NIPS 2007

NEURAL
INFORMATION
PROCESSING
SYSTEMS
CONFERENCE

Abstracts of Papers

TUTORIALS

December 3, 2007
Hyatt Regency
Vancouver, BC, Canada

CONFERENCE SESSIONS

December 3-6, 2007
Hyatt Regency
Vancouver, BC, Canada

WORKSHOPS

December 7-8, 2007
Westin Resort & Spa
Hilton Whistler Resort & Spa
Whistler, BC, Canada

NIPS 2007

Sponsored by the Neural Information Processing Systems Foundation, Inc.

There are 6 invited talks and 217 accepted papers selected from a total of 975 submissions 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 20, edited by John Platt, Daphne Koller, Yoram Singer, and Sam Roweis, to be published by NIPS in 2008.

Contents

Authors!

Please, see the last two pages of the book.

Organizing Committee

General Chair	John C. Platt, Microsoft Research
Program Co-Chairs	Daphne Koller, Stanford University
	Yoram Singer, Google
Tutorial Chair	Chris Burges, Microsoft Research
Workshop Co-Chairs	Adrienne Fairhall, University of Washington
	Charles Isbell, Georgia Institute of Technology
	Bob Williamson, National ICT Australia
Demonstration Co-Chairs	Giacomo Indiveri, INI ETH Zurich
	Xubo Song, Oregon Graduate Institute
Publications Chair	Sam Roweis, University of Toronto
Publicity Chair	Sumit Basu, Microsoft Research
Online Proceedings Chair	Andrew McCallum, University of Massachusetts, Amherst
Volunteers Chair	Fernando Perez Cruz, Princeton Univeristy

Program Committee

Program Co-Chairs	Daphne Koller, Stanford University
	Yoram Singer, Google
Program Committee Area Chairs	Francis Bach, Ecole des Mines de Paris
	Michael Black, Brown University
	Nicolo Cesa-Bianchi, Universita degli Studi di Milano
	Olivier Chapelle, Yahoo Research
	Sanjoy Dasgupta, UC San Diego
	Virginia de Sa, UC San Diego
	David Fleet, University of Toronto
	Isabelle Guyon, ClopiNet
	Bert Kappen, University of Nijmegen
	Dan Klein, UC Berkeley
	Chih-Jen Lin, National Taiwan University
	Kevin Murphy, University of British Columbia
	William Noble, University of Washington
	Stefan Schaal, University of Southern California
	Dale Schuurmans, University of Alberta
	Odelia Schwartz, Salk Institute & Albert Einstein College
	Fei Sha, UC Berkeley
	Mark Steyvers, UC Irvine
	Alan Stocker, New York University
	Yee Whye Teh, Gatsby Unit, UCL
	Nikos Vlassis, Technical University of Crete
	Ulrike von Luxburg, MPI for Biological Cybernetics
	Chris Williams, University of Edinburgh
	Andrew Zisserman, University of Oxford

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 2007 conference. The financial support enabled us to sponsor student travel and participation, the outstanding student paper awards, the demonstration track and the opening buffet.

Google
IBM Research
Intel
Microsoft Research
Pascal Network
Springer Machine Learning Journal
Toyota Technical Institute
Two Sigma Investments
Yahoo Research

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:

Rajat Raina (Stanford)
Chris Adams
Lee Campbell
Sarah Cercone
Chris Hiestand
Sheri Leone
Kristen Michener
Bryan Nielsen
Rosemary Miller
Mary Ellen Perry

Outstanding Student Paper Awards

Markov Chain Monte Carlo with People
Adam Sanborn and Thomas Griffiths

New Outer Bounds on the Marginal Polytope
David Sontag and Tommi Jaakkola

Student Paper Honorable Mentions

Inferring Elapsed Time from Stochastic Neural Processes
Maneesh Sahani and Misha Ahrens

Efficient Inference for Distributions on Permutations
Jonathan Huang, Carlos Guestrin and Leonidas Guibas

An Analysis of Convex Relaxations for MAP Estimation
Pawan Mudigonda, Vladimir Kolmogorov and Philip Torr

Program Highlights

Monday, December 3rd

7:30am–9:00am	Breakfast: *Regency A — Conference Level*
8:00am–6:00pm	Registration Desk Open
8:00am–6:00pm	Internet Access Room Open
9:30am–5:30pm	**Tutorials** 15
1:00pm–6:30pm	Poster/Demo Setup and Preview
6:30pm–7:30pm	Buffet and Opening Remarks (Dinner buffet open until 9pm; dessert buffet until 10pm.)
7:30pm–8:00pm	***Human Computation*** Luis von Ahn
8:10pm–8:25pm	**Oral Session: Opening Spotlights** 15
8:30pm–midnight	**Poster Sessions** 17
8:30pm–midnight	**Demo Sessions** 22

Tuesday, December 4th

7:30am–9:00am	Breakfast: *Perspectives Level*
8:00am–9:00am	Breakfast: also in *Regency A — Conference Level*
8:00am–6:00pm	Registration Desk Open
8:00am–6:00pm	Internet Access Room Open
8:30am–9:30am	**Invited Talk: *Statistical Models for Social Networks with Application to HIV Epidemiology*** Mark Handcock
9:30am–10:00am	**Oral Session: Structured Statistical Models** 23
10:00am–10:30am	Break
10:30am–12:00pm	**Oral Session: Probabilistic Optimization** 24
12:00pm–2:00pm	Break for Lunch
1:00pm–6:00pm	Poster/Demo Setup and Preview
2:00pm–3:00pm	**Invited Talk: *Projection Methods: Algorithmic Structures, Bregman Projections, and Acceleration Techniques*** Yair Censor

Wednesday, December 5th

Thursday, December 6th – NeuroThursday

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: Neuroscience I** 46
10:10am–10:40am	Break
10:40am–11:00am	**Oral Session: Neuroscience II** 46
11:00am–12:00pm	**Invited Talk:** ***Population coding of object images based on visual features and its relevance to view invariant representation*** Manabu Tanifuji
12:00pm	Close of Main Conference
2:00pm–3:30pm	Buses depart for Workshops

Schedule of Presentations

Monday, December 3rd

9:30am–11:30am **Tutorial Session 1**

1:00pm–3:00pm **Tutorial Session 2**

3:30pm–5:30pm **Tutorial Session 3**

6:30pm–7:30pm **Buffet and Opening Remarks**
(Dinner buffet open until 9pm; dessert buffet until 10pm.)

Monday, December 3rd

7:30PM–8:25PM Oral Session — Opening Spotlights (Chair: Dale Schuurmans, University of Alberta)

8:10PM **Spotlights**

7:30PM–Midnight Posters

7:30pm–Midnight Demos

3 ***Contraction of VLSI Spiking Neurons,*** Emre Neftci, Institute of Neuroinformatics, UNI-ETH Zurich, Elisabetta Chicca, Institute of Neuroinformatics, UNI-ETH Zurich, Giacomo Indiveri, Institute of Neuroinformatics, UNI-ETH Zurich, Jean-Jeacques Slotine, MIT , Rodney Douglas, Institute of Neuroinformatics, UNI-ETH Zurich

. . . . 84

4 ***Elefant,*** Kishor Gawande, National ICT Australia, Alex Smola, National ICT Australia, Vishwanathan S V N, National ICT Australia, Li Cheng, National ICT Australia, Simon Guenter, National ICT Australia 84

5 ***Gender and Age Recognition,*** Wei Xu, NEC Labs America, Kai Yu, NEC Labs America, Yihong Gong, NEC Labs America

. . . . 85

6 ***Learning To Race by Model-Based Reinforcement Learning with Adaptive Abstraction,*** Thore Graepel, Microsoft Research, Phil Trelford, Microsoft Research, Ralf Herbrich, Microsoft Research, Mykel Kochenderfer, Massachusetts Institute of Technology

. . . . 85

7 ***Predicting Human Gaze Using Low-level Saliency Combined with Face Detection,*** Moran Cerf, California Institute of Technology, Christof Koch, California Institute of Technology

. . . . 85

8 ***Robotic Eye Model with Learning of Pulse-Step Saccades,*** Per-Erik Forssen, University of British Columbia, Dinesh Pai, University of British Columbia

. . . . 85

Tuesday, December 4th

8:30AM–10:20AM Oral Session — Structured Statistical Models (Chair: Michael Black, Brown University)

8:30AM **Invited Talk:** ***Statistical Models for Social Networks with Application to HIV Epidemiology,*** Mark Handcock, University of Washington . . 87

9:30AM ***Probabilistic Matrix Factorization,*** Ruslan Salakhutdinov, University of Toronto, Department of Computer Science and Andriy Mnih, University of Toronto 87

9:50AM **Spotlights**

10:50AM–12:00PM Oral Session — Probabilistic Optimization (Chair: Francis Bach, Ecole Normale Superieure)

3:00PM–3:30PM Oral Session — Optimization for Learning (Chair: John Platt, Microsoft Research)

Consistent Minimization of Clustering Objective Functions, Ulrike von Luxburg, MPI for Biological Cybernetics, Sebastien Bubeck, INRIA futurs, Stefanie Jegelka, MPI for Biological Cybernetics and Michael Kaufmann, Universitat Tubingen 103

McRank: Learning to Rank Using Multiple Classification and Gradient Boosting, Ping Li, Cornell University, Christopher Burges, Microsoft Research and Qiang Wu, Microsoft Research 110

A Kernel Statistical Test of Independence, Arthur Gretton, MPI for Biological Cybernetics, Choon Hui Teo, SML, NICTA, Kenji Fukumizu, Institute of Statistical Mathematics, Le Song, NICTA and School of Information Technologies, the University of Sydney, Bernhard Scholkopf, MPI for Biological Cybernetics and Alex Smola, NICTA 123

Anytime Induction of Cost-sensitive Trees, Saher Esmeir, Computer Science Department, Technion-IIT and Shaul Markovitch, Computer Science Department, Technion-IIT 109

Iterative Non-linear Dimensionality Reduction with Manifold Sculpting, Michael Gashler, Brigham Young University, Dan Ventura, Brigham Young University and Tony Martinez, Brigham Young University 104

Random Features for Large-Scale Kernel Machines, Ali Rahimi, Intel Research and Benjamin Recht, California Institute of Technology 108

Boosting Algorithms for Maximizing the Soft Margin, Manfred Warmuth, UC Santa Cruz, Karen Glocer, UC Santa Cruz and Gunnar Ratsch, Friedrich Miescher Laboratory, Max Planck Society 102

Bundle Methods for Machine Learning, Alex Smola, NICTA, S V N Vishwanathan, Statistical Machine Learning Program, National ICT Australia and Quoc Le, Statistical Machine Learning, NICTA 102

Classification via Minimum Incremental Coding Length (MICL), John Wright, University of Illinois at Urbana-Champaign, Yangyu Tao, Microsoft Research, Zhouchen Lin, Microsoft Research, Yi Ma, Electrical & Computer Engineering Department, University of Illinois at Urbana-Champaign and Heung-Yeung Shum, Microsoft Research 111

4:00PM–5:30PM Oral Session — Theory and Sequential Decision Making (Chair: Sanjoy Dasgupta, University of California San Diego)

4:00PM ***FilterBoost: Regression and Classification on Large Datasets,*** Joseph K Bradley, Carnegie Mellon University and Robert Schapire, Princeton University 94

4:20PM ***The Price of Bandit Information for Online Optimization,*** Varsha Dani, University of Chicago, Thomas Hayes, Toyota Technological Institute at Chicago and Sham Kakade, Toyota Technological Institute 94

7:30PM–Midnight Posters

1 ***Stable Dual Dynamic Programming,*** Tao Wang, University of Alberta, Daniel Lizotte, U. Alberta, Michael Bowling, University of Alberta and Dale Schuurmans, University of Alberta ... 97

2 ***Bayes-Adaptive POMDPs,*** Stephane Ross, McGill University, School of Computer Science, Joelle Pineau, McGill University and Brahim Chaib-draa, Computer Science Department, Laval University ... 97

3 ***What makes some POMDP problems easy to approximate?,*** David Hsu, National University of Singapore, Wee Sun Lee, National University of Singapore and Nan Rong, Cornell University ... 97

4 ***Incremental Natural Actor-Critic Algorithms,*** Shalabh Bhatnagar, Indian Institute of Science, Richard Sutton, University of Alberta, Mohammad Ghavamzadeh, University of Alberta and Mark Lee, University of Alberta ... 98

5 ***Reinforcement Learning in Continuous Action Spaces through Sequential Monte Carlo Methods,*** Alessandro Lazaric, Politecnico di Milano, Marcello Restelli, Politecnico di Milano and Andrea Bonarini, AI&Robotics Lab - Politecnico di Milano ... 98

6 ***Fitted Q-iteration in continuous action-space MDPs,*** Csaba Szepesvari, University of Alberta, Andras Antos, Computer and Automation Research Institute, of the Hungarian Academy of Science and Remi Munos, INRIA . 99

7 ***Temporal Difference Updating without a Learning Rate,*** Marcus Hutter, RSISE(ANU) and SML(NICTA) and Shane Legg, IDSIA ... 99

8 ***Managing Power Consumption and Performance of Computing Systems Using Reinforcement Learning,*** Gerald Tesauro, IBM TJ Watson Research Center, Rajarshi Das, IBM Austin Research Laboratory, Hoi Chan, IBM Austin Research Laboratory, Jeffrey Kephart, IBM Research, David Levine, IBM Austin Research Laboratory, Freeman Rawson, IBM Austin Research Laboratory and Charles Lefurgy, IBM ... 100

9 ***Hierarchical Apprenticeship Learning with Application to Quadruped Locomotion,*** J. Zico Kolter, Stanford University, Pieter Abbeel, Stanford University and Andrew Ng, Stanford University ... 100

10 ***A Game-Theoretic Approach to Apprenticeship Learning,*** Umar Syed, Princeton University and Robert Schapire, Princeton University ... 94

11 ***The Price of Bandit Information for Online Optimization,*** Varsha Dani, University of Chicago, Thomas Hayes, Toyota Technological Institute at Chicago and Sham Kakade, Toyota Technological Institute ... 94

12 ***Adaptive Online Gradient Descent,*** Peter Bartlett, UC Berkeley, Alexander Rakhlin, UC Berkeley and Elad Hazan, Princeton University ... 92

7:30PM–Midnight Demos

Wednesday, December 5th

8:30AM–10:30AM Oral Session — Probabilistic Models and Methods (Chair: William Noble, University of Washington)

10:30AM–12:00PM Oral Session — Probabilistic Representations and Learning (Chair: YeeWhye Teh, Gatsby Computational Neuroscience Unit)

4:00PM–5:30PM Oral Session — Systems and Applications (Chair: Fei Sha, University of California Berkeley)

7:30PM–Midnight Posters

Thursday, December 6th

8:30AM–9:50AM Oral Session — Neuroscience I (Chair: Alan Stocker, New York University)

10:40AM–12:00PM Oral Session — Neuroscience II (Chair: Odelia Schwartz, Albert Einstein College of Medicine)

Abstract of Tutorials

Tutorial Session 1, 9:30am–11:30am

Tutorial: *Sensory Coding and Hierarchical Representations*

Michael Lewicki `http://www.cs.cmu.edu/~lewicki`
Carnegie Mellon University

The sensory and perceptual capabilities of biological organisms are still well beyond what we have been able to emulate with machines, and the brain devotes far more neural resources to the problems of sensory coding and early perception than we give credit in our algorithms. What is it all doing? Although a great deal has been learned about anatomical structure and physiological properties, insights into the underlying information processing algorithms have been difficult to obtain. Recent work, however, that has begun to elucidate some of the underlying computational principles and processes that biology uses to transform the raw sensory signal into a hierarchy of representations that subserve higher-level perceptual tasks. A central hypothesis in this work is that biological representations are optimal from the viewpoint of statistical information processing, and adapt to the statistics of the natural sensory environment. In this tutorial, I will review work on learning sensory codes that are optimal for the statistics of the natural sensory environment and show how these results provide theoretical explanations for a variety of physiological data in both the auditory and visual systems. This will include work that that has extended these results to provide functional explanations for many non-linear aspects of early auditory and visual processing. I will focus on work on the auditory and visual systems but also emphasize the generality of these approaches and how they can be applied to any sensory domain. I will also discuss work that generalizes the basic theory and shows how neural representations optimally compensate for sensory distortion and noise in neural populations. Finally, I will review work that goes beyond sensory coding and investigates the computational problems involved in computing more abstract sensory properties and invariant features that can subserve higher-level tasks such as perceptual organization and analysis of complex, natural scenes.

Dr. Lewicki an Associate professor in the Computer Science Department at Carnegie Mellon University and in the CMU-University of Pittsburgh Center for the Neural Basis of Cognition. He received his BS degree in mathematics and cognitive science from Carnegie Mellon University, his PhD degree in computation and neural systems from the California Institute of Technology, and did postdoctoral studies in the Computational Neurobiology Laboratory at the Salk Institute. His research involves theoretical and computational approaches to understanding the biological representation, processing, and learning of pattern structure in natural visual and acoustic environments.

Tutorial Session 1, 9:30am–11:30am

Tutorial: *Theory and Applications of Boosting*

Robert Schapire http://www.cs.princeton.edu/~schapire/
Princeton University

Boosting is a general method for producing a very accurate classification rule by combining rough and moderately inaccurate "rules of thumb." While rooted in a theoretical framework of machine learning, boosting has been found to perform quite well empirically. This tutorial will introduce the boosting algorithm AdaBoost, and explain the underlying theory of boosting, including explanations that have been given as to why boosting often does not suffer from overfitting, as well as some of the myriad other theoretical points of view that have been taken on this algorithm. Some practical applications and extensions of boosting will also be described.

Robert Schapire received his ScB in math and computer science from Brown University in 1986, and his SM (1988) and PhD (1991) from MIT under the supervision of Ronald Rivest. After a short post-doc at Harvard, he joined the technical staff at AT&T Labs (formerly AT&T Bell Laboratories) in 1991 where he remained for eleven years. At the end of 2002, he became a Professor of Computer Science at Princeton University. His awards include the 1991 ACM Doctoral Dissertation Award, the 2003 Godel Prize and the 2004 Kanelakkis Theory and Practice Award (both of the last two with Yoav Freund). His main research interest is in theoretical and applied machine learning.

Tutorial Session 2, 1:00pm–3:00pm

Tutorial: *Learning Using Many Examples*

Leon Bottou http://leon.bottou.org/
NEC Laboratories of America
Andrew Moore http://www.cs.cmu.edu/~awm/
Google

The statistical learning theory suggests to choose large capacity models that barely avoid over-fitting the training data. In that perspective, all datasets are small. Things become more complicated when one considers the computational cost of processing large datasets. Computationally challenging training sets appear when one want to emulate intelligence: biological brains learn quite efficiently from the continuous streams of perceptual data generated by our six senses, using limited amounts of sugar as a source of power. Computationally challenging training sets also appear when one wants to analyze the masses of data that describe the life of our computerized society. The more data we understand, the more we enjoy competitive advantages. The first part of the tutorial clarifies the relation between the statistical efficiency, the design of learning algorithms and their computational cost. The second part makes a detailed exploration of specific learning algorithms and of their implementation, with both simple and complex examples. The third part considers algorithms that learn with a single pass over the data. Certain algorithms have optimal properties but are often too costly. Workarounds are discussed. Finally, the fourth part shows how active example selection provides greater speed and reduces the feedback pressure that constrain parallel implementations.

Leon Bottou received a Diploe from l'Ecole Polytechnique, Paris in 1987, a Magistere en Mathematiques Fondamentales et Appliquees et Informatiques from Ecole Normale Superieure, Paris in 1988, and a PhD in Computer Science from Universite de Paris-Sud in 1991. He joined AT&T Bell Labs from 1991 to 1992 and AT&T Labs from 1995 to 2002. Between 1992 and 1995 he was chairman of Neuristique in Paris, a small company pioneering machine learning for data mining applications. He has been with NEC Labs America in Princeton since 2002. Leon's primary research interest is machine learning. His contributions to this field address theory, algorithms and large scale applications. Leon's secondary research interest is data compression and coding. His best known contribution in this field is the DjVu document compression technology (http://www.djvu.org.) Leon published over 70 papers and is serving on the boards of JMLR and IEEE TPAMI. He also serves on the scientific advisory board of Kxen Inc.

Andrew Moore is currently responsible for growing a new Google office on CMU's campus in Pittsburgh. The office focuses on numerous statistical and large scale systems issues in structured data extraction, Google's infrastructure, internet advertising, and fraud prevention. Prior to joining Google in January 2006, Andrew was a Professor of Robotics and Computer Science at the School of Computer Science, Carnegie Mellon University. Andrew began his career writing video-games for an obscure British personal computer (http://www.oric.org/index.php?page=software&fille=detail&num_log=2). He rapidly became a thousandaire and retired to academia, where he received a PhD from the University of Cambridge in 1991. His main research interest is data mining: statistical algorithms for finding all the potentially useful and statistically meaningful patterns in large sources of data.

Tutorial Session 2, 1:00pm–3:00pm

Tutorial: *Visual Recognition in Primates and Machines*

Tomaso Poggio http://cbcl.mit.edu/
McGovern Institute for Brain Research at MIT

Understanding the processing of information in our cortex is a significant part of understanding how the brain works and of understanding intelligence itself, arguably one of the greatest problems in science today. In particular, our visual abilities are computationally amazing and we are still far from imitating them with computers. Thus, visual cortex may well be a good proxy for the rest of the cortex and indeed for intelligence itself. But despite enormous progress in the physiology and anatomy of the visual cortex, our understanding of the underlying computations remains fragmentary. I will briefly review the anatomy and the physiology of primate visual cortex and then describe a class of quantitative models of the ventral stream for object recognition, which, heavily constrained by physiology and biophysics, have been developed during the last two decades and which have been recently shown to be quite successful in explaining several physiological data across different visual areas. I will discuss their performance and architecture from the point of view of state-of-the-art computer vision system. Surprisingly, such models also mimic the level of human performance in difficult rapid image categorization tasks in which human vision is forced to operate in a feedforward mode. I will then focus on the key limitations of such hierarchical feedforward models for object recognition, discuss why they are incomplete models of vision and suggest possible alternatives focusing on the computational role of attention and its likely substrate – cortical backprojections. Finally, I will outline a program of research to attack the broad challenge of understanding in terms of brain circuits the process of image inference and in particular recognition tasks beyond simple scene classification.

Tomaso A. Poggio, is the Eugene McDermott Professor at the Department of Brain and Cognitive Sciences; Co-Director, Center for Biological and Computational Learning; Member for the last 25 years of the Computer Science and Artificial Intelligence Laboratory at MIT; since 2000, member of the faculty of the McGovern Institute for Brain Research and member of the steering committee of the Center for Collective Intelligence. He is author or co-author of over 400 papers in the fields of learning theory, computer science, computational neuroscience, and nonlinear systems theory and he belongs to the editorial board of several scientific journals. He is an honorary member of the Neuroscience Research Program, a member of the American Academy of Arts and Sciences and a Founding Fellow of AAAI. He received several awards such as the Otto-Hahn-Medaille Award of the Max-Planck-Society, the Max Planck Research Award (with M. Fahle), from the Alexander von Humboldt Foundation, the MIT 50K Entrepreneurship Competition Award, the Laurea Honoris Causa in Ingegneria Informatica for the Bicentenario dell'Invenzione della Pila from the University of Pavia and the 2003 Gabor Award. His research has been interdisciplinary, between brains and computers. It is now focused on the mathematics of learning theory, the applications of learning techniques to computer vision, bioinformatics, computer graphics and especially on computational neuroscience of the visual cortex in close collaboration with several physiology labs.

Tutorial Session 3, 3:30pm–5:30pm

Tutorial: *Deep Belief Nets*

Geoffrey Hinton http://www.cs.toronto.edu/~hinton
University of Toronto

Complex probabilistic models of unlabeled data can be created by combining simpler models. Mixture models are obtained by averaging the densities of simpler models and "products of experts" are obtained by multiplying the densities together and renormalizing. A far more powerful type of combination is to form a "composition of experts" by treating the values of the latent variables of one model as the data for learning the next model. The first half of the tutorial will show how deep belief nets – directed generative models with many layers of hidden variables – can be learned one layer at a time by composing simple, undirected, product of expert models that only have one hidden layer. It will also explain why composing directed models does not work. Deep belief nets are trained as generative models on large, unlabeled datasets, but once multiple layers of features have been created by unsupervised learning, they can be fine-tuned to give excellent discrimination on small, labeled datasets. The second half of the tutorial will describe applications of deep belief nets to several tasks including object recognition, non-linear dimensionality reduction, document retrieval, and the interpretation of medical images. It will also show how the learning procedure for deep belief nets can be extended to high-dimensional time series and hierarchies of Conditional Random Fields.

Geoffrey Hinton received his PhD in Artificial Intelligence from the University of Edinburgh in 1978. He did postdoctoral work at the University of California San Diego and spent five years as a faculty member at Carnegie-Mellon University. He then moved to the University of Toronto where he is a University Professor in the Department of Computer Science and directs the program in Neural Computation and Adaptive Perception for the Canadian Institute for Advanced Research. From 1998 to 2001 he set up the Gatsby Computational Neuroscience Unit at University College London. Geoffrey Hinton is a fellow of the Royal Society, an honorary foreign member of the American Academy of Arts and Sciences, and a fellow of the Royal Society of Canada. He has received an honorary doctorate from the University of Edinburgh, the first David E. Rumelhart prize, and the IJCAI award for research excellence. Geoffrey Hinton and his research collaborators introduced Boltzmann machines, backpropagation, distributed representations, mixtures of experts, variational inference, variational Bayes, products of experts and deep belief nets.

Tutorial Session 3, 3:30pm–5:30pm

Tutorial: *Structured Prediction*

BEN TASKAR http://www.cis.upenn.edu/~taskar/
University of Pennsylvania

Structured prediction is a framework for solving problems of classification or regression in which the output variables are mutually dependent or constrained. These dependencies and constraints reflect sequential, spatial or combinatorial structure in the problem domain, and capturing such interactions is often as important as capturing input-output dependencies. Many such problems, including natural language parsing, machine translation, object segmentation, gene prediction, protein alignment and numerous other tasks in computational linguistics, speech, vision, biology, are not new. However, recent advances have brought about a unified view, efficient methodology and more importantly, significant accuracy improvements for both classical and novel problems. This tutorial will explain the fundamental computational and statistical challenges arising from the high dimensionality of the inputs and the exponential explosion of the number of possible joint outcomes. I will describe the confluence of developments in several areas in resolving these challenges for broad classes of problems: large margin and online methods for prediction, variational methods for graphical model inference, and large scale combinatorial and convex optimization. I will also outline several open issues of particular difficulty in structured prediction, including asymptotic consistency, the effects of approximate inference, semisupervised and weakly supervised learning.

Ben Taskar received his bachelor's and doctoral degree in Computer Science from Stanford University. After a postdoc at the University of California at Berkeley, he joined the faculty at the University of Pennsylvania Computer and Information Science Department in 2007. His research interests include machine learning, graphical models, large-scale and distributed convex optimization, applications in natural language processing, computer vision, and computational biology. His work on structured prediction has received awards at NIPS and EMNLP conferences.

Abstract of Presentations

Monday, December 3rd

Oral Session — Opening Spotlights (Chair: Dale Schuurmans, University of Alberta):

7:30PM Invited Talk: *Human Computation*

Luis von Ahn biglou@cs.cmu.edu
Carnegie Mellon University

Construction of the Empire State Building: 7 million human-hours. The Panama Canal: 20 million human-hours. Estimated number of human-hours spent playing computer solitaire around the world in one year: billions. A problem with today's computer society? No, an opportunity. What if this time and energy could be channeled into useful work? What if people could play computer games and accomplish work without even realizing it? What if billions of people collaborated to solve important problems for humanity or generate training data for computers? My work aims at a general paradigm for doing exactly that: utilizing human processing power to solve computational problems in a distributed manner. In particular, I focus on harnessing human time and energy for addressing problems that computers cannot yet solve. Although computers have advanced dramatically in many respects over the last 50 years, they still do not possess the basic conceptual intelligence or perceptual capabilities that most humans take for granted. By leveraging human skills and abilities in a novel way, I want to solve large-scale computational problems and/or collect training data to teach computers many of these human talents. To this end, I treat human brains as processors in a distributed system, each performing a small part of a massive computation. Unlike computer processors, however, humans require an incentive in order to become part of a collective computation. Among other things, I use online games as a means to encourage participation in the process.

8:10PM Spotlights

Discriminative Log-Linear Grammars with Latent Variables
Slav Petrov, University of California, Berkeley and Dan Klein, UC Berkeley.
See abstract, page 80.

Regret Minimization in Games with Incomplete Information
Martin Zinkevich, University of Alberta, Michael Johanson, University of Alberta, Michael Bowling, University of Alberta and Carmelo Piccione, University of Alberta.
See abstract, page 56.

Learning Monotonic Transformations for Classification
Andrew Howard, Columbia University and Tony Jebara, Columbia University.
See abstract, page 67.

Structured Learning with Approximate Inference
ALEX KULESZA, University of Pennsylvania and FERNANDO PEREIRA, Computer and Information Science, University of Pennsylvania.
See abstract, page 72.

Convex Clustering with Exemplar-Based Models
DANIAL LASHKARI, Massachusetts Institute of Technology and POLINA GOLLAND, Massachusetts Institute of Technology.
See abstract, page 62.

Feature Selection Methods for Improving Protein Structure Prediction with Rosetta
MICHAEL JORDAN, University of California, Berkeley, BEN BLUM, University of California, Berkeley, DAVID BAKER, University of Washington, PHILIP BRADLEY, MIT, RHIJU DAS, University of Washington and DAVID KIM, University of Washington.
See abstract, page 60.

A Bayesian LDA-based model for semi-supervised part-of-speech tagging
KRISTINA TOUTANOVA, Microsoft Research and MARK JOHNSON, Cognitive and Linguistic Sciences, Box 1978.
See abstract, page 81.

The Value of Labeled and Unlabeled Examples when the Model is Imperfect
KAUSHIK SINHA, Dept of Computer Science and Engineering, Ohio State University and MIKHAIL BELKIN, Ohio State University, Computer Science and Engineering.
See abstract, page 64.

Statistical Analysis of Semi-Supervised Regression
JOHN LAFFERTY, Carnegie Mellon and LARRY WASSERMAN, Carnegie Mellon.
See abstract, page 67.

Boosting the Area under the ROC Curve
PHIL LONG, Google and ROCCO SERVEDIO, Department of Computer Science, Columbia University.
See abstract, page 68.

Discovering Weakly-Interacting Factors in a Complex Stochastic Process
CHARLIE FROGNER, Harvard University and AVI PFEFFER, Harvard.
See abstract, page 76.

Convex Learning with Invariances
CHOON HUI TEO, Statistical Machine Learning Program, NICTA, Australian National University, AMIR GLOBERSON, CSAIL, MIT, SAM ROWEIS, University of Toronto and ALEX SMOLA, NICTA.
See abstract, page 70.

Estimating divergence functionals and the likelihood ratio by penalized convex risk minimization
XUANLONG NGUYEN, SAMSI, Duke University, MARTIN WAINWRIGHT, Department of EECS, Department of Statistics and MICHAEL JORDAN, University of California, Berkeley.
See abstract, page 70.

Bayesian Policy Learning with Trans-Dimensional MCMC

Ajay Jasra, Imperial College London, Arnaud Doucet, University of British Columbia, Matthew Hoffman, University of British Columbia and Nando de Freitas, University of British Columbia.

See abstract, page 58.

Learning with Transformation Invariant Kernels

Olivier Chapelle, Yahoo Research and Christian Walder, Max Planck Institute.

See abstract, page 71.

Posters:

1 *Regret Minimization in Games with Incomplete Information*

Martin Zinkevich maz@cs.ualberta.ca
University of Alberta
Michael Johanson johanson@cs.ualberta.ca
University of Alberta
Michael Bowling bowling@cs.ualberta.ca
University of Alberta
Carmelo Piccione carm@cs.ualberta.ca
University of Alberta

Extensive games are a powerful model of multiagent decision-making scenarios with incomplete information. Finding a Nash equilibrium for very large instances of these games has received a great deal of recent attention. In this paper, we describe a new technique for solving large games based on regret minimization. In particular, we introduce the notion of counterfactual regret, which exploits the degree of incomplete information in an extensive game. We show how minimizing counterfactual regret minimizes overall regret, and therefore in self-play can be used to compute a Nash equilibrium. We demonstrate this technique in the domain of poker, showing we can solve abstractions of limit Texas Hold'em with as many as $10^1 2$ states, two orders of magnitude larger than previous methods.

Spotlight presentation, Monday, 8:10PM.

2 *Computing Robust Counter-Strategies*

Michael Johanson johanson@cs.ualberta.ca
University of Alberta
Martin Zinkevich maz@cs.ualberta.ca
University of Alberta
Michael Bowling bowling@cs.ualberta.ca
University of Alberta

Adaptation to other initially unknown agents often requires computing an effective counter-strategy. In the Bayesian paradigm, one must find a good counter-strategy to the inferred posterior of the other agents' behavior. In the experts paradigm, one may want to choose experts that are good counter-strategies to the other agents' expected behavior. In this paper we introduce a technique for computing robust counter-strategies for adaptation in multiagent scenarios under a variety of paradigms. The strategies can take advantage of a suspected tendency in the decisions of the other agents, while bounding the worst-case performance when the tendency is not observed. The technique involves solving a modified game, and therefore can make use of recently developed algorithms for solving very large extensive games. We demonstrate the effectiveness of the technique in two-player Texas Hold'em. We show that the computed poker strategies are substantially more robust than best response counter-strategies, while still exploiting a suspected tendency. We also compose the generated strategies in an experts algorithm showing a dramatic improvement in performance over using simple best responses.

3 *Computational Equivalence of Fixed Points and No Regret Algorithms, and Convergence to Equilibria*

Satyen Kale satyen@cs.princeton.edu
Princeton University, Microsoft Research
Elad Hazan hazan@us.ibm.com
IBM Almaden Research Center

We study the relation between notions of game-theoretic equilibria which are based on stability under a set of deviations, and empirical equilibria which are reached by rational players. Rational players are modelled by players using no regret algorithms, which guarantee that their payoff in the long run is almost as much as the most they could hope to achieve by consistently deviating from the algorithm's suggested action. We show that for a given set of deviations over the strategy set of a player, it is possible to efficiently approximate fixed points of a given deviation if and only if there exist efficient no regret algorithms resistant to the deviations. Further, we show that if all players use a no regret algorithm, then the empirical distribution of their plays converges to an equilibrium.

4 *Competition Adds Complexity*

Judy Goldsmith goldsmit@cs.uky.edu
University of Kentucky, Dept of Computer Science
Martin Mundhenk mundhenk@cs.uni-jena.de
Friedrich-Schiller-Universitaet Jena

It is known that determinining whether a DEC-POMDP, namely, a cooperative partially observable stochastic game (POSG), has a cooperative strategy with positive expected reward is complete for NEXP. It was not known until now how cooperation affected that complexity. We show that, for competitive POSGs, the complexity of determining whether one team has a positive-expected-reward strategy is complete for the class NEXP with an oracle for NP.

5 *Receding Horizon Differential Dynamic Programming*

Yuval Tassa yuval.tassa@gmail.com
The Hebrew University
Tom Erez etom@cse.wustl.edu
Washington University in St. Louis
William Smart wds@cse.wustl.edu
Washington University in St. Louis

The control of high-dimensional, continuous, non-linear systems is a key problem in reinforcement learning and control. Local, trajectory-based methods, using techniques such as Differential Dynamic Programming (DDP) are not directly subject to the curse of dimensionality, but generate only local controllers. In this paper, we introduce Receding Horizon DDP (RH-DDP), an extension to the classic DDP algorithm, which allows us to construct stable and robust controllers based on a library of local-control trajectories. We demonstrate the effectiveness of our approach on a series of high-dimensional control problems using a simulated multi-link swimming robot. These experiments show that our approach effectively circumvents dimensionality issues, and is capable of dealing effectively with problems with (at least) 34 state and 14 action dimensions.

6 *Random Sampling of States in Dynamic Programming*

Chris Atkeson cga@cmu.edu
CMU Robotics Institute
Benjamin Stephens bstephens@cmu.edu
Carnegie Mellon University, Robotics Institute

We combine two threads of research on approximate dynamic programming: random sampling of states and using local trajectory optimizers to globally optimize a policy and associated value function. This combination allows us to replace a dense multidimensional grid with a much sparser adaptive sampling of states. Our focus is on finding steady state policies for the deterministic time invariant discrete time control problems with continuous states and actions often found in robotics. In this paper we show that we can now solve problems we couldn't solve previously with regular grid-based approaches.

7 *Bayesian Policy Learning with Trans-Dimensional MCMC*

Ajay Jasra ajay.jasra@imperial.ac.uk
Imperial College London
Arnaud Doucet arnaud@cs.ubc.ca
University of British Columbia
Matthew Hoffman hoffmanm@cs.ubc.ca
University of British Columbia
Nando de Freitas nando@cs.ubc.ca
University of British Columbia

A recently proposed formulation of the stochastic planning and control problem as one of parameter estimation for suitable artificial statistical models has led to the adoption of inference algorithms for this notoriously hard problem. At the algorithmic level, the focus has been on developing Expectation-Maximization (EM) algorithms. In this paper, we begin by making the crucial observation that the stochastic control problem can be reinterpreted as one of trans-dimensional inference. With this new understanding, we are able to propose a novel reversible jump Markov chain Monte Carlo (MCMC) algorithm that is more efficient than its EM counterparts. Moreover, it enables us to carry out full Bayesian policy search, without the need for gradients and with one single Markov chain. The new approach involves sampling directly from a distribution that is proportional to the reward and, consequently, performs better than classic simulations methods in situations where the reward is a rare event.

Spotlight presentation, Monday, 8:10PM.

8 *Theoretical Analysis of Heuristic Search Methods for Online POMDPs*

Stephane Ross stephane.ross@mail.mcgill.ca
McGill University, School of Computer Science
Brahim Chaib-draa chaib@ift.ulaval.ca
Computer Science Department, Laval University
Joelle Pineau jpineau@cs.mcgill.ca
McGill University

Planning in partially observable environments remains a challenging problem, despite significant recent advances in offline approximation techniques. A few online methods have also been proposed recently, and proven to be remarkably scalable, but without the theoretical guarantees of their offline counterparts. Thus it seems natural to try to unify offline and online techniques, preserving the theoretical properties of the former, and exploiting the scalability of the latter. In this paper, we provide theoretical guarantees on an anytime algorithm for POMDPs which aims to reduce the error made by approximate offline value iteration algorithms through the use of an efficient online searching procedure. The algorithm uses search heuristics based on an error analysis of lookahead search, to guide the online search towards reachable beliefs with the most potential to reduce error. We provide a general theorem showing that these search heuristics are admissible, and lead to complete and epsilon-optimal algorithms. This is, to the best of our knowledge, the strongest theoretical result available for online POMDP solution methods. We also provide empirical evidence showing that our approach is also practical, and can find (provably) near-optimal solutions in reasonable time.

9 *Optimistic Linear Programming gives Logarithmic Regret for Irreducible MDPs*

Ambuj Tewari ambuj@berkeley.edu
CS Division, UC Berkeley
Peter Bartlett bartlett@cs.berkeley.edu
UC Berkeley

We present an algorithm called Optimistic Linear Programming (OLP) for learning to optimize average reward in an irreducible but otherwise unknown Markov decision process (MDP). OLP uses its experience so far to estimate the MDP. It chooses actions by optimistically maximizing estimated future rewards over a set of next-state transition probabilities that are close to the estimates: a computation that corresponds to solving linear programs. We show that the total expected reward obtained by OLP up to time T is within $C(P) \log T$ of the reward obtained by the optimal policy, where $C(P)$ is an explicit, MDP-dependent constant. OLP is closely related to an algorithm proposed by Burnetas and Katehakis with four key differences: OLP is simpler, it does not require knowledge of the supports of transition probabilities and the proof of the regret bound is simpler, but our regret bound is a constant factor larger than the regret of their algorithm. OLP is also similar in flavor to an algorithm recently proposed by Auer and Ortner. But OLP is simpler and its regret bound has a better dependence on the size of the MDP.

