. However, for a wide class of densities that satisfy a certain sparsity condition, our estimator runs in probabilistic polynomial time and adapts to the unknown sparsity of the underlying density in two key ways: (1) it attains near-minimax mean-squared error, and (2) the computational complexity is lower for sparser densities. Our method also allows for flexible control of the trade-off between mean-squared error and computational complexity.

Spotlight presentation, Tuesday, 3:20pm.

T32 *Support Vector Machines with a Reject Option*

Yves Grandvalet yves.grandvalet@utc.fr
Heudiasyc, CNRS and Université de Technologie de Compiègne, France
Alain Rakotomamonjy alain.rakoto@insa-rouen.fr
LITIS EA 4108, INSA de Rouen, Saint Etienne du Rouvray, France
Joseph Keshet jkeshet@idiap.ch
Idiap Research Institute, Martigny, Switzerland
Stephane Canu stephane.canu@insa-rouen.fr
LITIS EA 4108, INSA de Rouen, Saint Etienne du Rouvray, France

We consider the problem of binary classification where the classifier may abstain instead of classifying each observation. The Bayes decision rule for this setup, known as Chow's rule, is defined by two thresholds on posterior probabilities. From simple desiderata, namely the consistency and the sparsity of the classifier, we derive the double hinge loss function that focuses on estimating conditional probabilities only in the vicinity of the threshold points of the optimal decision rule. We show that, for suitable kernel machines, our approach is universally consistent. We cast the problem of minimizing the double hinge loss as a quadratic program akin to the standard SVM optimization problem and propose an active set method to solve it efficiently. We finally provide preliminary experimental results illustrating the interest of our constructive approach to devising loss functions.

T33 *Sparse Online Learning via Truncated Gradient*

John Langford jl@yahoo-inc.com
Yahoo, White Plains, NY, United States
Lihong Li lihong@cs.rutgers.edu
Department of Computer Science, Rutgers University, Piscataway, NJ, United States
Tong Zhang tongz@rci.rutgers.edu
Department of Statistics, Rutgers University, Piscataway, NJ, United States

We propose a general method called truncated gradient to induce sparsity in the weights of online-learning algorithms with convex loss. This method has several essential properties. First, the degree of sparsity is continuous—a parameter controls the rate of sparsification from no sparsification to total sparsification. Second, the approach is theoretically motivated, and an instance of it can be regarded as an online counterpart of the popular L_1-regularization method in the batch setting. We prove that small rates of sparsification result in only small additional regret with respect to typical online-learning guarantees. Finally, the approach works well empirically. We apply it to several datasets and find that for datasets with large numbers of features, substantial sparsity is discoverable.

Spotlight presentation, Tuesday, 5:20pm.

T34 *On the Generalization Ability of Online Strongly Convex Programming Algorithms*

Sham M Kakade sham@tti-c.org
Ambuj Tewari tewari@tti-c.org
Toyota Technological Institute, Chicago, IL, United States

This paper examines the generalization properties of online convex programming algorithms when the loss function is Lipschitz and strongly convex. Our main result is a sharp bound, that holds with high probability, on the excess risk of the output of an online algorithm in terms of the average regret. This allows one to use recent algorithms with logarithmic cumulative regret guarantees to achieve fast convergence rates for the excess risk with high probability. The bound also solves an open problem regarding the convergence rate of Pegasos, a recently proposed method for solving the SVM optimization problem.

Spotlight presentation, Tuesday, 5:20pm.

T35 *Weighted Sums of Random Kitchen Sinks: Replacing minimization with randomization in learning*

Ali Rahimi ali.rahimi@intel.com
Intel Research, Berkeley, CA, United States
Benjamin Recht brecht@ist.caltech.edu
California Institute of Technology, Pasadena, CA, United States

Randomized neural networks are immortalized in this AI Koan:

In the days when Sussman was a novice, Minsky once came to him as he sat hacking at the PDP-6.

"What are you doing?" asked Minsky. "I am training a randomly wired neural net to play tic-tac-toe," Sussman replied. "Why is the net wired randomly?" asked Minsky. Sussman replied, "I do not want it to have any preconceptions of how to play."

Minsky then shut his eyes. "Why do you close your eyes?" Sussman asked his teacher. "So that the room will be empty," replied Minsky. At that moment, Sussman was enlightened.

We analyze shallow random networks with the help of concentration of measure inequalities. Specifically, we consider architectures that compute a weighted sum of their inputs after passing them through a bank of arbitrary randomized nonlinearities. We identify conditions under which these networks exhibit good classification performance, and bound their test error in terms of the size of the dataset and the number of random nonlinearities.

Spotlight presentation, Tuesday, 5:20pm.

T36 *On the Design of Loss Functions for Classification: theory, robustness to outliers, and SavageBoost*

Hamed Masnadi-Shirazi hamed4444@yahoo.com
UCSD, San Diego, CA, United States
Nuno Vasconcelos nuno@ucsd.edu
University of California San Diego, La Jolla, CA, United States

The machine learning problem of classifier design is studied from the perspective of probability elicitation, in statistics. This shows that the standard approach of proceeding from the specification of a loss, to the minimization of conditional risk is overly restrictive. It is shown that a better alternative is to start from the specification of a functional form for the minimum conditional risk, and derive the loss function. This has various consequences of practical interest, such as showing that 1) the widely adopted practice of relying on convex loss functions is unnecessary, and 2) many new losses can be derived for classification problems. These points are illustrated by the derivation of a new loss which is not convex, but does not compromise the computational tractability of classifier design, and is robust to the contamination of data with outliers. A new boosting algorithm, SavageBoost, is derived for the minimization of this loss. Experimental results show that it is indeed less sensitive to outliers than conventional methods, such as Ada, Real, or LogitBoost, and converges in fewer iterations.

Spotlight presentation, Tuesday, 5:20pm.

T37 *Linear Classification and Selective Sampling Under Low Noise Conditions*

Giovanni Cavallanti cavallanti@dsi.unimi.it
Nicolò Cesa-Bianchi cesa-bianchi@dsi.unimi.it
Università degli Studi di Milano, Milano, Italy
Claudio Gentile claudio.gentile@uninsubria.it
DICOM, Universita dell' Insubria, Varese, Italy

We provide a new analysis of an efficient margin-based algorithm for selective sampling in classification problems. Using the so-called Tsybakov low noise condition to parametrize the instance distribution, we show bounds on the convergence rate to the Bayes risk of both the fully supervised and the selective sampling versions of the basic algorithm. Our analysis reveals that, excluding logarithmic factors, the average risk of the selective sampler converges to the Bayes risk at rate $n^{-(1+\alpha)/(3+\alpha)}$, with labels being sampled at the same rate (here n denotes the sample size, and $\alpha{>}0$ is the exponent in the low noise condition). We compare this convergence rate to the rate $n^{-(1+\alpha)/(2+\alpha)}$ achieved by the fully supervised algorithm using all labels. Experiments on textual data reveal that simple variants of the proposed selective sampler perform much better than popular and similarly efficient competitors.

T38 *From Online to Batch Learning with Cutoff-Averaging*

Ofer Dekel oferd@microsoft.com
Microsoft Research, Redmond, WA, United States

We present "cutoff averaging", a technique for converting any conservative online learning algorithm into a batch learning algorithm. Most online-to-batch conversion techniques work well with certain types of online learning algorithms and not with others, whereas cutoff averaging explicitly tries to adapt to the characteristics of the online algorithm being converted. An attractive property of our technique is that it preserves the efficiency of the original online algorithm, making it approporiate for large-scale learning problems. We provide a statistical analysis of our technique and back our theoretical claims with experimental results.

T39 *Adaptive Martingale Boosting*

Phil Long plong@google.com
Google, Mountain View, CA, United States
Rocco A. Servedio rocco@cs.columbia.edu
Department of Computer Science, Columbia University, New York, NY, United States

In recent work Long and Servedio [LS05short] presented a "martingale boosting" algorithm that works by constructing a branching program over weak classifiers and has a simple analysis based on elementary properties of random walks. [LS05short] showed that this martingale booster can tolerate random classification noise when it is run with a noise-tolerant weak learner; however, a drawback of the algorithm is that it is not *adaptive*, i.e. it cannot effectively take advantage of variation in the quality of the weak classifiers it receives.

In this paper we present a variant of the original martingale boosting algorithm and prove that it is adaptive. This adaptiveness is achieved by modifying the original algorithm so that the random walks that arise in its analysis have different step size depending on the quality of the weak learner at each stage. The new algorithm inherits the desirable properties of the original [LS05short] algorithm, such as random classification noise tolerance, and has several other advantages besides adaptiveness: it requires polynomially fewer calls to the weak learner than the original algorithm, and it can be used with confidence-rated weak hypotheses that output real values rather than Boolean predictions.

Spotlight presentation, Tuesday, 5:20pm.

T40 *Mind the Duality Gap: Logarithmic regret algorithms for online optimization*

Shai Shalev-Shwartz shai@tti-c.org
Sham M Kakade sham@tti-c.org
Toyota Technological Institute, Chicago, IL, United States

We describe a primal-dual framework for the design and analysis of online strongly convex optimization algorithms. Our framework yields the tightest known logarithmic regret bounds for Follow-The-Leader and for the gradient descent algorithm proposed in [HazanKaKaAg06]. We then show that one can interpolate between these two extreme cases. In particular, we derive a new algorithm that shares the computational simplicity of gradient descent but achieves lower regret in many practical situations. Finally, we further extend our framework for generalized strongly convex functions.

Spotlight presentation, Tuesday, 11:50am.

T41 *On Bootstrapping the ROC Curve*

Patrice Bertail patrice.bertail@ensae.fr
Modalx, Université Paris X & CREST, France
Stéphan J.M. Clémençon stephan.clemencon@telecom-paristech.fr
LTCI, Telecom ParisTech and CNRS, Institut Telecom, PARIS, France
Nicolas Vayatis nicolas.vayatis@gmail.com
Ecole Normale Supérieure de Cachan, CMLA - CNRS & UniverSud, Cachan, France

This paper is devoted to thoroughly investigating how to bootstrap the ROC curve, a widely used visual tool for evaluating the accuracy of test/scoring statistics in the bipartite setup. The issue of confidence bands for the ROC curve is considered and a resampling procedure based on a smooth version of the empirical distribution called the "smoothed bootstrap" is introduced. Theoretical arguments and simulation results are presented to show that the "smoothed bootstrap" is preferable to a "naive" bootstrap in order to construct accurate confidence bands.

T42 *DiscLDA: Discriminative Learning for Dimensionality Reduction and Classification*

Simon Lacoste-Julien slacoste@cs.berkeley.edu
University of California, Berkeley, CA, United States
Fei Sha feisha@usc.edu
University of Southern California, Los Angeles, CA, United States
Michael Jordan jordan@cs.berkeley.edu
University of California, Berkeley, CA, United States

Probabilistic topic models (and their extensions) have become popular as models of latent structures in collections of text documents or images. These models are usually treated as generative models and trained using maximum likelihood estimation, an approach which may be suboptimal in the context of an overall classification problem. In this paper, we describe DiscLDA, a discriminative learning framework for such models as Latent Dirichlet Allocation (LDA) in the setting of dimensionality reduction with supervised side information. In DiscLDA, a class-dependent linear transformation is introduced on the topic mixture proportions. This parameter is estimated by maximizing the conditional likelihood using Monte Carlo EM. By using the transformed topic mixture proportions as a new representation of documents, we obtain a supervised dimensionality reduction algorithm that uncovers the latent structure in a document collection while preserving predictive power for the task of classification. We compare the predictive power of the latent structure of DiscLDA with unsupervised LDA on the 20 Newsgroup document classification task.

T43 *Posterior Consistency of the Silverman g-prior in Bayesian Model Choice*

zhihua zhang zhzhang@gmail.com
Michael I. Jordan jordan@cs.berkeley.edu
University of California, Berkeley, Berkeley, CA, United States
Dit-Yan Yeung dyyeung@cse.ust.hk
Hong Kong University of Science & Technology, Hong Kong, China

Kernel supervised learning methods can be unified by utilizing the tools from regularization theory. The duality between regularization and prior leads to interpreting regularization methods in terms of maximum a posteriori estimation and has motivated Bayesian interpretations of kernel methods. In this paper we pursue a Bayesian interpretation of sparsity in the kernel setting by making use of a mixture of a point-mass distribution and prior that we refer to as "Silverman's g-prior." We provide a theoretical analysis of the posterior consistency of a Bayesian model choice procedure based on this prior. We also establish the asymptotic relationship between this procedure and the Bayesian information criterion.

Spotlight presentation, Tuesday, 3:20pm.

T44 *Dimensionality Reduction for Data in Multiple Feature Representations*

Yen-Yu Lin yylin@iis.sinica.edu.tw
Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan
Tyng-Luh Liu liutyng@iis.sinica.edu.tw
IIS, Academia Sinica, Nankang, Taipei, Taiwan
Chiou-Shann Fuh fuh@csie.ntu.edu.tw
Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan

In solving complex visual learning tasks, adopting multiple descriptors to more precisely characterize the data has been a feasible way for improving performance. These representations are typically high dimensional and assume diverse forms. Thus finding a way to transform them into a unified space of lower dimension generally facilitates the underlying tasks, such as object recognition or clustering. We describe an approach that incorporates multiple kernel learning with dimensionality reduction (MKL-DR). While the proposed framework is flexible in simultaneously tackling data in various feature representations, the formulation itself is general in that it is established upon graph embedding. It follows that any dimensionality reduction techniques explainable by graph embedding can be generalized by our method to consider data in multiple feature representations.

T45 *Natural Image Denoising with Convolutional Networks*

Viren Jain viren@mit.edu
Sebastian Seung seung@mit.edu
MIT, Cambridge, MA, United States

We present an approach to low-level vision that combines two main ideas: the use of convolutional networks as an image processing architecture and an unsupervised learning procedure that synthesizes training samples from specific noise models. We demonstrate this approach on the challenging problem of natural image denoising. Using a test set with a hundred natural images, we find that convolutional networks provide comparable and in some cases superior performance to state of the art wavelet and Markov random field (MRF) methods. Moreover, we find that a convolutional network offers similar performance in the blind denoising setting as compared to other techniques in the non-blind setting. We also show how convolutional networks are mathematically related to MRF approaches by presenting a mean field theory for an MRF specially designed for image denoising. Although these approaches are related, convolutional networks avoid computational difficulties in MRF approaches that arise from probabilistic learning and inference. This makes it possible to learn image processing architectures that have a high degree of representational power (we train models with over 15,000 parameters), but whose computational expense is significantly less than that associated with inference in MRF approaches with even hundreds of parameters.

T46 *Resolution Limits of Sparse Coding in High Dimensions*

Alyson Fletcher alyson@eecs.berkeley.edu
University of California, Berkeley, Berkeley, CA, United States
Sundeep Rangan srangan@qualcomm.com
Qualcomm, Bridgewater, NJ, United States
Vivek Goyal vgoyal@mit.edu
Massachusetts Institute of Technology, Cambridge, MA, United States

This paper addresses the problem of sparsity pattern detection for unknown k-sparse n-dimensional signals observed through m noisy, random linear measurements. Sparsity pattern recovery arises in a number of settings including statistical model selection, pattern detection, and image acquisition. The main results in this paper are necessary and sufficient conditions for asymptotically-reliable sparsity pattern recovery in terms of the dimensions m, n and k as well as the signal-to-noise ratio (SNR) and the minimum-to-average ratio (MAR) of the nonzero entries of the signal. With the SNR and MAR fixed, it is shown that the scaling $m = O(k \log(n - k))$ is necessary for any algorithm to succeed, and also sufficient for a computationally-trivial correlation algorithm. Moreover, the trivial algorithm requires at most 4(1+SNR) times more measurements than maximum likelihood and 4/MAR times more measurements than the well-known lasso method. This provides insight on the precise value and limitations of convex programming-based algorithms.

T47 *Efficient Exact Inference in Planar Ising Models*

Nicol N. Schraudolph nic@schraudolph.org
Dmitry Kamenetsky dmitry.kamenetsky@nicta.com.au
National ICT Australia (NICTA), Canberra, Australia

We give polynomial-time algorithms for the exact computation of lowest-energy states, worst margin violators, partition functions, and marginals in certain binary undirected graphical models. Our approach provides an interesting alternative to the well-known graph cut paradigm in that it does not impose any submodularity constraints; instead we require planarity to establish a correspondence with perfect matchings in an expanded dual graph. Maximum-margin parameter estimation for a boundary detection task shows our approach to be efficient and effective.

T48 *Improving on Expectation Propagation*

Manfred Opper opperm@cs.tu-berlin.de
TU Berlin, Berlin, Germany
Ulrich Paquet ulrich@cantab.net
University of Cambridge, Cambridge, United Kingdom
Ole Winther owi@imm.dtu.dk
IMM, DTU, Bioinformatics Center, University of Copenhagen, Lyngby, Denmark

A series of corrections is developed for the fixed points of Expectation Propagation (EP), which is one of the most popular methods for approximate probabilistic inference. These corrections can lead to improvements of the inference approximation or serve as a sanity check, indicating when EP yields unrealiable results.

Spotlight presentation, Tuesday, 3:20pm.

T49 *Reconciling Real Scores with Binary Comparisons: A New Logistic Based Model for Ranking*

Nir Ailon nailon@google.com
Google Research NY, New York, NY, United States

The problem of ranking arises ubiquitously in almost every aspect of life. A statistical model for ranking predicts how humans rank subsets V of some universe U. In this work we define a statistical model for ranking that has the following desirable properties: (i) It is gracefully parametrized by a real valued function ("score") on U, (ii) The score parameters can be efficiently learnt from sample data consisting of a sequence of comparison bits (either "A precedes B" or "B precedes A"), (iii) For any subset V of U, the ranking which sorts the elements by score is the mode (most probable outcome) of the model distribution, (iv) It predicts the empirical strong independence between human response to comparison questions and the context in which the questions are asked, and (v) It admits an efficient $O(n \log n)$ time sampling algorithm for subsets of size n.

In the first part of this work, a family of ranking models is built axiomatically based on very simple desirable properties defined locally for comparisons, and automatically implies the existence of a global score function serving as a natural model parameter which can be efficiently fitted to data by solving a convex optimization problem.

In the second part , motivated by IR/ML application, we present a previously unstudied member of the model family defined in the first part, and compare it to other well known models from the econometric/statistical literature which were originally motivated by other problems.

T50 *Efficient Sampling for Gaussian Process Inference using Control Variables*

Michalis Titsias mtitsias@cs.man.ac.uk
University of Manchester, Manchester, United Kingdom
Neil D Lawrence neill@cs.man.ac.uk
School of Computer Science, University of Manchester, Manchester
Magnus Rattray magnus@cs.man.ac.uk
University of Manchester, Manchester

Sampling functions in Gaussian process (GP) models is challenging because of the highly correlated posterior distribution. We describe an efficient Markov chain Monte Carlo algorithm for sampling from the posterior process of the GP model. This algorithm uses control variables which are auxiliary function values that provide a low dimensional representation of the function. At each iteration, the algorithm proposes new values for the control variables and generates the function from the conditional GP prior. The control variable input locations are found by minimizing an objective function. We demonstrate the algorithm on regression and classification problems and we use it to estimate the parameters of a differential equation model of gene regulation.

Spotlight presentation, Tuesday, 3:20pm.

T51 *MAS: a multiplicative approximation scheme for probabilistic inference*

Ydo Wexler and Christopher Meek, Microsoft Research.
Oral presentation, Tuesday, 3:00pm. See abstract, page 102.

T52 *The Gaussian Process Density Sampler*

Ryan Adams rpa23@cam.ac.uk
Cavendish Laboratory University of Cambridge Cambridge, UK
Iain Murray murray@cs.toronto.edu
Department of Computer Science University of Toronto Toronto, ON, Canada
David MacKay mackay@mrao.cam.ac.uk
Cavendish Laboratory University of Cambridge Cambridge, UK

We present the Gaussian Process Density Sampler (GPDS), an exchangeable generative model for use in nonparametric Bayesian density estimation. Samples drawn from the GPDS are consistent with exact, independent samples from a fixed density function that is a transformation of a function drawn from a Gaussian process prior. Our formulation allows us to infer an unknown density from data using Markov chain Monte Carlo, which gives samples from the posterior distribution over density functions and from the predictive distribution on data space. We can also infer the hyperparameters of the Gaussian process. We compare this density modeling technique to several existing techniques on a toy problem and a skull-reconstruction task.

Spotlight presentation, Tuesday, 3:20pm.

T53 *Bayesian Experimental Design of Magnetic Resonance Imaging Sequences*

Matthias W Seeger mseeger@mmci.uni-saarland.de
Max Planck Institute for Informatics, Saarland University, Saarbruecken, Germany
Hannes Nickisch hn@tuebingen.mpg.de
Rolf Pohmann rolf.pohmann@tuebingen.mpg.de
Bernhard Schölkopf bs@tuebingen.mpg.de
MPI for Biological Cybernetics, Tübingen, Germany

We show how improved sequences for magnetic resonance imaging can be found through automated optimization of Bayesian design scores. Combining recent advances in approximate Bayesian inference and natural image statistics with high-performance numerical computation, we propose the first scalable Bayesian experimental design framework for this problem of high relevance to clinical and brain research. Our solution requires approximate inference for dense, non-Gaussian models on a scale seldom addressed before. We propose a novel scalable variational inference algorithm, and show how powerful methods of numerical mathematics can be modified to compute primitives in our framework. Our approach is evaluated on a realistic setup with raw data from a 3T MR scanner.

Spotlight presentation, Tuesday, 9:50am.

T54 *Bounds on marginal probability distributions*

Joris M Mooij joris.mooij@tuebingen.mpg.de
Max Planck Institute for Biological Cybernetics, Tübingen, Germany
Hilbert J Kappen b.kappen@science.ru.nl
Radboud University, Nijmegen, Netherlands

We propose a novel bound on single-variable marginal probability distributions in factor graphs with discrete variables. The bound is obtained by propagating bounds (convex sets of probability distributions) over a subtree of the factor graph, rooted in the variable of interest. By construction, the method not only bounds the exact marginal probability distribution of a variable, but also its approximate Belief Propagation marginal ("belief"). Thus, apart from providing a practical means to calculate bounds on marginals, our contribution also lies in providing a better understanding of the error made by Belief Propagation. We show that our bound outperforms the state-of-the-art on some inference problems arising in medical diagnosis.

Spotlight presentation, Tuesday, 3:20pm.

T55 *Clusters and Coarse Partitions in LP Relaxations*

David Sontag dsontag@csail.mit.edu
Massachusetts Institute of Technology, Cambridge, MA, United States
Amir Globerson gamir@csail.mit.edu
The Hebrew University of Jerusalem, Jerusalem, Israel
Tommi Jaakkola tommi@csail.mit.edu
Massachusetts Institute of Technology, Cambridge, MA, United States

We propose a new class of consistency constraints for Linear Programming (LP) relaxations for finding the most probable (MAP) configuration in graphical models. Usual cluster-based LP relaxations enforce joint consistency of the beliefs of a cluster of variables, with computational cost increasing exponentially with the size of the clusters. By *partitioning* the state space of a cluster and enforcing consistency only across partitions, we obtain a class of constraints which, although less tight, are computationally feasible for large clusters. We show how to solve the cluster selection and partitioning problem monotonically in the dual LP, using the current beliefs to guide these choices. We obtain a dual message-passing algorithm and apply it to protein design problems where the variables have large state spaces and the usual cluster-based relaxations are very costly.

Spotlight presentation, Tuesday, 3:20pm.

T56 *MCBoost: Multiple Classifier Boosting for Perceptual Co-clustering of Images and Visual Features*

Tae-Kyun Kim tkk22@cam.ac.uk
Roberto Cipolla rc10001@cam.ac.uk
University of Cambridge, Cambridge, United Kingdom

We present a new co-clustering problem of images and visual features. The problem involves a set of non-object images in addition to a set of object images and features to be co-clustered. Co-clustering is performed in a way that maximises discrimination of object images from non-object images, thus emphasizing discriminative features. This provides a way of obtaining perceptual joint-clusters of object images and features. We tackle the problem by simultaneously boosting multiple strong classifiers which compete for images by their expertise. Each boosting classifier is an aggregation of weak-learners, i.e. simple visual features. The obtained classifiers are useful for multi-category and multi-view object detection tasks. Experiments on a set of pedestrian images and a face data set demonstrate that the method yields intuitive image clusters with associated features and is much superior to conventional boosting classifiers in object detection tasks.

T57 *Learning Hybrid Models for Image Annotation with Partially Labeled Data*

Xuming He hexm@stat.ucla.edu
University of California, Los Angeles, Los Angeles, CA, United States
Richard Zemel zemel@cs.toronto.edu
University of Toronto, Toronto, ON, Canada

Extensive labeled data for image annotation systems, which learn to assign class labels to image regions, is difficult to obtain. We explore a hybrid model framework for utilizing partially labeled data that integrates a generative topic model for image appearance with discriminative label prediction. We propose three alternative formulations for imposing a spatial smoothness prior on the image labels. Tests of the new models and some baseline approaches on two real image datasets demonstrate the effectiveness of incorporating the latent structure.

T58 *Grouping Contours Via a Related Image*

Praveen Srinivasan psrin@seas.upenn.edu
GRASP Laboratory, University of Pennsylvania, Philadelphia, PA, United States
Liming Wang wanglm@fudan.edu.cn
Department of Computer Science and Engineering, Fudan University, Shanghai, China
Jianbo Shi jshi@cis.upenn.edu
GRASP Laboratory, University of Pennsylvania, Philadelphia, PA, United States

Contours have been established in the biological and computer vision literatures as a compact yet descriptive representation of object shape. While individual contours provide structure, they lack the large spatial support of region segments (which lack internal structure). We present a method for further grouping of contours in an image using their relationship to the contours of a second, related image. Stereo, motion, and similarity all provide cues that can aid this task; contours that have similar transformations relating them to their matching contours in the second image likely belong to a single group. To find matches for contours, we rely only on shape, which applies directly to all three modalities without modification, in constrant to the specialized approaches developed for each independently. Visually salient contours are extracted in each image, along with a set of candidate transformations for aligning subsets of them. For each transformation, groups of contours with matching shape across the two images are identified to provide a context for evaluating matches of individual contour points across the images. The resulting

contexts of contours are used to perform a final grouping on contours in the original image while simultaneously finding matches in the related image, again by shape matching. We demonstrate grouping results on image pairs consisting of stereo, motion, and similar images. Our method also produces qualitatively better results against a baseline method that does not use the inferred contexts.

T59 *Improved Moves for Truncated Convex Models*

M Pawan Kumar pawan@robots.ox.ac.uk
University of Oxford, Oxford, United Kingdom
Philip H S Torr philiptorr@brookes.ac.uk
Oxford Brookes University, Oxford, United Kingdom

We consider the problem of obtaining the approximate maximum a posteriori estimate of a discrete random field characterized by pairwise potentials that form a truncated convex model. For this problem, we propose an improved st-mincut based move making algorithm. Unlike previous move making approaches, which either provide a loose bound or no bound on the quality of the solution (in terms of the corresponding Gibbs energy), our algorithm achieves the same guarantees as the standard linear programming (LP) relaxation. Compared to previous approaches based on the LP relaxation, e.g. interior-point algorithms or tree-reweighted message passing (TRW), our method is faster as it uses only the efficient st-mincut algorithm in its design. Furthermore, it directly provides us with a primal solution (unlike TRW and other related methods which solve the dual of the LP). We demonstrate the effectiveness of the proposed approach on both synthetic and standard real data problems.

Our analysis also opens up an interesting question regarding the relationship between move making algorithms (such as α-expansion and the algorithms presented in this paper) and the randomized rounding schemes used with convex relaxations. We believe that further explorations in this direction would help design efficient algorithms for more complex relaxations.

Spotlight presentation, Tuesday, 3:20pm.

