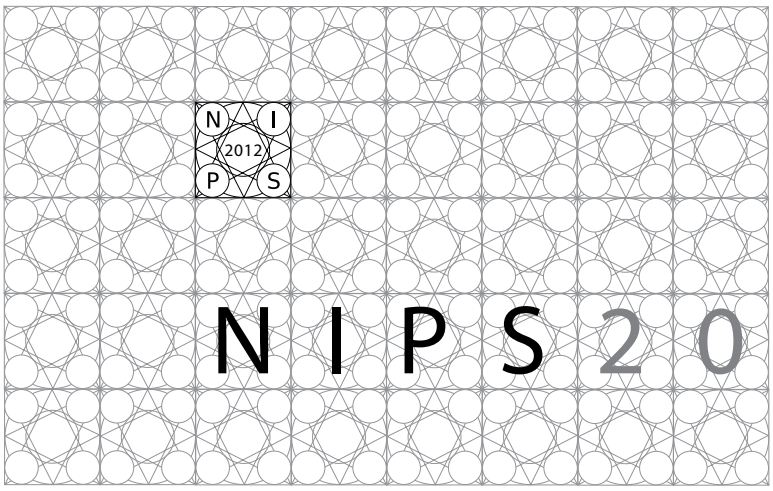
2012 CONFERENCE BOOK

Neural Information Processing Systems



NIPS 2012

Abstracts of Papers

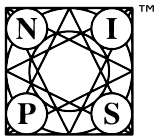
TUTORIALS
December 3, 2012
Harrah's & Harveys
Lake Tahoe, Nevada

CONFERENCE SESSIONS
December 4 - 6, 2012
Harrah's & Harveys
Lake Tahoe, Nevada

WORKSHOPS
December 7 - 8, 2012
Harrah's & Harveys
Lake Tahoe, Nevada

Sponsored by the Neural Information Processing System Foundation, Inc

The technical program includes 6 invited talks and 370 accepted papers, selected from a total of 1,452 submissions considered by the program committee. Because the conference stresses interdisciplinary interactions, there are no parallel sessions. Papers presented at the conference will appear in "Advances in Neural Information Processing Systems 25," edited by Peter Bartlett, Fernando Pereira, Léon Bottou, Chris Burges and Killian Weinberger



Neural Information Processing Systems Foundation

TABLE OF CONTENTS

Neural Information Processing Systems Foundation

ORGANIZING COMMITTEE

General Chairs:	**Peter Bartlett** (Queensland University of Technology and University of California, Berkeley); **Fernando Pereira** (Google Research)
Program Chairs:	**Leon Bottou** (Microsoft Research); **Chris J.C. Burges** (Microsoft Research)
Tutorials Chair:	**Tom Griffiths** (University of California, Berkeley)
Workshop Chairs:	**Raquel Urtasun** (TTI-Chicago); **Máté Lengyel** (University of Cambridge)
Demonstration Chair:	**Thore Graepel** (Microsoft Research)
Publications Chair and Electronic Proceedings Chair:	**Kilian Weinberger** (Washington University in St. Louis)
Program Manager:	**Jesper Lind** (Microsoft Research)

PROGRAM COMMITTEE

Shai Ben-David (U. Waterloo)
Samy Bengio (Google)
Matthias Bethge (Max Planck Institute Tübingen)
Alexandre Bouchard-Côté (U. British Columbia)
Sebastien Bubeck (Princeton)
Tiberio Caetano (NICTA Canberra)
Lawrence Carin (Duke)
Ronan Collobert (IDIAP Martigny)
Marco Cuturi (Kyoto University)
Pedro Domingos (U. Washington)
Tina Eliassi-Rad (Rutgers)
Rob Fergus (New York University)
François Fleuret (IDIAP Martigny)
Paolo Frasconi (U. Firenze)
Amir Globerson (Hebrew University)
Geoffrey Goodhill (U. of Queensland)
Hans-Peter Graf (NEC Labs Princeton)
Yves Grandvalet (Université de Technologie de Compiègne)
Kristen Grauman (U. Texas)
Patrick Haffner (AT&T Labs - Research)
Tamir Hazan (Toyota Technological Institute Chicago)
Katherine Heller (MIT)
Daniel Hsu (MSR New England)
Shiro Ikeda (ISM Tokyo)
Prateek Jain (MSR Bangalore)
Thorsten Joachims (Cornell)
Neil Lawrence (U. Sheffield)
Ping Li (Cornell)
Phil Long (Google)
Klaus-Robert Mueller (TU Berlin)
Remi Munos (Inria Lille)
Noboru Murata (Waseda University)
Iain Murray (U. Edinburgh)
Sebastian Nowozin (MSR Cambridge)
Klaus Obermayer (TU Berlin)
Aude Oliva (MIT)
Cheng Soon Ong (ETH Zürich)
Barak Pearlmutter (National University of Ireland Maynooth)
Doina Precup (McGill Montreal)
Gunnar Raetsch (Memorial Sloan Kettering)
Marc'Aurelio Ranzato (Google)
Cynthia Rudin (MIT)
Ruslan Salakhutdinov (MIT)
Ashutosh Saxena (Cornell)
Cordelia Schmid (Inria Alpes)
Fei Sha (U. Southern California)
Shai Shalev-Shwartz (Hebrew University)
Aarti Singh (CMU)
Alex Smola (Yahoo! Research)
Masaaki Sugiyama (Tokyo Institute of Technology)
Csaba Szepervari (U. Alberta)
Ambuj Tewari (U. Texas)
Raquel Urtasun (Toyota Technological Institute Chicago)
Jean-Philippe Vert (Mines ParisTeche)
Sethu Vijayakumar (Edinburgh)
Killian Weinberger (U. Washington)
Lin Xiao (MSR Redmond)
Eric Xing (CMU)
Jieping Ye (Arizona State U.)
Angela Yu (UCSD)
Kai Yu (NEC Labs Cupertino)
Xiaojin "Jerry" Zhu (U. Wisconsin)

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

NIPS FOUNDATION OFFICERS & BOARD MEMBERS

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

CORE LOGISTICS TEAM

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

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

SPONSORS

NIPS gratefully acknowledges the generosity of those individuals and organizations who have provided financial support for the NIPS 2012 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™

facebook

Microsoft Research

D E Shaw & Co

TWO SIGMA

IBM Research

GREENPLUM EMC²

HUAWEI

JANE STREET

ebay Research Labs

TOYOTA

United Technologies Research Center

YAHOO! LABS

SKYTREE

Springer
Machine Learning Journal

QUALCOMM

PROGRAM HIGHLIGHTS

MONDAY DEC 3RD

7:30 am – 6:30 pm
Registration Desk Open
Harveys Convention Center Floor, CC

8:00 am – 9:30 am
Breakfast, See map page 8

9:30 am – 5:30 pm
Tutorials
Harveys Convention Center Floor, CC

6:30 – 6:55 pm
Opening Remarks, Awards and Reception
Harrah's Special Events Center, 2nd Floor

7:00 – 11:59 pm
Poster Session
Harrah's Special Events Center, 2nd Floor

TUESDAY DEC 4TH

7:30 am – 9:30 am
Breakfast sponsored by Winton Capital
See map page 8

8 am – 5:30 pm
Registration Desk Open
Harveys Convention Center Floor, CC

9:00 – 10:10 am
Oral Session 1
Quantum information and the Brain
Invited Talk: Scott Aaronson

TCA: High Dimensional Principal Component Analysis for non-Gaussian Data
F. Han, H. Liu

10:10 –10:30 am
Spotlights Session 1

10:30 – 11:00 am - Coffee Break

11:00 –11:40 am
Oral Session 2
Spectral Learning of General Weighted Automata via Constrained Matrix Completion
B. Balle, M. Mohri

Relax and Randomize : From Value to Algorithms Completion
A. Rakhlin, O. Shamir, K. Sridharan

TUESDAY DEC 4TH

11:40 – 12:05 pm
Spotlights Session 2

12:05 – 2:00 pm - Lunch Break

2:00 – 3:30 pm
Oral Session 3
Classification with Deep Invariant Scattering Networks
Invited Talk: Stephane Mallat

A Stochastic Gradient Method with an Exponential Convergence Rate with Finite Training Sets
N. Le Roux, M. Schmidt, F. Bach

Approximating Concavely Parameterized Optimization Problems
J. Giesen, J. Mueller, S. Laue, S. Swiercy

3:30 – 3:50 pm
Spotlights Session 3

3:50 – 4:20 pm - Coffee Break

4:20 – 5:40 pm
Oral Session 4:
Spectral learning of linear dynamics from generalised linear observations with application to neural population data
L. Buesing, J. Macke, M. Sahani

High-Order Multi-Task Feature Learning to Identify Longitudinal Phenotypic Markers for Alzheimer Disease Progression Prediction
H. Wang, F. Nie, H. Huang, J. Yan, S. Kim, S. Risacher, A. Saykin, L. Shen

Multimodal Learning with Deep Boltzmann Machines
N. Srivastava, R. Salakhutdinov

Discriminative Learning of Sum-Product Networks
R. Gens, P. Domingos

5:40 – 6:00 pm
Spotlight Session 4

5:45 – 11:59 pm
Poster Session
Harrah's Special Events Center, 2nd Floor

PROGRAM HIGHLIGHTS

WEDNESDAY DEC 5TH

7:30 am – 9:30 am
Breakfast, See map page 8

8 am – 5:30 pm
Registration Desk Open
Harveys Convention Center Floor, CC

9:00 – 10:10 am
Oral Session 5
Challenges for Machine Learning in Computational Sustainability
Posner Lecture: T. Dietterich

Augmented-SVM: Automatic space partitioning for combining multiple non-linear dynamics
A. Shukla, A. Billard

10:10 –10:30 am
Spotlights Session 5

10:30 – 11:00 am - Coffee Break

11:00 –11:40 am
Oral Session 2
On the Use of Non-Stationary Policies for Stationary Infinite-Horizon Markov Decision Processes
B. Scherrer, B. Lesner

A Unifying Perspective of Parametric Policy Search Methods for Markov Decision Processes
T. Furmston, D. Barber

11:40 – 12:05 pm
Spotlights Session 6

12:05 – 2:00 pm - Lunch Break

2:00 – 3:30 pm
Oral Session 7
Signatures of Conscious Processing in the Human Brain
Invited Talk: Stanislas Dehaene

Privacy Aware Learning
J. Duchi, M. Jordan, M. Wainwright

On the (Non-)existence of Convex, Calibrated Surrogate Losses for Ranking
C. Calauzènes, N. Usunier, P. Gallinari

3:30 – 3:50 pm
Spotlights Session 7

3:50 – 4:20 pm - Coffee Break

4:20 – 5:40 pm
Oral Session 8
Graphical Models via Generalized Linear Models
E. Yang, P. Ravikumar, G. Allen, z. Liu

No voodoo here! Learning discrete graphical models via inverse covariance estimation
P. Loh, M. Wainwright

Near-Optimal MAP Inference for Determinantal Point Processes
A. Kulesza, J. Gillenwater, B. Taskar

Bayesian nonparametric models for bipartite graphs
F. Caron

5:40 – 6:00 pm
Spotlight Session 8

5:45 – 11:59 pm
Poster Session
Harrah's Special Events Center, 2nd Floor

THURSDAY DEC 6TH

7:30 am – 9:30 am
Breakfast, See map page 8

9:00 – 10:10 am
Oral Session 9
Suspicious Coincidences in the Brain
Posner Lecture: T. Sejnowski

Strategic Impatience in Go/NoGo versus Forced-Choice Decision-Making
P. Shenoy, A. Yu

10:10 –10:30 am
Spotlights Session 9

10:30 – 11:00 am - Coffee Break

11:00 –11:40 am
Oral Session 10
Fast Algorithms, Matrix Compression and Design by Simulation
Invited Talk: Leslie Greengard

Gradient Weights help Nonparametric Regressors
S. Kpotufe, A. Boularias

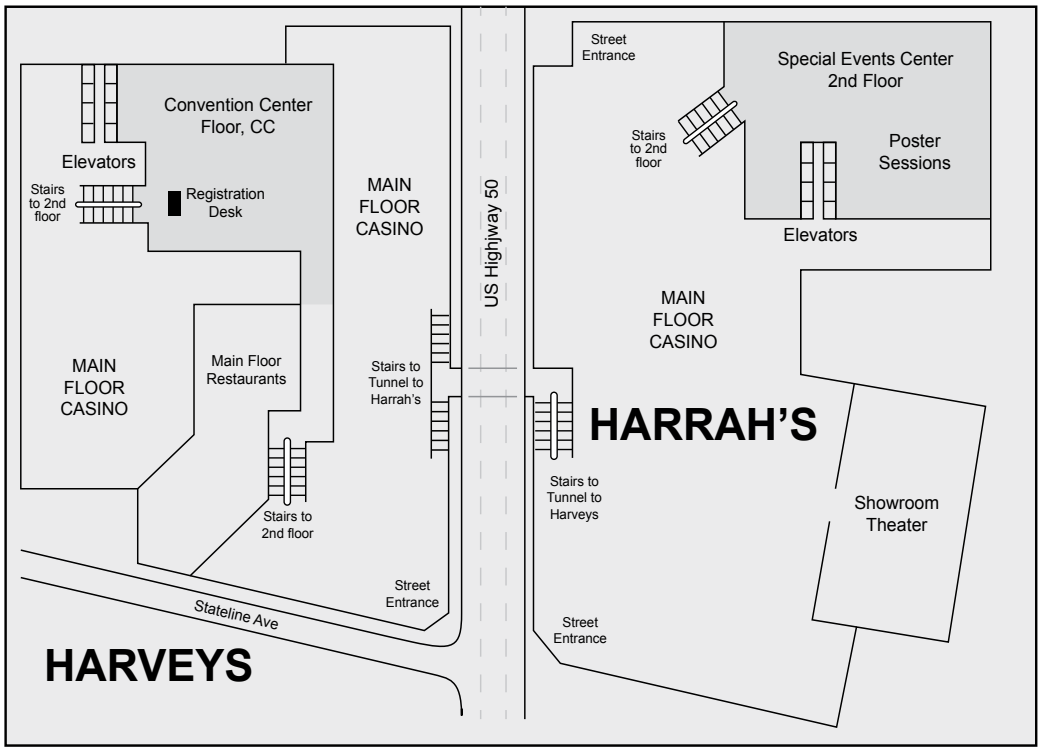
11:40 – 12:05 pm
Spotlights Session 6

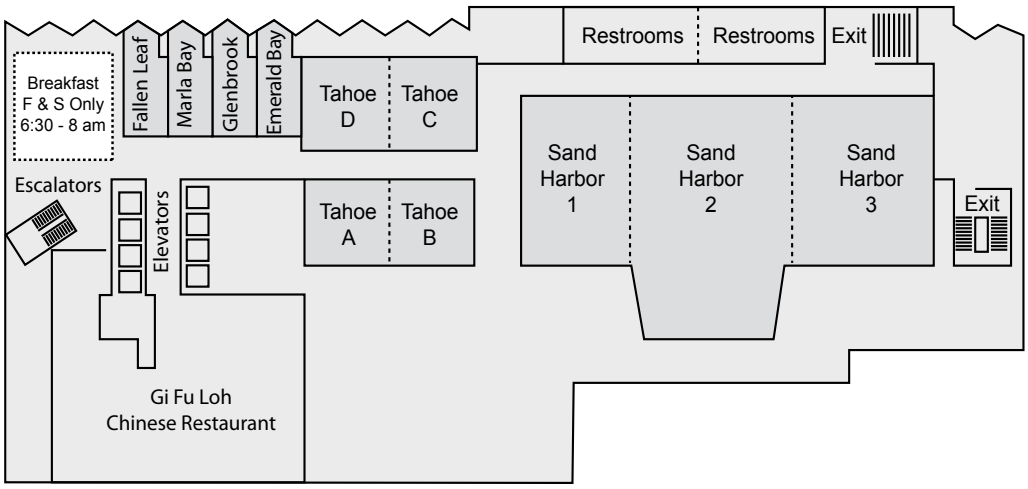
12:20 – 12:30 pm - Closing remarks

12:30 – 2:00 pm - Lunch Break

2:00 – 6:00 pm
Poster Session
Harrah's Special Events Center, 2nd Floor

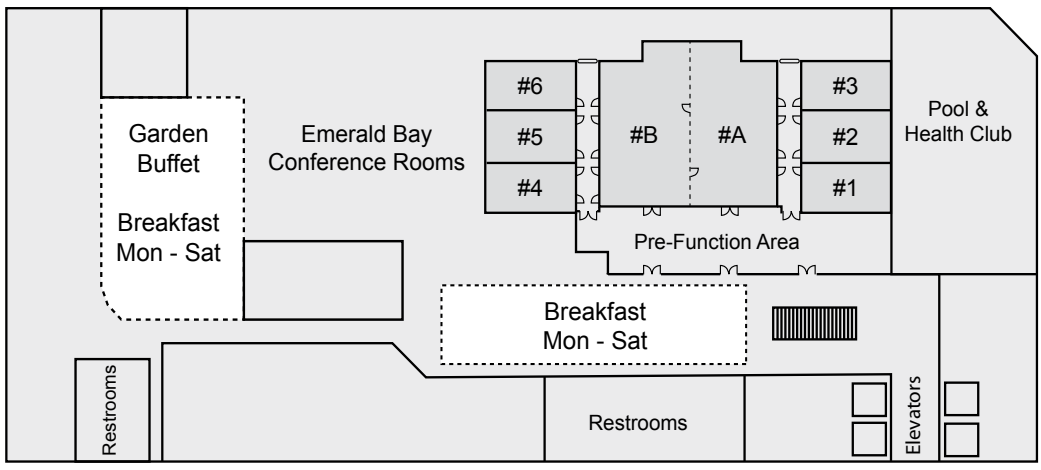
HARRAH'S/HARVEYS MAPS





HARRAH'S SPECIAL EVENTS CENTER 2ND FLOOR

3 Breakfast Areas
(dotted areas)
1 Breakfast area at
Top Of The Wheel
Restaurant in
Harvey's 12th Floor



HARVEYS CONVENTION CENTER FLOOR, CC

MONDAY

MONDAY TUTORIALS

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

Exact Approximate Learning
Paul Fearnhead
Location: Emerald Bay A, Harveys

Representation, Inference and Learning in Structured Statistical Models
Lise Getoor
Location: Emerald Bay B, Harveys

1:00 – 3:00 PM - **Tutorial Session 2**

User-Friendly Tools for Studying Random Matrices
Joel Tropp
Location: Emerald Bay A, Harveys

Machine Learning for Student Learning
Emma Brunskill, Geoffrey Gordon
Location: Emerald Bay B, Harveys

3:30 PM – 5:30 PM - Tutorial Session 3

Tutorial: Stochastic Search and Optimization
James C. Spall
Location: Emerald Bay A, Harveys

Tutorial: Consciousness and Information Theory
Giulio Tononi, Christof Koch
Location: Emerald Bay B, Harveys

Tutorial Session 1, 9:30 – 11:30am

Exact Approximate Learning

Paul Fearnhead, Lancaster University

There are many natural approximations that can be used within statistical learning. For example, in MCMC we could use a numerical or Monte Carlo approximation to the acceptance probability in cases where the target distribution cannot be written down (even up to a constant of proportionality). Or when sampling from an infinite-dimensional distribution, for example in Bayesian non-parametrics, we can use a finite-dimensional approximation (e.g. by truncating the tail of the true distribution). Recent work has shown that, in some cases, we can make these "approximations" and yet the underlying methods will still be "exact". So our MCMC algorithm will still have the correct target distribution, or we will still be drawing samples from the true infinite dimensional distributions. Informally, the key idea behind these "exact approximate" methods is that we are able to randomise the approximation so as to average it away. This tutorial will cover the two main examples of "exact approximate" methods: the pseudo-marginal approach and retrospective sampling. The ideas will be demonstrated on examples taken from Bayesian non-parametrics, changepoint detection and diffusions.

Paul Fearnhead *is Professor of Statistics at Lancaster University. He received his DPhil in Statistics from the University of Oxford in 1998; was a postdoctoral researcher at the University of Oxford until 2001; and then moved to the University of Lancaster, initially as a Lecturer in Statistics. He has worked on Monte Carlo methods within Bayesian statistics, including applications in population genetics, changepoint detection and inference for diffusions. He was awarded the Royal Statistical Society's Guy medal in Bronze in 2007, and Cambridge University's Adams Prize in 2006.*

ABSTRACTS OF TUTORIALS

Tutorial Session 1, 9:30 – 11:30am

Representation, Inference and Learning in Structured Statistical Models

Lise Getoor, University of Maryland

Addressing inherent uncertainty and exploiting structure are fundamental to understanding, designing and making predictions in large-scale information, biological and socio-technical systems. Statistical relational learning (SRL) builds on principles from probability theory and statistics to address uncertainty while incorporating tools from logic to represent structure. SRL methods are especially well-suited to domains where the input is best described as a large multi-relational network, such as online social media and communication networks, and we need to make structured predictions.

The first part of the tutorial will provide an introduction to key SRL concepts, including relational feature construction and representation, inference and learning methods for "lifted graphical models." The second part of the tutorial will describe three important challenges in network analysis: graph identification (inferring a graph from noisy observations), graph alignment (mapping components in one graph to another) and graph summarization (clustering the nodes and edges in a graph). I will overview approaches to these problems based on SRL methods, describe available datasets, and highlight opportunities for future research.

Throughout, I will pay particular attention to scaling and make connections to related areas of machine learning such as structured prediction and latent factor models.

Lise Getoor *is an Associate Professor in the Computer Science Department and the Institute for Advanced Computer Studies at the University of Maryland, College Park. Her research areas include machine learning, reasoning under uncertainty, and database management. She is co-editor with Ben Taskar of the book 'An Introduction to Statistical Relational Learning', MIT Press, 2007. She is a board member of the International Machine Learning Society, and has served as Machine Learning Journal Action Editor, Associate Editor for the ACM Transactions of Knowledge Discovery from Data, JAIR Associate Editor, and on the AAAI Council. She is a recipient of several best paper awards, an NSF Career Award and a National Physical Sciences Consortium Fellowship. She received her PhD from Stanford University, her Master's degree from the University of California, Berkeley, and her undergraduate degree from the University of California, Santa Barbara.*

ABSTRACTS OF TUTORIALS

Tutorial Session 2, 1:00 - 3:00pm

User-Friendly Tools for Studying Random Matrices

Joel Tropp, California Institute of Technology

Random matrices have come to play a significant role in computational mathematics and statistics. Established methods from random matrix theory have led to striking advances in these areas, but ongoing research has generated difficult questions that cannot be addressed without new tools. The purpose of this tutorial is to introduce some recent techniques, collectively called matrix concentration inequalities, that can simplify the study of many types of random matrices. These results parallel classical tail bounds for scalar random variables, such as the Bernstein inequality, but they apply directly to matrices. In particular, matrix concentration inequalities can be used to control the spectral norm of a sum of independent random matrices by harnessing basic properties of the summands. Many variants and extensions are now available, and the outlines of a larger theory are starting to emerge. These new techniques have already led to advances in many areas, including partial covariance estimation, randomized schemes for low-rank matrix decomposition, relaxation and rounding methods for combinatorial optimization, construction of maps for dimensionality reduction, techniques for subsampling large matrices, analysis of sparse approximation algorithms, and many others.

Joel A. Tropp *is Professor of Applied & Computational Mathematics at California Institute of Technology. He earned the Ph.D. degree in Computational Applied Mathematics from the University of Texas at Austin in 2004. Prof. Tropp's work lies at the interface of applied mathematics, electrical engineering, computer science, and statistics. The bulk of this research concerns the theoretical and computational aspects of sparse approximation, compressive sampling, and randomized linear algebra. He has also worked extensively on the properties of structured random matrices. Prof. Tropp has received several major awards for young researchers, including the 2007 ONR Young Investigator Award and the 2008 Presidential Early Career Award for Scientists and Engineers. He is also winner of the 32nd annual award for Excellence in Teaching from the Associated Students of the California Institute of Technology.*

Tutorial Session 2, 1:00 - 3:00pm

Machine Learning for Student Learning

Emma Brunskill, Carnegie Mellon University
Geoffrey Gordon, Carnegie Mellon University

Intelligent tutoring systems and online classes have the potential to revolutionize education. Realizing this potential requires tackling a large number of challenges that can be framed as machine learning problems. We will first provide a survey of several machine learning problems in education, such as modeling a student's thought process as she solves a problem, constructing the atoms of knowledge, and automated problem design. We will then discuss cognitive modeling and instructional policy construction in more depth, and describe state-of-the-art methods as well as ongoing challenges. Throughout the tutorial we will highlight where student learning results in opportunities for new algorithmic and theoretical advances in machine learning.

Emma Brunskill *is an Assistant Professor of Computer Science and an Affiliated Assistant Professor of Machine Learning at Carnegie Mellon University. Prior to this, she completed her PhD at the Massachusetts Institute of Technology and was a NSF Mathematical Sciences Postdoctoral Fellow at UC Berkeley. Her primary research is on sequential decision making under uncertainty, and she is particularly excited about applications of this work to intelligent tutoring systems and healthcare. Emma is also interested in how information technology can be used to help address challenges that arise in low resource areas. She is a Rhodes Scholar and was recently selected as a Microsoft Faculty Fellow.*

Dr. Geoffrey Gordon *is an Associate Research Professor in the Department of Machine Learning at Carnegie Mellon University, and co-director of the Department's Ph. D. program. He works on multi-robot systems, statistical machine learning, game theory, and planning in probabilistic, adversarial, and general-sum domains. His previous appointments include Visiting Professor at the Stanford Computer Science Department and Principal Scientist at Burning Glass Technologies in San Diego. Dr. Gordon received his B.A. in Computer Science from Cornell University in 1991, and his Ph.D. in Computer Science from Carnegie Mellon University in 1999*

ABSTRACTS OF TUTORIALS

Tutorial Session 3, 3:30 – 5:30 pm

Stochastic Search and Optimization

James C. Spall, Johns Hopkins University

Stochastic search and optimization (SS&O) methods are widely used in many areas of computational science. Online algorithms, such as stochastic gradient descent, are a prominent example of SS&O. The speaker will discuss some general issues related to how SS&O contributes to the analysis and control of modern systems as a way of: (i) coping with inherent system noise, (ii) providing algorithms that are relatively insensitive to modeling uncertainty, and (iii) providing algorithms that are able to find a global solution from among multiple local solutions. As a specific example of SS&O, the speaker will discuss the simultaneous perturbation stochastic approximation (SPSA) algorithm for difficult multivariate optimization problems arising in stochastic systems. The essential feature of SPSA, which accounts for its power and relative ease of use in difficult multivariate optimization problems, is the underlying gradient approximation that requires only two objective function measurements regardless of the dimension of the optimization problem. This talk will focus on the basic ideas and motivation behind SPSA without dwelling on the mathematical details. As time permits, the speaker will also include some discussion on contrasts with other algorithms (genetic algorithms, simulated annealing, etc.) and will briefly discuss some recent advances in areas such as discrete optimization and adaptive (second-order) search with or without stochastic gradients.

James C. Spall *is a member of the Principal Professional Staff at the Johns Hopkins Applied Physics Laboratory, a Research Professor in the JHU Department of Applied Mathematics and Statistics, and the Chairman of the Applied and Computational Mathematics Program within the JHU Engineering Programs for Professionals. Dr. Spall has published extensively in the areas of control systems and statistics and holds two U.S. patents for inventions in control systems, both licensed to U.S. companies. He is the editor and coauthor of the book Bayesian Analysis of Time Series and Dynamic Models (CRC Press) and the author of Introduction to Stochastic Search and Optimization (Wiley). Dr. Spall is one of the inaugural Senior Editors for the IEEE Transactions on Automatic Control and is a Contributing Editor for the Current Index to Statistics. He was the Program Chair for the 2007 IEEE Conference on Decision and Control and is a Fellow of IEEE.*

Tutorial Session 3, 3:30 - 5:30pm

Consciousness and Information Theory

Giulio Tononi, University of Wisconsin
Christof Koch, Allen Institute for Brain Science

Discovering the material basis of subjective experience, the heart of the ancient mind-body problem, is a quest pursued by many clinical and neuroscience laboratories. Yet discovering the neuronal correlates of consciousness leaves the question of the exact relationship between excitable (brain) matter and consciousness open. It has frequently been surmised that information theory can link the objective world of physics to the subjective world of our everyday experiences. In this tutorial, we introduce the audience to the modern study of consciousness and then focus on the integrated information theory (IIT). IIT stems from Gedanken experiments that lead to phenomenological axioms and ontological postulates and provides a quantitative framework from the perspective of information theory. This framework can be used to compute the complexity of any system of causally interacting parts, such as brains, computers or the internet. Many observations concerning the neural substrate of consciousness fall naturally into place within the IIT framework. Among them are the association of consciousness with certain neural systems rather than with others; the fact that neural processes underlying consciousness can influence or be influenced by neural processes that remain unconscious; the reduction of consciousness during dreamless sleep and generalized epileptic seizures; and the distinct role of different cortical architectures in affecting the quality of experience. The theory has significant implications for our view of nature.