10 *The Epoch-Greedy Algorithm for Multi-armed Bandits with Side Information*

John Langford jl@yahoo-inc.com
Yahoo Research
Tong Zhang tzhang@stat.rutgers.edu
Rutgers University

We present Epoch-Greedy, an algorithm for multi-armed bandits with observable side information. Epoch-Greedy has the following properties: No knowledge of a time horizon T is necessary. The regret incurred by Epoch-Greedy is controlled by a sample complexity bound for a hypothesis class. The regret scales as $O(T^{2/3}S^{1/3})$ or better (sometimes, much better). Here S is the complexity term in a sample complexity bound for standard supervised learning.

11 *Online Linear Regression and Its Application to Model-Based Reinforcement Learning*

Alexander Strehl strehl@cs.rutgers.edu
Yahoo Research, Rutgers University
Michael Littman mlittman@cs.rutgers.edu
Rutgers, Department of Computer Science

We provide a provably efficient algorithm for learning Markov Decision Processes (MDPs) with continuous state and action spaces in the online setting. Specifically, we take a model-based approach and show that a special type of online linear regression allows us to learn MDPs with (possibly kernalized) linearly parameterized dynamics. This result builds on Kearns and Singh's work that provides a provably efficient algorithm for finite state MDPs. Our approach is not restricted to the linear setting, and is applicable to other classes of continuous MDPs.

12 *Scan Strategies for Meteorological Radars*

Victoria Manfredi vmanfred@cs.umass.edu
University of Massachusetts Amherst
Jim Kurose kurose@cs.umass.edu
University of Massachusetts Amherst

We address the problem of adaptive sensor control in dynamic resource-constrained sensor networks. We focus on a meteorological sensing network comprising radars that can perform sector scanning rather than always scanning 360 degrees. We compare three sector scanning strategies. The sit-and-spin strategy always scans 360 degrees. The limited lookahead strategy additionally uses the expected environmental state K decision epochs in the future, as predicted from Kalman filters, in its decision-making. The full lookahead strategy uses all expected future states by casting the problem as a Markov decision process and using reinforcement learning to estimate the optimal scan strategy. We show that the main benefits of using a lookahead strategy are when there are multiple meteorological phenomena in the environment, and when the maximum radius of any phenomenon is sufficiently smaller than the radius of the radars. We also show that there is a tradeoff between the average quality with which a phenomenon is scanned and the number of decision epochs before which a phenomenon is rescanned.

13 *Feature Selection Methods for Improving Protein Structure Prediction with Rosetta*

Michael Jordan jordan@cs.berkeley.edu
University of California, Berkeley
Ben Blum bblum@cs.berkeley.edu
University of California, Berkeley
David Baker dabaker@u.washington.edu
University of Washington
Philip Bradley pbradley@alum.mit.edu
MIT
Rhiju Das rhiju@u.washington.edu
University of Washington

David Kim dekim@u.washington.edu
University of Washington

Rosetta is one of the leading algorithms for protein structure prediction today. It is a Monte Carlo energy minimization method requiring many random restarts to find structures with low energy. In this paper we present a resampling technique for structure prediction of small alpha/beta proteins using Rosetta. From an initial round of Rosetta sampling, we learn properties of the energy landscape that guide a subsequent round of sampling toward lower-energy structures. Rather than attempt to fit the full energy landscape, we use feature selection methods–L1-regularized linear regression–to identify structural features that give rise to low energy. We then enrich these structural features in the second sampling round. Results are presented across a benchmark set of nine small alpha/beta proteins demonstrating that our methods seldom impair, and frequently improve, Rosetta's performance.

Spotlight presentation, Monday, 8:10PM.

14 *A learning framework for nearest neighbor search*

Lawrence Cayton lcayton@cs.ucsd.edu
University of California, San Diego
Sanjoy Dasgupta dasgupta@cs.ucsd.edu
University of California San Diego

Can we leverage learning techniques to build a fast nearest-neighbor (NN) retrieval data structure? We present a general learning framework for the NN problem in which sample queries are used to learn the parameters of a data structure that minimize the retrieval time and/or the miss rate. We explore the potential of this novel framework through two popular NN data structures: KD-trees and the rectilinear structures employed by locality sensitive hashing. We derive a generalization theory for these data structure classes and present simple learning algorithms for both. Experimental results reveal that learning often improves on the already strong performance of these data structures.

15 *Topmoumoute Online Natural Gradient Algorithm*

Nicolas Le Roux nicolas.le.roux@umontreal.ca
Universite de Montreal
Pierre-Antoine Manzagol manzagop@iro.umontreal.ca
University of Montreal
Yoshua Bengio yoshua.bengio@umontreal.ca
Universite de Montreal

Guided by the goal of obtaining an optimization algorithm that is both fast and yielding good generalization, we study the descent direction maximizing the decrease in generalization error or the probability of not increasing generalization error. The surprising result is that from both the Bayesian and frequentist perspectives this can yield the natural gradient direction. Although that direction can be very expensive to compute we develop an efficient, general, online approximation to the natural gradient descent which is suited to large scale problems. We report experimental results showing much faster convergence in computation time and in number of iterations with TONGA (Topmoumoute Online natural Gradient Algorithm) than with stochastic gradient descent, even on very large datasets.

16 *Ultrafast Monte Carlo for Statistical Summations*

Michael Holmes mph@cc.gatech.edu
Georgia Tech, College of Computing
Alexander Gray agray@cc.gatech.edu
Georgia Institute of Technology, College of Computing
Charles Isbell isbell@cc.gatech.edu
Georgia Institute of Technology

Machine learning contains many computational bottlenecks in the form of nested summations over datasets. Computation of these summations is typically $O(n^2)$ or higher, which severely limits application to large datasets. We present a multi-stage stratified Monte Carlo method for approximating such summations with probabilistic relative error control. The essential idea is fast approximation by sampling in trees. This method differs from many previous scalability techniques (such as multi-tree methods) in that its error is stochastic, but we derive conditions for error control and demonstrate that they work. Further, we give a theoretical sample complexity for the method that is independent of dataset size, and show that this appears to hold in experiments, where speedups reach as high as 10^{14}, many orders of magnitude beyond the previous state of the art.

17 *Convex Clustering with Exemplar-Based Models*

Danial Lashkari danial@mit.edu
Massachusetts Institute of Technology
Polina Golland polina@csail.mit.edu
Massachusetts Institute of Technology

Clustering is often formulated as the maximum likelihood estimation of a mixture model that explains the data. The EM algorithm widely used to solve the resulting optimization problem is inherently a gradient-descent method and is sensitive to initialization. The resulting solution is a local optimum in the neighborhood of the initial guess. This sensitivity to initialization presents a significant challenge in clustering large data sets into many clusters. In this paper, we present a different approach to approximate mixture fitting for clustering. We introduce an exemplar-based likelihood function that approximates the exact likelihood. This formulation leads to a convex minimization problem and an efficient algorithm with *guaranteed convergence to the globally optimal solution*. The resulting clustering can be thought of as a probabilistic mapping of the data points to the set of exemplars that minimizes the average distance and the information-theoretic cost of mapping. We present experimental results illustrating the performance of our algorithm and its comparison with the conventional approach to mixture model clustering.

Spotlight presentation, Monday, 8:10PM.

18 *Random Projections for Manifold Learning*

Richard Baraniuk richb@rice.edu
Rice University
Michael Wakin wakin@acm.caltech.edu
California Institute of Technology
Chinmay Hegde ch3@rice.edu
Rice University

We propose a novel method for *linear* dimensionality reduction of manifold modeled data. First, we show that with a small number M of *random projections* of sample points in R^N

belonging to an unknown K-dimensional Euclidean manifold, the intrinsic dimension (ID) of the sample set can be estimated to high accuracy. Second, we rigorously prove that using only this set of random projections, we can estimate the structure of the underlying manifold. In both cases, the number random projections required is linear in K and logarithmic in N, meaning that $K < M \ll N$. To handle practical situations, we develop a greedy algorithm to estimate the smallest size of the projection space required to perform manifold learning. Our method is particularly relevant in distributed sensing systems and leads to significant potential savings in data acquisition, storage and transmission costs.

19 *Learning the structure of manifolds using random projections*

Yoav Freund yfreund@ucsd.edu
University of California, San Diego, Computer Science and Engineering
Sanjoy Dasgupta dasgupta@cs.ucsd.edu
University of California San Diego
Mayank Kabra mkabra@ucsd.edu
University of California, San Diego
Nakul Verma naverma@cs.ucsd.edu
CSE, University of California, San Diego

We present a simple variant of the k-d tree which automatically adapts to intrinsic low dimensional structure in data.

20 *A Unified Near-Optimal Estimator For Dimension Reduction in l_α ($0 < \alpha \leq 2$) Using Stable Random Projections*

Ping Li pingli@cornell.edu
Cornell University
Trevor Hastie hastie@stanford.edu
Stanford University

Many tasks (e.g., clustering) in machine learning only require the l_α distances instead of the original data. For dimension reductions in the l_α norm ($0 < \alpha \leq 2$), the method of *stable random projections* can efficiently compute the l_α distances in massive datasets (e.g., the Web or massive data streams) in one pass of the data. The estimation task for *stable random projections* has been an interesting topic. We propose a simple estimator based on the *fractional power* of the samples (projected data), which is surprisingly near-optimal in terms of the asymptotic variance. In fact, it achieves the Cramér-Rao bound when $\alpha = 2$ and $\alpha = 0+$. This new result will be useful when applying *stable random projections* to distance-based clustering, classifications, kernels, massive data streams etc.

21 *Modeling homophily and stochastic equivalence in symmetric relational data*

Peter Hoff hoff@stat.washington.edu
University of Washington

This article discusses a latent variable model for inference and prediction of symmetric relational data. The model, based on the idea of the eigenvalue decomposition, represents the relationship between two nodes as the weighted inner-product of node-specific vectors of latent characteristics. This "eigenmodel" generalizes other popular latent variable models, such as latent class and distance models: It is shown mathematically that any latent

class or distance model has a representation as an eigenmodel, but not vice-versa. The practical implications of this are examined in the context of three real datasets, for which the eigenmodel has as good or better out-of-sample predictive performance than the other two models.

22 *The Value of Labeled and Unlabeled Examples when the Model is Imperfect*

KAUSHIK SINHA sinha.39@osu.edu
Dept of Computer Science and Engineering, Ohio State University
MIKHAIL BELKIN mbelkin@cse.ohio-state.edu
Ohio State University, Computer Science and Engineering

Semi-supervised learning, i.e. learning from both labeled and unlabeled data has received significant attention in the machine learning literature in recent years. Still our understanding of the theoretical foundations of the usefulness of unlabeled data remains somewhat limited. The simplest and the best understood situation is when the data is described by an identifiable mixture model, and where each class comes from a pure component. This natural setup and its implications ware analyzed previously. One important result was that in certain regimes, labeled data becomes exponentially more valuable than unlabeled data. However, in most realistic situations, one would not expect that the data comes from a parametric mixture distribution with identifiable components. There have been recent efforts to analyze the non-parametric situation, for example, "cluster" and "manifold" assumptions have been suggested as a basis for analysis. Still, a satisfactory and fairly complete theoretical understanding of the nonparametric problem, similar to that studied previously has not yet been developed. In this paper we investigate an intermediate situation, when the data comes from a probability distribution, which can be modeled, but not perfectly, by an identifiable mixture distribution. This seems applicable to many situation, when, for example, a mixture of Gaussians is used to model the data. the contribution of this paper is an analysis of the role of labeled and unlabeled data depending on the amount of imperfection in the model.

Spotlight presentation, Monday, 8:10PM.

23 *Discriminative Batch Mode Active Learning*

YUHONG GUO yuhong@cs.ualberta.ca
U. Alberta
DALE SCHUURMANS dale@cs.ualberta.ca
University of Alberta

Active learning sequentially selects unlabeled instances to label with the goal of reducing the effort needed to learn a good classifier. Most previous studies in active learning have focused on selecting one unlabeled instance to label at a time while retraining in each iteration. Recently a few batch mode active learning approaches have been proposed that select a set of most informative unlabeled instances in each iteration under the guidance of heuristic scores. In this paper, we propose a discriminative batch mode active learning approach that formulates the instance selection task as a continuous optimization problem over auxiliary instance selection variables. The optimization is formulated to maximize the discriminative classification performance of the target classifier, while also taking the unlabeled data into account. Although the objective is not convex, we can manipulate a quasi-Newton method to obtain a good local solution. Our empirical studies on UCI

datasets show that the proposed active learning is more effective than current state-of-the art batch mode active learning algorithms.

24 *Multiple-Instance Active Learning*

Burr Settles bsettles@cs.wisc.edu
University of Wisconsin
Mark Craven craven@biostat.wisc.edu
University of Wisconsin
Soumya Ray sray@eecs.oregonstate.edu
Oregon State University

We present a framework for active learning in the multiple-instance (MI) setting. In an MI learning problem, instances are naturally organized into bags and it is the bags, instead of individual instances, that are labeled for training. MI learners assume that every instance in a bag labeled negative is actually negative, whereas at least one instance in a bag labeled positive is actually positive. We consider the particular case in which an MI learner is allowed to selectively query unlabeled instances from positive bags. This approach is well motivated in domains in which it is inexpensive to acquire bag labels and possible, but expensive, to acquire instance labels. We describe a method for learning from labels at mixed levels of granularity, and introduce two active query selection strategies motivated by the MI setting. Our experiments show that learning from instance labels can significantly improve performance of a basic MI learning algorithm in two multiple-instance domains: content-based image retrieval and text classification.

25 *Active Preference Learning with Discrete Choice Data*

Brochu Eric ebrochu@cs.ubc.ca
University of British COlumbia
Nando de Freitas nando@cs.ubc.ca
University of British Columbia
Abhijeet Ghosh ghosh@cs.ubc.ca
University of British Columbia

We propose an active learning algorithm that learns a continuous valuation model from discrete preferences. The algorithm automatically decides what items are best presented to an individual in order to find the item that they value highly in as few trials as possible, and exploits quirks of human psychology to minimize time and cognitive burden. To do this, our algorithm maximizes the expected improvement at each query without accurately modelling the entire valuation surface, which would be needlessly expensive. The problem is particularly difficult because the space of choices is infinite. We demonstrate the effectiveness of the new algorithm compared to related active learning methods. We also embed the algorithm within a decision making tool for assisting digital artists in rendering materials. The tool finds the best parameters while minimizing the number of queries.

26 *The Tradeoffs of Large Scale Learning*

Olivier Bousquet obousquet@gmail.com
Google

Leon Bottou leon@bottou.org
NEC Labs America

This contribution develops a theoretical framework that takes into account the effect of approximate optimization on learning algorithms. The analysis shows distinct tradeoffs for the case of small-scale and large-scale learning problems. Small-scale learning problems are subject to the usual approximation–estimation tradeoff. Large-scale learning problems are subject to a qualitatively different tradeoff involving the computational complexity of the underlying optimization algorithms in non-trivial ways.

27 *Learning Bounds for Domain Adaptation*

John Blitzer blitzer@cis.upenn.edu
University of Pennsylvania
Koby Crammer crammer@cis.upenn.edu
University of Pennsylvania
Alex Kulesza kulesza@cis.upenn.edu
University of Pennsylvania
Fernando Pereira pereira@cis.upenn.edu
Computer and Information Science, University of Pennsylvania
Jennifer Wortman wortmanj@seas.upenn.edu
University of Pennsylvania

Empirical risk minimization offers well-known learning guarantees when training and test data come from the same domain. In the real world, though, we often wish to adapt a classifier from a source domain with a large amount of training data to different target domain with very little training data. In this work we give uniform convergence bounds for algorithms that minimize a convex combination of source and target empirical risk. The bounds explicitly model the inherent trade-off between training on a large but inaccurate source data set and a small but accurate target training set. Our theory also gives results when we have multiple source domains, each of which may have a different number of instances, and we exhibit cases in which minimizing a non-uniform combination of source risks can achieve much lower target error than standard empirical risk minimization.

28 *Stability Bounds for Non-i.i.d. Processes*

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

The notion of algorithmic stability has been used effectively in the past to derive tight generalization bounds. A key advantage of these bounds is that they are de- signed for specific learning algorithms, exploiting their particular properties. But, as in much of learning theory, existing stability analyses and bounds apply only in the scenario where the samples are independently and identically distributed (i.i.d.). In many machine learning applications, however, this assumption does not hold. The observations received by the learning algorithm often have some inherent temporal dependence, which is clear in system diagnosis or time series prediction problems. This paper studies the scenario where the observations are drawn from a station- ary beta-mixing sequence, which implies a dependence between observations that weaken over time. It proves novel stability-based generalization bounds that hold even with this more general setting. These bounds strictly generalize the bounds given in the i.i.d. case. We also illustrate their application in the case of several general classes of learning algorithms, including Support Vector Regression and Kernel Ridge Regression.

29 *Learning Monotonic Transformations for Classification*

Andrew Howard ahoward@cs.columbia.edu
Columbia University
Tony Jebara jebara@cs.columbia.edu
Columbia University

A discriminative method is proposed for learning monotonic transformations of the training data jointly while estimating a large-margin classifier. Fixed monotonic transformations can be useful as a preprocessing step for many domains such as document classification, image histogram classification and gene microarray experiments. However, most classifiers only explore transformations through manual trial and error or via prior domain knowledge. The proposed method learns monotonic transformations automatically while training a large-margin classifier without any prior knowledge of the domain at hand. A monotonic piecewise linear function is learned which transforms data for subsequent processing by a linear hyperplane classifier. Two algorithmic implementations of the method are formalized. The first performs an alternating sequence of quadratic and linear programs to convergence until it obtains a locally optimal solution. A second algorithm is also provided using a convex semidefinite relaxation that overcomes initialization issues in the initial optimization problem. The effectiveness of these learned transformations on synthetic problems, text data and image data is demonstrated.

Spotlight presentation, Monday, 8:10PM.

30 *Compressed Regression*

Shuheng Zhou szhou@cs.cmu.edu
Carnegie Mellon University, Computer Science Department
John Lafferty lafferty@cs.cmu.edu
Carnegie Mellon
Larry Wasserman larry@stat.cmu.edu
Carnegie Mellon

Recent research has studied the role of sparsity in high dimensional regression and signal reconstruction, establishing theoretical limits for recovering sparse models from sparse data. In this paper we study a variant of this problem where the original n input variables are compressed by a random linear transformation to $m \ll n$ examples in p dimensions, and establish conditions under which a sparse linear model can be successfully recovered from the compressed data. A primary motivation for this compression procedure is to anonymize the data and preserve privacy by revealing little information about the original data. We characterize the number of random projections that are required for ℓ_1-regularized compressed regression to identify the nonzero coefficients in the true model with probability approaching one, a property called "sparsistence." In addition, we show that ℓ_1-regularized compressed regression asymptotically predicts as well as an oracle linear model, a property called "persistence." Finally, we characterize the privacy properties of the compression procedure in information-theoretic terms, establishing upper bounds on the rate of information communicated between the compressed and uncompressed data that decay to zero.

31 *Statistical Analysis of Semi-Supervised Regression*

John Lafferty lafferty@cs.cmu.edu
Carnegie Mellon
Larry Wasserman larry@stat.cmu.edu
Carnegie Mellon

Semi-supervised methods use unlabeled data in addition to labeled data to construct predictors. While existing semi-supervised methods have shown some promising empirical performance, their development has been based largely based on heuristics. In this paper we study semi-supervised learning from the viewpoint of minimax theory. Our first result shows that some common methods based on manifold regularization and graph Laplacians do not lead to faster minimax rates of convergence. Thus, the estimators that use the unlabeled data do not have smaller risk than the estimators that use only labeled data. We then develop several new approaches that provably lead to improved performance. The statistical tools of minimax analysis are thus used to offer some new perspective on the problem of semi-supervised learning.

Spotlight presentation, Monday, 8:10PM.

32 *Boosting the Area under the ROC Curve*

Phil Long plong@google.com
Google
Rocco Servedio rocco@cs.columbia.edu
Department of Computer Science, Columbia University

We show that any weak ranker that can achieve an area under the ROC curve slightly better than 1/2 (which can be achieved by random guessing) can be efficiently boosted to achieve an area under the ROC curve arbitrarily close to 1. We further show that this boosting can be performed even in the presence of independent misclassification noise, given access to a noise-tolerant weak ranker.

Spotlight presentation, Monday, 8:10PM.

33 *Optimal ROC Curve for a Combination of Classifiers*

Marco Barreno barreno@cs.berkeley.edu
University of California Berkeley
Alvaro Cardenas cardenas@cs.berkeley.edu
University of California, Berkeley
Doug Tygar doug.tygar@gmail.com
University of California, Berkeley

We present a new analysis for the combination of binary classifiers. Our analysis makes use of the Neyman-Pearson lemma as a theoretical basis to analyze combinations of classifiers. In particular, we give a method for finding the optimal decision rule for a combination of classifiers and prove that it has the optimal ROC curve. We show how our method generalizes and improves previous work on combining classifiers and generating ROC curves.

34 *Catching Change-points with Lasso*

Zaid Harchaoui zharchao@enst.fr
GET/Telecom Paris
Celine Levy-LeDuc levyledu@enst.fr
GET/Telecom, CNRS/LTCI

We propose a new approach for dealing with the estimation of the location of change-points in one-dimensional piecewise constant signals observed in white noise. Our approach consists in reframing this task in a variable selection context. We use a penalized least-squares criterion with a l1-type penalty for this purpose. We prove that, in an appropriate asymptotic framework, this method provides consistent estimators of the change-points. Then, we explain how to implement this method in practice by combining the LAR algorithm and a reduced version of the dynamic programming algorithm and we apply it to synthetic and real data.

35 *A New View of Automatic Relevance Determination*

David Wipf david.wipf@mrsc.ucsf.edu
Biomagnetic Imaging Lab, University of California, San Francisco
Srikantan Nagarajan sri.nagarajan@mrsc.ucsf.edu
Biomagnetic Imaging Lab, University of California, San Francisco

Automatic relevance determination (ARD), and the closely-related sparse Bayesian learning (SBL) framework, are effective tools for pruning large numbers of irrelevant features. However, popular update rules used for this process are either prohibitively slow in practice and/or heuristic in nature without proven convergence properties. This paper furnishes an alternative means of optimizing a general ARD cost function using an auxiliary function that can naturally be solved using a series of re-weighted L1 problems. The result is an efficient algorithm that can be implemented using standard convex programming toolboxes and is guaranteed to converge to a stationary point unlike existing methods. The analysis also leads to additional insights into the behavior of previous ARD updates as well as the ARD cost function. For example, the standard fixed-point updates of MacKay (1992) are shown to be iteratively solving a particular min-max problem, although they are not guaranteed to lead to a stationary point. The analysis also reveals that ARD is exactly equivalent to performing MAP estimation using a particular feature- and noise-dependent *non-factorial* weight prior with several desirable properties over conventional priors with respect to feature selection. In particular, it provides a tighter approximation to the L0 quasi-norm sparsity measure than the L1 norm. Overall these results suggests alternative cost functions and update procedures for selecting features and promoting sparse solutions.

36 *Support Vector Machine Classification with Indefinite Kernels*

Ronny Luss rluss@princeton.edu
ORFE, Princeton
Alexandre d'Aspremont aspremon@princeton.edu
ORFE, Princeton University

In this paper, we propose a method for support vector machine classification using indefinite kernels. Instead of directly minimizing or stabilizing a nonconvex loss function, our method simultaneously finds the support vectors and a proxy kernel matrix used in computing the loss. This can be interpreted as a robust classification problem where the

indefinite kernel matrix is treated as a noisy observation of the true positive semidefinite kernel. Our formulation keeps the problem convex and relatively large problems can be solved efficiently using the analytic center cutting plane method. We compare the performance of our technique with other methods on several data sets.

37 *Discriminative Keyword Selection Using Support Vector Machines*

William Campbell wcampbell@ll.mit.edu
MIT Lincoln Laboratory
Fred Richardson frichard@ll.mit.edu
MIT Lincoln Laboratory

Many tasks in speech processing involve classification of long term characteristics of a speech segment such as language, speaker, dialect, or topic. A natural technique for determining these characteristics is to first convert the input speech into a sequence of tokens such as words, phones, etc. From these tokens, we can then look for distinctive phrases, keywords, that characterize the speech. In many applications, a set of distinctive keywords may not be known a priori. In this case, an automatic method of building up keywords from short context units such as phones is desirable. We propose a method for construction of keywords based upon Support Vector Machines. We cast the problem of keyword selection as a feature selection problem for n-grams of phones. We propose an alternating filter-wrapper method that builds successively longer keywords. Application of this method on a language recognition task shows that the technique produces interesting and significant qualitative and quantitative results.

38 *Convex Learning with Invariances*

Choon Hui Teo choonhui.teo@gmail.com
Statistical Machine Learning Program, NICTA, Australian National University
Amir Globerson gamir@csail.mit.edu
CSAIL, MIT
Sam Roweis roweis@cs.toronto.edu
University of Toronto
Alex Smola alex.smola@gmail.com
NICTA

Incorporating invariances into a learning algorithm is a common problem in machine learning. We provide a convex formulation which can deal with arbitrary loss functions and arbitrary losses. In addition, it is a drop-in replacement for most optimization algorithms for kernels, including solvers of the SVMStruct family. The advantage of our setting is that it relies on column generation instead of modifying the underlying optimization problem directly.

Spotlight presentation, Monday, 8:10PM.

39 *Estimating divergence functionals and the likelihood ratio by penalized convex risk minimization*

XuanLong Nguyen xuanlong.nguyen@duke.edu
SAMSI, Duke University

MARTIN WAINWRIGHT wainwrig@eecs.berkeley.edu
Department of EECS, Department of Statistics
MICHAEL JORDAN jordan@eecs.berkeley.edu
University of California, Berkeley

We develop and analyze an algorithm for nonparametric estimation of divergence functionals and the density ratio of two probability distributions. Our method is based on a non-asymptotic variational characterization of f-divergences, which turns the estimation problem into a penalized convex risk minimization problem. We present a derivation of our kernel-based estimation algorithm and an analysis of convergence rates for the estimator. Our simulation results demonstrate the convergence behavior of our method, which compares favorably with existing methods in the literature.

Spotlight presentation, Monday, 8:10PM.

40 *Learning with Transformation Invariant Kernels*

OLIVIER CHAPELLE chapelle@mac.com
Yahoo Research
CHRISTIAN WALDER christian.walder@tuebingen.mpg.de
Max Planck Institute

This paper considers kernels invariant to translation, rotation and dilation. We show that no non-trivial positive definite (p.d.) kernels exist which are radial and dilation invariant, only conditionally positive definite (c.p.d.) ones. Accordingly, we discuss the c.p.d. case and provide some novel analysis, including an elementary derivation of a c.p.d. representer theorem. On the practical side, we give a support vector machine (s.v.m.) algorithm for arbitrary c.p.d. kernels. For the thinplate kernel this leads to a classifier with only one parameter (the amount of regularisation), which we demonstrate to be as effective as an s.v.m. with the Gaussian kernel, even though the Gaussian involves a second parameter (the length scale).

Spotlight presentation, Monday, 8:10PM.

41 *Using Deep Belief Nets to Learn Covariance Kernels for Gaussian Processes*

RUSLAN SALAKHUTDINOV rsalakhu@cs.utoronto.ca
University of Toronto, Department of Computer Science
GEOFFREY HINTON hinton@cs.toronto.edu
University of Toronto

We show how to use unlabeled data and a deep belief net (DBN) to learn a good covariance kernel for a Gaussian process. We first learn a deep generative model of the unlabeled data using the fast, greedy algorithm introduced by Hinton et.al. If the data is high-dimensional and highly-structured, a Gaussian kernel applied to the top layer of features in the DBN works much better than a similar kernel applied to the raw input. Performance at both regression and classification can then be further improved by using backpropagation through the DBN to discriminatively fine-tune the covariance kernel.

42 *Efficient multiple hyperparameter learning for log-linear models*

Chuong Do chuongdo@cs.stanford.edu
Stanford University
Chuan-Sheng Foo csfoo@cs.stanford.edu
Stanford University
Andrew Ng ang@cs.stanford.edu
Stanford University

Using multiple regularization hyperparameters is an effective method for managing model complexity in problems where input features have varying amounts of noise. While algorithms for choosing multiple hyperparameters are often used in neural networks and support vector machines, they are not common in structured prediction tasks, such as sequence labeling or parsing. In this paper, we consider the problem of learning regularization hyperparameters for log-linear models, a class of probabilistic models for structured prediction tasks which includes conditional random fields (CRFs). Using an implicit differentiation trick, we derive an efficient gradient-based method for learning Gaussian regularization priors with multiple hyperparameters. In both simulations and the real-world task of computational RNA secondary structure prediction, we find that multiple hyperparameter learning provides a significant boost in accuracy compared to models learned using only a single regularization hyperparameter.

43 *Structured Learning with Approximate Inference*

Alex Kulesza kulesza@cis.upenn.edu
University of Pennsylvania
Fernando Pereira pereira@cis.upenn.edu
Computer and Information Science, University of Pennsylvania

In many structured prediction problems, the highest-scoring labeling is hard to compute exactly, leading to the use of approximate inference methods. However, when inference is used in a learning algorithm, a good approximation of the score may not be sufficient. We show in particular that learning can fail even with an approximate inference method with rigorous approximation guarantees. There are two reasons for this. First, approximate methods can effectively reduce the expressivity of an underlying model by making it impossible to choose parameters that reliably give good predictions. Second, approximations can respond to parameter changes in such a way that standard learning algorithms are misled. In contrast, we give two positive results in the form of learning bounds for the use of LP-relaxed inference in structured perceptron and empirical risk minimization settings. We argue that without understanding combinations of inference and learning, such as these, that are appropriately compatible, learning performance under approximate inference cannot be guaranteed.

Spotlight presentation, Monday, 8:10PM.

44 *Convex Relaxations of Latent Variable Training*

Yuhong Guo yuhong@cs.ualberta.ca
U. Alberta
Dale Schuurmans dale@cs.ualberta.ca
University of Alberta

We investigate a new, convex relaxation of an expectation-maximization (EM) variant that approximates a standard objective while eliminating local minima. First, a cautionary result is presented, showing that any convex relaxation of EM over hidden variables

must give trivial results if any dependence on the missing values is retained. Although this appears to be a strong negative outcome, we then demonstrate how the problem can be bypassed by using equivalence relations instead of value assignments over hidden variables. In particular, we develop new algorithms for estimating exponential conditional models that only require equivalence relation information over the variable values. This reformulation leads to an exact expression for EM variants in a wide range of problems. We then develop a semidefinite relaxation that yields global training by eliminating local minima.

45 *Efficient Principled Learning of Thin Junction Trees*

Anton Chechetka antonc@cs.cmu.edu
Carnegie Mellon University
Carlos Guestrin guestrin@cs.cmu.edu
Carnegie Mellon University

We present the first truly polynomial algorithm for learning the structure of bounded-treewidth junction trees – an attractive subclass of probabilistic graphical models that permits both the compact representation of probability distributions and efficient exact inference. For a constant treewidth, our algorithm has polynomial time and sample complexity, and provides strong theoretical guarantees in terms of KL divergence from the true distribution. We also present a lazy extension of our approach that leads to very significant speed ups in practice, and demonstrate the viability of our method empirically, on several real world datasets. One of our key new theoretical insights is a method for bounding the conditional mutual information of arbitrarily large sets of random variables with only a polynomial number of mutual information computations on fixed-size subsets of variables, when the underlying distribution can be approximated by a bounded treewidth junction tree.

46 *Linear programming analysis of loopy belief propagation for weighted matching*

Sujay Sanghavi sanghavi@mit.edu
MIT
Dmitry Malioutov dmm@mit.edu
MIT
Alan Willsky willsky@mit.edu
MIT

Loopy belief propagation has been employed in a wide variety of applications with great empirical success, but it comes with few theoretical guarantees. In this paper we investigate the use of the max-product form of belief propagation for weighted matching problems on general graphs. We show that max-product converges to the correct answer if the linear programming (LP) relaxation of the weighted matching problem is tight and does not converge if the LP relaxation is loose. This provides an exact characterization of max-product performance and reveals connections to the widely used optimization technique of LP relaxation. In addition, we demonstrate that max-product is effective in solving practical weighted matching problems in a distributed fashion by applying it to the problem of self-organization in sensor networks.

47 *Fixing Max-Product: Convergent Message Passing Algorithms for MAP LP-Relaxations*

Amir Globerson gamir@csail.mit.edu
CSAIL, MIT
Tommi Jaakkola tommi@csail.mit.edu
Massachusetts Institute of Technology

We present a novel message passing algorithm for approximating the MAP problem in graphical models. The algorithm is similar in structure to max-product but unlike max-product it always converges, and can be proven to find the exact MAP solution in various settings. The algorithm is derived via block coordinate descent in a dual of the LP relaxation of MAP, but does not require any tunable parameters such as step size or tree weights. We also describe a generalization of the method to cluster based potentials. The new method is tested on synthetic and real-world problems, and compares favorably with previous approaches.

48 *Loop Series and Bethe Variational Bounds in Attractive Graphical Models*

Erik Sudderth sudderth@alum.mit.edu
University of California, Berkeley
Martin Wainwright wainwrig@eecs.berkeley.edu
Department of EECS, Department of Statistics
Alan Willsky willsky@mit.edu
MIT

Variational methods are frequently used to approximate or bound the partition or likelihood function of a Markov random field. Methods based on mean field theory are guaranteed to provide lower bounds, whereas certain types of convex relaxations provide upper bounds. In general, loopy belief propagation (BP) provides (often accurate) approximations, but not bounds. We prove that for a class of attractive binary models, the value specified by any fixed point of loopy BP always provides a lower bound on the true likelihood. Empirically, this bound is much better than the naive mean field bound, and requires no further work than running BP. We establish these lower bounds using a loop series expansion due to Chertkov and Chernyak, which we show can be derived as a consequence of the tree reparameterization characterization of BP fixed points.

49 *Local Algorithms for Approximate Inference in Minor-Excluded Graphs*

Kyomin Jung kmjung@mit.edu
MIT
Devavrat Shah devavrat@mit.edu
Assistant Professor

We present a new local approximation algorithm for computing MAP and log-partition function for arbitrary exponential family distribution represented by a finite-valued pairwise Markov random field (MRF), say G. Our algorithm is based on decomposing G into appropriately chosen small components; computing estimates locally in each of these components and then producing a good global solution. We prove that the algorithm can provide approximate solution within arbitrary accuracy when G excludes some finite sized graph as its minor and G has bounded degree: all Planar graphs with bounded

degree are examples of such graphs. The running time of the algorithm is $\Theta(n)$ (n is the number of nodes in G), with constant dependent on accuracy, degree of graph and size of the graph that is excluded as a minor (constant for Planar graphs). Our algorithm for minor-excluded graphs uses the decomposition scheme of Klein, Plotkin and Rao (1993). In general, our algorithm works with any decomposition scheme and provides quantifiable approximation guarantee that depends on the decomposition scheme.

50 *Adaptive Embedded Subgraph Algorithms using Walk-Sum Analysis*

Venkat Chandrasekaran venkatc@mit.edu
Massachusetts Institute of Technology
Jason Johnson jasonj@mit.edu
MIT
Alan Willsky willsky@mit.edu
MIT

We consider the estimation problem in Gaussian graphical models with arbitrary structure. We analyze the Embedded Trees algorithm, which solves a sequence of problems on tractable subgraphs thereby leading to the solution of the estimation problem on an intractable graph. Our analysis is based on the recently developed walk-sum interpretation of Gaussian estimation. We show that non-stationary iterations of the Embedded Trees algorithm using any sequence of subgraphs converge in walk-summable models. Based on walk-sum calculations, we develop adaptive methods that optimize the choice of subgraphs used at each iteration with a view to achieving maximum reduction in error. These adaptive procedures provide a significant speedup in convergence over stationary iterative methods, and also appear to converge in a larger class of models.

51 *CPR for CSPs: A Probabilistic Relaxation of Constraint Propagation*

Luis E. Ortiz leortiz@ece.uprm.edu
University of Puerto Rico, Mayaguez

This paper proposes constraint propagation relaxation (CPR), a probabilistic approach to classical constraint propagation that provides another view on the whole parametric family of survey propagation algorithms SP(ρ), ranging from belief propagation ($\rho = 0$) to (pure) survey propagation($\rho = 1$). More importantly, the approach elucidates the implicit, but fundamental assumptions underlying SP(ρ), thus shedding some light on its effectiveness and leading to applications beyond k-SAT.

52 *Fast Variational Inference for Large-scale Internet Diagnosis*

John Platt john.platt@microsoft.com
Microsoft Research
Emre Kiciman emrek@microsoft.com
Microsoft Research
David Maltz dmaltz@microsoft.com
Microsoft Research

Web servers on the Internet need to maintain high reliability, but the cause of intermittent failures of web transactions is non-obvious. We use Bayesian inference to diagnose

problems with web services. This diagnosis problem is far larger than any previously attempted: it requires inference of 10^4 possible faults from 10^5 observations. Further, such inference must be performed in less than a second. Inference can be done at this speed by combining a variational approximation and the use of stochastic gradient descent to optimize a variational cost function. We use this fast inference to diagnose a time series of anomalous HTTP requests taken from a real web service. The inference is fast enough to analyze network logs with billions of entries in a matter of hours.

53 *Discovering Weakly-Interacting Factors in a Complex Stochastic Process*

Charlie Frogner frogner@deas.harvard.edu
Harvard University
Avi Pfeffer avi@eecs.harvard.edu
Harvard

Dynamic Bayesian networks are structured representations of stochastic processes. Despite their structure, exact inference in DBNs is generally intractable. One approach to approximate inference involves grouping the variables in the process into smaller factors and keeping independent beliefs over these factors. In this paper we present several techniques for decomposing a dynamic Bayesian network automatically to enable factored inference. We examine a number of features of a DBN that capture different types of dependencies that will cause error in factored inference. An empirical comparison shows that the most useful of these is a heuristic that estimates the mutual information introduced between factors by one step of belief propagation. In addition to features computed over entire factors, for efficiency we explored scores computed over pairs of variables. We present search methods that use these features, pairwise and not, to find a factorization, and compare their results on several datasets. Automatic factorization extends the applicability of factored inference to large, complex models that are undesirable to factor by hand. Moreover, tests on real DBNs show that automatic factorization can achieve significantly lower error in some cases.

Spotlight presentation, Monday, 8:10PM.

54 *On Ranking in Survival Analysis: Bounds on the Concordance Index*

Vikas Raykar vikasraykar@gmail.com
Siemens Medical Solutions, USA
Harald Steck harald.steck@siemens.com
Siemens Medical Solutions, Computer-Aided Diagnosis and Therapy
Balaji Krishnapuram balaji.krishnapuram@siemens.com
Computer Aided Diagnosis & Therapy Group, Siemens Medical Solutions, USA
Cary Dehing-Oberije cary.dehing@maastro.nl
Maastro
Philippe Lambin philippe.lambin@maastro.nl
Maastro

In this paper, we show that classical survival analysis involving censored data can naturally be cast as a ranking problem. The concordance index (CI), which quantifies the quality of rankings, is the standard performance measure for model *assessment* in survival analysis. In contrast, the standard approach to *learning* the popular proportional hazard (PH)

model is based on Cox's partial likelihood. In this paper we devise two bounds on CI–one of which emerges directly from the properties of PH models–and optimize them *directly*. Our experimental results suggest that both methods perform about equally well, with our new approach giving slightly better results than the Cox's method. We also explain why a method designed to maximize the Cox's partial likelihood also ends up (approximately) maximizing the CI.

55 *Sparse Feature Learning for Deep Belief Networks*

Marc'Aurelio Ranzato ranzato@cs.nyu.edu
Courant Institute - New York University
Y-Lan Boureau ylan@cims.nyu.edu
INRIA, Courant Institute - New York University
Yann LeCun yann@cs.nyu.edu
Courant Institute, New York University

Unsupervised learning algorithms aim to discover the structure hidden in the data, and to learn representations that are more suitable as input to a supervised machine than the raw input. Many unsupervised methods are based on reconstructing the input from the representation, while constraining the representation to have certain desirable properties (e.g. low dimension, sparsity, etc). Others are based on approximating density by stochastically reconstructing the input from the representation. We describe a novel and efficient algorithm to learn sparse representations, and compare it theoretically and experimentally with a similar machines trained probabilistically, namely a Restricted Boltzmann Machine. We propose a simple criterion to compare and select different unsupervised machines based on the trade-off between the reconstruction error and the information content of the representation. We demonstrate this method by extracting features from a dataset of handwritten numerals, and from a dataset of natural image patches. We show that by stacking multiple levels of such machines and by training sequentially, high-order dependencies between the input variables can be captured.