T60 *Exploring Large Feature Spaces with Hierarchical Multiple Kernel Learning*

Francis Bach francis.bach@mines.org
INRIA - Ecole Normale Superieure, Paris, France

For supervised and unsupervised learning, positive definite kernels allow to use large and potentially infinite dimensional feature spaces with a computational cost that only depends on the number of observations. This is usually done through the penalization of predictor functions by Euclidean or Hilbertian norms. In this paper, we explore penalizing by sparsity-inducing norms such as the ℓ^1-norm or the block ℓ^1-norm. We assume that the kernel decomposes into a large sum of individual basis kernels which can be embedded in a directed acyclic graph; we show that it is then possible to perform kernel selection through a hierarchical multiple kernel learning framework, in polynomial time in the number of selected kernels. This framework is naturally applied to non linear variable selection; our extensive simulations on synthetic datasets and datasets from the UCI repository show that efficiently exploring the large feature space through sparsity-inducing norms leads to state-of-the-art predictive performance.

T61 *Nonlinear causal discovery with additive noise models*

Patrik O. Hoyer patrik.hoyer@helsinki.fi
University of Helsinki, Helsinki Institute for Information Technology, Helsinki, Finland
Dominik Janzing dominik.janzing@tuebingen.mpg.de
Joris M Mooij joris.mooij@tuebingen.mpg.de
Jonas Peters jonas.peters@tuebingen.mpg.de
Bernhard Schölkopf bs@tuebingen.mpg.de
Max-Planck Institute for Biological Cybernetics, Tübingen, Germany

The discovery of causal relationships between a set of observed variables is a fundamental problem in science. For continuous-valued data *linear* acyclic causal models are often used because these models are well understood and there are well-known methods to fit them to data. In reality, of course, many causal relationships are more or less *nonlinear*, raising some doubts as to the applicability and usefulness of purely linear methods. In this contribution we show that in fact the basic linear framework can be generalized to nonlinear models with additive noise. In this extended framework, nonlinearities in the data-generating process are in fact a blessing rather than a curse, as they typically provide information on the underlying causal system and allow more aspects of the true data-generating mechanisms to be identified. In addition to theoretical results we show simulations and some simple real data experiments illustrating the identification power provided by nonlinearities.

Spotlight presentation, Tuesday, 5:20pm.

T62 *Predicting the Geometry of Metal Binding Sites from Protein Sequence*

Paolo Frasconi p-f@dsi.unifi.it
Università degli Studi di Firenze, Firenze, Italy
Andrea Passerini passerini@disi.unitn.it
Università degli Studi di Trento, Trento, Italy

Metal binding is important for the structural and functional characterization of proteins. Previous prediction efforts have only focused on bonding state, i.e. deciding which protein residues act as metal ligands in some binding site. Identifying the geometry of metal-binding sites, i.e. deciding which residues are jointly involved in the coordination of a metal ion is a new prediction problem that has been never attempted before from protein sequence alone. In this paper, we formulate it in the framework of learning with structured outputs. Our solution relies on the fact that, from a graph theoretical perspective, metal binding has the algebraic properties of a matroid, enabling the application of greedy algorithms for learning structured outputs. On a data set of 199 non-redundant metalloproteins, we obtained precision/recall levels of 75%/46% correct ligand-ion assignments, which improves to 88%/88% in the setting where the metal binding state is known.

Spotlight presentation, Tuesday, 9:50am.

T63 *A mixture model for the evolution of gene expression in non-homogeneous datasets*

Gerald Quon gerald.quon@utoronto.ca
University of Toronto, Toronto, Canada
Yee Whye Teh ywteh@gatsby.ucl.ac.uk
Gatsby Computational Neuroscience Unit, UCL, London, United Kingdom
Esther Chan esther.chan@utoronto.ca
Timothy Hughes t.hughes@utoronto.ca
Molecular Genetics, University of Toronto, Toronto, ON, Canada
Michael Brudno brudno@cs.toronto.edu
Quaid Morris quaid.morris@utoronto.ca
Department of Computer Science, University of Toronto &
Donnelly CCBR, Toronto, ON, Canada

We address the challenge of assessing conservation of gene expression in complex, non-homogeneous datasets. Recent studies have demonstrated the success of probabilistic models in studying the evolution of gene expression in simple eukaryotic organisms such as yeast, for which measurements are typically scalar and independent. Models capable of studying expression evolution in much more complex organisms such as vertebrates are particularly important given the medical and scientific interest in species such as human and mouse. We present a statistical model that makes a number of significant extensions to previous models to enable characterization of changes in expression among highly complex organisms. We demonstrate the efficacy of our method on a microarray dataset containing diverse tissues from multiple vertebrate species. We anticipate that the model will be invaluable in the study of gene expression patterns in other diverse organisms as well, such as worms and insects.

T64 *Probabilistic detection of short events, with application to critical care monitoring*

Norm Aleks norm@cs.berkeley.edu
Stuart Russell russell@cs.berkeley.edu
Computer Science Divison, University of California, Berkeley, CA, United States
Michael G Madden michael.madden@nuigalway.ie
College of Engineering & Informatics, NUI Galway, Galway, Ireland
Diane Morabito morabitod@neurosurg.ucsf.edu
Department of Neurosurgery, University of California, San Francisco, CA, United States
Kristan Staudenmayer kstaudenmayer@sfghsurg.ucsf.edu
Stanford University, Stanford, CA, United States
Mitchell Cohen mcohen@sfghsurg.ucsf.edu
Department of Surgery, University of California, San Francisco, CA, United States
Geoffrey T. Manley manleyg@neurosurg.ucsf.edu
Department of Neurosurgery, University of California, San Francisco, CA, United States

We describe an application of probabilistic modeling and inference technology to the problem of analyzing sensor data in the setting of an intensive care unit (ICU). In particular, we consider the arterial-line blood pressure sensor, which is subject to frequent data artifacts that cause false alarms in the ICU and make the raw data almost useless for automated decision making. The problem is complicated by the fact that the sensor data are acquired at fixed intervals whereas the events causing data artifacts may occur at any time and have durations that may be significantly shorter than the data collection interval. We show that careful modeling of the sensor, combined with a general technique for detecting sub-interval events and estimating their duration, enables effective detection of artifacts and accurate estimation of the underlying blood pressure values.

T65 *Kernel-ARMA for Hand Tracking and Brain-Machine interfacing During 3D Motor Control*

Lavi Shpigelman, Hagai Lalazar and Eilon Vaadia, Hebrew University.
Oral presentation, Tuesday, 9:30am. See abstract, page 97.

T66 *Playing Pinball with non-invasive BCI*

Michael W Tangermann (ne Schröder) schroedm@cs.tu-berlin.de
Matthias Krauledat kraulem@cs.tu-berlin.de
Konrad Grzeska konradg@cs.tu-berlin.de
Max Sagebaum max.sagebaum@first.fraunhofer.de
Benjamin Blankertz blanker@cs.tu-berlin.de
Carmen Vidaurre vidcar@cs.tu-berlin.de
Klaus-Robert Müller krm@cs.tu-berlin.de
Berlin Institute of Technology, ML Lab, BBCI Group, Berlin, Germany

Compared to invasive Brain-Computer Interfaces (BCI), non-invasive BCI systems based on Electroencephalogram (EEG) signals have not been applied successfully for control tasks in complex environments. In the present study, however, we demonstrate that this is possible and report on the interaction of human subjects with a complex real device: a pinball machine. First results in this initial study clearly show that fast and well-timed control well beyond chance level is possible, even though the environment is extremely rich and requires predictive behavior. Using machine learning methods for mental state decoding, BCI-based pinball control is possible within the first session without the necessity to employ lengthy subject training. The current study clearly shows that very compelling control with excellent timing and dynamics is possible for a non-invasive BCI.

T67 *A Massively Parallel Digital Learning Processor*

Hans Peter Graf hpg@nec-labs.com
Srihari Cadambi cadambi@nec-labs.com
Igor Durdanovic igord@nec-labs.com
Venkata Jakkula jakkula@nec-labs.com
Murugan Sankaradass murugs@nec-labs.com
Eric Cosatto cosatto@nec-labs.com
Srimat Chakradhar chak@nec-labs.com
NEC Laboratories America, Princeton, NJ, United States

We present a new, massively parallel architecture for accelerating machine learning algorithms, based on arrays of vector processing elements (VPE) with variable-resolution arithmetic. Groups of VPEs operate in SIMD (single instruction multiple data) mode, and each group is connected to an independent memory bank. In this way memory bandwidth scales with the number of VPE, and the main data flows are local, keeping power dissipation low. With 256 VPEs, implemented on two FPGA (field programmable gate array) chips, we obtain a sustained speed of 19 GMACS (billion multiply-accumulate per sec.) for SVM training, and 86 GMACS for SVM classification. This performance is more than an order of magnitude higher than that of any FPGA implementation reported so far. The speed on one FPGA is similar to the fastest speeds published on a Graphics Processor for the MNIST problem, despite a clock rate of the FPGA that is six times lower. Tests with Convolutional Neural Networks show similar compute performances. This massively parallel architecture is particularly attractive for embedded applications, where low power dissipation is critical.

Spotlight presentation, Tuesday, 9:50am.

T68 *Effects of Stimulus Type and of Error-Correcting Code Design on BCI Speller Performance*

Jeremy Hill jez@tuebingen.mpg.de
MPI for Biological Cybernetics, Tübingen, Germany
Jason Farquhar J.Farquhar@donders.ru.nl
Donders Institute for Brain, Cognition and Behaviour,
Radboud University Nijmegen, The Netherlands
Suzanna Martens smartens@tuebingen.mpg.de
Felix Biessmann fbiessma@tuebingen.mpg.de
Bernhard Schölkopf bs@tuebingen.mpg.de
MPI for Biological Cybernetics, Tübingen, Germany

From an information-theoretic perspective, a noisy transmission system such as a visual Brain-Computer Interface (BCI) speller could benefit from the use of error-correcting codes. However, optimizing the code solely according to the maximal minimum-Hamming-distance criterion tends to lead to an overall increase in target frequency of target stimuli, and hence a significantly reduced average target-to-target interval (TTI), leading to difficulties in classifying the individual event-related potentials (ERPs) due to overlap and refractory effects. Clearly any change to the stimulus setup must also respect the possible psychophysiological consequences. Here we report new EEG data from experiments in which we explore stimulus types and codebooks in a within-subject design, finding an interaction between the two factors. Our data demonstrate that the traditional, row-column code has particular spatial properties that lead to better performance than one would expect from its TTIs and Hamming-distances alone, but nonetheless error-correcting codes can improve performance provided the right stimulus type is used.

Spotlight presentation, Tuesday, 9:50am.

T69 *Analyzing human feature learning as nonparametric Bayesian inference*

JOSEPH AUSTERWEIL joseph.austerweil@gmail.com
THOMAS L GRIFFITHS tom_griffiths@berkeley.edu
University of California, Berkeley, Berkeley, CA, United States

Almost all successful machine learning algorithms and cognitive models require powerful representations capturing the features that are relevant to a particular problem. We draw on recent work in nonparametric Bayesian statistics to define a rational model of human feature learning that forms a featural representation from raw sensory data without prespecifying the number of features. By comparing how the human perceptual system and our rational model use distributional and category information to infer feature representations, we seek to identify some of the forces that govern the process by which people separate and combine sensory primitives to form features.

Spotlight presentation, Tuesday, 9:50am.

T70 *Stress, noradrenaline, and realistic prediction of mouse behaviour using reinforcement learning*

GEDIMINAS LUKSYS, CARMEN SANDI and WULFRAM GERSTNER, EPFL, Lausanne, Switzerland.
Oral presentation, Tuesday, 10:30am. *See abstract, page 99.*

T71 *Extracting State Transition Dynamics from Multiple Spike Trains with Correlated Poisson HMM*

KENTARO KATAHIRA katahira@mns.k.u-tokyo.ac.jp
University of Tokyo, Kashiwa, Chiba, Japan, and RIKEN Brain Science Institute, Wako, Saitama, Japan
JUN NISHIKAWA j_kohaku@sc5.so-net.ne.jp
KAZUO OKANOYA okanoya@brain.riken.jp
RIKEN Brain Science Institute, Wako, Saitama, Japan
MASATO OKADA okada@k.u-tokyo.ac.jp
University of Tokyo, Kashiwa, Chiba, Japan, and RIKEN Brain Science Institute, Wako, Saitama, Japan

Neural activity is non-stationary and varies across time. Hidden Markov Models (HMMs) have been used to track the state transition among quasi-stationary discrete neural states. Within this context, an independent Poisson model has been used for the output distribution of HMMs; hence, the model is incapable of tracking the change in correlation without modulating the firing rate. To achieve this, we applied a multivariate Poisson distribution with a correlation term for the output distribution of HMMs. We formulated a Variational Bayes (VB) inference for the model. The VB could automatically determine the appropriate number of hidden states and correlation types while avoiding the overlearning problem. We developed an efficient algorithm for computing posteriors using the recursive relationship of a multivariate Poisson distribution. We demonstrated the performance of our method on synthetic data and a real spike train recorded from a songbird.

T72 *Structure Learning in Human Sequential Decision-Making*

Daniel Acuna acuna002@umn.edu
Paul R. Schrater schrater@umn.edu
University of Minnesota, Minneapolis, MN, United States

We use graphical models and structure learning to explore how people learn policies in sequential decision making tasks. Studies of sequential decision-making in humans frequently find suboptimal performance relative to an ideal actor that knows the graph model that generates reward in the environment. We argue that the learning problem humans face also involves learning the graph structure for reward generation in the environment. We formulate the structure learning problem using mixtures of reward models, and solve the optimal action selection problem using Bayesian Reinforcement Learning. We show that structure learning in one and two armed bandit problems produces many of the qualitative behaviors deemed suboptimal in previous studies. Our argument is supported by the results of experiments that demonstrate humans rapidly learn and exploit new reward structure.

T73 *Dependent Dirichlet Process Spike Sorting*

Jan Gasthaus j.gasthaus@gatsby.ucl.ac.uk
Frank Wood fwood@gatsby.ucl.ac.uk
Dilan Gorur dilan@gatsby.ucl.ac.uk
Yee Whye Teh ywteh@gatsby.ucl.ac.uk
Gatsby Computational Neuroscience Unit, UCL, London, United Kingdom

In this paper we propose a new incremental spike sorting model that automatically eliminates refractory period violations, accounts for action potential waveform drift, and can handle "appearance" and "disappearance" of neurons. Our approach is to augment a known time-varying Dirichlet process that ties together a sequence of infinite Gaussian mixture models, one per action potential waveform observation, with an interspike-interval-dependent likelihood that prohibits refractory period violations. We demonstrate this model by showing results from sorting two publicly available neural data recordings for which a partial ground truth labeling is known.

T74 *Optimal Response Initiation: Why Recent Experience Matters*

Matt Jones mcj@colorado.edu
Michael Mozer mozer@colorado.edu
Sachiko Kinoshita skinoshi@maccs.mq.edu.au
University of Colorado, Boulder, CO, United States

In most cognitive and motor tasks, speed-accuracy tradeoffs are observed: Individuals can respond slowly and accurately, or quickly yet be prone to errors. Control mechanisms governing the initiation of behavioral responses are sensitive not only to task instructions and the stimulus being processed, but also to the recent stimulus history. When stimuli can be characterized on an easy-hard dimension (e.g., word frequency in a naming task), items preceded by easy trials are responded to more quickly, and with more errors, than items preceded by hard trials. We propose a rationally motivated mathematical model of this sequential adaptation of control, based on a diffusion model of the decision process in

which difficulty corresponds to the drift rate for the correct response. The model assumes that responding is based on the posterior distribution over which response is correct, conditioned on the accumulated evidence. We derive this posterior as a function of the drift rate, and show that higher estimates of the drift rate lead to (normatively) faster responding. Trial-by-trial tracking of difficulty thus leads to sequential effects in speed and accuracy. Simulations show the model explains a variety of phenomena in human speeded decision making. We argue this passive statistical mechanism provides a more elegant and parsimonious account than extant theories based on elaborate control structures.

Spotlight presentation, Tuesday, 9:50am.

T75 *How memory biases affect information transmission: A rational analysis of serial reproduction*

Jing Xu jing.xu@berkeley.edu
Thomas L Griffiths tom_griffiths@berkeley.edu
University of California, Berkeley, Berkeley, CA, United States

Many human interactions involve pieces of information being passed from one person to another, raising the question of how this process of information transmission is affected by the capacities of the agents involved. In the 1930s, Sir Frederic Bartlett explored the influence of memory biases in “serial reproduction” of information, in which one person’s reconstruction of a stimulus from memory becomes the stimulus seen by the next person. These experiments were done using relatively uncontrolled stimuli such as pictures and stories, but suggested that serial reproduction would transform information in a way that reflected the biases inherent in memory. We formally analyze serial reproduction using a Bayesian model of reconstruction from memory, giving a general result characterizing the effect of memory biases on information transmission. We then test the predictions of this account in two experiments using simple one-dimensional stimuli. Our results provide theoretical and empirical justification for the idea that serial reproduction reflects memory biases.

Spotlight presentation, Tuesday, 9:50am.

T76 *Sequential effects: Superstition or rational behavior?*

Angela J Yu ajyu@ucsd.edu
University of California, San Diego, La Jolla, CA, United States
Jonathan D. Cohen jdc@princeton.edu
Department of Psychology and Princeton Neuroscience Institute,
Princeton University, Princeton, NJ, United States

In a variety of behavioral tasks, subjects exhibit an automatic and apparently sub-optimal sequential effect: they respond more rapidly and accurately to a stimulus if it reinforces a local pattern in stimulus history, such as a string of repetitions or alternations, compared to when it violates such a pattern. This is often the case even if the local trends arise by chance in the context of a randomized design, such that stimulus history has no predictive power. In this work, we use a normative Bayesian framework to examine the hypothesis that such idiosyncrasies may reflect the inadvertent engagement of fundamental mechanisms critical for adapting to changing statistics in the natural environment. We show that prior belief in non-stationarity can induce experimentally observed sequential effects

in an otherwise Bayes-optimal algorithm. The Bayesian algorithm is shown to be well approximated by linear-exponential filtering of past observations, a feature also apparent in the behavioral data. We derive an explicit relationship between the parameters and computations of the exact Bayesian algorithm and those of the approximate linear-exponential filter. Since the latter is equivalent to a leaky-integration process, a commonly used model of neuronal dynamics underlying perceptual decision-making and trial-to-trial dependencies, our model provides a principled account of why such dynamics are useful. We also show that near-optimal tuning of the leaky-integration process is possible, using stochastic gradient descent based only on the noisy binary inputs. This is a proof of concept that not only can neurons implement near-optimal prediction based on standard neuronal dynamics, but that they can also learn to tune the processing parameters without explicitly representing probabilities.

Spotlight presentation, Tuesday, 5:20pm.

T77 *A rational model of preference learning and choice prediction by children*

Chris Lucas clucas@berkeley.edu
Thomas L Griffiths tom_griffiths@berkeley.edu
University of California, Berkeley, CA, United States
Fei Xu fei@psych.ubc.ca
University of British Columbia, Vancouver, B.C., Canada
Christine Fawcett fawcett@berkeley.edu
University of California, Berkeley, CA, United States

Young children demonstrate the ability to make inferences about the preferences of other agents based on their choices. However, there exists no overarching account of what children are doing when they learn about preferences or how they use that knowledge. We use a rational model of preference learning, drawing on ideas from economics and computer science, to explain the behavior of children in several recent experiments. Specifically, we show how a simple econometric model can be extended to capture two- to four-year-olds' use of statistical information in inferring preferences, and their generalization of these preferences.

Spotlight presentation, Tuesday, 9:50am.

T78 *Learning to Use Working Memory in Partially Observable Environments through Dopaminergic Reinforcement*

Michael T Todd, Yael Niv and Jonathan D. Cohen, Princeton University.
Oral presentation, Tuesday, 10:50am. *See abstract, page 99.*

T79 *Dependence of Orientation Tuning on Recurrent Excitation and Inhibition in a Network Model of V1*

Klaus Wimmer klaus@cs.tu-berlin.de
Marcel Stimberg mst@cs.tu-berlin.de
Robert Martin rm@cs.tu-berlin.de
Bernstein Center for Computational Neuroscience, Berlin, and Dept. of Electrical Engineering and Computer Science, Technische Universität Berlin, Germany
Lars Schwabe lars.schwabe@uni-rostock.de
Dept. of Computer Science and Electrical Engineering, University of Rostock, Germany
Jorge Mariño xurxo@cdf.udc.es
Dept. of Medicine, Neuroscience, and Motor Control Group, University A Coruña, Spain
James Schummers schummej@mit.edu
Dept. of Brain and Cognitive Sciences, and Picower Center for Learning and Memory, MIT, Cambridge, MA, United States
David C. Lyon dclyon@uci.edu
Dept. of Anatomy and Neurobiology, University of California, Irvine, CA, United States
Mriganka Sur msur@mit.edu
Dept. of Brain and Cognitive Sciences, and Picower Center for Learning and Memory, MIT, Cambridge, MA, United States
Klaus Obermayer oby@cs.tu-berlin.de
Bernstein Center for Computational Neuroscience, Berlin, and Dept. of Electrical Engineering and Computer Science, Technische Universität Berlin, Germany

One major role of primary visual cortex (V1) in vision is the encoding of the orientation of lines and contours. The role of the local recurrent network in these computations is, however, still a matter of debate. To address this issue, we analyze intracellular recording data of cat V1, which combine measuring the tuning of a range of neuronal properties with a precise localization of the recording sites in the orientation preference map. For the analysis, we consider a network model of Hodgkin-Huxley type neurons arranged according to a biologically plausible two-dimensional topographic orientation preference map. We then systematically vary the strength of the recurrent excitation and inhibition relative to the strength of the afferent input. Each parametrization gives rise to a different model instance for which the tuning of model neurons at different locations of the orientation map is compared to the experimentally measured orientation tuning of membrane potential, spike output, excitatory, and inhibitory conductances. A quantitative analysis shows that the data provides strong evidence for a network model in which the afferent input is dominated by strong, balanced contributions of recurrent excitation and inhibition. This recurrent regime is close to a regime of "instability", where strong, self-sustained activity of the network occurs. The firing rate of neurons in the best-fitting network is particularly sensitive to small modulations of model parameters, which could be one of the functional benefits of a network operating in this particular regime.

T80 *Psychiatry: Insights into depression through normative decision-making models*

Quentin JM Huys qhuys@gatsby.ucl.ac.uk
Gatsby Computational Neuroscience Unit and Medical School, University College London, London, United Kingdom, and Center for Theoretical Neuroscience, Columbia University, New York, NY, United States
Joshua Vogelstein joshuav@jhu.edu

Johns Hopkins, Baltimore, MD, United States
Peter Dayan dayan@gatsby.ucl.ac.uk
Gatsby Computational Neuroscience Unit, UCL, London, United Kingdom

Decision making lies at the very heart of many psychiatric diseases. It is also a central theoretical concern in a wide variety of fields and has undergone detailed, in-depth, analyses. We take as an example Major Depressive Disorder (MDD), applying insights from a Bayesian reinforcement learning framework. We focus on anhedonia and helplessness. Helplessness—a core element in the conceptualizations of MDD that has lead to major advances in its treatment, pharmacological and neurobiological understanding—is formalized as a simple prior over the outcome entropy of actions in uncertain environments. Anhedonia, which is an equally fundamental aspect of the disease, is related to the effective reward size. These formulations allow for the design of specific tasks to measure anhedonia and helplessness behaviorally. We show that these behavioral measures capture explicit, questionnaire-based cognitions. We also provide evidence that these tasks may allow classification of subjects into healthy and MDD groups based purely on a behavioral measure and avoiding any verbal reports.

Demonstrations (7:30pm–12:00am)

1a *A Pencil Balancing Robot Using Only Spike-based Visual Input*

Jorg Conradt conradt@ini.ethz.ch
Patrick Lichtsteiner patrick@ini.ethz.ch
Raphael Berner raphael@ini.ethz.ch
Tobi Delbruck tobi@ini.ethz.ch
Rodney J. Douglas rjd@ini.ethz.ch
Matthew Cook cook@ini.ethz.ch
Institute of Neuroinformatics, ETH, Zurich, Switzerland

http://www.ini.unizh.ch/~conradt/projects/PencilBalancer

Animals by far outperform current technology when reacting to visual stimuli in low processing requirements, demonstrating astonishingly fast reaction times to changes. Current real-time vision based robotic control approaches, in contrast, typically require high computational resources to extract relevant information from sequences of images provided by a video camera. Most of the information contained in consecutive images is redundant, which often turns the vision processing algorithms into a limiting factor in high-speed robot control. As an example, robotic pole balancing with large objects is a well known exercise in current robotics research, but balancing arbitrary small poles (such as a pencil, which is too small for a human to balance) has not yet been achieved due to limitations in vision processing. At the Institute of Neuroinformatics we developed an analog silicon retina (http://siliconretina.ini.uzh.ch), which, in contrast to current video cameras, only reports individual events ("spikes") from individual pixels when the illumination changes within the pixel's field of view. Transmitting only the "on" and "off" spike events, instead of transmitting full vision frames, drastically reduces the amount of data processing required to react to environmental changes. This information encoding is directly inspired by the spike based information transfer from the human eye to visual cortex. In our demonstration, we address the challanging problem of balancing an arbitrary standard

pencil, based solely on visual information. A stereo pair of silicon retinas reports vision events caused by the moving pencil, which is standing on its tip on an actuated table. Then our processing algorithm extracts the pencil position and angle without ever using a "full scene" visual representation, but simply by processing only the spikes relevant to the pencil's motion. Our system uses neurally inspired hardware and a neurally inspired form of communication to achieve a difficult goal. Thus, it is truly a Neural Information Processing System.

2a *High-Accuracy 3D Sensing for Mobile Manipulators*

Stephen Gould sgould@cs.stanford.edu
Morgan Quigley mquigley@cs.stanford.edu
Siddarth Batra sidbatra@cs.stanford.edu
Ellen Klingbiel ellenrk@cs.stanford.edu
Quoc Le quocle@cs.stanford.edu
Andrew Y. Ng ang@cs.stanford.edu
Stanford University, Stanford, CA, United States

Our demonstration shows the advantages of high-accuracy 3d scanning to the tasks of object detection and door opening. Specifically we show how to obtain depth data from a 3d line scanner directly in the image frame of reference. This allows us to associate a "depth channel" with each image pixel, i.e. each pixel is represented as a four dimensional vector comprising red, green, blue and depth. We then show how, using this representation, we can get near perfect door opening and object detection performance for a selection of common household/office objects. This level of accuracy on a mobile robot in uncontrolled environments have never been achieved before.

3a *Vision Toolkit Based on Hierarchical Temporal Memory*

Dileep George dgeorge@numenta.com
Jeff Edwards jedwards@numenta.com
Jamie Niemasik jniemasik@numenta.com
Ron Marianetti rmarianetti@numenta.com
Bill Saphir bsaphir@numenta.com
Subutai Ahmad sahmad@numenta.com
Numenta, Inc., Menlo Park, CA, United States

http://www.numenta.com

We demonstrate a Vision Toolkit based on Hierarchical Temporal Memory (HTM). Hierarchical Temporal Memory is a technology being developed by Numenta based on a theory of operation of the neocortex. HTMs model data by learning spatial co-occurrences and temporal sequences at multiple levels in a hierarchy. More details on the technology can be obtained from http://www.numenta.com/for-developers/education/general-overview-htm.php. The vision toolkit has capabilities for object recognition, top-down attention and shape similarity search. The vision toolkit demonstrates robust invariant recognition of gray-scale objects in the presence of noise and large amounts of clutter. The HTM network behind this learns its invariant representation at every level through exposure to video sequences of moving objects. Inference in the network is done using Bayesian belief propagation. We demonstrate a network that handles 25 categories of images with large intra category variations. In addition to object recognition, the toolkit can also do visual similarity search. Traditionally, multiple objects in a scene are recognized using scan windows of different sizes. The HTM vision network can recognize and localize multiple objects in a scene using a top-down attention mechanism. The feed-forward pass on a scene gives a set of hypotheses about the objects present in the scene. The network localizes objects through the feedback propagation of the top hypotheses. One important aspect of the vision toolkit is the ease of use. The vision toolkit is a self-contained system that enables a user to create a recognition network with very little programming. The vision toolkit is built on top of NuPIC - Numenta Platform for intelligent computing. NuPIC allows for rapid prototyping using hierarchical networks. It uses python as the scripting language with math libraries written in C++ for faster execution. Parallelization capabilities are built in. We also demonstrate the generic capabilities of NuPIC in learning higher-order spatio-temporal models of data. We demonstrate unsupervised learning of action models using motion-capture data and hierarchical sequence prediction using web-navigation data.