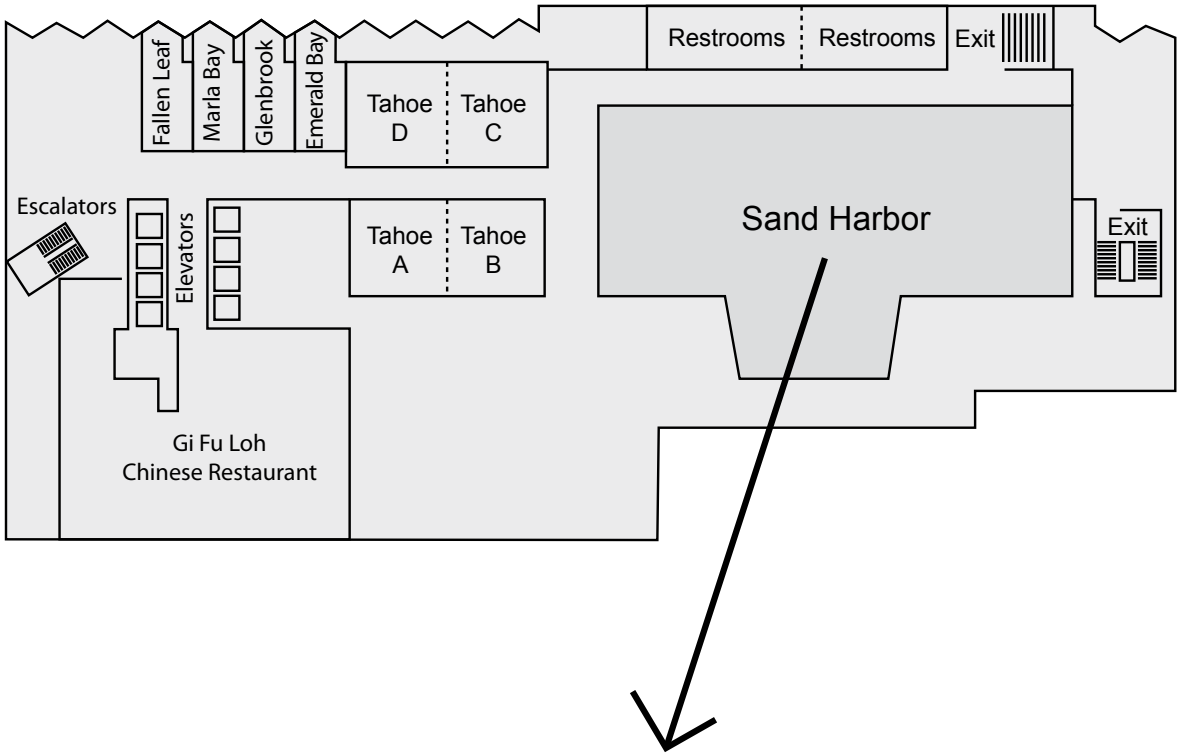
Giulio Tononi *is a psychiatrist and neuroscientist who has held faculty positions in Pisa, New York, San Diego and Madison, Wisconsin. The main focus of his work has been the scientific understanding of consciousness. His integrated information theory is a comprehensive theory of what consciousness is, how it can be measured, and how it is realized in the brain. The theory is being tested with neuroimaging, transcranial magnetic stimulation, and computer models. The other main focus of his work is to understand the function of sleep. He and collaborators study species ranging from fruit flies to humans, from the molecular and cellular level to the systems level. This research has led to the synaptic homeostasis hypothesis, according to which sleep is needed to renormalize synapses, counteracting the progressive increase in synaptic strength that occurs during wakefulness due to learning. The hypothesis has implications for understanding the effects of sleep deprivation and for developing diagnostic and therapeutic approaches to sleep disorders and neuropsychiatric disorders.*

Born in the American Midwest, ***Christof Koch*** *grew up in Holland, Germany, Canada, and Morocco, where he graduated from the Lycèe Descartes. He studied Physics and Philosophy at the University of Tübingen in Germany and was awarded his Ph.D. in Biophysics in 1982. After four years at MIT, Dr. Koch joined Caltech in 1986, where he is the Lois and Victor Troendle Professor of Cognitive and Behavioral Biology. In 2011, he became the CSO of the Allen Institute for Brain Science in Seattle to lead a large scale, focused and high-throughout, ten year effort to understand coding in the visual neocortex. The author of more than three hundred scientific papers and journal articles, patents and books, Dr. Koch studies the biophysics of computation, and the neuronal basis of visual perception, attention, and consciousness. Together with Francis Crick, with whom he worked for 16 years, he is one of the pioneers of the neurobiological approach to consciousness. His latest book is Consciousness: Confessions of a Romantic Reductionist (MIT Press, 2012). He loves dogs, Apple Computers, rock-climbing, trailing running in the mountains and biking.*

MONDAY POSTER FLOORPLAN

HARRAH'S

2ND FLOOR SPECIAL EVENTS CENTER



SAND HARBOR

89	88	81	80	73	72	65	64	57	56	49
90	87	82	79	74	71	66	63	58	55	50
91	86	83	78	75	70	67	62	59	54	51
93	92	85	84	77	76	69	68	61	60	53	52

48	41	40	33	32	25	24	17	16	09	08	01
47	42	39	34	31	26	23	18	15	10	07	02
46	43	38	35	30	27	22	19	14	11	06	03
45	44	37	36	29	28	21	20	13	12	05	04

MONDAY, DECEMBER 4TH

6:30 – 6:40PM - OPENING REMARKS, AWARDS & RECEPTION
Harrah's Special Events Center 2nd Floor

POSTER SESSION

AND RECEPTION - 7:00 – 11:59 PM

M1 Multiresolution analysis on the symmetric group
R. Kondor, W. Dempsey

M2 A Simple and Practical Algorithm for Differentially Private Data Release
F. McSherry, K. Ligett, M. Hardt

M3 Assessing Blinding in Clinical Trials
O. Arandjelovic

M4 Deep Neural Networks Segment Neuronal Membranes in Electron Microscopy Images
D. Ciresan, A. Giusti, I. Gambardella, J. Schmidhuber

M5 Waveform Driven Plasticity in BiFeO3 Memristive Devices: Model and Implementation
C. Mayr, P. Stärke, J. Partzsch, L. Cederstroem, R. Schüffny, Y. Shuai, N. DU, H. Schmidt

M6 Analog readout for optical reservoir computers
A. Smerieri, F. Duport, Y. Paquot, M. Haelterman, S. Massar

M7 Multi-scale Hyper-time Hardware Emulation of Human Motor Nervous System Based on Spiking Neurons using FPGA
C. Niu, S. Nandyala, W. Sohn, T. Sanger

M8 Learned Prioritization for Trading Off Accuracy and Speed
J. Jiang, A. Teichert, H. Daume III, J. Eisner

M9 Algorithms for Learning Markov Field Policies
A. Boularias, O. Kroemer, J. Peters

M10 Symbolic Dynamic Programming for Continuous State and Observation POMDPs
Z. Zamani, S. Sanner, P. Poupart, K. Kersting

M11 Tractable Objectives for Robust Policy Optimization
K. Chen, M. Bowling

M12 Efficient Reinforcement Learning for High Dimensional Linear Quadratic Systems
M. Ibrahimi, A. Javanmard, B. Van Roy

M13 One Permutation Hashing
P. Li, A. Owen, C. Zhang

M14 Isotropic Hashing
W. Kong, W. Li

M15 Generalization Bounds for Domain Adaptation
C. Zhang, J. Ye, L. Zhang

M16 The representer theorem for Hilbert spaces: a necessary and sufficient condition
F. Dinuzzo, B. Schölkopf

M17 Scalable nonconvex inexact proximal splitting
S. Sra

M18 Ordered Rules for Classification: A Discrete Optimization Approach to Associative Classification
A. Chang, D. Bertsimas, C. Rudin

M19 Submodular Bregman Divergences with Applications
R. Iyer, J. Bilmes

M20 High-dimensional Nonparanormal Graph Estimation via Smooth-projected Neighborhood Pursuit
T. Zhao, K. Roeder, H. Liu

M21 Sparse Approximate Manifolds for Differential Geometric MCMC
B. Calderhead, M. Sustik

M22 Persistent Homology for Learning Densities with Bounded Support
F. Pokorny, C. Ek, H. Kjellström, D. Kragic

M23 Meta-Gaussian Information Bottleneck
M. Rey, V. Roth

M24 The Coloured Noise Expansion and Parameter Estimation of Diffusion Processes
S. Lyons, A. Storkey, S. Sarkka

M25 Nonconvex Penalization, Levy Processes and Concave Conjugates
Z. Zhang, B. Tu

M26 Scalable imputation of genetic data with a discrete fragmentation-coagulation process
L. Elliott, Y. Teh

M27 The Time-Marginalized Coalescent Prior for Hierarchical Clustering
L. Boyles, M. Welling

M28 A Nonparametric Conjugate Prior Distribution for the Maximizing Argument of a Noisy Function
P. Ortega, T. Genewein, J. Grau-Moya, D. Balduzzi, D. Braun

M29 Slice sampling normalized kernel-weighted completely random measure mixture models
N. Foti, S. Williamson

M30 MAP Inference in Chains using Column Generation
D. Belanger, A. Passos, S. Riedel, A. McCallum

M31 Learning curves for multi-task Gaussian process regression
P. Sollich, S. Ashton

M32 Practical Bayesian Optimization of Machine Learning Algorithms
J. Snoek, H. Larochelle, R. Adams

M33 Iterative Thresholding Algorithm for Sparse Inverse Covariance Estimation
B. Rolfs, B. Rajaratnam, D. Guillot, A. Maleki, I. Wong

M34 A Divide-and-Conquer Method for Sparse Inverse Covariance Estimation
C. Hsieh, I. Dhillon, P. Ravikumar, A. Banerjee

M35 Approximate Message Passing with Consistent Parameter Estimation and Applications to Sparse Learning
U. Kamilov, S. Rangan, A. Fletcher, M. Unser

M36 The Bethe Partition Function of Log-supermodular Graphical Models
N. Ruozzi

M37 Convergence Rate Analysis of MAP Coordinate Minimization Algorithms
O. Meshi, T. Jaakkola, A. Globerson

M38 Bayesian Probabilistic Co-Subspace Addition
L. Shi

M39 Learning the Architecture of Sum-Product Networks Using Clustering on Variables
A. Dennis, D. Ventura

M40 Efficient Sampling for Bipartite Matching Problems
M. Volkovs, R. Zemel

M41 Projection Retrieval for Classification
M. Fiterau, A. Dubrawski

M42 Learning Multiple Tasks using Shared Hypotheses
K. Crammer, Y. Mansour

M43 Optimal kernel choice for large-scale two-sample tests
A. Gretton, B. Sriperumbudur, D. Sejdinovic, H. Strathmann, S. Balakrishnan, M. Pontil, K. Fukumizu

M44 Supervised Learning with Similarity Functions
P. Kar, P. Jain

M45 Density-Difference Estimation
M. Sugiyama, T. Kanamori, T. Suzuki, M. Plessis, S. Liu, I. Takeuchi

M46 The Lovasz θ function, SVMs and finding large dense subgraphs
V. Jethava, A. Martinsson, C. Bhattacharyya, D. Dubhashi

M47 Perceptron Learning of SAT
A. Flint, M. Blaschko

M48 A Polynomial-time Form of Robust Regression
Y. Yu, O. Aslan, D. Schuurmans

M49 Feature Clustering for Accelerating Parallel Coordinate Descent
C. Scherrer, A. Tewari, M. Halappanavar, D. Haglin

M50 Mixing Properties of Conditional Markov Chains with Unbounded Feature Functions
M. Sinn, B. Chen

M51 Efficient Monte Carlo Counterfactual Regret Minimization in Games with Many Player Actions
R. Gibson, M. Lanctot, N. Burch, D. Szafron

M52 Tight Bounds on Redundancy and Distinguishability of Label-Invariant Distributions
J. Acharya, H. Das, A. Orlitsky

M53 Exponential Concentration for Mutual Information Estimation with Application to Forests
H. Liu, J. Lafferty, L. Wasserman

M54 Bayesian estimation of discrete entropy with mixtures of stick-breaking priors
E. Archer, J. Pillow, I. Park

M55 Learning Halfspaces with the Zero-One Loss: Time-Accuracy Tradeoffs
A. Birnbaum, S. Shalev-Shwartz

M56 A Scalable CUR Matrix Decomposition Algorithm: Lower Time Complexity and Tighter Bound
S. Wang, Z. Zhang

M57 Online L1-Dictionary Learning with Application to Novel Document Detection
S. Kasiviswanathan, H. Wang, A. Banerjee, P. Melville

M58 Ensemble weighted kernel estimators for multivariate entropy estimation
K. Sricharan, A. Hero

M59 Dip-means: an incremental clustering method for estimating the number of clusters
A. Kalogeratos, A. Likas

M60 Convergence and Energy Landscape for Cheeger Cut Clustering
X. Bresson, T. Laurent, D. Uminsky, J. von Brecht

M61 Cardinality Restricted Boltzmann Machines
K. Swersky, D. Tarlow, I. Sutskever, R. Zemel, R. Salakhutdinov, R. Adams

M62 Controlled Recognition Bounds for Visual Learning and Exploration
V. Karasev, A. Chiuso, S. Soatto

M63 Clustering Aggregation as Maximum-Weight Independent Set
N. Li, L. Latecki

M64 Clustering by Nonnegative Matrix Factorization Using Graph Random Walk
Z. Yang, T. Hao, O. Dikmen, X. Chen, E. Oja

M65 Angular Quantization based Binary Codes for Fast Similarity Search
Y. Gong, S. Kumar, V. Verma, S. Lazebnik

M66 Near-optimal Differentially Private Principal Components
K. Chaudhuri, A. Sarwate, K. Sinha

M67 On the Sample Complexity of Robust PCA
M. Coudron, G. Lerman

M68 Learning the Dependency Structure of Latent Factors
Y. He, Y. Qi, k. kavukcuoglu, H. Park

M69 Bayesian Nonparametric Maximum Margin Matrix Factorization for Collaborative Prediction
M. Xu, J. Zhu, B. Zhang

M70 Identifiability and Unmixing of Latent Parse Trees
P. Liang, S. Kakade, D. Hsu

M71 How They Vote: Issue-Adjusted Models of Legislative Behavior
S. Gerrish, D. Blei

M72 3D Gaze Concurrences from Head-mounted Cameras
H. Park, e. Jain, Y. Sheikh

M73 Compressive Sensing MRI with Wavelet Tree Sparsity
C. Chen, J. Huang

M74 Recognizing Activities by Attribute Dynamics
W. Li, N. Vasconcelos

M75 Fusion with Diffusion for Robust Visual Tracking
Y. Zhou, X. Bai, W. Liu, L. Latecki

M76 Learning visual motion in recurrent neural networks
M. Pachitariu, M. Sahani

M77 Burn-in, bias, and the rationality of anchoring
F. Lieder, T. Griffiths, N. Goodman

M78 On the connections between saliency and tracking
V. Mahadevan, N. Vasconcelos

M79 Action-Model Based Multi-agent Plan Recognition
H. Zhuo, Q. Yang, S. Kambhampati

M80 Rational inference of relative preferences
N. Srivastava, P. Schrater

M81 Why MCA? Nonlinear Spike-and-slab Sparse Coding for Neurally Plausible Image Encoding
J. Shelton, P. Sterne, J. Bornschein, A. Sheikh, J. Lucke

M82 A System for Predicting Action Content On-Line and in Real Time before Action Onset in Humans – an Intracranial Study
U. Maoz, S. Ye, I. Ross, A. Mamelak, C. Koch

M83 A lattice filter model of the visual pathway
K. Gregor, D. Chklovskii

M84 Q-MKL: Matrix-induced Regularization in Multi-Kernel Learning with Applications to Neuroimaging
C. Hinrichs, V. Singh, J. Peng, S. Johnson

M85 Wavelet based multi-scale shape features on arbitrary surfaces for cortical thickness discrimination
W. Kim, D. Pachauri, C. Hatt, M. Chung, S. Johnson, V. Singh

M86 A P300 BCI for the Masses: Prior Information Enables Instant Unsupervised Spelling
P. Kindermans, H. Verschore, D. Verstraeten, B. Schrauwen

M87 Towards a learning-theoretic analysis of spike-timing dependent plasticity
D. Balduzzi, M. Besserve

M88 Synchronization can Control Regularization in Neural Systems via Correlated Noise Processes
J. Bouvrie, J. Slotine

M89 Homeostatic plasticity in Bayesian spiking networks as Expectation Maximization with posterior constraints
S. Habenschuss, J. Bill, B. Nessler

M90 Neurally Plausible Reinforcement Learning of Working Memory Tasks
J. Rombouts, S. Bohte, P. Roelfsema

M91 Spiking and saturating dendrites differentially expand single neuron computation capacity.
R. Cazé, M. Humphries, B. Gutkin

M92 Coding efficiency and detectability of rate fluctuations with non-Poisson neuronal firing
S. Koyama

M93 Efficient coding connects prior and likelihood function in perceptual Bayesian inference
X. Wei, A. Stocker

M1 Multiresolution Analysis on the Symmetric Group

Risi Kondor risi@uchicago.edu
Walter Dempsey wdempsey@uchicago.edu
University of Chicago

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

M2 A Simple and Practical Algorithm for Differentially Private Data Release

Frank McSherry mcsherry@microsoft.com
Silicon Valley, Microsoft Research
Katrina Ligett katrina@caltech.edu
Caltech
Moritz Hardt mhardt@us.ibm.com
IBM Almaden Research

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

M3 Assessing Blinding in Clinical Trials

Ognjen Arandjelovic ognjen.arandjelovic@gmail.com
Deakin University

The interaction between the patient's expected outcome of an intervention and the inherent effects of that intervention can have extraordinary effects. Thus in clinical trials an effort is made to conceal the nature of the administered intervention from the participants in the trial i.e. to blind it. Yet, in practice perfect blinding is impossible to ensure or even verify. The current standard is follow up the trial with an auxiliary questionnaire, which allows trial participants to express their belief concerning the assigned intervention and which is used to compute a measure of the extent of blinding in the trial. If the estimated extent of blinding exceeds a threshold the trial is deemed sufficiently blinded; otherwise, the trial is deemed to have failed. In this paper we make several important contributions. Firstly, we identify a series of fundamental problems of the aforesaid practice and discuss them in context of the most commonly used blinding measures. Secondly, motivated by the highlighted problems, we formulate a novel method for handling imperfectly blinded trials. We too adopt a post-trial feedback questionnaire but interpret the collected data using an original approach, fundamentally different from those previously proposed. Unlike previous approaches, ours is void of any ad hoc free parameters, is robust to small changes in auxiliary data and is not predicated on any strong assumptions used to interpret participants' feedback.

M4 Deep Neural Networks Segment Neuronal Membranes in Electron Microscopy Images

Dan Ciresan dan@idsia.ch
Alessandro Giusti alessandrog@idsia.ch
luca Maria Gambardella luca@idsia.ch
Juergen Schmidhuber juergen@idsia.ch
IDSIA

We address a central problem of neuroanatomy, namely, the automatic segmentation of neuronal structures depicted in stacks of electron microscopy (EM) images. This is necessary to efficiently map 3D brain structure and connectivity. To segment {\em biological} neuron membranes, we use a special type of deep {\em artificial} neural network as a pixel classifier. The label of each pixel (membrane or non-membrane) is predicted from raw pixel values in a square window centered on it. The input layer maps each window pixel to a neuron. It is followed by a succession of convolutional and max-pooling layers which preserve 2D information and extract features with increasing levels of abstraction. The output layer produces a calibrated probability for each class. The classifier is trained by plain gradient descent on a 512×512×30 stack with known ground truth, and tested on a stack of the same size (ground truth unknown to the authors) by the organizers of the ISBI 2012 EM Segmentation Challenge. Even without problem-specific post-processing, our approach outperforms competing techniques by a large margin in all three considered metrics, i.e. \emph{rand error}, \emph{warping error} and \emph{pixel error}. For pixel error, our approach is the only one outperforming a second human observer.

M5 Waveform Driven Plasticity in BiFeO3 Memristive Devices: Model and Implementation

Christian Mayr christian.mayr@tu-dresden.de
Paul Stärke paul.staerke@mailbox.tu-dresden.de
Johannes Partzsch johannes.partzsch@tu-dresden.de
Rene Schüffny rene.schueffny@tu-dresden.de
TU Dresden
Love Cederstroem love.cederstroem@zmdi.com
ZMD AG
Yao Shuai y.shuai@hzdr.de
Helmholtz-Zentrum Dresden-Rossendorf e.
V.NAN DU dunan@gatech.edu
Georgia Institute of Technology
Heidemarie Schmidt
Heidemarie.Schmidt@etit.tu-chemnitz.de
TU Chemnitz

Memristive devices have recently been proposed as efficient implementations of plastic synapses in neuromorphic systems. The plasticity in these memristive devices, i.e. their resistance change, is defined by

the applied waveforms. This behavior resembles biological synapses, whose plasticity is also triggered by mechanisms that are determined by local waveforms. However, learning in memristive devices has so far been approached mostly on a pragmatic technological level. The focus seems to be on finding any waveform that achieves spike-timing-dependent plasticity (STDP), without regard to the biological veracity of said waveforms or to further important forms of plasticity. Bridging this gap, we make use of a plasticity model driven by neuron waveforms that explains a large number of experimental observations and adapt it to the characteristics of the recently introduced BiFeO3 memristive material. Based on this approach, we show STDP for the first time for this material, with learning window replication superior to previous memristor-based STDP implementations. We also demonstrate in measurements that it is possible to overlay short and long term plasticity at a memristive device in the form of the well-known triplet plasticity. To the best of our knowledge, this is the first implementations of triplet plasticity on any physical memristive device.

M6 Analog readout for optical reservoir computers

Anteo Smerieri anteo.smerieri@gmail.com
Yvan Paquot ypaquot@ulb.ac.be
François Duport Francois.DUPORT@ulb.ac.be
Marc Haelterman mhaelter@ulb.ac.be
Serge Massar smassar@ulb.ac.be
Université libre de Bruxelles

Reservoir computing is a new, powerful and flexible machine learning technique that is easily implemented in hardware. Recently, by using a time-multiplexed architecture, hardware reservoir computers have reached performance comparable to digital implementations. Operating speeds allowing for real time information operation have been reached using optoelectronic systems. At present the main performance bottleneck is the readout layer which uses slow, digital postprocessing. We have designed an analog readout suitable for time-multiplexed optoelectronic reservoir computers, capable of working in real time. The readout has been built and tested experimentally on a standard benchmark task. Its performance is better than non-reservoir methods, with ample room for further improvement. The present work thereby overcomes one of the major limitations for the future development of hardware reservoir computers.

M7 Multi-scale Hyper-time Hardware Emulation of Human Motor Nervous System Based on Spiking Neurons using FPGA

C. Minos Niu minos.niu@sangerlab.net
Sirish Nandyala nandyala@usc.edu
Won Joon Sohn wonjsohn@gmail.com
Terence Sanger terry@sangerlab.net
BME, USC

Our central goal is to quantify the long-term progression of pediatric neurological diseases, such as a typical 10-15 years progression of child dystonia. To this purpose, quantitative models are convincing only if they can provide multi-scale details ranging from neuron spikes to limb biomechanics. The models also need to be evaluated in hyper-time, i.e. significantly faster than real-time, for producing useful predictions. We designed a platform with digital VLSI hardware for multi-scale hyper-time emulations of human motor nervous systems. The platform is constructed on a scalable, distributed array of Field Programmable Gate Array (FPGA) devices. All devices operate asynchronously with 1 millisecond time granularity, and the overall system is accelerated to 365x real-time. Each physiological component is implemented using models from well documented studies and can be flexibly modified. Thus the validity of emulation can be easily advised by neurophysiologists and clinicians. For maximizing the speed of emulation, all calculations are implemented in combinational logic instead of clocked iterative circuits. This paper presents the methodology of building FPGA modules in correspondence to components of a monosynaptic spinal loop. Results of emulated activities are shown. The paper also discusses the rationale of approximating neural circuitry by organizing neurons with sparse interconnections. In conclusion, our platform allows introducing various abnormalities into the neural emulation such that the emerging motor symptoms can be analyzed. It compels us to test the origins of childhood motor disorders and predict their long-term progressions.

M8 Learned Prioritization for Trading Off Accuracy and Speed

Jiarong Jiang jiarong@umiacs.umd.edu
Hal Daume III me@hal3.name
University of Maryland
Adam Teichert teichert@jhu.edu
Jason Eisner jason@cs.jhu.edu
Computer Science, Johns Hopkins University

Users want natural language processing (NLP) systems to be both fast and accurate, but quality often comes at the cost of speed. The field has been manually exploring various speed-accuracy tradeoffs (for particular problems and datasets). We aim to explore this space automatically, focusing here on the case of agenda-based syntactic parsing \cite{kay-1986}. Unfortunately, off-the-shelf reinforcement learning techniques fail to learn good policies: the state space is simply too large to explore naively. An attempt to counteract this by applying imitation learning algorithms also fails: the ``teacher'' is far too good to successfully imitate with our inexpensive features. Moreover, it is not specifically tuned for the known reward function. We propose a hybrid reinforcement/ apprenticeship learning algorithm that, even with only a few inexpensive features, can automatically learn weights that achieve competitive accuracies at significant improvements in speed over state-of-the-art baselines.

M9 Algorithms for Learning Markov Field Policies

Abdeslam Boularias boularias@tuebingen.mpg.de
Schölkopf, Max Planck Institute for Intelligent Systems
Oliver Kroemer oliverkro@googlemail.com
Jan Peters mail@jan-peters.net
Technische Universitaet Darmstadt

We present a new graph-based approach for incorporating domain knowledge in reinforcement learning applications. The domain knowledge is given as a weighted graph, or a kernel matrix, that loosely indicates which states should have similar optimal actions. We first introduce a bias into the policy search process by deriving a distribution on policies such that policies that disagree with the provided graph have low probabilities. This distribution corresponds to a Markov Random Field. We then present a reinforcement and an apprenticeship learning algorithms for finding such policy distributions. We also illustrate the advantage of the proposed approach on three problems: swing-up cart-balancing with nonuniform and smooth frictions, gridworlds, and teaching a robot to grasp new objects.

M10 Symbolic Dynamic Programming for Continuous State and Observation POMDPs

Zahra Zamani zahra.zamani@anu.edu.au
ANU and NICTA
Scott Sanner ssanner@gmail.com
Statistical Machine Learning, Nicta
Pascal Poupart ppoupart@cs.uwaterloo.ca
Computer Science, University of Waterloo
Kristian Kersting kristian.kersting@iais.fraunhofer.de
University of Bonn and Fraunhofer IAIS

Partially-observable Markov decision processes (POMDPs) provide a powerful model for real-world sequential decision-making problems. In recent years, point- based value iteration methods have proven to be extremely effective techniques for finding (approximately) optimal dynamic programming solutions to POMDPs when an initial set of belief states is known. However, no point-based work has provided exact point-based backups for both continuous state and observation spaces, which we tackle in this paper. Our key insight is that while there may be an infinite number of possible observations, there are only a finite number of observation partitionings that are relevant for optimal decision-making when a finite, fixed set of reachable belief states is known. To this end, we make two important contributions: (1) we show how previous exact symbolic dynamic pro- gramming solutions for continuous state MDPs can be generalized to continuous state POMDPs with discrete observations, and (2) we show how this solution can be further extended via recently developed symbolic methods to continuous state and observations to derive the minimal relevant observation partitioning for potentially correlated, multivariate observation spaces. We demonstrate proof-of- concept results on uni- and multi-variate state and observation steam plant control.

M11 Tractable Objectives for Robust Policy Optimization

Katherine Chen kchen4@cs.ualberta.ca
Michael Bowling bowling@cs.ualberta.ca
University of Alberta

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

M12 Efficient Reinforcement Learning for High Dimensional Linear Quadratic Systems

Morteza Ibrahimi ibrahimi@stanford.edu
Adel Javanmard adelj@stanford.edu
Benjamin Van Roy bvr@stanford.edu
Stanford University

We study the problem of adaptive control of a high dimensional linear quadratic (LQ) system. Previous work established the asymptotic convergence to an optimal controller for various adaptive control schemes. More recently, an asymptotic regret bound of O˜(T) was shown for T≫p where p is the dimension of the state space. In this work we consider the case where the matrices describing the dynamic of the LQ system are sparse and their dimensions are large. We present an adaptive control scheme that for p≫1 and T≫\polylog(p) achieves a regret bound of O˜(pT). In particular, our algorithm has an average cost of (1+\eps) times the optimum cost after T=\polylog(p)O(1/\eps2). This is in comparison to previous work on the dense dynamics where the algorithm needs Ω(p) samples before it can estimate the unknown dynamic with any significant accuracy. We believe our result has prominent applications in the emerging area of computational advertising, in particular targeted online advertising and advertising in social networks.