56 *Modeling image patches with a directed hierarchy of Markov random fields*

Simon Osindero osindero@cs.toronto.edu
University of Toronto
Geoffrey Hinton hinton@cs.toronto.edu
University of Toronto

We describe an efficient learning procedure for multilayer generative models that combine the best aspects of Markov random fields and deep, directed belief nets. The generative models can be learned one layer at a time and when learning is complete they have a very fast inference procedure for computing a good approximation to the posterior distribution in all of the hidden layers. Each hidden layer has its own MRF whose energy function is modulated by the top-down directed connections from the layer above. To generate from the model, each layer in turn must settle to equilibrium given its top-down input. We show that this type of model is good at capturing the statistics of patches of natural images.

57 *People Tracking with the Laplacian Eigenmaps Latent Variable Model*

Lu Zhengdong zhengdon@csee.ogi.edu
Oregon Health & Science University
Miguel Carreira-Perpinan mcarreira-perpinan@ucmerced.edu
University of California, Merced, School of Engineering
Cristian Sminchisescu crismin@nagoya.uchicago.edu
University of Chicago

Reliably recovering 3D human pose from monocular video requires constraints that bias the estimates towards typical human poses and motions. We define priors for people tracking using a Laplacian Eigenmaps Latent Variable Model (LELVM). LELVM is a probabilistic dimensionality reduction model that naturally combines the advantages of latent variable models—definining a multimodal probability density for latent and observed variables, and globally differentiable nonlinear mappings for reconstruction and dimensionality reduction—with those of spectral manifold learning methods—no local optima, ability to unfold highly nonlinear manifolds, and good practical scaling to latent spaces of high dimension. LELVM is computationally efficient, simple to learn from sparse training data, and compatible with standard probabilistic trackers such as particle filters. We analyze the performance of a LELVM-based probabilistic sigma point mixture tracker in several real and synthetic human motion sequences and demonstrate that LELVM provides sufficient constraints for robust operation in the presence of missing, noisy and ambiguous image measurements.

58 *Configuration Estimates Improve Pedestrian Finding*

Duan Tran ddtran2@uiuc.edu
University of Illinois at Urbana-Champaign
David Forsyth daf@uiuc.edu
University of Illinois at Urbana-Champaign

Fair discriminative pedestrian finders are now available. In fact, these pedestrian finders make most errors on pedestrians in configurations that are uncommon in the training data, for example, mounting a bicycle. This is undesirable. However, the human configuration can itself be estimated discriminatively using structure learning. We demonstrate a pedestrian finder which first finds the most likely human pose in the window using a discriminative procedure trained with structure learning on a small dataset. We then present features (local histogram of oriented gradient and local PCA of gradient) based on that configuration to an SVM classifier. We show, using the INRIA Person dataset, that estimates of configuration significantly improve the accuracy of a discriminative pedestrian finder.

59 *Rapid Inference on a Novel AND/OR graph for Object Detection, Segmentation and Parsing*

Long Zhu lzhu@stat.ucla.edu
UCLA
Hongjiang Zhang hjzhang@microsoft.com
Microsoft Research
Yuanhao Chen yhchen4@ustc.edu
USTC

Chenxi Lin chenxi.lin@microsoft.com
Microsoft Research Asia
Alan Yuille yuille@stat.ucla.edu
UCLA

In this paper we formulate a novel AND/OR graph representation capable of describing the different configurations of deformable articulated objects such as horses. The representation makes use of the *summarization principle* so that lower level nodes in the graph only pass on summary statistics to the higher level nodes. The probability distributions are invariant to position, orientation, and scale. We develop a novel inference algorithm that combined a bottom-up process for proposing configurations for horses together with a top-down process for refining and validating these proposals. The strategy of surround suppression is applied to ensure that the inference time is polynomial in the size of input data. The algorithm was applied to the tasks of detecting, segmenting and parsing horses. We demonstrate that the algorithm is fast and comparable with the state of the art approaches.

60 *Spatial Latent Dirichlet Allocation*

Xiaogang Wang xgwang@csail.mit.edu
Massachusetts Institute of Technology
Eric Grimson welg@csail.mit.edu
Massachusetts Institute of Technology

In recent years, the language model Latent Dirichlet Allocation (LDA), which clusters co-occurring words into topics, has been widely appled in the computer vision field. However, many of these applications have difficulty with modeling the spatial and temporal structure among visual words, since LDA assumes that a document is a "bag-of-words". It is also critical to properly design "words" and "documents" when using a language model to solve vision problems. In this paper, we propose a topic model Spatial Latent Dirichlet Allocation (SLDA), which better encodes spatial structure among visual words that are essential for solving many vision problems. The spatial information is not encoded in the value of visual words but in the design of documents. Instead of knowing the partition of words into documents *a priori*, the word-document assignment becomes a random hidden variable in SLDA. There is a generative procedure, where knowledge of spatial structure can be flexibly added as a prior, grouping visual words which are close in space into the same document. We use SLDA to discover objects from a collection of images, and show it achieves better performance than LDA.

61 *Unsupervised Feature Selection for Accurate Recommendation of High-Dimensional Image Data*

Sabri Boutemedjet sabri.boutemedjet@usherbrooke.ca
Departement d'informatique, Universite de Sherbrooke
Djemel Ziou djemel.ziou@usherbrooke.ca
Dept Info, Universite de Sherbrooke
Nizar Bouguila bouguila@encs.concordia.ca
Concordia University

Content-based image suggestion (CBIS) targets the recommendation of products based on user preferences on the visual content of images. In this paper, we motivate both feature selection and model order identification as two key issues for a successful CBIS. We propose a generative model in which the visual features and users are clustered into separate classes. We identify the number of both user and image classes with the simultaneous selection of relevant visual features using the message length approach. The goal is to ensure an accurate prediction of ratings for multidimensional non-Gaussian and continuous image descriptors. Experiments on a collected data have demonstrated the merits of our approach.

62 *Multiple-Instance Pruning For Learning Efficient Cascade Detectors*

Cha Zhang chazhang@microsoft.com
Microsoft Research
Paul Viola viola@microsoft.com
Microsoft Live Labs

Cascade detectors have been shown to operate extremely rapidly, with high accuracy, and have important applications such as face detection. Driven by this success, cascade earning has been an area of active research in recent years. Nevertheless, there are still challenging technical problems during the training process of cascade detectors. In particular, determining the optimal target detection rate for each stage of the cascade remains an unsolved issue. In this paper, we propose the multiple instance pruning (MIP) algorithm for soft cascades. This algorithm computes a set of thresholds which aggressively terminate computation with no reduction in detection rate or increase in false positive rate on the training dataset. The algorithm is based on two key insights: i) examples that are destined to be rejected by the complete classifier can be safely pruned early; ii) face detection is a multiple instance learning problem. The MIP process is fully automatic and requires no assumptions of probability distributions, statistical independence, or ad hoc intermediate rejection targets. Experimental results on the MIT&CMU dataset demonstrate significant performance advantages.

63 *A Probabilistic Approach to Language Change*

Alexandre Bouchard-Cote bouchard@cs.berkeley.edu
UC Berkeley
Percy Liang pliang@cs.berkeley.edu
UC Berkeley
Thomas Griffiths tom_griffiths@berkeley.edu
UC Berkeley
Dan Klein klein@cs.berkeley.edu
UC Berkeley

We present a probabilistic approach to language change in which word forms are represented by phoneme sequences that undergo stochastic edits along the branches of a phylogenetic tree. Our framework combines the advantages of the classical comparative method with the robustness of corpus-based probabilistic models. We use this framework to explore the consequences of two different schemes for defining probabilistic models of phonological change, evaluating these schemes using the reconstruction of ancient word forms in Romance languages. The result is an efficient inference procedure for automatically inferring ancient word forms from modern languages, which can be generalized to support inferences about linguistic phylogenies.

64 *Discriminative Log-Linear Grammars with Latent Variables*

Slav Petrov petrov@cs.berkeley.edu
University of California, Berkeley
Dan Klein klein@cs.berkeley.edu
UC Berkeley

We demonstrate that log-linear grammars with latent variables can be practically trained using discriminative methods. Central to efficient discriminative training is a hierarchical pruning procedure which allows feature expectations to be efficiently approximated in a gradient-based procedure. We compare L1 and L2 regularization and show that L1 regularization is superior, requiring fewer iterations to converge, and yielding sparser solutions. On full-scale treebank parsing experiments, the discriminative latent models outperform both the comparable generative latent models as well as the discriminative non-latent baselines.

Spotlight presentation, Monday, 8:10PM.

65 *A Bayesian LDA-based model for semi-supervised part-of-speech tagging*

Kristina Toutanova kristout@microsoft.com
Microsoft Research
Mark Johnson mark_johnson@brown.edu
Cognitive and Linguistic Sciences, Box 1978

We present a novel Bayesian statistical model for semi-supervised part-of-speech tagging. Our model extends the Latent Dirichlet Allocation (LDA) model and incorporates the intuition that words' distributions over tags, $p(t|w)$, are sparse. In addition we introduce a model for determining the set of possible tags of a word which captures important dependencies in the ambiguity classes of words. Our model outperforms the best previously proposed model for this task on a standard dataset.

Spotlight presentation, Monday, 8:10PM.

66 *HM-BiTAM: Bilingual Topic Exploration, Word Alignment, and Translation*

Bing Zhao bzhao@cs.cmu.edu
Carnegie Mellon University
Eric P. Xing epxing@cs.cmu.edu
School of Computer Science

We present a novel paradigm for statistical machine translation (SMT), based on joint modeling of word alignment and the topical aspects underlying bilingual document pairs via a hidden Markov Bilingual Topic AdMixture (HM-BiTAM). In this new paradigm, parallel sentence-pairs from a parallel document-pair are coupled via a certain semantic-flow, to ensure coherence of topical context in the alignment of matching words between languages, during likelihood-based training of topic-dependent translational lexicons, as well as topic representations in each language. The resulting trained HM-BiTAM can not only display topic patterns like other methods such as LDA, but now for bilingual corpora; it also offers a principled way of inferring optimal translation in a context-dependent way. Our method integrates the conventional IBM Models based on HMM — a key component for most of the state-of-the-art SMT systems, with the recently proposed BiTAM

model, and we report an extensive empirical analysis (in many way complementary to the description-oriented of our method in three aspects: word alignment, bilingual topic representation, and translation.

67 *Supervised Topic Models*

David Blei blei@cs.princeton.edu
Princeton University
Jon McAuliffe jon@mcauliffe.com
Statistics Department, University of Pennsylvania, Wharton School

We introduce supervised latent Dirichlet allocation (sLDA), a statistical model of labelled documents. The model accommodates a variety of response types. We derive a maximum-likelihood procedure for parameter estimation, which relies on variational approximations to handle intractable posterior expectations. Prediction problems motivate this research: we use the fitted model to predict response values for new documents. We test sLDA on two real-world problems: movie ratings predicted from reviews, and web page popularity predicted from text descriptions. We illustrate the benefits of sLDA versus modern regularized regression, as well as versus an unsupervised LDA analysis followed by a separate regression.

68 *Mining Internet-Scale Software Repositories*

Erik Linstead elinstea@ics.uci.edu
University of California, Irvine
Paul Rigor prigor@ics.uci.edu
Donald Bren School of ICS, UC Irvine
Sushil Bajracharya sbajrach@ics.uci.edu
University of California, Irvine
Cristina Lopes lopes@ics.uci.edu
University of California Irvine
Pierre Baldi pfbaldi@ics.uci.edu
School of Information and Computer Sciences, University of California, Irvine

Large repositories of source code create new challenges and opportunities for statistical machine learning. Here we first develop an infrastructure for the automated crawling, parsing, and database storage of open source software. The infrastructure allows us to gather Internet-scale source code. For instance, in one experiment, we gather 4,632 java projects from SourceForge and Apache totaling over 38 million lines of code from 9,250 developers. Simple statistical analyses of the data first reveal robust power-law behavior for package, SLOC, and method call distributions. We then develop and apply unsupervised author-topic, probabilistic models to automatically discover the topics embedded in the code and extract topic-word and author-topic distributions. In addition to serving as a convenient summary for program function and developer activities, these and other related distributions provide a statistical and information-theoretic basis for quantifying and analyzing developer similarity and competence, topic scattering, and document tangling, with direct applications to software engineering. Finally, by combining software textual content with structural information captured by our CodeRank approach, we are able to significantly improve software retrieval performance, increasing the AUC metric to 0.86– roughly 10-30% better than previous approaches based on text alone.

69 *Evaluating Search Engines by Modeling the Relationship Between Relevance and Clicks*

Ben Carterette carteret@cs.umass.edu
University of Massachusetts Amherst
Rosie Jones jonesr@yahoo-inc.com
Yahoo Research

We propose a model that leverages the millions of clicks received by web search engines, to predict document relevance. This allows the comparison of ranking functions when clicks are available but complete relevance judgments are not. After an initial training phase using a set of relevance judgments paired with click data, we show that our model can predict the relevance score of documents that have not been judged. These predictions can be used to evaluate the performance of a search engine, using our novel formalization of the confidence of the standard evaluation metric discounted cumulative gain (DCG), so comparisons can be made across time and datasets. This contrasts with previous methods which can provide only pair-wise relevance judgements between results shown for the same query. When no relevance judgments are available, we can identify the better of two ranked lists up to 82% of the time, and with only two relevance judgments for each query, we can identify the better ranking up to 94% of the time. While our experiments are on sponsored search results, which is the financial backbone of web search, our method is general enough to be applicable to algorithmic web search results as well. Furthermore, we give an algorithm to guide the selection of additional documents to judge to improve confidence.

70 *Automatic Generation of Social Tags for Music Recommendation*

Douglas Eck eckdoug@iro.umontreal.ca
University of Montreal
Paul Lamere paul.lamere@sun.com
Sun MIcrosystems
Thierry Bertin-Mahieux thierry.bertin-mahieux@umontreal.ca
University of Montreal
Stephen Green stephen.green@sun.com
Sun MIcrosystems

Social tags are user-generated keywords associated with some resource on the Web. In the case of music, social tags have become an important component of "Web2.0" recommender systems, allowing users to generate playlists based on use-dependent terms such as "chill" or "jogging" that have been applied to particular songs. In this paper, we propose a method for predicting these social tags directly from MP3 files. Using a set of boosted classifiers, we map audio features onto social tags collected from the Web. The resulting automatic tags (or "autotags") furnish information about music that is otherwise untagged or poorly tagged, allowing for insertion of previously unheard music into a social recommender. This avoids the "cold-start problem" common in such systems. Autotags can also be used to smooth the tag space from which similarities and recommendations are made by providing a set of comparable baseline tags for all tracks in a recommender system.

Demos:

1 *Automatic Cameraman*

Yoav Freund
University of California, San Diego
Evan Ettinger
University of California, San Diego
Brian McFee
University of California, San Diego
Deborah Goshorn
University of California, San Diego
Shankar Shivappa
University of California, San Diego

http://www.cse.ucsd.edu/˜yfreund/

2 *Building a 3-D Model From a Single Still Image*

Ashutosh Saxena
Stanford University
Min Sun
Stanford University
Andrew Ng
Stanford University

http://make3d.stanford.edu/

3 *Contraction of VLSI Spiking Neurons*

Emre Neftci
Institute of Neuroinformatics, UNI-ETH Zurich
Elisabetta Chicca
Institute of Neuroinformatics, UNI-ETH Zurich
Giacomo Indiveri
Institute of Neuroinformatics, UNI-ETH Zurich
Jean-Jeacques Slotine
MIT
Rodney Douglas
Institute of Neuroinformatics, UNI-ETH Zurich

4 *Elefant*

Kishor Gawande
National ICT Australia
Alex Smola
National ICT Australia
Vishwanathan S V N
National ICT Australia
Li Cheng
National ICT Australia

Simon Guenter
National ICT Australia

http://elefant.developer.nicta.com.au/

5 *Gender and Age Recognition*

Wei Xu
NEC Labs America
Kai Yu
NEC Labs America
Yihong Gong
NEC Labs America

6 *Learning To Race by Model-Based Reinforcement Learning with Adaptive Abstraction*

Thore Graepel
Microsoft Research
Phil Trelford
Microsoft Research
Ralf Herbrich
Microsoft Research
Mykel Kochenderfer
Massachusetts Institute of Technology

http://research.microsoft.com/mlp/apg/

7 *Predicting Human Gaze Using Low-level Saliency Combined with Face Detection*

Moran Cerf
California Institute of Technology
Christof Koch
California Institute of Technology

8 *Robotic Eye Model with Learning of Pulse-Step Saccades*

Per-Erik Forssen
University of British Columbia
Dinesh Pai
University of British Columbia

Tuesday, December 4th

Oral Session — Structured Statistical Models (Chair: Michael Black, Brown University):

8:30AM Invited Talk: *Statistical Models for Social Networks with Application to HIV Epidemiology*

Mark Handcock handcock@stat.washington.edu
University of Washington

We review statistical exponential family models that recognize the complex dependencies within relational data structures. Such models are being used to modeling the contact network that underlies infectious disease transmission. These models make it possible to represent key structural parameters of real networks, and then use these parameters to simulate disease spread across a dynamic network with the observed structural features. In this talk we review model inference from complete and sampled network data. To represent a dynamic network, both the structural regularities in the network and the dynamics of partnership formation and dissolution must be addressed. We demonstrate such a network model and apply the method to HIV spread in sub-Saharan Africa.

9:30AM *Probabilistic Matrix Factorization*

Ruslan Salakhutdinov rsalakhu@cs.utoronto.ca
University of Toronto, Department of Computer Science
Andriy Mnih amnih@cs.toronto.edu
University of Toronto

Many existing approaches to collaborative filtering can neither handle very large datasets nor easily deal with users who have very few ratings. In this paper we present the Probabilistic Matrix Factorization (PMF) model which scales linearly with the number of observations and, more importantly, performs well on the large, sparse, and very imbalanced Netflix dataset. We further extend the PMF model to include an adaptive prior on the model parameters and show how the model capacity can be controlled automatically. Finally, we introduce a constrained version of the PMF model that is based on the assumption that users who have rated similar sets of movies are likely to have similar preferences. The resulting model is able to generalize considerably better for users with very few ratings. When the predictions of multiple PMF models are linearly combined with the predictions of Restricted Boltzmann Machines models, we achieve an error rate of 0.8861, that is nearly 7% better than the score of Netflix's own system.

9:50AM Spotlights

Infinite State Bayes-Nets for Structured Domains
Max Welling, University of California Irvine, Ian Porteous, University of California Irvine and Evgeniy Bart, Cal Tech.
See abstract, page 117.

Hidden Common Cause Relations in Relational Learning
RICARDO SILVA, Gatsby Computational Neuroscience Unit, UCL, WEI CHU, Center for Computational Learning Systems, Columbia University and ZOUBIN GHAHRAMANI, University of Cambridge & CMU.
See abstract, page 120.

COFI RANK - Maximum Margin Matrix Factorization for Collaborative Ranking
MARKUS WEIMER, TU Darmstadt, ALEXANDROS KARATZOGLOU, Vienna University of Technology, Statistics, QUOC LE, Statistical Machine Learning Program, National ICT Australia and ALEX SMOLA, NICTA.
See abstract, page 104.

SpAM: Sparse Additive Models
JOHN LAFFERTY, Carnegie Mellon, HAN LIU, Machine Learning Department, Carnegie Mellon University, PRADEEP RAVIKUMAR, Carnegie Mellon University and LARRY WASSERMAN, Carnegie Mellon.
See abstract, page 121.

TrueSkill Through Time: Revisiting the History of Chess
PIERRE DANGAUTHIER, INRIA Rhones-Alpes, RALF HERBRICH, Microsoft Research, Applied Games, TOM MINKA, Microsoft Research Ltd and THORE GRAEPEL, Microsoft Research Cambridge.
See abstract, page 114.

Heterogeneous Component Analysis
SHIGEYUKI OBA, Nara Institute of Science and Technology, MOTOAKI KAWANABE, Fraunhofer FIRST.IDA, KLAUS-ROBERT MÜLLER, Fraunhofer FIRST.IDA and SHIN ISHII, Dept. Syst. Sci. Grad. Sch. Info. Kyoto Univ., Nare Institute of Science and Technology.
See abstract, page 105.

Density Estimation under Independent Similarly Distributed Sampling Assumptions
TONY JEBARA, Columbia University, YINGBO SONG, Columbia University and KAPIL THADANI, Columbia University.
See abstract, page 118.

Semi-Supervised Multitask Learning
QIUHUA LIU, Duke University, XUEJUN LIAO, Duke University and LAWRENCE CARIN, Duke University.
See abstract, page 107.

10:00AM Break

Oral Session — Probabilistic Optimization (Chair: Francis Bach, Ecole Normale Superieure):

10:30AM *Colored Maximum Variance Unfolding*

LE SONG lesong@it.usyd.edu.au
NICTA and School of Information Technologies, the University of Sydney
ALEX SMOLA alex.smola@gmail.com
NICTA
KARSTEN BORGWARDT kmb51@cam.ac.uk
University of Cambridge
ARTHUR GRETTON arthur.gretton@tuebingen.mpg.de
MPI for Biological Cybernetics

Maximum variance unfolding (MVU) is an effective heuristic for dimensionality reduction. It produces a low-dimensional representation of the data by maximizing the variance of their embeddings while preserving the local distances of the original data. We show that MVU also optimizes a statistical dependence measure which aims to retain the identity of individual observations under the distance-preserving constraints. This general view allows us to design "colored" variants of MVU, which produce low-dimensional representations for a given task, e.g. subject to class labels or other side information.

10:50AM *An Analysis of Convex Relaxations for MAP Estimation*

PAWAN MUDIGONDA pkmudigonda@brookes.ac.uk
Oxford Brookes University
VLADIMIR KOLMOGOROV vnk@adastral.ucl.ac.uk
University College London
PHILIP TORR philiptorr@brookes.ac.uk
Oxford Brookes University

The problem of obtaining the maximum a posteriori estimate of a given Markov random field is known to be NP-hard in general. However, due to its central importance in many applications, several approximate algorithms have been proposed in the literature. In this paper, we present an analysis of three such algorithms based on convex relaxations: (i) LP-S: the linear programming (LP) relaxation proposed by Schlesinger for a special case and independently by Chekuri et al., Koster et al. and Wainwright et al. for the general case; (ii) QP-RL: the quadratic programming (QP) relaxation by Ravikumar and Lafferty; and (iii) SOCP-MS: the second order cone programming (SOCP) relaxation first proposed by Muramatsu and Suzuki for two label problems and later extended by Kumar et al. for a general label set. Specifically, we show that the SOCP-MS and QP-RL relaxations are equivalent. Furthermore, we prove that despite the flexibility in the form of the constraints and the objective function offered by QP and SOCP, the LP-S relaxation dominates (i.e. provides a better approximation than) QP-RL and SOCP-MS. We generalize these results by defining a large class of SOCP (and equivalent QP) relaxations which are dominated by the LP-S relaxation. Based on these results we propose some novel SOCP constraints which dominate the previous approaches. We develop efficient algorithms for solving the resulting relaxations and show that the empirical results conform with our analysis.

11:10AM *New Outer Bounds on the Marginal Polytope*

David Sontag dsontag@csail.mit.edu
Massachusetts Institute of Technology
Tommi Jaakkola tommi@csail.mit.edu
Massachusetts Institute of Technology

We give a new class of outer bounds on the marginal polytope, and propose a cutting-plane algorithm for efficiently optimizing over these constraints. When combined with a concave upper bound on the entropy, this gives a new variational inference algorithm for probabilistic inference in discrete Markov Random Fields (MRFs). Valid constraints on the marginal polytope are derived through a series of projections onto the cut polytope. As a result, we obtain tighter upper bounds on the log-partition function. We also show empirically that our approximations of the marginals are significantly more accurate. Moreover, we demonstrate the advantage from the new constraints in finding the MAP assignment for protein structure prediction.

11:30AM *Scene Segmentation with CRFs Learned from Partially Labeled Images*

Jakob Verbeek jakob.verbeek@inria.fr
INRIA Rhone-Alpes, Laboratoire Jean Kuntzmann
Bill Triggs bill.triggs@inrialpes.fr
INRIA France

Conditional Random Fields (CRFs) are an effective tool for a variety of different data segmentation and labelling tasks including visual scene interpretation, which seeks to partition images into their constituent semantic-level regions and assign appropriate class labels to each region. For accurate labelling it is important to capture the global context of the image as well as local information. We introduce a CRF based scene labelling model that incorporates both local features and features aggregated over the whole image or large sections of it. Secondly, traditional CRF learning requires fully labelled datasets. Complete labellings are typically costly and troublesome to produce. We introduce an algorithm that allows CRF models to be learned from datasets where a substantial fraction of the nodes are unlabeled. It works by marginalizing out the unknown labels so that the log-likelihood of the known ones can be maximized by gradient ascent. Loopy Belief Propagation is used to approximate the marginals needed for the gradient and log-likelihood calculations and the Bethe free-energy approximation to the log-likelihood is monitored to control the step size. Our experimental results show that incorporating top-down aggregate features significantly improves the segmentations and that effective models can be learned from fragmentary labellings. The resulting methods give scene segmentation results comparable to the state-of-the-art on three different image databases.

11:50AM Spotlights

Collapsed Variational Inference for HDP

Yee Whye Teh, Gatsby Computational Neuroscience Unit, UCL, Kenichi Kurihara, Tokyo Institute of Technology and Max Welling, University of California Irvine.
See abstract, page 116.

Hierarchical Penalization

Marie Szafranski, Heudiasyc - CNRS 6599, Compiegne University of

Technology, YVES GRANDVALET, IDIAP Research Institute, CNRS and PIERRE MORIZET-MAHOUDEAUX, IDIAP Research Institute, CNRS.
See abstract, page 112.

Collective Inference on Markov Models for Modeling Bird Migration
DANIEL SHELDON, Cornell University, M.A. SALEH ELMOHAMED, Cornell University and DEXTER KOZEN, Cornell University.
See abstract, page 114.

The Generalized FITC Approximation
ANDREW NAISH-GUZMAN, University of Cambridge, Computer Laboratory and SEAN HOLDEN, Computer Laboratory, Cambridge University.
See abstract, page 121.

Distributed Inference for Latent Dirichlet Allocation
DAVID NEWMAN, UC Irvine, ARTHUR ASUNCION, University of California, Irvine, PADHRAIC SMYTH, University of California Irvine and MAX WELLING, University of California Irvine.
See abstract, page 117.

Catching Up Faster in Bayesian Model Selection and Model Averaging
TIM VAN ERVEN, Centrum voor Wiskunde en Informatica, PETER GRUNWALD, Centrum voor Wiskunde en Informatica and STEVEN DE ROOIJ, Centrum voor Wiskunde en Informatica (CWI).
See abstract, page 119.

Message Passing for Max-weight Independent Set
SUJAY SANGHAVI, MIT, DEVAVRAT SHAH, Assistant Professor and ALAN WILLSKY, MIT.
See abstract, page 113.

Privacy-Preserving Belief Propagation and Sampling
MICHAEL KEARNS, University of Pennsylvania, JINSONG TAN, University of Pennsylvania and JENNIFER WORTMAN, University of Pennsylvania.
See abstract, page 113.

12:00PM Break

Oral Session — Optimization for Learning (Chair: John Platt, Microsoft Research):

2:00PM Invited Talk: *Projection Methods: Algorithmic Structures, Bregman Projections, and Acceleration Techniques*

Yair Censor yair@math.haifa.ac.il
University of Haifa, Israel

Recognizing that the Perceptron algorithm is a projection method inspires our interest in the class of projection methods. Are there other methods in that class which can be interpreted as, or modified into, machine learning algorithms? What else can methods from this class "do" for us? How do projection methods work? Can they employ different kinds of projections, such as Bregman projections? What are the algorithmic structures available in the class of projection methods? what can be said about convergence properties and/or experimental initial behavior patterns? When, why and how are projection methods preferable over other methods? How can they be accelerated? We will touch upon these questions and explain why this class of methods has witnessed great progress in recent years. Some significant real-world applications benefit from the use of projection methods. We will briefly describe how the fully-discretized model in image reconstruction from projections and the inverse problem of intensity-modulated radiation therapy (IMRT) lend themselves to such methods. Finally, recent work in machine learning is being cross-fertilized by various projection methods.

3:00PM *Adaptive Online Gradient Descent*

Peter Bartlett bartlett@cs.berkeley.edu
UC Berkeley
Alexander Rakhlin rakhlin@gmail.com
UC Berkeley
Elad Hazan ehazan@cs.princeton.edu
Princeton University

We study the rates of growth of the regret in online convex optimization. First, we show that a simple extension of the algorithm of Hazan et al eliminates the need for a priori knowledge of the lower bound on the second derivatives of the observed functions. We then provide an algorithm, Adaptive Online Gradient Descent, which interpolates between the results of Zinkevich for linear functions and of Hazan et al for strongly convex functions, achieving intermediate rates between $\sqrt{T}$ and $\log T$. Furthermore, we show strong optimality of the algorithm. Finally, we provide an extension of our results to general norms.

3:20PM Spotlights

Consistent Minimization of Clustering Objective Functions
Ulrike von Luxburg, MPI for Biological Cybernetics, Sebastien Bubeck, INRIA futurs, Stefanie Jegelka, MPI for Biological Cybernetics and Michael Kaufmann, Universitat Tubingen.
See abstract, page 103.

McRank: Learning to Rank Using Multiple Classification and Gradient Boosting
PING LI, Cornell University, CHRISTOPHER BURGES, Microsoft Research and QIANG WU, Microsoft Research.
See abstract, page 110.

A Kernel Statistical Test of Independence
ARTHUR GRETTON, MPI for Biological Cybernetics, CHOON HUI TEO, SML, NICTA, KENJI FUKUMIZU, Institute of Statistical Mathematics, LE SONG, NICTA and School of Information Technologies, the University of Sydney, BERNHARD SCHOLKOPF, MPI for Biological Cybernetics and ALEX SMOLA, NICTA.
See abstract, page 123.

Anytime Induction of Cost-sensitive Trees
SAHER ESMEIR, Computer Science Department, Technion-IIT and SHAUL MARKOVITCH, Computer Science Department, Technion-IIT.
See abstract, page 109.

Iterative Non-linear Dimensionality Reduction with Manifold Sculpting
MICHAEL GASHLER, Brigham Young University, DAN VENTURA, Brigham Young University and TONY MARTINEZ, Brigham Young University.
See abstract, page 104.

Random Features for Large-Scale Kernel Machines
ALI RAHIMI, Intel Research and BENJAMIN RECHT, California Institute of Technology.
See abstract, page 108.

Boosting Algorithms for Maximizing the Soft Margin
MANFRED WARMUTH, UC Santa Cruz, KAREN GLOCER, UC Santa Cruz and GUNNAR RATSCH, Friedrich Miescher Laboratory, Max Planck Society.
See abstract, page 102.

Bundle Methods for Machine Learning
ALEX SMOLA, NICTA, S V N VISHWANATHAN, Statistical Machine Learning Program, National ICT Australia and QUOC LE, Statistical Machine Learning, NICTA.
See abstract, page 102.

Classification via Minimum Incremental Coding Length (MICL)
JOHN WRIGHT, University of Illinois at Urbana-Champaign, YANGYU TAO, Microsoft Research, ZHOUCHEN LIN, Microsoft Research, YI MA, Electrical & Computer Engineering Department, University of Illinois at Urbana-Champaign and HEUNG-YEUNG SHUM, Microsoft Research.
See abstract, page 111.

3:30PM Break

Oral Session — Theory and Sequential Decision Making (Chair: Sanjoy Dasgupta, University of California San Diego):

4:00PM *FilterBoost: Regression and Classification on Large Datasets*

JOSEPH K BRADLEY jkbradle@cs.cmu.edu
Carnegie Mellon University
ROBERT SCHAPIRE schapire@cs.princeton.edu
Princeton University

We study boosting in the filtering setting, where the booster draws examples from an oracle instead of using a fixed training set and so may train efficiently on very large datasets. Our algorithm, which is based on a logistic regression technique proposed by Collins, Schapire, & Singer, represents the first boosting-by-filtering algorithm which is truly adaptive and does not need the less realistic assumptions required by previous work. Moreover, we give the first proof that the algorithm of Collins et al. is a strong PAC learner, albeit within the filtering setting. Our proofs demonstrate the algorithm's strong theoretical properties for both classification and conditional probability estimation, and we validate these results through extensive experiments. Empirically, our algorithm proves more robust to noise and overfitting than batch boosters in conditional probability estimation and proves competitive in classification.

4:20PM *The Price of Bandit Information for Online Optimization*

VARSHA DANI varsha@cs.uchicago.edu
University of Chicago
THOMAS HAYES hayest@tti-c.org
Toyota Technological Institute at Chicago
SHAM KAKADE sham@tti-c.org
Toyota Technological Institute

In the online linear optimization problem, a learner must choose, in each round, a decision from a set $D \subset R^n$ in order to minimize an (unknown and changing) linear cost function. We present sharp rates of convergence (with respect to additive regret) for both the full information setting (where the cost function is revealed at the end of each round) and in the bandit setting (where only the scalar cost incurred is revealed). In particular, this paper is concerned with the *price of bandit information* — how much worse the regret is in the bandit case as compared to the full information case. For the full information case, the upper bound on the regret is $O^*(\sqrt{nT})$, where n is the ambient dimension and T is the time horizon. For the bandit case, we present an algorithm which achieves $O^*(n^{3/2}\sqrt{T})$ regret — all previous (nontrivial) bounds here were $O(\text{poly}(n)T^{2/3})$ or worse. It is striking that the convergence rate for the bandit setting is only a factor of n worse than in the full information case — in stark contrast to the K-arm bandit setting, where the gap in the dependence on K is exponential ($\sqrt{TK}$ vs. $\sqrt{T \log K}$). We also present lower bounds showing that this gap is at least $\sqrt{n}$, which we conjecture to be the correct order. The bandit algorithm we present can be implemented efficiently in special cases of particular interest, such as path planning and Markov Decision Problems.

4:40PM *A Game-Theoretic Approach to Apprenticeship Learning*

Umar Syed usyed@cs.princeton.edu
Princeton University
Robert Schapire schapire@cs.princeton.edu
Princeton University

We study the problem of an apprentice learning to behave in an environment with an unknown reward function by observing the behavior of an expert. We follow on the work of Abbeel and Ng who considered a framework in which the true reward function is assumed to be a linear combination of a set of known and observable features. We give a new algorithm that, like theirs, is guaranteed to learn a policy that is nearly as good as the expert's, given enough examples. However, unlike their algorithm, we show that ours may produce a policy that is substantially better than the expert's. Moreover, our algorithm is computationally much faster, is easy to implement, and can be applied even in the absence of an expert. The method is based on a game-theoretic view of the problem, which leads naturally to a direct application of the multiplicative-weights algorithm of Freund and Schapire for playing repeated matrix games. In addition to our formal presentation and analysis of the new algorithm, we sketch how the method can be applied when the transition function itself is unknown, and we provide an experimental demonstration of the algorithm on a toy video-game environment.

5:00PM *Cluster Stability for Finite Samples*

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

Cluster stability has recently received growing attention as a cluster validation criterion in a sample-based framework. However, recent work has shown that as the sample size increases to infinity, any clustering model will usually become asymptotically stable. This led to the conclusion that stability is lacking as a theoretical and practical tool. The discrepancy between this conclusion and the success of stability in practice has remained an open question, which we attempt to address. Our theoretical approach is that stability, as used by cluster validation algorithms, is similar in certain respects to measures of generalization in a model-selection framework. In such cases, the model chosen governs the convergence rate of generalization bounds. By arguing that these rates are more important than the sample size, we are led to the prediction that stability-based cluster validation algorithms should not degrade with increasing sample size, despite the asymptotic universal stability. This prediction is substantiated by a theoretical analysis as well as several empirical results. We conclude that stability remains a meaningful theoretical and practical criterion for cluster validity over finite samples.

5:20PM Spotlights

Reinforcement Learning in Continuous Action Spaces through Sequential Monte Carlo Methods

Alessandro Lazaric, Politecnico di Milano, Marcello Restelli, Politecnico di Milano and Andrea Bonarini, AI&Robotics Lab - Politecnico di Milano.
See abstract, page 98.

Selecting Observations against Adversarial Objectives
Andreas Krause, Computer Science Department, Carnegie Mellon University, Brendan McMahan, Google, Inc., Carlos Guestrin, Carnegie Mellon University and Anupam Gupta, Carnegie Mellon University.
See abstract, page 122.

A general agnostic active learning algorithm
Daniel Hsu, University of California San Diego, Sanjoy Dasgupta, University of California San Diego and Claire Monteleoni, University of California San Diego.
See abstract, page 101.

Managing Power Consumption and Performance of Computing Systems Using Reinforcement Learning
Gerald Tesauro, IBM TJ Watson Research Center, Rajarshi Das, IBM Austin Research Laboratory, Hoi Chan, IBM Austin Research Laboratory, Jeffrey Kephart, IBM Research, David Levine, IBM Austin Research Laboratory, Freeman Rawson, IBM Austin Research Laboratory and Charles Lefurgy, IBM.
See abstract, page 100.

Incremental Natural Actor-Critic Algorithms
Shalabh Bhatnagar, Indian Institute of Science, Richard Sutton, University of Alberta, Mohammad Ghavamzadeh, University of Alberta and Mark Lee, University of Alberta.
See abstract, page 98.

Stable Dual Dynamic Programming
Tao Wang, University of Alberta, Daniel Lizotte, U. Alberta, Michael Bowling, University of Alberta and Dale Schuurmans, University of Alberta.
See abstract, page 97.

Bayes-Adaptive POMDPs
Stephane Ross, McGill University, School of Computer Science, Joelle Pineau, McGill University and Brahim Chaib-draa, Computer Science Department, Laval University.
See abstract, page 97.

What makes some POMDP problems easy to approximate?
David Hsu, National University of Singapore, Wee Sun Lee, National University of Singapore and Nan Rong, Cornell University.
See abstract, page 97.

Hierarchical Apprenticeship Learning with Application to Quadruped Locomotion
J. Zico Kolter, Stanford University, Pieter Abbeel, Stanford University and Andrew Ng, Stanford University.
See abstract, page 100.

5:30PM Break

Posters:

1 *Stable Dual Dynamic Programming*

Tao Wang trysi@cs.ualberta.ca
University of Alberta
Daniel Lizotte dlizotte@cs.ualberta.ca
U. Alberta
Michael Bowling bowling@cs.ualberta.ca
University of Alberta
Dale Schuurmans dale@cs.ualberta.ca
University of Alberta

Recently, we have introduced a novel approach to dynamic programming and reinforcement learning that is based on maintaining explicit representations of stationary distributions instead of value functions. In this paper, we investigate the convergence properties of these dual algorithms both theoretically and empirically, and show how they can be scaled up by incorporating function approximation.

Spotlight presentation, Tuesday, 5:20PM.

2 *Bayes-Adaptive POMDPs*

Stephane Ross stephane.ross@mail.mcgill.ca
McGill University, School of Computer Science
Joelle Pineau jpineau@cs.mcgill.ca
McGill University
Brahim Chaib-draa chaib@damas.ift.ulaval.ca
Computer Science Department, Laval University

Bayesian Reinforcement Learning has generated substantial interest recently, as it provides an elegant solution to the exploration-exploitation trade-off in reinforcement learning. However most investigations of Bayesian reinforcement learning to date focus on the standard Markov Decision Processes (MDPs). Our goal is to extend these ideas to the more general Partially Observable MDP (POMDP) framework, where the state is a hidden variable. To address this problem, we introduce a new mathematical model, the Bayes-Adaptive POMDP. This new model allows one to (1) improve knowledge of the POMDP domain through interaction with the environment, and (2) plan optimal sequences of actions which can trade-off between improving the model, identifying the state, and gathering reward. We show how the model can be finitely approximated while preserving the value function. We describe approximations for belief tracking and planning in this model. Empirical results on two domains show that the model estimate and agent's return improve over time, as the agent learns better model estimates.