4a *Play Brain-Pong in 10 Minutes*

Cristian Grozea cristian.grozea@first.fraunhofer.de
Siamac Fazli siamac.fazli@first.fraunhofer.de
Márton Danóczy marton78@gmail.com
Florin Popescu florin.popescu@first.fraunhofer.de
Fraunhofer FIRST, IDA Group, Berlin, Germany

We propose to demonstrate a ready to use BCI system employing our patented dry electrodes cap with 6 electrodes [Popescu F, Fazli S, Badower Y, Blankertz B, Müller K *Single Trial Classification of Motor Imagination Using 6 Dry EEG Electrodes.* PLoS ONE 2(7)]. People from the public will be allowed to try it and be able to perform feedback session after a short setup period of 10 minutes.

5a *Worio: A Web-Scale Machine Learning System*

Nando de Freitas nando@cs.ubc.ca
Department of Computer Science, University of British Columbia, Vancouver, BC, Canada
Ali Davar ali@worio.com
Worio, Vancouver, BC, Canada

http://www.worio.com

Our team has created a unique technology that demonstrates how machine learning ideas can be applied successfully at web-scale. Worio has developed and integrated: (i) novel classifiers that automatically apply millions of tags to billions of documents accurately and within a day, (ii) a massive web-scale recommender system for millions of users and billions of items, and (iii) topic modeling of users, documents and queries. The technology opens up room for new products in knowledge discovery and information retrieval, filtering and organization. It is also a platform for developing more sophisticated-scale natural language applications.

6a *MLPACK: Scalable Machine Learning Software*

Alexander Gray agray@cc.gatech.edu
Georgia Institute of Technology, College of Computing, Atlanta, GA, United States

We will unveil MLPACK, a major open source machine learning software project intended to serve machine learning in the same way that LAPACK serves linear algebra. MLPACK is a software collection which will soon cover most or all of the major methods of machine learning. It is built using FASTlib, a well-designed C++ library with special attention to large datasets and the data structures which support them, memory and CPU efficiency, and seamless integration on top of LAPACK. We believe MLPACK fills a major need in machine learning. The initial release contains state-of-the-art fast algorithms for a moderate number of common machine learning methods.

Wednesday, December 10th

Oral Session (8:30am–9:50am): "Neural Coding"

8:30am Invited Talk: *Connectome: the quest to deconstruct the brain*

Sebastian Seung seung@mit.edu
Massachusetts Institute of Technology, Cambridge, MA, United States

Recent innovations in 3d nanoscale imaging are expected to produce teravoxel and petavoxel-sized images of the brain's neural networks. These datasets will only become useful for neuroscience if computer scientists can develop algorithms for automated image analysis. Chief among the challenges is accurate tracing of the "wires" of the brain, its axons and dendrites, through the 3d images. Achieving the necessary accuracy will require the use of machine learning, rather than hand-designed algorithms. If the tracing problem is solved, it will become possible to create automated systems that take a sample of brain tissue as input and generate its "wiring diagram" or "connectome". Such systems would revolutionize neuroscience by giving rise to a new field called "connectomics," defined by the high- throughput generation of data about neural connectivity, and the subsequent mining of that data for knowledge about the brain. I will discuss the impact that connectomics could have on our understanding of how the brain wires and rewires itself, the dynamics of activity in neural networks, and the neuropathological basis of mental disorders.

Sebastian Seung is Professor of Computational Neuroscience in the Department of Brain and Cognitive Sciences and the Department of Physics at the Massachusetts Institute of Technology, and Investigator of the Howard Hughes Medical Institute. He studied theoretical physics with David Nelson at Harvard University and completed postdoctoral training with Haim Sompolinsky at the Hebrew University of Jerusalem. Before joining the MIT faculty, he was a member of the Theoretical Physics Department at Bell Laboratories. He has been a Packard Fellow, Sloan Fellow, and McKnight Scholar.

9:30am *Modeling Short-term Noise Dependence of Spike Counts in Macaque Prefrontal Cortex*

Arno Onken aonken@cs.tu-berlin.de
Steffen Grünewälder gruenew@cs.tu-berlin.de
Technische Universität Berlin, Berlin, Germany
Matthias Munk matthias.munk@tuebingen.mpg.de
Max Planck Institute for Biological Cybernetics, Tübingen, Germany
Klaus Obermayer oby@cs.tu-berlin.de
Technische Universität Berlin, Berlin, Germany

Correlations between spike counts are often used to analyze neural coding. The noise is typically assumed to be Gaussian. Yet, this assumption is often inappropriate, especially for low spike counts. In this study, we present copulas as an alternative approach. With copulas it is possible to use arbitrary marginal distributions such as Poisson or negative binomial that are better suited for modeling noise distributions of spike counts. Furthermore, copulas place a wide range of dependence structures at the disposal and can be used to analyze higher order interactions. We develop a framework to analyze spike count data by means of copulas. Methods for parameter inference based on maximum likelihood estimates and for computation of Shannon entropy are provided. We apply the method to our data recorded from macaque prefrontal cortex. The data analysis leads to three significant findings: (1) copula-based distributions provide better fits than discretized multivariate normal distributions; (2) negative binomial margins fit the data better than Poisson margins; and (3) a dependence model that includes only pairwise interactions overestimates the information entropy by at least 19% compared to the model with higher order interactions.

Spotlights (9:50am–10:00am)

- ***Characterizing neural dependencies with Poisson copula models***
  Pietro Berkes, Brandeis University, Frank Wood and Jonathan Pillow, Gatsby Computational Neuroscience Unit, UCL.
  See abstract, page 157.

- ***A general framework for investigating how far the decoding process in the brain can be simplified***
  Masafumi Oizumi, The University of Tokyo, Toshiyuki Ishii, RIKEN Brain Science Institute, Kazuya Ishibashi, The University of Tokyo, Toshihiko Hosoya, RIKEN Brain Science Institute, and Masato Okada, The University of Tokyo.
  See abstract, page 156.

- ***Estimating the Location and Orientation of Complex, Correlated Neural Activity using MEG***
  David Wipf, Julia Owen, University of California, San Francisco, Hagai T Attias, Kensuke Sekihara, Golden Metallic Inc, and Srikantan S Nagarajan, University of California, San Francisco.
  See abstract, page 159.

- ***Nonparametric sparse hierarchical models describe V1 fMRI responses to natural images***
  Pradeep Ravikumar, Vincent Q Vu, Bin Yu, Thomas Naselaris, Kendrick Kay and Jack Gallant, University of California, Berkeley.
  See abstract, page 186.

- ***Artificial Olfactory Brain for Mixture Identification***
  Mehmet K Muezzinoglu, Alexander Vergara, Ramon Huerta, University of California San Diego, Thomas Nowotny, University of Sussex, Nikolai Rulkov, Henry Abarbanel, Allen Selverston and Mikhail Rabinovich, University of California San Diego.
  See abstract, page 159.

- ***Spike Feature Extraction Using Informative Samples***
  Zhi Yang, Qi Zhao and Wentai Liu, University of California at Santa Cruz.
  See abstract, page 156.

- ***The Conjoint Effect of Divisive Normalization and Orientation Selectivity on Redundancy Reduction***
  Fabian H. Sinz and Matthias Bethge, MPI for Biological Cybernetics.
  See abstract, page 158.

- ***An improved estimator of Variance Explained in the presence of noise***
  Ralf M Haefner, National Eye Institute, NIH, and Bruce G Cumming, Laboratory of Sensorimotor Research, NEI.
  See abstract, page 157.

- ***Interpreting the neural code with Formal Concept Analysis***
  Dominik Endres and Peter Foldiak, University of St Andrews.
  See abstract, page 158.

- ***Dynamic visual attention: searching for coding length increments***
  Xiaodi Hou, California Institute of Technology, and Liqing Zhang, Shanghai Jiao Tong University.
  See abstract, page 185.

Oral Session (10:30am–11:50am):
"Complex Dynamical Systems: Modeling and Estimation"

10:30am ***Modeling the effects of memory on human online sentence processing with particle filters***

Roger P. Levy rlevy@ling.ucsd.edu
University of California, San Diego, La Jolla, CA, United States
Florencia Reali floreali@berkeley.edu
Thomas L Griffiths tom_griffiths@berkeley.edu
University of California, Berkeley, CA, United States

Language comprehension in humans is significantly constrained by memory, yet rapid, highly incremental, and capable of utilizing a wide range of contextual information to resolve ambiguity and form expectations about future input. In contrast, most of the leading psycholinguistic models and fielded algorithms for natural language parsing are non-incremental, have run time superlinear in input length, and/or enforce structural locality constraints on probabilistic dependencies between events. We present a new limited-memory model of sentence comprehension which involves an adaptation of the particle filter, a sequential Monte Carlo method, to the problem of incremental parsing. We show that this model can reproduce classic results in online sentence comprehension, and that it naturally provides the first rational account of an outstanding problem in psycholinguistics, in which the preferred alternative in a syntactic ambiguity seems to grow more attractive over time even in the absence of strong disambiguating information.

10:50am ***On Computational Power and the Order-Chaos Phase Transition in Reservoir Computing***

Benjamin Schrauwen bschrauw@elis.ugent.be
Ghent University, Ghent, Belgium
Lars Buesing lars@igi.tu-graz.ac.at
Robert Legenstein robert.legenstein@igi.tugraz.at
TU Graz, Graz, Austria

Randomly connected recurrent neural circuits have proven to be very powerful models for online computations when a trained memoryless readout function is appended. Such Reservoir Computing (RC) systems are commonly used in two flavors: with analog or binary (spiking) neurons in the recurrent circuits. Previous work showed a fundamental difference between these two incarnations of the RC idea. The performance of a RC system built from binary neurons seems to depend strongly on the network connectivity structure. In networks of analog neurons such dependency has not been observed. In this article we investigate this apparent dichotomy in terms of the in-degree of the circuit nodes. Our analyses based amongst others on the Lyapunov exponent reveal that the phase transition between ordered and chaotic network behavior of binary circuits qualitatively differs from the one in analog circuits. This explains the observed decreased computational performance of binary circuits of high node in-degree. Furthermore, a novel mean-field predictor for computational performance is introduced and shown to accurately predict the numerically obtained results.

11:10am *Nonrigid Structure from Motion in Trajectory Space*

Ijaz Akhter akhter@lums.edu.pk
LUMS School of Science and Engineering, Lahore, Pakistan
Yaser Sheikh yaser@cs.cmu.edu
Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, United States
Sohaib Khan sohaib@lums.edu.pk
LUMS School of Science and Engineering, Lahore, Pakistan
Takeo Kanade tk@cs.cmu.edu
Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, United States

Existing approaches to nonrigid structure from motion assume that the instantaneous 3D shape of a deforming object is a linear combination of basis shapes, which have to be estimated anew for each video sequence. In contrast, we propose that the evolving 3D structure be described by a linear combination of basis trajectories. The principal advantage of this approach is that we do not need to estimate any basis vectors during computation. Instead, we show that generic bases over trajectories, such as the Discrete Cosine Transform (DCT) basis, can be used to compactly describe most real motions. This results in a significant reduction in unknowns, and corresponding stability in estimation. We report empirical performance, quantitatively using motion capture data and qualitatively on several video sequences exhibiting nonrigid motions including piece-wise rigid motion, articulated motion, partially nonrigid motion (such as a facial expression), and highly nonrigid motion (such as a person dancing).

11:30am *Model selection and velocity estimation using novel priors for motion patterns*

Shuang Wu shuangw@stat.ucla.edu
Department of Statistics, University of California, Los Angeles, CA, United States
Hongjing Lu hongjing@ucla.edu
Department of Psychology and Department of Statistics, University of California, Los Angeles, CA, United States
Alan Yuille yuille@stat.ucla.edu
Department of Statistics, University of California, Los Angeles, CA, United States

Psychophysical experiments show that humans are better at perceiving rotation and expansion than translation. These findings are inconsistent with standard models of motion integration which predict best performance for translation [6]. To explain this discrepancy, our theory formulates motion perception at two levels of inference: we first perform model selection between the competing models (e.g. translation, rotation, and expansion) and then estimate the velocity using the selected model. We define novel prior models for smooth rotation and expansion using techniques similar to those in the slow-and-smooth model [17] (e.g. Green functions of differential operators). The theory gives good agreement with the trends observed in human experiments.

Spotlights (11:50am–12:00pm)

- ***Offline Handwriting Recognition with Multidimensional Recurrent Neural Networks***
  ALEX GRAVES, TU Munich, and JUERGEN SCHMIDHUBER, IDSIA and TU Munich.
  See abstract, page 186.

- ***Recursive Segmentation and Recognition Templates for 2D Parsing***
  LEO ZHU, CSAIL MIT, YUANHAO CHEN, USTC, YUAN LIN, SJTU, CHENXI LIN, Microsoft Research Asia, and ALAN YUILLE, UCLA.
  See abstract, page 182.

- ***Learning a discriminative hidden part model for human action recognition***
  YANG WANG and GREG MORI, Simon Fraser University.
  See abstract, page 183.

- ***Learning Transformational Invariants from Natural Movies***
  CHARLES CADIEU and BRUNO OLSHAUSEN, University of California, Berkeley.
  See abstract, page 184.

- ***Nonparametric Bayesian Learning of Switching Linear Dynamical Systems***
  EMILY FOX, Massachusetts Institute of Technology, ERIK SUDDERTH, MICHAEL JORDAN, University of California, Berkeley, and ALAN WILLSKY, Massachusetts Institute of Technology.
  See abstract, page 179.

- ***The Infinite Factorial Hidden Markov Model***
  JURGEN VAN GAEL, University of Cambridge, YEE WHYE TEH, Gatsby Computational Neuroscience Unit, UCL, and ZOUBIN GHAHRAMANI, CMU & University of Cambridge.
  See abstract, page 181.

- ***Mixed Membership Stochastic Blockmodels***
  EDOARDO M. AIROLDI, DAVID M. BLEI, Princeton University, STEPHEN E. FIENBERG and ERIC P. XING, Carnegie Mellon University.
  See abstract, page 176.

- ***Supervised Bipartite Graph Inference***
  YOSHIHIRO YAMANISHI, Mines ParisTech.
  See abstract, page 164.

- ***Stochastic Relational Models for Large-scale Dyadic Data using MCMC***
  SHENGHUO ZHU, KAI YU and YIHONG GONG, NEC Laboratories America, Inc..
  See abstract, page 177.

- ***Using matrices to model symbolic relationship***
  ILYA SUTSKEVER and GEOFFREY E HINTON, University of Toronto.
  See abstract, page 181.

Oral Session (2:00pm–3:20pm): "Physics and High Order Statistics"

2:00pm Invited Talk: *Machine Learning in High Energy Physics*

HARRISON PROSPER harry@hep.fsu.edu
Department of Physics, Florida State University, Tallahassee, FL, United States

I begin with a brief discussion of the nature of high energy physics, and follow with a review of a few real-world examples of the application of machine learning methods in this field. I focus on the common, but difficult task, of extracting small signals masked by enormous backgrounds. The talk ends with a discussion of the computational challenges we expect to face in the very near future at the Large Hadron Collider and an enumeration of what my colleagues and I see as open questions.

Harrison Prosper did his doctorate in particle physics from the University of Manchester, England, in 1980 and, from 1982-1986, was a post-doctoral fellow at the Rutherford Appleton Laboratory, but stationed at the Institut Laue Langevin, Grenoble, France. In 1988, after a brief stint at Virginia Tech, Blacksburg, he joined the Fermi National Accelerator Laboratory as an Associate Scientist. In 1993,he joined the faculty at Florida State University, became a full professor in 1998, was elected a fellow of the American Physical Society in 2002, and became the Kirby W. Kemper professor of physics in 2006. A principal interest of his is the application of machine learning and Bayesian methods to particle physics research.

3:00pm *Characteristic Kernels on Groups and Semigroups*

KENJI FUKUMIZU fukumizu@ism.ac.jp
Institute of Statistical Mathematics, Tokyo, Japan
BHARATH K SRIPERUMBUDUR bharath.k.sriperumbudur@tuebingen.mpg.de
University of California, San Diego, CA, United States, and Max Planck Institute for Biological Cybernetics, Tübingen, Germany
ARTHUR GRETTON arthur.gretton@tuebingen.mpg.de
BERNHARD SCHÖLKOPF bs@tuebingen.mpg.de
Max Planck Institute for Biological Cybernetics, Tübingen, Germany

Embeddings of random variables in reproducing kernel Hilbert spaces (RKHSs) may be used to conduct statistical inference based on higher order moments. For sufficiently rich (characteristic) RKHSs, each probability distribution has a unique embedding, allowing all statistical properties of the distribution to be taken into consideration. Necessary and sufficient conditions for an RKHS to be characteristic exist for $\mathbb{R}^n$. In the present work, conditions are established for an RKHS to be characteristic on groups and semigroups. Illustrative examples are provided, including characteristic kernels on periodic domains, rotation matrices, and $\mathbb{R}^n_+$.

Spotlights (3:20pm–3:30pm)

- ***Kernel Measures of Independence for non-iid Data***
  XINHUA ZHANG, NICTA and Australian National University, LE SONG, Carnegie Mellon University, ARTHUR GRETTON, Max Planck Institute for Biological Cybernetics, and ALEX J. SMOLA, Yahoo! Research.
  See abstract, page 170.

- ***Multi-label Multiple Kernel Learning***
  SHUIWANG JI, LIANG SUN, Arizona State University, RONG JIN, Michigan State University, and JIEPING YE, Arizona State University.
  See abstract, page 163.

- ***Scalable Algorithms for String Kernels with Inexact Matching***
  PAVEL P KUKSA, PAI-HSI HUANG and VLADIMIR PAVLOVIC, Rutgers University.
  See abstract, page 164.

- ***Kernelized Sorting***
  NOVI QUADRIANTO, Australian National University, LE SONG, CMU, and ALEX J. SMOLA, Yahoo! Research.
  See abstract, page 174.

- ***Learning with Consistency between Inductive Functions and Kernels***
  HAIXUAN YANG, IRWIN KING and MICHAEL LYU, The Chinese University of Hong Kong.
  See abstract, page 171.

- ***Performance analysis for L_2 kernel classification***
  JOOSEUK KIM and CLAYTON SCOTT, University of Michigan.
  See abstract, page 169.

- ***Clustered Multi-Task Learning: A Convex Formulation***
  LAURENT JACOB, Mines ParisTech- Institut Curie - INSERM, FRANCIS BACH, INRIA - Ecole Normale Superieure, and JEAN-PHILIPPE VERT, Mines ParisTech- Institut Curie - INSERM.
  See abstract, page 168.

- ***Domain Adaptation with Multiple Sources***
  YISHAY MANSOUR, School of Computer Science, MEHRYAR MOHRI, Google Research, and AFSHIN ROSTAMIZADEH, Courant Institute, New York University.
  See abstract, page 167.

- ***Phase transitions for high-dimensional joint support recovery***
  SAHAND NEGAHBAN and MARTIN WAINWRIGHT, University of California, Berkeley.
  See abstract, page 167.

- ***Translated Learning: Transfer Learning across Different Feature Spaces***
  WENYUAN DAI, YUQIANG CHEN, GUI-RONG XUE, Shanghai Jiao Tong University, QIANG YANG, Hong Kong University of Science and Technology, and YONG YU, Shanghai Jiao Tong University.
  See abstract, page 171.

Oral Session (4:00pm–5:20pm): "Sparse Learning and Ranking"

4:00pm *Adaptive Forward-Backward Greedy Algorithm for Sparse Learning with Linear Models*

TONG ZHANG tzhang@stat.rutgers.edu
Rutgers University, Piscataway, NJ, United States

Consider linear prediction models where the target function is a sparse linear combination of a set of basis functions. We are interested in the problem of identifying those basis functions with non-zero coefficients and reconstructing the target function from noisy observations. Two heuristics that are widely used in practice are forward and backward greedy algorithms. First, we show that neither idea is adequate. Second, we propose a novel combination that is based on the forward greedy algorithm but takes backward steps adaptively whenever beneficial. We prove strong theoretical results showing that this procedure is effective in learning sparse representations. Experimental results support our theory.

4:20pm *An interior-point stochastic approximation method and an L1-regularized delta rule*

PETER CARBONETTO pcarbo@cs.ubc.ca
MARK SCHMIDT schmidtm@cs.ubc.ca
NANDO DE FREITAS nando@cs.ubc.ca
University of British Columbia, Vancouver, BC, Canada

The stochastic approximation method is behind the solution to many important, actively-studied problems in machine learning. Despite its far-reaching application, there is almost no work on applying stochastic approximation to learning problems with constraints. The reason for this, we hypothesize, is that no robust, widely-applicable stochastic approximation method exists for handling such problems. We propose that interior-point methods are a natural solution. We establish the stability of a stochastic interior-point approximation method both analytically and empirically, and demonstrate its utility by deriving an on-line learning algorithm that also performs feature selection via L1 regularization.

4:40pm *Global Ranking Using Continuous Conditional Random Fields*

TAO QIN taoqin@microsoft.com
TIE-YAN LIU tyliu@microsoft.com
Microsoft Research Asia, Beijing, China
XU-DONG ZHANG zhangxd@mail.tsinghua.edu.cn
DE-SHENG WANG wangdsh_ee@mail.tsinghua.edu.cn
Department of Electronic Engineering, Tsinghua University, Beijing, China
HANG LI hangli@microsoft.com
Microsoft Research Asia, Beijing, China

This paper studies global ranking problem by learning to rank methods. Conventional learning to rank methods are usually designed for 'local ranking', in the sense that the ranking model is defined on a single object, for example, a document in information retrieval. For many applications, this is a very loose approximation. Relations always exist between objects and it is better to define the ranking model as a function on all the objects to be ranked (i.e., the relations are also included). This paper refers to the problem as global ranking and proposes employing a Continuous Conditional Random Fields (CRF) for conducting the learning task. The Continuous CRF model is defined as a conditional probability distribution over ranking scores of objects conditioned on the objects. It can naturally represent the content information of objects as well as the relation information between objects, necessary for global ranking. Taking two specific information retrieval tasks as examples, the paper shows how the Continuous CRF method can perform global ranking better than baselines.

5:00pm *Inferring rankings under constrained sensing*

Srikanth Jagabathula `jskanth@mit.edu`
Devavrat Shah `devavrat@mit.edu`
Massachusetts Institute of Technology, Cambridge, MA, United States.

Motivated by applications such as elections, web-page ranking, revenue maximization in gambling, etc. we consider the question of inferring popular rankings. In many of these applications, the number of distinct popular rankings are usually very few. However, the data available for inferring them is highly constrained. As the first main result of this paper, we provide a simple and novel algorithm to recover the popular rankings and their popularity exactly under a natural set of assumptions. Specifically, under a natural stochastic model, our algorithm recovers them exactly as long as the number of distinct popular rankings up to $O(n)$ over n candidates (or elements). In certain applications, like ranked election, the interest is only in recovering the most popular (or mode) ranking. As the second result, we provide an algorithm based on Fourier transform over symmetric group to recover the most popular ranking under natural majority condition. Interestingly enough, this algorithm becomes a maximum weight matching on an appropriately defined weighted bipartite graph.

The questions considered in this paper are thematically related to the currently popular topic of compressed sensing. The existing results in the literature do not apply to our problem as the information is constrained and can not be sampled randomly. Thus, the results of this paper correspond to constrained compressed sensing.

Spotlights (5:20pm–5:30pm)

- ***Nonparametric regression and classification with joint sparsity constraints***
  Han Liu, John Lafferty and Larry Wasserman, Carnegie Mellon University.
  See abstract, page 165.

- ***Robust Regression and Lasso***
  Huan Xu, McGill University, Constantine Caramanis, The University of Texas at Austin, and Shie Mannor, McGill University.
  See abstract, page 166.

- ***High-dimensional union support recovery in multivariate regression***
  Guillaume Obozinski, Martin Wainwright and
  Michael Jordan, University of California, Berkeley.
  See abstract, page 165.

- ***Sparse Signal Recovery Using Markov Random Fields***
  Volkan Cevher, Marco F Duarte, Chinmay
  Hegde and Richard Baraniuk, Rice University.
  See abstract, page 184.

- ***Sparse probabilistic projections***
  Cédric Archambeau, University College London, and
  Francis Bach, INRIA - Ecole Normale Superieure.
  See abstract, page 174.

- ***Deflation Methods for Sparse PCA***
  Lester Mackey, University of California, Berkeley.
  See abstract, page 173.

- ***Robust Kernel Principal Component Analysis***
  Minh Hoai Nguyen and Fernando De la Torre, Carnegie Mellon University.
  See abstract, page 172.

- ***Bayesian Exponential Family PCA***
  Shakir Mohamed, Katherine Heller, University of Cambridge,
  and Zoubin Ghahramani, University of Cambridge & CMU.
  See abstract, page 173.

- ***ICA based on a Smooth Estimation of the Differential Entropy***
  Lev Faivishevsky and Jacob Goldberger, Bar Ilan University.
  See abstract, page 172.

- ***Localized Sliced Inverse Regression***
  Qiang Wu, ISDS, Duke University, Sayan Mukherjee, Duke
  University, and Feng Liang, University of Illinois at Urbana-Champaign.
  See abstract, page 169.

Poster Session (7:30pm–12:00am)

W1 *Spike Feature Extraction Using Informative Samples*

Zhi Yang yangzhi@soe.ucsc.edu
Qi Zhao zhaoqi@soe.ucsc.edu
Wentai Liu wentai@soe.ucsc.edu
University of California, Santa Cruz, CA, United States

A new spike feature extraction algorithm which targets real-time spike sorting and facilitates miniaturized microchip implementation is presented. The proposed algorithm has been evaluated on synthesized waveforms and experimentally recorded sequences. When compared with many spike sorting approaches our algorithm demonstrates significantly improved speed, accuracy and allows unsupervised execution. A preliminary hardware implementation has been realized using an integrated microchip interfaced with a personal computer.

Spotlight presentation, Wednesday, 9:50am.

W2 *Modeling Short-term Noise Dependence of Spike Counts in Macaque Prefrontal Cortex*

Arno Onken, Steffen Grünewälder, Technische Universität Berlin, Matthias Munk, Max Planck Institute for Biological Cybernetics, and Klaus Obermayer, Technische Universität Berlin.
Oral presentation, Wednesday, 9:30am. *See abstract, page 145.*

W3 *On Computational Power and the Order-Chaos Phase Transition in Reservoir Computing*

Benjamin Schrauwen, Ghent University, Lars Buesing and Robert Legenstein, TU Graz.
Oral presentation, Wednesday, 10:50am. *See abstract, page 148.*

W4 *A general framework for investigating how far the decoding process in the brain can be simplified*

Masafumi Oizumi oizumi@mns.k.u-tokyo.ac.jp
The University of Tokyo, Chiba, Japan
Toshiyuki Ishii tishii@brain.riken.jp
RIKEN Brain Science Institute, Wako, Saitama, Japan
Kazuya Ishibashi kazuya@mns.k.u-tokyo.ac.jp
The University of Tokyo, Chiba, Japan
Toshihiko Hosoya hosoya@brain.riken.jp
RIKEN Brain Science Institute, Wako, Saitama, Japan
Masato Okada okada@k.u-tokyo.ac.jp
The University of Tokyo, Chiba, Japan

"How is information decoded in the brain?" is one of the most difficult and important questions in neuroscience. Whether neural correlation is important or not in decoding neural activities is of special interest. We have developed a general framework for investigating how far the decoding process in the brain can be simplified. First, we hierarchically construct simplified probabilistic models of neural responses that ignore more than Kth-order correlations by using a maximum entropy principle. Then, we compute how much information is lost when information is decoded using the simplified models, i.e., "mismatched decoders". We introduce an information theoretically correct quantity for evaluating the information obtained by mismatched decoders. We applied our proposed framework to spike data for vertebrate retina. We used 100-ms natural movies as stimuli and computed the information contained in neural activities about these movies. We found that the information loss is negligibly small in population activities of ganglion cells even if all orders of correlation are ignored in decoding. We also found that if we assume stationarity for long durations in the information analysis of dynamically changing stimuli like natural movies, pseudo correlations seem to carry a large portion of the information.