M13 One Permutation Hashing

Ping Li pingli@cornell.edu
Cornell
Art Owen owen@stanford.edu
Stanford University
Cun-Hui Zhang czhang@stat.rutgers.edu
Rutgers University

While minwise hashing is promising for large-scale learning in massive binary data, the preprocessing cost is prohibitive as it requires applying (e.g.,) k=500 permutations on the data. The testing time is also expensive if a new data point (e.g., a new document or a new image) has not been processed. In this paper, we develop a simple \textbf{one permutation hashing} scheme to address this important issue. While it is true that the preprocessing step can be parallelized, it comes at the cost of additional hardware and implementation. Also, reducing k permutations to just one would be much more \textbf{energy-efficient}, which might be an important perspective as minwise hashing is commonly deployed in the search industry. While the theoretical probability analysis is interesting, our experiments on similarity estimation and SVM \& logistic regression also confirm the theoretical results.

M14 Isotropic Hashing

Weihao Kong kongweihao@cs.sjtu.edu.cn
Wu-Jun Li liwujun@cs.sjtu.edu.cn
Shanghai Jiao Tong University

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

M15 Generalization Bounds for Domain Adaptation

Chao Zhang zhangchao1015@gmail.com
Jieping Ye jieping.ye@asu.edu
Arizona State University
Lei Zhang zhanglei.njust@yahoo.com.cn
Nanjing University of Science and Technology

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

M16 The representer theorem for Hilbert spaces: a necessary and sufficient condition

Francesco Dinuzzo francesco.dinuzzo@gmail.com
Bernhard Schölkopf bs@tuebingen.mpg.de
Max Planck Institute for Intelligent Systems

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

M17 Scalable nonconvex inexact proximal splitting

Suvrit Sra suvrit@gmail.com
AGBS, Max Planck Institute for Intelligent Systems

We study large-scale, nonsmooth, nonconconvex optimization problems. In particular, we focus on nonconvex problems with \emph{composite} objectives. This class of problems includes the extensively studied convex, composite objective problems as a special case. To tackle composite nonconvex problems, we introduce a powerful new framework based on asymptotically \ emph{nonvanishing} errors, avoiding the common convenient assumption of eventually vanishing errors. Within our framework we derive both batch and incremental nonconvex proximal splitting algorithms. To our knowledge, our framework is first to develop and analyze incremental \emph{nonconvex} proximal-splitting algorithms, even if we disregard the ability to handle nonvanishing errors. We illustrate our theoretical framework by showing how it applies to difficult large-scale, nonsmooth, and nonconvex problems.

M18 Ordered Rules for Classification: A Discrete Optimization Approach to Associative Classification

Allison Chang aachang@mit.edu
Dimitris Bertsimas dbertsim@mit.edu
Massachusetts Institute of Technology
Cynthia Rudin crudin@princeton.edu
Applied Math, Princeton University

We aim to design classifiers that have the interpretability of association rules yet have predictive power on par with the top machine learning algorithms for classification. We propose a novel mixed integer optimization (MIO) approach called Ordered Rules for Classification (ORC) for this task. Our method has two parts. The first part mines a particular frontier of solutions in the space of rules, and we show that this frontier contains the best rules according to a variety of interestingness measures. The second part learns an optimal ranking for the rules to build a decision list classifier that is simple and insightful. We report empirical evidence using several different datasets to demonstrate the performance of this method.

M19 Submodular Bregman Divergences with Applications

Rishabh Iyer rkiyer@u.washington.edu
Jeff Bilmes bilmes@ee.washington.edu
University of Washington

We introduce a class of discrete divergences on sets (equivalently binary vectors) that we call the submodular Bregman divergences. We consider two kinds, defined either from tight modular upper or tight modular lower bounds of a submodular function. We show that the properties of these divergences are analogous to the (standard continuous) Bregman divergence. Further, we demonstrate how they generalize many useful divergences, including the weighted Hamming distance, squared weighted Hamming, weighted precision, recall, conditional mutual information, and a generalized KL-divergence on sets. We also show that the lower bound submodular Bregman is actually a special case of the generalized Bregman divergence on the \lovasz{} extension of a submodular function which we call the \ lovasz{} Bregman divergence. We then point out a number of applications of the submodular Bregman divergences, and in particular show that a proximal algorithm defined through the submodular Bregman divergences provides a framework for many mirror-descent style algorithms related to submodular function optimization. We also show that a generalization of the k-means algorithm using the \lovasz{} Bregman divergence is natural in clustering scenarios where the ordering is important. A unique property of this algorithm is that computing the mean ordering is extremely efficient unlike the other order based distance measures. \extendedv{Finally we provide a clustering framework for the submodular Bregman, and we derive fast algorithms for clustering sets of binary vectors (equivalently sets of sets).

M20 High-dimensional Nonparanormal Graph Estimation via Smooth-projected Neighborhood Pursuit

Tuo Zhao tzhao5@jhu.edu
Computer Sceince, Johns Hopkins University
Kathryn Roeder roeder@stat.cmu.edu
Carnegie Mellon University
Han Liu hanliu@princeton.edu
Princeton University

We propose a new smooth-projected neighborhood pursuit method for estimating high dimensional undirected graphs. Our method can be viewed as a semiparametric extension of the popular neighborhood pursuit approach proposed by N. Meinshausen and P. B{ü}hlmann 2006 from Gaussian to Gaussian copula models (or the nonparanormal models as proposed by Liu et. al 2009). In terms of methodology and computation, we project a possibly indefinite symmetric matrix into the cone of positive semidefinite matrices. The projection is formulated as a smoothed element-wise ℓ_∞-norm minimization problem. We develop an efficient fast proximal gradient algorithm with a provable optimal rate of convergence \cO(1/ϵ), where ϵ is the desired accuracy for the objective value. In terms of theory, we provide an alternative view to analyze the trade-off between computational efficiency and statistical error. We give a sufficient condition to secure that the smooth-projected neighborhood pursuit estimator achieves graph estimation consistency. Empirically, we conduct real data experiments on stock and genomic datasets to illustrate the usefulness of the proposed method.

M21 Sparse Approximate Manifolds for Differential Geometric MCMC

Ben Calderhead b.calderhead@ucl.ac.uk
CoMPLEX, University College London
Matyas Sustik sustik@cs.utexas.edu
University of Texas at Austin

One of the enduring challenges in Markov chain Monte Carlo methodology is the development of proposal mechanisms to make moves distant from the current point, that are accepted with high probability and at low computational cost. The recent introduction of locally adaptive MCMC methods based on the natural underlying Riemannian geometry of such models goes some way to alleviating these problems for certain classes of models for which the metric tensor is analytically tractable, however computational efficiency is not assured due to the necessity of potentially high-dimensional matrix operations at each iteration. In this paper we firstly investigate a sampling-based approach for approximating the metric tensor and suggest a valid MCMC algorithm that extends the applicability of Riemannian Manifold MCMC methods to statistical models that do not admit an analytically computable metric tensor. Secondly, we show how the approximation scheme we consider naturally motivates the use of l1 regularisation to improve estimates and obtain a sparse approximate inverse of the metric, which enables stable and sparse approximations of the local geometry to be made. We demonstrate the application of this algorithm for inferring the parameters of a realistic system of ordinary differential equations using a biologically motivated robust student-t error model, for which the expected Fisher Information is analytically intractable.

M22 Persistent Homology for Learning Densities with Bounded Support

Florian Pokorny fpokorny@csc.kth.se
Carl Henrik Ek chek@csc.kth.se
Hedvig Kjellström hedvig@csc.kth.se
Danica Kragic danik@csc.kth.se
Royal Institute of Technology

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

M23 Meta-Gaussian Information Bottleneck

Melanie Rey melanierey1@gmail.com
Volker Roth volker.roth@unibas.ch
University of Basel

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

M24 The Coloured Noise Expansion and Parameter Estimation of Diffusion Processes

Simon Lyons simonlyons@gmail.com
Amos Storkey a.storkey@ed.ac.uk
School of Informatics, University of Edinburgh
Simo Sarkka simo.sarkka@aalto.fi
Aalto University

Stochastic differential equations (SDE) are a natural tool for modelling systems that are inherently noisy or contain uncertainties that can be modelled as stochastic processes. Crucial to the process of using SDE to build mathematical models is the ability to estimate parameters of those models from observed data. Over the past few decades, significant progress has been made on this problem, but we are still far from having a definitive solution. We describe a novel method of approximating a diffusion process that we show to be useful in Markov chain Monte-Carlo (MCMC) inference algorithms. We take the 'white' noise that drives a diffusion process and decompose it into two terms. The first is a 'coloured noise' term that can be deterministically controlled by a set of auxilliary variables. The second term is small and enables us to form a linear Gaussian 'small noise' approximation. The decomposition allows us to take a diffusion process of interest and cast it in a form that is amenable to sampling by MCMC methods. We explain why many state-of-the-art inference methods fail on highly nonlinear inference problems. We demonstrate experimentally that our method performs well in such situations. Our results show that this method is a promising new tool for use in inference and parameter estimation problems.

M25 Nonconvex Penalization, Levy Processes and Concave Conjugates

Zhihua Zhang zhzhang@gmail.com
Bojun Tu tubojun@gmail.com
Zhejiang University

In this paper we study sparsity-inducing nonconvex penalty functions using Levy processes. We define such a penalty as the Laplace exponent of a subordinator. Accordingly, we propose a novel approach for the construction of sparsity-inducing nonconvex penalties. Particularly, we show that the nonconvex logarithmic (LOG) and exponential (EXP) penalty functions are the Laplace exponents of Gamma and compound Poisson subordinators, respectively. Additionaly, we explore the concave conjugate of nonconvex penalties. We find that the LOG and EXP penalties are the concave conjugates of the negatives of Kullback-Leiber (KL) distance functions. Furthermore, the relationship between these two penalties is due to asymmetricity of the KL distance.

M26 Scalable imputation of genetic data with a discrete fragmentation-coagulation process

Lloyd Elliott elliott@gatsby.ucl.ac.uk
University College London
Yee Whye Teh teh@stats.ox.ac.uk
University of Oxford

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

M27 The Time-Marginalized Coalescent Prior for Hierarchical Clustering

Levi Boyles lboyles@uci.edu
UC Irvine
Max Welling welling.max@gmail.com
University of Amsterdam

We introduce a new prior for use in Nonparametric Bayesian Hierarchical Clustering. The prior is constructed by marginalizing out the time information of Kingman's coalescent, providing a prior over tree structures which we call the Time-Marginalized Coalescent (TMC). This allows for models which factorize the tree structure and times, providing two benefits: more flexible priors may be constructed and more efficient Gibbs type inference can be used. We demonstrate this on an example model for density estimation and show the TMC achieves competitive experimental results.

M28 A Nonparametric Conjugate Prior Distribution for the Maximizing Argument of a Noisy Function

Pedro Ortega pedro.ortega@gmail.com
Jordi Grau-Moya jordi.grau@tuebingen.mpg.de
Tim Genewein tim.genewein@mpg.tuebingen.de
Intelligent Systems, Max-Planck Institute
David Balduzzi david.balduzzi@inf.ethz.ch
Computer Science, ETH Zurich
Daniel Braun dab54@cam.ac.uk
Engineering, University of Cambridge

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

M29 Slice sampling normalized kernel-weighted completely random measure mixture models

Nick Foti nfoti@cs.dartmouth.edu
Computer Science, Dartmouth College
Sinead Williamson sineadannewilliamson@gmail.com
Machine Learning, Carnegie Mellon University

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

M30 MAP Inference in Chains using Column Generation

David Belanger belanger@cs.umass.edu
Alexandre Passos alexandre.tp@gmail.com
Sebastian Riedel sriedel@cs.umass.edu
Andrew McCallum mccallum@cs.umass.edu
UMass Amherst

Linear chains and trees are basic building blocks in many applications of graphical models. Although exact inference in these models can be performed by dynamic programming, this computation can still be prohibitively expensive with non-trivial target variable domain sizes due to the quadratic dependence on this size. Standard message-passing algorithms for these problems are inefficient because they compute scores on hypotheses for which there is strong negative local evidence. For this reason there has been significant previous interest in beam search and its variants; however, these methods provide only approximate inference. This paper presents new efficient exact inference algorithms based on the combination of it column generation and pre-computed bounds on the model's cost structure. Improving worst-case performance is impossible. However, our method substantially speeds real-world, typical-case inference in chains and trees. Experiments show our method to be twice as fast as exact Viterbi for Wall Street Journal part-of-speech tagging and over thirteen times faster for a joint part-of-speed and named-entity-recognition task. Our algorithm is also extendable to new techniques for approximate inference, to faster two-best inference, and new opportunities for connections between inference and learning.

M31 Learning curves for multi-task Gaussian process regression

Peter Sollich peter.sollich@kcl.ac.uk
Simon Ashton srfashton@gmail.com
King's College London

We study the average case performance of multi-task Gaussian process (GP) regression as captured in the learning curve, i.e.\ the average Bayes error for a chosen task versus the total number of examples n for all tasks. For GP covariances that are the product of an input-dependent covariance function and a free-form inter-task covariance matrix, we show that accurate approximations for the learning curve can be obtained for an arbitrary number of tasks T. We use these to study the asymptotic learning behaviour for large n. Surprisingly, multi-task learning can be asymptotically essentially useless: examples from other tasks only help when the degree of inter-task correlation, ρ, is near its maximal value ρ=1. This effect is most extreme for learning of smooth target functions as described by e.g.\ squared exponential kernels. We also demonstrate that when learning {\em many} tasks, the learning curves separate into an initial phase, where the Bayes error on each task is reduced down to a plateau value by ``collective learning'' even though most tasks have not seen examples, and a final decay that occurs only once the number of examples is proportional to the number of tasks.

M32 Practical Bayesian Optimization of Machine Learning Algorithms

Jasper Snoek jasper@cs.toronto.edu
University of Toronto
Hugo Larochelle hugo.larochelle@usherbrooke.ca
Université de Sherbrooke
Ryan Adams rpa@seas.harvard.edu
Harvard University

The use of machine learning algorithms frequently involves careful tuning of learning parameters and model hyperparameters. Unfortunately, this tuning is often a "black art" requiring expert experience, rules of thumb, or sometimes brute-force search. There is therefore great appeal for automatic approaches that can optimize the performance of any given learning algorithm to the problem at hand. In this work, we consider this problem through the framework of Bayesian optimization, in which a learning algorithm's generalization performance is modeled as a sample from a Gaussian process (GP). We show that certain choices for the nature of the GP, such as the type of kernel and the treatment of its hyperparameters, can play a crucial role in obtaining a good optimizer that can achieve expert-level performance. We describe new algorithms that take into account the variable cost (duration) of learning algorithm experiments and that can leverage the presence of multiple cores for parallel experimentation. We show that these proposed algorithms improve on previous automatic procedures and can reach or surpass human expert-level optimization for many algorithms including Latent Dirichlet Allocation, Structured SVMs and convolutional neural networks.

M33 Iterative Thresholding Algorithm for Sparse Inverse Covariance Estimation

Ben Rolfs benrolfs@stanford.edu
Bala Rajaratnam brajarat@stanford.edu
Dominique Guillot dguillot@stanford.edu
Arian Maleki arian.maleki@gmail.com
Ian Wong ihm.wong@gmail.com
Stanford University

Sparse graphical modelling/inverse covariance selection is an important problem in machine learning and has seen significant advances in recent years. A major focus has been on methods which perform model selection in high dimensions. To this end, numerous convex ℓ1 regularization approaches have been proposed in the literature. It is not however clear which of these methods are optimal in any well-defined sense. A major gap in this regard pertains to the rate of convergence of proposed optimization methods. To address this, an iterative thresholding algorithm for numerically solving the ℓ1-penalized maximum likelihood problem for sparse inverse covariance estimation is presented. The proximal gradient method considered in this paper is shown to converge at a linear rate, a result which is the first of its kind for numerically solving the sparse inverse covariance estimation problem. The convergence rate is provided in closed form, and is related to the condition number of the optimal point. Numerical results demonstrating the proven rate of convergence are presented.

M34 A Divide-and-Conquer Method for Sparse Inverse Covariance Estimation

Cho-Jui Hsieh cjhsieh@cs.utexas.edu
Inderjit Dhillon inderjit@cs.utexas.edu
Pradeep Ravikumar pradeepr@cs.utexas.edu
University of Texas, Austin
Arindam Banerjee abanerje@ece.utexas.edu
Univ. of Minnesota

In this paper, we consider the ℓ1 regularized sparse inverse covariance matrix estimation problem with a very large number of variables. Even in the face of this high dimensionality, and with limited number of samples, recent work has shown this estimator to have strong statistical guarantees in recovering the true structure of the sparse inverse covariance matrix, or alternatively the underlying graph structure of the corresponding Gaussian Markov Random Field. Our proposed algorithm divides the problem into smaller sub-problems, and uses the solutions of the sub-problems to build a good approximation for the original problem. We derive a bound on the distance of the approximate solution to the true solution. Based on this bound, we propose a clustering algorithm that attempts to minimize this bound, and in practice, is able to find effective partitions of the variables. We further use the approximate solution, i.e., solution resulting from solving the sub-problems, as an initial point to solve the original problem, and achieve a much faster computational procedure. As an example, a recent state-of-the-art method, QUIC requires 10 hours to solve a problem (with 10,000 nodes) that arises from a climate application, while our proposed algorithm, Divide and Conquer QUIC (DC-QUIC) only requires one hour to solve the problem.

M35 Approximate Message Passing with Consistent Parameter Estimation and Applications to Sparse Learning

Ulugbek Kamilov ulugbek.kamilov@epfl.ch
MIchael Unser michael.unser@epfl.ch
EPFL
Sundeep Rangan srangan@poly.edu
ECE, Polytechnic Institute of New York University
Alyson Fletcher alyson@eecs.berkeley.edu
University of California, Berkeley

We consider the estimation of an i.i.d.\ vector \xbf∈\Rn from measurements \ybf∈\Rm obtained by a general cascade model consisting of a known linear transform followed by a probabilistic componentwise (possibly nonlinear) measurement channel. We present a method, called adaptive generalized approximate message passing (Adaptive GAMP), that enables joint learning of the statistics of the prior and measurement channel along with estimation of the unknown vector \xbf. The proposed algorithm is a generalization of a recently-developed method by Vila and Schniter that uses expectation-maximization (EM) iterations where the posteriors in the E-steps are computed via approximate message passing. The techniques can be applied to a large class of learning problems including the learning of sparse priors in compressed sensing or identification of linear-nonlinear cascade models in dynamical systems and neural spiking processes. We prove that for large i.i.d.\ Gaussian transform matrices the asymptotic componentwise behavior of the adaptive GAMP algorithm is predicted by a simple set of scalar state evolution equations. This analysis shows that the adaptive GAMP method can yield asymptotically consistent parameter estimates, which implies that the algorithm achieves a reconstruction quality equivalent to the oracle algorithm that knows the correct parameter values. The adaptive GAMP methodology thus provides a systematic, general and computationally efficient method applicable to a large range of complex linear-nonlinear models with provable guarantees.

M36 The Bethe Partition Function of Log-supermodular Graphical Models

Nicholas Ruozzi nicholas.ruozzi@epfl.ch
EPFL

Sudderth, Wainwright, and Willsky conjectured that the Bethe approximation corresponding to any fixed point of the belief propagation algorithm over an attractive, pairwise binary graphical model provides a lower bound on the true partition function. In this work, we resolve this conjecture in the affirmative by demonstrating that, for any graphical model with binary variables whose potential functions (not necessarily pairwise) are all log-supermodular, the Bethe partition function always lower bounds the true partition function. The proof of this result follows from a new variant of the “four functions” theorem that may be of independent interest.

M37 Convergence Rate Analysis of MAP Coordinate Minimization Algorithms

Ofer Meshi meshi@cs.huji.ac.il
Hebrew University of Jerusalem
Tommi Jaakkola tommi@csail.mit.edu
EECS, Massachusetts Institute of Technology
Amir Globerson gamir@cs.huji.ac.il
Hebrew University

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

M38 Bayesian Probabilistic Co-Subspace Addition

Lei Shi shilei06@baidu.com
Natural Language Processing, Baidu.com, Inc

For modeling data matrices, this paper introduces Probabilistic Co-Subspace Addition (PCSA) model by simultaneously capturing the dependent structures among both rows and columns. Briefly, PCSA assumes that each entry of a matrix is generated by the additive combination of the linear mappings of two features, which distribute in the row-wise and column-wise latent subspaces. Consequently, it captures the dependencies among entries intricately, and is able to model the non-Gaussian and heteroscedastic density. Variational inference is proposed on PCSA for approximate Bayesian learning, where the updating for posteriors is formulated into the problem of solving Sylvester equations. Furthermore, PCSA is extended to tackling and filling missing values, to adapting its sparseness, and to modelling tensor data. In comparison with several state-of-art approaches, experiments demonstrate the effectiveness and efficiency of Bayesian (sparse) PCSA on modeling matrix (tensor) data and filling missing values.

M39 Learning the Architecture of Sum-Product Networks Using Clustering on Variables

Aaron Dennis adennis@byu.edu
Dan Ventura ventura@cs.byu.edu
Computer Science, Brigham Young University

The sum-product network (SPN) is a recently-proposed deep model consisting of a network of sum and product nodes, and has been shown to be competitive with state-of-the-art deep models on certain difficult tasks such as image completion. Designing an SPN network architecture that is suitable for the task at hand is an open question. We propose an algorithm for learning the SPN architecture from data. The idea is to cluster variables (as opposed to data instances) in order to identify variable subsets that strongly interact with one another. Nodes in the SPN network are then allocated towards explaining these interactions. Experimental evidence shows that learning the SPN architecture significantly improves its performance compared to using a previously-proposed static architecture.

M40 Efficient Sampling for Bipartite Matching Problems

Maksims Volkovs mvolkovs@cs.toronto.edu
Richard Zemel zemel@cs.toronto.edu
Computer Science, University of Toronto

Bipartite matching problems characterize many situations, ranging from ranking in information retrieval to correspondence in vision. Exact inference in real-world applications of these problems is intractable, making efficient approximation methods essential for learning and inference. In this paper we propose a novel {\it sequential matching} sampler based on the generalization of the Plackett-Luce model, which can effectively make large moves in the space of matchings. This allows the sampler to match the difficult target distributions common in these problems: highly multimodal distributions with well separated modes. We present experimental results with bipartite matching problems - ranking and image correspondence - which show that the sequential matching sampler efficiently approximates the target distribution, significantly outperforming other sampling approaches.

M41 Projection Retrieval for Classification

Madalina Fiterau mfiterau@cs.cmu.edu
Artur Dubrawski awd@cs.cmu.edu
Carnegie Mellon University

In many applications classification systems often require in the loop human intervention. In such cases the decision process must be transparent and comprehensible simultaneously requiring minimal assumptions on the underlying data distribution. To tackle this problem, we formulate it as an axis-alligned subspacefinding task under the assumption that query specific information dictates the complementary use of the subspaces. We develop a regression-based approach called RECIP that efficiently solves this problem by finding projections that minimize a nonparametric conditional entropy estimator. Experiments show that the method is accurate in identifying the informative projections of the dataset, picking the correct ones to classify query points and facilitates visual evaluation by users.

M42 Learning Multiple Tasks using Shared Hypotheses

Koby Crammer koby@ee.technion.ac.il
The Technion
Yishay Mansour mansour@cs.tau.ac.il
Tel-Aviv University

In this work we consider a setting where we have a very large number of related tasks with few examples from each individual task. Rather than either learning each task individually (and having a large generalization error) or learning all the tasks together using a single hypothesis (and suffering a potentially large inherent error), we consider learning a small pool of {\em shared hypotheses}. Each task is then mapped to a single hypothesis in the pool (hard association). We derive VC dimension generalization bounds for our model, based on the number of tasks, shared hypothesis and the VC dimension of the hypotheses class. We conducted experiments with both synthetic problems and sentiment of reviews, which strongly support our approach.

M43 Optimal kernel choice for large-scale two-sample tests

Arthur Gretton arthur.gretton@googlemail.com
Bharath Sriperumbudur bharathsv.ucsd@gmail.com
Heiko Strathmann heiko.strathmann@googlemail.com
Dino Sejdinovic dino.sejdinovic@gmail.com
Massimiliano Pontil m.pontil@cs.ucl.ac.uk
University College London
Sivaraman Balakrishnan the.seeing.stone@gmail.com
Carnegie Mellon University
Kenji Fukumizu fukumizu@ism.ac.jp
Institute of Statistical Mathematics

Abstract Given samples from distributions p and q, a two-sample test determines whether to reject the null hypothesis that p=q, based on the value of a test statistic measuring the distance between the samples. One choice of test statistic is the maximum mean discrepancy (MMD), which is a distance between embeddings of the probability distributions in a reproducing kernel Hilbert space. The kernel used in obtaining these embeddings is thus critical in ensuring the test has high power, and correctly distinguishes unlike distributions with high probability. A means of parameter selection for the two-sample test based on the MMD is proposed. For a given test level (an upper bound on the probability of making a Type I error), the kernel is chosen so as to maximize the test power, and minimize the probability of making a Type II error. The test statistic, test threshold, and optimization over the kernel parameters are obtained with cost linear in the sample size. These properties make the kernel selection and test procedures suited to data streams, where the observations cannot all be stored in memory. In experiments, the new kernel selection approach yields a more powerful test than earlier kernel selection heuristics.

M44 Supervised Learning with Similarity Functions

Puru Kar purushot@cse.iitk.ac.in
Indian Institute of Technology Kanpur
Prateek Jain prajain@microsoft.com
Microsoft Research Lab

We address the problem of general supervised learning when data can only be accessed through an (indefinite) similarity function between data points. Existing work on learning with indefinite kernels has concentrated solely on binary/multiclass classification problems. We propose a model that is generic enough to handle any supervised learning task and also subsumes the model previously proposed for classification. We give a ''goodness'' criterion for similarity functions w.r.t. a given supervised learning task and then adapt a well-known landmarking technique to provide efficient algorithms for supervised learning using ''good'' similarity functions. We demonstrate the effectiveness of our model on three important supervised learning problems: a) real-valued regression, b) ordinal regression and c) ranking where we show that our method guarantees bounded generalization error. Furthermore, for the case of real-valued regression, we give a natural goodness definition that, when used in conjunction with a recent result in sparse vector recovery, guarantees a sparse predictor with bounded generalization error. Finally, we report results of our learning algorithms on regression and ordinal regression tasks using non-PSD similarity functions and demonstrate the effectiveness of our algorithms, especially that of the sparse landmark selection algorithm that achieves significantly higher accuracies than the baseline methods while offering reduced computational costs.

M45 Density-Difference Estimation

Masashi Sugiyama sugi@cs.titech.ac.jp
Marthinus du Plessis christo@sg.cs.titech.ac.jp
Song Liu song@sg.cs.titech.ac.jp
Tokyo Institute of Technology
Takafumi Kanamori kanamori@is.nagoya-u.ac.jp
Nagoya university
Taiji Suzuki s-taiji@stat.t.u-tokyo.ac.jp
University of Tokyo
Ichiro Takeuchi takeuchi.ichiro@nitech.ac.jp
Nagoya Institute of Technology

We address the problem of estimating the difference between two probability densities. A naive approach is a two-step procedure of first estimating two densities separately and then computing their difference. However, such a two-step procedure does not necessarily work well because the first step is performed without regard to the second step and thus a small estimation error incurred in the first stage can cause a big error in the second stage. In this paper, we propose a single-shot procedure for directly estimating the density difference without separately estimating two densities. We derive a non-parametric finite-sample error bound for the proposed single-shot density-difference estimator and show that it achieves the optimal convergence rate. We then show how the proposed density-difference estimator can be utilized in L2-distance approximation. Finally, we experimentally demonstrate the usefulness of the proposed method in robust distribution comparison such as class-prior estimation and change-point detection.

M46 The Lovasz θ function, SVMs and finding large dense subgraphs

Vinay Jethava jethava@chalmers.se
Anders Martinsson andemar@student.chalmers.se
Devdatt Dubhashi dubhashi@chalmers.se
Chalmers University
Chiranjib Bhattacharyya
chiranjib.bhattacharyya@gmail.com
Indian Institute of Science

The Lovasz θ function of a graph, is a fundamental tool in combinatorial optimization and approximation algorithms. Computing θ involves solving a SDP and is extremely expensive even for moderately sized graphs. In this paper we establish that the Lovasz θ function is equivalent to a kernel learning problem related to one class SVM. This interesting connection opens up many opportunities bridging graph theoretic algorithms and machine learning. We show that there exist graphs, which we call SVM−θ graphs, on which the Lovasz θ function can be approximated well by a one-class SVM. This leads to a novel use of SVM techniques to solve algorithmic problems in large graphs e.g. identifying a planted clique of size Θ(n) in a random graph G(n,12). A classic approach for this problem involves computing the θ function, however it is not scalable due to SDP computation. We show that the random graph with a planted clique is an example of SVM−θ graph, and as a consequence a SVM based approach easily identifies the clique in large graphs and is competitive with the state-of-the-art. Further, we introduce the notion of a ''common orthogonal labeling'' which extends the notion of a ''orthogonal labelling of a single graph (used in defining the θ function) to multiple graphs. The problem of finding the optimal common orthogonal labelling is cast as a Multiple Kernel Learning problem and is used to identify a large common dense region in multiple graphs. The proposed algorithm achieves an order of magnitude scalability compared to the state of the art.