Spotlight presentation, Tuesday, 5:20PM.

3 *What makes some POMDP problems easy to approximate?*

David Hsu dyhsu@comp.nus.edu.sg
National University of Singapore
Wee Sun Lee leews@comp.nus.edu.sg
National University of Singapore

Nan Rong nan.rong@gmail.com
Cornell University

Point-based algorithms have been surprisingly successful in computing approximately optimal policies for partially observable Markov decision processes (POMDPs) in high dimensional belief spaces. In this work, we seek to understand the belief-space properties that allow some POMDP problems to be approximated efficiently and thus help to explain the point-based algorithms' success often observed in the experiments. We show that an approximately optimal POMDP solution can be computed in time polynomial in the covering number of a reachable belief space, the subset of the belief space reachable from a given belief point. We also show that under the weaker condition of having a small covering number for an optimal reachable space, the subset of the belief space reachable under an optimal policy, computing an approximately optimal solution is NP-hard. However, given a set of points from an optimal reachable space that covers it well, an approximate solution can be computed in polynomial time. The covering number highlights several interesting properties that help reduce the complexity of POMDP problems in practice, such as fully observed state variables, beliefs with sparse support, smooth beliefs, and circulant state-transition matrices.

Spotlight presentation, Tuesday, 5:20PM.

4 *Incremental Natural Actor-Critic Algorithms*

Shalabh Bhatnagar shalabh@csa.iisc.ernet.in
Indian Institute of Science
Richard Sutton sutton@cs.ualberta.ca
University of Alberta
Mohammad Ghavamzadeh mgh@cs.ualberta.ca
University of Alberta
Mark Lee mlee@cs.ualberta.ca
University of Alberta

We present four new reinforcement learning algorithms based on actor-critic and natural-gradient ideas, and provide their convergence proofs. Actor-critic reinforcement learning methods are online approximations to policy iteration in which the value-function parameters are estimated using temporal difference learning and the policy parameters are updated by stochastic gradient descent. Methods based on policy gradient in this way are of special interest because of their compatibility with function approximation methods, which are needed to handle large or infinite state spaces, and the use of temporal difference learning in this way is of interest because in many applications it dramatically reduces the variance of the policy gradient estimates. The use of the natural gradient is of interest because it can produce better conditioned parameterizations and has been shown to further reduce variance in some cases. Our results extend prior two-timescale convergence results for actor-critic methods by Konda et al. by using temporal difference learning in the actor and by incorporating natural gradients, and extend prior empirical studies of natural-gradient actor-critic methods by Peters et al. by providing the first convergence proofs and the first fully incremental algorithms.

Spotlight presentation, Tuesday, 5:20PM.

5 *Reinforcement Learning in Continuous Action Spaces through Sequential Monte Carlo Methods*

Alessandro Lazaric lazaric@elet.polimi.it
Politecnico di Milano
Marcello Restelli restelli@elet.polimi.it
Politecnico di Milano
Andrea Bonarini bonarini@elet.polimi.it
AI&Robotics Lab - Politecnico di Milano

Learning in real-world domains often requires to deal with continuous state and action spaces. Although many solutions have been proposed to apply Reinforcement Learning algorithms to continuous state problems, the same techniques can be hardly extended to continuous action spaces, where, besides the computation of a good approximation of the value function, a fast method for the identification of the highest-valued action is needed. In this paper, we propose a novel actor-critic approach in which the policy of the actor is estimated through sequential Monte Carlo methods. The importance sampling step is performed on the basis of the values learned by the critic, while the resampling step modifies the actor's policy. The proposed approach has been empirically compared to other learning algorithms into several domains; in this paper, we report results obtained in a control problem consisting of steering a boat across a river.

Spotlight presentation, Tuesday, 5:20PM.

6 *Fitted Q-iteration in continuous action-space MDPs*

Csaba Szepesvari szepesva@cs.ualberta.ca
University of Alberta
Andras Antos antos@szit.bme.hu
Computer and Automation Research Institute, of the Hungarian Academy of Science
Remi Munos remi.munos@polytechnique.fr
INRIA

We consider continuous state, continuous action batch reinforcement learning where the goal is to learn a good policy from a sufficiently rich trajectory generated by some policy. We study a variant of fitted Q-iteration, where the greedy action selection is replaced by searching for a policy in a restricted set of candidate policies by maximizing the average action values. We provide a rigorous analysis of this algorithm, proving what we believe is the first finite-time bound for value-function based algorithms for continuous state and action problems.

7 *Temporal Difference Updating without a Learning Rate*

Marcus Hutter marcus@hutter1.net
RSISE(ANU) and SML(NICTA)
Shane Legg shane@idsia.ch
IDSIA

We derive an equation for temporal difference learning from statistical first principles. Specifically, we start with the variational principle and then bootstrap to produce an updating rule for discounted state value estimates. The resulting equation is similar to the standard equation for temporal difference learning with eligibility traces, so called TD(λ), however it lacks the parameter α that specifies the learning rate. In the place of this free parameter there is now an equation for the learning rate that is specific to each state transition. We experimentally test this new learning rule against TD(λ) and find that it offers superior performance in various settings. Finally, we make some preliminary investigations into how to extend our new temporal difference algorithm to reinforcement learning. To do this we combine our update equation with both Watkins' Q(λ) and Sarsa(λ) and find that it again offers superior performance with fewer parameters.

8 *Managing Power Consumption and Performance of Computing Systems Using Reinforcement Learning*

Gerald Tesauro gtesauro@us.ibm.com
IBM TJ Watson Research Center
Rajarshi Das rajarshi@us.ibm.com
IBM Austin Research Laboratory
Hoi Chan hychan@us.ibm.com
IBM Austin Research Laboratory
Jeffrey Kephart kephart@us.ibm.com
IBM Research
David Levine dwl@us.ibm.com
IBM Austin Research Laboratory
Freeman Rawson frawson@us.ibm.com
IBM Austin Research Laboratory
Charles Lefurgy lefurgy@us.ibm.com
IBM

Electrical power management in large-scale IT systems such as commercial datacenters is an application area of rapidly growing interest from both an economic and ecological perspective, with billions of dollars and millions of metric tons of CO_2 emissions at stake annually. Businesses want to save power without sacrificing performance. This paper presents a reinforcement learning approach to simultaneous online management of both performance and power consumption. We apply RL in a realistic laboratory testbed using a Blade cluster and dynamically varying HTTP workload running on a commercial web applications middleware platform. We embed a CPU frequency controller in the Blade servers' firmware, and we train policies for this controller using a multi-criteria reward signal depending on both application performance and CPU power consumption. Our testbed scenario posed a number of challenges to successful use of RL, including multiple disparate reward functions, limited decision sampling rates, and pathologies arising when using multiple sensor readings as state variables. We describe innovative practical solutions to these challenges, and demonstrate clear performance improvements over both hand-designed policies as well as obvious "cookbook" RL implementations.

Spotlight presentation, Tuesday, 5:20PM.

9 *Hierarchical Apprenticeship Learning with Application to Quadruped Locomotion*

J. Zico Kolter kolter@cs.stanford.edu
Stanford University
Pieter Abbeel pabbeel@cs.stanford.edu
Stanford University
Andrew Ng ang@cs.stanford.edu
Stanford University

We consider apprenticeship learning — learning from expert demonstrations — in the setting of large, complex domains. Past work in apprenticeship learning requires that

the expert demonstrate complete trajectories through the domain, but in many problems where even an expert has difficulty controlling the system, this is infeasible. For example, consider the task of teaching a quadruped robot to navigate over extreme terrain; demonstrating an optimal policy (i.e., an optimal set of foot locations over the entire terrain) is a highly non-trivial task, even for an expert. In this paper we propose a method for hierarchical apprenticeship learning, which allows the algorithm to accept isolated advice at different hierarchical levels of the control task. This type of advice is often feasible for experts to give, even if the expert is unable to demonstrate full trajectories. This thus allows us to extend the apprenticeship learning paradigm to much larger, more challenging domains. In particular, in this paper we apply the hierarchical apprenticeship learning algorithm to the task of quadruped locomotion over extreme terrain, and achieve, to the best of our knowledge, results superior to any previously published work.

Spotlight presentation, Tuesday, 5:20PM.

10 *A Game-Theoretic Approach to Apprenticeship Learning*

Umar Syed, Princeton University and Robert Schapire, Princeton University.
Oral presentation, Tuesday, 4:40PM. *See abstract, page 94.*

11 *The Price of Bandit Information for Online Optimization*

Varsha Dani, University of Chicago, Thomas Hayes, Toyota Technological Institute at Chicago and Sham Kakade, Toyota Technological Institute.
Oral presentation, Tuesday, 4:20PM. *See abstract, page 94.*

12 *Adaptive Online Gradient Descent*

Peter Bartlett, UC Berkeley, Alexander Rakhlin, UC Berkeley and Elad Hazan, Princeton University.
Oral presentation, Tuesday, 3:00PM. *See abstract, page 92.*

13 *A general agnostic active learning algorithm*

Daniel Hsu djhsu@cs.ucsd.edu
University of California San Diego
Sanjoy Dasgupta dasgupta@cs.ucsd.edu
University of California San Diego
Claire Monteleoni cmontel@cs.ucsd.edu
University of California San Diego

We present an agnostic active learning algorithm for any hypothesis class of bounded VC dimension under arbitrary data distributions. Most previous work on active learning either makes strong distributional assumptions, or else is computationally prohibitive. Our algorithm extends the simple scheme of Cohn, Atlas, and Ladner to the agnostic setting, using reductions to supervised learning that harness generalization bounds in a simple but subtle manner. We provide a fall-back guarantee that bounds the algorithm's label complexity by the agnostic PAC sample complexity. Our analysis yields asymptotic label complexity improvements for certain hypothesis classes and distributions. We also demonstrate improvements experimentally.

Spotlight presentation, Tuesday, 5:20PM.

14 *Progressive mixture rules are deviation suboptimal*

Jean-Yves Audibert audibert@certis.enpc.fr
Ecole Nationale des Ponts et Chaussees

We consider the learning task consisting in predicting as well as the best function in a finite reference set G up to the smallest possible additive term. If $R(g)$ denotes the generalization error of a prediction function g, under reasonable assumptions on the loss function (typically satisfied by the least square loss when the output is bounded), it is known that the progressive mixture rule g_n satisfies $ER(g_n) < \min_{g \in G} R(g) + Cst(\log |G|)/n$ where n denotes the size of the training set, E denotes the expectation wrt the training set distribution. This work shows that, surprisingly, for appropriate reference sets G, the deviation convergence rate of the progressive mixture rule is only no better than $Cst/\sqrt{n}$, and not the expected Cst/n. It also provides an algorithm which does not suffer from this drawback.

15 *Boosting Algorithms for Maximizing the Soft Margin*

Manfred Warmuth manfred@soe.ucsc.edu
UC Santa Cruz
Karen Glocer kag@soe.ucsc.edu
UC Santa Cruz
Gunnar Ratsch gunnar.raetsch@tuebingen.mpg.de
Friedrich Miescher Laboratory, Max Planck Society

We present a novel boosting algorithm, called Softboost, designed for sets of binary labeled examples that are not necessarily separable by convex combinations of base hypotheses. Our algorithm aims to achieve robustness by *capping* the distributions on the examples. Our update of the distribution is motivated by minimizing a relative entropy subject to the capping constraints and constraints on the edges of the obtained base hypotheses. The capping constraints imply a soft margin in the dual optimization problem and our algorithm produces a convex combination of hypotheses whose *soft margin* is within δ of the optimum. We employ relative entropy projection methods to prove an $O(\frac{\ln N}{\delta^2})$ iteration bound for our algorithm, where N is number of examples.

Spotlight presentation, Tuesday, 3:20PM.

16 *A Risk Minimization Principle for a Class of Parzen Estimators*

Kristiaan Pelckmans kpelckma@gmail.com
ESAT - SCD/sista, KULeuven
Johan Suykens johan.suykens@esat.kuleuven.be
ESAT - SCD/sista, KULeuven
Bart De Moor bart.demoor@esat.kuleuven.be
ESAT - SCD/sista, KULeuven

This paper explores the use of a Maximal Average Margin (MAM) optimality principle for the design of learning algorithms. It is shown that the application of this risk minimization principle results in a class of (computationally) simple learning machines similar to the classical Parzen window classifier. A direct relation with the Rademacher complexities is established, as such facilitating analysis and providing a notion of certainty of prediction. This analysis is related to Support Vector Machines by means of a margin transformation. The power of the MAM principle is illustrated further by application to ordinal regression tasks, resulting in an $O(n)$ algorithm able to process large datasets in reasonable time.

17 *Bundle Methods for Machine Learning*

Alex Smola alex.smola@gmail.com
NICTA
S V N Vishwanathan svn.vishwanathan@nicta.com.au
Statistical Machine Learning Program, National ICT Australia
Quoc Le quoc.le@anu.edu.au
Statistical Machine Learning, NICTA

We present a globally convergent method for regularized risk minimization problems. Our method applies to Support Vector estimation, regression, Gaussian Processes, and any other regularized risk minimization setting which leads to a convex optimization problem. SVMPerf can be shown to be a special case of our approach. In addition to the unified framework we present tight convergence bounds, which show that our algorithm converges in $O(1/\epsilon)$ steps to ϵ precision for general convex problems and in $O(\log \epsilon)$ steps for continuously differentiable problems. We demonstrate in experiments the performance of our approach.

Spotlight presentation, Tuesday, 3:20PM.

18 *Simulated Annealing: Rigorous finite-time guarantees for optimization on continuous domains*

Andrea Lecchini-Visintini alv1@leicester.ac.uk
University of Leicester
John Lygeros lygeros@control.ee.ethz.ch
ETH Zurich
Jan Maciejowski jmm@eng.cam.ac.uk
Cambridge University

Simulated annealing is a popular method for approaching the solution of a global optimization problem. Existing results on its performance apply to discrete combinatorial optimization where the optimization variables can assume only a finite set of possible values. We introduce a new general formulation of simulated annealing which allows one to guarantee finite-time performance in the optimization of functions of continuous variables. The results hold universally for any optimization problem on a bounded domain and establish a connection between simulated annealing and up-to-date theory of convergence of Markov chain Monte Carlo methods on continuous domains. This work is inspired by the concept of finite-time learning with known accuracy and confidence developed in statistical learning theory.

Spotlight presentation, Wednesday, 9:50AM.

19 *Consistent Minimization of Clustering Objective Functions*

Ulrike von Luxburg ulrike.luxburg@tuebingen.mpg.de
MPI for Biological Cybernetics
Sebastien Bubeck sebastien.bubeck@inria.fr
INRIA futurs
Stefanie Jegelka stefanie.jegelka@tuebingen.mpg.de
MPI for Biological Cybernetics
Michael Kaufmann mk@informatik.uni-tuebingen.de
Universitat Tubingen

Clustering is often formulated as a discrete optimization problem. The objective is to find, among all partitions of the data set, the best one according to some quality measure. However, in the statistical setting where we assume that the finite data set has been sampled from some underlying space, the goal is not to find the best partition of the given sample, but to approximate the true partition of the underlying space. We argue that the discrete optimization approach usually does not achieve this goal. As an alternative, we suggest the paradigm of "nearest neighbor clustering". Instead of selecting the best out of all partitions of the sample, it only considers partitions in some restricted function class. Using tools from statistical learning theory we prove that nearest neighbor clustering is statistically consistent. Moreover, its worst case complexity is polynomial by construction, and it can be implemented with small average case complexity using branch and bound.

Spotlight presentation, Tuesday, 3:20PM.

20 *Cluster Stability for Finite Samples*

Ohad Shamir, The Hebrew University of Jerusalem and Naftali Tishby, Hebrew University.

Oral presentation, Tuesday, 5:00PM. See abstract, page 95.

21 *Probabilistic Matrix Factorization*

Ruslan Salakhutdinov, University of Toronto, Department of Computer Science and Andriy Mnih, University of Toronto.
Oral presentation, Tuesday, 9:30AM. See abstract, page 87.

22 *COFI RANK - Maximum Margin Matrix Factorization for Collaborative Ranking*

Markus Weimer mweimer@tk.informatik.tu-darmstadt.de
TU Darmstadt
Alexandros Karatzoglou alexis@ci.tuwien.ac.at
Vienna University of Technology, Statistics
Quoc Le quoc.le@nicta.com.au
Statistical Machine Learning Program, National ICT Australia
Alex Smola alex.smola@gmail.com
NICTA

In this paper, we consider collaborative filtering as a ranking problem. We present a method which uses Maximum Margin Matrix Factorization and optimizes ranking instead of rating. We use structured output prediction to optimize for specific non-uniform ranking scores. Experimental results show that our method gives very good ranking scores and scales well on collaborative filtering tasks.

Spotlight presentation, Tuesday, 9:50AM.

23 *Colored Maximum Variance Unfolding*

Le Song, NICTA and School of Information Technologies, the University of Sydney, Alex Smola, NICTA, Karsten Borgwardt, University of Cambridge and Arthur Gretton, MPI for Biological Cybernetics.
Oral presentation, Tuesday, 10:30AM. See abstract, page 89.

24 *Iterative Non-linear Dimensionality Reduction with Manifold Sculpting*

Michael Gashler mikegashler@gmail.com
Brigham Young University
Dan Ventura ventura@cs.byu.edu
Brigham Young University
Tony Martinez martinez@cs.byu.edu
Brigham Young University

Many algorithms have been recently developed for reducing dimensionality by projecting data onto an intrinsic non-linear manifold. Unfortunately, existing algorithms often lose significant precision in this transformation. Manifold Sculpting is a new algorithm that iteratively reduces dimensionality by simulating surface tension in local neighborhoods. We present several experiments that show Manifold Sculpting yields more accurate results than existing algorithms with both generated and natural data-sets. Manifold Sculpting is also able to benefit from both prior dimensionality reduction efforts.

Spotlight presentation, Tuesday, 3:20PM.

25 *The Distribution Family of Similarity Distances*

Gertjan Burghouts gertjan.burghouts@tno.nl
TNO - Electro Optics
Arnold Smeulders smeulders@science.uva.nl
University of Amsterdam
Jan-Mark Geusebroek mark@science.uva.nl
University of Amsterdam

Assessing similarity between features is a key step in object recognition and scene categorization tasks. We argue that knowledge on the distribution of distances generated by similarity functions is crucial in deciding whether features are similar or not. Intuitively one would expect that similarities between features could arise from any distribution. In this paper, we will derive the contrary, and report the theoretical result that L_p-norms –a class of commonly applied distance metrics– from one feature vector to other vectors are Weibull-distributed if the feature values are correlated and non-identically distributed. Besides these assumptions being realistic for images, we experimentally show them to hold for various popular feature extraction algorithms, for a diverse range of images. This fundamental insight opens new directions in the assessment of feature similarity, with projected improvements in object and scene recognition algorithms.

26 *Heterogeneous Component Analysis*

Shigeyuki Oba shige-o@is.naist.jp
Nara Institute of Science and Technology
Motoaki Kawanabe motoaki.kawanabe@first.fraunhofer.de
Fraunhofer FIRST.IDA
Klaus-Robert Müller klaus@first.fraunhofer.de
Fraunhofer FIRST.IDA
Shin Ishii ishii@is.naist.jp
Dept. Syst. Sci. Grad. Sch. Info. Kyoto Univ., Nare Institute of Science and Technology

In bioinformatics it is often desirable to combine data from various measurement sources and thus structured feature vectors are to be analyzed that possess different intrinsic blocking characteristics (e.g., different patterns of missing values, observation noise levels, effective intrinsic dimensionalities). We propose a new machine learning tool, heterogeneous component analysis (HCA), for feature extraction in order to better understand the factors that underlie such complex structured heterogeneous data. HCA is a linear block-wise sparse Bayesian PCA based not only on a probabilistic model with block-wise residual variance terms but also on a Bayesian treatment of a block-wise sparse factor-loading matrix. We study various algorithms that implement our HCA concept extracting sparse heterogeneous structure by obtaining common components for the blocks and specific components within each block. Simulations on toy and bioinformatics data underline the usefulness of the proposed structured matrix factorization concept.

Spotlight presentation, Tuesday, 9:50AM.

27 *Discriminative K-means for Clustering*

Jieping Ye jieping.ye@asu.edu
Department of Computer Science & Engineering, Arizona State University
Zheng Zhao zhaozheng@asu.edu
Arizona State University
Mingrui Wu mingrui.wu@tuebingen.mpg.de
Max Planck Institute

We present a theoretical study on the discriminative clustering framework, recently proposed for simultaneous subspace selection via linear discriminant analysis (LDA) and clustering. Empirical results have shown its favorable performance in comparison with several other popular clustering algorithms. However, the inherent relationship between subspace selection and clustering in this framework is not well understood, due to the iterative nature of the algorithm. We show in this paper that this iterative subspace selection and clustering is equivalent to kernel K-means with a specific kernel Gram matrix. This provides significant and new insights into the nature of this subspace selection procedure. Based on this equivalence relationship, we propose the Discriminative K-means (DisKmeans) algorithm for simultaneous LDA subspace selection and clustering, as well as an automatic parameter estimation procedure. We also present the nonlinear extension of DisKmeans using kernels. We show that the learning of the kernel matrix over a convex set of pre-specified kernel matrices can be incorporated into the clustering formulation. The connection between DisKmeans and several other clustering algorithms is also analyzed. The presented theories and algorithms are evaluated through experiments on a collection of benchmark data sets.

28 *DIFFRAC: a discriminative and flexible framework for clustering*

Francis Bach francis.bach@mines.org
INRIA - Ecole Normale Superieure
Zaid Harchaoui zharchao@enst.fr
GET/Telecom Paris

We present a novel linear clustering framework (Diffrac) which relies on a linear discriminative cost function and a convex relaxation of a combinatorial optimization problem. The large convex optimization problem is solved through a sequence of lower dimensional singular value decompositions. This framework has several attractive properties: (1) although apparently similar to K-means, it exhibits superior clustering performance than K-means, in particular in terms of robustness to noise. (2) It can be readily extended to non linear clustering if the discriminative cost function is based on positive definite kernels, and can then be seen as an alternative to spectral clustering. (3) Prior information on the partition is easily incorporated, leading to state-of-the-art performance for semi-supervised learning, for clustering or classification. We present empirical evaluations of our algorithms on synthetic and real medium-scale datasets.

29 *Regularized Boost for Semi-Supervised Learning*

KE CHEN kchen@ieee.org
School of Computer Science, The University of Manchester
SHIHAI WANG shihai.wang@postgrad.manchester.ac.uk
School of Computer Science, The University of

Semi-supervised inductive learning concerns how to learn a decision rule from a data set containing both labeled and unlabeled data. Several boosting algorithms have been extended to semi-supervised learning with various strategies. To our knowledge, however, none of them takes local smoothness constraints among data into account during ensemble learning. In this paper, we introduce a local smoothness regularizer to semi-supervised boosting algorithms based on the universal optimization framework of margin cost functionals. Our regularizer is applicable to existing semi-supervised boosting algorithms to improve their generalization and speed up their training. Comparative results on synthetic, benchmark and real world tasks demonstrate the effectiveness of our local smoothness regularizer. We discuss relevant issues and relate our regularizer to previous work.

30 *Semi-Supervised Multitask Learning*

QIUHUA LIU ql@ee.duke.edu
Duke University
XUEJUN LIAO xjliao@ee.duke.edu
Duke University
LAWRENCE CARIN lcarin@ee.duke.edu
Duke University

A semi-supervised multitask learning (MTL) framework is presented, in which M parameterized semi-supervised classifiers, each associated with one of M partially labeled data manifolds, are learned jointly under the constraint of a soft-sharing prior imposed over the parameters of the classifiers. The unlabeled data are utilized by basing classifier learning on neighborhoods, induced by a Markov random walk over a graph representation of each manifold. Experimental results on real data sets demonstrate that semi-supervised MTL yields significant improvements in generalization performance over either semi-supervised single-task learning (STL) or supervised MTL.

Spotlight presentation, Tuesday, 9:50AM.

31 *Efficient Convex Relaxation for Transductive Support Vector Machine*

IRWIN KING king@cse.cuhk.edu.hk
The Chinese University of Hong Kong

Jianke Zhu jkzhu@cse.cuhk.edu.hk
Cse.Cuhk
Rong Jin rongjin@cse.msu.edu
Michigan State University
Zenglin Xu zlxu@cse.cuhk.edu.hk
The Chinese University of Hong Kong
Michael Lyu lyu@cse.cuhk.edu.hk
The Chinese University of Hong Kong

We consider the problem of Support Vector Machine transduction, which involves a combinatorial problem with exponential computational complexity in the number of unlabeled examples. Although several studies are devoted to Transductive SVM, they suffer either from the high computation complexity or from the solutions of local optimum. To address this problem, we propose solving Transductive SVM via a convex relaxation, which converts the NP-hard problem to a semi-definite programming. Compared with the other SDP relaxation for Transductive SVM, the proposed algorithm is computationally more efficient with the number of free parameters reduced from O(n2) to O(n) where n is the number of examples. Empirical study with several benchmark data sets shows the promising performance of the proposed algorithm in comparison with other state-of-the-art implementations of Transductive SVM.

32 *A Randomized Algorithm for Large Scale Support Vector Learning*

Krishnan Kumar krishnan.suresh@gmail.com
Indian Institute of Science
Chiru Bhattacharya chiru@csa.iisc.ernet.in
Indian Institute of Science
Ramesh Hariharan ramesh@strandls.com
Strand Life Sciences

We propose a randomized algorithm for large scale SVM learning which solves the problem by iterating over random subsets of the data. Crucial to the algorithm for scalability is the size of the subsets chosen. In the context of text classification we show that, by using ideas from random projections, a sample size of O(log n) can be used to obtain a solution which is close to the optimal with a high probability. Experiments done on synthetic and real life data sets demonstrate that the algorithm scales up SVM learners, without loss in accuracy.

33 *Random Features for Large-Scale Kernel Machines*

Ali Rahimi ali.rahimi@intel.com
Intel Research
Benjamin Recht brecht@ist.caltech.edu
California Institute of Technology

To accelerate the training of kernel machines, we propose to map the input data to a randomized low-dimensional feature space and then apply existing fast linear methods. The features are designed so that the inner products of the transformed data are approximately equal to those in the feature space of a user specified shift-invariant kernel. We explore two sets of random features, provide convergence bounds on their ability to approximate various radial basis kernels, and show that in large-scale classification and regression tasks

linear machine learning algorithms applied to these features outperform state-of-the-art large-scale kernel machines.

Spotlight presentation, Tuesday, 3:20PM.

34 *Multi-Task Learning via Conic Programming*

Tsuyoshi Kato kato-tsuyoshi@cb.k.u-tokyo.ac.jp
University of Tokyo, Graduate School of Frontier Sciences
Hisashi Kashima kashi_pong@yahoo.co.jp
IBM Research, Tokyo Research Laboratory
Masashi Sugiyama sugi@cs.titech.ac.jp
Tokyo Institute of Technology, Department of Computer Science
Kiyoshi Asai asai@cbrc.jp
CBRC Japan

When we have several related tasks, solving them simultaneously is shown to be more effective than solving them individually. This approach is called multi-task learning (MTL) and has been studied extensively. Existing approaches to MTL often treat all the tasks as *uniformly related* to each other and the relatedness of the tasks is controlled globally. For this reason, the existing methods can lead to undesired solutions when some tasks are not highly related to each other, and some pairs of related tasks can have significantly different solutions. In this paper, we propose a novel MTL algorithm that can overcome these problems. Our method makes use of a task network, which describes the relation structure among tasks. This allows us to deal with intricate relation structures in a systematic way. Furthermore, we control the relatedness of the tasks locally, so all pairs of related tasks are guaranteed to have similar solutions. We apply the above idea to support vector machines (SVMs) and show that the optimization problem can be cast as a second order cone program, which is convex and can be solved efficiently. The usefulness of our approach is demonstrated through simulations with protein super-family classification and ordinal regression problems.

35 *Anytime Induction of Cost-sensitive Trees*

Saher Esmeir esaher@cs.technion.ac.il
Computer Science Department, Technion-IIT
Shaul Markovitch shaulm@cs.technion.ac.il
Computer Science Department, Technion-IIT

Machine learning techniques are increasingly being used to produce a wide-range of classifiers for complex real-world applications that involve nonuniform testing costs and misclassification costs. As the complexity of these applications grows, the management of resources during the learning and classification processes becomes a challenging task. In this work we introduce ACT Anytime Cost-sensitive Trees), a novel framework for operating in such environments. ACT is an anytime algorithm that allows trading computation time for lower classification costs. It builds a tree top-down and exploits additional time resources to obtain better estimations for the utility of the different candidate splits. Using sampling techniques ACT approximates for each candidate split the cost of the subtree under it and favors the one with a minimal cost. Due to its stochastic nature ACT is expected to be able to escape local minima, into which greedy methods may be trapped. Experiments with a variety of datasets were conducted to compare the performance of ACT to that of the state of the art cost-sensitive tree learners. The results show that

for most domains ACT produces trees of significantly lower costs. ACT is also shown to exhibit good anytime behavior with diminishing returns.

Spotlight presentation, Tuesday, 3:20PM.

36 *Parallelizing Support Vector Machines on Distributed Computers*

Edward Chang edchang@google.com
Google Research, University of California
Kaihua Zhu zkh@google.com
Google Inc.
Hao Wang whao@google.com
Google Inc
Hongjie Bai hongjie@google.com
Google Inc
Jian Li lijian@google.com
Google Inc
Zhihuan Qiu zhihuan@google.com
Google Inc
Hang Cui hang@google.com
Google Inc

Support Vector Machines (SVMs) suffer from a widely recognized scalability problem in both memory use and computational time. To improve scalability, we have developed a parallel SVM algorithm (PSVM), which reduces memory use through performing a row-based, approximate matrix factorization, and which loads only essential data to each machine to perform parallel computation. Let n denote the number of training instances, p the reduced matrix dimension after factorization (p is significantly smaller than n), and m the number of machines. PSVM reduces the memory requirement from $O(n^2)$ to $O(np/m)$, and improves computation time to $O(np^2/m)$. Empirical studies on up to 500 computers shows PSVM to be effective.

37 *FilterBoost: Regression and Classification on Large Datasets*

Joseph K Bradley, Carnegie Mellon University
and Robert Schapire, Princeton University.
Oral presentation, Tuesday, 4:00PM. *See abstract, page 94.*

38 *McRank: Learning to Rank Using Multiple Classification and Gradient Boosting*

Ping Li pingli@cornell.edu
Cornell University
Christopher Burges cburges@microsoft.com
Microsoft Research
Qiang Wu qiangwu@microsoft.com
Microsoft Research

We cast the ranking problem as (1) multiple classification ("Mc") (2) multiple ordinal classification, which lead to computationally tractable learning algorithms for relevance ranking in Web search. We consider the DCG criterion (discounted cumulative gain),

a standard quality measure in information retrieval. Our approach is motivated by the fact that perfect classifications result in perfect DCG scores and the DCG errors are bounded by classification errors. We propose using the *Expected Relevance* to convert class probabilities into ranking scores. The class probabilities are learned using a gradient boosting tree algorithm. Evaluations on large-scale datasets show that our approach can improve *LambdaRank* and the regressions-based ranker, in terms of the (normalized) DCG scores. An efficient implementation of the boosting tree algorithm is also presented.

Spotlight presentation, Tuesday, 3:20PM.

39 *A General Boosting Method and its Application to Learning Ranking Functions for Web Search*

Zhaohui Zheng zhaohui@yahoo-inc.com
Yahoo Inc.
Hongyuan Zha zha@cc.gatech.edu
Georgia Institute of Technology
Gordon Sun gzsun@yahoo-inc.com
Yahoo Research
Olivier Chapelle chap@yaho-inc.com
Yahoo Research
Keke Chen kchen@yahoo-inc.com
Yahoo Research
Tong Zhang tzhang@stat.rutgers.edu
Rutgers University

We present a general boosting method extending functional gradient boosting to optimize complex loss functions that are encountered in many machine learning problems. Our approach is based on optimization of quadratic upper bounds of the loss functions which allows us to present a rigorous convergence analysis of the algorithm. More importantly, this general framework enables us to use a standard regression base learner such as decision trees for fitting any loss function. We illustrate an application of the proposed method in learning ranking functions for Web search by combining both preference data and labeled data for training. We present experimental results for Web search using data from a commercial search engine that show significant improvements of our proposed methods over some existing methods.

40 *Classification via Minimum Incremental Coding Length (MICL)*

John Wright jnwright@uiuc.edu
University of Illinois at Urbana-Champaign
Yangyu Tao f-yyutao@msrchina.research.microsoft.com
Microsoft Research
Zhouchen Lin zhoulin@microsoft.com
Microsoft Research
Yi Ma yima@uiuc.edu
Electrical & Computer Engineering Department, University of Illinois at Urbana-Champaign
Heung-Yeung Shum hshum@microsoft.com
Microsoft Research

We present a simple new criterion for classification, based on principles from lossy data compression. The criterion assigns a test sample to the class that uses the minimum

number of additional bits to code the test sample, subject to an allowable distortion. We rigorously prove asymptotic optimality of this criterion for Gaussian (normal) distributions and analyze its relationships to classical classifiers. The theoretical results provide new insights into the relationships among a variety of popular classifiers such as MAP, RDA, k-NN, and SVM, as well as unsupervised methods based on lossy coding. Our formulation induces several good effects on the resulting classifier. First, minimizing the lossy coding length induces a regularization effect which stabilizes the (implicit) density estimate in a small sample setting. Second, compression provides a uniform means of handling classes of varying dimension. The new criterion and its kernel and local versions perform competitively on synthetic examples, as well as on real imagery data such as handwritten digits and face images. On these problems, the performance of our simple classifier approaches the best reported results, without using domain-specific information. All MATLAB code and classification results will be made publicly available for peer evaluation.

Spotlight presentation, Tuesday, 3:20PM.

41 *On higher-order perceptron algorithms*

Claudio Gentile claudio.gentile@uninsubria.it
Universita dell' Insubria
Fabio Vitale fabiovdk@yahoo.com
DICOM, Universita dell' Insubria
Cristian Brotto cristianbrotto@inwind.it
DICOM, Universita dell' Insubria

A new algorithm for on-line learning linear-threshold functions is proposed which efficiently combines second-order statistics about the data with the "logarithmic behavior" of multiplicative/dual-norm algorithms. An initial theoretical analysis is provided suggesting that our algorithm might be viewed as a standard Perceptron algorithm operating on a transformed sequence of examples with improved margin properties. We also report on experiments carried out on datasets from diverse domains, with the goal of comparing to known Perceptron algorithms (first-order, second-order, additive, multiplicative). Our learning procedure seems to generalize quite well, and converges faster than the corresponding multiplicative baseline algorithms.

42 *Hierarchical Penalization*

Marie Szafranski marie.szafranski@hds.utc.fr
Heudiasyc - CNRS 6599, Compiegne University of Technology
Yves Grandvalet yves.grandvalet@utc.fr
IDIAP Research Institute, CNRS
Pierre Morizet-Mahoudeaux pierre.morizet@utc.fr
IDIAP Research Institute, CNRS

This article presents hierarchical penalization, a generic framework for incorporating prior information in the fitting of statistical models, when the explicative variables are organized in a hierarchical structure. The penalizer, derived from an adaptive penalization formulation, is a convex functional that performs soft selection at the group level, and that shrinks variables within each group, to favor solutions with few leading terms in the final combination. The framework, originally derived for taking into account prior knowledge, is shown to be useful in kernel regression, when several parameters are used to model the influence of features.

Spotlight presentation, Tuesday, 11:50AM.

43 *New Outer Bounds on the Marginal Polytope*

David Sontag, Massachusetts Institute of Technology and
Tommi Jaakkola, Massachusetts Institute of Technology.
Oral presentation, Tuesday, 11:10AM. *See abstract, page 89.*

44 *An Analysis of Convex Relaxations for MAP Estimation*

Pawan Mudigonda, Oxford Brookes University, Vladimir Kolmogorov,
University College London and Philip Torr, Oxford Brookes University.
Oral presentation, Tuesday, 10:50AM. *See abstract, page 89.*

45 *Message Passing for Max-weight Independent Set*

Sujay Sanghavi sanghavi@mit.edu
MIT
Devavrat Shah devavrat@mit.edu
Assistant Professor
Alan Willsky willsky@mit.edu
MIT

We investigate the use of message-passing algorithms for the problem of finding the max-weight independent set (MWIS) in a graph. First, we study the performance of loopy max-product belief propagation. We show that, if it converges, the quality of the estimate is closely related to the tightness of an LP relaxation of the MWIS problem. We use this relationship to obtain sufficient conditions for correctness of the estimate. We then develop a modification of max-product – one that converges to an optimal solution of the dual of the MWIS problem. We also develop a simple iterative algorithm for estimating the max-weight independent set from this dual solution. We show that the MWIS estimate obtained using these two algorithms is conjunction correct when the graph is bipartite and the MWIS is unique. Finally, we show that any problem of MAP estimation for probability distributions over finite domains can be reduced to an MWIS problem. We believe this reduction will yield new insights and algorithms for MAP estimation.

Spotlight presentation, Tuesday, 11:50AM.

46 *Privacy-Preserving Belief Propagation and Sampling*

Michael Kearns mkearns@cis.upenn.edu
University of Pennsylvania
Jinsong Tan jinsong@seas.upenn.edu
University of Pennsylvania
Jennifer Wortman wortmanj@seas.upenn.edu
University of Pennsylvania

We provide provably privacy-preserving versions of belief propagation, Gibbs sampling, and other local algorithms — distributed multiparty protocols in which each party or vertex learns only its final local value, and absolutely nothing else.

Spotlight presentation, Tuesday, 11:50AM.

47 *Scene Segmentation with CRFs Learned from Partially Labeled Images*

Jakob Verbeek, INRIA Rhone-Alpes, Laboratoire Jean Kuntzmann and Bill Triggs, INRIA France.

Oral presentation, Tuesday, 11:30AM. See abstract, page 90.

48 *Fast and Scalable Training of Semi-Supervised CRFs with Application to Activity Recognition*

Maryam Mahdaviani maryam@cs.ubc.ca
University of British Columbia
Tanzeem Choudhury tanzeem.choudhury@intel.com
Intel Research

We present a new and efficient semi-supervised training method for parameter estimation and feature selection in conditional random fields (CRFs). In real-world applications such as activity recognition, unlabeled sensor traces are relatively easy to obtain whereas labeled examples are expensive and tedious to collect. Furthermore, the ability to automatically select a small subset of discriminatory features from a large pool can be advantageous in terms of computational speed as well as accuracy. In this paper, we introduce the semi-supervised virtual evidence boosting (sVEB) algorithm for training CRFs – a semi-supervised extension to the recently developed virtual evidence boosting (VEB) method for feature selection and parameter learning. The objective function of sVEB combines the unlabeled conditional entropy with labeled conditional pseudo-likelihood. It reduces the overall system cost as well as the human labeling cost required during training, which are both important considerations in building real-world inference systems. Experiments on synthetic data and real activity traces collected from wearable sensors, illustrate that sVEB benefits from both the use of unlabeled data and automatic feature selection, and outperforms other semi-supervised approaches.

49 *Collective Inference on Markov Models for Modeling Bird Migration*

Daniel Sheldon dsheldon@cs.cornell.edu
Cornell University
M.A. Saleh Elmohamed saleh@cam.cornell.edu
Cornell University
Dexter Kozen kozen@cs.cornell.edu
Cornell University

We investigate a family of inference problems on Markov models, where many sample paths are drawn from a Markov chain and partial information is revealed to an observer who attempts to reconstruct the sample paths. We present algorithms and hardness results for several variants of this problem which arise by revealing different information to the observer and imposing different requirements for the reconstruction of sample paths. Our algorithms are analogous to the classical Viterbi algorithm for Hidden Markov Models, which finds single most probable sample path given a sequence of observations. Our work is motivated by an important application in ecology: inferring bird migration paths from a large database of observations.

Spotlight presentation, Tuesday, 11:50AM.