Spotlight presentation, Wednesday, 9:50am.

W5 *Characterizing neural dependencies with Poisson copula models*

PIETRO BERKES berkes@brandeis.edu
Brandeis University, Waltham, MA, United States
FRANK WOOD fwood@gatsby.ucl.ac.uk
JONATHAN PILLOW pillow@gatsby.ucl.ac.uk
Gatsby Computational Neuroscience Unit, UCL, London, United Kingdom

The coding of information by neural populations depends critically on the statistical dependencies between neuronal responses. However, there is no simple model that combines the observations that (1) marginal distributions over single-neuron spike counts are often approximately Poisson; and (2) joint distributions over the responses of multiple neurons are often strongly dependent. Here, we show that both marginal and joint properties of neural responses can be captured using Poisson copula models. Copulas are joint distributions that allow random variables with arbitrary marginals to be combined while incorporating arbitrary dependencies between them. Different copulas capture different kinds of dependencies, allowing for a richer and more detailed description of dependencies than traditional summary statistics, such as correlation coefficients. We explore a variety of Poisson copula models for joint neural response distributions, and derive an efficient maximum likelihood procedure for estimating them. We apply these models to neuronal data collected in the macaque motor cortex, and quantify the improvement in coding accuracy afforded by incorporating the dependency structure between pairs of neurons.

Spotlight presentation, Wednesday, 9:50am.

W6 *An improved estimator of Variance Explained in the presence of noise*

RALF M HAEFNER haefnerr@nei.nih.gov
Laboratory for Sensorimotor Research, National Eye Institute, NIH, Bethesda, MD, United States
BRUCE G CUMMING bgc@lsr.nei.nih.gov
Laboratory of Sensorimotor Research, NEI, Bethesda, MD, United States

A crucial part of developing mathematical models of how the brain works is the quantification of their success. One of the most widely-used metrics yields the percentage of the variance in the data that is explained by the model. Unfortunately, this metric is biased due to the intrinsic variability in the data. This variability is in principle unexplainable by the model. We derive a simple analytical modification of the traditional formula that significantly improves its accuracy (as measured by bias) with similar or better precision (as measured by mean-square error) in estimating the true underlying Variance Explained by the model class. Our estimator advances on previous work by a) accounting for the uncertainty in the noise estimate, b) accounting for overfitting due to free model parameters mitigating the need for a separate validation data set and c) adding a conditioning term. We apply our new estimator to binocular disparity tuning curves of a set of macaque V1 neurons and find that on a population level almost all of the variance unexplained by Gabor functions is attributable to noise.

Spotlight presentation, Wednesday, 9:50am.

W7 *The Conjoint Effect of Divisive Normalization and Orientation Selectivity on Redundancy Reduction*

Fabian H. Sinz fabee@tuebingen.mpg.de
Matthias Bethge mbethge@tuebingen.mpg.de
MPI for Biological Cybernetics, Tuebingen, Germany

Bandpass filtering, orientation selectivity, and contrast gain control are prominent features of sensory coding at the level of V1 simple cells. While the effect of bandpass filtering and orientation selectivity can be assessed within a linear model, contrast gain control is an inherently nonlinear computation. Here we employ the class of L_p elliptically contoured distributions to investigate the extent to which the two features—orientation selectivity and contrast gain control—are suited to model the statistics of natural images. Within this framework we find that contrast gain control can play a significant role for the removal of redundancies in natural images. Orientation selectivity, in contrast, has only a very limited potential for redundancy reduction.

Spotlight presentation, Wednesday, 9:50am.

W8 *Interpreting the neural code with Formal Concept Analysis*

Dominik Endres dme2@st-andrews.ac.uk
School of Psychology, University of St Andrews, St Andrews, Fife, United Kingdom
Peter Foldiak pf2@st-andrews.ac.uk
University of St Andrews, St Andrews, United Kingdom

We propose a novel application of Formal Concept Analysis (FCA) to neural decoding: instead of just trying to figure out which stimulus was presented, we demonstrate how to explore the semantic relationships between the neural representation of large sets of stimuli. FCA provides a way of displaying and interpreting such relationships via concept lattices. We explore the effects of neural code sparsity on the lattice. We then analyze neurophysiological data from high-level visual cortical area STSa, using an exact Bayesian approach to construct the formal context needed by FCA. Prominent features of the resulting concept lattices are discussed, including indications for a product-of-experts code in real neurons.

Spotlight presentation, Wednesday, 9:50am.

W9 *Artificial Olfactory Brain for Mixture Identification*

Mehmet K Muezzinoglu mmuezzin@ucsd.edu
Alexander Vergara vergara@ucsd.edu
Ramon Huerta rhuerta@ucsd.edu
Institute for Nonlinear Science, University of California San Diego, La Jolla, CA, United States
Thomas Nowotny t.nowotny@sussex.ac.uk
University of Sussex, Brighton, United Kingdom
Nikolai Rulkov nrulkov@ucsd.edu
Henry Abarbanel hdia@ucsd.edu
Allen Selverston aselverston@ucsd.edu
Mikhail Rabinovich mrabinovich@ucsd.edu
Institute for Nonlinear Science, University of California San Diego, La Jolla, CA, United States

The odor transduction process has a large time constant and is susceptible to various types of noise. Therefore, the olfactory code at the sensor/receptor level is in general a slow and highly variable indicator of the input odor in both natural and artificial situations. Insects overcome this problem by using a neuronal device in their Antennal Lobe (AL), which transforms the identity code of olfactory receptors to a spatio-temporal code. This transformation improves the decision of the Mushroom Bodies (MBs), the subsequent classifier, in both speed and accuracy. Here we propose a rate model based on two intrinsic mechanisms in the insect AL, namely integration and inhibition. Then we present a MB classifier model that resembles the sparse and random structure of insect MB. A local Hebbian learning procedure governs the plasticity in the model. These formulations not only help to understand the signal conditioning and classification methods of insect olfactory systems, but also can be leveraged in synthetic problems. Among them, we consider here the discrimination of odor mixtures from pure odors. We show on a set of records from metal-oxide gas sensors that the cascade of these two new models facilitates fast and accurate discrimination of even highly imbalanced mixtures from pure odors.

Spotlight presentation, Wednesday, 9:50am.

W10 *Estimating the Location and Orientation of Complex, Correlated Neural Activity using MEG*

David Wipf david.wipf@mrsc.ucsf.edu
Biomagnetic Imaging Lab, University of California, San Francisco, San Francisco, CA, United States
Julia Owen julia.owen@ucsf.edu
University of California, San Francisco, CA, United States
Hagai T Attias htattias@gmail.com
Kensuke Sekihara ksekiha@cc.tmit.ac.jp
Golden Metallic Inc, San Francisco, CA, United States
Srikantan S Nagarajan sri@radiology.ucsf.edu
University of California, San Francisco, CA, United States

The synchronous brain activity measured via MEG (or EEG) can be interpreted as arising from a collection (possibly large) of current dipoles or sources located throughout the cortex. Estimating the number, location, and orientation of these sources remains a challenging task, one that is significantly compounded by the effects of source correlations and the presence of interference from spontaneous brain activity, sensor noise, and other artifacts. This paper derives an empirical Bayesian method for addressing each of these issues in a principled fashion. The resulting algorithm guarantees descent of a cost function uniquely designed to handle unknown orientations and arbitrary correlations. Robust interference suppression is also easily incorporated. In a restricted setting, the proposed method is shown to have theoretically zero bias estimating both the location and orientation of multi-component dipoles even in the presence of correlations, unlike a variety of existing Bayesian localization methods or common signal processing techniques such as beamforming and sLORETA. Empirical results on both simulated and real data sets verify the efficacy of this approach.

Spotlight presentation, Wednesday, 9:50am.

W11 *A computational model of hippocampal function in trace conditioning*

Elliot A Ludvig elliot@cs.ualberta.ca
University of Alberta, Edmonton, Alberta, AB, Canada
Richard s Sutton sutton@cs.ualberta.ca
Eric Verbeek everbeek@cs.ualberta.ca
James Kehoe jkehoe@psy.unsw.edu.au
University of Alberta, Edmonton, Canada

We present a new reinforcement-learning model for the role of the hippocampus in classical conditioning, focusing on the differences between trace and delay conditioning. In the model, all stimuli are represented both as unindividuated wholes and as a series of temporal elements with varying delays. These two stimulus representations interact, producing different patterns of learning in trace and delay conditioning. The model proposes that hippocampal lesions eliminate long-latency temporal elements, but preserve short-latency temporal elements. For trace conditioning, with no contiguity between stimulus and reward, these long-latency temporal elements are vital to learning adaptively timed responses. For delay conditioning, in contrast, the continued presence of the stimulus supports conditioned responding, and the short-latency elements suppress responding early in the stimulus. In accord with the empirical data, simulated hippocampal damage impairs trace conditioning, but not delay conditioning, at medium-length intervals. With longer intervals, learning is impaired in both procedures, and, with shorter intervals, in neither. In addition, the model makes novel predictions about the response topography with extended stimuli or post-training lesions. These results demonstrate how temporal contiguity, as in delay conditioning, changes the timing problem faced by animals, rendering it both easier and less susceptible to disruption by hippocampal lesions.

W12 *Characterizing response behavior in multisensory perception with conflicting cues*

Rama Natarajan rama@cs.toronto.edu
Iain Murray murray@cs.toronto.edu
University of Toronto, Toronto, ON, Canada
Ladan Shams ladan@psych.ucla.edu
University of California, Los Angeles CA, United States
Richard Zemel zemel@cs.toronto.edu
University of Toronto, Toronto, ON, Canada

We explore a recently proposed mixture model approach to understanding interactions between conflicting sensory cues. Alternative model formulations, differing in their sensory noise models and inference methods, are compared based on their fit to experimental data. Heavy-tailed sensory likelihoods yield a better description of the subjects' response behavior than standard Gaussian noise models. We study the underlying case for this result, and then present several testable predictions of these models.

W13 *Load and Attentional Bayes*

Peter Dayan, Gatsby Computational Neuroscience Unit.
Oral presentation, Thursday, 10:40am. See abstract, page 195.

W14 *Human Active Learning*

Rui M Castro rmcastro@ee.columbia.edu
Columbia University, New York, NY, United States
Charles Kalish cwkalish@wisc.edu
Robert Nowak nowak@ece.wisc.edu
Ruichen Qian qian2@wisc.edu
Tim Rogers ttrogers@wisc.edu
Xiaojin Zhu jerryzhu@cs.wisc.edu
University of Wisconsin-Madison, Madison, WI, United States

We investigate a topic at the interface of machine learning and cognitive science. Human active learning, where learners can actively query the world for information, is contrasted with passive learning from random examples. Furthermore, we compare human active learning performance with predictions from statistical learning theory. We conduct a series of human category learning experiments inspired by a machine learning task for which active and passive learning error bounds are well understood, and dramatically distinct. Our results indicate that humans are capable of actively selecting informative queries, and in doing so learn better and faster than if they are given random training data, as predicted by learning theory. However, the improvement over passive learning is not as dramatic as that achieved by machine active learning algorithms. To the best of our knowledge, this is the first quantitative study comparing human category learning in active versus passive settings.

W15 *Modeling the effects of memory on human online sentence processing with particle filters*

Roger P. Levy, University of California, San Diego, Florencia Reali and Thomas L Griffiths, University of California, Berkeley.
Oral presentation, Wednesday, 10:30am. See abstract, page 148.

W16 *A spatially varying two-sample recombinant coalescent, with applications to HIV escape response*

Alexander Braunstein braunsf@wharton.upenn.edu
Statistics Department, University of Pennsylvania, Wharton School, Philadelphia, PA, United States
Zhi Wei zhiwei@njit.edu
Department of Computer Science, New Jersey Institute of Technology, Newark, NJ, United States
Shane T. Jensen stjensen@wharton.upenn.edu
Jon McAuliffe jon@mcauliffe.com
Statistics Department, University of Pennsylvania, Wharton School, Philadelphia, PA, United States

Statistical evolutionary models provide an important mechanism for describing and understanding the escape response of a viral population under a particular therapy. We present a new hierarchical model that incorporates spatially varying mutation and recombination rates at the nucleotide level. It also maintains separate parameters for treatment and control groups, which allows us to estimate treatment effects explicitly. We use the model to investigate the sequence evolution of HIV populations exposed to a recently developed antisense gene therapy, as well as a more conventional drug therapy. The detection of biologically relevant and plausible signals in both therapy studies demonstrates the effectiveness of the method.

W17 *Global Ranking Using Continuous Conditional Random Fields*

Tao Qin, Tie-Yan Liu, Microsoft Research Asia, Xu-Dong Zhang, De-Sheng Wang, Tsinghua University, and Hang Li, Microsoft Research Asia.
Oral presentation, Wednesday, 4:40pm. See abstract, page 153.

W18 *Bayesian Synchronous Grammar Induction*

Phil Blunsom pblunsom@inf.ed.ac.uk
Trevor Cohn tcohn@inf.ed.ac.uk
Miles Osborne miles@inf.ed.ac.uk
University of Edinburgh, Edinburgh, United Kingdom

We present a novel method for inducing synchronous context free grammars (SCFGs) from a corpus of parallel string pairs. SCFGs can model equivalence between strings in terms of substitutions, insertions and deletions, and the reordering of sub-strings. We develop a non-parametric Bayesian model and apply it to a machine translation task, using priors to replace the various heuristics commonly used in this field. Using a variational Bayes training procedure, we learn the latent structure of translation equivalence through the induction of synchronous grammar categories for phrasal translations, showing improvements in translation performance over previously proposed maximum likelihood models.

W19 *Estimating Robust Query Models with Convex Optimization*

Kevyn Collins-Thompson kct@cs.cmu.edu
Carnegie Mellon University, Pittsburgh, PA, United States

Query expansion is a long-studied approach for improving retrieval effectiveness by enhancing the user's original query with additional related terms. Current algorithms for automatic query expansion have been shown to consistently improve retrieval accuracy on average, but are highly unstable and have bad worst-case performance for individual queries. We introduce a novel risk framework that formulates query model estimation as a constrained labeling problem on a graph of term relations. The model combines assignment costs based on a baseline feedback algorithm, edge weights based on term similarity, and simple constraints to enforce aspect balance, query coverage, and term centrality. Results across multiple standard test collections show consistent and significant reductions in the number and magnitude of expansion failures, while retaining the strong positive gains of the baseline algorithm.

W20 *Online Models for Content Optimization*

Deepak Agarwal dagarwal@yahoo-inc.com
Bee-Chung Chen beechun@yahoo-inc.com
Pradheep Elango elango@yahoo-inc.com
Yahoo! Research, Santa Clara, CA, United States
Nitin Motgi nmotgi@yahoo-inc.com
Yahoo! Inc, Sunnyvale, CA,United States
Seung-Taek Park parkst@yahoo-inc.com
Raghu Ramakrishnan ramakris@yahoo-inc.com
Yahoo! Research, Santa Clara, CA, United States
Scott Roy sroy@yahoo-inc.com
Joe Zachariah joez@yahoo-inc.com
Yahoo! Inc, Sunnyvale, CA, United States

We describe a new content publishing system that selects articles to serve to a user, choosing from an editorially programmed pool that is frequently refreshed. It is now deployed on a major Internet portal, and selects articles to serve to hundreds of millions of user visits per day, significantly increasing the number of user clicks over the original manual approach, in which editors periodically selected articles to display. Some of the challenges we face include a dynamic content pool, short article lifetimes, non-stationary click-through rates, and extremely high traffic volumes. The fundamental problem we must solve is to quickly identify which items are popular(perhaps within different user segments), and to exploit them while they remain current. We must also explore the underlying pool constantly to identify promising alternatives, quickly discarding poor performers. Our approach is based on tracking per article performance in near real time through online models. We describe the characteristics and constraints of our application setting, discuss our design choices, and show the importance and effectiveness of coupling online models with a simple randomization procedure. We discuss the challenges encountered in a production online content-publishing environment and highlight issues that deserve careful attention. Our analysis of this application also suggests a number of future research avenues.

W21 *Multi-label Multiple Kernel Learning*

Shuiwang Ji shuiwang.ji@asu.edu
Liang Sun sun.liang@asu.edu
Arizona State University, Tempe, AZ, United States
Rong Jin rongjin@cse.msu.edu
Michigan State University, East Lansing, MI, United States
Jieping Ye jieping.ye@asu.edu
Arizona State University, Tempe, AZ, United States

We present a multi-label multiple kernel learning (MKL) formulation in which the data are embedded into a low-dimensional space directed by the instance-label correlations encoded into a hypergraph. We formulate the problem in the kernel-induced feature space and propose to learn the kernel matrix as a linear combination of a given collection of kernel matrices in the MKL framework. The proposed learning formulation leads to a non-smooth min-max problem, which can be cast into a semi-infinite linear program (SILP). We further propose an approximate formulation with a guaranteed error bound which involves an unconstrained convex optimization problem. In addition, we show that the objective function of the approximate formulation is differentiable with Lipschitz continuous gradient, and hence existing methods can be employed to compute the optimal solution efficiently. We apply the proposed formulation to the automated annotation of Drosophila gene expression pattern images, and promising results have been reported in comparison with representative algorithms.

Spotlight presentation, Wednesday, 3:20pm.

W22 *Supervised Bipartite Graph Inference*

Yoshihiro Yamanishi yoshihiro.yamanishi@ensmp.fr
Mines ParisTech, Fontainebleau, France

We formulate the problem of bipartite graph inference as a supervised learning problem, and propose a new method to solve it from the viewpoint of distance metric learning. The method involves the learning of two mappings of the heterogeneous objects to a unified Euclidean space representing the network topology of the bipartite graph, where the graph is easy to infer. The algorithm can be formulated as an optimization problem in a reproducing kernel Hilbert space. We report encouraging results on the problem of compound-protein interaction network reconstruction from chemical structure data and genomic sequence data.

Spotlight presentation, Wednesday, 11:50am.

W23 *Scalable Algorithms for String Kernels with Inexact Matching*

Pavel P Kuksa pkuksa@cs.rutgers.edu
Pai-Hsi Huang paihuang@cs.rutgers.edu
Vladimir Pavlovic vladimir@cs.rutgers.edu
Rutgers University, Department of Computer Science, Piscataway, NJ, United States

We present a new family of linear time algorithms based on sufficient statistics for string comparison with mismatches under the string kernels framework. Our algorithms improve theoretical complexity bounds of existing approaches while scaling well with respect to the sequence alphabet size, the number of allowed mismatches and the size of the dataset. In particular, on large alphabets with loose mismatch constraints our algorithms are several orders of magnitude faster than the existing algorithms for string comparison under the mismatch similarity measure. We evaluate our algorithms on synthetic data and real applications in music genre classification, protein remote homology detection and protein fold prediction. The scalability of the algorithms allows us to consider complex sequence transformations, modeled using longer string features and larger numbers of mismatches, leading to a state-of-the-art performance with significantly reduced running times.

Spotlight presentation, Wednesday, 3:20pm.

W24 *Inferring rankings under constrained sensing*

Srikanth Jagabathula and Devavrat Shah, MIT.
Oral presentation, Wednesday, 5:00pm. See abstract, page 154.

W25 *High-dimensional union support recovery in multivariate regression*

Guillaume Obozinski gobo@stat.berkeley.edu
Martin Wainwright wainwrig@stat.berkeley.edu
Michael Jordan jordan@stat.berkeley.edu
Statistics Department, University of California, Berkeley, CA, United States

We study the behavior of block L1/L2 regularization for multivariate regression, where a K-dimensional response vector is regressed upon a fixed set of p covariates. We are interested in sparse multivariate regression; specifically, our goal is to recover the subset of covariates that are active in at least one of the regression problems. We refer to this as the union support recovery problem. Studying this problem under high-dimensional scaling (where the number p of problem parameters as well as sample size n tend to infinity simultaneously), our main result is to show that exact recovery is possible once the sample complexity parameter given by $\theta(n; p; s, B*) := n/(\psi(B*) \log(p - s))$ exceeds a critical threshold. Here n is the sample size, p is the ambient dimension of the regression model, s is the size of the union of supports, and $psi(B*)$ is a sparsity-overlap function that measures a combination of the sparsities and overlaps of the K-regression coefficient vectors that constitute the model. This sparsity-overlap function reveals that under some structural assumptions, block L1/L2 regularization for multivariate regression never harms performance relative to a naive L1-approach, and can yield substantial improvements in sample complexity (up to a factor of K) when the regression vectors are suitably orthogonal. Although our statements are made in an asymptotic form for clarity, the analysis itself is non-asymptotic in nature, yielding explicit bounds on loss for finite choices of the triple $(n; p; s)$. We complement our theoretical results with simulations that demonstrate the sharpness of the result, even for relatively small problems (e.g.,$p = 32$).

Spotlight presentation, Wednesday, 5:20pm.

W26 *Nonparametric regression and classification with joint sparsity constraints*

Han Liu hanliu@cs.cmu.edu
John Lafferty lafferty@cs.cmu.edu
Larry Wasserman larry@stat.cmu.edu
Carnegie Mellon University, Pittsburgh, PA 15213, United States

We propose new families of models and algorithms for high-dimensional nonparametric learning with joint sparsity constraints. Our approach is based on a regularization method that enforces common sparsity patterns across different function components in a nonparametric additive model. The algorithms employ a coordinate descent approach that is based on a functional soft-thresholding operator. The framework yields several new models, including multi-task sparse additive models, multi-response sparse additive models, and sparse additive multi-category logistic regression. The methods are illustrated with experiments on synthetic data and gene microarray data.

Spotlight presentation, Wednesday, 5:20pm.

W27 *Robust Regression and Lasso*

Huan Xu xuhuan@cim.mcgill.ca
McGill University, Montreal, QC, Canada
Constantine Caramanis cmcaram@ece.utexas.edu
The University of Texas at Austin, Austin, TX, United States
Shie Mannor shie.mannor@mcgill.ca
McGill University, Montreal, Canada

We consider robust least-squares regression with feature-wise disturbance. We show that this formulation leads to tractable convex optimization problems, and we exhibit a particular uncertainty set for which the robust problem is equivalent to ℓ_1 regularized regression (Lasso). This provides an interpretation of Lasso from a robust optimization perspective. We generalize this robust formulation to consider more general uncertainty sets, which all lead to tractable convex optimization problems. Therefore, we provide a new methodology for designing regression algorithms, which generalize known formulations. The advantage is that robustness to disturbance is a physical property that can be exploited: in addition to obtaining new formulations, we use it directly to show sparsity properties of Lasso, as well as to prove a general consistency result for robust regression problems, including Lasso, from a unified robustness perspective.

Spotlight presentation, Wednesday, 5:20pm.

W28 *An interior-point stochastic approximation method and an L1-regularized delta rule*

Peter Carbonetto, Mark Schmidt and Nando de Freitas, University of British Columbia.
Oral presentation, Wednesday, 4:20pm. *See abstract, page 153.*

W29 *Non-parametric Regression Between Manifolds*

Florian Steinke steinke@tuebingen.mpg.de
MPI for Biological Cybernetics, Tübingen, Germany
Matthias Hein hein@cs.uni-sb.de
Saarland University, Saarbruecken, Germany

This paper discusses non-parametric regression between Riemannian manifolds. This learning problem arises frequently in many application areas ranging from signal processing, computer vision, over robotics to computer graphics. We present a new algorithmic scheme for the solution of this general learning problem based on regularized empirical risk minimization. The regularization functional takes into account the geometry of input and output manifold, and we show that it implements a prior which is particularly natural. Moreover, we demonstrate that our algorithm performs well in a difficult surface registration problem.

W30 *Phase transitions for high-dimensional joint support recovery*

Sahand Negahban sahand_n@eecs.berkeley.edu
Department of Electrical Engineering and Computer Sciences,
University of California, Berkeley, CA, United States
Martin Wainwright wainwrig@eecs.berkeley.edu
Department of Electrical Engineering and Computer Sciences, and Department of Statistics, University of California, Berkeley, CA, United States

We consider the following instance of transfer learning: given a pair of regression problems, suppose that the regression coefficients share a partially common support, parameterized by the overlap fraction α between the two supports. This set-up suggests the use of $\ell_{1,\infty}$-regularized linear regression for recovering the support sets of both regression vectors. Our main contribution is to provide a sharp characterization of the sample complexity of this $\ell_{1,\infty}$ relaxation, exactly pinning down the minimal sample size n required for joint support recovery as a function of the model dimension p, support size s and overlap $\alpha \in [0,1]$. For measurement matrices drawn from standard Gaussian ensembles, we prove that the joint $\ell_{1,\infty}$-regularized method undergoes a phase transition characterized by order parameter $\theta_{1,\infty}(n,p,s,\alpha) = \frac{n}{(4-3\alpha)s\log(p-(2-\alpha)s)}$. More precisely, the probability of successfully recovering both supports converges to 1 for scalings such that $\theta_{1,\infty}>1$, and converges to 0 for scalings for which $\theta_{1,\infty}<1$. An implication of this threshold is that the use of the $\ell_{1,\infty}$-regularization leads to gains in sample complexity if the overlap parameter is large enough ($\alpha>2/3$), but performs worse than a naive approach if $\alpha<2/3$. We illustrate the close agreement between these theoretical predictions, and the actual behavior in simulations. Thus, our results illustrate both the benefits and dangers associated with block-$\ell_{1,\infty}$ regularization in high-dimensional inference.

Spotlight presentation, Wednesday, 3:20pm.

W31 *Domain Adaptation with Multiple Sources*

Yishay Mansour mansour@tau.ac.il
School of Computer Science, Tel Aviv, Israel
Mehryar Mohri mohri@cims.nyu.edu
Courant Institute of Mathematical Sciences, Google Research, New York, NY, United States
Afshin Rostamizadeh rostami@cs.nyu.edu
Courant Institute, New York University, New York, NY, United States

This paper presents a theoretical analysis of the problem of adaptation with multiple sources. For each source domain, the distribution over the input points as well as a hypothesis with error at most ϵ are given. The problem consists of combining these hypotheses to derive a hypothesis with small error with respect to the target domain. We present several theoretical results relating to this problem. In particular, we prove that standard convex combinations of the source hypotheses may in fact perform very poorly and that, instead, combinations weighted by the source distributions benefit from favorable theoretical guarantees. Our main result shows that, remarkably, for any fixed target function, there exists a distribution weighted combining rule that has a loss of at most ϵ with respect to *any* target mixture of the source distributions. We further generalize the setting from a single target function to multiple consistent target functions and show the existence of a combining rule with error at most 3ϵ. Finally, we report empirical results for a multiple source adaptation problem with a real-world dataset.

Spotlight presentation, Wednesday, 3:20pm.

W32 *Relative Margin Machines*

Pannagadatta Shivaswamy pks2103@cs.columbia.edu
Tony Jebara jebara@cs.columbia.edu
Columbia University, New York, NY, United States

In classification problems, Support Vector Machines maximize the margin of separation between two classes. While the paradigm has been successful, the solution obtained by SVMs is dominated by the directions with large data spread and biased to separate the classes by cutting along large spread directions. This article proposes a novel formulation to overcome such sensitivity and maximizes the margin relative to the spread of the data. The proposed formulation can be efficiently solved and experiments on digit datasets show drastic performance improvements over SVMs.

W33 *Clustered Multi-Task Learning: A Convex Formulation*

Laurent Jacob laurent.jacob@mines-paristech.fr
Mines ParisTech, Centre for Computational Biology, Fontainebleau, France - Institut Curie, Paris, France - INSERM U900, Paris, France
Francis Bach francis.bach@mines.org
INRIA - WILLOW Project-Team, Ecole Normale Superieure, Paris, France
Jean-Philippe Vert jean-philippe.vert@mines-paristech.fr
Mines ParisTech, Centre for Computational Biology, Fontainebleau, France - Institut Curie, Paris, France - INSERM U900, Paris, France

In multi-task learning several related tasks are considered simultaneously, with the hope that by an appropriate sharing of information across tasks, each task may benefit from the others. In the context of learning linear functions for supervised classification or regression, this can be achieved by including a priori information about the weight vectors associated with the tasks, and how they are expected to be related to each other. In this paper, we assume that tasks are clustered into groups, which are unknown beforehand, and that tasks within a group have similar weight vectors. We design a new spectral norm that encodes this a priori assumption, without the prior knowledge of the partition of tasks into groups, resulting in a new convex optimization formulation for multi-task learning. We show in simulations on synthetic examples and on the iedb MHC-I binding dataset, that our approach outperforms well-known convex methods for multi-task learning, as well as related non convex methods dedicated to the same problem.