M47 Perceptron Learning of SAT

Alex Flint alexf@robots.ox.ac.uk
Oxford University
Matthew Blaschko matthew.blaschko@inria.fr
Ecole Centrale Paris

Boolean satisfiability (SAT) as a canonical NP-complete decision problem is one of the most important problems in computer science. In practice, real-world SAT sentences are drawn from a distribution that may result in efficient algorithms for their solution. Such SAT instances are likely to have shared characteristics and substructures. This work approaches the exploration of a family of SAT solvers as a learning problem. In particular, we relate polynomial time solvability of a SAT subset to a notion of margin between sentences mapped by a feature function into a Hilbert space. Provided this mapping is based on polynomial time computable statistics of a sentence, we show that the existance of a margin between these data points implies the existance of a polynomial time solver for that SAT subset based on the Davis-Putnam-Logemann-Loveland algorithm. Furthermore, we show that a simple perceptron-style learning rule will find an optimal SAT solver with a bounded number of training updates. We derive a linear time computable set of features and show analytically that margins exist for important polynomial special cases of SAT. Empirical results show an order of magnitude improvement over a state-of-the-art SAT solver on a hardware verification task.

M48 A Polynomial-time Form of Robust Regression

Yao-Liang Yu yaoliang@cs.ualberta.ca
Ozlem Aslan ozlemmaslan@gmail.com
Dale Schuurmans dale@cs.ualberta.ca
Department of Computing Science, University of Alberta

Despite the variety of robust regression methods that have been developed, current regression formulations are either NP-hard, or allow unbounded response to even a single leverage point. We present a general formulation for robust regression --Variational M-estimation--that unifies a number of robust regression methods while allowing a tractable approximation strategy. We develop an estimator that requires only polynomial-time, while achieving certain robustness and consistency guarantees. An experimental evaluation demonstrates the effectiveness of the new estimation approach compared to standard methods.

M49 Feature Clustering for Accelerating Parallel Coordinate Descent

Chad Scherrer chad.scherrer@gmail.com
Independent Consultant
Ambuj Tewari ambujtewari@gmail.com
University of Michigan
Mahantesh Halappanavar
Mahantesh.Halappanavar@pnnl.gov
David Haglin David.Haglin@pnnl.gov
Pacific Northwest National Laboratory

Large scale ℓ1-regularized loss minimization problems arise in numerous applications such as compressed sensing and high dimensional supervised learning, including classification and regression problems. High performance algorithms and implementations are critical to efficiently solving these problems. Building upon previous work on coordinate descent algorithms for ℓ1 regularized problems, we introduce a novel family of algorithms called block-greedy coordinate descent that includes, as special cases, several existing algorithms such as SCD, Greedy CD, Shotgun, and Thread-greedy. We give a unified convergence analysis for the family of block-greedy algorithms. The analysis suggests that block-greedy coordinate descent can better exploit parallelism if features are clustered so that the maximum inner product between features in different blocks is small. Our theoretical convergence analysis is supported with experimental results using data from diverse real-world applications. We hope that algorithmic approaches and convergence analysis we provide will not only advance the field, but will also encourage researchers to systematically explore the design space of algorithms for solving large-scale ℓ1-regularization problems.

M50 Mixing Properties of Conditional Markov Chains with Unbounded Feature Functions

Mathieu Sinn mathsinn@ie.ibm.com
IBM Research - Ireland
Bei Chen bei.chen@math.mcmaster.ca
McMaster University

Conditional Markov Chains (also known as Linear-Chain Conditional Random Fields in the literature) are a versatile class of discriminative models for the distribution of a sequence of hidden states conditional on a sequence of observable variables. Large-sample properties of Conditional Markov Chains have been first studied by Sinn and Poupart [1]. The paper extends this work in two directions: first, mixing properties of models with unbounded feature functions are being established; second, necessary conditions for model identifiability and the uniqueness of maximum likelihood estimates are being given.

M51 Efficient Monte Carlo Counterfactual Regret Minimization in Games with Many Player Actions

Richard Gibson rggibson@cs.ualberta.ca
Marc Lanctot lanctot@ualberta.ca
Neil Burch nburch@ualberta.ca
Duane Szafron dszafron@ualberta.ca
Computing Science, University of Alberta

Counterfactual Regret Minimization (CFR) is a popular, iterative algorithm for computing strategies in extensive-form games. The Monte Carlo CFR (MCCFR) variants reduce the per iteration time cost of CFR by traversing a sampled portion of the tree. The previous most effective instances of MCCFR can still be very slow in games with many player actions since they sample every action for a given player. In this paper, we present a new MCCFR algorithm, Average Strategy Sampling (AS), that samples a subset of the player's actions according to the player's average strategy. Our new algorithm is inspired by a new, tighter bound on the number of iterations required by CFR to converge to a given solution quality. In addition, we prove a similar, tighter bound for AS and other popular MCCFR variants. Finally, we validate our work by demonstrating that AS converges faster than previous MCCFR algorithms in both no-limit poker and Bluff.

M52 Tight Bounds on Redundancy and Distinguishability of Label-Invariant Distributions

Jayadev Acharya jacharya@ucsd.edu
Hirakendu Das hdas@ucsd.edu
University of California, San Diego
Alon Orlitsky alon@ucsd.edu

The minimax KL-divergence of any distribution from all distributions in a given collection has several practical implications. In compression, it is the least additional number of bits over the entropy needed in the worst case to encode the output of a distribution in the collection. In online estimation and learning, it is the lowest expected log-loss regret when guessing a sequence of random values. In hypothesis testing, it upper bounds the largest number of distinguishable distributions in the collection. Motivated by problems ranging from population estimation to text classification and speech recognition, several machine-learning and information-theory researchers have recently considered label-invariant distributions and properties of \iid-drawn samples. Using techniques that reveal and exploit the structure of these distributions, we improve on a sequence of previous works and show that the minimax KL-divergence of the collection of label-invariant distributions over length-n \iid sequences is between 0.3·n1/3 and n1/3log2n.

M53 Exponential Concentration for Mutual Information Estimation with Application to Forests

Han Liu hanliu@princeton.edu
Princeton University
John Lafferty lafferty@gmail.com
University of Chicago
Larry Wasserman larry@stat.cmu.edu
Carnegie Mellon University

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

M54 Bayesian estimation of discrete entropy with mixtures of stick-breaking priors

Evan Archer earcher@utexas.edu
Jonathan Pillow pillow@mail.utexas.edu
Il Memming Park memming@austin.utexas.edu
University of Texas at Austin

We consider the problem of estimating Shannon's entropy H in the under-sampled regime, where the number of possible symbols may be unknown or countably infinite. Pitman-Yor processes (a generalization of Dirichlet processes) provide tractable prior distributions over the space of countably infinite discrete distributions, and have found major applications in Bayesian non-parametric statistics and machine learning. Here we show that they also provide natural priors for Bayesian entropy estimation, due to the remarkable fact that the moments of the induced posterior distribution over H can be computed analytically. We derive formulas for the posterior mean (Bayes' least squares estimate) and variance under such priors. Moreover, we show that a fixed Dirichlet or Pitman-Yor process prior implies a narrow prior on H, meaning

the prior strongly determines the entropy estimate in the under-sampled regime. We derive a family of continuous mixing measures such that the resulting mixture of Pitman-Yor processes produces an approximately flat (improper) prior over H. We explore the theoretical properties of the resulting estimator, and show that it performs well on data sampled from both exponential and power-law tailed distributions.

M55 Learning Halfspaces with the Zero-One Loss: Time-Accuracy Tradeoffs

Aharon Birnbaum aharon.birnbaum@mail.huji.ac.il
Shai Shalev-Shwartz shai.shwartz@gmail.com
Hebrew University

Given α,ϵ, we study the time complexity required to improperly learn a halfspace with misclassification error rate of at most (1+α)Lγ∗+ϵ, where Lγ∗ is the optimal γ-margin error rate. For α=1/γ, polynomial time and sample complexity is achievable using the hinge-loss. For α=0, \cite{ShalevShSr11} showed that \poly(1/γ) time is impossible, while learning is possible in time exp(O˜(1/γ)). An immediate question, which this paper tackles, is what is achievable if α∈(0,1/γ). We derive positive results interpolating between the polynomial time for α=1/γ and the exponential time for α=0. In particular, we show that there are cases in which α=o(1/γ) but the problem is still solvable in polynomial time. Our results naturally extend to the adversarial online learning model and to the PAC learning with malicious noise model.

M56 A Scalable CUR Matrix Decomposition Algorithm: Lower Time Complexity and Tighter Bound

Shusen Wang wssatzju@gmail.com
Zhihua Zhang zhzhang@gmail.com
CS, Zhejiang University

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

M57 Online L1-Dictionary Learning with Application to Novel Document Detection

Shiva Kasiviswanathan kasivisw@gmail.com
GE Global Research
Huahua Wang huwang@cs.umn.edu
Arindam Banerjee abanerje@ece.utexas.edu
Univ. of Minnesota
Prem Melville prem.melville@gmail.com
IBM Watson Research

Given their pervasive use, social media, such as Twitter, have become a leading source of breaking news. A key task in the automated identification of such news is the detection of novel documents from a voluminous stream of text documents in a scalable manner. Motivated by this challenge, we introduce the problem of online L1-dictionary learning where unlike traditional dictionary learning, which uses squared loss, the L1-penalty is used for measuring the reconstruction error. We present an efficient online algorithm for this problem based on alternating directions method of multipliers, and establish a sublinear regret bound for this algorithm. Empirical results on news-stream and Twitter data, shows that this online L1-dictionary learning algorithm for novel document detection gives more than an order of magnitude speedup over the previously known batch algorithm, without any significant loss in quality of results. Our algorithm for online L1-dictionary learning could be of independent interest.

M58 Ensemble weighted kernel estimators for multivariate entropy estimation

Kumar Sricharan kksreddy@umich.edu
Alfred Hero hero@umich.edu
EECS, University of Michigan

The problem of estimation of entropy functionals of probability densities has received much attention in the information theory, machine learning and statistics communities. Kernel density plug-in estimators are simple, easy to implement and widely used for estimation of entropy. However, kernel plug-in estimators suffer from the curse of dimensionality, wherein the MSE rate of convergence is glacially slow - of order O(T−γ/d), where T is the number of samples, and γ>0 is a rate parameter. In this paper, it is shown that for sufficiently smooth densities, an ensemble of kernel plug-in estimators can be combined via a weighted convex combination, such that the resulting weighted estimator has a superior parametric MSE rate of convergence of order O(T−1). Furthermore, it is shown that these optimal weights can be determined by solving a convex optimization problem which does not require training data or knowledge of the underlying density, and therefore can be performed offline. This novel result is remarkable in that, while each of the individual kernel plug-in estimators belonging to the ensemble suffer from the curse of dimensionality, by appropriate ensemble averaging we can achieve parametric convergence rates.

M59 Dip-means: an incremental clustering method for estimating the number of clusters

Argyris Kalogeratos akaloger@cs.uoi.gr
Aristidis Likas arly@cs.uoi.gr
Computer Science, University of Ioannina

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

M60 Convergence and Energy Landscape for Cheeger Cut Clustering

Xavier Bresson xbresson@cityu.edu.hk
City University of Hong Kong
Thomas Laurent laurent@math.ucr.edu
UC Riverside
David Uminsky duminsky@usfca.edu
University of San Francisco
James von Brecht jub@math.ucla.edu
UCLA

Unsupervised clustering of scattered, noisy and high-dimensional data points is an important and difficult problem. Continuous relaxations of balanced cut problems yield excellent clustering results. This paper provides rigorous convergence results for two algorithms that solve the relaxed Cheeger Cut minimization. The first algorithm is a new steepest descent algorithm and the second one is a slight modification of the Inverse Power Method algorithm \cite{pro:HeinBuhler10OneSpec}. While the steepest descent algorithm has better theoretical convergence properties, in practice both algorithm perform equally. We also completely characterize the local minima of the relaxed problem in terms of the original balanced cut problem, and relate this characterization to the convergence of the algorithms.

M61 Cardinality Restricted Boltzmann Machines

Kevin Swersky kswersky@cs.toronto.edu
Danny Tarlow dtarlow@cs.toronto.edu
Ilya Sutskever ilya@cs.utoronto.ca
Richard Zemel zemel@cs.toronto.edu
Russ Salakhutdinov rsalakhu@mit.edu
University of Toronto
Ryan Adams rpa@seas.harvard.edu
Harvard University

The Restricted Boltzmann Machine (RBM) is a popular density model that is also good for extracting features. A main source of tractability in RBM models is the model's assumption that given an input, hidden units activate independently from one another. Sparsity and competition in the hidden representation is believed to be beneficial, and while an RBM with competition among its hidden units would acquire some of the attractive properties of sparse coding, such constraints are not added due to the widespread belief that the resulting model would become intractable. In this work, we show how a dynamic programming algorithm developed in 1981 can be used to implement exact sparsity in the RBM's hidden units. We then expand on this and show how to pass derivatives through a layer of exact sparsity, which makes it possible to fine-tune a deep belief network (DBN) consisting of RBMs with sparse hidden layers. We show that sparsity in the RBM's hidden layer improves the performance of both the pre-trained representations and of the fine-tuned model.

M62 Controlled Recognition Bounds for Visual Learning and Exploration

Vasiliy Karasev karasev00@ucla.edu
Stefano Soatto soatto@cs.ucla.edu
UCLA
Alessandro Chiuso chiuso@dei.unipd.it
University of Padova

We describe the tradeoff between the performance in a visual recognition problem and the control authority that the agent can exercise on the sensing process. We focus on the problem of "visual search" of an object in an otherwise known and static scene, propose a measure of control authority, and relate it to the expected risk and its proxy (conditional entropy of the posterior density). We show this analytically, as well as empirically by simulation using the simplest known model that captures the phenomenology of image formation, including scaling and occlusions. We show that a "passive" agent given a training set can provide no guarantees on performance beyond what is afforded by the priors, and that an "omnipotent" agent, capable of infinite control authority, can achieve arbitrarily good performance (asymptotically).

M63 Clustering Aggregation as Maximum-Weight Independent Set

Nan Li Nan.Li@temple.edu
Longin Jan Latecki latecki@temple.edu
Temple University

We formulate clustering aggregation as a special instance of Maximum-Weight Independent Set (MWIS) problem. For a given dataset, an attributed graph is constructed from the union of the input clusterings generated by different underlying clustering algorithms with different parameters. The vertices, which represent the distinct clusters, are weighted by an internal index measuring both cohesion and separation. The edges connect the vertices whose corresponding clusters overlap. Intuitively, an optimal aggregated clustering can be obtained by selecting an optimal subset of non-overlapping clusters partitioning the dataset together. We formalize this intuition as the MWIS problem on the attributed graph, i.e., finding the heaviest subset of mutually non-adjacent vertices. This MWIS problem exhibits a special structure. Since the clusters of each input clustering form a partition of the dataset, the vertices corresponding to each clustering form a maximal independent set (MIS) in the attributed graph. We propose a variant of simulated annealing method that takes advantage of this special structure. Our algorithm starts from each MIS, which is close to a distinct local optimum of the MWIS problem, and utilizes a local search heuristic to explore its neighborhood in order to find the MWIS. Extensive experiments on many challenging datasets show that: 1. our approach to clustering aggregation automatically decides the optimal number of clusters; 2. it does not require any parameter tuning for the underlying clustering algorithms; 3. it can combine the advantages of different underlying clustering algorithms to achieve superior performance; 4. it is robust against moderate or even bad input clusterings.

M64 Clustering by Nonnegative Matrix Factorization Using Graph Random Walk

Zhirong Yang zhirong.yang@aalto.fi
Tele Hao tele.hao@aalto.fi
Onur Dikmen onur.dikmen@aalto.fi
Erkki Oja erkki.oja@tkk.fi
Aalto University
Xi Chen xichen@cs.cmu.edu
Carnegie Mellon University

Nonnegative Matrix Factorization (NMF) is a promising relaxation technique for clustering analysis. However, conventional NMF methods that directly approximate the pairwise similarities using the least square error often yield mediocre performance for data in curved manifolds because they can capture only the immediate similarities between data samples. Here we propose a new NMF clustering method which replaces the approximated matrix with its smoothed version using random walk. Our method can thus accommodate farther relationships between data samples. Furthermore, we introduce a novel regularization in the proposed objective function in order to improve over spectral clustering. The new learning objective is optimized by a multiplicative Majorization-Minimization algorithm with a scalable implementation for learning the factorizing matrix. Extensive experimental results on real-world datasets show that our method has strong performance in terms of cluster purity.

M65 Angular Quantization based Binary Codes for Fast Similarity Search

Yunchao Gong yunchao@cs.unc.edu
Vishal Verma verma@cs.unc.edu
Chapel Hill
Sanjiv Kumar sanjivk@google.com
Research, NY, Google
Svetlana Lazebnik lazebnik@cs.unc.edu
UIUC

This paper focuses on the problem of learning binary embeddings for efficient retrieval of high-dimensional non-negative data. Such data typically arises in a large number of vision and text applications where counts or frequencies are used as features. Also, cosine distance is commonly used as a measure of dissimilarity between such vectors. In this work, we introduce a novel spherical quantization scheme to generate binary embedding of such data and analyze its properties. The number of quantization landmarks in this scheme grows exponentially with data dimensionality resulting in low-distortion quantization. We propose a very efficient method for computing the binary embedding using such large number of landmarks. Further, a linear transformation is learned to minimize the quantization error by adapting the method to the input data resulting in improved embedding. Experiments on image and text retrieval applications show superior performance of the proposed method over other existing state-of-the-art methods.

M66 Near-optimal Differentially Private Principal Components

Kamalika Chaudhuri kamalika@cs.ucsd.edu
UC San Diego
Anand Sarwate asarwate@ttic.edu
N/A, TTI Chicago
Kaushik Sinha sinhak@cse.ohio-state.edu
Ohio State University

Principal components analysis (PCA) is a standard tool for identifying good low-dimensional approximations to data sets in high dimension. Many current data sets of interest contain private or sensitive information about individuals. Algorithms which operate on such data should be sensitive to the privacy risks in publishing their outputs. Differential privacy is a framework for developing tradeoffs between privacy and the utility of these outputs. In this paper we investigate the theory and empirical performance of differentially private approximations to PCA and propose a new method which explicitly optimizes the utility of the output. We demonstrate that on real data, there this a large performance gap between the existing methods and our method. We show that the sample complexity for the two procedures differs in the scaling with the data dimension, and that our method is nearly optimal in terms of this scaling.

M67 On the Sample Complexity of Robust PCA

Matthew Coudron mcoudron@mit.edu
Massachusetts Institute of Technology
Gilad Lerman lerman@umn.edu
University of Minnesota

We estimate the sample complexity of a recent robust estimator for a generalized version of the inverse covariance matrix. This estimator is used in a convex algorithm for robust subspace recovery (i.e., robust PCA). Our model assumes a sub-Gaussian underlying distribution and an i.i.d.~sample from it. Our main result shows with high probability that the norm of the difference between the generalized inverse covariance of the underlying distribution and its estimator from an i.i.d.~sample of size N is of order O(N−0.5+\eps) for arbitrarily small \ eps>0 (affecting the probabilistic estimate); this rate of convergence is close to one of direct covariance and inverse covariance estimation, i.e., O(N−0.5). Our precise probabilistic estimate implies for some natural settings that the sample complexity of the generalized inverse covariance estimation when using the Frobenius norm is O(D2+δ) for arbitrarily small δ>0 (whereas the sample complexity of direct covariance estimation with Frobenius norm is O(D2)). These results provide similar rates of convergence and sample complexity for the corresponding robust subspace recovery algorithm, which are close to those of PCA. To the best of our knowledge, this is the only work analyzing the sample complexity of any robust PCA algorithm.

M68 Learning the Dependency Structure of Latent Factors

Yunlong He he.yunlong@gmail.com
Haesun Park hpark@cc.gatech.edu
Georgia Institute of Technology
Yanjun Qi yanjun@nec-labs.com
koray kavukcuoglu koray@nec-labs.com
NEC Laboratories, America

In this paper, we study latent factor models with the dependency structure in the latent space. We propose a general learning framework which induces sparsity on the undirected graphical model imposed on the vector of latent factors. A novel latent factor model SLFA is then proposed as a matrix factorization problem with a special regularization term that encourages collaborative reconstruction. The main benefit (novelty) of the model is that we can simultaneously learn the lower-dimensional representation for data and model the pairwise relationships between latent factors explicitly. An on-line learning algorithm is devised to make the model feasible for large-scale learning problems. Experimental results on two synthetic data and two real-world data sets demonstrate that pairwise relationships and latent factors learned by our model provide a more structured way of exploring high-dimensional data, and the learned representations achieve the state-of-the-art classification performance.

M69 Bayesian Nonparametric Maximum Margin Matrix Factorization for Collaborative Prediction

Minjie Xu chokkyvista06@gmail.com
Jun Zhu jjzhunet9@hotmail.com
Tsinghua University
Bo Zhang bozhang@fairisaac.com
Fair Isaac Corp.

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

M70 Identifiability and Unmixing of Latent Parse Trees

Percy Liang pliang@cs.stanford.edu
Stanford University
Sham Kakade skakade@microsoft.com
Daniel Hsu danielhsu@gmail.com
Microsoft Research

This poster has been moved to Wednesday. Please see poster W93 on page 90 for abstract description.

M71 How They Vote: Issue-Adjusted Models of Legislative Behavior

Sean Gerrish sean.gerrish@gmail.com
David Blei blei@cs.princeton.edu
Princeton University

We develop a probabilistic model of legislative data that uses the text of the bills to uncover lawmakers' positions on specific political issues. Our model can be used to explore how a lawmaker's voting patterns deviate from what is expected and how that deviation depends on what is being voted on. We derive approximate posterior inference algorithms based on variational methods. Across 12 years of legislative data, we demonstrate both improvement in heldout predictive performance and the model's utility in interpreting an inherently multi-dimensional space.

M72 3D Gaze Concurrences from Head-mounted Cameras

Hyun Soo Park hyunsoop@cs.cmu.edu
Eakta Jain eakta@cmu.edu
Yaser Sheikh yaser@cs.cmu.edu
Carnegie Mellon University

A gaze concurrence is a point in 3D where the gaze directions of two or more people intersect. It is a strong indicator of social saliency because the attention of the participating group is focused on that point. In scenes occupied by large groups of people, multiple concurrences may occur and transition over time. In this paper, we present a method to locate multiple gaze concurrences that occur in a social scene from videos taken by head-mounted cameras. We model the gaze as a cone-shaped distribution emanating from the center of the eyes, capturing the variation of eye-in-head motion. We calibrate the parameters of this distribution by exploiting the fixed relationship between the primary gaze ray and the head-mounted camera pose. The resulting gaze model enables us to build a social saliency field in 3D. We estimate the number and 3D locations of the gaze concurrences via provably convergent mode-seeking in the social saliency field. Our algorithm is applied to reconstruct multiple gaze concurrences in several real world scenes and evaluated quantitatively against motion-captured ground truth.

M73 Compressive Sensing MRI with Wavelet Tree Sparsity

Chen Chen cchen@mavs.uta.edu
Uni. of Texas at Arlington

Junzhou Huang jzhuang@uta.edu
University of Texas at Arlington

In Compressive Sensing Magnetic Resonance Imaging (CS-MRI), one can reconstruct a MR image with good quality from only a small number of measurements. This can significantly reduce MR scanning time. According to structured sparsity theory, the measurements can be further reduced to O(K+logn) for tree-sparse data instead of O(K+Klogn) for standard K-sparse data with length n. However, few of existing algorithms has utilized this for CS-MRI, while most of them use Total Variation and wavelet sparse regularization. On the other side, some algorithms have been proposed for tree sparsity regularization, but few of them has validated the benefit of tree structure in CS-MRI. In this paper, we propose a fast convex optimization algorithm to improve CS-MRI. Wavelet sparsity, gradient sparsity and tree sparsity are all considered in our model for real MR images. The original complex problem is decomposed to three simpler subproblems then each of the subproblems can be efficiently solved with an iterative scheme. Numerous experiments have been conducted and show that the proposed algorithm outperforms the state-of-the-art CS-MRI algorithms, and gain better reconstructions results on real MR images than general tree based solvers or algorithms.

M74 Recognizing Activities by Attribute Dynamics

Weixin Li wel017@ucsd.edu
Nuno Vasconcelos nuno@ece.ucsd.edu
UC San Diego

In this work, we consider the problem of modeling the dynamic structure of human activities in the attributes space. A video sequence is first represented in a semantic feature space, where each feature encodes the probability of occurrence of an activity attribute at a given time. A generative model, denoted the binary dynamic system (BDS), is proposed to learn both the distribution and dynamics of different activities in this space. The BDS is a non-linear dynamic system, which extends both the binary principal component analysis (PCA) and classical linear dynamic systems (LDS), by combining binary observation variables with a hidden Gauss-Markov state process. In this way, it integrates the representation power of semantic modeling with the ability of dynamic systems to capture the temporal structure of time-varying processes. An algorithm for learning BDS parameters, inspired by a popular LDS learning method from dynamic textures, is proposed. A similarity measure between BDSs, which generalizes the Binet-Cauchy kernel for LDS, is then introduced and used to design activity classifiers. The proposed method is shown to outperform similar classifiers derived from the kernel dynamic system (KDS) and state-of-the-art approaches for dynamics-based or attribute-based action recognition.

M75 Fusion with Diffusion for Robust Visual Tracking

Yu Zhou zhouyu.hust@gmail.com
Xiang Bai xiang.bai@gmail.com
Wenyu Liu liuwy@hust.edu.cn
Huazhong University of Science and Technology
Longin Jan Latecki latecki@temple.edu
Temple University

A weighted graph is used as an underlying structure of many algorithms like semi-supervised learning and spectral clustering. The edge weights are usually determined by a single similarity measure, but it often hard if not impossible to capture all relevant aspects of similarity when using a single similarity measure. In par-ticular, in the case of visual object matching it is beneficial to integrate different similarity measures that focus on different visual representations. In this paper, a novel approach to integrate multiple similarity measures is pro-posed. First pairs of similarity measures are combined with a diffusion process on their tensor product graph (TPG). Hence the diffused similarity of each pair of ob-jects becomes a function of joint diffusion of the two original similarities, which in turn depends on the neighborhood structure of the TPG. We call this process Fusion with Diffusion (FD). However, a higher order graph like the TPG usually means significant increase in time complexity. This is not the case in the proposed approach. A key feature of our approach is that the time complexity of the dif-fusion on the TPG is the same as the diffusion process on each of the original

graphs, Moreover, it is not necessary to explicitly construct the TPG in our frame-work. Finally all diffused pairs of similarity measures are combined as a weighted sum. We demonstrate the advantages of the proposed approach on the task of visual tracking, where different aspects of the appearance similarity between the target object in frame t and target object candidates in frame t+1 are integrated. The obtained method is tested on several challenge video sequences and the experimental results show that it outperforms state-of-the-art tracking methods.

M76 Learning visual motion in recurrent neural networks

Marius Pachitariu marius@gatsby.ucl.ac.uk
Maneesh Sahani maneesh@gatsby.ucl.ac.uk
Gatsby Computational Neuroscience Unit

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

M77 Burn-in, bias, and the rationality of anchoring

Falk Lieder falk.lieder@gmail.com
ITET, ETH Zurich
Tom Griffiths tom_griffiths@berkeley.edu
University of California, Berkeley
Noah Goodman ngoodman@stanford.edu
Stanford University

Bayesian inference provides a unifying framework for addressing problems in machine learning, artificial intelligence, and robotics, as well as the problems facing the human mind. Unfortunately, exact Bayesian inference is intractable in all but the simplest models. Therefore minds and machines have to approximate Bayesian inference. Approximate inference algorithms can achieve a wide range of time-accuracy tradeoffs, but what is the optimal tradeoff? We investigate time-accuracy tradeoffs using the Metropolis-Hastings algorithm as a metaphor for the mind's inference algorithm(s). We find that reasonably accurate decisions are possible long before the Markov chain has converged to the posterior distribution, i.e. during the period known as burn-in. Therefore the strategy that is optimal subject to the mind's bounded processing speed and opportunity costs may perform so few iterations that the resulting samples are biased towards the initial value. The resulting cognitive process model provides a rational basis for the anchoring-and-adjustment heuristic. The model's quantitative predictions are tested against published data on anchoring in numerical estimation tasks. Our theoretical and empirical results suggest that the anchoring bias is consistent with approximate Bayesian inference.