50 *TrueSkill Through Time: Revisiting the History of Chess*

Pierre Dangauthier pierre.dangauthier@imag.fr
INRIA Rhones-Alpes
Ralf Herbrich rherb@microsoft.com
Microsoft Research, Applied Games
Tom Minka minka@microsoft.com
Microsoft Research Ltd
Thore Graepel thoreg@microsoft.com
Microsoft Research Cambridge

We extend the Bayesian skill rating system TrueSkill to infer entire time series of skills of players by smoothing through time instead of filtering. The skill of each participating player, say, every year is represented by a latent skill variable which is affected by the relevant game outcomes that year, and coupled with the skill variables of the previous and subsequent year. Inference in the resulting factor graph is carried out by approximate message passing (EP) along the time series of skills. As before the system tracks the uncertainty about player skills, explicitly models draws, can deal with any number of competing entities and can infer individual skills from team results. We extend the system to estimate player-specific draw margins. Based on these models we present an analysis of the skill curves of important players in the history of chess over the past 150 years. Results include plots of players' lifetime skill development as well as the ability to compare the skills of different players across time. Our results indicate that a) the overall playing strength has increased over the past 150 years, and b) that modelling a player's ability to force a draw provides significantly better predictive power.

Spotlight presentation, Tuesday, 9:50AM.

51 *Regulator Discovery from Gene Expression Time Series of Malaria Parasites: a Hierachical Approach*

Jose Miguel Hernandez-Lobato josemiguel.hernandez@uam.es
Universidad Autonoma de Madrid
Tjeerd Dijkstra t.dijkstra@science.ru.nl
IRIS Radboud University Nijmegen
Tom Heskes t.heskes@science.ru.nl
Radboud University

We introduce a hierarchical Bayesian model for the discovery of putative regulators from gene expression data only. The hierarchy incorporates the knowledge that there are just a few regulators that by themselves only regulate a handful of genes. This is implemented through a so-called spike-and-slab prior, a mixture of Gaussians with different widths, with mixing weights from a hierarchical Bernoulli model. For efficient inference we implemented expectation propagation. Running the model on a malaria parasite data set, we found four genes with significant homology to transcription factors in an amoebe, one RNA regulator and three genes of unknown function (out of the top ten genes considered).

52 *Variational Inference for Diffusion Processes*

Cedric Archambeau c.archambeau@cs.ucl.ac.uk
University College London
Manfred Opper opperm@cs.tu-berlin.de
TU Berlin

Yuan Shen y.shen2@aston.ac.uk
Neural Computing Research Group, Aston University
Dan Cornford d.cornford@aston.ac.uk
Neural Computing Research Group, Aston University
John Shawe-Taylor j.shawe-taylor@cs.ucl.ac.uk
University College London

Diffusion processes are a family of continuous-time continuous-state stochastic processes that are in general only partially observed. The joint estimation of the forcing parameters and the system noise (volatility) in these dynamical systems is a crucial, but non-trivial task, especially when the system is nonlinear and multi-modal. We propose a variational treatment of diffusion processes, which allows us to estimate these parameters by simple gradient techniques and which is computationally less demanding than most MCMC approaches. Furthermore, our parameter inference scheme does not break down when the time step gets smaller, unlike most current approaches. Finally, we show how a cheap estimate of the posterior over the parameters can be constructed based on the variational free energy.

53 *Variational inference for Markov jump processes*

Manfred Opper opperm@cs.tu-berlin.de
TU Berlin
Guido Sanguinetti guido@dcs.shef.ac.uk
University of Sheffield

Markov jump processes play an important role in a large number of application domains. However, realistic systems are analytically intractable and they have traditionally been analysed using simulation based techniques, which do not provide a framework for statistical inference. We propose a mean field approximation to perform posterior inferencet and parameter estimation.The approximation allows a practical solution to the inference problem, while still retaining a good degree of accuracy. We illustrate our approach on two biologically motivated systems.

54 *Collapsed Variational Inference for HDP*

Yee Whye Teh ywteh@gatsby.ucl.ac.uk
Gatsby Computational Neuroscience Unit, UCL
Kenichi Kurihara kurihara@mi.cs.titech.ac.jp
Tokyo Institute of Technology
Max Welling welling@ics.uci.edu
University of California Irvine

A wide variety of Dirichlet-multinomial 'topic' models have found interesting applications in recent years. While Gibbs sampling remains an important method of inference in such models, variational techniques have certain advantages such as easy assessment of convergence, easy optimization without the need to maintain detailed balance, a bound on the marginal likelihood, and side-stepping of issues with topic-identifiability. The most accurate variational technique thus far, namely collapsed variational LDA (CV-LDA), did not deal with model selection nor did it include inference for the hyper-parameters. We generalize their technique to the HDP to address these issues. The result is a collapsed variational inference technique that can add topics indefinitely, for instance through split and merge heuristics, while the algorithm will automatically remove clusters which are

not supported by the data. Experiments show a very significant improvement in accuracy relative to CV-LDA.

Spotlight presentation, Tuesday, 11:50AM.

55 *Distributed Inference for Latent Dirichlet Allocation*

David Newman newman@uci.edu
UC Irvine
Arthur Asuncion asuncion@uci.edu
University of California, Irvine
Padhraic Smyth pjsmyth@uci.edu
University of California Irvine
Max Welling mwelling@uci.edu
University of California Irvine

We investigate the problem of learning a widely-used latent-variable model – the Latent Dirichlet Allocation (LDA) or topic model – using distributed computation, where each of P processors only sees 1/P of the total data set. We propose two distributed inference schemes that are motivated from different perspectives. The first scheme uses local Gibbs sampling on each processor with periodic updates—it is simple to implement and can be viewed as an approximation to a single processor implementation of Gibbs sampling. The second scheme relies on a hierarchical Bayesian extension of the standard LDA model to directly account for the fact that data are distributed across P processors—it has a theoretical guarantee of convergence but is more complex to implement than the approximate method. Using three real-world text corpora we show that distributed learning works very well for LDA models, i.e., perplexity and precision-recall scores for distributed learning are indistinguishable from those obtained with single-processor learning. Our extensive experimental results include large-scale distributed computation on 1000 virtual processors; and speedup experiments of learning topics in a 100-million word corpus using 16 processors.

Spotlight presentation, Tuesday, 11:50AM.

56 *Infinite State Bayes-Nets for Structured Domains*

Max Welling welling@ics.uci.edu
University of California Irvine
Ian Porteous iporteou@ics.uci.edu
University of California Irvine
Evgeniy Bart bart@caltech.edu
Cal Tech

A general modeling framework is proposed that unifies nonparametric-Bayesian models, topic-models and Bayesian networks. This class of infinite state Bayes nets (ISBN) can be viewed as directed networks of 'hierarchical Dirichlet processes' (HDPs) where the domain of the variables can be structured (e.g. words in documents or features in images). To model the structure and to share groups between them we use 'cascades' of Dirichlet priors. We show that collapsed Gibbs sampling can be done efficiently in these models by leveraging the structure of the Bayes net and using the forward-filtering-backward-sampling algorithm for junction trees. Existing models, such as nested-DP, Pachinko allocation, mixed membership models etc. are described as ISBNs. Two experiments have been implemented to illustrate these ideas.

Spotlight presentation, Tuesday, 9:50AM.

57 *The Infinite Gamma-Poisson Feature Model*

Michalis Titsias michalis_titsias@yahoo.gr
University of Thessaly

We address the problem of factorial learning which associates a set of latent causes or features with the observed data. Factorial models usually assume that each feature has a single occurrence in a given data point. However, there are data such as images where latent features have multiple occurrences, e.g. a visual object class can have multiple instances shown in the same image. To deal with such cases, we present a probability model over non-negative integer valued matrices with possibly unbounded number of columns. This model can play the role of the prior in an nonparametric Bayesian learning scenario where both the latent features and the number of their occurrences are unknown. We use this prior together with a likelihood model for unsupervised learning from images using a Markov Chain Monte Carlo inference algorithm.

58 *Sparse Overcomplete Latent Variable Decomposition of Counts Data*

Madhusudana Shashanka shashanka@cns.bu.edu
Mars Inc., Boston University
Bhiksha Raj bhiksha@merl.com
Mitsubishi Electric Research Laboratories
Paris Smaragdis paris@alum.mit.edu
Adobe Systems Inc.

An important problem in many fields is the analysis of counts data to extract meaningful latent components. Methods like Probabilistic Latent Semantic Analysis (PLSA) and Latent Dirichlet Allocation (LDA) have been proposed for this purpose. However, they are limited in the number of components they can extract and also do not have a provision to control the "expressiveness" of the extracted components. In this paper, we present a learning formulation to address these limitations by employing the notion of sparsity. We start with the PLSA framework and use an entropic prior in a maximum a posteriori formulation to enforce sparsity. We show that this allows the extraction of overcomplete sets of latent components which better characterize the data. We present experimental evidence of the utility of such representations.

59 *Density Estimation under Independent Similarly Distributed Sampling Assumptions*

Tony Jebara jebara@cs.columbia.edu
Columbia University
Yingbo Song yingbo@cs.columbia.edu
Columbia University
Kapil Thadani kapil@cs.columbia.edu
Columbia University

A method is proposed for semiparametric estimation where parametric and nonparametric criteria are exploited in density estimation and unsupervised learning. This is accomplished by making sampling assumptions on a dataset that smoothly interpolate between the extreme of independently distributed (or *id*) sample data (as in nonparametric kernel density estimators) to the extreme of independent *identically* distributed (or *iid*) sample data. This article makes independent *similarly* distributed (or *isd*) sampling assumptions and interpolates between these two using a scalar parameter. The parameter controls a Bhattacharyya affinity penalty between pairs of distributions on samples. Surprisingly, the *isd* method maintains certain consistency and unimodality properties akin to maximum likelihood estimation. The proposed *isd* scheme is an alternative for handling nonstationarity in data without making drastic hidden variable assumptions which often make estimation difficult and laden with local optima. Experiments in density estimation on a variety of datasets confirm the superiority of *isd* over *iid* estimation, *id* estimation and mixture modeling.

Spotlight presentation, Tuesday, 9:50AM.

60 *Direct Importance Estimation with Model Selection and Its Application to Covariate Shift Adaptation*

Motoaki Kawanabe motoaki.kawanabe@first.fraunhofer.de
Fraunhofer FIRST.IDA
Paul von Buenau buenau@cs.tu-berlin.de
TU Berlin
Hisashi Kashima kashi_pong@yahoo.co.jp
IBM Research, Tokyo Research Laboratory
Masashi Sugiyama sugi@cs.titech.ac.jp
Tokyo Institute of Technology, Department of Computer Science
Shinichi Nakajima nakajima.s@nikon.co.jp
Nikon Corporation

When training and test samples follow different input distributions (i.e., the situation called *covariate shift*), the maximum likelihood estimator is known to lose its consistency. For regaining consistency, the log-likelihood terms need to be weighted according to the *importance* (i.e., the ratio of test and training input densities). Thus, accurately estimating the importance is one of the key tasks in covariate shift adaptation. A naive approach is to first estimate training and test input densities and then estimate the importance by the ratio of the density estimates. However, since density estimation is a hard problem, this approach tends to perform poorly especially in high dimensional cases. In this paper, we propose a direct importance estimation method that does not require the input density estimates. Our method is equipped with a natural model selection procedure so tuning parameters such as the kernel width can be objectively optimized. This is an advantage over a recently developed method of direct importance estimation. Simulations illustrate the usefulness of our approach.

61 *Catching Up Faster in Bayesian Model Selection and Model Averaging*

Tim van Erven tim.van.erven@cwi.nl
Centrum voor Wiskunde en Informatica
Peter Grunwald pdg@cwi.nl
Centrum voor Wiskunde en Informatica
Steven de Rooij rooij@cwi.nl
Centrum voor Wiskunde en Informatica (CWI)

Bayesian model averaging, model selection and their approximations such as BIC are generally statistically consistent, but sometimes achieve slower rates of convergence than other methods such as AIC and leave-one-out cross-validation. On the other hand, these other methods can be inconsistent. We identify the "catch-up phenomenon" as a novel explanation for the slow convergence of Bayesian methods. Based on this analysis we define the switch-distribution, a modification of the Bayesian marginal distribution. We prove that in many situations model selection and prediction based on the switch-distribution is both consistent and achieves optimal convergence rates, thereby resolving the AIC/BIC dilemma. The method is practical; we give an efficient implementation.

Spotlight presentation, Tuesday, 11:50AM.

62 *Hidden Common Cause Relations in Relational Learning*

Ricardo Silva rbas@gatsby.ucl.ac.uk
Gatsby Computational Neuroscience Unit, UCL
Wei Chu chuwei@cs.columbia.edu
Center for Computational Learning Systems, Columbia University
Zoubin Ghahramani zoubin@eng.cam.ac.uk
University of Cambridge & CMU

When predicting class labels for objects within a relational database, it is often helpful to consider a model for relationships: this allows for information between class labels to be shared and to improve prediction performance. However, there are different ways by which objects can be related within a relational database. One traditional way corresponds to a Markov network structure: each existing relation is represented by an undirected edge. This encodes that, conditioned on input features, each object label is independent of other object labels given its neighbors in the graph. However, there is no reason why Markov networks should be the only representation of choice for symmetric dependence structures. Here we discuss the case when relationships are postulated to exist due to hidden common causes. We discuss how the resulting graphical model differs from Markov networks, and how it describes different types of real-world relational processes. A Bayesian nonparametric classification model is built upon this graphical representation and evaluated with several empirical studies.

Spotlight presentation, Tuesday, 9:50AM.

63 *Predictive Matrix-Variate t Models*

Shenghuo Zhu shenghuo@gmail.com
NEC Laboratories America, Inc.
Kai Yu kai.yu.cool@gmail.com
NEC Laboratories America, Inc.
Yihong Gong ygong@sv.nec-labs.com
NEC Laboratories

It is becoming increasingly important to learn from a partially-observed random matrix and predict its missing elements. We assume that the entire matrix is a single sample drawn from a matrix-variate t distribution and suggest a matrix-variate t model (MVTM) to predict those missing elements. We show that MVTM generalizes a range of known probabilistic models, and automatically performs model selection to encourage sparse predictive models. Due to the non-conjugacy of its prior, it is difficult to make predictions by computing the mode or mean of the posterior distribution. We suggest an optimization

method that sequentially minimizes a convex upper-bound of the log-likelihood, which is very efficient and scalable. The experiments on a toy data and EachMovie dataset show a good predictive accuracy of the model.

64 *SpAM: Sparse Additive Models*

John Lafferty lafferty@cs.cmu.edu
Carnegie Mellon
Han Liu hanliu@cs.cmu.edu
Machine Learning Department, Carnegie Mellon University
Pradeep Ravikumar pradeepr@cs.cmu.edu
Carnegie Mellon University
Larry Wasserman larry@stat.cmu.edu
Carnegie Mellon

We present a new class of models for high-dimensional nonparametric regression and classification called sparse additive models (SpAM). Our methods combine ideas from sparse linear modeling and additive nonparametric regression. We derive a method for fitting the models that is effective even when the number of covariates is larger than the sample size. A statistical analysis of the properties of SpAM is given together with empirical results on synthetic and real data, showing that SpAM can be effective in fitting sparse nonparametric models in high dimensional data.

Spotlight presentation, Tuesday, 9:50AM.

65 *The Generalized FITC Approximation*

Andrew Naish-Guzman agpn2@cam.ac.uk
University of Cambridge, Computer Laboratory
Sean Holden sbh11@cl.cam.ac.uk
Computer Laboratory, Cambridge University

We present an efficient generalization of the sparse pseudo-input Gaussian process (SPGP) model developed by Snelson and Ghahramani (2005), applying it to binary classification problems. By taking advantage of the SPGP prior covariance structure, we derive a numerically stable algorithm with $O(NM^2)$ training complexity—asymptotically the same as related sparse methods such as the informative vector machine (Lawrence et al. (2003)), but which more faithfully represents the posterior. We present experimental results for several benchmark problems showing that in many cases this allows an exceptional degree of sparsity without compromising accuracy. Following Snelson and Ghahramani, we locate pseudo-inputs by gradient ascent on the marginal likelihood, but exhibit occasions when this is likely to fail, for which we suggest alternative solutions.

Spotlight presentation, Tuesday, 11:50AM.

66 *Gaussian Process Models for Link Analysis and Transfer Learning*

Kai Yu kai.yu.cool@gmail.com
NEC Laboratories America, Inc.

Wei Chu chuwei@cs.columbia.edu
Center for Computational Learning Systems, Columbia University

In this paper we develop a Gaussian process (GP) framework to model a collection of reciprocal random variables defined on the *edges* of a network. We show how to construct GP priors, i.e., covariance functions, on the edges of directed, undirected, and bipartite graphs. The model suggests an intimate connection between *link prediction* and *transfer learning*, which were traditionally considered two separate research topics. Though a straightforward GP inference has a very high complexity, we develop an efficient learning algorithm that can handle a large number of observations. The experimental results on several real-world data sets verify superior learning capacity.

67 *Robust Regression with Twinned Gaussian Processes*

Andrew Naish-Guzman agpn2@cam.ac.uk
University of Cambridge, Computer Laboratory
Sean Holden sbh11@cl.cam.ac.uk
Computer Laboratory, Cambridge University

We propose a Gaussian process (GP) framework for robust inference in which a GP prior on the mixing weights of a two-component noise model augments the standard process over latent function values. This approach is a generalization of the mixture likelihood used in traditional robust GP regression, and a specialization of the GP mixture models suggested by Tresp (2000) and Rasmussen and Ghahramani (2002). The value of this restriction is in its tractable expectation propagation updates, which allow for faster inference and model selection, and better convergence than the standard mixture. An additional benefit over the latter method lies in our ability to incorporate knowledge of the noise domain to influence predictions, and to recover with the predictive distribution information about the outlier distribution via the gating process. The model has asymptotic complexity equal to that of conventional robust methods, but yields more confident predictions on benchmark problems than classical heavy-tailed models and exhibits improved stability for data with clustered corruptions, for which they fail altogether. We show further how our approach can be used without adjustment for more smoothly heteroscedastic data, and suggest how it could be extended to more general noise models. We also address similarities with the work of Goldberg et al. (1998), and the more recent contributions of Tresp, and Rasmussen and Ghahramani.

68 *Selecting Observations against Adversarial Objectives*

Andreas Krause krausea@cs.cmu.edu
Computer Science Department, Carnegie Mellon University
Brendan McMahan brendan.mcmahan@gmail.com
Google, Inc.
Carlos Guestrin guestrin@cs.cmu.edu
Carnegie Mellon University
Anupam Gupta anupamg@cs.cmu.edu
Carnegie Mellon University

In many applications, one has to actively select among a set of expensive observations before making an informed decision. Often, we want to select observations which perform well when evaluated with an objective function chosen by an adversary. Examples include minimizing the maximum posterior variance in Gaussian Process regression, robust experimental design, and sensor placement for outbreak detection. In this paper, we present the

Submodular Saturation algorithm, a simple and efficient algorithm with strong theoretical approximation guarantees for the case where the possible objective functions exhibit submodularity, an intuitive diminishing returns property. Moreover, we prove that better approximation algorithms do not exist unless NP-complete problems admit efficient algorithms. We evaluate our algorithm on several real-world problems. For Gaussian Process regression, our algorithm compares favorably with state-of-the-art heuristics described in the geostatistics literature, while being simpler, faster and providing theoretical guarantees. For robust experimental design, our algorithm performs well compared to SDP-based algorithms.

Spotlight presentation, Tuesday, 5:20PM.

69 *A Kernel Statistical Test of Independence*

ARTHUR GRETTON arthur@tuebingen.mpg.de
MPI for Biological Cybernetics
CHOON HUI TEO choonhui.teo@rsise.anu.edu.au
SML, NICTA
KENJI FUKUMIZU fukumizu@ism.ac.jp
Institute of Statistical Mathematics
LE SONG lesong@it.usyd.edu.au
NICTA and School of Information Technologies, the University of Sydney
BERNHARD SCHOLKOPF bs@tuebingen.mpg.de
MPI for Biological Cybernetics
ALEX SMOLA alex.smola@gmail.com
NICTA

Whereas kernel measures of independence have been widely applied in machine learning (notably in kernel ICA), there is as yet no method to determine whether they have detected statistically significant dependence. We provide a novel test of the independence hypothesis for one particular kernel independence measure, the Hilbert-Schmidt independence criterion (HSIC). The resulting test costs $O(m^2)$, where m is the sample size. We demonstrate that this test outperforms established contingency table-based tests. Finally, we show the HSIC test also applies to text (and to structured data more generally), for which no other independence test presently exists.

Spotlight presentation, Tuesday, 3:20PM.

70 *Testing for Homogeneity with Kernel Fisher Discriminant Analysis*

ZAID HARCHAOUI zharchao@tsi.enst.fr
GET/Telecom Paris
FRANCIS BACH francis.bach@mines.org
INRIA - Ecole Normale Superieure
MOULINES ERIC moulines@tsi.enst.fr
GET/Telecom-Paris

We propose to test for the homogeneity of two samples by using Kernel Fisher discriminant Analysis. This provides us with a consistent nonparametric test statistic, for which we derive the asymptotic distribution under the null hypothesis. We give experimental evidence of the relevance of our method on both artificial and real datasets.

Demos:

9 *Adaptive Bottle*

Kamil Adiloglu
Berlin University of Technology
Robert Annies
Berlin University of Technology
Yon Visell
McGill University
Karmen Franinovic
University of Applied Sciences & Arts
Carlo Drioli
University of Verona

10 *Basal-ganglia-inspired Hierarchical Reinforcement Learning in an AIBO robot*

Robert Hearn
Dartmouth College
Richard Granger
Dartmouth College

11 *CLOP: a Matlab Learning Object Package*

Amir Reza Saffari Azar Alamdari
Graz University of Technology
Isabelle Guyon
Clopinet
Hugo Escalante
Instituto Nacional de Astrofisica, Optica y Electronica
Gokhan Bakir
MPI for Biological Cybernetics
Gavin Cawley
University of East Anglia

http://www.agnostic.inf.ethz.ch/models.php

12 *Holistic Scene Understanding from Visual and Range Data*

Stephen Gould
Stanford University
Morgan Quigley
Stanford University
Andrew Ng
Stanford University
Daphne Koller
Stanford University

http://ai.stanford.edu/˜sgould/vision/3d

13 *Robust Biped Locomotion Using Simple Low-dimensional Control Policies*

Michiel van de Panne
University of British Columbia
Kang Yin
University of British Columbia
Stelian Coros
University of British Columbia
Kevin Loken
Electronic Arts

14 *Tekkotsu Cognitive Robotics*

David Touretzky
Carnegie Mellon University
Ethan Tira-Thompson
Carnegie Mellon University

http://tekkotsu.org

15 *Visualization of DepthMotion Perception by Model Cortical Neurons*

Eric Kong-Chau Tsang
Hong Kong University of Science and Technology
Stanley Yiu Man Lam
Hong Kong University of Science and Technology
Bertram Shi
Hong Kong University of Science and Technology

Wednesday, December 5th

Oral Session — Probabilistic Models and Methods (Chair: William Noble, University of Washington):

8:30AM Invited Talk: *Computational and Statistical Problems in Population Genetics*

NICK PATTERSON nickp@broad.mit.edu
Broad Institute of Harvard and MIT

We very briefly will introduce some key ideas of population genetics and indicate why the basic models are simple to explain but hard to use for analysis. We discuss some modern analytic tools and indicate why they are not completely satisfactory. Finally we will give a list of some problems that are interesting to me and for which even partial solutions would advance the field.

9:30AM *Bayesian Agglomerative Clustering with Coalescents*

YEE WHYE TEH ywteh@gatsby.ucl.ac.uk
Gatsby Computational Neuroscience Unit, UCL
HAL DAUME III me@hal3.name
University of Utah
DANIEL ROY droy@mit.edu
Massachusetts Institute of Technology, CSAIL

We introduce a new Bayesian model for hierarchical clustering based on a prior over trees called Kingman's coalescent. We develop novel greedy and sequential Monte Carlo inferences which operate in a bottom-up agglomerative fashion. We show experimentally the superiority of our algorithms over others, and demonstrate our approach in document clustering and phylolinguistics.

9:50AM Spotlights

Augmented Functional Time Series Representation and Forecasting with Gaussian Processes
NICOLAS CHAPADOS, University of Montreal, ApSTAT Technologies Inc. and YOSHUA BENGIO, University of Montreal.
See abstract, page 144.

Cooled and Relaxed Survey Propagation for MRFs
HAI LEONG CHIEU, National University of Singapore, WEE SUN LEE, National University of Singapore and YEE WHYE TEH, Gatsby Computational Neuroscience Unit, UCL.
See abstract, page 141.

Simulated Annealing: Rigorous finite-time guarantees for optimization on continuous domains
ANDREA LECCHINI-VISINTINI, University of Leicester, JOHN LYGEROS, ETH Zurich and JAN MACIEJOWSKI, Cambridge University.
See abstract, page 103.

Multi-task Gaussian Process Prediction
Chris Williams, Univeristy of Edinburgh, Kian Ming Chai, University of Edinburgh and Edwin Bonilla, School of Informatics, University of Edinburgh.
See abstract, page 143.

Bayesian Co-Training
Shipeng Yu, Siemens Medical Solutions USA, Inc, Balaji Krishnapuram, Computer Aided Diagnosis & Therapy Group, Siemens Medical Solutions, USA, Romer Rosales, Siemens, Harald Steck, Siemens Medical Solutions, Computer-Aided Diagnosis and Therapy and R. Bharat Rao, Siemens Medical Solutions.
See abstract, page 142.

Adaptive Bayesian Inference
Ozgur Sumer, University of Chicago, Umut Acar, Toyota Technological Institute, Alexander T. Ihler, Toyota Technological Insitute at Chicago and Ramgopal R. Mettu, University of Massachusetts Amherst.
See abstract, page 142.

An Analysis of Inference with the Universum
Fabian Sinz, Max Planck Institute f Biological Cybernetics, Olivier Chapelle, Yahoo Research, Alekh Agarwal, EECS, University of California, Berkeley and Bernhard Scholkopf, MPI for Biological Cybernetics.
See abstract, page 143.

Combined discriminative and generative articulated pose and non-rigid shape estimation
Leonid Sigal, Brown University, Alexandru Balan, Brown University and Michael Black, Brown University.
See abstract, page 152.

Agreement-Based Learning
Percy Liang, UC Berkeley, Dan Klein, UC Berkeley and Michael Jordan, University of California, Berkeley.
See abstract, page 140.

Expectation Maximization and Posterior Constraints
Kuzman Ganchev, University of Pennsylvania, Joao Graca, L2F INESC-ID Lisboa and Ben Taskar, University of Pennsylvania.
See abstract, page 140.

10:00AM Break

Oral Session — Probabilistic Representations and Learning (Chair: YeeWhye Teh, Gatsby Computational Neuroscience Unit):

10:30AM *Learning with Tree-Averaged Densities and Distributions*

Sergey Kirshner sergey@cs.ualberta.ca
University of Alberta

We utilize the ensemble of trees framework, a tractable mixture over super-exponential number of tree-structured distributions, to develop a new model for multivariate density estimation. The model is based on the construction of tree-structured copulas – multivariate distributions with uniform on $[0, 1]$ marginals. By averaging over all possible tree structures, the new model can approximate distributions with complex variable dependencies. We propose an EM algorithm to estimate the parameters for these tree-averaged models for both the real-valued and the categorical case. Based on the tree-averaged framework, we propose a new model for joint precipitation amounts data on networks of rain stations.

10:50AM *Non-parametric Modeling of Partially Ranked Data*

Guy Lebanon lebanon@stat.purdue.edu
Purdue University
Yi Mao ymao@ecn.purdue.edu
Purdue University

Statistical models on full and partial rankings of n items are often of limited practical use for large n due to computational consideration. We explore the use of non-parametric models for partially ranked data and derive efficient procedures for their use for large n. The derivations are largely possible through combinatorial and algebraic manipulations based on the lattice of partial rankings. In particular, we demonstrate for the first time a non-parametric coherent and consistent model capable of efficiently aggregating partially ranked data of different types.

11:10AM *Efficient Inference for Distributions on Permutations*

Jonathan Huang jch1@cs.cmu.edu
Carnegie Mellon University
Carlos Guestrin guestrin@cs.cmu.edu
Carnegie Mellon University
Leonidas Guibas guibas@stanford.edu
Stanford University

Permutations are ubiquitous in many real world problems, such as voting, rankings and data association. Representing uncertainty over permutations is challenging, since there are $n!$ possibilities, and typical compact representations, such as graphical models, cannot capture the mutual exclusivity constraints associated with permutations. In this paper, we use the "low-frequency" terms of a Fourier decomposition to represent such distributions compactly. We present *Kronecker conditioning*, a new general, efficient approach for maintaining these distributions directly in the Fourier domain. Low order Fourier-based approximations can lead to functions that do not correspond to valid distributions. To address this problem, we present an efficient quadratic program defined directly in the Fourier domain to project the approximation onto the polytope of legal marginal distributions.

We demonstrate the effectiveness of our approach on a real camera-based multi-people tracking setting.

11:30AM *Exponential Family Predictive Representations of State*

David Wingate wingated@umich.edu
University of Michigan
Satinder Singh Baveja baveja@umich.edu
University of Michigan

In order to represent state in controlled, partially observable, stochastic dynamical systems, some sort of sufficient statistic for history is necessary. Predictive representations of state (PSRs) capture state as statistics of the future. We introduce a new model of such systems called the "Exponential family PSR," which defines as state the time-varying parameters of an exponential family distribution which models n sequential observations in the future. This choice of state representation explicitly connects PSRs to state-of-the-art probabilistic modeling, which allows us to take advantage of current efforts in high-dimensional density estimation, and in particular, graphical models and maximum entropy models. We present a parameter learning algorithm based on maximum likelihood, and we show how a variety of current approximate inference methods apply. We evaluate the quality of our model with reinforcement learning by directly evaluating the control performance of the model.

11:50AM Spotlights

Ensemble Clustering using Semidefinite Programming
Vikas Singh, University of Wisconsin Madison, Lopamudra Mukherjee, Department of Computer Science and Engineerin, University at Buffalo, Jiming Peng, UIUC and Jinhui Xu, State University of New York at Buffalo.
See abstract, page 137.

Kernel Measures of Conditional Dependence
Kenji Fukumizu, Institute of Statistical Mathematics, Arthur Gretton, MPI for Biological Cybernetics, Xiaohai Sun, MPI for Biological Cybernetics and Bernhard Scholkopf, MPI for Biological Cybernetics.
See abstract, page 137.

Better than least squares: comparison of objective functions for estimating linear-nonlinear models
Tatyana Sharpee, The Salk Institute for Biological Studies.
See abstract, page 138.

One-Pass Boosting
Zafer Barutcuoglu, Princeton University, Phil Long, Google and Rocco Servedio, Department of Computer Science, Columbia University.
See abstract, page 137.

Transfer Learning using Kolmogorov Complexity: Basic Theory and Empirical Evaluations
M. M. Mahmud, University of Illinois at Urbana Champaign and Sylvian Ray, University of Illinois at Urbana-Champaign.
See abstract, page 139.

How SVMs can estimate quantiles and the median
Andreas Christmann, Vrije Universiteit Brussel and Ingo Steinwart, CCS-3, Los Alamos National Laboratory.
See abstract, page 138.

A Spectral Regularization Framework for Multi-Task Structure Learning
Andreas Argyriou, University College London, Charles A. Micchelli, SUNY Albany, Massimiliano Pontil, UCL and Yiming Ying, Bristol University.
See abstract, page 139.

Nearest-Neighbor-Based Active Learning for Rare Category Detection
Jingrui He, Carnegie Mellon University, School of Computer Science and Jaime Carbonell, Carnegie Mellon University.
See abstract, page 140.

Experience-Guided Search: A Theory of Attentional Control
Michael Mozer, University of Colorado, Boulder and David Baldwin, Indiana University.
See abstract, page 147.

12:00PM Break

Oral Session — Cognitive Processes (Chair: Mark Steyvers, University of California Irvine):

2:00PM Invited Talk: *Core Knowledge of Number and Geometry*

ELIZABETH SPELKE spelke@wjh.harvard.edu
Harvard University

Formal mathematics is a cultural and historical invention, passed from adults to most children through extensive formal instruction. Nevertheless, mathematics builds on systems of knowledge that emerge independently of instruction or culture. Studies of human infants and nonhuman primates can tell us about the properties of those "core systems" and their interactions. Studies of children at the brink of formal schooling suggest how those systems are harnessed to permit learning of symbolic mathematics.

3:00PM *Markov Chain Monte Carlo with People*

ADAM SANBORN asanborn@gatsby.ucl.ac.uk
Gatsby Computational Neuroscience Unit
THOMAS GRIFFITHS tom_griffiths@berkeley.edu
UC Berkeley

Many formal models of cognition implicitly use subjective probability distributions to capture the assumptions of human learners. Most applications of these models determine these distributions indirectly. We propose a method for directly determining the assumptions of human learners by sampling from subjective probability distributions. Using a correspondence between a model of human choice and Markov chain Monte Carlo (MCMC), we describe a method for sampling from the distributions over objects that people associate with different categories. In our task, subjects choose whether to accept or reject a proposed change to an object. The task is constructed so that these decisions follow an MCMC acceptance rule, defining a Markov chain for which the stationary distribution is the category distribution. We test this procedure for both artificial categories acquired in the laboratory, and natural categories acquired from experience.

3:20PM Spotlights

Object Recognition by Scene Alignment
BRYAN RUSSELL, MIT CSAIL, ANTONIO TORRALBA, CSAIL MIT, CE LIU, Massachusetts Institute of Technology, ROB FERGUS, CSAIL, MIT and WILLIAM FREEMAN, MIT.
See abstract, page 151.

Subspace-Based Face Recognition in Analog VLSI
MIGUEL FIGUEROA, Universidad de Concepcion, GONZALO CARVAJAL, Universidad de Concepcion and WALDO VALENZUELA, Universidad de Concepcion.
See abstract, page 166.

Comparing Bayesian models for multisensory cue combination without mandatory integration
ULRIK BEIERHOLM, Caltech, KONRAD KORDING, Northwestern University, Pysiology and PMnR, LADAN SHAMS, UCLA, Caltech and WEI JI MA, University of Rochester.
See abstract, page 146.

Learning Visual Attributes
VITTORIO FERRARI, University of Oxford and ANDREW ZISSERMAN, University of Oxford.
See abstract, page 151.

Congruence between model and human attention reveals unique signatures of critical visual events
ROBERT PETERS, University of Southern California, Department of Computer Science and LAURENT ITTI, University of Southern California.
See abstract, page 146.

A Bayesian Framework for Cross-Situational Word-Learning
MICHAEL FRANK, Massachusetts Institute of Technology, NOAH GOODMAN, MIT and JOSHUA TENENBAUM, MIT.
See abstract, page 147.

Retrieved context and the discovery of semantic structure
MARC HOWARD, Syracuse University and VINAYAK RAO, Syracuse University.
See abstract, page 149.

Theoretical Analysis of Learning with Reward-Modulated Spike-Timing-Dependent Plasticity
ROBERT LEGENSTEIN, TU Graz, DEJAN PECEVSKI, Graz University of Technology and WOLFGANG MAASS, TU Graz.
See abstract, page 160.

Sequential Hypothesis Testing under Stochastic Deadlines
PETER FRAZIER, ORFE, Princeton University and ANGELA YU, Princeton University.
See abstract, page 148.

3:30PM Break

Oral Session — Systems and Applications (Chair: Fei Sha, University of California Berkeley):

4:00PM *A Constraint Generation Approach to Learning Stable Linear Dynamical Systems*

Sajid Siddiqi siddiqi@cs.cmu.edu
Robotics Institute, Carnegie Mellon University
Byron Boots beb@cs.cmu.edu
Computer Science Dept., Carnegie Mellon University
Geoffrey Gordon ggordon@cs.cmu.edu
CMU, Machine Learning Department

Stability is a desirable characteristic for linear dynamical systems, but it is often ignored by algorithms that learn these systems from data. We propose a novel method for learning stable linear dynamical systems: we formulate an approximation of the problem as a convex program, start with a solution to a relaxed version of the program, and incrementally add constraints to improve stability. Rather than continuing to generate constraints until we reach a feasible solution, we test stability at each step; because the convex program is only an approximation of the desired problem, this early stopping rule can yield a higher-quality solution. We apply our algorithm to the task of learning dynamic textures from image sequences as well as to modeling biosurveillance drug-sales data. The constraint generation approach leads to noticeable improvement in the quality of simulated sequences. We compare our method to those of Lacy and Bernstein, with positive results in terms of accuracy, quality of simulated sequences, and efficiency.

4:20PM *The Infinite Markov Model*

Daichi Mochihashi daichi@cslab.kecl.ntt.co.jp
NTT Communication Science Laboratories
Eiichiro Sumita eiichiro.sumita@atr.jp
ATR Japan

We present a nonparametric Bayesian method of estimating variable order Markov processes up to a theoretically infinite order. By extending a stick-breaking prior, which is usually defined on a unit interval, “vertically” to the trees of infinite depth associated with a hierarchical Chinese restaurant process, our model directly infers the hidden orders of Markov dependences from which each symbol originates. Experiments on character and word sequences in natural language showed that the model has a comparative performance with an exponentially large full-order model, while computationally much efficient in both time and space. We expect that this basic model will also extend to the variable order hierarchical clustering of general data.

4:40PM *A probabilistic model for generating realistic lip movements from speech*

Gwenn Englebienne ge@cs.man.ac.uk
University of Manchester
Tim Cootes tim.cootes@manchester.ac.uk
University of Manchester

MAGNUS RATTRAY magnus.rattray@manchester.ac.uk
University of Manchester

The present work aims to model the correspondence between facial motion and speech. The face and sound are modelled separately, with phonemes being the link between both. We propose a sequential model and evaluate its suitability for the generation of the facial animation from a sequence of phonemes, which we obtain from speech. We evaluate the results both by computing the error between generated sequences and real video, as well as with a rigorous double-blind test with human sub jects. Experiments show that our model compares favourably to other existing methods and that the sequences generated are comparable to real video sequences.

5:00PM *Blind channel identification for speech dereverberation using l1-norm sparse learning*

YUANQING LIN linyuanq@seas.upenn.edu
University of Pennsylvania
JINGDONG CHEN jingdong@research.bell-labs.com
Bell Labs, Alcatel-Lucent
YOUNGMOO KIM ykim@drexel.edu
Drexel University, Electrical & Computer Engineering Dept.
DANIEL LEE ddlee@seas.upenn.edu
University of Pennsylvania

Speech dereverberation remains an open problem after more than three decades of research. The most challenging step in speech dereverberation is blind channel identification (BCI). Although many BCI approaches have been developed, their performance is still far from satisfactory for practical applications. The main difficulty in BCI lies in finding an appropriate acoustic model, which not only can effectively resolve solution degeneracies due to the lack of knowledge of the source, but also robustly models real acoustic environments. This paper proposes a sparse acoustic room impulse response (RIR) model for BCI, that is, an acoustic RIR can be modeled by a sparse FIR filter. Under this model, we show how to formulate the BCI of a single-input multiple-output (SIMO) system into a l1-norm regularized least squares (LS) problem, which is convex and can be solved efficiently with guaranteed global convergence. The sparseness of solutions is controlled by l1-norm regularization parameters. We propose a sparse learning scheme that infers the optimal l1-norm regularization parameters directly from microphone observations under a Bayesian framework. Our results show that the proposed approach is effective and robust, and it yields source estimates in real acoustic environments with high fidelity to anechoic chamber measurements.

5:20PM Spotlights

Extending position/phase-shift tuning to motion energy neurons improves velocity discrimination
YIU MAN LAM, Hong Kong University of Science & Technology and BERTRAM SHI, Hong Kong University of Science & Technology.
See abstract, page 160.

Invariant Common Spatial Patterns: Alleviating Nonstationarities in Brain-Computer Interfacing
BENJAMIN BLANKERTZ, Technical University of Berlin, Fraunhofer FIRST (IDA), MOTOAKI KAWANABE, Fraunhofer FIRST.IDA, RYOTA TOMIOKA,

Fraunhofer FIRST.IDA, Friederike Hohlefeld, Neurophysics Group, Charite Berlin, Vadim Nikulin, Neurophysics Group, Charite, Berlin, Bernstein Center, Berlin and Klaus-Robert Müller, Fraunhofer FIRST.IDA.
See abstract, page 161.