Spotlight presentation, Wednesday, 3:20pm.

W34 *Characteristic Kernels on Groups and Semigroups*

Kenji Fukumizu, Institute of Statistical Mathematics, Bharath K Sriperumbudur, University of California, San Diego and Max Planck Institute for Biological Cybernetics, Arthur Gretton and Bernhard Schölkopf, Max Planck Institute for Biological Cybernetics.
Oral presentation, Wednesday, 3:00pm. *See abstract, page 151.*

W35 *Localized Sliced Inverse Regression*

Qiang Wu qiang@stat.duke.edu
ISDS, Duke University, Durham, NC, United States
Sayan Mukherjee sayan@stat.duke.edu
Duke University, Durham, United States
Feng Liang liangf@uiuc.edu
University of Illinois at Urbana-Champaign, Champaign, IL, United States

We developed localized sliced inverse regression for supervised dimension reduction. It has the advantages of preventing degeneracy, increasing estimation accuracy, and automatic subclass discovery in classification problems. A semisupervised version is proposed for the use of unlabeled data. The utility is illustrated on simulated as well as real data sets.

Spotlight presentation, Wednesday, 5:20pm.

W36 *Performance analysis for L_2 kernel classification*

JooSeuk Kim stannum@umich.edu
Clayton Scott clayscot@umich.edu
University of Michigan, Ann Arbor, MI, United States

We provide statistical performance guarantees for a recently introduced kernel classifier that optimizes the L_2 or integrated squared error (ISE) of a difference of densities. The classifier is similar to a support vector machine (SVM) in that it is the solution of a quadratic program and yields a sparse classifier. Unlike SVMs, however, the L_2 kernel classifier does not involve a regularization parameter. We prove a distribution free concentration inequality for a cross-validation based estimate of the ISE, and apply this result to deduce an oracle inequality and consistency of the classifier on the sense of both ISE and probability of error. Our results can also be specialized to give performance guarantees for an existing method of L_2 kernel density estimation.

Spotlight presentation, Wednesday, 3:20pm.

W37 *Adaptive Forward-Backward Greedy Algorithm for Sparse Learning with Linear Models*

Tong Zhang, Rutgers University.

Oral presentation, Wednesday, 4:00pm. *See abstract, page 153.*

W38 *Empirical performance maximization for linear rank statistics*

Stéphan J.M. Clémençon stephan.clemencon@telecom-paristech.fr
LTCI, Telecom ParisTech and CNRS, Institut Telecom, PARIS, France
Nicolas Vayatis nicolas.vayatis@gmail.com
Ecole Normale Supérieure de Cachan, CMLA - CNRS & UniverSud, Cachan, France

The ROC curve is known to be the golden standard for measuring performance of a test/scoring statistic regarding its capacity of discrimination between two populations in a wide variety of applications, ranging from anomaly detection in signal processing to information retrieval, through medical diagnosis. Most practical performance measures used in scoring applications such as the AUC, the local AUC, the p-norm push, the DCG and others, can be seen as summaries of the ROC curve. This paper highlights the fact that many of these empirical criteria can be expressed as (conditional) linear rank statistics. We investigate the properties of empirical maximizers of such performance criteria and provide preliminary results for the concentration properties of a novel class of random variables that we will call a linear rank process.

W39 *Kernel Measures of Independence for non-iid Data*

Xinhua Zhang xinhua.zhang@anu.edu.au
CSL, Australian National University and NICTA, Canberra, Australia
Le Song lesong@cs.cmu.edu
Carnegie Mellon University, Pittsburgh, PA, United States
Arthur Gretton arthur@tuebingen.mpg.de
Max Planck Institute for Biological Cybernetics, Tübingen, Germany
Alex J. Smola alex@smola.org
Yahoo! Research, Santa-Clara, CA, United States

Many machine learning algorithms can be formulated in the framework of statistical independence such as the Hilbert Schmidt Independence Criterion. In this paper, we extend this criterion to deal with structured and interdependent observations. This is achieved by modeling the structures using undirected graphical models and comparing the Hilbert space embeddings of distributions. We apply this new criterion to independent component analysis and sequence clustering.

Spotlight presentation, Wednesday, 3:20pm.

W40 *Fast Rates for Regularized Objectives*

KARTHIK SRIDHARAN karthik@tti-c.org
SHAI SHALEV-SHWARTZ shai@tti-c.org
NATHAN SREBRO nati@uchicago.edu
Toyota Technical Institute, Chicago, IL, United States

We show that the empirical minimizer of a stochastic strongly convex objective, where the stochastic component is linear, converges to the population minimizer with rate $O(1/n)$. The result applies, in particular, to the SVM objective. Thus, we get a rate of $O(1/n)$ on the convergence of the SVM objective to its infinite data limit. We demonstrate how this is essential for obtaining tight oracle inequalities for SVMs. The results extend also to strong convexity with respect to other ℓ_p norms, and so also to objectives regularized using other norms.

W41 *Learning with Consistency between Inductive Functions and Kernels*

HAIXUAN YANG hxyang@cse.cuhk.edu.hk
IRWIN KING king@cse.cuhk.edu.hk
MICHAEL LYU lyu@cse.cuhk.edu.hk
Department of Computer Science & Engineering, The Chinese University of Hong Kong, Shatin, Hong Kong

Regularized Least Squares (RLS) algorithms have the ability to avoid over-fitting problems and to express solutions as kernel expansions. However, we observe that the current RLS algorithms cannot provide a satisfactory interpretation even on a constant function. On the other hand, while kernel-based algorithms have been developed in such a tendency that almost all learning algorithms are kernelized or being kernelized, a basic fact is often ignored: The learned function from the data and the kernel fits the data well, but may not be consistent with the kernel. Based on these considerations and on the intuition that a good kernel-based inductive function should be consistent with both the data and the kernel, a novel learning scheme is proposed. The advantages of this scheme lie in its corresponding Representer Theorem, its strong interpretation ability about what kind of functions should not be penalized, and its promising accuracy improvements shown in a number of experiments. Furthermore, we provide a detailed technical description about heat kernels, which serves as an example for the readers to apply similar techniques for other kernels. Our work provides a preliminary step in a new direction to explore the varying consistency between inductive functions and kernels under various distributions.

Spotlight presentation, Wednesday, 3:20pm.

W42 *Translated Learning: Transfer Learning across Different Feature Spaces*

WENYUAN DAI dwyak@apex.sjtu.edu.cn
YUQIANG CHEN yuqiangchen@apex.sjtu.edu.cn
GUI-RONG XUE grxue@apex.sjtu.edu.cn
Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China
QIANG YANG qyang@cse.ust.hk
Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong
YONG YU yyu@apex.sjtu.edu.cn
Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China

This paper investigates a new machine learning strategy called *translated learning*. Unlike many previous learning tasks, we focus on how to use labeled data from one feature space to enhance the classification of other entirely different learning spaces. For example, we might wish to use labeled text data to help learn a model for classifying image data, when the labeled images are difficult to obtain. An important aspect of translated learning is to build a "bridge" to link one feature space (known as the "source space") to another space (known as the "target space") through a translator in order to migrate the knowledge from source to target. The translated learning solution uses a language model to link the class labels to the features in the source spaces, which in turn is *translated* to the features in the target spaces. Finally, this chain of linkages is completed by tracing back to the instances in the target spaces. We show that this path of linkage can be modeled using a Markov chain and risk minimization. Through experiments on the text-aided image classification and cross-language classification tasks, we demonstrate that our translated learning framework can greatly outperform many state-of-the-art baseline methods.

Spotlight presentation, Wednesday, 3:20pm.

W43 *ICA based on a Smooth Estimation of the Differential Entropy*

LEV FAIVISHEVSKY levtemp@gmail.com
JACOB GOLDBERGER goldbej@eng.biu.ac.il
Bar Ilan University, Ramat-Gan, Israel

In this paper we introduce the MeanNN approach for estimation of main information theoretic measures such as differential entropy, mutual information and divergence. As opposed to other nonparametric approaches the MeanNN results in smooth differentiable functions of the data samples with clear geometrical interpretation. Then we apply the proposed estimators to the ICA problem and obtain a smooth expression for the mutual information that can be analytically optimized by gradient descent methods. The improved performance on the proposed ICA algorithm is demonstrated on standard tests in comparison with state-of-the-art techniques.

Spotlight presentation, Wednesday, 5:20pm.

W44 *Robust Kernel Principal Component Analysis*

Minh Hoai Nguyen minhhoai@gmail.com
Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, United States
Fernando De la Torre ftorre@cs.cmu.edu
Carnegie Mellon University, Pittsburgh, PA, United States

Kernel Principal Component Analysis (KPCA) is a popular generalization of linear PCA that allows non-linear feature extraction. In KPCA, data in the input space is mapped to higher (usually) dimensional feature space where the data can be linearly modeled. The feature space is typically induced implicitly by a kernel function, and linear PCA in the feature space is performed via the kernel trick. However, due to the implicitness of the feature space, some extensions of PCA such as robust PCA cannot be directly generalized to KPCA. This paper presents a technique to overcome this problem, and extends it to a unified framework for treating noise, missing data, and outliers in KPCA. Our method is based on a novel cost function to perform inference in KPCA. Extensive experiments, in both synthetic and real data, show that our algorithm outperforms existing methods.

Spotlight presentation, Wednesday, 5:20pm.

W45 *Bayesian Exponential Family PCA*

Shakir Mohamed sm694@eng.cam.ac.uk
Katherine Heller kheller@gmail.com
University of Cambridge, Cambridge, United Kingdom
Zoubin Ghahramani zoubin@eng.cam.ac.uk
University of Cambridge & CMU, Cambridge, United Kingdom

Principal Components Analysis (PCA) has become established as one of the key tools for dimensionality reduction when dealing with real valued data. Approaches such as exponential family PCA and non-negative matrix factorisation have successfully extended PCA to non-Gaussian data types, but these techniques fail to take advantage of Bayesian inference and can suffer from problems of overfitting and poor generalisation. This paper presents a fully probabilistic approach to PCA, which is generalised to the exponential family, based on Hybrid Monte Carlo sampling. We describe the model which is based on a factorisation of the observed data matrix, and show performance of the model on both synthetic and real data.

Spotlight presentation, Wednesday, 5:20pm.

W46 *Deflation Methods for Sparse PCA*

Lester Mackey lmackey@cs.berkeley.edu
University of California, Berkeley, Berkeley, CA

In analogy to the PCA setting, the sparse PCA problem is often solved by iteratively alternating between two subtasks: cardinality-constrained rank-one variance maximization and matrix deflation. While the former has received a great deal of attention in the literature, the latter is seldom analyzed and is typically borrowed without justification from the PCA context. In this work, we demonstrate that the standard PCA deflation procedure is seldom appropriate for the sparse PCA setting. To rectify the situation, we first develop several deflation alternatives better suited to the cardinality-constrained context. We then reformulate the sparse PCA optimization problem to explicitly reflect the maximum *additional* variance objective on each round. The result is a generalized deflation procedure that typically outperforms more standard techniques on real-world datasets.

Spotlight presentation, Wednesday, 5:20pm.

W47 *Sparse probabilistic projections*

CÉDRIC ARCHAMBEAU c.archambeau@cs.ucl.ac.uk
University College London, United Kingdom
FRANCIS BACH francis.bach@mines.org
INRIA - Ecole Normale Superieure, Paris, France

We present a generative model for performing sparse probabilistic projections, which includes sparse principal component analysis and sparse canonical correlation analysis as special cases. Sparsity is enforced by means of automatic relevance determination or by imposing appropriate prior distributions, such as generalised hyperbolic distributions. We derive a variational Expectation-Maximisation algorithm for the estimation of the hyperparameters and show that our novel probabilistic approach compares favourably to existing techniques. We illustrate how the proposed method can be applied in the context of cryptoanalysis as a pre-processing tool for the construction of template attacks.

Spotlight presentation, Wednesday, 5:20pm.

W48 *Kernelized Sorting*

NOVI QUADRIANTO novi.quad@gmail.com
NICTA - Statistical Machine Learning, Australian National University, Canberra ACT, Australia
LE SONG lesong@cs.cmu.edu
Carnegie Mellon University, Pittsburgh, PA, United States
ALEX J. SMOLA alex@smola.org
Yahoo! Research, Santa-Clara, CA, United States

Object matching is a fundamental operation in data analysis. It typically requires the definition of a similarity measure between the classes of objects to be matched. Instead, we develop an approach which is able to perform matching by requiring a similarity measure only within each of the classes. This is achieved by maximizing the dependency between matched pairs of observations by means of the Hilbert Schmidt Independence Criterion. This problem can be cast as one of maximizing a quadratic assignment problem with special structure and we present a simple algorithm for finding a locally optimal solution.

Spotlight presentation, Wednesday, 3:20pm.

W49 *QUIC-SVD: Fast SVD Using Cosine Trees*

Michael Holmes mph@cc.gatech.edu
Alex Gray agray@cc.gatech.edu
Charles Isbell isbell@cc.gatech.edu
College of Computing, Georgia Tech, Atlanta, GA, United States

The Singular Value Decomposition is a key operation in many machine learning methods. Its computational cost, however, makes it unscalable and impractical for the massive-sized datasets becoming common in applications. We present a new Monte Carlo method, QUIC-SVD, for fast approximation of the full SVD with automatic sample size minimization and empirical relative error control. Previous Monte Carlo approaches have not addressed the full SVD nor benefited from the efficiency of automatic, empirically-driven sample sizing. Our empirical tests show speedups of several orders of magnitude over exact SVD. Such scalability should enable QUIC-SVD to meet the needs of a wide array of methods and applications.

W50 *Continuously-adaptive discretization for message-passing algorithms*

Michael Isard misard@microsoft.com
Microsoft Research (SVC), Mountain View, CA, United States
John MacCormick jmac@dickinson.edu
Dickinson College, Carlisle, PA, United States
Kannan Achan kannan.achan@gmail.com
Yahoo!, Mountain View, United States

Continuously-Adaptive Discretization for Message-Passing (CAD-MP) is a new message-passing algorithm employing adaptive discretization. Most previous message-passing algorithms approximated arbitrary continuous probability distributions using either: a family of continuous distributions such as the exponential family; a particle-set of discrete samples; or a fixed, uniform discretization. In contrast, CAD-MP uses a discretization that is (i) non-uniform, and (ii) adaptive. The non-uniformity allows CAD-MP to localize interesting features (such as sharp peaks) in the marginal belief distributions with time complexity that scales logarithmically with precision, as opposed to uniform discretization which scales at best linearly. We give a principled method for altering the non-uniform discretization according to information-based measures. CAD-MP is shown in experiments on simulated data to estimate marginal beliefs much more precisely than competing approaches for the same computational expense.

W51 *Variational Mixture of Gaussian Process Experts*

Chao Yuan chao.yuan@siemens.com
Claus Neubauer claus.neubauer@siemens.com
Siemens Corporate Research, Princeton, NJ, United States

Mixture of Gaussian processes models extended a single Gaussian process with ability of modeling multi-modal data and reduction of training complexity. Previous inference algorithms for these models are mostly based on Gibbs sampling, which can be very slow, particularly for large-scale data sets. We present a new generative mixture of experts model. Each expert is still a Gaussian process but is reformulated by a linear model. This breaks the dependency among training outputs and enables us to use a much faster variational Bayesian algorithm for training. Our gating network is more flexible than previous generative approaches as inputs for each expert are modeled by a Gaussian mixture model. The number of experts and number of Gaussian components for an expert are inferred automatically. A variety of tests show the advantages of our method.

W52 *A Convex Upper Bound on the Log-Partition Function for Binary Distributions.*

Laurent El Ghaoui elghaoui@eecs.berkeley.edu
University of California at Berkeley, Berkeley, CA94720, United States
Assane Gueye agueye@eecs.berkeley.edu
University of California at Berkeley, Berkeley, CA 94720, United States

We consider the problem of bounding from above the log-partition function corresponding to second-order Ising models for binary distributions. We introduce a new bound, the cardinality bound, which can be computed via convex optimization. The corresponding error on the log-partition function is bounded above by twice the distance, in model parameter space, to a class of "standard" Ising models, for which variable inter-dependence is described via a simple mean field term. In the context of maximum-likelihood, using the new bound instead of the exact log-partition function, while constraining the distance to the class of standard Ising models, leads not only to a good approximation to the log-partition function, but also to a model that is parsimonious, and easily interpretable. We compare our bound with the log-determinant bound introduced by Wainwright and Jordan (2006), and show that when the l_1-norm of the model parameter vector is small enough, the latter is outperformed by the new bound.

W53 *Mixed Membership Stochastic Blockmodels*

Edoardo M. Airoldi eairoldi@princeton.edu
Computer Science Department, Lewis-Sigler Institute for Integrative Genomics, Princeton, NJ, United States
David M. Blei blei@cs.princeton.edu
Computer Science Department, Princeton University, Princeton, NJ, United States
Stephen E. Fienberg fienberg@stat.cmu.edu
Department of Statistics, School of computer Science, Carnegie Mellon University, Pittsburgh, PA, United States
Eric P. Xing epxing@cs.cmu.edu
School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, United States

Consider data consisting of pairwise measurements, such as presence or absence of links between pairs of objects. These data arise, for instance, in the study of protein interactions and gene regulatory networks, collections of author-recipient email, and social networks. Analyzing pairwise measurements with probabilistic models requires special assumptions, since the usual independence or exchangeability assumptions no longer hold. Here we introduce a class of variance allocation models for pairwise measurements: mixed membership stochastic blockmodels. These models combine global parameters that instantiate dense patches of connectivity (blockmodel) with local parameters that instantiate node-specific variability in the connections (mixed membership). We develop a general variational inference algorithm for fast approximate posterior inference. We demonstrate the advantages of mixed membership stochastic blockmodels with applications to social networks and protein interaction networks.

Spotlight presentation, Wednesday, 11:50am.

W54 *Stochastic Relational Models for Large-scale Dyadic Data using MCMC*

Shenghuo Zhu zsh@sv.nec-labs.com
Kai Yu kyu@sv.nec-labs.com
Yihong Gong ygong@sv.nec-labs.com
NEC Laboratories America, Inc., Cupertino, CA, United States

Stochastic relational models (SRMs) provide a rich family of choices for learning and predicting dyadic data between two sets of entities. The models generalize matrix factorization to a supervised learning problem that utilizes attributes of entities in a hierarchical Bayesian framework. Previously empirical Bayes inference was applied for SRMs, which is, however, not scalable when the size of either entity set grows to tens of thousands. In this paper, we introduce a Markov chain Monte Carlo (MCMC) algorithm for equivalent models of SRMs in order to scale the computation to very large dyadic data sets. Both superior scalability and predictive accuracy are demonstrated on a collaborative filtering problem, which involves tens of thousands users and half million items.

Spotlight presentation, Wednesday, 11:50am.

W55 *The Mondrian Process*

Daniel M Roy, CSAIL/MIT, and Yee Whye Teh,
Gatsby Computational Neuroscience Unit, UCL.
Oral presentation, Thursday, 8:30am. *See abstract, page 193.*

W56 *Asynchronous Distributed Learning of Topic Models*

Arthur Asuncion asuncion@ics.uci.edu
Padhraic Smyth smyth@ics.uci.edu
Max Welling welling@ics.uci.edu
University of California, Irvine, CA, United States

Distributed learning is a problem of fundamental interest in machine learning and cognitive science. In this paper, we present asynchronous distributed learning algorithms for two well-known unsupervised learning frameworks: Latent Dirichlet Allocation (LDA) and Hierarchical Dirichlet Processes (HDP). In the proposed approach, the data are distributed across P processors, and processors independently perform Gibbs sampling on their local data and communicate their information in a local asynchronous manner with other processors. We demonstrate that our asynchronous algorithms are able to learn global topic models that are statistically as accurate as those learned by the standard LDA and HDP samplers, but with significant improvements in computation time and memory. We show speedup results on a 730-million-word text corpus using 32 processors, and we provide perplexity results for up to 1500 virtual processors. As a stepping stone in the development of asynchronous HDP, a parallel HDP sampler is also introduced.

W57 *Online Optimization in X-Armed Bandits*

Sébastien Bubeck sebastien.bubeck@inria.fr
Remi Munos remi.munos@inria.fr
INRIA Lille - Nord Europe, SequeL project, France
Gilles Stoltz gilles.stoltz@ens.fr
ENS Paris, CNRS, HEC Paris, France
Csaba Szepesvari szepesva@cs.ualberta.ca
University of Alberta, Edmonton, AB, Canada

We consider a generalization of stochastic bandit problems where the set of arms, X, is allowed to be a generic topological space. We constraint the mean-payoff function with a dissimilarity function over X in a way that is more general than Lipschitz. We construct an arm selection policy whose regret improves upon previous result for a large class of problems. In particular, our results imply that if X is the unit hypercube in a Euclidean space and the mean-payoff function has a finite number of global maxima around which the behavior of the function is locally Hölder with a known exponent, then the expected regret is bounded up to a logarithmic factor by $\sqrt{n}$, i.e., the rate of the growth of the regret is independent of the dimension of the space. Moreover, we prove the minimax optimality of our algorithm for the class of mean-payoff functions we consider.

W58 *Bayesian Kernel Shaping for Learning Control*

Jo-Anne Ting joanneti@usc.edu
Mrinal Kalakrishnan kalakris@usc.edu
University of Southern California, Los Angeles, CA, United States
Sethu Vijayakumar sethu.vijayakumar@ed.ac.uk
University of Edinburgh, Edinburgh, United Kingdom
Stefan Schaal sschaal@usc.edu
University of Southern California, Los Angeles, CA, United States, and
ATR Computational Neuroscience Laboratories, Kyoto, Japan

In kernel-based regression learning, optimizing each kernel individually is useful when the data density, curvature of regression surfaces (or decision boundaries) or magnitude of output noise (i.e., heteroscedasticity) varies spatially. Unfortunately, it presents a complex computational problem as the danger of overfitting is high and the individual optimization of every kernel in a learning system may be overly expensive due to the introduction of too many open learning parameters. Previous work has suggested gradient descent techniques or complex statistical hypothesis methods for local kernel shaping, typically requiring some amount of manual tuning of meta parameters. In this paper, we focus on nonparametric regression and introduce a Bayesian formulation that, with the help of variational approximations, results in an EM-like algorithm for simultaneous estimation of regression and kernel parameters. The algorithm is computationally efficient (suitable for large data sets), requires no sampling, automatically rejects outliers and has only one prior to be specified. It can be used for nonparametric regression with local polynomials or as a novel method to achieve nonstationary regression with Gaussian Processes. Our methods are particularly useful for learning control, where reliable estimation of local tangent planes is essential for adaptive controllers and reinforcement learning. We evaluate our methods on several synthetic data sets and on an actual robot which learns a task-level control law.

W59 *Nonparametric Bayesian Learning of Switching Linear Dynamical Systems*

Emily Fox ebfox@mit.edu
Massachusetts Institute of Technology, Cambridge, MA, United States
Erik Sudderth sudderth@eecs.berkeley.edu
Michael Jordan jordan@cs.berkeley.edu
University of California, Berkeley, Berkeley, CA, United States
Alan Willsky willsky@mit.edu
Massachusetts Institute of Technology, Cambridge, MA, United States

Many nonlinear dynamical phenomena can be effectively modeled by a system that switches among a set of conditionally linear dynamical modes. We consider two such models: the switching linear dynamical system (SLDS) and the switching vector autoregressive (VAR) process. In this paper, we present a nonparametric approach to the learning of an unknown number of persistent, smooth dynamical modes by utilizing a hierarchical Dirichlet process prior. We develop a sampling algorithm that combines a truncated approximation to the Dirichlet process with efficient joint sampling of the mode and state sequences. The utility and flexibility of our model are demonstrated on synthetic data, sequences of dancing honey bees, and the IBOVESPA stock index.

Spotlight presentation, Wednesday, 11:50am.

W60 *Adapting to a Market Shock: Optimal Sequential Market-Making*

Sanmay Das sanmay@cs.rpi.edu
Malik Magdon-Ismail magdon@cs.rpi.edu
Rensselaer Polytechnic Institute, Troy, NY, United States

We study the profit-maximization problem of a monopolistic market-maker who sets two-sided prices in an asset market. The sequential decision problem is hard to solve because the state space is a function. We demonstrate that the belief state is well approximated by a Gaussian distribution. We prove a key monotonicity property of the Gaussian state update which makes the problem tractable, yielding the first optimal sequential market-making algorithm in an established model. The algorithm leads to a surprising insight: an optimal monopolist can provide more liquidity than perfectly competitive market-makers in periods of extreme uncertainty, because a monopolist is willing to absorb initial losses in order to learn a new valuation rapidly so she can extract higher profits later.

W61 *Automatic online tuning for fast Gaussian summation*

Vlad I Morariu morariu@umd.edu
Balaji Vasan Srinivasan balajiv@umiacs.umd.edu
University of Maryland, College Park, MD, United States
Vikas C Raykar vikas.raykar@siemens.com
Siemens Medical Solutions Inc., USA, King of Prussia, United States
Ramani Duraiswami ramani@umiacs.umd.edu
Larry Davis lsd@cs.umd.edu
University of Maryland, College Park, MD, United States

Many machine learning algorithms require the summation of Gaussian kernel functions, an expensive operation if implemented straightforwardly. Several methods have been proposed to reduce the computational complexity of evaluating such sums, including tree and analysis based methods. These achieve varying speedups depending on the bandwidth, dimension, and prescribed error, making the choice between methods difficult for machine learning tasks. We provide an algorithm that combines tree methods with the Improved Fast Gauss Transform (IFGT). As originally proposed the IFGT suffers from two problems: (1) the Taylor series expansion does not perform well for very low bandwidths, and (2) parameter selection is not trivial and can drastically affect performance and ease of use. We address the first problem by employing a tree data structure, resulting in four evaluation methods whose performance varies based on the distribution of sources and targets and input parameters such as desired accuracy and bandwidth. To solve the second problem, we present an online tuning approach that results in a black box method that automatically chooses the evaluation method and its parameters to yield the best performance for the input data, desired accuracy, and bandwidth. In addition, the new IFGT parameter selection approach allows for tighter error bounds. Our approach chooses the fastest method at negligible additional cost, and has superior performance in comparisons with previous approaches.

W62 *Sparse Convolved Gaussian Processes for Multi-output Regression*

Mauricio Alvarez alvarezm@cs.man.ac.uk
University of Manchester, School of Computer Science, Manchester, United Kingdom
Neil D Lawrence neill@cs.man.ac.uk
School of Computer Science, University of Manchester, Manchester

We present a sparse approximation approach for dependent output Gaussian processes (GP). Employing a latent function framework, we apply the convolution process formalism to establish dependencies between output variables, where each latent function is represented as a GP. Based on these latent functions, we establish an approximation scheme using a conditional independence assumption between the output processes, leading to an approximation of the full covariance which is determined by the locations at which the latent functions are evaluated. We show results of the proposed methodology for synthetic data and real world applications on pollution prediction and a sensor network.

W63 *The Infinite Factorial Hidden Markov Model*

Jurgen Van Gael jv279@cam.ac.uk
University of Cambridge, Cambridge, United Kingdom
Yee Whye Teh ywteh@gatsby.ucl.ac.uk
Gatsby Computational Neuroscience Unit, UCL, London, United Kingdom
Zoubin Ghahramani zoubin@eng.cam.ac.uk
CMU & University of Cambridge, Cambridge, United Kingdom

We introduces a new probability distribution over a potentially infinite number of binary Markov chains which we call the Markov Indian buffet process. This process extends the IBP to allow temporal dependencies in the hidden variables. We use this stochastic process to build a nonparametric extension of the factorial hidden Markov model. After working out an inference scheme which combines slice sampling and dynamic programming we demonstrate how the infinite factorial hidden Markov model can be used for blind source separation.