M78 On the connections between saliency and tracking

Vijay Mahadevan vijay.mahadevan@gmail.com
Labs, Yahoo!
Nuno Vasconcelos nuno@ece.ucsd.edu
UC San Diego

A model connecting visual tracking and saliency has recently been proposed. This model is based on the saliency hypothesis for tracking which postulates that tracking is achieved by the top-down tuning, based on target features, of discriminant center-surround saliency mechanisms over time. In this work, we identify three main predictions that must hold if the hypothesis were true: 1) tracking reliability should be larger for salient than for non-salient targets, 2) tracking reliability should have a dependence on the defining variables of saliency, namely feature contrast and distractor heterogeneity, and must replicate the dependence of saliency on these variables, and 3) saliency and tracking can be implemented with common low level neural mechanisms. We confirm that the first two predictions hold by reporting results from a set of human behavior studies on the connection between saliency and tracking. We also show that the third prediction holds by constructing a common neurophysiologically plausible architecture that can computationally solve both saliency and tracking. This architecture is fully compliant with the standard physiological models of V1 and MT, and with what is known about attentional control in area LIP, while explaining the results of the human behavior experiments.

M79 Action-Model Based Multi-agent Plan Recognition

Hankz Hankui Zhuo zhuohank@gmail.com
Sun Yat-sen University
Qiang Yang qyang@cse.ust.hk
Hong Kong University of Science and Technology
Subbarao Kambhampati rao@asu.edu

Multi-Agent Plan Recognition (MAPR) aims to recognize dynamic team structures and team behaviors from the observed team traces (activity sequences) of a set of intelligent agents. Previous MAPR approaches required a library of team activity sequences (team plans) be given as input. However, collecting a library of team plans to ensure adequate coverage is often difficult and costly. In this paper, we relax this constraint, so that team plans are not required to be provided beforehand. We assume instead that a set of action models are available. Such models are often already created to describe domain physics; i.e., the preconditions and effects of effects actions. We propose

a novel approach for recognizing multi-agent team plans based on such action models rather than libraries of team plans. We encode the resulting MAPR problem as a \emph{satisfiability problem} and solve the problem using a state-of-the-art weighted MAX-SAT solver. Our approach also allows for incompleteness in the observed plan traces. Our empirical studies demonstrate that our algorithm is both effective and efficient in comparison to state-of-the-art MAPR methods based on plan libraries.

M80 Rational inference of relative preferences

Nisheeth Srivastava nisheeths@gmail.com
Paul Schrater schrater@umn.edu
University of Minnesota

Statistical decision theory axiomatically assumes that the relative desirability of different options that humans perceive is well described by assigning them option-specific scalar utility functions. However, this assumption is refuted by observed human behavior, including studies wherein preferences have been shown to change systematically simply through variation in the set of choice options presented. In this paper, we show that interpreting desirability as a relative comparison between available options at any particular decision instance results in a rational theory of value-inference that explains heretofore intractable violations of rational choice behavior in human subjects. Complementarily, we also characterize the conditions under which a rational agent selecting optimal options indicated by dynamic value inference in our framework will behave identically to one whose preferences are encoded using a static ordinal utility function.

M81 Why MCA? Nonlinear Spike-and-slab Sparse Coding for Neurally Plausible Image Encoding

Jacquelyn Shelton shelton@fias.uni-frankfurt.de
Philip Sterne sterne@fias.uni-frankfurt.de
Abdul Saboor Sheikh sheikh@fias.uni-frankfurt.de
Jorg Bornschein bornschein@fias.uni-frankfurt.de
Jorg Lucke luecke@fias.uni-frankfurt.de
Frankfurt Institute for Advanced Studies

Modelling natural images with sparse coding (SC) has faced two main challenges: flexibly representing varying pixel intensities and realistically representing low-level components, e.g. edges. This paper proposes a novel multiple-cause generative model of low-level image statistics that generalizes the standard SC model in two crucial points: (1) it uses a spike-and-slab prior distribution for a more realistic representation of component absence/intensity, and (2) the model uses the highly nonlinear combination rule of maximal causes analysis (MCA). The major challenge is parameter optimization because a model with either (1) or (2) results in a strongly multimodal posterior. We show for the first time that a model combining both improvements can be trained efficiently while retaining the rich structure of the posterior. We design an exact piecewise Gibbs sampling method and combine this with a variational method based on preselection of latent dimensions. This combined training scheme tackles both analytical and computational intractability and enables application of the model to a large number of observed and hidden dimensions. Applying the model to image patches we study the optimal encoding of images by simple cells in V1 and compare the model's predictions with in vivo neural recordings. In contrast to standard SC, we find that the optimal prior favors asymmetric, bimodal, and sparse activity of simple cells. Testing our model for consistency we find that the average posterior is approximately equal to the prior. Furthermore, due to the nonlinearity, the model predicts a large number of globular receptive fields (RFs), another significant difference from standard SC. The inferred prior and the high proportion of predicted globular fields make the model more consistent with neural data than previous SC models, suggesting closer tuning of simple cells to visual stimuli than has been predicted until now.

M82 A System for Predicting Action Content On-Line and in Real Time before Action Onset in Humans – an Intracranial Study

Uri Maoz urim@cs.huji.ac.il
Shengxuan Ye sye@caltech.edu
Caltech
Ian Ross ianrossmd@aol.com
Huntington Memorial Hospital
Adam Mamelak adam.mamelak@cshs.org
Cedars Sinai Medical Center
Christof Koch koch.christof@gmail.com
Allen Institute for Brain Science

The ability to predict action content from neural signals in real time before action onset has been long sought in the neuroscientific study of decision-making, agency and volition. On-line real-time (ORT) prediction is important for understanding the relation between neural correlates of decision-making and conscious, voluntary action. Here, epilepsy patients, implanted with intracranial depth microelectrodes or subdural grid electrodes for clinical purposes, participated in a "matching-pennies" game against either the experimenter or a computer. In each trial, subjects were given a 5s countdown, after which they had to raise their left or right hand immediately as the "go" signal appeared on a computer screen. They won a fixed amount of money if they raised a different hand than their opponent and lost that amount otherwise. The working hypothesis of this experiment was that neural precursors of the subjects' decisions precede action onset and potentially also the awareness of the decision to move, and that these signals could be detected in intracranial local field potentials (LFP). We found that low-frequency LFP signals from a combination of 10 channels, especially bilateral anterior cingulate cortex and supplementary motor area, were predictive of the intended left-/right-hand movements before the onset of the go signal. Our ORT system predicted which hand the patient would raise 0.5s before the go signal with 68±3% accuracy in two patients. Based on these results, we constructed an ORT system that tracked up to 30 channels simultaneously, and tested it on retrospective data from 6 patients. On average, we could predict the correct hand choice in 80% of the

trials, which rose to 90% correct if we let the system drop about 1/3 of the trials on which it was less confident. Our system demonstrates – for the first time – the feasibility of accurately predicting a binary action in real time for patients with intracranial recordings, well before the action occurs.

M83 A lattice filter model of the visual pathway

Karol Gregor karol.gregor@gmail.com
NYU
Dmitri Chklovskii chklovskiid@janelia.hhmi.org
HHMI

Early stages of visual processing are thought to decorrelate, or whiten, the incoming temporally varying signals. Because the typical correlation time of natural stimuli, as well as the extent of temporal receptive fields of lateral geniculate nucleus (LGN) neurons, is much greater than neuronal time constants, such decorrelation must be done in stages combining contributions of multiple neurons. We propose to model temporal decorrelation in the visual pathway with the lattice filter, a signal processing device for stage-wise decorrelation of temporal signals. The stage-wise architecture of the lattice filter maps naturally onto the visual pathway (photoreceptors -> bipolar cells -> retinal ganglion cells -> LGN) and its filter weights can be learned using Hebbian rules in a stage-wise sequential manner. Moreover, predictions of neural activity from the lattice filter model are consistent with physiological measurements in LGN neurons and fruit fly second-order visual neurons. Therefore, the lattice filter model is a useful abstraction that may help unravel visual system function.

M84 Q-MKL: Matrix-induced Regularization in Multi-Kernel Learning with Applications to Neuroimaging

Chris Hinrichs hinrichs@cs.wisc.edu
Vikas Singh vsingh@biostat.wisc.edu
Sterling Johnson scj@medicine.wisc.edu
University of Wisconsin Madison
Jiming Peng pengj@uiuc.edu
University of Illinois

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

M85 Wavelet based multi-scale shape features on arbitrary surfaces for cortical thickness discrimination

Won Hwa Kim wonhwa@cs.wisc.edu
Moo. K Chung mkchung@wisc.edu
Sterling Johnson scj@medicine.wisc.edu
Deepti Pachauri pachauri@cs.wisc.edu
Vikas Singh vsingh@biostat.wisc.edu
Charles Hatt hatt@wisc.edu
University of Wisconsin - Madison

Hypothesis testing on signals defined on surfaces (such as the cortical surface) is a fundamental component of a variety of studies in Neuroscience. The goal here is to identify regions that exhibit changes as a function of the clinical condition under study. As the clinical questions of interest move towards identifying very early signs of diseases, the corresponding statistical differences at the group level invariably become weaker and increasingly hard to identify. Indeed, after a multiple comparisons correction is adopted (to account for correlated statistical tests over all surface points), very few regions may survive. In contrast to hypothesis tests on point-wise measurements, in this paper, we make the case for performing statistical analysis on multi-scale shape descriptors that characterize the local topological context of the signal around each surface vertex. Our descriptors are based on recent results from harmonic analysis, that show how wavelet theory extends to non-Euclidean settings (i.e., irregular weighted graphs). We provide strong evidence that these descriptors successfully pick up group-wise differences, where traditional methods either fail or yield unsatisfactory results. Other than this primary application, we show how the framework allows performing cortical surface smoothing in the native space without mappint to a unit sphere.

M86 A P300 BCI for the Masses: Prior Information Enables Instant Unsupervised Spelling

Pieter-Jan Kindermans Pieterjan.Kindermans@ugent.be
Hannes Verschore hv1989@gmail.com
David Verstraeten David.Verstraeten@ugent.be
Benjamin Schrauwen Benjamin.Schrauwen@ugent.be
Ghent University

The usability of Brain Computer Interfaces (BCI) based on the P300 speller is severely hindered by the need for long training times and many repetitions of the same stimulus. In this contribution we introduce a set of unsupervised hierarchical probabilistic models that tackle both problems simultaneously by incorporating prior knowledge from two sources: information from other training subjects (through transfer learning) and information about the words being spelled (through language models). We show, that due to this prior knowledge, the performance of the unsupervised models parallels and in some cases even surpasses that of supervised models, while eliminating the tedious training session.

M87 Towards a learning-theoretic analysis of spike-timing dependent plasticity

David Balduzzi david.balduzzi@inf.ethz.ch
ETH Zurich
Michel Besserve michel.besserve@tuebingen.mpg.de
MPI for Intelligent Systems

This paper suggests a learning-theoretic perspective on how synaptic plasticity benefits global brain functioning. We introduce a model, the selectron, that (i) arises as the fast time constant limit of leaky integrate-and-fire neurons equipped with spiking timing dependent plasticity (STDP) and (ii) is amenable to theoretical analysis. We show that the selectron encodes reward estimates into spikes and that an error bound on spikes is controlled by a spiking margin and the sum of synaptic weights. Moreover, the efficacy of spikes (their usefulness to other reward maximizing selectrons) also depends on total synaptic strength. Finally, based on our analysis, we propose a regularized version of STDP, and show the regularization improves the robustness of neuronal learning when faced with multiple stimuli.

M88 Synchronization can Control Regularization in Neural Systems via Correlated Noise Processes

Jake Bouvrie jvb@csail.mit.edu
Jean-Jacques Slotine jjs@mit.edu
Massachusetts Institute of Technology

To learn reliable rules that can generalize to novel situations, the brain must be capable of imposing some form of regularization. Here we suggest, through theoretical and computational arguments, that the combination of noise with synchronization provides a plausible mechanism for regularization in the nervous system. The functional role of regularization is considered in a general context in which coupled computational systems receive inputs corrupted by correlated noise. Noise on the inputs is shown to impose regularization, and when synchronization upstream induces time-varying correlations across noise variables, the degree of regularization can be calibrated over time. The resulting qualitative behavior matches experimental data from visual cortex.

M89 Homeostatic plasticity in Bayesian spiking networks as Expectation Maximization with posterior constraints

Stefan Habenschuss habenschuss@igi.tugraz.at
Johannes Bill bill@igi.tugraz.at
Bernhard Nessler nessler@igi.tugraz.at
Graz University of Technology

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

M90 Neurally Plausible Reinforcement Learning of Working Memory Tasks

Jaldert Rombouts J.O.Rombouts@cwi.nl
Sander Bohte nips@bohte.com
Centrum Wiskunde Informatica
Pieter Roelfsema p.roelfsema@nin.knaw.nl
Netherlands Institute for Neuroscience

A key function of brains is undoubtedly the abstraction and maintenance of information from the environment for later use. Neurons in association cortex play an important role in this process: during learning these neurons become tuned to relevant features and represent the information that is required later as a persistent elevation of their activity. It is however not well known how these neurons acquire their task-relevant tuning. Here we introduce a biologically plausible learning scheme that explains how neurons

become selective for relevant information when animals learn by trial and error. We propose that the action selection stage feeds back attentional signals to earlier processing levels. These feedback signals interact with feedforward signals to form synaptic tags at those connections that are responsible for the stimulus-response mapping. A globally released neuromodulatory signal interacts with these tagged synapses to determine the sign and strength of plasticity. The learning scheme is generic because it can train networks in different tasks, simply by varying inputs and rewards. It explains how neurons in association cortex learn to (1) temporarily store task-relevant information in non-linear stimulus-response mapping tasks and (2) learn to optimally integrate probabilistic evidence for perceptual decision making.

M91 Spiking and saturating dendrites differentially expand single neuron computation capacity.

Romain Cazé romain.caze@ens.fr
ENS
Mark Humphries m.d.humphries@sheffield.ac.uk
INSERM U960
Boris Gutkin boris.gutkin@ens.fr
ENS, Paris, France

The integration of excitatory inputs in dendrites is non-linear: multiple excitatory inputs can produce a local depolarization departing from the arithmetic sum of each input's response taken separately. If this depolarization is bigger than the arithmetic sum, the dendrite is spiking; if the depolarization is smaller, the dendrite is saturating. Decomposing a dendritic tree into independent dendritic spiking units greatly extends its computational capacity, as the neuron then maps onto a two layer neural network, enabling it to compute linearly non-separable Boolean functions (lnBFs). How can these lnBFs be implemented by dendritic architectures in practise? And can saturating dendrites equally expand computational capacity? To adress these questions we use a binary neuron model and Boolean algebra. First, we confirm that spiking dendrites enable a neuron to compute lnBFs using an architecture based on the disjunctive normal form (DNF). Second, we prove that saturating dendrites as well as spiking dendrites also enable a neuron to compute lnBFs using an architecture based on the conjunctive normal form (CNF). Contrary to the DNF-based architecture, a CNF-based architecture leads to a dendritic unit tuning that does not imply the neuron tuning, as has been observed experimentally. Third, we show that one cannot use a DNF-based architecture with saturating dendrites. Consequently, we show that an important family of lnBFs implemented with a CNF-architecture can require an exponential number of saturating dendritic units, whereas the same family implemented with either a DNF-architecture or a CNF-architecture always require a linear number of spiking dendritic unit. This minimization could explain why a neuron spends energetic resources to make its dendrites spike.

M92 Coding efficiency and detectability of rate fluctuations with non-Poisson neuronal firing

Shinsuke Koyama skoyama@ism.ac.jp
The Institute of Statistical Mathematics

Statistical features of neuronal spike trains are known to be non-Poisson. Here, we investigate the extent to which the non-Poissonian feature affects the efficiency of transmitting information on fluctuating firing rates. For this purpose, we introduce the Kullbuck-Leibler (KL) divergence as a measure of the efficiency of information encoding, and assume that spike trains are generated by time-rescaled renewal processes. We show that the KL divergence determines the lower bound of the degree of rate fluctuations below which the temporal variation of the firing rates is undetectable from sparse data. We also show that the KL divergence, as well as the lower bound, depends not only on the variability of spikes in terms of the coefficient of variation, but also significantly on the higher-order moments of interspike interval (ISI) distributions. We examine three specific models that are commonly used for describing the stochastic nature of spikes (the gamma, inverse Gaussian (IG) and lognormal ISI distributions), and find that the time-rescaled renewal process with the IG distribution achieves the largest KL divergence, followed by the lognormal and gamma distributions.

M93 Efficient coding connects prior and likelihood function in perceptual Bayesian inference

Xue-Xin Wei weixxpku@gmail.com
Alan Stocker astocker@sas.upenn.edu
University of Pennsylvania

A common challenge for Bayesian approaches in modeling perceptual behavior is the fact that the two fundamental components of a Bayesian model, the prior distribution and the likelihood function, are formally unconstrained. Here we argue that a neural system that emulates Bayesian inference naturally imposes constraints by way of how it represents sensory information in populations of neurons. More specifically, we propose an efficient encoding principle that constrains both the likelihood and the prior based on low-level environmental statistics. The resulting Bayesian estimates can show biases away from the peaks of a prior distribution, a behavior seemingly at odds with the traditional view of Bayesian estimates yet one that has indeed been reported in human perception of visual orientation. We demonstrate that our framework correctly predicts these biases, and show that the efficient encoding characteristics of the model neural population matches the reported orientation tuning characteristics of neurons in primary visual cortex. Our results suggest that efficient coding can be a promising hypothesis in constraining neural implementations of Bayesian inference.

TUESDAY

ORAL SESSION

SESSION 1 - 9:00 – 10:10 AM

Session Chair: Raquel Urtasun

INVITED TALK: Quantum information and the Brain

Scott Aaronson aaronson@csail.mit.edu
Massachusetts Institute of Technology

Ever since quantum mechanics was discovered nearly a century ago, famous scientists from Eddington to Wigner to Compton to Eccles to Penrose have speculated about possible connections to the brain – a quest often parodied as "quantum mechanics is mysterious, the brain is mysterious, ergo they must be related somehow." In this talk, I'll offer a critical survey of these ideas from the modern standpoint of quantum information theory, pointing out the huge conceptual and experimental problems that have plagued most concrete proposals. However, I'll also explain why I think some role for quantum mechanics in cognition is not yet excluded, and discuss what sorts of advances in neuroscience and physics might help settle the question.

Scott Aaronson *received his PhD in computer science from UC Berkeley in 2004, then did postdocs at the Institute for Advanced Study, Princeton and at the University of Waterloo. He's now TIBCO Career Development Associate Professor at MIT. His research interests center around the limitations of quantum computers, and computational complexity theory more generally. He also writes a popular blog (www.scottaaronson.com/blog) and created the Complexity Zoo (www.complexityzoo.com), an online encyclopedia of computational complexity theory. He received the National Science Foundation's Alan T. Waterman Award in 2012.*

TCA: High Dimensional Principal Component Analysis for non-Gaussian Data

Fang Han fhan@jhsph.edu
Johns Hopkins University
Han Liu hanliu@cs.jhu.edu
Princeton University

We propose a high dimensional semiparametric scale-invariant principle component analysis, named TCA, by utilize the natural connection between the elliptical distribution family and the principal component analysis. Elliptical distribution family includes many well-known multivariate distributions like multivariate t and logistic and it is extended to the meta-elliptical by Fang (2002) using the copula techniques. In this paper we extend the meta-elliptical distribution family to a even larger family, called transelliptical. We prove that TCA can obtain a near-optimal s(log d/n)^{1/2} estimation consistency rate in the transelliptical distribution family, even if the distributions are very heavy-tailed, have infinite second moments, do not have densities and possess arbitrarily continuous marginal distributions. A feature selection result with explicit rate is also provided. TCA is also implemented in both numerical simulations and large-scale stock data to illustrate its empirical performance. Both theories and experiments confirm that TCA can achieve model flexibility, estimation accuracy and robustness at almost no cost.

SPOTLIGHT SESSION

SESSION 1 - 10:10 – 10:30 AM

- Putting Bayes to sleep
  W. Koolen, D. Adamskiy, Royal Holloway, University of London; M. Warmuth, Univ. of Calif. at Santa Cruz
  See abstract T57, page 62

- Confusion-Based Online Learning and a Passive-Aggressive Scheme
  L. Ralaivola, Aix-Marseille University
  See abstract W77, page 100

- Volume Regularization for Binary Classification
  K. Crammer, The Technion; T. Wagner, Weizmann Institute
  See abstract T46, page 59

- Fast Resampling Weighted v-Statistics
  C. Zhou, NIH; j. Park, ; Y. Fu, Northeastern University
  See abstract T49, page 60

- Pointwise Tracking the Optimal Regression Function
  Y. Wiener, R. El-Yaniv, Technion
  See abstract Th56, page 124

ORAL SESSION

SESSION 2 - 11:00 – 11:40 AM

Session Chair: Sebastien Bubeck

Spectral Learning of General Weighted Automata via Constrained Matrix Completion

Borja Balle bballe@lsi.upc.edu
Universitat Politecnica de Catalunya (UPC)
Mehryar Mohri mohri@google.com
Courant Institute & Google Research

Many tasks in text and speech processing and computational biology involve functions from variable-length strings to real numbers. A wide class of such functions can be computed by weighted automata. Spectral methods based on singular value decompositions of Hankel matrices have been recently proposed for learning probability distributions over strings that can be computed by weighted automata. In this paper we show how this method can be applied to the problem of learning a general weighted automata from a sample of string-label pairs generated by an arbitrary distribution. The main obstruction to this approach is that in general some entries of the Hankel matrix that needs to be decomposed

may be missing. We propose a solution based on solving a constrained matrix completion problem. Combining these two ingredients, a whole new family of algorithms for learning general weighted automata is obtained. Generalization bounds for a particular algorithm in this class are given. The proofs rely on a stability analysis of matrix completion and spectral learning.

Relax and Randomize : From Value to Algorithms

Sasha Rakhlin rakhlin@gmail.com
Karthik Sridharan karthik.sridharan@gmail.com
University of Pennsylvania
Ohad Shamir ohadshamir@gmail.com
Microsoft Research

We show a principled way of deriving online learning algorithms from a minimax analysis. Various upper bounds on the minimax value, previously thought to be non-constructive, are shown to yield algorithms. This allows us to seamlessly recover known methods and to derive new ones, also capturing such ''unorthodox'' methods as Follow the Perturbed Leader and the R^2 forecaster. Understanding the inherent complexity of the learning problem thus leads to the development of algorithms. To illustrate our approach, we present several new algorithms, including a family of randomized methods that use the idea of a ''random play out''. New versions of the Follow-the-Perturbed-Leader algorithms are presented, as well as methods based on the Littlestone's dimension, efficient methods for matrix completion with trace norm, and algorithms for the problems of transductive learning and prediction with static experts.

SPOTLIGHT SESSION

SESSION 2 - 11:40 – 12:05 AM

- **A Convex Formulation for Learning Scale-Free Networks via Submodular Relaxation**
  A. Defazio, ANU; T. Caetano, NICTA/ANU
  See abstract W47, page 92

- **Active Comparison of Prediction Models**
  C. Sawade, N. Landwehr, University of Potsdam; T. Scheffer, Universität Potsdam
  See abstract T25, page 54

- **Entangled Monte Carlo**
  S. Jun, A. Bouchard-Côté, UBC; L. Wang, University of Western Ontario
  See abstract T28, page 55

- **Modelling Reciprocating Relationships**
  C. Blundell, Gatsby Unit, UCL; K. Heller, Duke University; J. Beck, University of Rochester
  See abstract W40, page 91

- **Continuous Relaxations for Discrete Hamiltonian Monte Carlo**
  Z. Ghahramani, University of Cambridge; Y. Zhang, C. Sutton, A. Storkey, University of Edinburgh
  See abstract T42, page 58

- **Efficient high dimensional maximum entropy modeling via symmetric partition functions**
  P. Vernaza, D. Bagnell, Carnegie Mellon University
  See abstract, page 48 (W51)

ORAL SESSION

SESSION 3 - 2:00 – 3:30 AM

Session Chair: Lin Xiao

INVITED TALK: Classification with Deep Invariant Scattering Networks

Stephane Mallat mallat@cmap.polytechnique.fr
Ecole Normale Supérieure

High-dimensional data representation is in a confused infancy compared to statistical decision theory. How to optimize kernels or so called feature vectors? Should they increase or reduce dimensionality? Surprisingly, deep neural networks have managed to build kernels accumulating experimental successes. This lecture shows that invariance emerges as a central concept to understand high-dimensional representations, and deep network mysteries.

Intra-class variability is the curse of most high-dimensional signal classifications. Fighting it means finding informative invariants. Standard mathematical invariants are either non-stable for signal classification or not sufficiently discriminative. We explain how convolution networks compute stable informative invariants over any group such as translations, rotations or frequency transpositions, by scattering data in high dimensional spaces, with wavelet filters. Beyond groups, invariants over manifolds can also be learned with unsupervised strategies that involve sparsity constraints. Applications will be discussed and shown on images and sounds.

Stéphane Mallat *received the Ph.D. degree in electrical engineering from the University of Pennsylvania, in 1988. He was then Professor at the Courant Institute of Mathematical Sciences. In 1995, he became Professor in Applied Mathematics at Ecole Polytechnique, Paris. From 2001 to 2007 he was co-founder and CEO of a semiconductor start-up company. In 2012 he joined the Computer Science Department of Ecole Normale Supérieure, in Paris.*

Stéphane Mallat's research interests include signal processing, computer vision, harmonic analysis and learning. He wrote a "Wavelet tour of signal processing: the sparse way". In 1997, he received the Outstanding Achievement Award from the SPIE Society and was a plenary lecturer at the International Congress of Mathematicians in 1998. He also received the 2004 European IST Grand prize, the 2004 INIST-CNRS prize for most cited French researcher in engineering and computer science, and the 2007 EADS grand prize of the French Academy of Sciences.

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

Nicolas Le Roux nicolas@le-roux.name
Criteo
Mark Schmidt mark.schmidt@inria.fr
NRIA - SIERRA Project Team
Francis Bach francis.bach@mines.org
Ecole Normale Superieure

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

Approximating Concavely Parameterized Optimization Problems

Joachim Giesen joachim.giesen@uni-jena.de
Jens Mueller jkm@informatik.uni-jena.de
Soeren Laue soeren.laue@uni-jena.de
Sascha Swiercy sascha.swiercy@googlemail.com
Friedrich Schiller University Jena

We consider an abstract class of optimization problems that are parameterized concavely in a single parameter, and show that the solution path along the parameter can always be approximated with accuracy $\varepsilon>0$ by a set of size $O(1/\varepsilon)$. A lower bound of size $\Omega(1/\varepsilon)$ shows that the upper bound is tight up to a constant factor. We also devise an algorithm that calls a step-size oracle and computes an approximate path of size $O(1/\varepsilon)$. Finally, we provide an implementation of the oracle for soft-margin support vector machines, and a parameterized semi-definite program for matrix completion.

SPOTLIGHT SESSION

SESSION 3 - 3:30 – 3:50 AM

- **A Quasi-Newton Proximal Splitting Method**
  S. Becker, Paris-6/CNRS; J. Fadili, CNRS-ENSICAEN-Univ. Caen
  See abstract Th35, page 120

- **Factoring nonnegative matrices with linear programs**
  B. Recht, UW-Madison; C. Re, University of Wisconsin; J. Tropp, California Institute of Technology; V. Bittorf, University of Wisconsin–Madison
  See abstract Th30, page 118

- **Scaled Gradients on Grassmann Manifolds for Matrix Completion**
  T. Ngo, Y. Saad, University of Minnesota
  See abstract T13, page 51

- **Multi-criteria Anomaly Detection using Pareto Depth Analysis**
  K. Hsiao, K. Xu, J. Calder, A. Hero, University of Michigan
  See abstract T68, page 64

- **Semi-Supervised Domain Adaptation with Non-Parametric Copulas**
  D. Lopez-Paz, B. Schölkopf, Max Planck Institute for Intelligent Systems; J. Hernández-Lobato, Cambridge University
  See abstract T11, page 51

ORAL SESSION

SESSION 4 - 4:20 – 5:40 AM

Session Chair: Gunnar Raetsch

Spectral learning of linear dynamics from generalised-linear observations with application to neural population data

Lars Buesing lars@gatsby.ucl.ac.uk
Maneesh Sahani maneesh@gatsby.ucl.ac.uk
University College London
Jakob Macke Jakob.Macke@gmail.com
Max Planck Institute and Bernstein Center Tübingen

Latent linear dynamical systems with generalised-linear observation models arise in a variety of applications, for example when modelling the spiking activity of populations of neurons. Here, we show how spectral learning methods for linear systems with Gaussian observations (usually called subspace identification in this context) can be extended to estimate the parameters of dynamical system models observed through non-Gaussian noise models. We use this approach to obtain estimates of parameters for a dynamical model of neural population data, where the observed spike-counts are Poisson-distributed with log-rates determined by the latent dynamical process, possibly driven by external inputs. We show that the extended system identification algorithm is consistent and accurately recovers the correct parameters on large simulated data sets with much smaller computational cost than approximate expectation-maximisation (EM) due to the non-iterative nature of subspace identification. Even on smaller data sets, it provides an effective initialization for EM, leading to more robust performance and faster convergence. These benefits are shown to extend to real neural data.