Inferring Neural Firing Rates from Spike Trains Using Gaussian Processes
John Cunningham, Stanford University, Department of Electrical Engineering, Byron Yu, Stanford University, Maneesh Sahani, Gatsby Computational Neuroscience Unit, UCL and Krishna Shenoy, Stanford.
See abstract, page 161.

An in-silico Neural Model of Dynamic Routing through Neuronal Coherence
Devarajan Sridharan, Neurosciences Program, Stanford University, Brian Percival, Stanford University, John Arthur, Stanford University and Kwabena Boahen, Stanford.
See abstract, page 164.

Contraction Properties of VLSI Cooperative Competitive Neural Networks of Spiking Neurons
Emre Neftci, Institute of Neuroinformatics, UNIZ - ETHZ, Elisabetta Chicca, Institute of Neuroinformatics, UNIZ - ETHZ, Giacomo Indiveri, UZH-ETH Zurich, Institute of Neuroinformatics, Jean-Jeacques Slotine, MIT and Rodney Douglas, Institute of Neuroinformatics, UNIZ - ETHZ.
See abstract, page 165.

Measuring Neural Synchrony by Message Passing
Tomasz Rutkowski, Brain Science Institute RIKEN, Advanced Brain Signal Processing Lab, Andrzej Cichocki, Riken BSI, François Vialatte, RIKEN BSI, L.ABSP and Justin Dauwels, RIKEN Brain Science Institute.
See abstract, page 159.

Near-Maximum Entropy Models for Binary Neural Representations of Natural Images
Matthias Bethge, MPI Tuebingen and Philipp Berens, MPI for Biological Cybernetics.
See abstract, page 153.

The rat as particle filter
Nathaniel Daw, Center for Neural Science & Psychology Dept., New York University and Aaron Courville, University of Montreal.
See abstract, page 149.

5:30PM Break

Posters:

1 *One-Pass Boosting*

Zafer Barutcuoglu zbarutcu@princeton.edu
Princeton University
Phil Long plong@google.com
Google
Rocco Servedio rocco@cs.columbia.edu
Department of Computer Science, Columbia University

This paper studies boosting algorithms that make a single pass over a set of base classifiers. We first analyze a one-pass algorithm in the setting of boosting with diverse base classifiers. Our guarantee is the same as the best proved for any boosting algorithm, but our one-pass algorithm is much faster than previous approaches. We next exhibit a random source of examples for which a "picky" variant of AdaBoost that skips poor base classifiers can outperform the standard AdaBoost algorithm, which uses every base classifier, by an exponential factor. Experiments with Reuters and synthetic data show that one-pass boosting can substantially improve on the accuracy of Naive Bayes, and that picky boosting can sometimes lead to a further improvement in accuracy.

Spotlight presentation, Wednesday, 11:50AM.

2 *Ensemble Clustering using Semidefinite Programming*

Vikas Singh vsingh@biostat.wisc.edu
University of Wisconsin Madison
Lopamudra Mukherjee lm37@cse.buffalo.edu
Department of Computer Science and Engineerin, University at Buffalo
Jiming Peng pengj@uiuc.edu
UIUC
Jinhui Xu jinhui@cse.buffalo.edu
State University of New York at Buffalo

We consider the ensemble clustering problem where the task is to 'aggregate' multiple clustering solutions into a single consolidated clustering that maximizes the shared information among given clustering solutions. We obtain several new results for this problem. First, we note that the notion of agreement under such circumstances can be better captured using an agreement measure based on a $2D$ string encoding rather than voting strategy based methods proposed in literature. Using this generalization, we first derive a nonlinear optimization model to maximize the new agreement measure. We then show that our optimization problem can be transformed into a strict 0-1 Semidefinite Program (SDP) via novel convexification techniques which can subsequently be relaxed to a polynomial time solvable SDP. Our experiments indicate improvements not only in terms of the proposed agreement measure but also the existing agreement measures based on voting strategies. We discuss extensive evaluations of the algorithm on clustering and image segmentation databases.

Spotlight presentation, Wednesday, 11:50AM.

3 *Kernel Measures of Conditional Dependence*

Kenji Fukumizu fukumizu@ism.ac.jp
Institute of Statistical Mathematics
Arthur Gretton arthur.gretton@tuebingen.mpg.de
MPI for Biological Cybernetics
Xiaohai Sun xiaohai.sun@tuebingen.mpg.de
MPI for Biological Cybernetics
Bernhard Scholkopf bs@tuebingen.mpg.de
MPI for Biological Cybernetics

We propose a new measure of conditional dependence of random variables, based on normalized cross-covariance operators on reproducing kernel Hilbert spaces. Unlike previous kernel dependence measures, the proposed criterion does not depend on the choice of kernel in the limit of infinite data, for a wide class of kernels. At the same time, it has a straightforward empirical estimate with good convergence behaviour. We discuss the theoretical properties of the measure, and demonstrate its application in experiments.

Spotlight presentation, Wednesday, 11:50AM.

4 *How SVMs can estimate quantiles and the median*

Andreas Christmann christmann.andreas@gmail.com
Vrije Universiteit Brussel
Ingo Steinwart ingosteinwart@yahoo.com
CCS-3, Los Alamos National Laboratory

We investigate kernel-based quantile regression based on the pinball loss and support vector regression based on the eps-insensitive loss. Conditions are given which quarantee that the set of exact minimizers contains only one function. Some results about oracle inequalities and learning rates of these methods are presented.

Spotlight presentation, Wednesday, 11:50AM.

5 *Better than least squares: comparison of objective functions for estimating linear-nonlinear models*

Tatyana Sharpee sharpee@phy.ucsf.edu
The Salk Institute for Biological Studies

This paper compares a family of methods for characterizing neural feature selectivity with natural stimuli in the framework of the linear-nonlinear model. In this model, the neural firing rate is a nonlinear function of a small number of relevant stimulus components. The relevant stimulus dimensions can be found by maximizing one of the family of objective functions, Rényi divergences of different orders. We show that maximizing one of them, Rényi divergence of order 2, is equivalent to least-square fitting of the linear-nonlinear model to neural data. Next, we derive reconstruction errors in relevant dimensions found by maximizing Rényi divergences of arbitrary order in the asymptotic limit of large spike numbers. We find that the smallest errors are obtained with Rényi divergence of order 1, also known as Kullback-Leibler divergence. This corresponds to finding relevant dimensions by maximizing mutual information. Finally, we numerically test how these optimization schemes perform in the regime of low signal-to-noise ratio (small number of spikes and increasing neural noise) for model visual neurons. We find that optimization schemes based on either least square fitting or information maximization perform well

even when number of spikes is small. Information maximization provides slightly, but significantly, better reconstructions than least square fitting. This makes the problem of finding relevant dimensions one of the examples where information-theoretic measures are no more data limited than those derived from least squares.

Spotlight presentation, Wednesday, 11:50AM.

6 *Transfer Learning using Kolmogorov Complexity: Basic Theory and Empirical Evaluations*

M. M. Mahmud mmmahmud@uiuc.edu
University of Illinois at Urbana Champaign
Sylvian Ray ray@cs.uiuc.edu
University of Illinois at Urbana-Champaign

In transfer learning we aim to solve new problems using fewer examples using information gained from solving related problems. Transfer learning has been successful in practice, and extensive PAC analysis of these methods has been developed. However it is not yet clear how to define relatedness between tasks. This is considered as a major problem as it is conceptually troubling and it makes it unclear how much information to transfer and when and how to transfer it. In this paper we propose to measure the amount of information one task contains about another using conditional Kolmogorov complexity between the tasks. We show how existing theory neatly solves the problem of measuring relatedness and transferring the 'right' amount of information in sequential transfer learning in a Bayesian setting. The theory also suggests that, in a very formal and precise sense, no other reasonable transfer method can do much better than our Kolmogorov Complexity theoretic transfer method, and that sequential transfer is always justified. We also develop a practical approximation to the method and use it to transfer information between 8 arbitrarily chosen databases from the UCI ML repository.

Spotlight presentation, Wednesday, 11:50AM.

7 *A Spectral Regularization Framework for Multi-Task Structure Learning*

Andreas Argyriou a.argyriou@cs.ucl.ac.uk
University College London
Charles A. Micchelli cam@math.albany.edu
SUNY Albany
Massimiliano Pontil m.pontil@cs.ucl.ac.uk
UCL
Yiming Ying enxyy@bristol.ac.uk
Bristol University

Learning the common structure shared by a set of supervised tasks is an important practical and theoretical problem. Knowledge of this structure may lead to better generalization performance on the tasks and may also facilitate learning new tasks. We propose a framework for solving this problem, which is based on regularization with spectral functions of matrices. This class of regularization problems exhibits appealing computational properties and can be optimized efficiently by an alternating minimization algorithm. In addition, we provide a necessary and sufficient condition for convexity of the regularizer. We analyze concrete examples of the framework, which are equivalent to regularization with L_p

matrix norms. Experiments on two real data sets indicate that the algorithm scales well with the number of tasks and improves on state of the art statistical performance.

Spotlight presentation, Wednesday, 11:50AM.

8 *Nearest-Neighbor-Based Active Learning for Rare Category Detection*

JINGRUI HE jingruih@cs.cmu.edu
Carnegie Mellon University, School of Computer Science
JAIME CARBONELL jgc@cs.cmu.edu
Carnegie Mellon University

Rare category detection is an open challenge for active learning, especially in the de-novo case (no labeled examples), but of significant practical importance for data mining - e.g. detecting new financial transaction fraud patterns, where normal legitimate transactions dominate. This paper develops a new method for detecting an instance of each minority class via an unsupervised local-density-differential sampling strategy. Essentially a variable-scale nearest neighbor process is used to optimize the probability of sampling tightly-grouped minority classes, subject to a local smoothness assumption of the majority class. Results on both synthetic and real data sets are very positive, detecting each minority class with only a fraction of the actively sampled points required by random sampling and by Pelleg's Interleave method, the prior best technique in the sparse literature on this topic.

Spotlight presentation, Wednesday, 11:50AM.

9 *Agreement-Based Learning*

PERCY LIANG pliang@cs.berkeley.edu
UC Berkeley
DAN KLEIN klein@cs.berkeley.edu
UC Berkeley
MICHAEL JORDAN jordan@cs.berkeley.edu
University of California, Berkeley

The learning of probabilistic models with many hidden variables and non-decomposable dependencies is an important but challenging problem. In contrast to traditional approaches based on approximate inference in a single intractable model, our approach is to train a set of tractable component models by encouraging them to agree on the hidden variables. This allows us to capture non-decomposable aspects of the data while still maintaining tractability. We exhibit an objective function for our approach, derive EM-style algorithms for parameter estimation, and demonstrate their effectiveness on three challenging real-world learning tasks.

Spotlight presentation, Wednesday, 9:50AM.

10 *Expectation Maximization and Posterior Constraints*

KUZMAN GANCHEV kuzman@cis.upenn.edu
University of Pennsylvania
JOAO GRACA joao.graca@l2f.inesc-id.pt
L2F INESC-ID Lisboa

Ben Taskar taskar@cis.upenn.edu
University of Pennsylvania

The expectation maximization (EM) algorithm is a widely used maximum likelihood estimation procedure for statistical models when the values of some of the variables in the model are not observed. Very often, however, our aim is primarily to find a model that assigns values to the latent variables that have intended meaning for our data and maximizing expected likelihood only sometimes accomplishes this. Unfortunately, it is typically difficult to add even simple a-priori information about latent variables in graphical models without making the models overly complex or intractable. In this paper, we present an efficient, principled way to inject rich constraints on the posteriors of latent variables into the EM algorithm. Our method can be used to learn tractable graphical models that satisfy additional, otherwise intractable constraints. Focusing on clustering and the alignment problem for statistical machine translation, we show that simple, intuitive posterior constraints can greatly improve the performance over standard baselines and be competitive with more complex, intractable models.

Spotlight presentation, Wednesday, 9:50AM.

11 *Bayesian Agglomerative Clustering with Coalescents*

Yee Whye Teh, Gatsby Computational Neuroscience Unit, UCL, Hal Daume III, University of Utah and Daniel Roy, Massachusetts Institute of Technology, CSAIL.
Oral presentation, Wednesday, 9:30AM. *See abstract, page 127.*

12 *Learning with Tree-Averaged Densities and Distributions*

Sergey Kirshner, University of Alberta.
Oral presentation, Wednesday, 10:30AM. *See abstract, page 129.*

13 *Bayesian binning beats approximate alternatives: estimating peri-stimulus time histograms*

Dominik Endres dme2@st-andrews.ac.uk
School of Psychology, University of St Andrews
Mike Oram mwo@st-andrews.ac.uk
School of Psychology, University of St Andrews
Johannes Schindelin js108@st-andrews.ac.uk
School of Psychology, University of St Andrews
Peter Foldiak pf2@st-andrews.ac.uk
School of Psychology, University of St Andrews

The peristimulus time historgram (PSTH) and its more continuous cousin, the spike density function (SDF) are staples in the analytic toolkit of neurophysiologists. The former is usually obtained by binning spiketrains, whereas the standard method for the latter is smoothing with a Gaussian kernel. Selection of a bin with or a kernel size is often done in an relatively arbitrary fashion, even though there have been recent attempts to remedy this situation. We develop an exact Bayesian, generative model approach to estimating PSHTs and demonstate its superiority to competing methods. Further advantages of our scheme include automatic complexity control and error bars on its predictions.

14 *Cooled and Relaxed Survey Propagation for MRFs*

Hai Leong Chieu haileong@nus.edu.sg
National University of Singapore
Wee Sun Lee leews@comp.nus.edu.sg
National University of Singapore
Yee Whye Teh ywteh@gatsby.ucl.ac.uk
Gatsby Computational Neuroscience Unit, UCL

We describe a new algorithm, Relaxed Survey Propagation (RSP), for finding MAP configurations in Markov random fields. We compare its performance with state-of-the-art algorithms including the max-product belief propagation, its sequential tree-reweighted variant, residual (sum-product) belief propagation, and tree-structured expectation propagation. We show that it outperforms all approaches for Ising models with mixed couplings, as well as on a web person disambiguation task formulated as a supervised clustering problem.

Spotlight presentation, Wednesday, 9:50AM.

15 *Bayesian Co-Training*

Shipeng Yu shipeng.yu@siemens.com
Siemens Medical Solutions USA, Inc
Balaji Krishnapuram balaji.krishnapuram@siemens.com
Computer Aided Diagnosis & Therapy Group, Siemens Medical Solutions, USA
Romer Rosales romer.rosales@siemens.com
Siemens
Harald Steck harald.steck@siemens.com
Siemens Medical Solutions, Computer-Aided Diagnosis and Therapy
R. Bharat Rao bharat.rao@siemens.com
Siemens Medical Solutions

We propose a Bayesian undirected graphical model for co-training, or more generally for semi-supervised multi-view learning. This makes explicit the previously unstated assumptions of a large class of co-training type algorithms, and also clarifies the circumstances under which these assumptions fail. Building upon new insights from this model, we propose an improved method for co-training, which is a novel co-training kernel for Gaussian process classifiers. The resulting approach is convex and avoids local-maxima problems, unlike some previous multi-view learning methods. Furthermore, it can automatically estimate how much each view should be trusted, and thus accommodate noisy or unreliable views. Experiments on toy data and real world data sets illustrate the benefits of this approach.

Spotlight presentation, Wednesday, 9:50AM.

16 *Adaptive Bayesian Inference*

Ozgur Sumer osumer@cs.uchicago.edu
University of Chicago
Umut Acar umut@tti-c.org
Toyota Technological Institute
Alexander T. Ihler ihler@tti-c.org
Toyota Technological Insitute at Chicago

RAMGOPAL R. METTU mettu@ecs.umass.edu
University of Massachusetts Amherst

Motivated by stochastic systems in which observed evidence and conditional dependencies between states of the network change over time, and certain quantities of interest (marginal distributions, likelihood estimates etc.) must be updated, we study the problem of dynamic inference in tree-structured Bayesian networks. We describe an algorithm for dynamic inference that handles a broad range of changes to the network and is able to maintain marginal distributions, MAP estimates, and data likelihoods in all expected logarithmic time. We give an implementation of our algorithm and provide experiments that show that the algorithm can yield up to two orders of magnitude speedups on answering queries and responding to dynamic changes over the sum-product algorithm.

Spotlight presentation, Wednesday, 9:50AM.

17 *An Analysis of Inference with the Universum*

FABIAN SINZ fabee@tuebingen.mpg.de
Max Planck Institute f Biological Cybernetics
OLIVIER CHAPELLE chap@yaho-inc.com
Yahoo Research
ALEKH AGARWAL alekhagarwal@gmail.com
EECS, University of California, Berkeley
BERNHARD SCHOLKOPF bs@tuebingen.mpg.de
MPI for Biological Cybernetics

We study a pattern classification algorithm which has recently been proposed by Vapnik and coworkers. It builds on a new inductive principle which assumes that in addition to positive and negative data, a third class of data is available, termed the Universum. We assay the behavior of the algorithm by establishing links with Fisher discriminant analysis and oriented PCA, as well as with an SVM in a projected subspace (or, equivalently, with a data-dependent reduced kernel). We also provide experimental results.

Spotlight presentation, Wednesday, 9:50AM.

18 *Non-parametric Modeling of Partially Ranked Data*

GUY LEBANON, Purdue University and YI MAO, Purdue University.
Oral presentation, Wednesday, 10:50AM. *See abstract, page 129.*

19 *Efficient Inference for Distributions on Permutations*

JONATHAN HUANG, Carnegie Mellon University, CARLOS GUESTRIN, Carnegie Mellon University and LEONIDAS GUIBAS, Stanford University.
Oral presentation, Wednesday, 11:10AM. *See abstract, page 129.*

20 *Multi-task Gaussian Process Prediction*

CHRIS WILLIAMS ckiw@inf.ed.ac.uk
Univeristy of Edinburgh
KIAN MING CHAI k.m.a.chai@sms.ed.ac.uk
University of Edinburgh

Edwin Bonilla edwin.bonilla@ed.ac.uk
School of Informatics, University of Edinburgh

In this paper we investigate multi-task learning in the context of Gaussian Processes (GP). We propose a model that learns a shared covariance function on input-dependent features and a "free-form" covariance matrix over tasks. This allows for good flexibility when modelling inter-task dependencies while avoiding the need for large amounts of data for training. We show that under the assumption of noise-free observations and block design, predictions for a given task only depend on its target values and therefore a cancellation of inter-task transfer occurs. We evaluate the benefits of our model on two practical applications: a compiler performance prediction problem and an exam score prediction task. Additionally, we make use of GP approximations and properties of our model in order to provide scalability to large data sets.

Spotlight presentation, Wednesday, 9:50AM.

21 *Augmented Functional Time Series Representation and Forecasting with Gaussian Processes*

Nicolas Chapados chapados@iro.umontreal.ca
University of Montreal, ApSTAT Technologies Inc.
Yoshua Bengio bengioy@iro.umontreal.ca
University of Montreal

We introduce a functional representation of time series which allows forecasts to be performed over an unspecified horizon with progressively-revealed information sets. By virtue of using Gaussian processes, a complete covariance matrix between forecasts at several time-steps is available. This information is put to use in an application to actively trade price spreads between commodity futures contracts. The approach delivers impressive out-of-sample risk-adjusted returns after transaction costs on a portfolio of 30 spreads.

Spotlight presentation, Wednesday, 9:50AM.

22 *A Constraint Generation Approach to Learning Stable Linear Dynamical Systems*

Sajid Siddiqi, Robotics Institute, Carnegie Mellon University,
Byron Boots, Computer Science Dept., Carnegie Mellon University
and Geoffrey Gordon, CMU, Machine Learning Department.
Oral presentation, Wednesday, 4:00PM. *See abstract, page 134.*

23 *Exponential Family Predictive Representations of State*

David Wingate, University of Michigan and Satinder Singh Baveja, University of Michigan.
Oral presentation, Wednesday, 11:30AM. *See abstract, page 130.*

24 *The Infinite Markov Model*

Daichi Mochihashi, NTT Communication Science Laboratories and Eiichiro Sumita, ATR Japan.
Oral presentation, Wednesday, 4:20PM. *See abstract, page 134.*

25 *Unconstrained On-line Handwriting Recognition with Recurrent Neural Networks*

ALEX GRAVES alex.graves@gmail.com
Dalle Molle Institue for AI, Technical University of Munich
SANTIAGO FERNANDEZ santiago@idsia.ch
IDSIA
JUERGEN SCHMIDHUBER juergen@idsia.ch
IDSIA, Switzerland, TU Munich, Germany
MARCUS LIWICKI liwicki@iam.unibe.ch
Institute of Computer Science & Applied Math, University of Bern, Switzerland
HORST BUNKE bunke@iam.unibe.ch
Institute of Computer Science & Applied Math, University of Bern, Switzerland

On-line handwriting recognition is unusual among sequence labelling tasks in that the underlying generator of the observed data, i.e. the movement of the pen, is recorded directly. However, the raw data can be difficult to interpret because each letter is spread over many pen locations. As a consequence, sophisticated pre-processing is required to obtain inputs suitable for conventional sequence labelling algorithms, such as HMMs. In this paper we describe a system capable of directly transcribing raw on-line handwriting data. The system consists of a recurrent neural network trained for sequence labelling, combined with a probabilistic language model. In experiments on an unconstrained on-line database, we record excellent results using either raw or pre-processed data, well outperforming a benchmark HMM in both cases.

26 *A probabilistic model for generating realistic lip movements from speech*

GWENN ENGLEBIENNE, University of Manchester, TIM COOTES, University of Manchester and MAGNUS RATTRAY, University of Manchester.
Oral presentation, Wednesday, 4:40PM. See abstract, page 134.

27 *Blind channel identification for speech dereverberation using l1-norm sparse learning*

YUANQING LIN, University of Pennsylvania, JINGDONG CHEN, Bell Labs, Alcatel-Lucent, YOUNGMOO KIM, Drexel University, Electrical & Computer Engineering Dept. and DANIEL LEE, University of Pennsylvania.
Oral presentation, Wednesday, 5:00PM. See abstract, page 135.

28 *Modeling Natural Sounds with Modulation Cascade Processes*

RICHARD TURNER turner@gatsby.ucl.ac.uk
Gatsby Computational Neuroscience Unit, UCL
MANEESH SAHANI maneesh@gatsby.ucl.ac.uk
Gatsby Computational Neuroscience Unit, UCL

Natural sounds are structured on many time-scales. A typical segment of speech, for example, contains features that span four orders of magnitude: Sentences (˜1s); phonemes (˜0.1s); glottal pulses (˜0.01s); and formants (<0.001s). The auditory system uses information from each of these time-scales to solve complicated tasks such as auditory scene analysis. One route toward understanding how auditory processing accomplishes this analysis is to build neuroscience-inspired algorithms which solve similar tasks and to compare

the properties of these algorithms with properties of auditory processing. There is however a discord: Current machine-audition algorithms largely concentrate on the shorter time-scale structures in sounds, and the longer structures are ignored. The reason for this is two-fold. Firstly, it is a difficult technical problem to construct an algorithm that utilises both sorts of information. Secondly, it is computationally demanding to simultaneously process data both at high resolution (to extract short temporal information) and for long duration (to extract long temporal information). The contribution of this work is to develop a new statistical model for natural sounds that captures structure across a wide range of time-scales, and to provide efficient learning and inference algorithms. We demonstrate the success of this approach on a missing data task.

29 *Comparing Bayesian models for multisensory cue combination without mandatory integration*

Ulrik Beierholm beierh@caltech.edu
Caltech
Konrad Kording konrad@koerding.com
Northwestern University, Pysiology and PMnR
Ladan Shams ladan@psych.ucla.edu
UCLA, Caltech
Wei Ji Ma weijima@gmail.com
University of Rochester

Bayesian models of multisensory perception traditionally address the problem of estimating a variable that is assumed to be the underlying cause of two sensory signals. The brain, however, has to solve a more general problem: it has to establish which signals come from the same source and should be integrated, and which ones do not and should be segregated. In the last couple of years, a few models have been proposed to solve this problem in a Bayesian fashion. One of these has the strength that it formalizes the causal structure of sensory signals. We describe these models and conduct an experiment to test human performance in an auditory-visual spatial localization task in which integration is not mandatory. We find that the causal Bayesian inference model accounts for the data better than other models.

Spotlight presentation, Wednesday, 3:20PM.

30 *Congruence between model and human attention reveals unique signatures of critical visual events*

Robert Peters rjpeters@klab.caltech.edu
University of Southern California, Department of Computer Science
Laurent Itti itti@usc.edu
University of Southern California

Current computational models of bottom-up and top-down components of attention are predictive of eye movements across a range of stimuli and of simple, fixed visual tasks (such as visual search for a target among distractors). However, to date there exists no computational framework which can reliably mimic human gaze behavior in more complex environments and tasks, such as driving a vehicle through traffic. Here, we develop a hybrid computational/behavioral framework, combining simple models for bottom-up salience and top-down relevance, and looking for changes in the predictive power of these components at different critical event times during 4.7 hours (500,000 video frames) of observers playing car racing and flight combat video games. This approach is motivated by

our observation that the predictive strengths of the salience and relevance models exhibit reliable temporal signatures during critical event windows in the task sequence—for example, when the game player directly engages an enemy plane in a flight combat game, the predictive strength of the salience model increases significantly, while that of the relevance model decreases significantly. Our new framework combines these temporal signatures to implement several event detectors. Critically, we find that an event detector based on fused behavioral and stimulus information (in the form of the model's predictive strength) is much stronger than detectors based on behavioral information alone (eye position) or image information alone (model saliency maps). This approach to event detection, based on eye tracking combined with computational models applied to the visual input, may have useful applications as a less-invasive alternative to other event detection approaches based on neural signatures derived from EEG or fMRI recordings.

Spotlight presentation, Wednesday, 3:20PM.

31 *Experience-Guided Search: A Theory of Attentional Control*

Michael Mozer mozer@cs.colorado.edu
University of Colorado, Boulder
David Baldwin baldwind@indiana.edu
Indiana University

People perform a remarkable range of tasks that require search of the visual environment for a target item among distractors. The Guided Search model (Wolfe, 1994, 2007), or GS, is perhaps the best developed psychological account of human visual search. To prioritize search, GS assigns saliency to locations in the visual field. Saliency is a linear combination of activations from retinotopic maps representing primitive visual features. GS includes heuristics for setting the gain coefficient associated with each map. Variants of GS have formalized the notion of optimization as a principle of attentional control (e.g., Baldwin & Mozer, 2006; Navalpakkam & Itti, 2006; Rao et al., 2002), but every GS-like model must be 'dumbed down' to match human data, e.g., by corrupting the saliency map with noise and by imposing arbitrary restrictions on gain modulation. We propose a principled probabilistic formulation of GS, called Experience-Guided Search (EGS), based on a generative model of the environment that makes three claims: (1) Feature detectors produce Poisson spike trains whose rates are conditioned on feature type and whether the feature belongs to a target or distractor; (2) the environment and/or task is nonstationary and can change over a sequence of trials; and (3) a prior specifies that features are more likely to be present for target than for distractors. Through experience, EGS infers latent environment variables that determine the gains for guiding search. Control is thus cast as probabilistic inference, not optimization. We show that EGS can replicate a range of human data from visual search, including data that GS does not address.

Spotlight presentation, Wednesday, 11:50AM.

32 *Markov Chain Monte Carlo with People*

Adam Sanborn, Gatsby Computational Neuroscience Unit and Thomas Griffiths, UC Berkeley.
Oral presentation, Wednesday, 3:00PM. *See abstract, page 132.*

33 *A Bayesian Framework for Cross-Situational Word-Learning*

Michael Frank mcfrank@mit.edu
Massachusetts Institute of Technology
Noah Goodman ndg@mit.edu
MIT
Joshua Tenenbaum jbt@mit.edu
MIT

For infants, early word learning is a chicken-and-egg problem. One way to learn a word is to observe that it co-occurs with a particular referent across different situations. Another way is to use the social context of an utterance to infer the intended referent of a word. Here we present a Bayesian model of cross-situational word learning, and an extension of this model that also learns which social cues are relevant to determining reference. We test our model on a small corpus of mother-infant interaction and find it performs better than competing models. Finally, we show that our model accounts for experimental phenomena including mutual exclusivity, fast-mapping, and generalization from social cues.

Spotlight presentation, Wednesday, 3:20PM.

34 *Learning and using relational theories*

Charles Kemp ckemp@mit.edu
MIT
Joshua Tenenbaum jbt@mit.edu
MIT
Noah Goodman ndg@mit.edu
MIT

Much of human knowledge is organized into sophisticated systems that are often called intuitive theories. We propose that intuitive theories are mentally represented in a logical language, and that the subjective complexity of a theory is determined by the length of its representation in this language. This complexity measure helps to explain how theories are learned from relational data, and how theories support inductive inferences about unobserved relations. We describe two behavioral experiments that test our approach, and show that our measure accounts better for subjective complexity than a measure developed by Nelson Goodman.

35 *Sequential Hypothesis Testing under Stochastic Deadlines*

Peter Frazier pfrazier@princeton.edu
ORFE, Princeton University
Angela Yu ajyu@princeton.edu
Princeton University

Most models of decision-making in neuroscience assume an infinite horizon, which yields an optimal solution that integrates evidence up to a fixed decision threshold. However, under most experimental as well as naturalistic behavioral settings, the decision has to be made before some finite deadline, which is often experienced as a stochastic quantity, either due to variable external constraints or internal timing uncertainty. In this work, we formulate this problem as sequential hypothesis testing under a stochastic horizon. We use dynamic programming tools to show that, for a large class of deadline distributions, the Bayes-optimal solution requires integrating evidence up to a threshold that declines

monotonically over time. We will use numerical simulations to illustrate the optimal policy in the special cases of a fixed deadline and one that is drawn from a gamma distribution.

Spotlight presentation, Wednesday, 3:20PM.

36 *Retrieved context and the discovery of semantic structure*

Marc Howard marc@memory.syr.edu
Syracuse University
Vinayak Rao vsrao@syr.edu
Syracuse University

Semantic memory refers to our knowledge of facts and relationships between concepts. A successful semantic memory depends on inferring relationships between items that are not explicitly taught. Recent mathematical modeling of episodic memory argues that episodic recall relies on retrieval of a gradually-changing representation of temporal context. We show that retrieved context enables the development of a global memory space that reflects relationships between all items that have been previously learned. When newly-learned information is integrated into this structure, it is placed in some relationship to all other items, even if that relationship has not been explicitly learned. We demonstrate this effect for global semantic structures shaped topologically as a ring, and as a two-dimensional sheet. We also examined the utility of this learning algorithm for learning a more realistic semantic space by training it on a large pool of synonym pairs. Retrieved context enabled the model to "infer" relationships between synonym pairs that had not yet been presented.

Spotlight presentation, Wednesday, 3:20PM.

37 *The rat as particle filter*

Nathaniel Daw daw@cns.nyu.edu
Center for Neural Science & Psychology Dept., New York University
Aaron Courville aaron.courville@gmail.com
University of Montreal

The core tenet of Bayesian modeling is that subjects represent beliefs as distributions over possible hypotheses. Such models have fruitfully been applied to the study of learning in the context of animal conditioning experiments (and anologously designed human learning tasks), where they explain phenomena such as retrospective revaluation that seem to demonstrate that subjects entertain multiple hypotheses simultaneously. However, a recent quantitative analysis of individual subject records by Gallistel and colleagues cast doubt on a very broad family of conditioning models by showing that all of the key features the models capture about even simple learning curves are artifacts of averaging over subjects. Rather than smooth learning curves (which Bayesian models interpret as revealing the gradual tradeoff from prior to posterior as data accumulate), subjects acquire suddenly, and their predictions continue to fluctuate abruptly. These data demand revisiting the model of the individual versus the ensemble, and also raise the worry that more sophisticated behaviors thought to support Bayesian models might also emerge artifactually from averaging over the simpler behavior of individuals. We suggest that the suddenness of changes in subjects' beliefs (as expressed in conditioned behavior) can be modeled by assuming they are conducting inference using sequential Monte Carlo sampling with a small number of samples — one, in our simulations. Ensemble behavior resembles exact Bayesian models since, as in particle filters, it averages over many samples. Further,

the model is capable of exhibiting sophisticated behaviors like retrospective revaluation at the ensemble level, even given minimally sophisticated individuals that do not track uncertainty from trial to trial. These results point to the need for more sophisticated experimental analysis to test Bayesian models, and refocus theorizing on the individual, while at the same time clarifying why the ensemble may be of interest.

Spotlight presentation, Wednesday, 5:20PM.

38 *The Noisy-Logical Distribution and its Application to Causal Inference*

HongJing Lu hongjing@hkucc.hku.hk
Psychology Department, UCLA
Alan Yuille yuille@stat.ucla.edu
UCLA

We describe a novel noisy-logical distribution for representing the distribution of a binary output variable conditioned on multiple binary input variables. The distribution is represented in terms of noisy-or's and noisy-and-not's of causal features which are conjunctions of the binary inputs. The standard noisy-or and noisy-and-not models, used in causal reasoning and artificial intelligence, are special cases of the noisy-logical distribution. We prove that the noisy-logical distribution is complete in the sense that it can represent all conditional distributions provided a sufficient number of causal factors are used. We illustrate the noisy-logical distribution by showing that it can account for new experimental findings on how humans perform causal reasoning in more complex contexts. Finally, we speculate on the use of the noisy-logical distribution for causal reasoning and artificial intelligence.

39 *Optimal models of sound localization by barn owls*

Brian Fischer fischerb@caltech.edu
California Institute of Technology

Sound localization by barn owls is commonly modeled as a matching procedure where localization cues derived from auditory inputs are compared to stored templates. While the matching models can explain properties of neural responses, no model explains how the owl resolves spatial ambiguity in the localization cues to produce accurate localization near the center of gaze. Here, we examine two models for the barn owl's sound localization behavior. First, we consider a maximum likelihood estimator in order to further evaluate the cue matching model. Second, we consider a maximum a posteriori estimator to test if a Bayesian model with a prior that emphasizes directions near the center of gaze can reproduce the owl's localization behavior. We show that the maximum likelihood estimator can not reproduce the owl's behavior, while the maximum a posteriori estimator is able to match the behavior. This result suggests that the standard cue matching model will not be sufficient to explain sound localization behavior in the barn owl. The Bayesian model provides a new framework for analyzing sound localization in the barn owl and leads to predictions about the owl's localization behavior.

40 *Modelling motion primitives and their timing in biologically executed movements*

Marc Toussaint mtoussai@cs.tu-berlin.de
TU Berlin
Ben Williams ben.williams@ed.ac.uk
School of Informatics,University of Edinburgh
Amos Storkey a.storkey@ed.ac.uk
University of Edinburgh

Biological movement is built up of sub-blocks or motion primitives. Such primitives provide a compact representation of movement which is also desirable in robotic control applications. We analyse handwriting data to gain a better understanding of use of primitives and their timings in biological movements. Inference of the shape and the timing of primitives can be done using a factorial HMM based model, allowing the handwriting to be represented in primitive timing space. This representation provides a distribution of spikes corresponding to the primitive activations, which can also be modelled using HMM architectures. We show how the coupling of the low level primitive model, and the higher level timing model during inference can produce good reconstructions of handwriting, with shared primitives for all characters modelled. This coupled model also captures the variance profile of the dataset which is accounted for by spike timing jitter. The timing code provides a compact representation of the movement while generating a movement without an explicit timing model produces a scribbling style of output.

41 *Object Recognition by Scene Alignment*

Bryan Russell brussell@csail.mit.edu
MIT CSAIL
Antonio Torralba torralba@csail.mit.edu
CSAIL MIT
Ce Liu celiu@csail.mit.edu
Massachusetts Institute of Technology
Rob Fergus fergus@csail.mit.edu
CSAIL, MIT
William Freeman billf@csail.mit.edu
MIT

Current object recognition systems can only recognize a limited number of object categories; scaling up to many categories is the next challenge in object recognition. We seek to build a system to recognize and localize many different object categories in complex scenes. We achieve this through a deceptively simple approach: by matching the input image, in an appropriate representation, to images in a large training set of labeled images (LabelMe). This provides us with a set of retrieval images, providing hypotheses for object identities and locations. We then transfer the labelings from the retrieval set. We demonstrate the effectiveness of this approach and study algorithm component contributions using held-out test sets from the LabelMe database.

Spotlight presentation, Wednesday, 3:20PM.

42 *Learning Visual Attributes*

Vittorio Ferrari ferrari@robots.ox.ac.uk
University of Oxford

Andrew Zisserman az@robots.ox.ac.uk
University of Oxford

We present a probabilistic generative model of visual attributes, together with an efficient learning algorithm. Attributes are visual qualities of objects, such as 'red', 'striped', or 'spotted'. The model sees attributes as patterns of image segments, repeatedly sharing some characteristic properties. These can be any combination of appearance, shape, or the layout of segments within the pattern. Moreover, attributes with general appearance are taken into account, such as the pattern of alternation of *any* two colors which is characteristic for stripes. To enable learning from unsegmented training images, the model is learnt discriminatively, by optimizing a likelihood ratio. As demonstrated in the experimental evaluation, our model can learn in a weakly supervised setting and encompasses a broad range of attributes. We show that attributes can be learnt starting from a text query to Google image search, and can then be used to recognize the attribute and determine its spatial extent in novel real-world images.

Spotlight presentation, Wednesday, 3:20PM.

43 *Combined discriminative and generative articulated pose and non-rigid shape estimation*

Leonid Sigal ls@cs.brown.edu
Brown University
Alexandru Balan alb@cs.brown.edu
Brown University
Michael Black black@cs.brown.edu
Brown University

Estimation of three-dimensional articulated human pose and motion from images is a central problem in computer vision. Much of the previous work has been limited by the use of crude generative models of humans represented as articulated collections of simple parts such as cylinders. Automatic initialization of such models has proved difficult and most approaches assume that the size and shape of the body parts are known a priori. In this paper we propose a method for automatically recovering a detailed parametric model of non-rigid body shape and pose from monocular imagery. Specifically, we represent the body using a parameterized triangulated mesh model that is learned from a database of human range scans. We demonstrate a discriminative method to directly recover the model parameters from monocular images using a mixture of regressors. This predicted pose and shape are used to initialize a generative model for more detailed pose and shape estimation. The resulting approach allows fully automatic pose and shape recovery from monocular and multi-camera imagery. Experimental results show that our method is capable of robustly recovering articulated pose, shape and biometric measurements (e.g. height, weight, etc.) in both calibrated and uncalibrated camera environments.

Spotlight presentation, Wednesday, 9:50AM.

44 *Learning the 2-D Topology of Images*

Nicolas Le Roux nicolas.le.roux@umontreal.ca
Universite de Montreal
Yoshua Bengio yoshua.bengio@umontreal.ca
Universite de Montreal
Pascal Lamblin lamblinp@iro.umontreal.ca
Universite de Montreal

MARC JOLIVEAU marc.joliveau@ecp.fr
MAS - Ecole Centrale Paris
BALAZS KEGL balazs.kegl@gmail.com
University of Paris-Sud / CNRS

We study the following question: is the two-dimensional structure of images a very strong prior or is it something that can be learned with a few examples of natural images? If someone gave us a learning task involving images for which the two-dimensional topology of pixels was not known, could we discover it automatically and exploit it? For example suppose that the pixels had been permuted in a fixed but unknown way, could we recover the relative two-dimensional location of pixels on images? The surprising result presented here is that not only the answer is yes but that about as few as a thousand images are enough to approximately recover the relative locations of about a thousand pixels. This is achieved using a manifold learning algorithm applied to pixels associated with a measure of distributional similarity between pixel intensities. We compare different topology-extraction approaches and show how having the two-dimensional topology can be exploited.