Spotlight presentation, Wednesday, 11:50am.

W64 *Using matrices to model symbolic relationship*

Ilya Sutskever ilya@cs.utoronto.ca
Geoffrey E Hinton hinton@cs.toronto.edu
University of Toronto, Toronto, ON, Canada

We describe a way of learning matrix representations of objects and relationships. The goal of learning is to allow multiplication of matrices to represent symbolic relationships between objects and symbolic relationships between relationships, which is the main novelty of the method. We demonstrate that this leads to excellent generalization in two different domains: modular arithmetic and family relationships. We show that the same system can learn first-order propositions such as $(2, 5) \in +3$ or $(Christopher, Penelope) \in has\_wife$, and higher-order propositions such as $(3, +3) \in plus$ and $(+3, -3) \in inverse$ or $(has\_husband, has\_wife) \in higher\_oppsex$. We further demonstrate that the system understands how higher-order propositions are related to first-order ones by showing that it can correctly answer questions about first-order propositions involving the relations $+3$ or has_wife even though it has not been trained on any first-order examples involving these relations.

Spotlight presentation, Wednesday, 11:50am.

W65 *Covariance Estimation for High Dimensional Data Vectors Using the Sparse Matrix Transform*

Guangzhi Cao gcao@purdue.edu
Charles Bouman bouman@purdue.edu
School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN, United States

Covariance estimation for high dimensional vectors is a classically difficult problem in statistical analysis and machine learning due to limited sample size. In this paper, we propose a new approach to covariance estimation, which is based on constrained maximum likelihood (ML) estimation of the covariance. Specifically, the covariance is constrained to have an eigen decomposition which can be represented as a sparse matrix transform (SMT). The SMT is formed by a product of pairwise coordinate rotations known as Givens rotations. Using this framework, the covariance can be efficiently estimated using greedy minimization of the log likelihood function, and the number of Givens rotations can be efficiently computed using a cross-validation procedure. The estimator obtained using this method is always positive definite and well-conditioned even with limited sample size. Experiments on hyperspectral data show that SMT covariance estimation results in consistently better estimates of the covariance for a variety of different classes and sample sizes compared to traditional shrinkage estimators.

W66 *Cascaded Classification Models: Combining Models for Holistic Scene Understanding*

Geremy Heitz, Stephen Gould, Ashutosh Saxena and Daphne Koller, Stanford University.
Oral presentation, Thursday, 9:10am. See abstract, page 193.

W67 *Multi-Level Active Prediction of Useful Image Annotations for Recognition*

Sudheendra Vijayanarasimhan and Kristen Grauman, University of Texas at Austin.
Oral presentation, Thursday, 9:30am. See abstract, page 194.

W68 *Nonrigid Structure from Motion in Trajectory Space*

Ijaz Akhter, LUMS School of Science and Engineering, Yaser Sheikh, Carnegie Mellon University, Sohaib Khan, LUMS School of Science and Engineering, and Takeo Kanade, Carnegie Mellon University.
Oral presentation, Wednesday, 11:10am. See abstract, page 148.

W69 *Recursive Segmentation and Recognition Templates for 2D Parsing*

Leo Zhu leozhu@csail.mit.edu
CSAIL MIT, Cambridge, MA, United States
Yuanhao Chen yhchen4@ustc.edu
USTC, Hefei, China
Yuan Lin loirey@sjtu.edu.cn
SJTU, Shanghai, China
Chenxi Lin chenxil@microsoft.com
Microsoft Research Asia, BEIJING, China
Alan Yuille yuille@stat.ucla.edu
UCLA, Los Angeles, CA, United States

Language and image understanding are two major goals of artificial intelligence which can both be conceptually formulated in terms of parsing the input signal into a hierarchical representation. Natural language researchers have made great progress by exploiting the 1D structure of language to design efficient polynomial-time parsing algorithms. By contrast, the two-dimensional nature of images makes it much harder to design efficient image parsers and the form of the hierarchical representations is also unclear. Attempts to adapt representations and algorithms from natural language have only been partially successful.

In this paper, we propose a Hierarchical Image Model (HIM) for 2D image parsing which outputs image segmentation and object recognition. This HIM is represented by recursive segmentation and recognition templates in multiple layers and has advantages for representation, inference, and learning. Firstly, the HIM has a coarse-to-fine representation which is capable of capturing long-range dependency and exploiting different levels of contextual information. Secondly, the structure of the HIM allows us to design a rapid inference algorithm, based on dynamic programming, which enables us to parse the image rapidly in polynomial time. Thirdly, we can learn the HIM efficiently in a discriminative manner from a labeled dataset. We demonstrate that HIM outperforms other state-of-the-art methods by evaluation on the challenging public MSRC image dataset. Finally, we sketch how the HIM architecture can be extended to model more complex image phenomena.

Spotlight presentation, Wednesday, 11:50am.

W70 *Learning a discriminative hidden part model for human action recognition*

Yang Wang ywang12@cs.sfu.ca
Greg Mori mori@cs.sfu.ca
Simon Fraser University, Burnaby, Canada

We present a discriminative part-based approach for human action recognition from video sequences using motion features. Our model is based on the recently proposed hidden conditional random field∼(hCRF) for object recognition. Similar to hCRF for object recognition, we model a human action by a flexible constellation of parts conditioned on image observations. Different from object recognition, our model combines both large-scale global features and local patch features to distinguish various actions. Our experimental results show that our model is comparable to other state-of-the-art approaches in action

recognition. In particular, our experimental results demonstrate that combining large-scale global features and local patch features performs significantly better than directly applying hCRF on local patches alone.

Spotlight presentation, Wednesday, 11:50am.

W71 *Sparse Signal Recovery Using Markov Random Fields*

Volkan Cevher volkan@rice.edu
Marco F Duarte duarte@rice.edu
Chinmay Hegde ch3@rice.edu
Richard Baraniuk richb@rice.edu
Electrical and Computer Engineering, Rice University, Houston, TX, United States

Compressive Sensing (CS) combines sampling and compression into a single sub-Nyquist linear measurement process for sparse and compressible signals. In this paper, we extend the theory of CS to include signals that are concisely represented in terms of a graphical model. In particular, we use Markov Random Fields (MRFs) to represent sparse signals whose nonzero coefficients are clustered. Our new model-based reconstruction algorithm, dubbed Lattice Matching Pursuit (LaMP), stably recovers MRF-modeled signals using many fewer measurements and computations than the current state-of-the-art algorithms.

Spotlight presentation, Wednesday, 5:20pm.

W72 *Extended Grassmann Kernels for Subspace-Based Learning*

Jihun Hamm jhham@seas.upenn.edu
Daniel D Lee ddlee@seas.upenn.edu
University of Pennsylvania, Philadelphia, PA, United States

Subspace-based learning problems involve data whose elements are linear subspaces of a vector space. To handle such data structures, Grassmann kernels have been proposed and used previously. In this paper, we analyze the relationship between Grassmann kernels and probabilistic similarity measures. Firstly, we show that the KL distance in the limit yields the Projection kernel on the Grassmann manifold, whereas the Bhattacharyya kernel becomes trivial in the limit and is suboptimal for subspace-based problems. Secondly, based on our analysis of the KL distance, we propose extensions of the Projection kernel which can be extended to the set of affine as well as scaled subspaces. We demonstrate the advantages of these extended kernels for classification and recognition tasks with Support Vector Machines and Kernel Discriminant Analysis using synthetic and real image databases.

W73 *Learning Transformational Invariants from Natural Movies*

Charles Cadieu cadieu@berkeley.edu
Bruno Olshausen baolshausen@berkeley.edu
Helen Wills Neuroscience Institute, University of California, Berkeley, CA, United States

We describe a hierarchical, probabilistic model that learns to extract complex motion from movies of the natural environment. The model consists of two hidden layers: the first layer produces a sparse representation of the image that is expressed in terms of local amplitude and phase variables. The second layer learns the higher-order structure among the time-varying phase variables. After training on natural movies, the top layer units discover the structure of phase-shifts within the first layer. We show that the top layer units encode transformational invariants: they are selective for the speed and direction of a moving pattern, but are invariant to its spatial structure (orientation/spatial-frequency). The diversity of units in both the intermediate and top layers of the model provides a set of testable predictions for representations that might be found in V1 and MT. In addition, the model demonstrates how feedback from higher levels can influence representations at lower levels as a by-product of inference in a graphical model.

Spotlight presentation, Wednesday, 11:50am.

W74 *Shared Segmentation of Natural Scenes Using Dependent Pitman-Yor Processes*

Erik Sudderth and Michael Jordan, University of California, Berkeley.
Oral presentation, Thursday, 8:50am. *See abstract, page 193.*

W75 *Model selection and velocity estimation using novel priors for motion patterns*

Shuang Wu, Hongjing Lu and Alan Yuille, University of California, Los Angeles.
Oral presentation, Wednesday, 11:30am. *See abstract, page 149.*

W76 *Using Bayesian Dynamical Systems for Motion Template Libraries*

Silvia Chiappa silvia.chiappa@tuebingen.mpg.de
Jens Kober jens.kober@tuebingen.mpg.de
Jan Peters mail@jan-peters.net
MPI for Biological Cybernetics, Tuebingen, Germany

Motor primitives or motion templates have become an important concept for both modeling human motor control as well as generating robot behaviors using imitation learning. Recent impressive results range from humanoid robot movement generation to timing models of human motions. The automatic generation of skill libraries containing multiple motion templates is an important step in robot learning. Such a skill learning system needs to cluster similar movements together and represent each resulting motion template as a generative model which is subsequently used for the execution of the behavior by a robot system. In this paper, we show how human trajectories captured as multidimensional time-series can be clustered using Bayesian mixtures of linear Gaussian state-space models based on the similarity of their dynamics. The appropriate number of templates is automatically determined by enforcing a parsimonious parametrization. As the resulting model is intractable, we introduce a novel approximation method based on variational Bayes, which is especially designed to enable the use of efficient inference algorithms. On recorded human Balero movements, this method is not only capable of finding reasonable motion templates but also yields a generative model which works well in the execution of this complex task on a simulated anthropomorphic SARCOS arm.

W77 *Dynamic visual attention: searching for coding length increments*

XIAODI HOU xiaodi.hou@gmail.com
California Institute of Technology, Pasadena, CA, United States
LIQING ZHANG zhang-lq@cs.sjtu.edu.cn
Shanghai Jiao Tong University, Shanghai, China

A visual attention system should respond placidly when common stimuli are presented, while at the same time keep alert to anomalous visual inputs. In this paper, a dynamic visual attention model based on the rarity of features is proposed. We introduce the Incremental Coding Length (ICL) to measure the perspective entropy gain of each feature. The objective of our model is to maximize the entropy of the sampled visual features. In order to optimize energy consumption, the limit amount of energy of the system is re-distributed amongst features according to their Incremental Coding Length. By selecting features with large coding length increments, the computational system can achieve attention selectivity in both static and dynamic scenes. We demonstrate that the proposed model achieves superior accuracy in comparison to mainstream approaches in static saliency map generation. Moreover, we also show that our model captures several less-reported dynamic visual search behaviors, such as attentional swing and inhibition of return.

Spotlight presentation, Wednesday, 9:50am.

W78 *Offline Handwriting Recognition with Multidimensional Recurrent Neural Networks*

ALEX GRAVES alex.graves@gmail.com
TU Munich, Garching bei München, Germany
JUERGEN SCHMIDHUBER juergen@idsia.ch
IDSIA, Manno-Lugano, Switzerland, and TU Munich, Garching bei München, Germany

Offline handwriting recognition—the transcription of images of handwritten text—is an interesting task, in that it combines computer vision with sequence learning. In most systems the two elements are handled separately, with sophisticated preprocessing techniques used to extract the image features and sequential models such as HMMs used to provide the transcriptions. By combining two recent innovations in neural networks—multidimensional recurrent neural networks and connectionist temporal classification—this paper introduces a globally trained offline handwriting recogniser that takes raw pixel data as input. Unlike competing systems, it does not require any alphabet specific preprocessing, and can therefore be used unchanged for any language. Evidence of its generality and power is provided by data from a recent international Arabic recognition competition, where it outperformed all entries (91.4% accuracy compared to 87.2% for the competition winner) despite the fact that neither author understands a word of Arabic.

Spotlight presentation, Wednesday, 11:50am.

W79 *Nonparametric sparse hierarchical models describe V1 fMRI responses to natural images*

Pradeep Ravikumar pradeepr@stat.berkeley.edu
Vincent Q Vu vqv@stat.berkeley.edu
Bin Yu binyu@stat.berkeley.edu
Thomas Naselaris tnaselar@berkeley.edu
Kendrick Kay kendrick@berkeley.edu
Jack Gallant gallant@berkeley.edu
University of California, Berkeley, CA, United States

We propose a novel hierarchical, nonlinear model that predicts brain activity in area V1 evoked by natural images. In the study reported here brain activity was measured by means of functional magnetic resonance imaging (fMRI), a noninvasive technique that provides an indirect measure of neural activity pooled over a small volume ($\sim$ 2mm cube) of brain tissue. Our model, which we call the SpAM V1 model, is based on the reasonable assumption that fMRI measurements reflect the (possibly nonlinearly) pooled, rectified output of a large population of simple and complex cells in V1. It has a hierarchical filtering stage that consists of three layers: model simple cells, model complex cells, and a third layer in which the complex cells are linearly pooled (called "pooled-complex" cells). The pooling stage then obtains the measured fMRI signals as a sparse additive model (SpAM) in which a sparse nonparametric (nonlinear) combination of model complex cell and model pooled-complex cell outputs are summed. Our results show that the SpAM V1 model predicts fMRI responses evoked by natural images better than a benchmark model that only provides linear pooling of model complex cells. Furthermore, the spatial receptive fields, frequency tuning and orientation tuning curves of the SpAM V1 model estimated for each voxel appears to be consistent with the known properties of V1, and with previous analyses of this data set. A visualization procedure applied to the SpAM V1 model shows that most of the nonlinear pooling consists of simple compressive or saturating nonlinearities.

Spotlight presentation, Wednesday, 9:50am.

W80 *Reducing statistical dependencies in natural signals using radial Gaussianization*

Siwei Lyu, University at Albany, SUNY, and Eero Simoncelli, New York University.
Oral presentation, Thursday, 9:50am. *See abstract, page 194.*

Demonstrations (7:30pm–12:00am)

1b *Visualizing NIPS Cooperations using Multiple Maps t-SNE*

Laurens van der Maaten l.vandermaaten@micc.unimaas.nl
Maastricht University, MICC-IKAT, Maastricht, Netherlands
Geoffrey Hinton hinton@cs.toronto.edu
Department of Computer Science, University of Toronto, Toronto, Ontario, Canada

Our demonstration shows visualizations of NIPS co-authorships that were constructed by the multiple maps version of t-SNE. We gathered a dataset of the co-authorships in all NIPS papers (volume 1-20). The co-authorships are represented in a square matrix in which each row (and column) corresponds to a single author. The element $(i,\ j)$ in the matrix represents the number of papers that author i and author j wrote together. After normalization, this matrix may form the input into a multidimensional scaling technique, such as t-SNE (L.J.P. van der Maaten and G.E. Hinton. *Visualizing Data using t-SNE.* Journal of Machine Learning Research, 2008), to construct a two-dimensional map that visualizes the pairwise similarities between the authors. Unfortunately, it is impossible for multidimensional scaling techniques to construct an appropriate visualization of the similarity data, because of the following fundamental problem: if author A wrote papers together with author B and author B wrote papers together with author C, but author C never wrote papers with author A, the similarity relations cannot be modeled in a two-dimensional metric map due to the triangle inequality: in the map, author A is then modeled close to author B, and author B is modeled close to author C, as a result of which A is close to C as well, which is wrong. In order to alleviate this problem, we developed a multiple maps version of t-SNE that creates a collection of multiple maps (based on ideas proposed in J.A. Cook, I. Sutskever, A. Mnih, and G.E. Hinton. *Visualizing similarity data with a mixture of maps.* Proceedings of AI*STATS-07, 2007). Each author has a copy in each of the maps, and each copy is weighted by a mixing proportion (the mixing proportions for a single author over all maps sum up to 1). The multiple maps version of t-SNE can perfectly model the example above, and as a result, it is very good at visualizing NIPS co-authorship data. We ran the multiple maps version of t-SNE on the NIPS dataset, and constructed a visualization tool that shows the resulting maps in a visually appealing way. The visualization clearly shows the research clusters within the NIPS community, and reveals cooperations between these clusters (or people that have moved from one research cluster to another during their career). The visualization tool allows the user to quickly switch between maps, zoom and scroll in maps, search for authors, find out if authors are modeled in other maps as well, lay over a cooperation graph over the maps, etc.

2b *Real-time Topology Learning*

Kynan Eng kynan@ini.phys.ethz.ch
Tobi Delbruck tobi@ini.ethz.ch
Matthias Schrag mschrag@gmx.net
Martin Boerlin martin_boerlin@bluewin.ch
Institute of Neuroinformatics, ETH, Zurich, Switzerland

http://www.ini.ethz.ch/~kynan/projects/topology

The demonstration shows the first implementation (in any technology) of a real-time system for learning the spatial topological arrangements of cellular elements using an STDP-based unsupervised learning rule. Given a stream of address-event labelled inputs from the cellular elements, the algorithm learns their adjacency and automatically detects edge elements within a short time. The demonstration graphically shows the progress of learning during user interaction. Related paper: Boerlin, M and Delbruck, T and Eng, K (2008). *Getting to Know Your Neighbors: Unsupervised Learning of Topography from Real-World*, Event-Based Input. Neural Computation.

3b *Infer.NET: Software for Graphical Models*

Tom Minka minka@microsoft.com
John Winn jwinn@microsoft.com
John Guiver joguiver@microsoft.com
Microsoft Research, Machine Learning and Perception Group, Redmond, WA, United States
Anitha Kannan ankannan@microsoft.com
Microsoft Research, Search Labs, Mountain View, CA, United States

http://research.microsoft.com/mlp/ml/Infer/Infer.htm

This is a demo of Infer.NET which is a .NET library for inference in graphical models. Unlike black-box machine learning packages, Infer.NET allows you to piece together complex probabilistic models via a straightforward programming interface. Compared to existing inference tools like BUGS and VIBES, Infer.NET is unique in that it can perform either Expectation Propagation or Variational Message Passing on models with discrete and continuous variables. It supports arbitrary mixture models via factor graphs with gates. It is also uniquely structured as a compiler, taking a .NET program with random variables as input and producing a specially-optimized inference program on output.

4b *RL-Glue: From Grid Worlds to Sensor Rich Robots*

Brian Tanner btanner@cs.ualberta.ca
Adam White awhite@cs.ualberta.ca
Richard S. Sutton sutton@cs.ualberta.ca
University of Alberta, Computing Science, Edmonton, Alberta, Canada

http://glue.rl-community.org/nips-demo-2008

We propose to demonstrate a suite of software and robotics projects for reinforcement learning (RL) that have been in development for several years in Dr. Richard Sutton's group at the University of Alberta. Specifically, the three projects that we intend to showcase are: RL-Glue interface, the CritterBot robotic platform, and RL-Viz experimentation platform. The demonstration will illustrate how these projects allow researchers to develop learning agents that can be evaluated in a graphical simulation (RL-Viz) and on a mobile robot (CritterBot). RL-Glue is a language- and platform-independent protocol for evaluating reinforcement learning agents with environment programs. RL-Glue separates the agent- and environment-development process so that each can be written in different languages and even executed over the Internet from different computers. RL-Glue have had a significant influence on the way empirical comparisons are done in reinforcement learning. RL-Glue has been used to evaluate agents in four international competitions at high profile machine learning conferences. The most recent competition, held in conjunction with ICML 08, attracted over 150 teams. The final test phase of the competition included over 20 teams comprised of more than 40 participants. RL-Glue has been used by several university instructors, in several countries, to teach reinforcement learning. Several researchers have used RL-Glue to benchmark their agents in papers published in top international conferences, including NIPS. The CritterBot is an ongoing project at the University of Alberta whose goal is to add a further robotics effort to challenge, direct, and inspire the research on grounded artificial intelligence. This robot is small and mobile and outfitted with an unusually rich set of sensors, including sensors for touch, acceleration, motion, sound, vision, and several kinds of proximity. The initial objective is for the robot to form an extended multi-level model of the relationships among its sensors and between its sensors and its actuators. We have proposed that higher-level knowledge can be grounded in raw data of sensations and actions; this robotic platform will challenge and inspire us to see if it can really be done. We also plan to use this platform as a test case for rapid learning and for the use of reinforcement learning by non-experts. We would like a person whose has no training to be able to teach the system new ways of behaving in an intuitive manner much as one might train a particularly cooperative dog. Learning agents can interact with the CritterBot through RL-Glue just like with any other RL-Glue environment. RL-Viz provides the reinforcement learning community for the first time ever with a flexible, general, standardized, cross language and cross platform protocol/framework for managing and visualizing the interaction between agents and environments in reinforcement learning experiments. The RL-Viz project includes several state-of-the-art tasks used in learning research including, Tetris, a remote-controlled helicopter simulator provided by Andrew Ng's team at Stanford, keep-away soccer and a real-time strategy engine. RL-Viz is a protocol and library layered on top of RL-Glue. RL-Viz supports advanced features such as visualization of environments and agents and run-time loading of agents and environments. The software for most recent RL competition (mentioned above) was based on RL-Viz. We will present the latest developments in the RL-Glue project and demonstrate how RL-Glue provides a novel, unified architecture for developing reinforcement learning algorithms for simulation and physical experiments. This framework makes it easier to compare the performance of agents in a variety of simulated and physical tasks.

5b *The Chiara: A New Robotic Platform for Education and Research*

David S. Touretzky dst@cs.cmu.edu
Computer Science Department, Carnegie Mellon University, Pittsburgh, PA, United States
Ethan J. Tira-Thompson ejt@andrew.cmu.edu
Gaku Sato gsato@alumni.cmu.edu
Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, United States
David P. Rice dprice@andrew.cmu.edu
National Robotics Engineering Center, Carnegie Mellon University, Pittsburgh, PA, United States

http://www.chiara-robot.com

The Chiara is a new robot designed for education and research, intended to fill the niche left by the Sony AIBO. It's a hexapod robot with an arm and gripper, a webcam on a pan/tilt mount, and a Pico-ITX processor (1GHz clock speed, 1GB RAM, and 80GB hard drive) that runs Ubuntu Linux. The robot is programmed in C++ using our Tekkotsu robotics software framework that includes an innovative "dual coding" vision system, a kinematics solver, and extensive teleoperation and monitoring tools. The Chiara was awarded the second place prize at the 17th Annual AAAI (Association for the Advancement of Artificial Intelligence) Mobile Robot Exhibition in Chicago in July 2008, and is featured in a two-page article in the November/December issue of Robot magazine. It is being commercialized by RoPro Design, Inc., of Beaver, PA, and should be available for purchase by the end of the year. However, the design is entirely open source, both hardware and software, so users will be free to modify the robot or even make their own if they wish.

6b *Machine Perception for Human Machine Interaction*

Paul Ruvolo paul@mplab.ucsd.edu
Marian Stewart Bartlett mbartlett@ucsd.edu
Nicholas Butko nick@mplab.ucsd.edu
Claudia Lainscsek clainscsek@uscd.edu
Gwen Littleworth gwen@mplab.ucsd.edu
Jacob Whitehill jake@mplab.ucsd.edu
Tingfan Wu ting@mplab.ucsd.edu
Javier Movellan movellan@mplab.ucsd.edu
Computer Science & Engineering, University of California, San Diego, CA, United States

We present four live demonstrations of state-of-the-art machine perception technologies for use in real-time interaction between humans and machines. (1) Computer Expression Recognition Toolbox (CERT). CERT is the first system for fully automated coding of the Facial Action Coding System (FACS). FACS is a method from experimental psychology which decomposes facial expressions into 46 component movements, and enables investigations of new relationships between facial movement and internal state. Machine learning applied to the outputs of CERT has been shown to differentiate fake from real pain expressions with greater accuracy than naive human subjects, and to detect driver fatigue by predicting a crash 60-seconds prior in driving simulators. (2) Real-time detection of auditory moods. Detection of auditory phenomena (e.g. the emotion in a speaker's voice, the presence of laughing or music) provides robots with a valuable tool for inferring the current social mood. The ability to sense these dynamics in real-time opens the door for robots that interact more seamlessly with humans. Our approach adapts state of the art object recognition algorithms to the task of auditory category recognition, achieving excellent performance at little computational cost. (3) Real Time Infomax approach to visual saliency. While there has been an explosion in recent years in modeling human visual attention in task-free conditions, all or nearly all existing models are inappropriate for use in robots because they rely on complicated, slow calculations. We will demonstrate a real time active camera that orients itself towards the most informative regions of the visual scene. The approach is based on a Bayesian model of visual saliency that models well human eye movements in open-ended tasks. The approach runs efficiently (100FPS) on a modern low-end computer. (4) Real-time facial expression analysis for automated tutoring systems. Using the CERT toolbox (1) we investigate machine perception techniques to estimate a subject's perceived level of difficulty and his/her desired playback speed of a video lecture on a second-by-second basis. The demo we showcase is capable of running in real-time and automatically modulating the speed of a video lecture to suit the subject's current grasp of the material being presented.

Thursday, December 11th

Oral Session (8:30am–10:10am):
"Nonparametric Processes, Scene Processing, and Image Statistics"

8:30am *The Mondrian Process*

DANIEL M ROY droy@mit.edu
Computer Science and Artificial Intelligence Laboratory, MIT, Cambridge, MA, USA
YEE WHYE TEH ywteh@gatsby.ucl.ac.uk
Gatsby Computational Neuroscience Unit, UCL, London, United Kingdom

We describe a novel stochastic process called the Mondrian process and use it to define multidimensional and hierarchical generalizations of Sethuraman's stick-breaking construction of the Dirichlet process. We apply the Mondrian process as a nonparametric prior for relational data modeling by appealing to the Aldous-Hoover representation for infinitely and partially exchangeable arrays.

8:50am *Shared Segmentation of Natural Scenes Using Dependent Pitman-Yor Processes*

ERIK SUDDERTH sudderth@eecs.berkeley.edu
MICHAEL JORDAN jordan@cs.berkeley.edu
University of California, Berkeley, Berkeley, CA, United States

We develop a statistical framework for the simultaneous, unsupervised segmentation and discovery of visual object categories from image databases. Examining a large set of manually segmented scenes, we show that object frequencies and segment sizes both follow power law distributions, which are well modeled by the Pitman-Yor (PY) process. This nonparametric prior distribution leads to learning algorithms which discover an unknown set of objects, and segmentation methods which automatically adapt their resolution to each image. Generalizing previous applications of PY processes, we use Gaussian processes to discover spatially contiguous segments which respect image boundaries. Using a novel family of variational approximations, our approach produces segmentations which compare favorably to state-of-the-art methods, while simultaneously discovering categories shared among natural scenes.

9:10am *Cascaded Classification Models: Combining Models for Holistic Scene Understanding*

Geremy Heitz gaheitz@stanford.edu
Stephen Gould sgould@stanford.edu
Ashutosh Saxena asaxena@stanford.edu
Daphne Koller koller@cs.stanford.edu
Stanford University, Stanford, CA, United States

One of the original goals of computer vision was to fully understand a natural scene. This requires solving several problems simultaneously, including object detection, labeling of meaningful regions, and 3d reconstruction. While great progress has been made in tackling each of these problems in isolation, only recently have researchers again been considering the difficult task of assembling various methods to the mutual benefit of all. We consider learning a set of such classification models in such a way that they both solve their own problem and help each other. We develop a framework known as Cascaded Classification Models (CCM), where repeated instantiations of these classifiers are coupled by their input/output variables in a cascade that improves performance at each level. Our method requires only a limited "black box" interface with the models, allowing us to use very sophisticated, state-of-the-art classifiers without having to look under the hood. We demonstrate the effectiveness of our method on a large set of natural images by combining the subtasks of scene categorization, object detection, multiclass image segmentation, and 3d scene reconstruction.