High-Order Multi-Task Feature Learning to Identify Longitudinal Phenotypic Markers for Alzheimer Disease Progression Prediction

Hua Wang huawangcs@gmail.com
Feiping Nie feipingnie@gmail.com
Heng Huang heng@uta.edu
University of Texas Arlington
Jingwen Yan jingyan@iupui.edu
Sungeun Kim sk31@iupui.edu
Shannon Risacher srisache@iupui.edu
Andrew Saykin asaykin@iupui.edu
Li Shen shenli@iupui.edu
Indiana University School of Medicine

Alzheimer disease (AD) is a neurodegenerative disorder characterized by progressive impairment of memory and other cognitive functions. Regression analysis has been studied to relate neuroimaging measures to cognitive status. However, whether these measures have further predictive power to infer a trajectory of cognitive performance over time is still an under-explored but important topic in AD research. We propose a novel high-order multi-task learning model to address this issue. The proposed model explores the temporal correlations existing in data features and regression tasks by the structured sparsity-inducing norms. In addition, the sparsity of the model enables the selection of a small number of MRI measures while maintaining high prediction accuracy. The empirical studies, using the baseline MRI and serial cognitive data of the ADNI cohort, have yielded promising results.

Multimodal Learning with Deep Boltzmann Machines

Nitish Srivastava nitish@cs.toronto.edu
Russ Salakhutdinov rsalakhu@mit.edu
University of Toronto

We propose a Deep Boltzmann Machine for learning a generative model of multimodal data. We show how to use the model to extract a meaningful representation of multimodal data. We find that the learned representation is useful for classification and information retreival tasks, and hence conforms to some notion of semantic similarity. The model defines a probability density over the space of multimodal inputs. By sampling from the conditional distributions over each data modality, it possible to create the representation even when some data modalities are missing. Our experimental results on bi-modal data consisting of images and text show that the Multimodal DBM can learn a good generative model of the joint space of image and text inputs that is useful for information retrieval from both unimodal and multimodal queries. We further demonstrate that our model can significantly outperform SVMs and LDA on discriminative tasks. Finally, we compare our model to other deep learning methods, including autoencoders and deep belief networks, and show that it achieves significant gains.

Discriminative Learning of Sum-Product Networks

Robert Gens rcg@cs.washington.edu
Pedro Domingos pedrod@cs.washington.edu
University of Washington

Sum-product networks are a new deep architecture that can perform fast, exact inference on high-treewidth models. Only generative methods for training SPNs have been proposed to date. In this paper, we present the first discriminative training algorithms for SPNs, combining the high accuracy of the former with the representational power and tractability of the latter. We show that the class of tractable discriminative SPNs is broader than the class of tractable generative ones, and propose an efficient backpropagation-style algorithm for computing the gradient of the conditional log likelihood. Standard gradient descent suffers from the diffusion problem, but networks with many layers can be learned reliably using ''hard'' gradient descent, where marginal inference is replaced by MPE inference (i.e., inferring the most probable state of the non-evidence variables). The resulting updates have a simple and intuitive form. We test discriminative SPNs on standard image classification tasks. We obtain the best results to date on the CIFAR-10 dataset, using fewer features than prior methods with an SPN architecture that learns local image structure discriminatively. We also report the highest published test accuracy on STL-10 even though we only use the labeled portion of the dataset.

SPOTLIGHT SESSION

SESSION 4 - 5:40 – 6:00 PM

- **ImageNet Classification with Deep Convolutional Neural Networks**
  A. Krizhevsky, I. Sutskever, G. Hinton, University of Toronto
  See abstract Th25, page 117

- **Searching for objects driven by context**
  B. Alexe, ETH ZURICH; N. Heess, University College London; Y. Teh, University of Oxford; V. Ferrari, University of Edinburgh
  See abstract Th80, page 129

- **3D Object Detection and Viewpoint Estimation with a Deformable 3D Cuboid Model**
  S. Fidler, S. Dickinson, University of Toronto; R. Urtasun, TTI-Chicago
  See abstract Th82, page 68

- **Analyzing 3D Objects in Cluttered Images**
  M. Hejrati, University of California, Irvine; D. Ramanan,
  See abstract T79, page 67

- **Discriminatively Trained Sparse Code Gradients for Contour Detection**
  X. Ren, L. Bo, Intel Labs
  See abstract T78, page 67

POSTER SESSION

AND RECEPTION - 7:00 – 11:59 PM

T1 **Minimizing Uncertainty in Pipelines**
N. Dalvi, A. Parameswaran, V. Rastogi

T2 **Tensor Decomposition for Fast Parsing with Latent-Variable PCFGs**
S. Cohen, M. Collins

T3 **Causal discovery with scale-mixture model for spatiotemporal variance dependencies**
Z. Chen, K. Zhang, L. CHAN

T4 **Learning Partially Observable Models Using Temporally Abstract Decision Trees**
E. Talvitie

T5 **On-line Reinforcement Learning Using Incremental Kernel-Based Stochastic Factorization**
A. Barreto, D. Precup, J. Pineau

T6 **Multimodal Learning with Deep Boltzmann Machines**
N. Srivastava, R. Salakhutdinov

T7 **A Better Way to Pre-Train Deep Boltzmann Machines**
R. Salakhutdinov, G. Hinton

T8 **Emergence of Object-Selective Features in Unsupervised Feature Learning**
A. Coates, A. Karpathy, A. Ng

T9 **Deep Representations and Codes for Image Auto-Annotation**
R. Kiros, C. Szepesvari

T10 **High Dimensional Semiparametric Scale-invariant Principal Component Analysis**
F. Han, H. Liu

T11 **Semi-Supervised Domain Adaptation with Non-Parametric Copulas**
D. Lopez-Paz, J. Hernández-Lobato, B. Schölkopf

T12 **Finite Sample Convergence Rates of Zero-Order Stochastic Optimization Methods**
J. Duchi, M. Jordan, M. Wainwright, A. Wibisono

T13 **Scaled Gradients on Grassmann Manifolds for Matrix Completion**
T. Ngo, Y. Saad

T14 **Minimizing Sparse High-Order Energies by Submodular Vertex-Cover**
A. Delong, O. Veksler, A. Osokin, Y. Boykov

T15 **Accelerated Training for Matrix-norm Regularization: A Boosting Approach**
X. Zhang, Y. Yu, D. Schuurmans

T16 **Approximating Concavely Parameterized Optimization Problems**
J. Giesen, J. Mueller, S. Laue, S. Swiercy

T17 **Finding Exemplars from Pairwise Dissimilarities via Simultaneous Sparse Recovery**
E. Elhamifar, G. Sapiro, R. Vidal

T18 **Convex Multi-view Subspace Learning**
M. White, Y. Yu, X. Zhang, D. Schuurmans

T19 **A Polylog Pivot Steps Simplex Algorithm for Classification**
E. Hazan, Z. Karnin

T20 **Stochastic optimization and sparse statistical recovery: Optimal algorithms for high dimensions**
A. Agarwal, S. Negahban, M. Wainwright

T21 **Stochastic Gradient Descent with Only One Projection**
M. Mahdavi, T. Yang, R. Jin, S. Zhu

T22 **Optimal Regularized Dual Averaging Methods for Stochastic Optimization**
X. Chen, Q. Lin, J. Pena

T23 **A Stochastic Gradient Method with an Exponential Convergence Rate for Finite Training Sets**
N. Le Roux, M. Schmidt, F. Bach

T24 **On Lifting the Gibbs Sampling Algorithm**
D. Venugopal, V. Gogate

T25 **Active Comparison of Prediction Models**
C. Sawade, N. Landwehr, T. Scheffer

T26 **Multi-Task Averaging**
S. Feldman, M. Gupta, B. Frigyik

T27 **Multiplicative Forests for Continuous-Time Processes**
J. Weiss, S. Natarajan, D. Page

T28 **Entangled Monte Carlo**
S. Jun, L. Wang, A. Bouchard-Côté

T29 **Fast Bayesian Inference for Non-Conjugate Gaussian Process Regression**
M. Khan, S. Mohamed, K. Murphy

T30 **Globally Convergent Dual MAP LP Relaxation Solvers using Fenchel-Young Margins**
A. Schwing, T. Hazan, M. Pollefeys, R. Urtasun

T31 **Fast Variational Inference in the Conjugate Exponential Family**
J. Hensman, M. Rattray, N. Lawrence

T32 **Minimization of Continuous Bethe Approximations: A Positive Variation**
J. Pacheco, E. Sudderth

T33 **Perfect Dimensionality Recovery by Variational Bayesian PCA**
S. Nakajima, R. Tomioka, M. Sugiyama, S. Babacan

T34 **Expectation Propagation in Gaussian Process Dynamical Systems**
M. Deisenroth, S. Mohamed

T35 **Random function priors for exchangeable graphs and arrays**
J. Lloyd, D. Roy, P. Orbanz, Z. Ghahramani

T36 **Forward-Backward Activation Algorithm for Hierarchical Hidden Markov Models**
K. Wakabayashi, T. Miura

T37 **Active Learning of Model Evidence Using Bayesian Quadrature**
M. Osborne, D. Duvenaud, R. Garnett, C. Rasmussen, S. Roberts, Z. Ghahramani

T38 **Phoneme Classification using Constrained Variational Gaussian Process Dynamical System**
H. Park, J. Kim, S. Park, S. Yun, C. Yoo

T39 **Density Propagation and Improved Bounds on the Partition Function**
S. Ermon, C. Gomes, A. Sabharwal, B. Selman

T40 **High Dimensional Transelliptical Graphical Models**
H. Liu, F. Han

T41 **Scaling Constrained Continuous Markov Random Fields with Consensus Optimization**
S. Bach, L. Getoor, M. Broecheler

T42 **Continuous Relaxations for Discrete Hamiltonian Monte Carlo**
Z. Ghahramani, Y. Zhang, C. Sutton, A. Storkey

T43 **Calibrated Elastic Regularization in Matrix Completion**
C. Zhang, T. Sun

T44 **Latent Graphical Model Selection: Efficient Methods for Locally Tree-like Graphs**
A. Anandkumar, R. Valluvan

T45 **Slice Normalized Dynamic Markov Logic Networks**
T. Papai, H. Kautz, D. Stefankovic

T46 **Volume Regularization for Binary Classification**
K. Crammer, T. Wagner

T47 **Spectral Learning of General Weighted Automata via Constrained Matrix Completion**
B. Balle, M. Mohri

T48 **Learning Probability Measures with respect to Optimal Transport Metrics**
G. Canas, L. Rosasco

T49 **Fast Resampling Weighted v-Statistics**
C. Zhou, j. Park, Y. Fu

T50 **Approximating Equilibria in Sequential Auctions with Incomplete Information and Multi-Unit Demand**
A. Greenwald, J. Li, E. Sodomka

T51 **Interpreting prediction markets: a stochastic approach**
N. Della Penna, M. Reid, R. Frongillo

T52 **Active Learning of Multi-Index Function Models**
H. Tyagi, V. Cevher

T53 **Hierarchical Optimistic Region Selection driven by Curiosity**
O. Maillard

T54 **Risk-Aversion in Multi-armed Bandits**
A. Sani, A. Lazaric, R. Munos

T55 **Online allocation and homogeneous partitioning for piecewise constant mean-approximation**
A. Carpentier, O. Maillard

T56 **Adaptive Stratified Sampling for Monte-Carlo integration of Differentiable functions**
A. Carpentier, R. Munos

T57 **Putting Bayes to sleep**
W. Koolen, D. Adamskiy, M. Warmuth

T58 **Online Sum-Product Computation**
M. Herbster, F. Vitale, S. Pasteris

T59 **Learning with Target Prior**
Z. Wang, S. Lyu, G. Schalk, Q. Ji

T60 **Learning High-Density Regions for a Generalized Kolmogorov-Smirnov Test in High-Dimensional Data**
A. Glazer, M. Lindenbaoum, S. Markovitch

T61 **Learning with Partially Absorbing Random Walks**
X. Wu, Z. Li, S. Chang, J. Wright, A. So

T62 **Small-Variance Asymptotics for Exponential Family Dirichlet Process Mixture Models**
K. Jiang, B. Kulis, M. Jordan

T63 **Repulsive Mixtures**
F. PETRALIA, V. Rao, D. Dunson

T64 **Training sparse natural image models with a fast Gibbs sampler of an extended state space**
L. Theis, J. Sohl-Dickstein, M. Bethge

T65 **Provable ICA with Unknown Gaussian Noise, with Implications for Gaussian Mixtures and Autoencoders**
S. Arora, R. Ge, A. Moitra, S. Sachdeva

T66 **TCA: High Dimensional Principal Component Analysis for non-Gaussian Data**
F. Han, H. Liu

T67 **Matrix reconstruction with the local max norm**
R. Foygel, N. Srebro, R. Salakhutdinov

T68 **Multi-criteria Anomaly Detection using Pareto Depth Analysis**
K. Hsiao, K. Xu, J. Calder, A. Hero

T69 **The Perturbed Variation**
M. Harel, S. Mannor

T70 **A Geometric take on Metric Learning**
S. Hauberg, O. Freifeld, M. Black

T71 **Semi-supervised Eigenvectors for Locally-biased Learning**
T. Jansen Hansen, M. Mahoney

T72 **LUCID: Locally Uniform Comparison Image Descriptor**
A. Ziegler, E. Christiansen, D. Kriegman, S. Belongie

T73 **Learning to Align from Scratch**
G. Huang, M. Mattar, H. Lee, E. Learned-Miller

T74 **From Deformations to Parts: Motion-based Segmentation of 3D Objects**
S. Ghosh, E. Sudderth, M. Loper, M. Black

T75 **Max-Margin Structured Output Regression for Spatio-Temporal Action Localization**
D. Tran, J. Yuan

T76 **Unsupervised template learning for fine-grained object recognition**
S. Yang, L. Bo, J. Wang, L. Shapiro

T77 **A Generative Model for Parts-based Object Segmentation**
S. Eslami, C. Williams

T78 **Discriminatively Trained Sparse Code Gradients for Contour Detection**
X. Ren, L. Bo

T79 **Analyzing 3D Objects in Cluttered Images**
M. Hejrati, D. Ramanan

T80 **Timely Object Recognition**
S. Karayev, T. Baumgartner, M. Fritz, T. Darrell

T81 **Semantic Kernel Forests from Multiple Taxonomies**
S. Hwang, K. Grauman, F. Sha

T82 **3D Object Detection and Viewpoint Estimation with a Deformable 3D Cuboid Model**
S. Fidler, S. Dickinson, R. Urtasun

T83 **Localizing 3D cuboids in single-view images**
J. Xiao, B. Russell, A. Torralba

T84 **Kernel Latent SVM for Visual Recognition**
W. Yang, Y. Wang, A. Vahdat, G. Mori

T85 **Graphical Gaussian Vector for Image Categorization**
T. Harada

T86 **Learning about Canonical Views from Internet Image Collections**
E. Mezuman, Y. Weiss

T87 **Diffusion Decision Making for Adaptive k-Nearest Neighbor Classification**
Y. Noh, F. Park, D. Lee

T88 **Modeling the Forgetting Process using Image Regions**
A. Khosla, J. Xiao, A. Torralba, A. Oliva

T89 **Bayesian Hierarchical Reinforcement Learning**
F. Cao, S. Ray

T90 **Spectral learning of linear dynamics from generalised-linear observations with application to neural population data**
L. Buesing, J. Macke, M. Sahani

T91 **A systematic approach to extracting semantic information from functional MRI data**
F. Pereira, M. Botvinick

T92 **Fully Bayesian inference for neural models with negative-binomial spiking**
J. Pillow, J. Scott

T93 **Bayesian active learning with localized priors for fast receptive field characterization**
M. Park, J. Pillow

DEMONSTRATIONS

DEMONSTRATIONS 7:00 - 11:59 PM

1A **A Stochastic Spiking Network Model of Sensorimotor Control**, A. Ghoreyshi, T. Sanger, J. Rocamora

2A **EVA: Engine for Visual Annotation**, J. Deng, J. Krause, Z. Huang, A. Berg, F. Li

3A **Gait analysis using the Kinect sensor**, M. Gabel, E. Renshaw, A. Schuster, R. Gilad - Bachrach

4A **Gesture recognition with Kinect**, I. Guyon

5A **NCS: A Large-Scale Brain Simulator**, L. Jayet Bray, D. Tanna, F. Harris, Jr

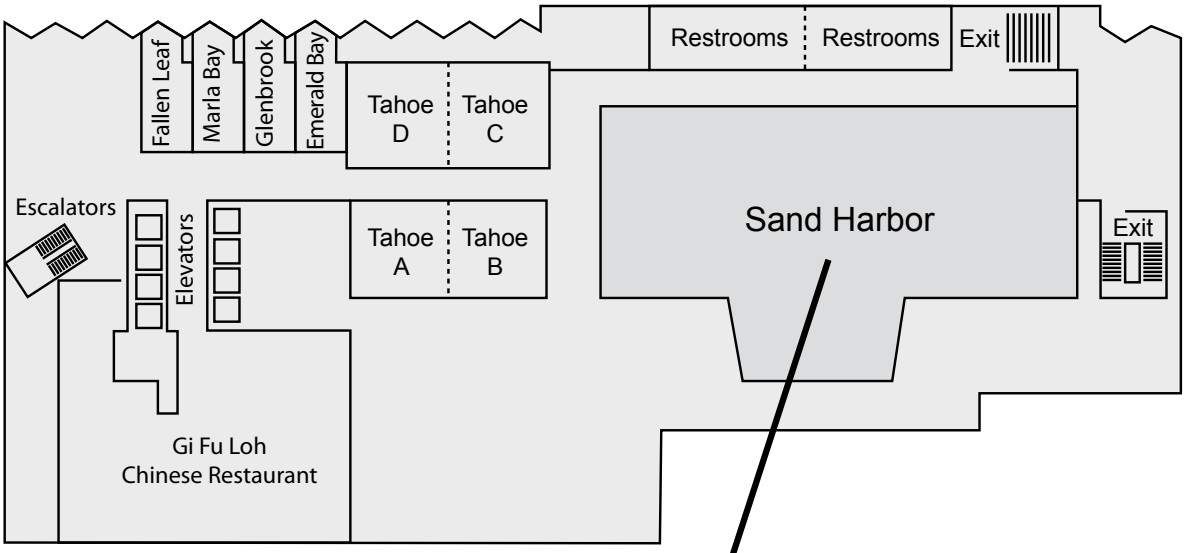
6A **Real-time Fusion/normalization of Multiple SVM with libMR**, T. Boult

7A **Ubiquitous Content: How musicians will search for every riff, musical phrase, and idea ever recorded**., J. LeBoeuf

TUESDAY POSTER FLOORPLAN

HARRAH'S

2ND FLOOR SPECIAL EVENTS CENTER



SAND HARBOR

	89	88	81	80	73	72	65	64	57	56	49
	90	87	82	79	74	71	66	63	58	55	50
	91	86	83	78	75	70	67	62	59	54	51
93	92	85	84	77	76	69	68	61	60	53	52

48	41	40	33	32	25	24	17	16	09	08	01
47	42	39	34	31	26	23	18	15	10	07	02
46	43	38	35	30	27	22	19	14	11	06	03
45	44	37	36	29	28	21	20	13	12	05	04

T1 Minimizing Uncertainty in Pipelines

Nilesh Dalvi nileshdalvi@gmail.com
Facebook
Aditya Parameswaran adityagp@cs.stanford.edu
Stanford University
Vibhor Rastogi vibhor.rastogi@gmail.com
Google

In this paper, we consider the problem of debugging large pipelines by human labeling. We represent the execution of a pipeline using a directed acyclic graph of AND and OR nodes, where each node represents a data item produced by some operator in the pipeline. We assume that each operator assigns a confidence to each of its output data. We want to reduce the uncertainty in the output by issuing queries to a human expert, where a query consists of checking if a given data item is correct. In this paper, we consider the problem of asking the optimal set of queries to minimize the resulting output uncertainty. We perform a detailed evaluation of the complexity of the problem for various classes of graphs. We give efficient algorithms for the problem for trees, and show that, for a general dag, the problem is intractable.

T2 Tensor Decomposition for Fast Parsing with Latent-Variable PCFGs

Shay Cohen scohen@cs.columbia.edu
Michael Collins mcollins@cs.columbia.edu
Columbia University

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

T3 Causal discovery with scale-mixture model for spatiotemporal variance dependencies

Zhitang Chen ztchen@cse.cuhk.edu.hk
Laiwan CHAN lwchan@cse.cuhk.edu.hk
The Chinese University of Hong Kong
Kun Zhang kun.zhang@tuebingen.mpg.de
Max Planck Institute for Biological Cybernetics

In conventional causal discovery, structural equation models (SEM) are directly applied to the observed variables, meaning that the causal effect can be represented as a function of the direct causes themselves. However, in many real world problems, there are significant dependencies in the variances or energies, which indicates that causality may possibly take place at the level of variances or energies. In this paper, we propose a probabilistic causal scale-mixture model with spatiotemporal variance dependencies to represent a specific type of generating mechanism of the observations. In particular, the causal mechanism including contemporaneous and temporal causal relations in variances or energies is represented by a Structural Vector AutoRegressive model (SVAR). We prove the identifiability of this model under the non-Gaussian assumption on the innovation processes. We also propose algorithms to estimate the involved parameters and discover the contemporaneous causal structure. Experiments on synthesis and real world data are conducted to show the applicability of the proposed model and algorithms.

T4 Learning Partially Observable Models Using Temporally Abstract Decision Trees

Erik Talvitie erik.talvitie@fandm.edu
Franklin & Marshall College

This paper introduces timeline trees, which are partial models of partially observable environments. Timeline trees are given some specific predictions to make and learn a decision tree over history. The main idea of timeline trees is to use temporally abstract features to identify and split on features of key events, spread arbitrarily far apart in the past (whereas previous decision-tree-based methods have been limited to a finite suffix of history). Experiments demonstrate that timeline trees can learn to make high quality predictions in complex, partially observable environments with high-dimensional observations (e.g. an arcade game).

T5 On-line Reinforcement Learning Using Incremental Kernel-Based Stochastic Factorization

Andre Barreto amsb@cs.mcgill.ca
Doina Precup dprecup@cs.mcgill.ca
Joelle Pineau jpineau@cs.mcgill.ca
School of McGill University

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

T6 Multimodal Learning with Deep Boltzmann Machines

Nitish Srivastava nitish@cs.toronto.edu
Russ Salakhutdinov rsalakhu@mit.edu
University of Toronto

We propose a Deep Boltzmann Machine for learning a generative model of multimodal data. We show how to use the model to extract a meaningful representation of multimodal data. We find that the learned representation is useful for classification and information retreival tasks, and hence conforms to some notion of semantic similarity. The model defines a probability density over the space of multimodal inputs. By sampling from the conditional distributions over each data modality, it possible to create the representation even when some data modalities are missing. Our experimental results on bi-modal data consisting of images and text show that the Multimodal DBM can learn a good generative model of the joint space of image and text inputs that is useful for information retrieval from both unimodal and multimodal queries. We further demonstrate that our model can significantly outperform SVMs and LDA on discriminative tasks. Finally, we compare our model to other deep learning methods, including autoencoders and deep belief networks, and show that it achieves significant gains.

T7 A Better Way to Pre-Train Deep Boltzmann Machines

Russ Salakhutdinov rsalakhu@mit.edu
Geoffrey Hinton hinton@cs.toronto.edu
University of Toronto

We describe how the pre-training algorithm for Deep Boltzmann Machines (DBMs) is related to the pre-training algorithm for Deep Belief Networks and we show that under certain conditions, the pre-training procedure improves the variational lower bound of a two-hidden-layer DBM. Based on this analysis, we develop a different method of pre-training DBMs that distributes the modelling work more evenly over the hidden layers. Our results on the MNIST and NORB datasets demonstrate that the new pre-training algorithm allows us to learn better generative models.

T8 Emergence of Object-Selective Features in Unsupervised Feature Learning

Adam Coates acoates@cs.stanford.edu
Andrew Ng ang@cs.stanford.edu
Andrej Karpathy andrej.karpathy@gmail.com
Stanford University

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

T9 Deep Representations and Codes for Image Auto-Annotation

Ryan Kiros rkiros@ualberta.ca
Csaba Szepesvari szepesva@cs.ualberta.ca
University of Alberta

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

T10 High Dimensional Semiparametric Scale-invariant Principal Component Analysis

Fang Han fhan@jhsph.edu
Johns Hopkins University
Han Liu hanliu@cs.jhu.edu
Princeton University

We propose a high dimensional semiparametric scale-invariant principal component analysis, named Copula Component Analysis (COCA). The semiparametric model assumes that, after unspecified marginally monotone transformations, the distributions are multivariate Gaussian. The COCA accordingly estimates the leading eigenvector of the correlation matrix of the latent Gaussian distribution. The robust nonparametric rank-based correlation coefficient estimator, Spearman's rho, is exploited in estimation. We prove that, although the marginal distributions can be arbitrarily continuous, the COCA estimators obtain fast estimation rates and are feature selection consistent in the setting where the dimension is nearly exponentially large relative to the sample size. Careful numerical experiments on the simulated data are conducted under both ideal and noisy settings, which suggest that the COCA loses little even when the data are truely Gaussian. The COCA is also implemented on a large-scale genomic data to illustrate its empirical usefulness.

T11 Semi-Supervised Domain Adaptation with Non-Parametric Copulas

David Lopez-Paz david.lopez.paz@gmail.com
Bernhard Schölkopf bs@tuebingen.mpg.de
Max Planck Institute for Intelligent Systems
Jose Miguel Hernández-Lobato jmh233@cam.ac.uk
Cambridge University

A new framework based on the theory of copulas is proposed to address semi-supervised domain adaptation problems. The presented method factorizes any multivariate density into a product of marginal distributions and bivariate copula functions. Therefore, changes in each of these factors can be detected and corrected to adapt a density model across different learning domains. Importantly, we introduce a novel vine copula model, which allows for this factorization in a non-parametric manner. Experimental results on regression problems with real-world data illustrate the efficacy of the proposed approach when compared to state-of-the-art techniques.

T12 Finite Sample Convergence Rates of Zero-Order Stochastic Optimization Methods

John Duchi jduchi@cs.berkeley.edu
Andre Wibisono wibisono@berkeley.edu
Michael Jordan jordan@cs.berkeley.edu
Martin Wainwright wainwrig@eecs.berkeley.edu
UC Berkeley

We consider derivative-free algorithms for stochastic optimization problems that use only noisy function values rather than gradients, analyzing their finite-sample convergence rates. We show that if pairs of function values are available, algorithms that use gradient estimates based on random perturbations suffer a factor of at most dim in convergence rate over traditional stochastic gradient methods, where dim is the dimension of the problem. We complement our algorithmic development with information-theoretic lower bounds on the minimax convergence rate of such problems, which show that our bounds are sharp with respect to all problem-dependent quantities: they cannot be improved by more than constant factors.

T13 Scaled Gradients on Grassmann Manifolds for Matrix Completion

Thanh Ngo thango@cs.umn.edu
Yousef Saad saad@cs.umn.edu
University of Minnesota

This paper describes gradient methods based on a scaled metric on the Grassmann manifold for low-rank matrix completion. The proposed methods significantly improve canonical gradient methods especially on ill-conditioned matrices, while maintaining established global convegence and exact recovery guarantees. A connection between a form of subspace iteration for matrix completion and the scaled gradient descent procedure is also established. The proposed conjugate gradient method based on the scaled gradient outperforms several existing algorithms for matrix completion and is competitive with recently proposed methods.

T14 Minimizing Sparse High-Order Energies by Submodular Vertex-Cover

Andrew Delong andrew.delong@gmail.com
Electrical & Computer University of Toronto
Olga Veksler olga@csd.uwo.ca
Yuri Boykov yuri@csd.uwo.ca
University of Western Ontario
Anton Osokin anton.osokin@gmail.com
Moscow State University

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

T15 Accelerated Training for Matrix-norm Regularization: A Boosting Approach

Xinhua Zhang xinhua.zhang.cs@gmail.com
National ICT Australia (NICTA)
Yao-Liang Yu yaoliang@cs.ualberta.ca
Dale Schuurmans dale@cs.ualberta.ca
University of Alberta

Sparse learning models typically combine a smooth loss with a nonsmooth penalty, such as trace norm. Although recent developments in sparse approximation have offered promising solution methods, current approaches either apply only to matrix-norm constrained problems or provide suboptimal convergence rates. In this paper, we propose a boosting method for regularized learning that guarantees ϵ accuracy within $O(1/\epsilon)$ iterations. Performance is further accelerated by interlacing boosting with fixed-rank local optimization---exploiting a simpler local objective than previous work. The proposed method yields state-of-the-art performance on large-scale problems. We also demonstrate an application to latent multiview learning for which we provide the first efficient weak-oracle.