45 *Kernels on Attributed Pointsets with Applications*

MEHUL PARSANA mehul.parsana@gmail.com
Indian Institute of Science
SOURANGSHU BHATTACHARYA sourangshu@gmail.com
Indian Institute of Science
CHIRU BHATTACHARYA chiru@csa.iisc.ernet.in
Indian Institute of Science
K. R. RAMAKRISHNAN krr@ee.iisc.ernet.in
Indian Institute of Science

This paper introduces kernels on attributed pointsets, which are sets of vectors embedded in an euclidean space. The embedding gives the notion of neighborhood, which is used to define positive semidefinite kernels on pointsets. Two novel kernels on neighborhoods are proposed, one evaluating the attribute similarity and the other evaluating shape similarity. Shape similarity function is motivated from spectral graph matching techniques. The kernels are tested on three real life applications: face recognition, photo album tagging, and shot annotation in video sequences, with encouraging results.

46 *Near-Maximum Entropy Models for Binary Neural Representations of Natural Images*

MATTHIAS BETHGE mbethge@tuebingen.mpg.de
MPI Tuebingen
PHILIPP BERENS berens@tuebingen.mpg.de
MPI for Biological Cybernetics

Maximum entropy analysis of binary variables provides an elegant way for studying the role of pairwise correlations in neural populations. Unfortunately, these approaches suffer from their poor scalability to high dimensions. In sensory coding, however, high-dimensional data is ubiquitous. Here, we introduce a new approach using a near-maximum entropy model, that makes this type of analysis feasible for very high-dimensional data—the model parameters can be derived in closed form and sampling is easy. We demonstrate its usefulness by studying a simple neural representation model of natural images. For the first time, we are able to directly compare predictions from a pairwise maximum entropy model not only in small groups of neurons, but also in larger populations of more than thousand units. Our results indicate that in such larger networks interactions exist that are not predicted by pairwise correlations, despite the fact that pairwise correlations explain the lower-dimensional marginal statistics extremely well up to the limit of dimensionality where estimation of the full joint distribution is feasible.

Spotlight presentation, Wednesday, 5:20PM.

47 *On Sparsity and Overcompleteness in Image Models*

Pietro Berkes berkes@gatsby.ucl.ac.uk
Gatsby Computational Neuroscience Unit, UCL
Richard Turner turner@gatsby.ucl.ac.uk
Gatsby Computational Neuroscience Unit, UCL
Maneesh Sahani maneesh@gatsby.ucl.ac.uk
Gatsby Computational Neuroscience Unit, UCL

Computational models of visual cortex, and in particular those based on sparse coding, have enjoyed much recent attention. Despite this currency, the question of how sparse or how over-complete a sparse representation should be, has gone without principled answer. Here, we use Bayesian model-selection methods to address these questions for a sparse-coding model based on a Student-t prior. Having validated our methods on toy data, we find that natural images are indeed best modelled by extremely sparse distributions; although for the Student-t prior, the associated optimal basis size is only modestly overcomplete.

48 *The discriminant center-surround hypothesis for bottom-up saliency*

Vijay Mahadevan vmahadev@ucsd.edu
University of California, San Diego
Dashan Gao dgao@ucsd.edu
University of California, San Diego
Nuno Vasconcelos nuno@ece.ucsd.edu
University of California San Diego

The classical hypothesis, that bottom-up saliency is a center-surround process, is combined with a more recent hypothesis that all saliency decisions are optimal in a decision-theoretic sense. The combined hypothesis is denoted as discriminant center-surround saliency, and the corresponding optimal saliency architecture is derived. This architecture equates the saliency of each image location to the discriminant power of a set of features with respect to the classification problem that opposes stimuli at center and surround, at that location. It is shown that the resulting saliency detector makes accurate quantitative predictions for various aspects of the psychophysics of human saliency, including non-linear properties beyond the reach of previous saliency models. Furthermore, it is shown that discriminant center-surround saliency can be easily generalized to various stimulus modalities (such as color, orientation and motion), and provides optimal solutions for many other saliency problems of interest for computer vision. Optimal solutions, under this hypothesis, are derived for a number of the former (including static natural images, dense motion fields, and even dynamic textures), and applied to a number of the latter (the prediction of human eye fixations, motion-based saliency in the presence of ego-motion, and motion-based saliency in the presence of highly dynamic backgrounds). In result, discriminant saliency

is shown to predict eye fixations better than previous models, and produce background subtraction algorithms that outperform the state-of-the-art in computer vision.

49 *Predicting human gaze using low-level saliency combined with face detection*

Moran Cerf moranc@caltech.edu
California Institute of Technology
Jonathan Harel harel@klab.caltech.edu
California Institute of Technology
Wolfgang Einhaeuser wolfgang.einhaeuser@inf.ethz.ch
ETH Zurich,Institute of Computational Science
Christof Koch koch@klab.caltech.edu
California Institute of Technology

Under natural viewing conditions, human observers shift their gaze to allocate processing resources to subsets of the visual input. Many computational models have aimed at predicting such voluntary attentional shifts. Although the importance of high level stimulus properties (higher order statistics, semantics) stands undisputed, most models are based on low-level features of the input alone. In this study we recorded eye-movements of human observers while they viewed photographs of natural scenes. About two thirds of the stimuli contained at least one person. We demonstrate that a combined model of face detection and low-level saliency clearly outperforms a low-level model in predicting locations humans fixate. This is reflected in our finding fact that observes, even when not instructed to look for anything particular, fixate on a face with a probability of over 80% within their first two fixations (500ms). Remarkably, the model's predictive performance in images that do not contain faces is not impaired by spurious face detector responses, which is suggestive of a bottom-up mechanism for face detection. In summary, we provide a novel computational approach which combines high level object knowledge (in our case: face locations) with low-level features to successfully predict the allocation of attentional resources.

50 *GRIFT: A graphical model for inferring visual classification features from human data*

Michael Ross mgross@mit.edu
University of Massachusetts Amherst
Andrew Cohen acohen@psych.umass.edu
University of Massachusetts Amherst

This paper describes a new model for human visual classification that enables the recovery of image features that explain human subjects' performance on different visual classification tasks. Unlike previous methods, this algorithm does not model their performance with a single linear classifier operating on raw image pixels. Instead, it models classification as the combination of multiple feature detectors. This approach extracts more information about human visual classification than has been previously possible with other methods and provides a foundation for further exploration.

51 *Sparse deep belief net model for visual area V2*

HONGLAK LEE hllee@cs.stanford.edu
Stanford University
CHAITANYA EKANADHAM chaitu@cs.stanford.edu
Stanford University
ANDREW NG ang@cs.stanford.edu
Stanford University

Motivated in part by the hierarchical organization of cortex, a number of algorithms have recently been proposed that try to learn hierarchical, or "deep," structure from unlabeled data. While several authors have formally or informally compared their algorithms to computations performed in visual area V1 (and the cochlea), little attempt has been made thus far to evaluate these algorithms in terms of their fidelity for mimicking computations at deeper levels in the cortical hierarchy. This paper presents an unsupervised learning model that faithfully mimics certain properties of visual area V2. Specifically, we develop a sparse variant of the deep belief networks of Hinton et al. (2006). We learn two layers of nodes in the network, and demonstrate that the first layer, similar to prior work on sparse coding and ICA, results in localized, oriented, edge filters, similar to the Gabor functions known to model V1 cell receptive fields. Further, the second layer in our model encodes correlations of the first layer responses in the data. Specifically, it picks up both collinear ("contour") features as well as corners and junctions. More interestingly, in a quantitative comparison, the encoding of these more complex "corner" features matches well with the results from the Ito & Komatsu's study of biological V2 responses. This suggests that our sparse variant of deep belief networks holds promise for modeling more higher-order features.

52 *Estimating disparity with confidence from energy neurons*

ERIC KONG-CHAU TSANG eeeric@ee.ust.hk
Hong Kong University of Science & Technology
BERTRAM SHI eebert@ust.hk
Hong Kong Univ. of Science and Technology

The peak location in a population of phase-tuned neurons has been shown to be a more reliable estimator for disparity than the peak location in a population of position-tuned neurons. Unfortunately, the disparity range covered by a phase-tuned population is limited by phase wraparound. Thus, a single population cannot cover the large range of disparities encountered in natural scenes unless the scale of the receptive fields is chosen to be very large, which results in very low resolution depth estimates. Here we describe a biologically plausible measure of the confidence that the stimulus disparity is inside the range covered by a population of phase-tuned neurons. Based upon this confidence measure, we propose an algorithm for disparity estimate that uses many populations of high-resolution phase-tuned neurons that are biased to different disparity ranges via position shifts between the left and right eye receptive fields. The population with the highest confidence is used to estimate the stimulus disparity. We show that this algorithm outperforms a previously proposed coarse-to-fine algorithm for disparity estimation, which uses disparity estimates from coarse scales to select the populations used at higher scales.

53 *Learning Horizontal Connections in a Sparse Coding Model of Natural Images*

Pierre Garrigues garrigue@eecs.berkeley.edu
UC Berkeley
Bruno Olshausen baolshausen@berkeley.edu
UC Berkeley

It has been shown that adapting a dictionary of basis functions to the statistics of natural images so as to maximize sparsity in the coefficients results in a set of dictionary elements whose spatial properties resemble those of V1 (primary visual cortex) receptive fields. However, the resulting sparse coefficients still exhibit pronounced statistical dependencies, thus violating the independence assumption of the sparse coding model. Here, we propose a model that attempts to capture the dependencies among the basis function coefficients by including a pairwise coupling term in the prior over the coefficient activity states. When adapted to the statistics of natural images, the coupling terms learn a combination of facilitatory and inhibitory interactions among neighboring basis functions. These learned interactions may offer an explanation for the function of horizontal connections in V1, and we discuss the implications of our findings for physiological experiments.

54 *A Bayesian Model of Conditioned Perception*

Alan Stocker alan@cns.nyu.edu
New York University, Howard Hughes Medical Institute
Eero Simoncelli eero@cns.nyu.edu
New York University, Howard Hughes Medical Institute

We propose an extended probabilistic model for human perception. We argue that in many circumstances, human observers simultaneously evaluate sensory evidence under different hypotheses regarding the underlying physical process that might have generated the sensory information. Within this context, inference can be optimal if the observer weighs each hypothesis according to the correct belief in that hypothesis. But if the observer commits to a particular hypothesis, the belief in that hypothesis is converted into subjective certainty, and subsequent perceptual behavior is suboptimal, conditioned only on the chosen hypothesis. We demonstrate that this framework can explain psychophysical data of a recently reported decision-estimation experiment. The model well accounts for the data, predicting the same estimation bias as a consequence of the preceding decision step. The power of the framework is that it has no free parameters except the degree of the observer's uncertainty about its internal sensory representation. All other parameters are defined by the particular experiment which allows us to make quantitative predictions of human perception to two modifications of the original experiment.

55 *Receptive Fields without Spike-Triggering*

Jakob Macke jakob@tuebingen.mpg.de
Max Planck Institute Biological Cybernetics
Guenther Zeck zeck@neuro.mpg.de
Max Planck Institute of Neurobiology
Matthias Bethge mbethge@tuebingen.mpg.de
MPI Tuebingen

Stimulus selectivity of sensory neurons is often characterized by estimating their receptive field properties such as orientation selectivity. Receptive fields are usually derived from the mean (or covariance) of the spike-triggered stimulus ensemble. This approach treats each spike as an independent message but does not take into account that information might be conveyed through patterns of neural activity that are distributed across space or time. Can we find a concise description for the processing of a whole population of neurons analogous to the receptive field for single neurons? Here, we present a generalization of the linear receptive field which is not bound to be triggered on individual spikes but can be meaningfully linked to distributed response patterns. More precisely, we seek to identify those stimulus features and the corresponding patterns of neural activity that are most reliably coupled. We use an extension of reverse-correlation methods based on canonical correlation analysis. The resulting population receptive fields span the subspace of stimuli that is most informative about the population response. We evaluate our approach using both neuronal models and multi-electrode recordings from rabbit retinal ganglion cells. We show how the model can be extended to capture nonlinear stimulus-response relationships using kernel canonical correlation analysis, which makes it possible to test different coding mechanisms. Our technique can also be used to calculate receptive fields from multi-dimensional neural measurements such as those obtained from dynamic imaging methods.

56 *An online Hebbian learning rule that performs Independent Component Analysis*

Claudia Clopath claudia.clopath@epfl.ch
Laboratory of Computational Neuroscience, EPFL
Wulfram Gerstner wulfram.gerstner@epfl.ch
EPFL
Andre Longtin alongtin@uottawa.ca
Physics Department, University of Ottawa

Independent component analysis (ICA) is a powerful method to decouple signals. Most of the algorithms performing ICA do not consider the temporal correlations of the signal, but only higher moments of its amplitude distribution. Moreover, they require some preprocessing of the data (whitening) so as to remove second order correlations. In this paper, we are interested in understanding the neural mechanism responsible for solving ICA. We present an online learning rule that exploits delayed correlations in the input. This rule performs ICA by detecting joint variations in the firing rates of pre- and postsynaptic neurons, similar to a local rate-based Hebbian learning rule.

57 *Bayesian Inference for Spiking Neuron Models with a Sparsity Prior*

Sebastian Gerwinn, Max Planck Institute f Biological Cybernetics, Jakob Macke, Max Planck Institute Biological Cybernetics, Matthias Seeger, Max Planck Institute for, Biological Cybernetics and Matthias Bethge, MPI Tuebingen.
Oral presentation, Thursday, 9:10AM. See abstract, page 168.

58 *A neural network implementing optimal state estimation based on dynamic spike train decoding*

Omer Bobrowski, Technion - Israel Institute of Technology, Ron Meir, Technion, Shy Shoham, Technion - Israel Institute of Technology and Yonina Eldar, Technion.
Oral presentation, Thursday, 8:50AM. See abstract, page 167.

59 *Hippocampal Contributions to Control: The Third Way*

Mate Lengyel, Collegium Budapest, Institute for Advanced Study
and Peter Dayan, Gatsby Computational Neuroscience Unit.
Oral presentation, Thursday, 10:40AM. See abstract, page 170.

60 *Measuring Neural Synchrony by Message Passing*

Tomasz Rutkowski tomek@brain.riken.jp
Brain Science Institute RIKEN, Advanced Brain Signal Processing Lab
Andrzej Cichocki cia@brain.riken.jp
Riken BSI
François Vialatte fvialatte@brain.riken.jp
RIKEN BSI, L.ABSP
Justin Dauwels justin@dauwels.com
RIKEN Brain Science Institute

A novel approach to measure the interdependence of two time series is proposed, referred to as "stochastic event synchrony" (SES); it quantifies the alignment of two point processes by means of the following parameters: time delay, standard deviation of the timing jitter, the fraction of "spurious" events, and the average similarity of the events. In contrast to the other measures, SES quantifies the synchrony of oscillatory events (instead of more conventional amplitude or phase synchrony). Pairwise alignment of the point processes is cast as a statistical inference problem, which is solved by applying the max-product algorithm on a graphical model. The SES parameters are determined from the resulting pairwise alignment by maximum a posteriori (MAP) estimation. The proposed interdependence measure is applied to the problem of detecting anomalies in EEG synchrony of Mild Cognitive Impairment (MCI) patients.

Spotlight presentation, Wednesday, 5:20PM.

61 *Neural characterization in partially observed populations of spiking neurons*

Jonathan Pillow, Gatsby Computational Neuroscience Unit, UCL
and Peter Latham, Gatsby Computational Neuroscience Unit.
Oral presentation, Thursday, 9:30AM. See abstract, page 168.

62 *Simplified Rules and Theoretical Analysis for Information Bottleneck Optimization and PCA with Spiking Neurons*

Lars Buesing lars@igi.tugraz.at
TU Graz
Wolfgang Maass wolfgang.maass@igi.tugraz.at
TU Graz

We show that under suitable assumptions (primarily linearization) a simple and perspicuous online learning rule for Information Bottleneck optimization with spiking neurons can be derived. This rule performs on common benchmark tasks as well as a rather complex rule that has previously been proposed. Furthermore, the transparency of this new learning rule makes a theoretical analysis of its convergence properties feasible. A variation of this learning rule (with sign changes) provides a theoretically founded method for performing Principal Component Analysis (PCA) with spiking neurons. By applying this rule to

an ensemble of neurons, different principal components of the input can be extracted. In addition, it is possible to preferentially extract those principal components from incoming signals X that are related or are not related to some additional target signal Y_T. In a biological interpretation, this target signal Y_T (also called relevance variable) could represent proprioceptive feedback, input from other sensory modalities, or top-down signals.

63 *Extending position/phase-shift tuning to motion energy neurons improves velocity discrimination*

Yiu Man Lam eelym@ust.hk
Hong Kong University of Science & Technology
Bertram Shi eebert@ee.ust.hk
Hong Kong University of Science & Technology

We extend position and phase-shift tuning, concepts already well established in the disparity energy neuron literature, to motion energy neurons. We show that Reichardt-like detectors can be considered examples of position tuning, and that motion energy filters whose complex valued spatio-temporal receptive fields are space-time separable can be considered examples of phase tuning. By combining these two types of detectors, we obtain an architecture for constructing motion energy neurons whose center frequencies can be adjusted by both phase and position shifts. Similar to recently described neurons in the primary visual cortex, these new motion energy neurons exhibit tuning that is between purely space-time separable and purely speed tuned. We propose a functional role for this intermediate level of tuning by demonstrating that comparisons between pairs of these motion energy neurons can reliably discriminate between inputs whose velocities lie above or below a given reference velocity

Spotlight presentation, Wednesday, 5:20PM.

64 *Theoretical Analysis of Learning with Reward-Modulated Spike-Timing-Dependent Plasticity*

Robert Legenstein legi@igi.tugraz.at
TU Graz
Dejan Pecevski dejan@igi.tugraz.at
Graz University of Technology
Wolfgang Maass maass@igi.tugraz.at
TU Graz

Reward-modulated spike-timing-dependent plasticity (STDP) has recently emerged as a candidate for a learning rule that could explain how local learning rules at single synapses support behaviorally relevant adaptive changes in complex networks of spiking neurons. However the potential and limitations of this learning rule could so far only be tested through computer simulations. This article provides tools for an analytic treatment of reward-modulated STDP, which allow us to predict under which conditions reward-modulated STDP will be able to achieve a desired learning effect. In particular, we can produce in this way a theoretical explanation and a computer model for a fundamental experimental finding on biofeedback in monkeys.

Spotlight presentation, Wednesday, 3:20PM.

65 *Inferring Elapsed Time from Stochastic Neural Processes*

Maneesh Sahani, Gatsby Computational Neuroscience Unit, UCL
and Misha Ahrens, Gatsby Computational Neuroscience Unit, UCL.
Oral presentation, Thursday, 8:30AM. *See abstract, page 167.*

66 *Inferring Neural Firing Rates from Spike Trains Using Gaussian Processes*

John Cunningham jcunnin@stanford.edu
Stanford University, Department of Electrical Engineering
Byron Yu byronyu@stanford.edu
Stanford University
Maneesh Sahani maneesh@gatsby.ucl.ac.uk
Gatsby Computational Neuroscience Unit, UCL
Krishna Shenoy shenoy@stanford.edu
Stanford

Neural signals present challenges to analytical efforts due to their noisy, spiking nature. Many studies of neuroscientific and neural prosthetic importance rely on a smoothed, denoised neural signal considered to be the spike train's underlying firing rate. Current techniques to find time varying firing rates require ad hoc choices of parameters, offer no confidence intervals on their estimates, and can obscure potentially important single trial variability. We present a new method, based on a Gaussian Process prior, for inferring probabilistically optimal estimates of firing rate functions underlying single or multiple neural spike trains. We simulate spike trains to test the performance of the method and demonstrate significant average error improvement over standard smoothing techniques.

Spotlight presentation, Wednesday, 5:20PM.

67 *Invariant Common Spatial Patterns: Alleviating Nonstationarities in Brain-Computer Interfacing*

Benjamin Blankertz benjamin.blankertz@first.fhg.de
Technical University of Berlin, Fraunhofer FIRST (IDA)
Motoaki Kawanabe nabe@first.fraunhofer.de
Fraunhofer FIRST.IDA
Ryota Tomioka ryota.tomioka@first.fraunhofer.de
Fraunhofer FIRST.IDA
Friederike Hohlefeld f.u.hohlefeld@web.de
Neurophysics Group, Charite Berlin
Vadim Nikulin vadim.nikulin@charite.de
Neurophysics Group, Charite, Berlin, Bernstein Center, Berlin
Klaus-Robert Müller klaus@first.fraunhofer.de
Fraunhofer FIRST.IDA

Brain-Computer Interfaces can suffer from a large variance of the subject conditions within and across sessions. For example vigilance fluctuations in the individual, variable task involvement, workload etc. alter the characteristics of EEG signals and thus challenge a stable BCI operation. In the present work we aim to define features based on a variant of the common spatial patterns (CSP) algorithm that are constructed invariant with respect to such nonstationarities. We enforce invariance properties by adding terms to the denominator of a Raleigh coefficient representation of CSP such as disturbance covariance

matrices from fluctuations in visual processing. In this manner physiological prior knowledge can be used to shape the classification engine for BCI. As a proof of concept we present a BCI classifier that is robust to changes in the level of parietal alpha-activity. In other words, the EEG decoding still works when there are lapses in vigilance.

Spotlight presentation, Wednesday, 5:20PM.

68 *EEG-Based Brain-Computer Interaction: Improved Accuracy by Automatic Single-Trial Error Detection*

Pierre Ferrez pierre.ferrez@idiap.ch
IDIAP Research Institute
Jose del R. Millan jose.millan@idiap.ch
IDIAP Research Institute

Brain-computer interfaces (BCIs), as any other interaction modality based on physiological signals and body channels (e.g., muscular activity, speech and gestures), are prone to errors in the recognition of subject's intent. An elegant approach to improve the accuracy of BCIs consists in a verification procedure directly based on the presence of error-related potentials (ErrP) in the EEG recorded right after the occurrence of an error. Six healthy volunteer subjects with no prior BCI experience participated in a new human-robot interaction experiment where they were asked to mentally move a cursor towards a target that can be reached within a few steps using motor imagination. This experiment confirms the previously reported presence of a new kind of ErrP. These "Interaction ErrP" exhibit a first sharp negative peak followed by a positive peak and a second broader negative peak (~290, ~350 and ~470 ms after the feedback, respectively). But in order to exploit these ErrP we need to detect them in each single trial using a short window following the feedback associated to the response of the classifier embedded in the BCI. We have achieved an average recognition rate of correct and erroneous single trials of 81.8% and 76.2%, respectively. Furthermore, we have achieved an average recognition rate of the subject's intent while trying to mentally drive the cursor of 73.1%. These results show that it's possible to simultaneously extract useful information for mental control to operate a brain-actuated device as well as cognitive states such as error potentials to improve the quality of the brain-computer interaction. Finally, using a well-known inverse model (sLORETA), we show that the main focus of activity at the occurrence of the ErrP are, as expected, in the pre-supplementary motor area and in the anterior cingulate cortex.

69 *Second Order Bilinear Discriminant Analysis for single trial EEG analysis*

Christoforos Christoforou cchristoforos@yahoo.com
The Graduate Center of the City University, The City College of the City University
Paul Sajda ps629@columbia.edu
Columbia University
Lucas C. Parra parra@ccny.cuny.edu
City College of New York

Traditional analysis methods for single-trial classification of electro-encephalography (EEG) focus on two types of paradigms: phase locked methods, in which the amplitude of the signal is used as the feature for classification, i.e. event related potentials; and second order methods, in which the feature of interest is the power of the signal, i.e event related (de)synchronization. The process of deciding which paradigm to use is ad hoc and is driven by knowledge of neurological findings. Here we propose a unified method in which the algorithm learns the best first and second order spatial and temporal features for classification of EEG based on a bilinear model. The efficiency of the method is demonstrated in simulated and real EEG from a benchmark data set for Brain Computer Interface.

70 *Predicting Brain States from fMRI Data: Incremental Functional Principal Component Regression*

Sennay Ghebreab ghebreab@science.uva.nl
University of Amsterdam
Arnold Smeulders smeulders@science.uva.nl
University of Amsterdam
Pieter Adriaans pietera@science.uva.nl
IvI, Universiteit van Amsterdam

We propose a method for reconstruction of human brain states directly from functional neuroimaging data. The method extends the traditional multivariate regression analysis of discretized fMRI data to the domain of stochastic functional measurements, facilitating evaluation of brain responses to naturalistic stimuli and boosting the power of functional imaging. The method searches for sets of voxel timecourses that optimize a multivariate functional linear model in terms of Rsquare-statistic. Population based incremental learning is used to search for spatially distributed voxel clusters, taking into account the variation in Haemodynamic lag across brain areas and among subjects by voxel-wise nonlinear registration of stimuli to fMRI data. The method captures spatially distributed brain responses to naturalistic stimuli without attempting to localize function. Application of the method for prediction of naturalistic stimuli from new and unknown fMRI data shows that the approach is capable of identifying distributed clusters of brain locations that are highly predictive of a specific stimuli.

71 *Locality and low-dimensions in the prediction of natural experience from fMRI*

Francois Meyer francois.meyer@colorado.edu
University of Colorado at Boulder
Greg Stephens gstephen@princeton.edu
Princeton University, Center For Brain, Mind and Behavior

Functional Magnetic Resonance Imaging (fMRI) provides an unprecedented window into the complex functioning of the human brain, typically detailing the activity of thousands of voxels during hundreds of sequential time points. Unfortunately, the interpretation of fMRI is complicated due both to the relatively unknown connection between the hemodynamic response and neural activity and the unknown spatiotemporal characteristics of the cognitive patterns themselves. Here, we use data from the Experience Based Cognition competition to compare global and local methods of prediction applying both linear and nonlinear techniques of dimensionality reduction. We build global low dimensional representations of an fMRI dataset, using linear and nonlinear methods. We learn a set of time series that are implicit functions of the fMRI data, and predict the values of these times series in the future from the knowledge of the fMRI data only. We find effective, low-dimensional models based on the principal components of cognitive activity in classically-defined anatomical regions, the Brodmann Areas. Furthermore for some of the stimuli, the top predictive regions were stable across subjects and episodes, including Wernicke's area for verbal instructions, visual cortex for facial and body features, and

visual-temporal regions (Brodmann Area 7) for velocity. These interpretations and the relative simplicity of our approach provide a transparent and conceptual basis upon which to build more sophisticated techniques for fMRI decoding. To our knowledge, this is the first time that classical areas have been used in fMRI for an effective prediction of complex natural experience.

72 *Continuous Time Particle Filtering for fMRI*

Lawrence Murray lawrence.murray@ed.ac.uk
School of Informatics, University of Edinburgh
Amos Storkey a.storkey@ed.ac.uk
University of Edinburgh

We construct a biologically motivated stochastic differential model of the neural and hemodynamic activity underlying the observed Blood Oxygen Level Dependent (BOLD) signal in Functional Magnetic Resonance Imaging (fMRI). The model poses a difficult parameter estimation problem, both theoretically due to the nonlinearity and divergence of the differential system, and computationally due to its time and space complexity. We adapt a particle filter and smoother to the task, and discuss some of the practical approaches used to tackle the difficulties, including use of sparse matrices and parallelisation. Results demonstrate the tractability of the approach in its application to an effective connectivity study.

73 *An in-silico Neural Model of Dynamic Routing through Neuronal Coherence*

Devarajan Sridharan dsridhar@stanford.edu
Neurosciences Program, Stanford University
Brian Percival bperci@stanford.edu
Stanford University
John Arthur jarthur@stanford.edu
Stanford University
Kwabena Boahen boahen@stanford.edu
Stanford

We describe a neurobiologically plausible model to implement dynamic routing using the concept of neuronal communication through neuronal coherence. The model has a three-tier architecture: a raw input tier, a routing control tier, and an invariant output tier. The correct mapping between input and output tiers is realized by an appropriate alignment of the phases of their respective background oscillations by the routing control units. We present an example architecture, implemented on a neuromorphic chip, that is able to achieve circular-shift invariance. A simple extension to our model can accomplish circular-shift dynamic routing with only $O(N)$ connections, compared to $O(N^2)$ connections required by traditional models.

Spotlight presentation, Wednesday, 5:20PM.

74 *Learning to classify complex patterns using a VLSI network of spiking neurons*

Srinjoy Mitra, Institute of Neuroinformatics, UNIZ - ETHZ,
Giacomo Indiveri, UZH-ETH Zurich, Institute of Neuroinformatics
and Stefano Fusi, Institute of Neuroinformatics, UNIZ - ETHZ.
Oral presentation, Thursday, 9:50AM. See abstract, page 169.

75 *A configurable analog VLSI neural network with spiking neurons and self-regulating plastic synapses*

Massimiliano Giulioni giulioni@roma2.infn.it
Italian National Institute Of Health
Mario Pannunzi mario.pannunzi@iss.it
National Institutes of Health, Universita Roma 3
Davide Badoni davide.badoni@roma2.infn.it
I.N.F.N. Sezione Roma Tor Vergata
Vittorio Dante dante@iss.infn.it
Istituto Superiore di Sanita Rome Italy
Paolo del Giudice paolo.delgiudice@iss.infn.it
Italian National Institute of Health, INFN

We summarize the implementation of an analog VLSI chip hosting a network of 32 integrate-and-fire (IF) neurons with spike-frequency adaptation and 2,048 Hebbian plastic bistable spike-driven stochastic synapses endowed with a self-regulating mechanism which stops unnecessary synaptic changes. The synaptic matrix can be flexibly configured and provides both recurrent and AER-based connectivity with external, AER compliant devices. We demonstrate the ability of the network to efficiently classify overlapping patterns, thanks to the self-regulating mechanism.

76 *Contraction Properties of VLSI Cooperative Competitive Neural Networks of Spiking Neurons*

Emre Neftci emre@ini.phys.ethz.ch
Institute of Neuroinformatics, UNIZ - ETHZ
Elisabetta Chicca chicca@ini.phys.ethz.ch
Institute of Neuroinformatics, UNIZ - ETHZ
Giacomo Indiveri giacomo@ini.phys.ethz.ch
UZH-ETH Zurich, Institute of Neuroinformatics
Jean-Jeacques Slotine jjs@mit.edu
MIT
Rodney Douglas rjd@ini.phys.ethz.ch
Institute of Neuroinformatics, UNIZ - ETHZ

A non–linear dynamic system is called contracting if initial conditions are forgotten exponentially fast, so that all trajectories converge to a single trajectory. We use contraction theory to derive an upper bound for the strength of recurrent connections that guarantees contraction for complex neural networks. Specifically, we apply this theory to a special class of recurrent networks, often called Cooperative Competitive Networks (CCNs), which are an abstract representation of the cooperative-competitive connectivity observed in cortex. This specific type of network is believed to play a major role in shaping cortical responses and selecting the relevant signal among distractors and noise. In this paper, we analyze contraction of combined CCNs of linear threshold units and verify the results of

our analysis in a hybrid analog/digital VLSI CCN comprising spiking neurons and dynamic synapses.

Spotlight presentation, Wednesday, 5:20PM.

77 *Subspace-Based Face Recognition in Analog VLSI*

Miguel Figueroa mfigueroa@gmail.com
Universidad de Concepcion
Gonzalo Carvajal gcarvaja@udec.cl
Universidad de Concepcion
Waldo Valenzuela waldovalenzuela@udec.cl
Universidad de Concepcion

We describe an analog-VLSI neural network for face recognition based on subspace methods. The system uses a dimensionality-reduction network whose coefficients can be either programmed or learned on-chip to perform PCA, or programmed to perform LDA. A second network with user-programmed coefficients performs classification with Manhattan distances. The system uses on-chip compensation techniques to reduce the effects of device mismatch. Using the ORL database with 12x12-pixel images, our circuit achieves up to 85% classification performance (98% of an equivalent software implementation).

Spotlight presentation, Wednesday, 3:20PM.

Thursday, December 6th

Oral Session — Neuroscience I (Chair: Alan Stocker, New York University):

8:30AM *Inferring Elapsed Time from Stochastic Neural Processes*

Maneesh Sahani maneesh@gatsby.ucl.ac.uk
Gatsby Computational Neuroscience Unit, UCL
Misha Ahrens ahrens@gatsby.ucl.ac.uk
Gatsby Computational Neuroscience Unit, UCL

Many perceptual processes and neural computations, such as speech recognition, motor control and learning, depend on the ability to measure and mark the passage of time. However, the neural mechanisms that make such temporal judgements possible are unknown. A number of different hypotheses have been advanced, all of which depend on the known evolution of a neural or psychological state, possibly through oscillations or the gradual decay of a memory trace. We suggest a new model, which instead exploits the fact that neural and sensory processes, even when their precise evolution is unpredictable, exhibit statistically structured changes. We show that this structure can be exploited for timing, and that reliable timing estimators can be derived from the statistics of the processes. This framework of decoding time from stochastic processes allows for a much wider array of neural implementations of time estimation than has been considered so far, and can simultaneously emulate several different behavioral findings, which so far have only been understood in psychological terms.

8:50AM *A neural network implementing optimal state estimation based on dynamic spike train decoding*

Omer Bobrowski bober@tx.technion.ac.il
Technion - Israel Institute of Technology
Ron Meir rmeir@ee.technion.ac.il
Technion
Shy Shoham sshoham@bm.technion.ac.il
Technion - Israel Institute of Technology
Yonina Eldar yonina@ee.technion.ac.il
Technion

It is becoming increasingly evident that organisms acting in uncertain dynamical environments often employ exact or approximate Bayesian statistical calculations in order to continuously estimate the environmental state, integrate information from multiple sensory modalities, form predictions and choose actions. What is less clear is how these putative computations are implemented by cortical neural networks. An additional level of complexity is introduced because these networks observe the world through spike trains received from primary sensory afferents, rather than directly. A recent line of research has described mechanisms by which such computations can be implemented using a network of neurons whose activity directly represents a probability distribution across the possible "world states". Much of this work, however, uses various approximations, which severely restrict the domain of applicability of these implementations. Here we make use of rigorous

mathematical results from the theory of continuous time point process filtering, and show how optimal real-time state estimation and prediction may be implemented in a general setting using linear neural networks. We demonstrate the applicability of the approach with several examples, and relate the required network properties to the statistical nature of the environment, thereby quantifying the compatibility of a given network with its environment.

9:10AM *Bayesian Inference for Spiking Neuron Models with a Sparsity Prior*

Sebastian Gerwinn sgerwinn@tuebingen.mpg.de
Max Planck Institute f Biological Cybernetics
Jakob Macke jakob@tuebingen.mpg.de
Max Planck Institute Biological Cybernetics
Matthias Seeger seeger@tuebingen.mpg.de
Max Planck Institute for, Biological Cybernetics
Matthias Bethge mbethge@tuebingen.mpg.de
MPI Tuebingen

Generalized linear models are the most commonly used tools to describe the stimulus selectivity of sensory neurons. Here we present a Bayesian treatment of such models. Using the expectation propagation algorithm, we are able to approximate the full posterior distribution over all weights. In addition, we use a Laplacian prior to favor sparse solutions. Therefore, stimulus features that do not critically influence neural activity will be assigned zero weights and thus be effectively excluded by the model. This feature selection mechanism facilitates both the interpretation of the neuron model as well as its predictive abilities. The posterior distribution can be used to obtain confidence intervals which makes it possible to assess the statistical significance of the solution. In neural data analysis, the available amount of experimental measurements is often limited whereas the parameter space is large. In such a situation, both regularization by a sparsity prior and uncertainty estimates for the model parameters are essential. We apply our method to multi-electrode recordings of retinal ganglion cells and use our uncertainty estimate to test the statistical significance of functional couplings between neurons. Furthermore we used the sparsity of the Laplace prior to select those filters from a spike-triggered covariance analysis that are most informative about the neural response.

9:30AM *Neural characterization in partially observed populations of spiking neurons*

Jonathan Pillow pillow@gatsby.ucl.ac.uk
Gatsby Computational Neuroscience Unit, UCL
Peter Latham pel@gatsby.ucl.ac.uk
Gatsby Computational Neuroscience Unit

Point process encoding models provide powerful statistical methods for understanding the responses of neurons to sensory stimuli. Although these models have been successfully applied to responses of neurons in the early sensory pathway, they have fared less well as a models of responses in deeper brain areas, as they do not easily take into account multiple stages of processing. Here we introduce a new twist on this approach: we include unobserved as well as observed spike trains. This provides us with a more powerful model, and thus more flexibility in fitting data. More importantly, it allows us to estimate connectivity patterns among neurons (both observed and unobserved), and so should give

insight into how networks process sensory input. We demonstrate the model on a simple toy network consisting of two neurons. The formalism, based on variational EM, can be easily extended to larger networks.

9:50AM *Learning to classify complex patterns using a VLSI network of spiking neurons*

Srinjoy Mitra srinjoy@ini.ch
Institute of Neuroinformatics, UNIZ - ETHZ
Giacomo Indiveri giacomo@ini.phys.ethz.ch
UZH-ETH Zurich, Institute of Neuroinformatics
Stefano Fusi fusi@ini.phys.ethz.ch
Institute of Neuroinformatics, UNIZ - ETHZ

Real time classification of complex patterns of trains of spikes is a difficult and important computational problem. Here we propose a compact, low power, fully analog neuromorphic device which can learn to classify complex patterns of mean firing rates. The chip implements a network of integrate-and-fire neurons connected by bistable plastic synapses. Learning is supervised by a teacher which simply provides an extra input to the output neurons during training. The synapses are modified only as long as the current generated by the plastic synapses does not match the output desired by the teacher (as in the perceptron learning rule). Our device has been designed to be able to learn linearly separable patterns and we show in a series of tests that it can classify uncorrelated random spatial patterns of mean firing rates.

10:10AM Break

Oral Session — Neuroscience II (Chair: Odelia Schwartz, Albert Einstein College of Medicine):

10:40AM *Hippocampal Contributions to Control: The Third Way*

Mate Lengyel lmate@gatsby.ucl.ac.uk
Collegium Budapest, Institute for Advanced Study
Peter Dayan dayan@gatsby.ucl.ac.uk
Gatsby Computational Neuroscience Unit

Recent experimental studies have focused on the specialization of different neural structures for different types of instrumental behavior. Recent theoretical work has provided normative accounts for why there should be more than one control system, and how the output of different controllers can be integrated. Two particlar controllers have been identified, one associated with a forward model and the prefrontal cortex and a second associated with computationally simpler, habitual, actor-critic methods and part of the striatum. We argue here for the normative appropriateness of an additional, but so far marginalized control system, associated with episodic memory, and involving the hippocampus and medial temporal cortices. We analyze in depth a class of simple environments to show that episodic control should be useful in a range of cases characterized by complexity and inferential noise, and most particularly at the very early stages of learning, long before habitization has set in. We interpret data on the transfer of control from the hippocampus to the striatum in the light of this hypothesis.

11:00AM Invited Talk: *Population coding of object images based on visual features and its relevance to view invariant representation*

Manabu Tanifuji insinfor@brain.riken.jp
RIKEN Brain Science Institute

The monkey inferotemporal cortex (IT) is the association cortex implicated in object perception and recognition, but little is known how neurons in this area represent object images. Intrinsic signal imaging revealed that visual images of objects activated columns in a distributed manner in IT. When a part of these object were removed (stimulus simplification), the simplified stimuli activated only a subset of columns elicited by the originals. This result and subsequent extracellular recordings of neuronal activity from these columns suggested that objects were represented by the combination of columns, each of which represents a visual feature of objects (the feature-based representation). These features can be local features such as protrusions, curvature, and rectangular shapes appeared in part of object images. To find activity related to spatial relationship among the local features, we segmented an object image into parts where local features are accommodated and investigated spatial relationship among these parts instead of that among particular local features. Intrinsic signal imaging revealed the spots that were activated by the object but not by either of the individual parts. Extracellular recording of neuronal activity from these spots revealed that neurons were not sensitive to shapes of the individual parts, but to spatial arrangements of two parts. This result indicates that some columns represent local features but other represent global arrangements of such local features. Unique representation of an object may be achieved by combining both types of feature columns together. We have shown that representation of object images in monkey IT cortex is based on population coding with visual features of object images. One advantage of this feature-based representation is that a variety of possible combinations of visual features

made it possible to represent many objects differently in IT cortex. If majority of visual features representing an object do not largely change in a certain range of different viewing angles, population coding would also help categorization of images of the object viewed from different angles. We examined this possibility by recording population activity elicited by various object images with a dense multiple-electrode array, and found that not single site activity but the population activity indeed helps to represent object images view invariantly for a certain range.