9:30am *Multi-Level Active Prediction of Useful Image Annotations for Recognition*

Sudheendra Vijayanarasimhan svnaras@gmail.com
Kristen Grauman grauman@cs.utexas.edu
University of Texas at Austin, Austin, TX, United States

We introduce a framework for actively learning visual categories from a mixture of weakly and strongly labeled image examples. We propose to allow the category-learner to strategically choose what annotations it receives—based on both the expected reduction in uncertainty as well as the relative costs of obtaining each annotation. We construct a multiple-instance discriminative classifier based on the initial training data. Then all remaining unlabeled and weakly labeled examples are surveyed to actively determine which annotation ought to be requested next. After each request, the current classifier is incrementally updated. Unlike previous work, our approach accounts for the fact that the optimal use of manual annotation may call for a combination of labels at multiple levels of granularity (e.g., a full segmentation on some images and a present/absent flag on others). As a result, it is possible to learn more accurate category models with a lower total expenditure of manual annotation effort.

9:50am *Reducing statistical dependencies in natural signals using radial Gaussianization*

Siwei Lyu lsw@cs.albany.edu
Computer Science Department, University at Albany, SUNY
Eero Simoncelli eero@cns.nyu.edu
Howard Hughes Medical Institute, Center for Neural Science, Courant Institute of Mathematical Sciences, New York University

We consider the problem of efficiently encoding a signal by transforming it to a new representation whose components are statistically independent. A widely studied linear solution, independent components analysis (ICA), exists for the case when the signal is generated as a linear transformation of independent non- Gaussian sources. Here, we examine a complementary case, in which the source is non-Gaussian but elliptically symmetric. In this case, no linear transform suffices to properly decompose the signal into independent components, but we show that a simple nonlinear transformation, which we call radial Gaussianization (RG), is able to remove all dependencies. We then demonstrate this methodology in the context of natural signal statistics. We first show that the joint distributions of bandpass filter responses, for both sound and images, are better described as elliptical than linearly transformed independent sources. Consistent with this, we demonstrate that the reduction in dependency achieved by applying RG to either pairs or blocks of bandpass filter responses is significantly greater than that achieved by PCA or ICA.

Oral Session (10:40am–12:00pm): "Attention and Mind"

10:40am *Load and Attentional Bayes*

Peter Dayan dayan@gatsby.ucl.ac.uk
Gatsby Computational Neuroscience Unit, University College London, London, United Kingdom

Selective attention is a most intensively studied psychological phenomenon, rife with theoretical suggestions and schisms. A critical idea is that of limited capacity, the allocation of which has produced half a century's worth of conflict about such phenomena as early and late selection. An influential resolution of this debate is based on the notion of perceptual load (Lavie, 2005, TICS, 9: 75), which suggests that low-load, easy tasks, because they underuse the total capacity of attention, mandatorily lead to the processing of stimuli that are irrelevant to the current attentional set; whereas high-load, difficult tasks grab all resources for themselves, leaving distractors high and dry. We argue that this theory presents a challenge to Bayesian theories of attention, and suggest an alternative, statistical, account of key supporting data.

11:00am Invited Talk: *Theory of Mind with fMRI*

Rebecca Saxe
Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, United States

Externally observable components of human actions carry only a tiny fraction of the information that matters. Human observers are vastly more interested in perceiving or inferring the mental states - the beliefs, desires and intentions - that lie behind the observable shell. If a person checks her watch, is she uncertain about the time, late for an appointment, or bored with the conversation? If a person shoots his friend on a hunting trip, did he intend revenge or just mistake his friend for a partridge? The mechanism people use to infer and reason about another person's states of mind is called a "Theory of Mind" (ToM). One of the most striking discoveries of recent human cognitive neuroscience is that there is a group of brain regions in human cortex that selectively and specifically underlie this mechanism. I will describe recent studies from my lab characterising the functional profile of one of these regions, the right temporo-parietal junction. The challenge for the future remains: to construct an adequate computational description of a neurally implemented mechanism, that could reason about another person's thoughts.

Rebecca Saxe studied Psychology and Philosophy at Oxford, and then received a PhD in Cognitive Science from MIT, working under the supervision of Nancy Kanwisher. For the next three years, she was a Junior Fellow at Harvard University, moonlighting in the developmental psychology lab of Susan Carey. Since 2006 she has been an assistant professor of Cognitive Neuroscience, back at MIT. Her work investigates the development and neural basis of human social cognition (saxelab.mit.edu).

Mini Symposia (1:30pm–4:30pm)

MS1 *Algebraic methods in machine learning*

Risi Kondor `risi@gatsby.ucl.ac.uk`
Gatsby Computational Neuroscience Unit, UCL, London, United Kingdom
Guy Lebanon `lebanon@cc.gatech.edu`
Georgia Institute of Technology, Atlanta, GA, United States
Jason Morton `jason@math.stanford.edu`
Stanford University, Stanford, CA, United States

There has recently been a surge of interest in algebraic methods in machine learning. This includes new approaches to ranking problems, the budding field of algebraic statistics and various applications of non-commutative Fourier transforms. The aim of the workshop is to bring together these distinct communities, explore connections, and showcase algebraic methods to the machine learning community at large. The symposium is intended to be accessible to researchers with no prior exposure to abstract algebra. The program includes three short tutorials that will cover the basic concepts necessary for understanding cutting edge research in the field.

MS2 *Computational Photography*

Bill Freeman wtf@mit.edu
Artificial Intelligence Lab, MIT, Cambridge, MA, United States
Bernhard Schölkopf bs@tuebingen.mpg.de
MPI for Biological Cybernetics, Tübingen, Germany

Computation will change photography. The sensor no longer has to record the final image, but only data that can lead to the final image. Computation can solve longstanding photographic problems (e.g. deblurring) and well as open the door for radical new designs and capabilities for image capture, processing, and viewing (e.g. lightfield cameras). Many of these possibilities other great machine learning problems, and much of the progress in computational photography will rely on solutions to these challenging machine learning problems. We have gathered 5 leading researchers in this new field to describe their work at the intersection of photography and machine learning.

MS3 *Machine Learning in Computational Biology*

Gal Chechik gal@ai.stanford.edu
Google, Mountain View, CA, United States
Christina Leslie cleslie@cbio.mskcc.org
Memorial Sloan-Kettering Cancer Center, New York, NY, United States
Quaid Morris quaid.morris@utoronto.ca
Banting and Best Department of Medical Research,
University of Toronto, Toronto, ON, Canada
William Noble noble@gs.washington.edu
Department of Genome Sciences, University of Washington, Seattle, WA,United States
Gunnar Raetsch Gunnar.Raetsch@tuebingen.mpg.de
Friedrich Miescher Laboratory, Max Planck Society, Tübingen, Germany

The field of computational biology has seen dramatic growth over the past few years, both in terms of new available data, new scientific questions, and new challenges for learning and inference. In particular, biological data is often relationally structured and highly diverse, well-suited to approaches that combine multiple weak evidence from heterogeneous sources. These data may include sequenced genomes of a variety of organisms, gene expression data from multiple technologies, protein expression data, protein sequence and 3D structural data, protein interactions, gene ontology and pathway databases, genetic variation data (such as SNPs), and an enormous amount of textual data in the biological and medical literature. New types of scientific and clinical problems require the development of novel supervised and unsupervised learning methods that can use these growing resources. The goal of this workshop is to present emerging problems and machine learning techniques in computational biology.

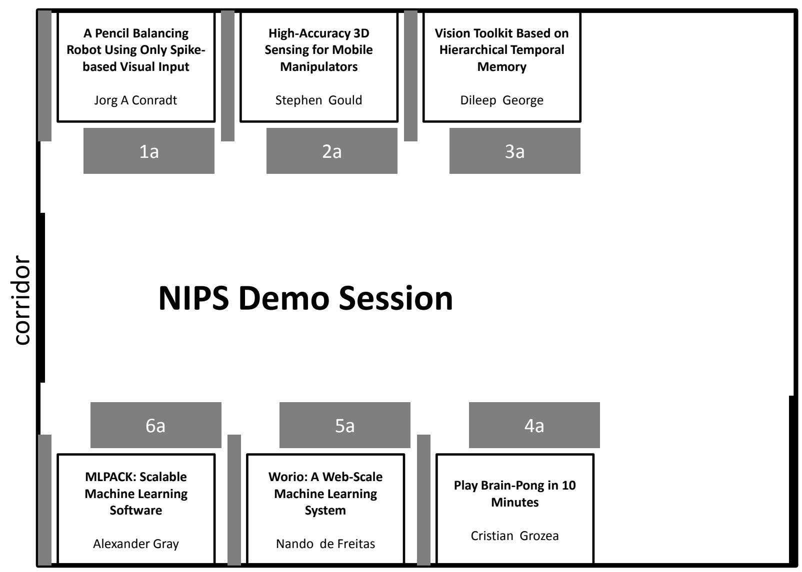
MS4 *Principled Theoretical Frameworks for the Perception-Action Cycle*

Daniel Polani d.polani@herts.ac.uk
School of Computer Science, University of Hertfordshire, Hatfield, United Kingdom
Naftali Tishby tishby@cs.huji.ac.il
Institute of Computer Science and Center for Neural Computation,
The Hebrew University, Jerusalem, Israel

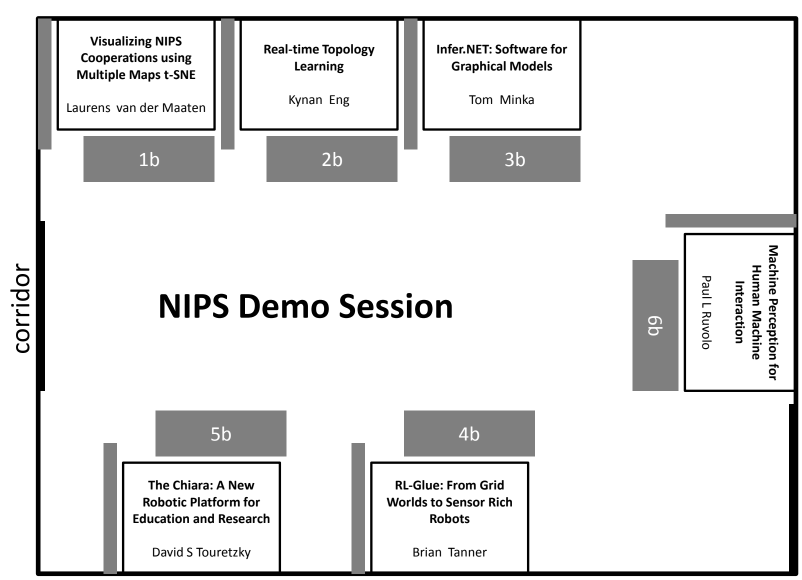
A significant emphasis in trying to achieve adaptation and learning in the perception-action cycle of agents lies in the development of suitable algorithms. While partly these algorithms result from mathematical constructions, in modern research much attention is given to methods that mimic biological processes. However, mimicking the apparent features of what appears to be a biologically relevant mechanism makes it difficult to separate the essentials of adaptation and learning from accidents of evolution. This is a challenge both for the understanding of biological systems as well as for the design of artificial ones. Therefore, recent work is increasingly concentrating on identifying general principles rather than individual mechanisms for biologically relevant information processing. One advantage is that a small selection of principles can give rise to a variety of - effectively equivalent - mechanisms. The ultimate goal is to attain a more transparent and unified view on the phenomena in question. Possible candidates for such principles governing the dynamics of the perception-action cycle include but are not limited to information theory, Bayesian models, energy-based concepts or group-theoretical principles.

Poster and Demonstrations Floor Plans

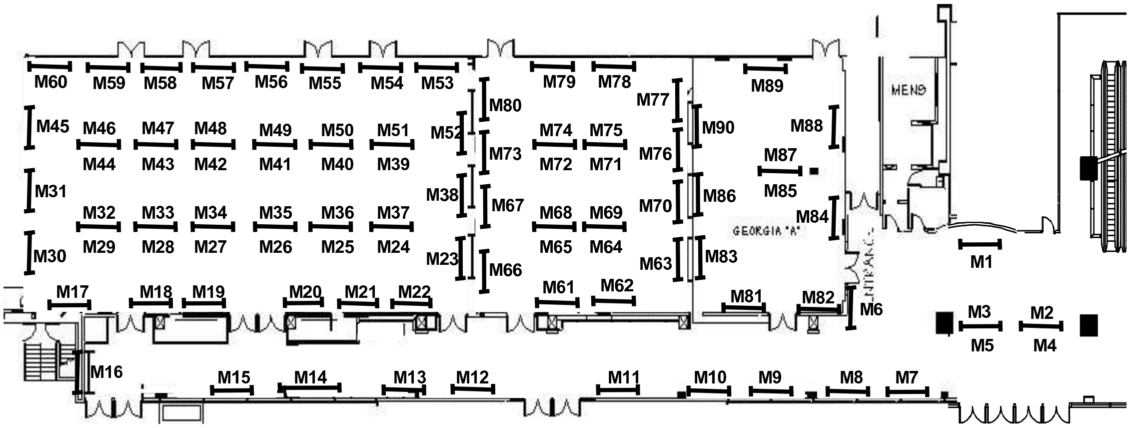
Tuesday Demonstrations



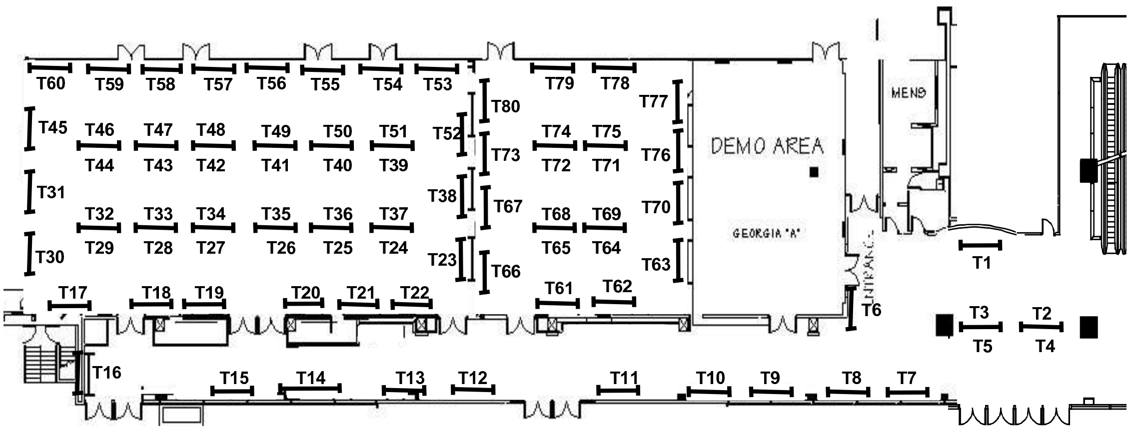
Wednesday Demonstrations



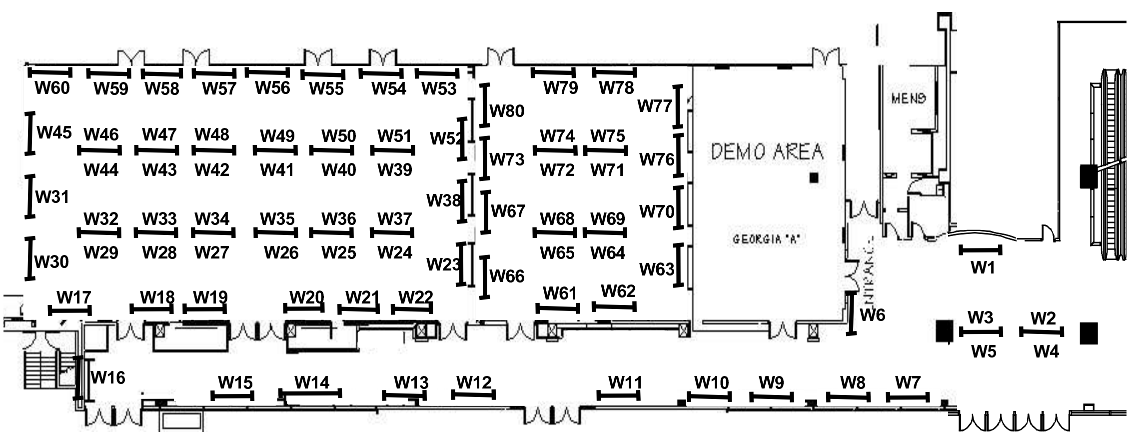
Monday Posters



Tuesday Posters



Wednesday Posters



Frequently Asked Questions

MEALS:

Are any meals covered in the registration price?
The following meals are covered in the registration fee:

At the Tutorials and Conference:

Breakfast	Monday	8:00 - 9:30 AM	Regency A, Level 3
Opening Banquet	Monday	6:30 - 8:30 PM	Regency Ballroom
Breakfast	Tuesday - Thursday	7:30 - 9:00 AM	Perspectives, 34th Floor

At the Workshops:

Breakfast	Friday, Saturday	6:30 - 8:00 AM	Emerald Ballroom
Closing Banquet	Saturday	7:00 - 11:00 PM	Emerald Ballroom

Note: There is a $45 charge to bring a family member or friend to the banquet at the Hyatt Regency Vancouver and $40 charge at the Westin Resort & Spa. You may purchase these tickets at the NIPS Registration Desk in Vancouver or Whistler. Children are not allowed to attend the banquet.

LOST AND FOUND
Is there a lost-and-found area at the NIPS Conference?
Yes. This is located at the NIPS registration desk.

BUSINESSES AND SERVICES IN THE AREA
Which businesses and services are available close to the conference area?
There is a photocopier in the business center in the Hyatt lobby downstairs, across from the front desk. There is a Staples store, a drugstore, and a food court in the shopping center on the first floor of the Hyatt, accessible through the lobby.

Buses and Transportation

BUS TICKETS:
I have purchased a bus ticket, but I cannot choose which bus I want to be on.
Visit https://nips.cc/bus to assign yourself to a bus.

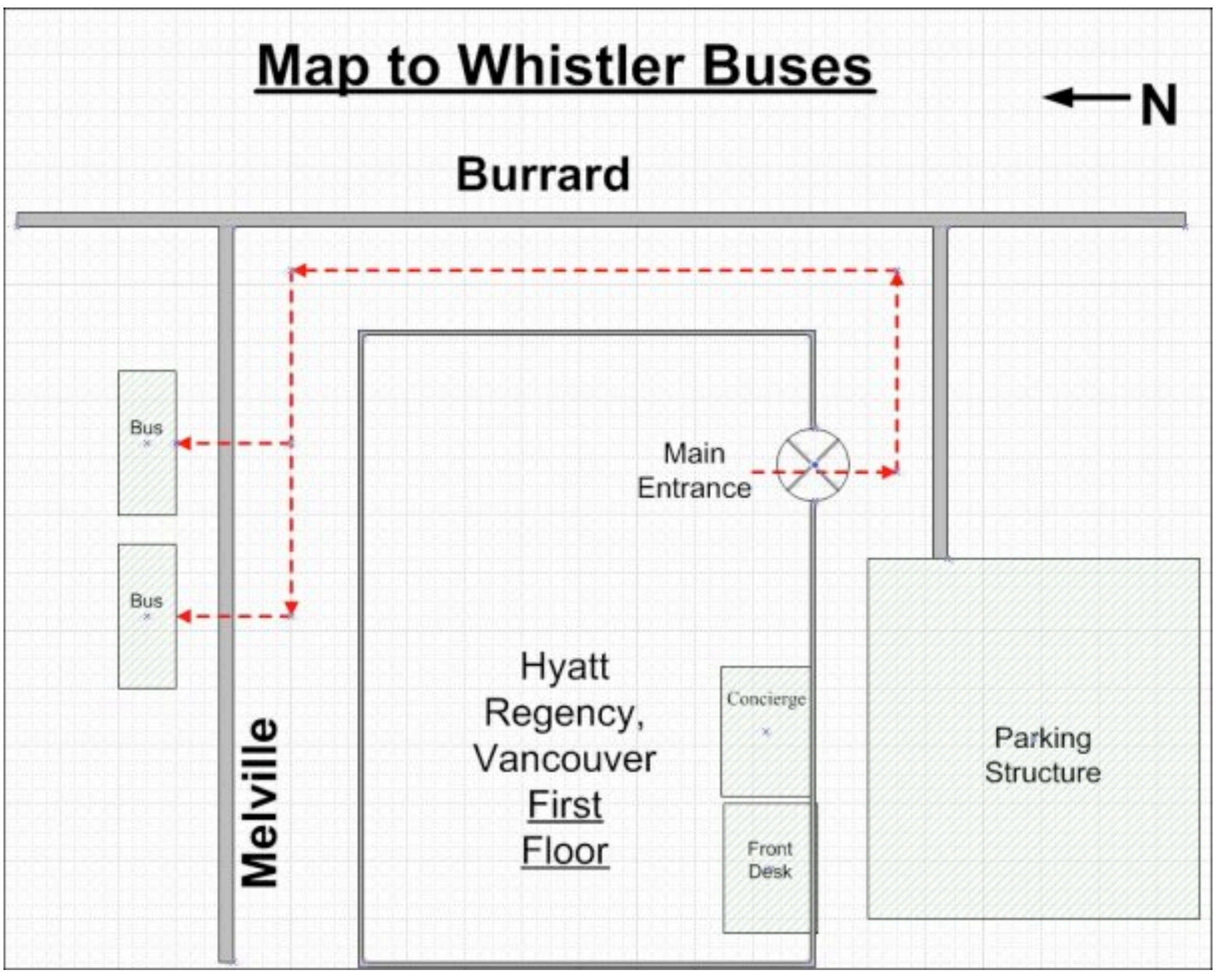
Does my bus ticket from Vancouver to Whistler include a ride back to Vancouver after the conference?
No. Each participant must make individual arrangements for the return from Whistler to the Vancouver International Airport.

Buses from Hyatt Regency Vancouver to Whistler

Buses will be available to transport Conference participants from the Hyatt Regency Vancouver to Whistler on Thursday, December 11. If you are planning to travel to Whistler by this means, please make advance reservations on the NIPS registration website when you register (http://nips.cc/register) for the Conference and/or Workshops. If you have already registered, please revisit your account online (http://nips.cc/register) and reserve a seat (http://nips.cc/bus). Spaces fill up quickly so register early so you will be to get on the bus of your choice. It will be possible to register for the bus to Whistler until 2pm on Wednesday, December10.

Buses will leave from the Hyatt Regency on Thursday, December 11, beginning at 2:00 pm and run until 6:00 pm. Buses will load 15 minutes prior to departure – please take this into account in your planning. You should consider picking up something to eat for the bus ride, you will arrive late in Whistler and dining options will be limited.

To reserve a bus seat go to : http://nips.cc/bus



NOTE: You must make your own arrangements for return transportation to Vancouver.

Each participant must make individual arrangements for the return from Whistler to the Vancouver International Airport. Reservations are required. Perimeter's Whistler Express picks up in the general area of the Westin Resort & Spa every 2 hours beginning at approximately 6:00 am and ending at approximately 6:00 pm. Your Perimeter reservation confirmation form will provide the exact time of your pick up. There is a special departure at 4:00 am on Sunday December 14, 2008 only.

The rate for NIPS 2008 participants is $50 CA (plus 6% GST, $5 fuel surchrge) (~$54 US). Remember to identify yourself as a participant of the NIPS Conference to receive the special group rate. Make reservations no later than 12:00 noon one day prior to departure time to reserve your space. Refundable if canceled one day prior to travel, there is a $15.00 cancellation fee.

Perimeter Phone Numbers:

US/Canada Toll Free	1-877-317-7788
Whistler	604-905-0041
Vancouver	604-266-538

For more information email Perimeter Reservations: res@perimeterbus.com
Reservations may also be made for Whistler Express bus service on the Perimeter website: http://www.perimeterbus.com/. Use the group code: NIPS 2008.

Driving Instructions from Vancouver to Whistler

Hyatt Regency Vancouver to Whistler: (Driving time: ~2-2.5 hours) From the Hyatt Regency on Burrard, head South on Burrard and turn right (West) on Georgia and follow it through Stanley Park and over the Lions Gate Bridge. Exit west off/under the bridge and into West Vancouver. Turn right on Taylor Way at Park Royal Shopping Centre. Follow Taylor Way half-a-mile up the hill to the Highway 1 (Trans-Canada) overpass. Join Highway 1 Westbound until you reach the junction with Highway 99 (Sea-to-Sky Highway). Exit right onto Highway 99. Follow this route for just over 100 kilometers along scenic Howe Sound, past Squamish to Whistler.

Ski Info:

Please note that pre-booking rates will only be available until December 11, 2008.

1. Call 1-800-766-0449 ; Our Reservation Center is open from 8:00 am –6:00 pm PST until November 16th, and from 7:00 am –7:00 pm between November 15th and December 11th, 2008.
2. Identify the name of your group to the Reservation agent.
3. Quote this **Discount Code#: 10646**

At the NIPS Workshop Mobile Unit, each participant will receive a Whistler Blackcomb Discount Card, which will entitle you to 10% off Retail purchases in Whistler Blackcomb shops and 15% off of food at GLC, Dusty’s, and Merlin’s.

Mobile Unit Hours (Westin Lobby)

Thursday, December 11	7pm - 10pm
Friday, December 12	8am – 9:30am

Hotel Contact Information:

Hyatt Regency
655 Burrard Street
Vancouver, British Columbia
Canada V6C 2R7
1-604-683-1234

The Hilton Whistler Resort & Spa
4050 Whistler Way
Whistler, British Columbia
Canada V0N 1B4
1-800-515-4050

Westin Whistler Resort and Spa
4090 Whistler Way
Whistler, British Columbia
Canada V0N 1B4
1-888-634-5577

Wireless Networking at the NIPS Conference

Wireless networking will be available in Vancouver on level three (Regency) where the Tutorial and main Conference are held and on level two (Plaza) where the Poster and Demonstration Sessions are located.

Wireless networking will be available in Whistler in all Workshop areas and guest rooms. You will need to supply your own 802.11b/g (a.k.a. Airport or WiFi) wireless network adapter for your laptop or PDA - unfortunately the popularity of this service precludes us supplying either wireless adapters or individual technical support for the wireless network.

Network Configuration

Your computer must be configured to use DHCP ("Obtain an IP automatically") with the following wireless network settings:

Hyatt Regency in Vancouver (Tutorials, Conference, Poster and Demonstrations):

- Wireless Network Name (SSID): NIPS
- Wireless Equivalent Privacy (WEP): disabled

Westin in Whistler (Workshops):

- Wireless Network Name (SSID): datavalet-Westin
- Wireless Equivalent Privacy (WEP): disabled

Hilton in Whistler (Workshops):

- Wireless Network Name (SSID): Wayport_Access
- Wireless Equivalent Privacy (WEP): disabled

Outbound email (SMTP) server: mail.nips.cc

RESTAURANT SUGGESTIONS

Listed below are some of Vancouver's best eateries. Please give us your comments and reviews, so we can keep updating our list.

Key: **$$$** = Expensive **$$** = Moderate

FEATURE RESTAURANT

Located on the 2nd level of Hyatt Regency Vancouver

For Reservations: 604 639-4770

Mosaic Bar & Grille, Hyatt Regency Vancouver's signature restaurant, oversees the vibrant downtown core of the city. Truly cosmopolitan in appeal, Mosaic's menu offers you flavours from around the world that will undoubtedly delight your senses.

Westcoast

Aqua Riva *200 Granville Street* **$$$** 604 683-5599
Rotisserie wood-fired grilling enhances the natural flavours of fresh, local and meats and extensive wine lists. Great view of Burrard Inlet and the North Shore mountains.

Raincity Grill *1193 Denman Street* **$$$** 604 685-7337
Very creative West-Coast bistro in English Bay. A fabulous wine list, with an emphasis on the Pacific Northwest, plus a good selection of local products makes this restaurant popular amongst visitors and local alike.

Bishops Restaurant *2188 West 4th Avenue* **$$$** 604 738-2025
Cozy, popular, bistro-style restaurant. Clientele includes several former Prime Ministers of Canada. Excellent food, a carefully selected wine list and attentive service make it the critics and peoples choice for the best North American food in Vancouver. Approximately 10-15 minutes from the hotel.