T16 Approximating Concavely Parameterized Optimization Problems

Joachim Giesen joachim.giesen@uni-jena.de
Jens Mueller jkm@informatik.uni-jena.de
Soeren Laue soeren.laue@uni-jena.de

Sascha Swiercy sascha.swiercy@googlemail.com
Friedrich-Schiller-Universitat Jena

We consider an abstract class of optimization problems that are parameterized concavely in a single parameter, and show that the solution path along the parameter can always be approximated with accuracy $\varepsilon>0$ by a set of size $O(1/\varepsilon)$. A lower bound of size $\Omega(1/\varepsilon)$ shows that the upper bound is tight up to a constant factor. We also devise an algorithm that calls a step-size oracle and computes an approximate path of size $O(1/\varepsilon)$. Finally, we provide an implementation of the oracle for soft-margin support vector machines, and a parameterized semi-definite program for matrix completion.

T17 Finding Exemplars from Pairwise Dissimilarities via Simultaneous Sparse Recovery

Ehsan Elhamifar ehsan@cis.jhu.edu
UC Berkeley
Guillermo Sapiro guillermo.sapiro@duke.edu
Duke University
Rene Vidal rvidal@cis.jhu.edu
Johns Hopkins University

Given pairwise dissimilarities between data points, we consider the problem of finding a subset of data points called representatives or exemplars that can efficiently describe the data collection. We formulate the problem as a row-sparsity regularized trace minimization problem which can be solved efficiently using convex programming. The solution of the proposed optimization program finds the representatives and the probability that each data point is associated to each one of the representatives. We obtain the range of the regularization parameter for which the solution of the proposed optimization program changes from selecting one representative to selecting all data points as the representatives. When data points are distributed around multiple clusters according to the dissimilarities, we show that the data in each cluster select only representatives from that cluster. Unlike metric-based methods, our algorithm does not require that the pairwise dissimilarities be metrics and can be applied to dissimilarities that are asymmetric or violate the triangle inequality. We demonstrate the effectiveness of the proposed algorithm on synthetic data as well as real-world datasets of images and text.

T18 Convex Multi-view Subspace Learning

Martha White whitem@ualberta.ca
Yao-Liang Yu yaoliang@cs.ualberta.ca
Dale Schuurmans dale@cs.ualberta.ca
University of Alberta
Xinhua Zhang xinhua.zhang.cs@gmail.com
National ICT Australia (NICTA)

Subspace learning seeks a low dimensional representation of data that enables accurate reconstruction. However, in many applications, data is obtained from multiple sources rather than a single source (e.g. an object might be viewed by cameras at different angles, or a document might consist of text and images). The conditional independence of separate sources imposes constraints on their shared latent representation, which, if respected, can improve the quality of the learned low dimensional representation. In this paper, we present a convex formulation of multi-view subspace learning that enforces conditional independence while reducing dimensionality. For this formulation, we develop an efficient algorithm that recovers an optimal data reconstruction by exploiting an implicit convex regularizer, then recovers the corresponding latent representation and reconstruction model, jointly and optimally. Experiments illustrate that the proposed method produces high quality results.

T19 A Polylog Pivot Steps Simplex Algorithm for Classification

Elad Hazan ehazan@ie.technion.ac.il
Technion
Zohar Karnin zkarnin@ymail.com
Yahoo! Research

We present a simplex algorithm for linear programming in a linear classification formulation. The paramount complexity parameter in linear classification problems is called the margin. We prove that for margin values of practical interest our simplex variant performs a polylogarithmic number of pivot steps in the worst case, and its overall running time is near linear. This is in contrast to general linear programming, for which no sub-polynomial pivot rule is known.

T20 Stochastic optimization and sparse statistical recovery: Optimal algorithms for high dimensions

Alekh Agarwal alekh@cs.berkeley.edu
Martin Wainwright wainwrig@eecs.berkeley.edu
UC Berkeley
Sahand Negahban sahand_n@eecs.berkeley.edu
University of California, Berkeley

We develop and analyze stochastic optimization algorithms for problems in which the expected loss is strongly convex, and the optimum is (approximately) sparse. Previous approaches are able to exploit only one of these two structures, yielding a \order(\pdim/T) convergence rate for strongly convex objectives in \pdim dimensions and \order(\ spindex(log\pdim)/T) convergence rate when the optimum is \spindex-sparse. Our algorithm is based on successively solving a series of ℓ1-regularized optimization problems using Nesterov's dual averaging algorithm. We establish that the error of our solution after T iterations is at most \order(\spindex(log\pdim)/T), with natural extensions to approximate sparsity. Our results apply to locally Lipschitz losses including the logistic, exponential, hinge and least-squares losses. By recourse to statistical minimax results, we show that our convergence rates are optimal up to constants. The effectiveness of our approach is also confirmed in numerical simulations where we compare to several baselines on a least-squares regression problem.

T21 Stochastic Gradient Descent with Only One Projection

Mehrdad Mahdavi mahdavim@msu.edu
Rong Jin rong+@cs.cmu.edu
Michigan State University
Tianbao Yang yangtia1@msu.edu
GE Global Research
Shenghuo Zhu zsh@nec-labs.com
NEC Laboratories America

Although many variants of stochastic gradient descent have been proposed for large-scale convex optimization, most of them require projecting the solution at {\ it each} iteration to ensure that the obtained solution stays within the feasible domain. For complex domains (e.g., positive semidefinite cone), the projection step can be computationally expensive, making stochastic gradient descent unattractive for large-scale optimization problems. We address this limitation by developing a novel stochastic gradient descent algorithm that does not need intermediate projections. Instead, only one projection at the last iteration is needed to obtain a feasible solution in the given domain. Our theoretical analysis shows that with a high probability, the proposed algorithms achieve an O(1/T) convergence rate for general convex optimization, and an O(lnT/T) rate for strongly convex optimization under mild conditions about the domain and the objective function.

T22 Optimal Regularized Dual Averaging Methods for Stochastic Optimization

Xi Chen xichen@cs.cmu.edu
Qihang Lin qihangl@andrew.cmu.edu
Javier Pena jfp@andrew.cmu.edu
Carnegie Mellon University

This paper considers a wide spectrum of regularized stochastic optimization problems where both the loss function and regularizer can be non-smooth. We develop a novel algorithm based on the regularized dual averaging (RDA) method, that can simultaneously achieve the optimal convergence rates for both convex and strongly convex loss. In particular, for strongly convex loss, it achieves the optimal rate of O(1N+1N2) for N iterations, which improves the best known rate O(logNN) of previous stochastic dual averaging algorithms. In addition, our method constructs the final solution directly from the proximal mapping instead of averaging of all previous iterates. For widely used sparsity-inducing regularizers (e.g., ℓ1-norm), it has the advantage of encouraging sparser solutions. We further develop a multi-stage extension using the proposed algorithm as a subroutine, which achieves the uniformly-optimal rate O(1N+exp{−N}) for strongly convex loss.

T23 A Stochastic Gradient Method with an Exponential Convergence Rate for Finite Training Sets

Nicolas Le Roux nicolas@le-roux.name
Criteo
Mark Schmidt mark.schmidt@inria.fr
Francis Bach francis.bach@mines.org
INRIA - Ecole Normale Superieure

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

T24 On Lifting the Gibbs Sampling Algorithm

Deepak Venugopal dxv021000@utdallas.edu
Vibhav Gogate vgogate@hlt.utdallas.edu
The University of Texas at Dallas

Statistical relational learning models combine the power of first-order logic, the de facto tool for handling relational structure, with that of probabilistic graphical models, the de facto tool for handling uncertainty. Lifted probabilistic inference algorithms for them have been the subject of much recent research. The main idea in these algorithms is to improve the speed, accuracy and scalability of existing graphical models' inference algorithms by exploiting symmetry in the first-order representation. In this paper, we consider blocked Gibbs sampling, an advanced variation of the classic Gibbs sampling algorithm and lift it to the first-order level. We propose to achieve this by partitioning the first-order atoms in the relational model into a set of disjoint clusters such that exact lifted inference is polynomial in each cluster given an assignment to all other atoms not in the cluster. We propose an approach for constructing such clusters and determining their complexity and show how it can be used to trade accuracy with computational complexity in a principled manner. Our experimental evaluation shows that lifted Gibbs sampling is superior to the propositional algorithm in terms of accuracy and convergence.

T25 Active Comparison of Prediction Models

Christoph Sawade sawade@cs.uni-potsdam.de
Niels Landwehr landwehr@cs.uni-potsdam.de
Tobias Scheffer scheffer@cs.uni-potsdam.de
University of Potsdam

We address the problem of comparing the risks of two given predictive models - for instance, a baseline model and a challenger - as confidently as possible on a fixed labeling budget. This problem occurs whenever models cannot be compared on held-out training data, possibly because the training data are unavailable or do not reflect the desired test distribution. In this case, new test instances have to be drawn and labeled at a cost. We devise an active comparison method that selects instances according to an instrumental sampling distribution. We derive the sampling distribution that maximizes the power of a statistical test applied to the observed empirical risks, and thereby minimizes the likelihood of choosing the inferior model. Empirically, we investigate model selection problems on several classification and regression tasks and study the accuracy of the resulting p-values.

T26 Multi-Task Averaging

Sergey Feldman sergeyfeldman@gmail.com
Maya Gupta gupta@ee.washington.edu
Bela Frigyik frigyik@uw.edu
Electrical University of Washington

We present a multi-task learning approach to jointly estimate the means of multiple independent data sets. The proposed multi-task averaging (MTA) algorithm results in a convex combination of the single-task averages. We derive the optimal amount of regularization, and show that it can be effectively estimated. Simulations and real data experiments demonstrate that MTA both maximum likelihood and James-Stein estimators, and that our approach to estimating the amount of regularization rivals cross-validation in performance but is more computationally efficient.

T27 Multiplicative Forests for Continuous-Time Processes

Jeremy Weiss jcweiss@cs.wisc.edu
David Page page@biostat.wisc.edu
University of Wisconsin - Madison
Sriraam Natarajan snataraj@wakehealth.edu
Wake Forest University Baptist Medical Center

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

T28 Entangled Monte Carlo

Seong-Hwan Jun s2jun.uw@gmail.com
Alexandre Bouchard-Côté bouchard@stat.ubc.ca
UBC
Liangliang Wang wang@stats.uwo.ca
University of Western Ontario

We propose a novel method for scalable parallelization of SMC algorithms, Entangled Monte Carlo simulation (EMC). EMC avoids the transmission of particles between nodes, and instead reconstructs them from the particle genealogy. In particular, we show that we can reduce the communication to the particle weights for each machine while efficiently maintaining implicit global coherence of the parallel simulation. We explain methods to efficiently maintain a genealogy of particles from which any particle can be reconstructed. We demonstrate using examples from Bayesian phylogenetic that the computational gain from parallelization using EMC significantly outweighs the cost of particle reconstruction. The timing experiments show that reconstruction of particles is indeed much more efficient as compared to transmission of particles.

T29 Fast Bayesian Inference for Non-Conjugate Gaussian Process Regression

Emtiyaz Khan emtiyaz@cs.ubc.ca
Shakir Mohamed shakirm@cs.ubc.ca
Kevin Murphy murphyk@gmail.com
University of British Columbia

We present a new variational inference algorithm for Gaussian processes with non-conjugate likelihood functions. This includes binary and multi-class classification, as well as ordinal regression. Our method constructs a convex lower bound, which can be optimized by using an efficient fixed point update method. We then show empirically that our new approach is much faster than existing methods without any degradation in performance.

T30 Globally Convergent Dual MAP LP Relaxation Solvers using Fenchel-Young Margins

Alex Schwing aschwing@inf.ethz.ch
Marc Pollefeys marc.pollefeys@inf.ethz.ch
ETH Zurich
Tamir Hazan tamir@ttic.edu
Raquel Urtasun rurtasun@ttic.edu
TTI-Chicago

While finding the exact solution for the MAP inference problem is intractable for many real-world tasks, MAP LP relaxations have been shown to be very effective in practice. However, the most efficient methods that perform block coordinate descent can get stuck in sub-optimal points as they are not globally convergent. In this work we propose to augment these algorithms with an ϵ-descent approach and present a method to efficiently optimize for a descent direction in the subdifferential using a margin-based extension of the Fenchel-Young duality theorem. Furthermore, the presented approach provides a methodology to construct a primal optimal solution from its dual optimal counterpart. We demonstrate the efficiency of the presented approach on spin glass models and protein interactions problems and show that our approach outperforms state-of-the-art solvers.

T31 Fast Variational Inference in the Conjugate Exponential Family

James Hensman james.hensman@gmail.com
Magnus Rattray magnus@cs.man.ac.uk
Neil Lawrence N.Lawrence@shef.ac.uk
University of Sheffield

We present a general method for deriving collapsed variational inference algorithms for probabilistic models in the conjugate exponential family. Our method unifies many existing approaches to collapsed variational inference. Our collapsed variational inference leads to a new lower bound on the marginal likelihood. We exploit the information geometry of the bound to derive much faster optimization methods based on conjugate gradients for these models. Our approach is very general and is easily applied to any model where the mean field update equations have been derived. Empirically we show significant speed-ups for probabilistic models optimized using our bound.

T32 Minimization of Continuous Bethe Approximations: A Positive Variation

Jason Pacheco pachecoj@cs.brown.edu
Erik Sudderth sudderth@cs.brown.edu
Brown University

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

T33 Perfect Dimensionality Recovery by Variational Bayesian PCA

Shinichi Nakajima shinnkj23@gmail.com
Nikon Corporation
Ryota Tomioka tomioka@mist.i.u-tokyo.ac.jp
University of Tokyo
Masashi Sugiyama sugi@cs.titech.ac.jp
Tokyo Institute of Technology
S. Derin Babacan dbabacan@gmail.com
University of Illinois at Urbana-Champaign

The variational Bayesian (VB) approach is one of the best tractable approximations to the Bayesian estimation, and it was demonstrated to perform well in many applications. However, its good performance was not fully understood theoretically. For example, VB sometimes produces a sparse solution, which is regarded as a practical advantage of VB, but such sparsity is hardly observed in the rigorous Bayesian estimation. In this paper, we focus on probabilistic PCA and give more theoretical insight into the empirical success of VB. More specifically, for the situation where the noise variance is unknown, we derive a sufficient condition for perfect recovery of the true PCA dimensionality in the large-scale limit when the size of an observed matrix goes to infinity. In our analysis, we obtain bounds for a noise variance estimator and simple closed-form solutions for other parameters, which themselves are actually very useful for better implementation of VB-PCA.

T34 Expectation Propagation in Gaussian Process Dynamical Systems

Marc Deisenroth marc@ias.tu-darmstadt.de
TU Darmstadt
Shakir Mohamed shakirm@cs.ubc.ca
University of British Columbia

Rich and complex time-series data, such as those generated from engineering sys- tems, financial markets, videos or neural recordings are now a common feature of modern data analysis. Explaining the phenomena underlying these diverse data sets requires flexible and accurate models. In this paper, we promote Gaussian process dynamical systems as a rich model class appropriate for such analysis. In particular, we present a message passing algorithm for approximate inference in GPDSs based on expectation propagation. By phrasing inference as a general mes- sage passing problem, we iterate forward-backward smoothing. We obtain more accurate posterior distributions over latent structures, resulting in improved pre- dictive performance compared to state-of-the-art GPDS smoothers, which are spe- cial cases of our general iterative message passing algorithm. Hence, we provide a unifying approach within which to contextualize message passing in GPDSs.

T35 Random function priors for exchangeable graphs and arrays

James Lloyd jrl44@cam.ac.uk
Dan Roy d.roy@eng.cam.ac.uk
Zoubin Ghahramani zoubin@eng.cam.ac.uk
University of Cambridge
Peter Orbanz porbanz@stat.columbia.edu
Columbia University

A fundamental problem in the analysis of relational data---graphs, matrices or higher-dimensional arrays---is to extract a summary of the common structure underlying relations between individual entities. A successful approach is latent variable modeling, which summarizes this structure as an embedding into a suitable latent space. Results in probability theory, due to Aldous, Hoover and Kallenberg, show that relational data satisfying an exchangeability property can be represented in terms of a random measurable function. In a Bayesian model, this function constitutes the natural model parameter, and we discuss how available latent variable models can be classified according to how they implicitly approximate this parameter. We obtain a flexible yet simple model for relational data by representing the parameter function as a Gaussian process. Efficient inference draws on the large available arsenal of Gaussian process algorithms; sparse approximations prove particularly useful. We demonstrate applications of the model to network data and clarify its relation to models in the literature, several of which emerge as special cases.

T36 Forward-Backward Activation Algorithm for Hierarchical Hidden Markov Models

Kei Wakabayashi kwakaba@slis.tsukuba.ac.jp
University of Tsukuba
Takao Miura miurat@hosei.ac.jp
Hosei University

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

T37 Active Learning of Model Evidence Using Bayesian Quadrature

Michael Osborne mosb@robots.ox.ac.uk
Stephen Roberts sjrob@robots.ox.ac.uk
University of Oxford
David Duvenaud dkd23@cam.ac.uk
Carl Edward Rasmussen cer54@cam.ac.uk
Zoubin Ghahramani zoubin@eng.cam.ac.uk
Dept of University of Cambridge
Roman Garnett rgarnett@cs.cmu.edu
Carnegie Mellon University

Numerical integration is an key component of many problems in scientific computing, statistical modelling, and machine learning. Bayesian Quadrature is a model-based method for numerical integration which, relative to standard Monte Carlo methods, offers increased sample efficiency and a more robust estimate of the uncertainty in the estimated integral. We propose a novel Bayesian Quadrature approach for numerical integration when the integrand is non-negative, such as the case of computing the marginal likelihood, predictive distribution, or normalising constant of a probabilistic model. Our approach approximately marginalises the quadrature model's hyperparameters in closed form, and introduces an active learning scheme to optimally select function evaluations, as opposed to using Monte Carlo samples. We demonstrate our method on both a number of synthetic benchmarks and a real scientific problem from astronomy.

T38 Phoneme Classification using Constrained Variational Gaussian Process Dynamical System

Hyunsin Park hs.park@kaist.ac.kr
Jongmin Kim kimjm0309@gmail.com
Sanghyuk Park shine0624@kaist.ac.kr
Sungrack Yun yunsungrack@kaist.ac.kr
Chang D. Yoo cdyoo@ee.kaist.ac.kr
KAIST

This paper describes a new acoustic model based on variational Gaussian process dynamical system (VGPDS) for phoneme classification. The proposed model overcomes the limitations of the classical HMM in modeling the real speech data, by adopting a nonlinear and nonparametric model. In our model, the GP prior on the dynamics function enables representing the complex dynamic structure of speech, while the GP prior on the emission function successfully models the global dependency over the observations. Additionally, we introduce variance constraint to the original VGPDS for mitigating sparse approximation error of the kernel matrix. The effectiveness of the proposed model is demonstrated with extensive experimental results including parameter estimation, classification performance on the synthetic and benchmark datasets.

T39 Density Propagation and Improved Bounds on the Partition Function

Stefano Ermon ermonste@cs.cornell.edu
Carla P. Gomes gomes@cs.cornell.edu
Bart Selman selman@cs.cornell.edu
Cornell University
Ashish Sabharwal ashish.sabharwal@us.ibm.com
IBM Watson Research Center

Given a probabilistic graphical model, its density of states is a function that, for any likelihood value, gives the number of configurations with that probability. We introduce a novel message-passing algorithm called Density Propagation (DP) for estimating this function. We show that DP is exact for tree-structured graphical models and is, in general, a strict generalization of both sum-product and max-product algorithms. Further, we use density of states and tree decomposition to introduce a new family of upper and lower bounds on the partition function. For any tree decompostion, the new upper bound based on finer-grained density of state information is provably at least as tight as previously known bounds based on convexity of the log-partition function, and strictly stronger if a general condition holds. We conclude with empirical evidence of improvement over convex relaxations and mean-field based bounds.

T40 High Dimensional Transelliptical Graphical Models

Han Liu hanliu@cs.jhu.edu
Princeton University
Fang Han fhan@jhsph.edu
Johns Hopkins University

We advocate the use of a new distribution family--the transelliptical--for robust inference of high dimensional graphical models. The transelliptical family is an extension of the nonparanormal family proposed by Liu et al. (2009). Just as the nonparanormal extends the normal by transforming the variables using univariate functions, the transelliptical extends the elliptical family in the same way. We propose a nonparametric rank-based regularization estimator which achieves the parametric rates of convergence for both graph recovery and parameter estimation. Such a result suggests that the extra robustness and flexibility obtained by the semiparametric transelliptical modeling incurs almost no efficiency loss. Thorough numerical experiments are provided to back up our theory.

T41 Scaling Constrained Continuous Markov Random Fields with Consensus Optimization

Stephen Bach bach@cs.umd.edu
Lise Getoor getoor@cs.umd.edu
Matthias Broecheler matthias@cs.umd.edu
University of Maryland

We study scaling a class of probabilistic graphical models well-suited to constrained, continuous domains. We show how to solve the most-probable-explanation problem for these models with a consensus-optimization framework. We derive closed-form solutions for consensus-optimization subproblems induced by several types of common dependencies. We improve the performance of consensus optimization by deriving an algorithm that can additionally find closed-form solutions to subproblems in certain cases, depending on the current optimization iterate, not just the subproblem itself. We demonstrate superior performance of our approach over commercial interior-point methods, the current state-of-the-art for the problems we study. In fact, in our evaluation our method scales linearly with the size of the problem.

T42 Continuous Relaxations for Discrete Hamiltonian Monte Carlo

Zoubin Ghahramani zoubin@eng.cam.ac.uk
University of Cambridge
Yichuan Zhang s0956889@sms.ed.ac.uk
Charles Sutton csutton@inf.ed.ac.uk
Amos Storkey a.storkey@ed.ac.uk
University of Edinburgh

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

T43 Calibrated Elastic Regularization in Matrix Completion

Cun-Hui Zhang czhang@stat.rutgers.edu
Rutgers University
Tingni Sun tingni@stat.rutgers.edu
University of Pennsylvania

This paper concerns the problem of matrix completion, which is to estimate a matrix from observations in a small subset of indices. We propose a calibrated spectrum elastic net method with a sum of the nuclear and Frobenius penalties and develop an iterative algorithm to solve the convex minimization problem. The iterative algorithm alternates between imputing the missing entries in the incomplete matrix by the current guess and estimating the matrix by a scaled soft-thresholding singular value decomposition of the imputed matrix until the resulting matrix converges. A calibration step follows to correct the bias caused by the Frobenius penalty. Under proper coherence conditions and for suitable penalties levels, we prove that the proposed estimator achieves an error bound of nearly optimal order and in proportion to the noise level. This provides a unified analysis of the noisy and noiseless matrix completion problems. Simulation results are presented to compare our proposal with previous ones.

T44 Latent Graphical Model Selection: Efficient Methods for Locally Tree-like Graphs

Anima Anandkumar a.anandkumar@uci.edu
Ragupathyraj Valluvan rvalluva@uci.edu
UC Irvine

Graphical model selection refers to the problem of estimating the unknown graph structure given observations at the nodes in the model. We consider a challenging instance of this problem when some of the nodes are latent or hidden. We characterize conditions for tractable graph estimation and develop efficient methods with provable guarantees. We consider the class of Ising models Markov on locally tree-like graphs, which are in the regime of correlation decay. We propose an efficient method for graph estimation, and establish its structural consistency when the number of samples n scales as n=Ω(θmin−δη(η+1)−2logp), where θmin is the minimum edge potential, δ is the depth (i.e., distance from a hidden node to the nearest observed nodes), and η is a parameter which depends on the minimum and maximum node and edge potentials in the Ising model. The proposed method is practical to implement and provides flexibility to control the number of latent variables and the cycle lengths in the output graph. We also present necessary conditions for graph estimation by any method and show that our method nearly matches the lower bound on sample requirements.

T45 Slice Normalized Dynamic Markov Logic Networks

Tivadar Papai papai@cs.rochester.edu
Henry Kautz kautz@cs.rochester.edu
Daniel Stefankovic stefanko@cs.rochester.edu
University of Rochester

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

T46 Volume Regularization for Binary Classification

Koby Crammer koby@ee.technion.ac.il
The Technion
Tal Wagner talw@tx.technion.ac.il
Weizmann Institute

We introduce a large-volume box classification for binary prediction, which maintains a subset of weight vectors, and specifically axis-aligned boxes. Our learning algorithm seeks for a box of large volume that contains ``simple'' weight vectors which most of are accurate on the training set. Two versions of the learning process are cast as convex optimization problems, and it is shown how to solve them efficiently. The formulation yields a natural PAC-Bayesian performance bound and it is shown to minimize a quantity directly aligned with it. The algorithm outperforms SVM and the recently proposed AROW algorithm on a majority of 30 NLP datasets and binarized USPS optical character recognition datasets.

T47 Spectral Learning of General Weighted Automata via Constrained Matrix Completion

Borja Balle bballe@lsi.upc.edu
Universitat Politecnica de Catalunya (UPC)
Mehryar Mohri mohri@google.com
Courant Institute & Google Research

Many tasks in text and speech processing and computational biology involve functions from variable-length strings to real numbers. A wide class of such functions can be computed by weighted automata. Spectral methods based on singular value decompositions of Hankel matrices have been recently proposed for learning probability distributions over strings that can be computed by weighted automata. In this paper we show how this method can be applied to the problem of learning a general weighted automata from a sample of string-label pairs generated by an arbitrary distribution. The main obstruction to this approach is that in general some entries of the Hankel matrix that needs to be decomposed may be missing. We propose a solution based on solving a constrained matrix completion problem. Combining these two ingredients, a whole new family of algorithms for learning general weighted automata is obtained. Generalization bounds for a particular algorithm in this class are given. The proofs rely on a stability analysis of matrix completion and spectral learning.

T48 Learning Probability Measures with respect to Optimal Transport Metrics

Guille Canas guilledc@MIT.EDU
CSAIL/BCS, IIT-MIT
Lorenzo Rosasco lrosasco@mit.edu
MIT and Italian Institute of Technology

We study the problem of estimating, in the sense of optimal transport metrics, a measure which is assumed supported on a manifold embedded in a Hilbert space. By establishing a precise connection between optimal transport metrics, optimal quantization, and learning theory, we derive new probabilistic bounds for the performance of a classic algorithm in unsupervised learning (k-means), when used to produce a probability measure derived from the data. In the course of the analysis, we arrive at new lower bounds, as well as probabilistic bounds on the convergence rate of the empirical law of large numbers, which, unlike existing bounds, are applicable to a wide class of measures.

T49 Fast Resampling Weighted v-Statistics

Chunxiao Zhou czhou4@gmail.com
NIH
jiseong Park jiseongp@gmail.com
Yun Fu yunfu@ece.neu.edu
Northeastern University

In this paper, a novel, computationally fast, and alternative algorithm for com- puting weighted v-statistics in resampling both univariate and multivariate data is proposed. To avoid any real resampling, we have linked this problem with finite group action and converted it into a problem of orbit enumeration. For further computational cost reduction, an efficient method is developed to list all orbits by their symmetry order and calculate all index function orbit sums and data function orbit sums recursively. The computational complexity analysis shows reduction in the computational cost from n! or nn level to low-order polynomial level.

T50 Approximating Equilibria in Sequential Auctions with Incomplete Information and Multi-Unit Demand

Amy Greenwald amy@cs.brown.edu
Jiacui Li Jiacui_Li@brown.edu
Eric Sodomka sodomka@cs.brown.edu
Brown University

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

T51 Interpreting prediction markets: a stochastic approach

Nicolas Della Penna nikete@gmail.com
Mark Reid mark.reid@anu.edu.au
Australian National University
Rafael Frongillo raf@cs.berkeley.edu
UC Berkeley

We strengthen recent connections between prediction markets and learning by showing that a natural class of market makers can be understood as performing stochastic mirror descent when trader demands are sequentially drawn from a fixed distribution. This provides new insights into how market prices (and price paths) may be interpreted as a summary of the market's belief distribution by relating them to the optimization problem being solved. In particular, we show that the stationary point of the stochastic process of prices generated by the market is equal to the market's Walrasian equilibrium of classic market analysis. Together, these results suggest how traditional market making mechanisms might be replaced with general purpose learning algorithms while still retaining guarantees about their behaviour.