12:00PM End of Conference

Reviewers

Pieter Abbeel
Jacob Abernethy
Vitaly Ablavsky
Kannan Achan
Arash Afkanpour
Felix Agakov
Amrudin Agovic
Yuan (Alan) Qi
Yasemin Altun
Charles Anderson
Galen Andrew
Cedric Archambeau
Panagiotis Artemiadis
Chris Atkeson
Hagai Attias
Jean-Julien Aucouturier
Jean-Yves Audibert
Peter Auer
Shai Avidan
Arik Azran
Boris Babenko
Drew Bagnell
Timothy Bailey
Chris Baker
Goekhan Bakir
Pierre Baldi
Arindam Banerjee
Amir Banihashemi
David Barber
Aharon Bar Hillel
Kobus Barnard
Andrew Barto
Sugato Basu
Sumit Basu
Sue Becker
Mikhail Belkin
Alexandre Belloni
Pashmina Bendale
Shai Ben-David
Samy Bengio
Yoshua Bengio
Asa Ben Hur
Matthias Bethge
Alina Beygelzimer
Jinbo Bi
Steffen Bickel
Steffen Bickel
Daniel Bikel
Mikhail Bilenko
Jeff Bilmes
Gilles Blanchard
Benjamin Blankertz
Matthew Blaschko
David Blei
John Blitzer
Leon Bottou
Olivier Bousquet
Michael Bowling
Xavier Boyen
David Brainard
Christoph Braun
Mikio Braun
Ulf Brefeld
Scott Brown
Michael Brueckner
Peter Brunner
Joachim Buhmann
Christopher Burges
Keith Bush
Stephane Canu
Nannan Cao
Andrea Caponnetto
Peter Carbonetto
Jose Carmena
Miguel Carreira-Perpinan
Gilles Casso
Gert Cauwenberghs
Gavin Cawley
Nicolo Cesa-Bianchi
Hakan Cevikalp
Olivier Chapelle
Sutton Charles
Erick Chastain
Gal Chechik
Kumar Chellapilla
Li Cheng
Max Chickering
Wei Chu
Michael Collins
Dorin Comaniciu
Adrian Corduneanu
Corinna Cortes
Eric Cosatto
Gary Cottrell
James Coughlan
Aaron Courville
Koby Crammer
Nello Cristianni

Aron Culotta
Marco Cuturi
Florence D'Alche-Buc
Marton Danoczy
David Danks
Denver Dash
Alexandre d'Aspremont
Hal Daume III
Jason Davis
Nathaniel Daw
Peter Dayan
Tijl De Bie
Fabio de Bona
Dennis DeCoste
Nando de Freitas
Marc Deisenroth
Ofer Dekel
Olivier Delalleau
Tobi Delbruck
Jose del R. Millan
Sophie Deneve
Li Deng
Francois Denis
Simon Dennis
Rachid Deriche
Inderjit Dhillon
Uwe Dick
Thomas Dietterich
Laura Dietz
Arnaud Doucet
Kenji Doya
Gerard Dreyfus
Greg Druck
Aaron D'Souza
Carl Edward Rasmussen
Alexei Efros
Jason Eisner
Gal Elidan
Charles Elkan
Daniel Ellis
Yaakov Engel
Barbara Engelhardt
Eleazar Eskin
Ralph Etienne-Cummings
Mark Everingham
Jason Farquhar
Siamac Fazli
Jacob Feldman
Rob Fergus
Vittorio Ferrari
Luciana Ferrer
Brian Ferris
Brian Ferris
Thomas Finley
Jozsef Fiser
Francois Fleuret
David Forsyth
Dieter Fox
Matthias Franz
William Freeman
Brendan Frey
Nir Friedman
Karl Friston
Qiang Fu
Kenji Fukumizu
Kenji Fukumizu
Glenn Fung
Thomas Gaertner
Gregory Gage
Surya Ganguli
Surya Ganguli
Artur Garcez
Christophe Garcia
Pierre Garrigues
Michael Gastpar
Peter Gehler
Claudio Gentile
Alborz Geramifard
Wulfram Gerstner
Sebastian Gerwinn
Zoubin Ghahramani
Mohammad Ghavamzadeh
Ali Ghodsi
Ran Gilad-Bachrach
Daniel Gildea
Mark Girolami
Jim Glass
Amir Globerson
Karen Glocer
Moises Goldszmidt
Sharon Goldwater
David Gondek
Noah Goodman
Geoffrey Gordon
Dilan Gorur
Thore Graepel
Yves Grandvalet
David Grangier
Russell Greiner
Tayfun Gurel
Arthur Gretton
Thomas Griffiths

Greg Grudic
Peter Grunwald
Carlos Guestrin
Yuhong Guo
Bernard Haasdonk
Patrick Haffner
Philipp Hafliger
Peter Haider
Yaroslav Halchenko
Greg Hamerly
Buhm Han
Zaid Harchaoui
David Hardoon
Stefan Harmeling
John Harris
Alexander Hartemink
Mark Hasegawa-Johnson
Andrea Hasenstaub
Nicho Hatsopoulos
John-Dylan Haynes
Martial Hebert
David Heckerman
Matthias Hein
Moritz Helias
Katherine Heller
Ralf Herbrich
Mark Herbster
Perfecto Herrera
Tomer Hertz
Aaron Hertzmann
Tom Heskes
Jeremy Hill
Thilo Hinterberger
Geoffrey Hinton
Michael Hirsch
Leigh Hochberg
Jesse Hoey
Matthew Hoffman
Thomas Hofmann
Derek Hoiem
Antti Honkela
Timothy Horiuchi
Eric Horvitz
Patrik Hoyer
Anne Hsu
Jianying Hu
David Huber
Martial Hue
Don Hush
Dirk Husmeier
Quentin Huys
Aapo Hyvarinen
Tsuyoshi Ide
Christian Igel
Christian Igel
Alexander Ihler
Shiro Ikeda
Giacomo Indiveri
Nathan Intrator
Nathan Intrator
Michael Isard
Shin Ishii
Tommi Jaakkola
Jeff Jackson
Robert Jacobs
Herbert Jaeger
Prateek Jain
Larsen Jan
Tony Jebara
Hilbert J Kappen
Thorsten Joachims
Mark Johnson
Matt Jones
Michael Jordan
Matti Kaariainen
Sham Kakade
Adam Kalai
Anitha Kannan
Ashish Kapoor
Rob Kass
Tsuyoshi Kato
Motoaki Kawanabe
Balazs Kegl
Sathiya Keerthi
Charles Kemp
Joseph Keshet
Dongmin Kim
Sung-Phil Kim
Scott Kirkpatrick
Sergey Kirshner
Jyrki Kivinen
Anssi Klapuri
Robert Kleinberg
James Knight
David Knill
Levente Kocsis
Daphne Koller
Vladimir Koltchinskii
Terry Koo
Konrad Kording
Balaji Krishnapuram
Dean Krusienski

Alex Kulesza
Brian Kulis
Dima Kuzmin
James Kwok
Simon Lacoste-Julien
Michail Lagoudakis
Christoph Lampert
John Langford
Patricio La Rosa
Pavel Laskov
Francois Laviolette
Neil Lawrence
Svetlana Lazebnik
John Lazzaro
Erik Learned-Miller
Guy Lebanon
Yann LeCun
Daniel Lee
Michael Lee
Tai-Sing Lee
Mate Lengyel
Christina Leslie
Lukasz Lew
Michael Lewicki
Rui Li
Xiao Li
Percy Liang
Ben Liblit
Jongwoo Lim
Ruei-Sung Lim
Hsuan-Tien Lin
Richard Lippmann
Shih-Chii Liu
Karen Livescu
Daniel Lizotte
Phil Long
Manuel Loth
Jorg Lucke
Elliot Ludvig
John Lynch
Wolfgang Maass
Jakob Macke
David Madigan
Sridhar Mahadevan
Shie Mannor
Vikash Mansinghka
Mario Marchand
Tim Marks
Lluis Marquez
Regis Martinez
Amir Massoud Farahmand
Irina Mateeva
Klaus Mathiak
David McAllester
Jon McAuliffe
Andrew McCallum
Ryan McDonald
Frank McSherry
Carsten Mehring
Marina Meila
Ron Meir
Juergen Mellinger
Roland Memisevic
David Meunier
Sebastian Mika
Kai Miller
David Mimno
Thomas Minka
Tom Minka
Andriy Mnih
Volodymyr Mnih
Mehryar Mohri
Gianluigi Mongillo
Claire Monteleoni
Claire Monteleoni
Joris Mooij
Quaid Morris
Anthony Mouraud
Michael Mozer
Pawan Mudigonda
Klaus-Robert Mueller
Viktor Muller
Remi Munos
Iain Murray
Roderick Murray-Smith
Ion Muslea
Gernot Muller-Putz
Jay Myung
Srikantan Nagarajan
Vinod Nair
Dan Navarro
Arye Nehorai
James Neufeld
David Newman
Tyler Neylon
Andrew Ng
XuanLong Nguyen
Duy Nguyen-Thong
Marija Nikolic
Mahesan Niranjan
Yael Niv
David Noelle

Sebastian Nowozin
Klaus Obermayer
Guillaume Obozinski
Jean-Marc Odobez
Paul O'Grady
Frans Oliehoek
Bruno Olshausen
Manfred Opper
Sean O'Rourke
Simon Osindero
Gyorgy Ottucsak
Chris Pal
Christopher Pal
Jason Palmer
Liam Paninski
Ronald Parr
Ofer Pasternak
Hanna Pasula
Helene Paugam-Moisy
Paul Pavlidis
Barak Pearlmutter
Jiming Peng
Fernando Pereira
Fernando Perez-Cruz
Amy Perfors
Jan Peters
Robert Peters
Avi Pfeffer
Gert Pfurtscheller
Petra Philips
Jonathan Pillow
Joelle Pineau
John Platt
Jean Ponce
Massimiliano Pontil
Pascal Poupart
Marco Prato
Doina Precup
Hubert Preissl
Thomas Preuss
Eric P. Xing
Till Quack
Joaquin Quinonero Candela
Filip Radlinski
Gunnar Raetsch
Rajat Rainal
Alexander Rakhlin
Alain Rakotomamonjy
Liva Ralaivola
Ananth Ranganathan
Christopher Raphael
Ashish Rastogi
Magnus Rattray
Vikas Raykar
James Rehg
Martin Rehn
Martin Rehn
Alex Renaux
Juan Restrepo
Gael Richard
Martin Riedmiller
Brian Roark
Mark Rogers
Simon Rogers
Khashayar Rohanimanesh
Richard Rohwer
Romer Rosales
Romer Rosales
Michal Rosen-Zvi
David Ross
Saharon Rosset
Afshin Rostamizadeh
Volker Roth
Stefan Rotter
Sam Roweis
Nicholas Roy
Matthew Rudary
Cynthia Rudin
Stefan Ruping
Bryan Russell
Mehran Sahami
Maneesh Sahani
Paul Sajda
Ruslan Salakhutdinov
Maryam Saleh
Adam Sanborn
Guido Sanguinetti
Pinaki Sarder
Lawrence Saul
Craig Saunders
Tobias Scheffer
Katya Scheinberg
Alois Schlogl
Mark Schmidt
Mark Schmidt
Jeff Schneider
Bernhard Scholkopf
Sebastian Schultheiss
Anton Schwaighofer
Gabriele Schweikert
Stan Sclaroff
Matthias Seeger

Florent Segonne
Thomas Serafini
Peggy Series
Fei Sha
Gregory Shakhnarovich
Shai Shalev-Shwartz
Hanhuai Shan
Benyah Shaparenko
Tatyana Sharpee
Madhusudana Shashanka
Blake Shaw
Pradeep Shenoy
Bertram Shi
Pannagadatta Shivaswamy
Aaron Shon
Lavi Shpigelman
Leonid Sigal
Ricardo Silva
John Simeral
Alex Simma
Vikas Sindhwani
Yoram Singer
Fabian Sinz
Josef Sivic
Noam Slonim
Paris Smaragdis
Cristian Sminchisescu
Jared Smith
Noah Smith
Trey Smith
Alex Smola
Edward Snelson
Stefano Soatto
Jascha Sohl-Dickstein
Artem Sokolov
Kemal Sonmez
Cheng Soon Ong
Finnegan Southey
Suvrit Sra
Nathan Srebro
Nisheeth Srivastava
Benjamin Staude
Harald Steck
Ingo Steinwart
Gilles Stoltz
Amos Storkey
Erik Sudderth
Masashi Sugiyama
Arun Surendran
Ilya Sutskever
Richard Sutton
Johan Suykens
Csaba Szepesvari
Ameet Talwalkar
Michael Tangermann
Brian Tanner
Ben Taskar
A. Taylan Cemgil
Russ Tedrake
Joshua Tenenbaum
Ambuj Tewari
Evangelos Theodorou
Romain Thibaux
Alexander Thomas
Sebastian Thrun
Tai-peng Tian
Jo-Anne Ting
Naftali Tishby
Emanuel Todorov
David Tolliver
Ryota Tomioka
Antonio Torralba
Marc Toussaint
Kristina Toutanova
Volker Tresp
Bill Triggs
Koji Tsuda
Gokhan Tur
Tinne Tuytelaars
Naonori Ueda
Patrick van der Smagt
Joost van de Weijer
Martijn van Otterlo
Wolf Vanpaemel
Andre van Schaik
Manik Varma
Theresa Vaughan
Nicolas Vayatis
Shankar Vembu
Jakob Verbeek
Jean-Philippe Vert
Sethu Vijayakumar
Pascal Vincent
Nikos Vlassis
S V N Vishwanathan
Eric-Jan Wagenmakers
Kiri Wagstaff
Martin Wainwright
Christian Walder
Xiang Wan
Shaojun Wang
Tao Wang

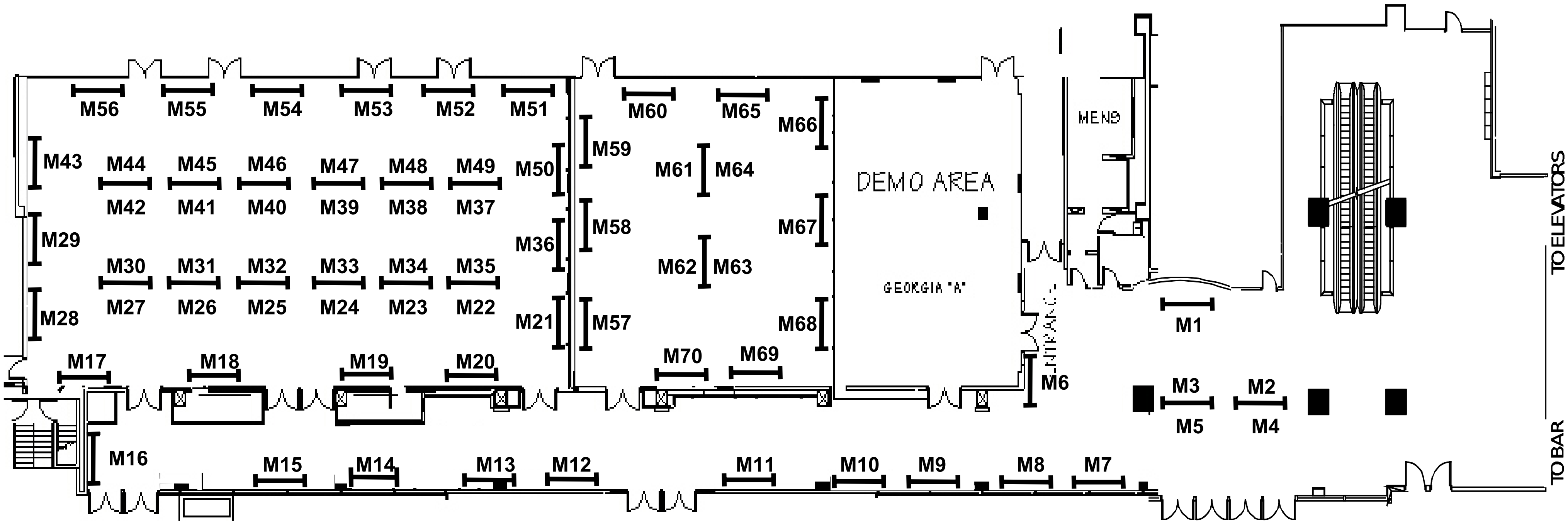
Xuerui Wang
Manfred Warmuth
Kilian Weinberger
Yair Weiss
Bastian Wemmenhove
Scott Wen-tau Yih
Wim Wiegerinck
Eric Wiewiora
Dana Wilkinson
David Wingate
John Winn
Ole Winther
Weng-Keen Wong
Frank Wood
Phil Woodland
Wei Wu
Huan Xu
Linli Xu
Stella X. Yu
Ming-Hsuan Yang
Junming Yin
Angela Yu
Byron Yu
Chen Yu
Chun-Nam Yu
Kai Yu
Shipeng Yu
Quan Yuan
Yisong Yue
Alan Yuille
Noah Zaitlen
Thorsten Zander
Luca Zanni
Georg Zeller
Richard Zemel
Hongyuan Zha
ChengXiang Zhai
Zhihua Zhang
Tong Zhang
Tong Zhao
Dengyong Zhou
Xiaojin Zhu
Alexander Zien
Onno Zoeter
Geoffrey Zweig

Notes

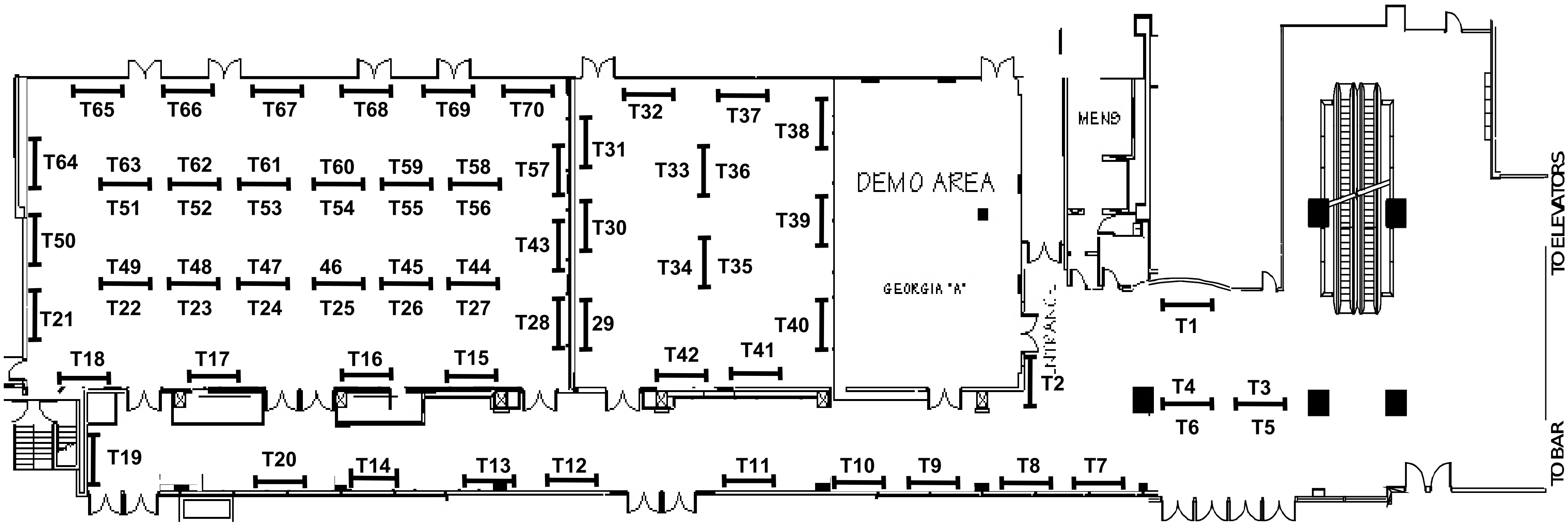
Subject Index

Author Index

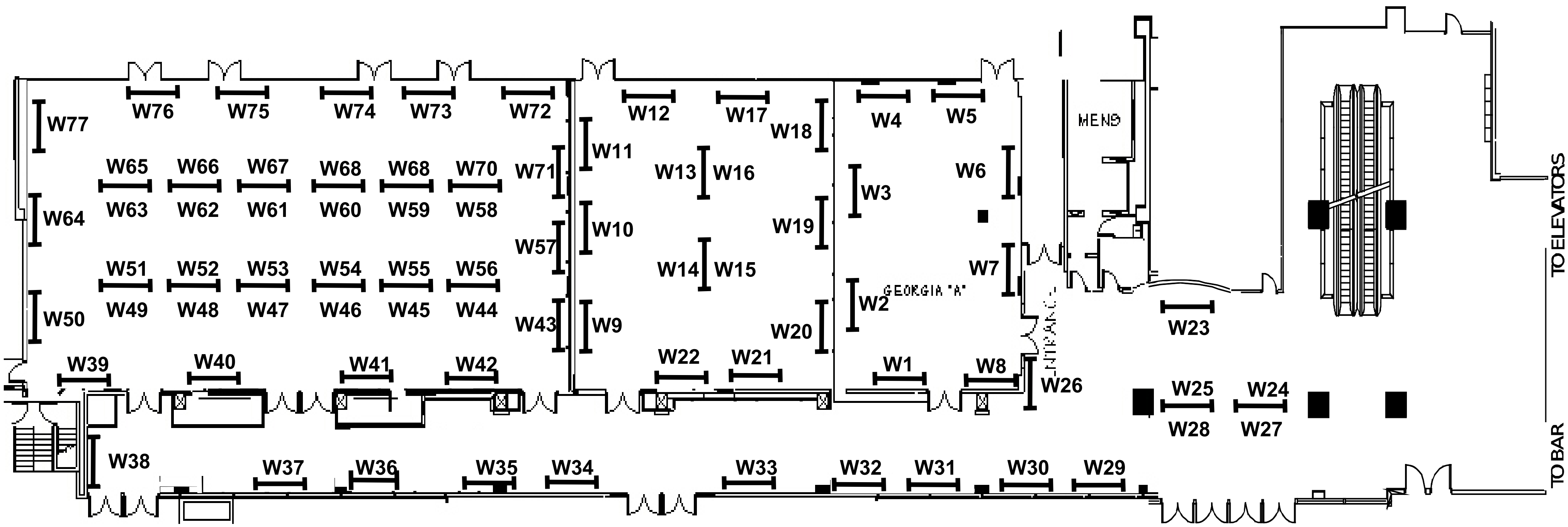
Monday Poster Session

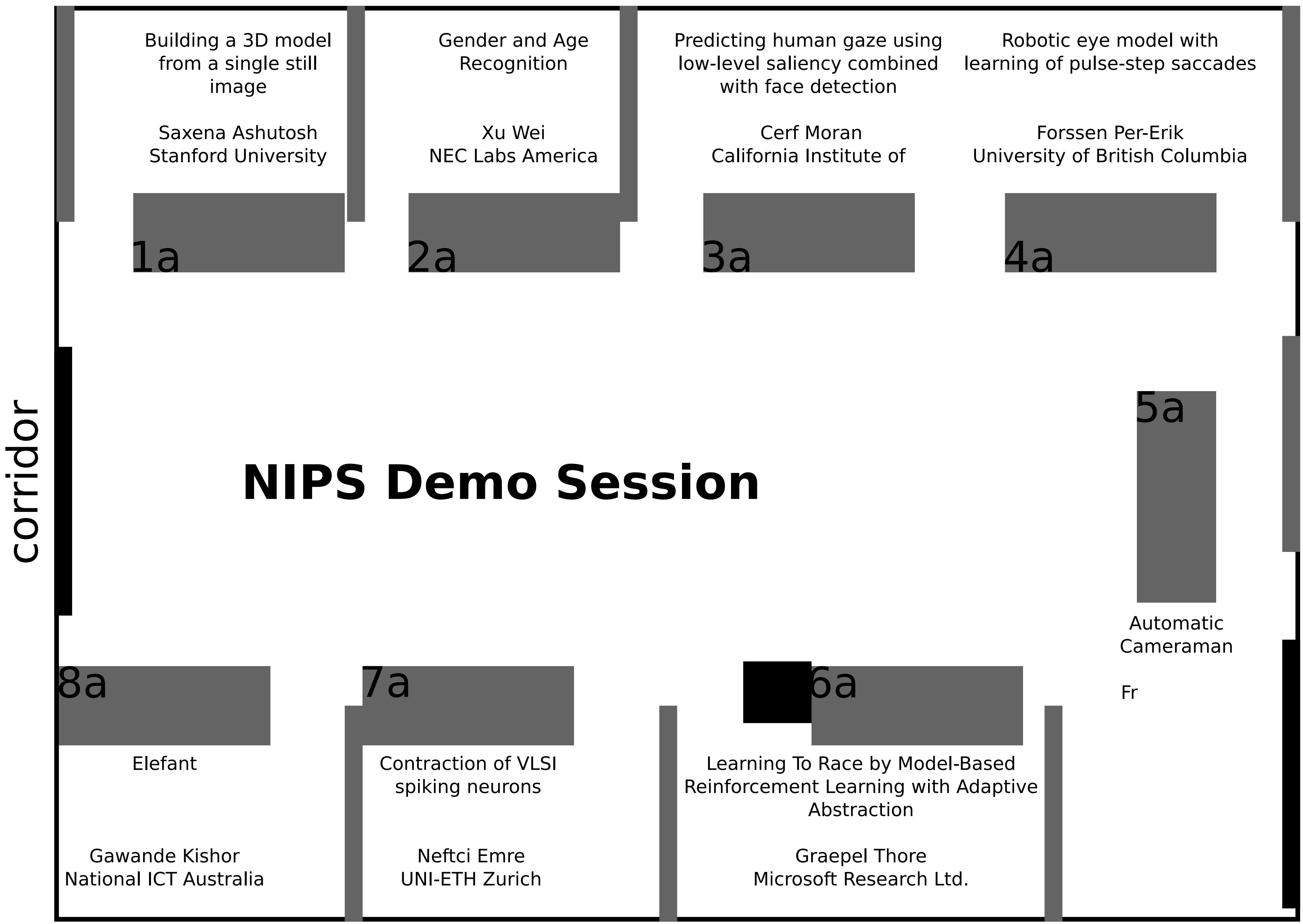


Tuesday Poster Session

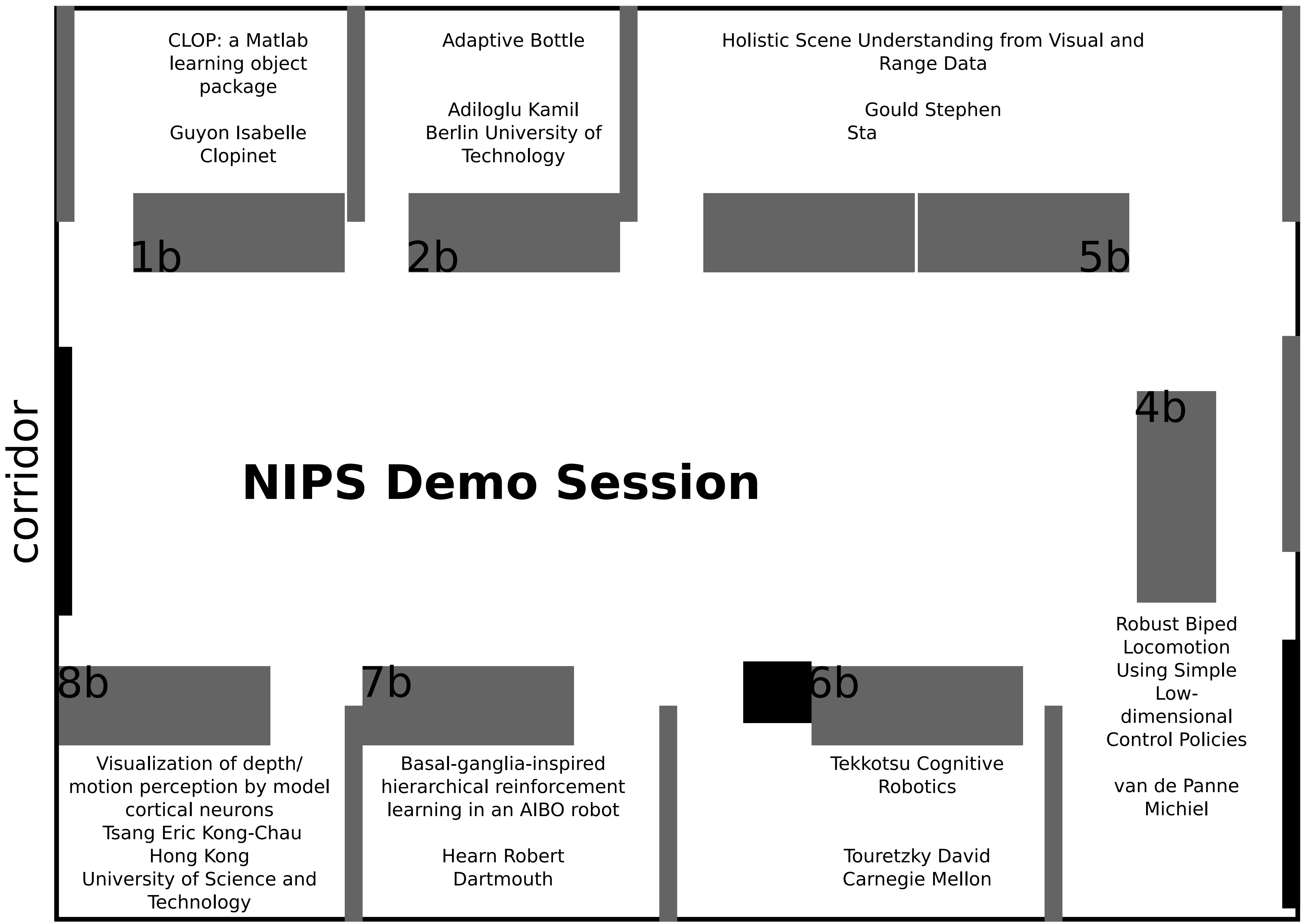


Wednesday Poster Session





Building a 3D model from a single still image
Saxena Ashutosh
Stanford University
1a
Gender and Age Recognition
Xu Wei
NEC Labs America
2a
Predicting human gaze using low-level saliency combined with face detection
Cerf Moran
California Institute of
3a
Robotic eye model with learning of pulse-step saccades
Forssen Per-Erik
University of British Columbia
4a
corridor
NIPS Demo Session
5a
Automatic Cameraman
Fr
8a
Elefant
Gawande Kishor
National ICT Australia
7a
Contraction of VLSI spiking neurons
Neftci Emre
UNI-ETH Zurich
6a
Learning To Race by Model-Based Reinforcement Learning with Adaptive Abstraction
Graepel Thore
Microsoft Research Ltd.


NIPS Demo Session
corridor
1b
CLOP: a Matlab learning object package
Guyon Isabelle
Clopinet
2b
Adaptive Bottle
Adiloglu Kamil
Berlin University of Technology
5b
Holistic Scene Understanding from Visual and Range Data
Gould Stephen
Sta
4b
Robust Biped Locomotion Using Simple Low-dimensional Control Policies
van de Panne Michiel
8b
Visualization of depth/motion perception by model cortical neurons
Tsang Eric Kong-Chau
Hong Kong University of Science and Technology
7b
Basal-ganglia-inspired hierarchical reinforcement learning in an AIBO robot
Hearn Robert
Dartmouth
6b
Tekkotsu Cognitive Robotics
Touretzky David
Carnegie Mellon

Useful Information

Frequently Asked Questions

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.

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	7:00 - 9:00 AM	Perspectives, 34th Floor

At the Workshops:

Opening Gathering (Snacks & Cash Bar)	Thursday	6:30 - 8:30 PM	Emerald Ballroom
Breakfast	Friday, Saturday	6:30 - 8:00 AM	Emerald Ballroom
Closing Banquet	Saturday	7:00 - 11:00 PM	Emerald Ballroom

Note: There is a $35 charge to bring a family member or friend to the banquet at the Hyatt Regency Vancouver or 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.

TUTORIALS REGISTRATION:
Does the fee for the NIPS Tutorials include all the tutorials or is the fee per tutorial?
The fee includes all Tutorials. Hotel accommodation is separate.

WORKSHOP REGISTRATION:

Does the Workshop registration fee include all the Workshops?

The Workshop registration fee includes all Workshops. Hotel accommodation is separate.

WIRELESS INTERNET:

Is wireless available at the NIPS Conference?

The Wireless information page on the NIPS website contains information on how to get network access in the Conference and Workshop areas of the host hotels in Vancouver and Whistler. The document changes every year to reflect updated information for the current meeting. To view the document, go to http://nips.cc, hover your mouse over the current meeting and look for a flyout called Wireless.

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.

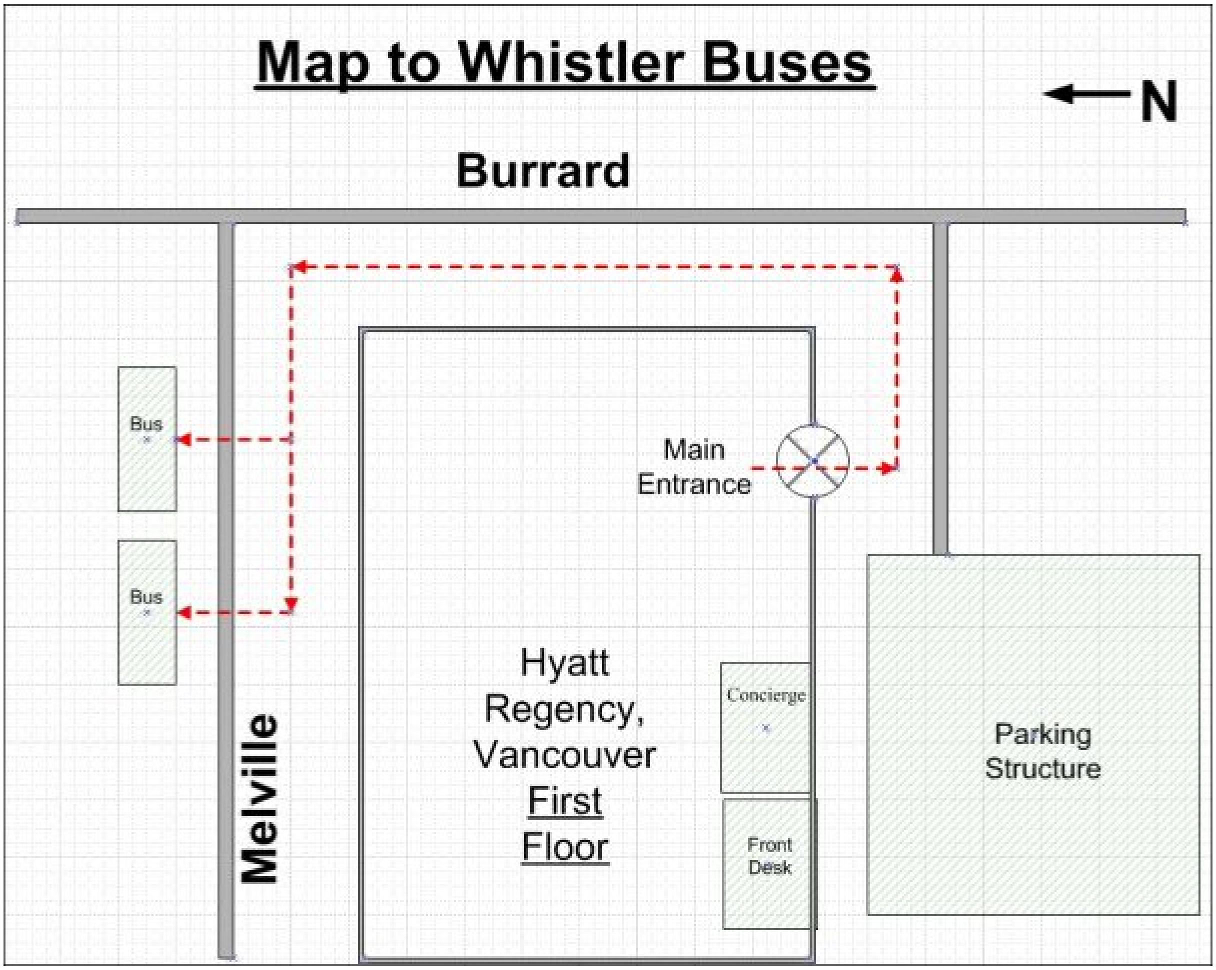
Buses and Transportation

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 6. 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, December 5.

Buses will leave from the Hyatt Regency on Thursday, December 6, beginning at 2:00 pm and running every half hour until 3:30 pm. Buses will load 15 minutes prior to departure – please take this into account in your planning.

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 9, 2007 only.

The rate for NIPS 2007 participants is $57 CA (plus 6% GST) (~$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 the Perimeter website: http://www.perimeterbus.com/. Use the group code: NIPS 2007.
Or fax Perimeter the registration form (http://media.nips.cc/Conferences/2007/Bus/Perimeter-WhistlerExpress-Fax.pdf)

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 5, 2007.

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 17th and December 5th, 2007.
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 6	5pm - 9pm
Friday, December 7	9am - 11am
Saturday, December 8	9am - 11am

Hotel Contact Information:

Hyatt Regency
655 Burrard Street
Vancouver, British Columbia
Canada V6C 2R7
1-604-683-1234

Westin Whistler Resort and Spa
4090 Whistler Way
Whistler, British Columbia
Canada V0N 1B4
1-888-634-5577

The Hilton Whistler Resort & Spa
4050 Whistler Way
Whistler, British Columbia
Canada V0N 1B4
1-800-515-4050

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

Recommended Restaurants in the Vancouver Area

Joe Fortes – 777 Thurlow Street

(2 Blocks from the Hyatt Regency Vancouver)
Vancouver's original oyster bar. Upscale restaurant specializing in seafood.
(Entrees: $18.00 - $30.00)

Raincity Grill – 1193 Denman Street

Located at English Bay with spectacular water views
Regional cuisine, fresh local fare is ingeniously prepared, often with organic ingredients.
(Entrees: $20.00 - $30.00)

Shanghai Bistro – 1128 Alberni Street

(2 Blocks from the Hyatt Regency Vancouver)
Excellent Szechwan and Cantonese cuisine in a bright and modern room. Watch for the noodle making show nightly

Sequoia Grill -Ferguson Point

(10 minutes by taxi from the Hyatt Regency Vancouver)
Enjoy a magnificent view of English Bay while dining on creative regional dishes.
(Entrees: $14.00 - $24.00)

Villa Del Lupo – 869 Hamilton Street

(8 minutes by taxi from the Hyatt Regency Vancouver)
Critics favorite year after year. Tucson Menu in a simple and elegant heritage décor.   Family Service
(Entrees: $16.00 - $27.00)

Cin Cin – 1154 Robson Street

(5 Blocks from the Hyatt Regency Vancouver)
Mediterranean style food and restaurant. A gorgeous Tucson room permeated by smoked alder, cherry and apple from the wood fire oven and grills.
(Entrees: $12.00 - $30.00)

Blue Water Café & Raw Bar – 1095 Hamilton Street

(10 minutes by taxi from the Hyatt Regency Vancouver)
Applauded for serving some of the city's freshest and best seafood
(Entrees: $16.00 - $27.00)

Nu Restaurant – 1661 Granville Street

Nu embodies the newest of dining styles in Vancouver.
(Entrées: Under $30.00)

The hotel concierge would be happy to help with reservations or more suggestions.

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 http://nips2007.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 have changed, click on the magnifying glass and modify the data for your accepted paper.

   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 Pacific Standard Time, Friday, January 11, 2008.

2. **Upload keywords:** The nips2007.confmaster.net website permits you to upload "Free Keywords". Please log onto the site and upload/check your paper's keywords. You can find suggested keywords on page 181 of this booklet, or use your own descriptive keywords. Please limit yourself to a maximum of 7 keywords. Please upload the keywords by Monday, January 7, 2008.

3. **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 `nips07@cs.stanford.edu`. 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 Monday January 7, 2008. If you do not have access to a scanner, you can fax your signed form to +1-650-725-1449 but scanning is the preferred method if at all possible.

4. 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 are available on the NIPS Instructions for Authors `http://nips07.stanford.edu/instructions.html`. Detailed formatting instructions can be found at `http://nips07.stanford.edu/nips2007.pdf`

5. 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.

6. 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.

7. 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 20

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: ____

____

____