LIFT Bar and Grill *333 Menchion Mews* **$$** 604 689-5438

Located on the seawall in Coal Harbour, just minutes away from the financial district, this dynamic and sophisticated oceanfront restaurant boasts stunning views of Stanley Park and the North Shore mountains.

FUEL Restaurant 1944 West 4th Avenue **$$$** 604-288-7905

Fuel Restaurant offers superb regional cuisine in an approachable and comfortable setting. Chef Robert Belcham's high-end regional fare utilizes cooking techniques from around the world that are suited to each dish. The menu changes often to be in sync with the seasons, showcasing expertly sourced ingredients from local farmers and purveyors. Wine is a key component of the dining experience at Fuel, featuring a collection of special vintages and producers chosen specifically to complement the menu.

Seafood

Blue Water Cafe *1095 Hamilton Street* **$$$** 604 688-8078

Housed in a brick and beam heritage warehouse conversion, the dynamic dining room provides views of the 'East meets West' kitchen and raw bar. Private wine room and heated patio. Popular with locals and out of towners.

Bridges *Granville Island* **$$** 604 687-4400

Located on Granville Island near the public market, overlooking False Creek. Fine dining upstairs and casual dining in the bistro downstairs. Nouvelle cuisine fish and game. Best views of the City and mountains.

The Fish House *8901 Stanley Park Drive* **$$$** 604 681-7275

Creative local chef and a very innovative menu. Lovely warm clubhouse atmosphere situated in Stanley Park.

Joe Fortes Seafood House *777 Thurlow Street* **$$** 604 669-1940

The favourite bar and grill of Vancouverites. A grand room in a bustling downtown location buzzing with personable, professional servers makes this a favourite spot with locals, visitors and celebrities. Good food presented in a lively atmosphere. Great oyster bar.

Salmon House on the Hill *2229 Folkstone Way, W.Van* **$$** 604 926-3212

Native Canadian style seafood. All tables have a beautiful view of Vancouver. Barbecued salmon over alder wood is the house specialty. Approx 25 minutes from the hotel.

Cardero's Restaurant *1583 Coal Harbour Quay* **$$** 604-669-7666

We are fish, chops & a wok. Casual, upscale West Coast Menu in a warm yet sophisticated atmosphere. Nestled amid the yachts and seaplanes of Coal Harbour Marina. Spectacular views of the local mountains and Burrard Inlet from the restaurant, pub and year-round patio.

Coast Restaurant *12457 Hamilton Street* **$$** 604-685-5010

Contemporary Zen-inspired décor, specialty lighting, selective use of reclaimed fir for the tabletops and abstract graphics combined with a 'wall of water' to create a comfortable and tranquil environment. A dramatic 17-foot, 2700 bottle wine display features a collection of local and international wines.

Goldfish Pacific Kitchen 1118 Mainland Street **$$** 604.689.8318

Located in Yaletown at the corner of Mainland and Helmcken, GoldFish Pacific Kitchen offers a fresh take on West Coast cooking. Focusing on seafood and drawing on flavours from the Pacific Rim, Executive Chef William Tse creates fresh, modern dishes designed for sharing. Guests enjoy a variety of wines from the West Coast, in addition to a selection of premium sakes and hand-crafted cocktails. The warm, vibrant room at GoldFish has ideal areas for accommodating both intimate dining experiences and large private events.

The Sandbar *1535 Johnston Street* **$$** 604-669-9030

The Sandbar Seafood on Granville Island- serves up Vancouver's freshest seafood in a warm, sophisticated atmosphere. Located in the heart of vibrant Granville Island overlooking False Creek. Fresh oyster bar, tapas, Hoshi's Sushi Bar, dinner and dancing. Vancouver's most popular spot to see and be seen.

Cannery *2205 Commissioner Street* **$$$** 604 254-9606

Located in the scenic Port of Vancouver, The Cannery Seafood Restaurant has been a prime destination for locals and visitors since 1971. The restaurant exemplifies the simple style of a traditional West Coast cannery with a mix of nautical artifacts that complement the breathtaking view of the harbour, North Shore Mountains and city skyline. Chef Frederic Couton offers seasonal menus, including a daily fresh sheet featuring one of the largest variety of fresh local and imported fish and seafood in Vancouver.

Continental

Mosaic Bar & Grille *Hyatt Regency Vancouver* **$$** 604 639-4770

Mosaic Bar & Grille, Hyatt Regency Vancouver's signature restaurant, oversees the vibrant downtown core of the city. Truly cosmopolitan in appeal, Mosaic's menu offers you flavours from around the world that will undoubtedly delight your senses.

The Sequoia Grill *Ferguson Point of Stanley Park* **$$** 604 669-3281

The drawing room and conservatory of this lovely garden home offers sunset views over Howe Sound. The drive to and from the restaurant through Stanley Park is an added bonus.

The Observatory *6400 Nancy Greene Way* **$$$** 604 984-0661

Nestled in a chalet atop Grouse Mountain 3500 ft above Vancouver, provides you a view from Bellingham to Vancouver Island. Complimenting the stunning surroundings is wonderful Continental cuisine and polished service. A great romantic destination.

Steak

Gotham Steakhouse *615 Seymour Street* **$$$** 604 605-8282

An American Style Steakhouse with stunning décor, superb service and an award winning wine list. Excellent, simple, well thought out menu with quality ingredients. USDA Prime Beef. Gotham is located in a beautifully restored heritage building, in the heart of the financial district and is walking distance from Vancouver's entertainment district.

Mortons *750 West Cordova Street* **$$$** 604 915-5105Morton's The Steakhouse, Vancouver is renowned for its signature menu and legendary hospitality. Now in its 22nd year, Morton's has set the standard for fine steakhouse dining.

Pinky's Steakhouse 1265 Hamilton Street **$$** 604.637.3135
Pinky's is an urban twist on the classic steakhouse with a menu built arond the highest quality Sterling Silver Beef, accompanied by signature "non-steak" entrees, snappy starters and yummy desserts.

Kobe Japanese Steak House *1042 Alberni Street* **$$** 604-684-2451
Canada's first Japanese Steak House offers an array of Fantastic Steaks, Succulent Seafood, Teriyaki Chicken and Garden Fresh Vegetables. Colourful Samurai style Chefs perform with flashing blades as you watch your dinner being cooked.

French/Belgian

Cafe de Paris *751 Denman Street* **$$** 604 687-1418
A French bistro that is a favourite spot with the locals, including other restaurateurs. Terrific Pepper Steak and Pommes Frites.

Le Crocodile *909 Burrard Street* **$$$** 604 669-4298
Very busy, but the ultimate in French bistro cuisine and service. Often the critics and peoples choice for the best French restaurant in the city.

Le Gavroche *1616 Alberni Street* **$$$** 604 685-3924
Located in a turn of the century house, this is an excellent choice for an intimate candlelight dinner. Good wine selection and reserve wines upon request.

L'Hermitage *1025 Robson Street* **$$$** 604 689-3237
Traditional French cuisine served in an intimate atmosphere by polished service staff. Outstanding wine list including some from the owner's own vineyard in France.

Chambar Belgian Restaurant 562 Beatty Street **$$$** 604-879-7119
Leave your day at the door & enjoy the essentials – food, service & great music. Experience stunning cocktails, an extensive Belgian beer list, and an interesting selection of affordable wines. Dessert is no mere afterthought – they are utterly delicious. Chambar - an unpretentious fling with fine dining.

Italian

Italian Kitchen *1037 Alberni Street* **$$-$$$** 604.687.2858
This modern two-level sophisticated Italian Restaurant located in the "trendy" area of downtown Vancouver has a street level entrance with a marbled terrace-like second floor with opening windows. The Italian Kitchen is a white table cloth restaurant offering excellent value.

Caffe De Medici *1025 Robson Street* **$$$** 604 669-9322
Northern Italian cuisine featuring veal, pasta & fresh seafood. Elegant surroundings. Centrally located in downtown Vancouver, just minutes away from all major hotels.

Cin Cin *1154 Robson Street* **$$** 604 688-7338

Very popular for its central location, reasonable prices, beautiful décor and a terrific wine selection. Excellent creative Mediterranean/Italian cuisine with a Northwest touch.

Cioppino's *1133 Hamilton Street* $$$ 604 688-7466
A cherry wood interior makes the dining room a warm setting for an imitate or lively occasion with French inspired Mediterranean signature dishes cooked to perfection in the open kitchen. An elegant wine room is available for private functions.

La Terrazza *1088 Cambie Street* **$$$** 604 899-4449

Warmth, comfort, authentic culinary flavours set in the ambience of a classical Italian villa come together at La Terrazza Restaurant. Experience fine dining with understated elegance. Wine lovers will be enticed by La Terrazza's unique Enoteca, a wine cellar and private dining room seating up to 22 people and suitable for corporate functions.

Il Giardino *1382 Hornby Street* **$$$** 604 669-2422

This Tuscan look alike Villa at the end of Hornby street hides an attractive jumble of four tile rooms and a vinedraped courtyard with a wood-burning oven. Try Umberto's famous 32-ounce Fiorentina for Two. Garden patio dining, private wine room, group functions welcome.

Chinese

Victoria Chinese Restaurant *Royal Centre Mall* **$$** 604 669-8383

Located in the Royal Centre Mall (adjacent to the Hyatt), Victoria Chinese Restaurant offers an excellent selection of Dim Sum and a variety of delicious Chinese dishes.

Imperial Chinese Seafood *355 Burrard Street* **$$$** 604-688-8191

Grandeur has its place and in Vancouver you can find it at the Imperial Chinese Seafood Restaurant, an elegant fine-dining restaurant established in 1989, renowned for providing the ultimate Cantonese dining experience.

Kirin Mandarin *1166 Alberni Street* **$$** 604 682-8833

Specializes in fresh seafood creations. Outstanding selection of traditional and unique Cantonese dishes in elegant ambience with downtown and mountain views. Freshly-made Dim-Sum served at Lunch.

Japanese

Tojo's *202 - 777 West Broadway* **$$$** 604 872-8050

Hidekazu Tojo prepares an amazing Japanese dining experience in this stylish ultramodern establishment. Tatami rooms are beautiful and North-facing windows offer downtown cityscapes. Critic's choice for the best Japanese food and best seafood in the city.

Yoshi *689 Denman Street* **$$$** 604 738-8226

Near Stanley Park and amazing Japanese restaurant in downtown. Good Sushi and 8 courses Japanese traditional dining.

West Coast/Fusion

glowbal grill & satay bar *1079 Mainland Street* **$$** 604 602-0835

Proven to be one of the finest restaurants in town - voted 'Best New Vancouver Restaurant 2002', People's Choice Award and Reader's Choice Award, Vancouver Magazine. Vancouver's hot mecca for food and fashion, this 100-seat room boasts an award winning Chef, sleek decor and an electric atmosphere.

Sanafir *1026 Granville Street* **$$** 604 678-1049

The 25-foot high ceiling, social seating, high energy kitchen and exotic lounge invite guests into a buzz of energy. The room has been designed to allow for a variety of dining options; one can enjoy comfortable and intimate dining in the main dining area on the first level or gather around the low-level tables or on Moroccan-style beds in the mezzanine level overlooking the restaurant.

Reviewers

Pieter Abbeel
Naoki Abe
Dimitris Achlioptas
Felix Agakov
Deepak Agarwal
Edoardo M. Airoldi
Shotaro Akaho
Cyril Allauzen
Pierre Alquier
Charles Anderson
Cédric Archambeau
Hideki Asoh
Chris Atkeson
Jean-Yves Audibert
Shai Avidan
Francis Bach
Drew Bagnell
Timothy Bailey
Goekhan Bakir
Maria-Florina Balcan
Arindam Banerjee
Aharon Bar Hillel
Ziv Bar-Joseph
Andreas Bartels
Andrew Barto
Sumit Basu
Mohsen Bayati
Sue Becker
Ulrik R. Beierholm
Mikhail Belkin
Asa Ben Hur
Shai Ben-David
Samy Bengio
Yoshua Bengio
Kristin Bennett
Alex Berg
Matthias Bethge
Jinbo Bi
Daniel M. Bikel
Mikhail Bilenko
Jeff A. Bilmes
Eric Blais
David Blei
John Blitzer
Karsten Borgwardt
Vivek Borkar
Leon Bottou
Matthew Botvinick
Guillaume Bouchard
Michael Bowling
Christoph Braun
Sébastien Bubeck
Joachim M. Buhmann
Christopher J. Burges
Daniel A. Butts
Stephane Canu
Andrea Caponnetto
Olivier Cappe
Francois Caron
Miguel A. Carreira-Perpinan
Xavier Carreras
Rui Castro
Gert Cauwenberghs
Gavin Cawley
Lawrence Cayton
A. Taylan Cemgil
Nicolò Cesa-Bianchi
Nicolas Chapados
Olivier Chapelle
Kamalika Chaudhuri
Gal Chechik
Li Cheng
Sonia Chernova
David Chiang
Silvia Chiappa
Andreas Christmann
Wei Chu
Stéphan J.M. Clémençon
Stephen Clark
Bertrand S. Clarke
Michael Collins
Vincent Conitzer
Antoine Cornuejols
Francesco Corona
Greg Corrado
Aaron Courville
Koby Crammer
Jacob Crandall
James Curran
James Cussens
Marco Cuturi
Florence D'Alche-Buc
Aaron D'Souza
Arnak Dalalyan
David Danks
Trevor Darrell
Anirban Dasgupta
Sanjoy Dasgupta

Denver Dash
Hal Daume III
Nathaniel D Daw
Peter Dayan
Tijl De Bie
Nando de Freitas
Virginia de Sa
Dennis DeCoste
Marc P. Deisenroth
Ofer Dekel
Olivier Delalleau
Li Deng
Inderjit Dhillon
Thomas Dietterich
Carlos Diuk
Adrian Dobra
Eizaburo Doi
Piotr Dollar
Arnaud Doucet
Kenji Doya
Mark Dredze
Petros Drineas
Jennifer Dy
Alexandre dÁspremont
Douglas Eck
Laurent El Ghaoui
Ran El-Yaniv
James Elder
Charles Elkan
Daniel P. W. Ellis
Yaakov Engel
Eleazar Eskin
Eyal Even-Dar
Amir M. Farahmand
Jason D. Farquhar
Li Fei-Fei
Jacob Feldman
Viktoria V Felmetsger
Rob Fergus
Alan Fern
Vittorio Ferrari
Sanja Fidler
Jenny Finkel
Jozsef Fiser
David Fleet
Francois Fleuret
Patrick Flynn
David Forsyth
Charless Fowlkes
Emily B Fox
Paolo Frasconi
Desobry Frederic
William Freeman
Nir Friedman
Karl J. Friston
Kenji Fukumizu
Thomas Gaertner
Aurélien Garivier
Peter V. Gehler
Claudio Gentile
Sebastian Gerwinn
Zoubin Ghahramani
Mohammad Ghavamzadeh
Ali Ghodsi
Ran Gilad-Bachrach
Andrew Gilpin
Mark Girolami
Amir Globerson
Jozien Goense
Jacob Goldberger
Geoffrey Gordon
Dilan Gorur
Thore Graepel
Yves Grandvalet
David Grangier
Kristen Grauman
Russell Greiner
Arthur Gretton
Jim Griffin
Thomas L. Griffiths
Moritz Grosse-Wentrup
Greg Grudic
Yuhong Guo
Maya Gupta
Patrick Haffner
Philipp Hafliger
Barbara Hammer
Zaid Harchaoui
David R. Hardoon
Stefan Harmeling
John Harris
Alexander J Hartemink
Andrea Hasenstaub
Elad Hazan
Matthias Hein
Katherine A Heller
Ralf Herbrich
Tomer Hertz
Aaron Hertzmann
Tom Heskes
Jeremy Hill
Geoffrey E. Hinton

Jesse Hoey
Heiko Hoffmann
Thomas Hofmann
Derek Hoiem
Eric Horvitz
Patrik O. Hoyer
Daniel J Hsu
Aapo Hyvarinen
Christian Igel
Alexander Ihler
Kazushi Ikeda
Shiro Ikeda
Alexander Ilin
Giacomo Indiveri
Masato Inoue
Michael Isard
Shin Ishii
Tommi Jaakkola
Laurent Jacob
Robert Jacobs
Herbert Jaeger
Michael James
Larsen Jan
Martin Jansche
Dominik Janzing
Tony Jebara
David D. Jensen
Michael B. Johanson
Mark Johnson
Nebojsa Jojic
Michael Jordan
Balázs Kégl
Toshihiro Kamishima
Takafumi Kanamori
Hyun Min Kang
Juha T. Karhunen
yan karklin
Hisashi Kashima
Samuel Kaski
Motoaki Kawanabe
Sathiya Keerthi
Charles Kemp
Kristian Kersting
Joseph Keshet
Sergey Kirshner
Jyrki Kivinen
Dan Klein
Robert Kleinberg
David C. Knill
Jens Kober
Vladimir Kolmogorov
Risi Kondor
Aryeh Kontorovich
Konrad Kording
Balaji Krishnapuram
Eyal Krupka
Alex Kulesza
Brian Kulis
Takio Kurita
Dima Kuzmin
James Kwok
John Lafferty
Michail G. Lagoudakis
Christoph H Lampert
Gert Lanckriet
Niels Landwehr
John Langford
Hugo Larochelle
Kate Larson
Pavel Laskov
Francois Laviolette
Neil D. Lawrence
Svetlana Lazebnik
Nicolas Le Roux
Guy Lebanon
Guillaume Lecue
Ann Lee
Daniel D. Lee
Michael Lee
Tai-Sing Lee
Wee Sun Lee
Robert Legenstein
Amaury Lendasse
Máté Lengyel
Vincent Lepetit
Christina Leslie
Roger Levy
Lihong Li
Percy S. Liang
Lin Liao
Hsuan-Tien Lin
Jennifer Listgarten
Ce Liu
Shih-Chii Liu
Karen Livescu
Daniel Lizotte
Phil Long
rosasco lorenzo
Elliot A Ludvig
Wolfgang Maass
Satyaki Mahalanabis
Jesus Malo

Gideon Mann
Christopher D Manning
Shie Mannor
Vikash Mansinghka
Yishay Mansour
Mario Marchand
Tim Marks
Lluis Marquez
Jiri Matas
Jon McAuliffe
Sam McClure
Josh McDermott
Ryan McDonald
H. Brendan McMahan
Marina Meila
Ron Meir
Roland Memisevic
Paul Merolla
Brian Milch
Bradley A. Minch
Vivienne L. Ming
Tom Minka
Michael Mistry
Daichi Mochihashi
Baback Moghaddam
Gianluigi Mongillo
Claire Monteleoni
Joris M. Mooij
Ray Mooney
Greg Mori
Jun Morimoto
Quaid Morris
Fabien Moutarde
Klaus-Robert Mueller
Sayan Mukherjee
Remi Munos
Noboru Murata
Kevin Murphy
Iain Murray
Ramesh Nallapati
Mukund Narasimhan
Dan Navarro
Gerhard Neumann
Jennifer Neville
Andrew Y. Ng
XuanLong Nguyen
Duy Nguyen-Tuong
Yael Niv
William Stafford Noble
David C. Noelle
Robert Nowak
Sebastian Nowozin
Ryan W. O'Donnell
Sean O'Rourke
Klaus Obermayer
Guillaume Obozinski
Uwe Ohler
Erkki Oja
Aude Oliva
Bruno Olshausen
Cheng Soon Ong
Takashi Onoda
Manfred Opper
Francesco Orabona
Christian Osendorfer
Simon Osindero
Tomoko Ozeki
David Page
Liam Paninski
Hyeyoung Park
Ronald Parr
Lucas C. Parra
Andrea Passerini
Hanna M. Pasula
Barak A. Pearlmutter
Dmitry Pechyony
Kristiaan Pelckmans
Marcello Pelillo
Jaakko Peltonen
Jiming Peng
Fernando Pereira
Francisco Pereira
Fernando Perez-Cruz
Amy Perfors
Jan Peters
Petra Philips
Joelle Pineau
John Platt
Jean Ponce
Massimiliano Pontil
Pascal Poupart
Doina Precup
Hubert Preissl
Yuan (Alan) Qi
Ariadna Quattoni
Joaquin Quiñonero Candela
Thomas Rückstiess
Gunnar Raetsch
Ali Rahimi
Tapani Raiko
Umesh Rajashekar
Alexander Rakhlin

Alain Rakotomamonjy
Liva Ralaivola
Deva Ramanan
Nathan Ratliff
Magnus Rattray
Pradeep Ravikumar
Stefan Riezler
Alessandro Rinaldo
Irina Rish
Brian Roark
Romer Rosales
David Rosenberg
David A. Ross
Michael G. Ross
Saharon Rosset
Afshin Rostamizadeh
Stefan Roth
Volker Roth
Yasser Roudi
Daniel M. Roy
Cynthia Rudin
Bryan Russell
Nicole Rust
Maneesh Sahani
Paul Sajda
Ruslan Salakhutdinov
Adam Sanborn
Ted Sandler
Sujay Sanghavi
Guido Sanguinetti
Craig Saunders
Bernhard Schölkopf
Stefan Schaal
Robert E. Schapire
Tobias Scheffer
Katya Scheinberg
Bernt Schiele
Juergen Schmidhuber
Jeff Schneider
Paul R. Schrater
Dale Schuurmans
Anton Schwaighofer
Clayton Scott
Michele Sebag
Matthias W. Seeger
Frank Sehnke
Walter Senn
Peggy Series
Thomas Serre
Rocco A. Servedio
Fei Sha
Gregory Shakhnarovich
Shai Shalev-Shwartz
Ohad Shamir
John Shawe-Taylor
Xiaotong Shen
Pradeep Shenoy
Bertram E. Shi
Jianbo Shi
Pannagadatta Shivaswamy
Lavi Shpigelman
Ilya Shpitser
Leonid Sigal
Ricardo Silva
David Silver
Ozgur Simsek
Vikas Sindhwani
Yoram Singer
Aarti Singh
Satinder Singh Baveja
Josef Sivic
Noam Slonim
Cristian Sminchisescu
Noah A. Smith
Alex J. Smola
Edward Snelson
Peter Sollich
Le Song
Finnegan Southey
Alessandro Sperduti
Peter Spirtes
Nathan Srebro
Karthik Sridharan
Bharath K. Sriperumbudur
Daniel Stefankovic
Ingo Steinwart
Susanne Still
Alan Stocker
Alexander Strehl
Nathan Sturtevant
Erik Sudderth
Masashi Sugiyama
Charles Sutton
Johan Suykens
Csaba Szepesvari
Prasad Tadepalli
Takashi Takenouchi
Junı́chi Takeuchi
Partha Pratim Talukdar
Erik Talvitie
Ameet S. Talwalkar
Marshall Tappen

Ben Taskar
Sekhar Tatikonda
Sharpee Tatyana
Matthew E. Taylor
Russ Tedrake
Yee Whye Teh
Joshua Tenenbaum
Gerald Tesauro
Evangelos Theodorou
Romain Thibaux
Jo-Anne Ting
Antonio Torralba
Mikhail Traskin
Volker Tresp
Bill Triggs
Ioannis Tsochantaridis
Koji Tsuda
Zhuowen Tu
Tinne Tuytelaars
Eiji Uchibe
Naonori Ueda
Raquel Urtasun
Giorgio Valentini
Antal van den Bosch
Patrick van der Smagt
Martijn van Otterlo
Manik Varma
Nuno Vasconcelos
Nicolas Vayatis
Jakob Verbeek
alessandro verri
Jean-Philippe Vert
Sethu Vijayakumar
Pascal Vincent
R. Jacob Vogelstein
Martin Wainwright
Christian Walder
Dirk B. Walther
Larry Wasserman
Kilian Q Weinberger
Eugene Weinstein
Yair Weiss
Max Welling
Wim Wiegerinck
Daniel Wierstra
Eric Wiewiora
Chris Williams
David Wingate
John Winn
Olivier Wintenberger
Ole Winther
Florentin Woergoetter
Frank Wood
Phil Woodland
Jennifer Wortman
Mingrui Wu
Abraham Wyner
Eric P. Xing
Linli Xu
Ming-Hsuan Yang
Chen Yanover
Scott Wen-tau Yih
Changwon Yoo
Byron Yu
Kai Yu
Shipeng Yu
Stella X. Yu
Mikhail Zaslavskiy
Richard Zemel
Hongyuan Zha
Tong Zhang
Lu Zhengdong
Ding-Xuan Zhou
Shuheng Zhou
Xiaojin Zhu
Alexander Zien
Martin Zinkevich
Onno Zoeter
Geoffrey Zweig

Subject Index

Author Index

Notes

Instructions for Authors

A lot of work needs to be done over a very short time between the arrival of the papers and the submission of the whole package. Please help us by reading and carefully following these instructions.

1. **Upload new paper/metadata:** To submit the camera-ready version of your paper, log into the confmaster web site `http://nips2008.confmaster.net`, click on "View Own Papers", click on the diskette with the red uparrow below it on the row that corresponds to your accepted paper, and upload your PostScript or PDF file. If the title or authors are different from those printed in this brochure, send an email to `nips@nec-labs.com` with the subject "Subject: NIPS paper *paperid* Title or Author Change", where *paperid* is the confmaster paper number.

   For document preparation instructions, see below. Remember to use the latest style files and to set them to generate a non-anonymous paper.

   All camera-ready papers and metadata must be uploaded to the website by 11:59PM Eastern Time, Friday, January 9, 2009.

2. **Author Agreement Form:** One author must complete and sign the author agreement form which appears on the last page of the NIPS program booklet (and is also available on the NIPS website). This year we are moving to electronic collection of these forms. Please scan the completed, signed form and convert it to PDF format. (Color is optional.) Email the PDF scan of the completed, signed form to `nips@nec-labs.com`. Use the subject "Subject: NIPS paper *paperid* Agreement Form", where *paperid* is the paper number assigned at papers.nips.cc. The signed, scanned author agreement form is due by Wednesday, January 7, 2009.

3. Make sure you have prepared your paper according to the NIPS style file. Use the official NIPS style files and resist temptations to redefine parameters such as `\textheight` and `\textwidth` to give yourself more space. There is a limit of 8 pages. Stay with the required font sizes. If you do not follow the formatting instructions we will ask you to resend your paper. This will only result in more work for yourself and for the publication chair. Links to the style files and detailled instructions are available on `http://ml.nec-labs.com/nips`.

4. Do not include in the front page of the paper any statement about "To appear in..." or about the NIPS category your paper belongs to. Do not force page numbers to appear.

5. Make sure your paper prints correctly in black and white. Do not rely on colors to convey information in figures. Avoid figures with dark grey curves on gray background.

6. Please use PostScript or PDF format with paper size "letter". Make sure that PDF files contain only Type-1 fonts using program `pdffonts` or using menu "File→Document Properties→Fonts" in Acrobat. Other fonts (like Type-3) might come from graphics files imported into the document. If you want to make sure your paper is printable, please use only Type-1 fonts.

   LaTeX users should try the following commands:

   ```
   dvips paper.dvi -o paper.ps -t letter -Ppdf -G0
   ps2pdf paper.ps
   ```

Contributing Authors' Letter of Agreement

(must be completed by one author per paper)

Title of article: ____

____

Author(s): ____

____

For the purpose of publication of the above article in the book tentatively entitled

Advances in Neural Information Processing Systems 21

I do warrant that we are the sole authors except for those portions shown to be in quotations;

I authorize the Neural Information Processing Systems Foundation to publish the article and to claim copyright in the book in which it is published for all editions, to license translations of the article as part of translations of the book as a whole, and also to license non exclusively reprints and electronic publications of the article;

I appoint the editor as our agent in everything pertaining to the publication of the article in this book;

I agree to read proof if requested to do so and otherwise to cooperate to the end of publication;

and I confirm that the article does not violate copyright or other proprietary right, that it contains no material that is libelous or in any way unlawful, that it has not previously been been published, in the whole or in part, except as I have advised the editor in writing, and that I will not now publish the article in any other edited volume without the consent of the Neural Information Processing Systems Foundation.

With this last exception, we retain the right to use the material in the article in any other writing of our own, whether in article or book form.

Signature: ____

Date: ____

Address: ____

____

____