T52 Active Learning of Multi-Index Function Models

Hemant Tyagi h83.tyagi@gmail.com
Ecole Polytechnique Federale de Lausanne
Volkan Cevher volkan.cevher@epfl.ch
STI-IEL-LIONS, EPFL

We consider the problem of actively learning \ textit{multi-index} functions of the form f(\vecx)=g(\matA\ vecx)=∑i=1kgi(\vecaiT\vecx) from point evaluations of f. We assume that the function f is defined on an ℓ2-ball in \Reald, g is twice continuously differentiable almost everywhere, and \matA∈Rk×d is a rank k matrix, where k≪d. We propose a randomized, active sampling scheme for estimating such functions with uniform approximation guarantees. Our theoretical developments leverage recent techniques from low rank matrix recovery, which enables us to derive an estimator of the function f along with sample complexity bounds. We also characterize the noise robustness of the scheme, and provide empirical evidence that the high-dimensional scaling of our sample complexity bounds are quite accurate.

T53 Hierarchical Optimistic Region Selection driven by Curiosity

Odalric-Ambrym Maillard
odalricambrym.maillard@gmail.com
Montanuniversität Leoben

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

T54 Risk-Aversion in Multi-armed Bandits

Amir Sani reachme@amirsani.com
Remi Munos remi.munos@inria.fr
Alessandro Lazaric alessandro.lazaric@inria.fr
INRIA Lille-Nord Europe

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

T55 Online allocation and homogeneous partitioning for piecewise constant mean-approximation

Alexandra Carpentier a.carpentier@statslab.cam.ac.uk
Cambridge University
Odalric-Ambrym Maillard
odalricambrym.maillard@gmail.com
Montanuniversität Leoben

In the setting of active learning for the multi-armed bandit, where the goal of a learner is to estimate with equal precision the mean of a finite number of arms, recent results show that it is possible to derive strategies based on finite-time confidence bounds that are competitive with the best possible strategy. We here consider an extension of this problem to the case when the arms are the cells of a finite partition P of a continuous sampling space X \subset \Real^d. Our goal is now to build a piecewise constant approximation of a noisy function (where each piece is one region of P and P is fixed beforehand) in order to maintain the local quadratic error of approximation on each cell equally low. Although this extension is not trivial, we show that a simple algorithm based on upper confidence bounds can be proved to be adaptive to the function itself in a near-optimal way, when |P| is chosen to be of minimax-optimal order on the class of \alpha-Hölder functions.

T56 Adaptive Stratified Sampling for Monte-Carlo integration of Differentiable functions

Alexandra Carpentier a.carpentier@statslab.cam.ac.uk
Cambridge University
Remi Munos remi.munos@inria.fr
NRIA Lille - Nord Europe

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

T57 Putting Bayes to sleep

Wouter Koolen wouter@cs.rhul.ac.uk
Dmitri Adamskiy adamskiy@cs.rhul.ac.uk
University of London
Manfred Warmuth manfred@cse.ucsc.edu
Univ. of Calif. at Santa Cruz

We consider sequential prediction algorithms that are given the predictions from a set of models as inputs. If the nature of the data is changing over time in that different models predict well on different segments of the data, then adaptivity is typically achieved by mixing into the weights in each round a bit of the initial prior (kind of like a weak restart). However, what if the favored models in each segment are from a small subset, i.e. the data is likely to be predicted well by models that predicted well before? Curiously, fitting such ''sparse composite models'' is achieved by mixing in a bit of all the past posteriors. This self-referential updating method is rather peculiar, but it is efficient and gives superior performance on many natural data sets. Also it is important because it introduces a long-term memory: any model that has done well in the past can be recovered quickly. While Bayesian interpretations can be found for mixing in a bit of the initial prior, no Bayesian interpretation is known for mixing in past posteriors. We build atop the ''specialist'' framework from the online learning literature to give the Mixing Past Posteriors update a proper Bayesian foundation. We apply our method to a well-studied multitask learning problem and obtain a new intriguing efficient update that achieves a significantly better bound.

T58 Online Sum-Product Computation

Mark Herbster m.herbster@cs.ucl.ac.uk
Stephen Pasteris stephen.pasteris@googlemail.com
University College London
Fabio Vitale fabio.vitale@unimi.it
Università degli Studi di Milano

We consider the problem of performing efficient sum-product computations in an online setting over a tree. A natural application of our methods is to compute the marginal distribution at a vertex in a tree-structured Markov random field. Belief propagation can be used to solve this problem. However, belief propagation requires time linear in the size of the tree. This is too slow in an online setting where we are continuously receiving new data and computing individual marginals. With our method we aim to update the data and compute marginals in time that is no more than logarithmic in the size of the tree, and is often significantly less. We accomplish this via a hierarchical covering structure that caches previous local sum-product computations. Our contribution is three-fold: we i) give a linear time algorithm to find an optimal hierarchical cover of a tree; ii) give a sum-product-like algorithm to efficiently compute marginals with respect to this cover; and iii) apply ``i'' and ``ii'' to find an efficient algorithm with a regret bound for the online {\em allocation} problem in a multi-task setting.

T59 Learning with Target Prior

Zuoguan Wang zuoguanwang@gmail.com
Qiang Ji qji@ecse.rpi.edu
ECSE, Rensselaer Polytechnic Institute
Siwei Lyu lsw@cs.albany.edu
University at Albany SUNY
Gerwin Schalk schalk@wadsworth.org
Wadsworth Center

In the conventional approaches for supervised parametric learning, relations between data and target variables are provided through training sets consisting of pairs of corresponded data and target variables. In this work, we describe a new learning scheme for parametric learning, in which the target variables \y can be modeled with a prior model p(\y) and the relations between data and target variables are estimated through p(\y) and a set of uncorresponded data \x in training. We term this method as learning with target priors (LTP). Specifically, LTP learning seeks parameter \t that maximizes the log likelihood of f\t(\x) on a uncorresponded training set with regards to p(\y). Compared to the conventional (semi) supervised learning approach, LTP can make efficient use of prior knowledge of the target variables in the form of probabilistic distributions, and thus removes/reduces the reliance on training data in learning. Compared to the Bayesian approach, the learned parametric regressor in LTP can be more efficiently implemented and deployed in tasks where running efficiency is critical, such as on-line BCI signal decoding. We demonstrate the effectiveness of the proposed approach on parametric regression tasks for BCI signal decoding and pose estimation from video.

T60 Learning High-Density Regions for a Generalized Kolmogorov-Smirnov Test in High-Dimensional Data

Assaf Glazer assafgr@cs.technion.ac.il
Michael Lindenbaoum mic@cs.technion.ac.il
Shaul Markovitch shaulm@cs.technion.ac.il
Technion

We propose an efficient, generalized, nonparametric, statistical Kolmogorov-Smirnov test for detecting distributional change in high-dimensional data. To implement the test, we introduce a novel, hierarchical, minimum-volume sets estimator to represent the distributions to be tested. Our work is motivated by the need to detect changes in data streams, and the test is especially efficient in this context. We provide the theoretical foundations of our test and show its superiority over existing methods.

T61 Learning with Partially Absorbing Random Walks

Xiao-Ming Wu xw2223@columbia.edu
Zhenguo Li zgli@ee.columbia.edu
Shih-Fu Chang sfchang@ee.columbia.edu
John Wright johnwright@ee.columbia.edu
Columbia University New York
Anthony Man-Cho So manchoso@se.cuhk.edu.hk
The Chinese University of Hong Kong

We propose a novel stochastic process that is with probability αi being absorbed at current state i, and with probability 1−αi follows a random edge out of it. We analyze its properties and show its potential for exploring graph structures. We prove that under proper absorption rates, a random walk starting from a set S of low conductance will be mostly absorbed in S. Moreover, the absorption probabilities vary slowly inside S, while dropping sharply outside S, thus implementing the desirable cluster assumption for graph-based learning. Remarkably, the partially absorbing process unifies many popular models arising in a variety of contexts, provides new insights into them, and makes it possible for transferring findings from one paradigm to another. Simulation results demonstrate its promising applications in graph-based learning.

T62 Small-Variance Asymptotics for Exponential Family Dirichlet Process Mixture Models

Ke Jiang jiangk@cse.ohio-state.edu
Brian Kulis brian.kulis@gmail.com
Ohio State U.
Michael Jordan jordan@cs.berkeley.edu
University of California

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

T63 Repulsive Mixtures

FRANCESCA PETRALIA fp12@duke.edu
STATISTICS, DUKE UNIVERSITY

Vinayak Rao vrao@gatsby.ucl.ac.uk
UCL
David Dunson dunson@stat.duke.edu
Duke University

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

T64 Training sparse natural image models with a fast Gibbs sampler of an extended state space

Lucas Theis lucas@bethgelab.org
University of Tübingen
Jascha Sohl-Dickstein jascha@berkeley.edu
UC Berkeley
Matthias Bethge matthias@bethgelab.org
Max Planck Institute for Biological Cybernetics

We present a new learning strategy based on an efficient blocked Gibbs sampler for sparse overcomplete linear models. Particular emphasis is placed on statistical image modeling, where overcomplete models have played an important role in discovering sparse representations. Our Gibbs sampler is faster than general purpose sampling schemes while also requiring no tuning as it is free of parameters. Using the Gibbs sampler and a persistent variant of expectation maximization, we are able to extract highly sparse distributions over latent sources from data. When applied to natural images, our algorithm learns source distributions which resemble spike-and-slab distributions. We evaluate the likelihood and quantitatively compare the performance of the overcomplete linear model to its complete counterpart as well as a product of experts model, which represents another overcomplete generalization of the complete linear model. In contrast to previous claims, we find that overcomplete representations lead to significant improvements, but that the overcomplete linear model still underperforms other models.

T65 Provable ICA with Unknown Gaussian Noise, with Implications for Gaussian Mixtures and Autoencoders

Sanjeev Arora arora@cs.princeton.edu
Rong Ge rongge@cs.princeton.edu
Princeton University
Ankur Moitra moitra@ias.edu
IAS
Sushant Sachdeva sachdeva@cs.princeton.edu

We present a new algorithm for Independent Component Analysis (ICA) which has provable performance guarantees. In particular, suppose we are given samples of the form y=Ax+η where A is an unknown n×n matrix and x is chosen uniformly at random from {+1,−1}n, η is an n-dimensional Gaussian random variable with unknown covariance Σ: We give an algorithm that provable recovers A and Σ up to an additive ϵ whose running time and sample complexity are polynomial in n and 1/ϵ. To accomplish this, we introduce a novel ``quasi-whitening'' step that may be useful in other contexts in which the covariance of Gaussian noise is not known in advance. We also give a general framework for finding all local optima of a function (given an oracle for approximately finding just one) and this is a crucial step in our algorithm, one that has been overlooked in previous attempts, and allows us to control the accumulation of error when we find the columns of A one by one via local search.

T66 TCA: High Dimensional Principal Component Analysis for non-Gaussian Data

Fang Han fhan@jhsph.edu
Johns Hopkins University
Han Liu hanliu@cs.jhu.edu
Princeton University

We propose a high dimensional semiparametric scale-invariant principle component analysis, named TCA, by utilize the natural connection between the elliptical distribution family and the principal component analysis. Elliptical distribution family includes many well-known multivariate distributions like multivariate t and logistic and it is extended to the meta-elliptical by Fang (2002) using the copula techniques. In this paper we extend the meta-elliptical distribution family to a even larger family, called transelliptical. We prove that TCA can obtain a near-optimal s(log d/n)^{1/2} estimation consistency rate in the transelliptical distribution family, even if the distributions are very heavy-tailed, have infinite second moments, do not have densities and possess arbitrarily continuous marginal distributions. A feature selection result with explicit rate is also provided. TCA is also implemented in both numerical simulations and large-scale stock data to illustrate its empirical performance. Both theories and experiments confirm that TCA can achieve model flexibility, estimation accuracy and robustness at almost no cost.

T67 Matrix reconstruction with the local max norm

Rina Foygel rinafb@stanford.edu
Stanford University
Nati Srebro nati@ttic.edu
TTI-Chicago
Russ Salakhutdinov rsalakhu@mit.edu
University of Toronto

We introduce a new family of matrix norms, the ''local max'' norms, generalizing existing methods such as the max norm, the trace norm (nuclear norm), and the weighted or smoothed weighted trace norms, which have been extensively used in the literature as regularizers for matrix reconstruction problems. We show that this new family can be used to interpolate between the (weighted or unweighted) trace norm and the more conservative max norm. We test this interpolation on simulated data and on the large-scale Netflix and MovieLens ratings data, and find improved accuracy relative to the existing matrix norms. We also provide theoretical results showing learning guarantees for some of the new norms.

T68 Multi-criteria Anomaly Detection using Pareto Depth Analysis

Mark Hsiao coolmark@umich.edu
Kevin Xu xukevin@umich.edu
Jeff Calder jcalder@umich.edu
Alfred Hero hero@umich.edu
University of Michigan

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

T69 The Perturbed Variation

Maayan Harel maayanga@tx.technion.ac.il
Shie Mannor shie@ee.technion.ac.il
Technion University

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

T70 A Geometric take on Metric Learning

Søren Hauberg soren.hauberg@tuebingen.mpg.de
Michael Black black@is.mpg.de
Max Planck Institute for Intelligent Systems
Oren Freifeld freifeld@dam.brown.edu
Brown University

Multi-metric learning techniques learn local metric tensors in different parts of a feature space. With such an approach, even simple classifiers can be competitive with the state-of-the-art because the distance measure locally adapts to the structure of the data. The learned distance measure is, however, non-metric, which has prevented multi-metric learning from generalizing to tasks such as dimensionality reduction and regression in a principled way. We prove that, with appropriate changes, multi-metric learning corresponds to learning the structure of a Riemannian manifold. We then show that this structure gives us a principled way to perform dimensionality reduction and regression according to the learned metrics. Algorithmically, we provide the first practical algorithm for computing geodesics according to the learned metrics, as well as algorithms for computing exponential and logarithmic maps on the Riemannian manifold. Together, these tools let many Euclidean algorithms take advantage of multi-metric learning. We illustrate the approach on regression and dimensionality reduction tasks that involve predicting measurements of the human body from shape data.

T71 Semi-supervised Eigenvectors for Locally-biased Learning

Toke Jansen Hansen tokejansenhansen@gmail.com
Technical University of Denmark
Michael Mahoney mahoneymw@gmail.com
Stanford

In many applications, one has information, e.g., labels that are provided in a semi-supervised manner, about a specific target region of a large data set, and one wants to perform machine learning and data analysis tasks nearby that pre-specified target region. Locally-biased problems of this sort are particularly challenging for popular eigenvector-based machine learning and data analysis tools. At root, the reason is that eigenvectors are inherently global quantities. In this paper, we address this issue by providing a methodology to construct semi-supervised eigenvectors of a graph Laplacian, and we illustrate how these locally-biased eigenvectors can be used to perform locally-biased machine learning. These semi-supervised eigenvectors capture successively-orthogonalized directions of maximum variance, conditioned on being well-correlated with an input seed set of nodes that is assumed to be provided in a semi-supervised manner. We also provide several empirical examples demonstrating how these semi-supervised eigenvectors can be used to perform locally-biased learning.

T72 LUCID: Locally Uniform Comparison Image Descriptor

Andrew Ziegler Andrewzieg@gmail.com
Eric Christiansen echristiansen@cs.ucsd.edu
David Kriegman kriegman@cs.ucsd.edu
Serge Belongie sjb@cs.ucsd.edu
UC San Diego

Keypoint matching between pairs of images using popular descriptors like SIFT or a faster variant called SURF is at the heart of many computer vision algorithms including recognition, mosaicing, and structure from motion. For real-time mobile applications, very fast but less accurate descriptors like BRIEF and related methods use a random sampling of pairwise comparisons of pixel intensities in an image patch. Here, we introduce Locally Uniform Comparison Image Descriptor (LUCID), a simple description method based on permutation distances between the ordering of intensities of RGB values between two patches. LUCID is computable in linear time with respect to patch size and does not require floating point computation. An analysis reveals an underlying issue that limits the potential of BRIEF and related approaches compared to LUCID. Experiments demonstrate that LUCID is faster than BRIEF, and its accuracy is directly comparable to SURF while being more than an order of magnitude faster.

T73 Learning to Align from Scratch

Gary Huang gbhuang@cs.umass.edu
Howard Hughes Medical Institute
Marwan Mattar mmattar@cs.umass.edu
Erik Learned-Miller elm@cs.umass.edu
UMass Amherst
Honglak Lee honglak@eecs.umich.edu
University of Michigan

Unsupervised joint alignment of images has been demonstrated to improve performance on recognition tasks such as face verification. Such alignment reduces undesired variability due to factors such as pose, while only requiring weak supervision in the form of poorly aligned examples. However, prior work on unsupervised alignment of complex, real world images has required the careful selection of feature representation based on hand-crafted image descriptors, in order to achieve an appropriate, smooth optimization landscape. In this paper, we instead propose a novel combination of unsupervised joint alignment with unsupervised feature learning. Specifically, we incorporate deep learning into the {\em congealing} alignment framework. Through deep learning, we obtain features that can represent the image at differing resolutions based on network depth, and that are tuned to the statistics of the specific data being aligned. In addition, we modify the learning algorithm for the restricted Boltzmann machine by incorporating a group sparsity penalty, leading to a topographic organization on the learned filters and improving subsequent alignment results. We apply our method to the Labeled Faces in the Wild database (LFW). Using the aligned images produced by our proposed unsupervised algorithm, we achieve a significantly higher accuracy in face verification than obtained using the original face images, prior work in unsupervised alignment, and prior work in supervised alignment. We also match the accuracy for the best available, but unpublished method.

T74 From Deformations to Parts: Motion-based Segmentation of 3D Objects

Soumya Ghosh sghosh@cs.brown.edu
Erik Sudderth sudderth@cs.brown.edu
Brown University
Matthew Loper mloper@tuebingen.mpg.de
Michael Black black@is.mpg.de
Max Planck Institute for Intelligent Systems

We develop a method for discovering the parts of an articulated object from aligned meshes capturing various three-dimensional (3D) poses. We adapt the distance dependent Chinese restaurant process (ddCRP) to allow nonparametric discovery of a potentially unbounded number of parts, while simultaneously guaranteeing a spatially connected segmentation. To allow analysis of datasets in which object instances have varying shapes, we model part variability across poses via affine transformations. By placing a matrix normal-inverse-Wishart prior on these affine transformations, we develop a ddCRP Gibbs sampler which tractably marginalizes over transformation uncertainty. Analyzing a dataset of humans captured in dozens of poses, we infer parts which provide quantitatively better motion predictions than conventional clustering methods.

T75 Max-Margin Structured Output Regression for Spatio-Temporal Action Localization

Du Tran trandu@gmail.com
Dartmouth College
Junsong Yuan jsyuan@ntu.edu.sg
Nanyang Technological University

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

T76 Unsupervised template learning for fine-grained object recognition

Shulin Yang yang@cs.washington.edu
University of Washington
Liefeng Bo liefengbo@gmail.com
Intel Labs

Jue Wang juewang@adobe.com
Adobe
Linda Shapiro shapiro@cs.washington.edu

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

T77 A Generative Model for Parts-based Object Segmentation

Ali Eslami s.m.eslami@sms.ed.ac.uk
Chris Williams ckiw@inf.ed.ac.uk
University of Edinburgh

The Shape Boltzmann Machine (SBM) has recently been introduced as a state-of-the-art model of foreground/ background object shape. We extend the SBM to account for the foreground object's parts. Our model, the Multinomial SBM (MSBM), can capture both local and global statistics of part shapes accurately. We combine the MSBM with an appearance model to form a fully generative model of images of objects. Parts-based image segmentations are obtained simply by performing probabilistic inference in the model. We apply the model to two challenging datasets which exhibit significant shape and appearance variability, and find that it obtains results that are comparable to the state-of-the-art.

T78 Discriminatively Trained Sparse Code Gradients for Contour Detection

Xiaofeng Ren xiaofeng.ren@intel.com
Liefeng Bo liefengbo@gmail.com
Intel Labs

Finding contours in natural images is a fundamental problem that serves as the basis of many tasks such as image segmentation and object recognition. At the core of contour detection technologies are a set of hand-designed gradient features, used by most existing approaches including the state-of-the-art Global Pb (gPb) operator. In this work, we show that contour detection accuracy can be significantly improved by computing Sparse Code Gradients (SCG), which measure contrast using patch representations automatically learned through sparse coding. We use K-SVD and Orthogonal Matching Pursuit for efficient dictionary learning and encoding, and use multi-scale pooling and power transforms to code oriented local neighborhoods before computing gradients and applying linear SVM. By extracting rich representations from pixels and avoiding collapsing them prematurely, Sparse Code Gradients effectively learn how to measure local contrasts and find contours. We improve the F-measure metric on the BSDS500 benchmark to 0.74 (up from 0.71 of gPb contours). Moreover, our learning approach can easily adapt to novel sensor data such as Kinect-style RGB-D cameras: Sparse Code Gradients on depth images and surface normals lead to promising contour detection using depth and depth+color, as verified on the NYU Depth Dataset. Our work combines the concept of oriented gradients with sparse representation and opens up future possibilities for learning contour detection and segmentation.

T79 Analyzing 3D Objects in Cluttered Images

Mohsen Hejrati shejrati@uci.edu
University of California, Irvine
Deva Ramanan dramanan@ics.uci.edu

We present an approach to detecting and analyzing the 3D configuration of objects in real-world images with heavy occlusion and clutter. We focus on the application of finding and analyzing cars. We do so with a two-stage model; the first stage reasons about 2D shape and appearance variation due to within-class variation(station wagons look different than sedans) and changes in viewpoint. Rather than using a view-based model, we describe a compositional representation that models a large number of effective views and shapes using a small number of local view-based templates. We use this model to propose candidate detections and 2D estimates of shape. These estimates are then refined by our second stage, using an explicit 3D model of shape and viewpoint. We use a morphable model to capture 3D within-class variation, and use a weak-perspective camera model to capture viewpoint. We learn all model parameters from 2D annotations. We demonstrate state-of-the-art accuracy for detection, viewpoint estimation, and 3D shape reconstruction on challenging images from the PASCAL VOC 2011 dataset.

T80 Timely Object Recognition

Sergey Karayev sergeyk@eecs.berkeley.edu
Trevor Darrell trevor@eecs.berkeley.edu
UC Berkeley
Tobias Baumgartner tobibaum@gmail.com
RWTH Aachen
Mario Fritz mfritz@mpi-inf.mpg.de
MPI Informatics

In a large visual multi-class detection framework, the timeliness of results can be crucial. Our method for timely multi-class detection aims to give the best possible performance at any single point after a start time; it is terminated at a deadline time. Toward this goal, we formulate a dynamic, closed-loop policy that infers the contents of the image in order to decide which detector to deploy next. In contrast to previous work, our method significantly diverges from the predominant greedy strategies, and is able to learn to take actions with deferred values. We evaluate our method with a novel timeliness measure, computed as the area under an Average Precision vs. Time curve. Experiments are conducted on the eminent PASCAL VOC object detection dataset. If execution is stopped when only half the detectors have been run, our method obtains 66% better AP than a random ordering, and 14% better performance than an intelligent baseline. On the timeliness measure, our method obtains at least 11% better performance. Our code, to be made available upon publication, is easily extensible as it treats detectors and classifiers as black boxes and learns from execution traces using reinforcement learning.

T81 Semantic Kernel Forests from Multiple Taxonomies

Sung Ju Hwang sjhwang@cs.utexas.edu
Kristen Grauman grauman@cs.utexas.edu
University of Texas at Austin
Fei Sha feisha@usc.edu
University of Southern California

When learning features for complex visual recognition problems, labeled image exemplars alone can be insufficient. While an \emph{object taxonomy} specifying the categories' semantic relationships could bolster the learning process, not all relationships are relevant to a given visual classification task, nor does a single taxonomy capture all ties that \emph{are} relevant. In light of these issues, we propose a discriminative feature learning approach that leverages \emph{multiple} hierarchical taxonomies representing different semantic views of the object categories (e.g., for animal classes, one taxonomy could reflect their phylogenic ties, while another could reflect their habitats). For each taxonomy, we first learn a tree of semantic kernels, where each node has a Mahalanobis kernel optimized to distinguish between the classes in its children nodes. Then, using the resulting \ emph{semantic kernel forest}, we learn class-specific kernel combinations to select only those relationships relevant to recognize each object class. To learn the weights, we introduce a novel hierarchical regularization term that further exploits the taxonomies' structure. We demonstrate our method on challenging object recognition datasets, and show that interleaving multiple taxonomic views yields significant accuracy improvements.

T82 3D Object Detection and Viewpoint Estimation with a Deformable 3D Cuboid Model

Sanja Fidler fidler@cs.toronto.edu
Sven Dickinson sven@cs.toronto.edu
University of Toronto
Raquel Urtasun rurtasun@ttic.edu
TTI-Chicago

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

T83 Localizing 3D cuboids in single-view images

Jianxiong Xiao jxiao@csail.mit.edu
Antonio Torralba torralba@csail.mit.edu
Massachusetts Institute of Technology
Bryan Russell brussell@csail.mit.edu
U Washington

In this paper we seek to detect rectangular cuboids and localize their corners in uncalibrated single-view images depicting everyday scenes. In contrast to recent approaches that rely on detecting vanishing points of the scene and grouping line segments to form cuboids, we build a discriminative parts-based detector that models the appearance of the cuboid corners and internal edges while enforcing consistency to a 3D cuboid model. Our model is invariant to the different 3D viewpoints and aspect ratios and is able to detect cuboids across many different object categories. We introduce a database of images with cuboid annotations that spans a variety of indoor and outdoor scenes and show qualitative and quantitative results on our collected database. Our model out-performs baseline detectors that use 2D constraints alone on the task of localizing cuboid corners.

T84 Kernel Latent SVM for Visual Recognition

Weilong Yang weilongyang@gmail.com
Google Inc.
Yang Wang kywang@gmail.com
University of Manitoba
Arash Vahdat avahdat@sfu.ca
Greg Mori mori@cs.sfu.ca
Simon Fraser University

Latent SVMs (LSVMs) are a class of powerful tools that have been successfully applied to many applications in computer vision. However, a limitation of LSVMs is that they rely on linear models. For many computer vision tasks, linear models are suboptimal and nonlinear models learned with kernels typically perform much better. Therefore it is desirable to develop the kernel version of LSVM. In this paper, we propose kernel latent SVM (KLSVM) -- a new learning framework that combines latent SVMs and kernel methods. We develop an iterative training algorithm to learn the model parameters. We demonstrate the effectiveness of KLSVM using three different applications in visual recognition. Our KLSVM formulation is very general and can be applied to solve a wide range of applications in computer vision and machine learning.

T85 Graphical Gaussian Vector for Image Categorization

Tatsuya Harada harada@isi.imi.i.u-tokyo.ac.jp
The University of Tokyo

This paper proposes a novel image representation called a Graphical Gaussian Vector, which is a counterpart of the codebook and local feature matching approaches. In our method, we model the distribution of local features

as a Gaussian Markov Random Field (GMRF) which can efficiently represent the spatial relationship among local features. We consider the parameter of GMRF as a feature vector of the image. Using concepts of information geometry, proper parameters and a metric from the GMRF can be obtained. Finally we define a new image feature by embedding the metric into the parameters, which can be directly applied to scalable linear classifiers. Our method obtains superior performance over the state-of-the-art methods in the standard object recognition datasets and comparable performance in the scene dataset. As the proposed method simply calculates the local auto-correlations of local features, it is able to achieve both high classification accuracy and high efficiency.

T86 Learning about Canonical Views from Internet Image Collections

Elad Mezuman elad.mezuman@mail.huji.ac.il
Yair Weiss yweiss@cs.huji.ac.il
Hebrew University

Although human object recognition is supposedly robust to viewpoint, much research on human perception indicates that there is a preferred or “canonical” view of objects. This phenomenon was discovered more than 30 years ago but the canonical view of only a small number of categories has been validated experimentally. Moreover, the explanation for why humans prefer the canonical view over other views remains elusive. In this paper we ask: Can we use Internet image collections to learn more about canonical views? We start by manually finding the most common view in the results returned by Internet search engines when queried with the objects used in psychophysical experiments. Our results clearly show that the most likely view in the search engine corresponds to the same view preferred by human subjects in experiments. We also present a simple method to find the most likely view in an image collection and apply it to hundreds of categories. Using the new data we have collected we present strong evidence against the two most prominent formal theories of canonical views and provide novel constraints for new theories.

T87 Diffusion Decision Making for Adaptive k-Nearest Neighbor Classification

Yung-Kyun Noh yungkyun.noh@gmail.com
Frank Park fcp@snu.ac.kr
Seoul National University
Daniel Lee ddlee@seas.upenn.edu
University of Pennsylvania

This paper sheds light on some fundamental connections of the diffusion decision making model of neuroscience and cognitive psychology with k-nearest neighbor classification. We show that conventional k-nearest neighbor classification can be viewed as a special problem of the diffusion decision model in the asymptotic situation. Applying the optimal strategy associated with the diffusion decision model, an adaptive rule is developed for determining appropriate values of k in k-nearest neighbor classification. Making use of the sequential probability ratio test (SPRT) and Bayesian analysis, we propose five different criteria for adaptively acquiring nearest neighbors. Experiments with both synthetic and real datasets demonstrate the effectivness of our classification criteria.

T88 Modeling the Forgetting Process using Image Regions

Aditya Khosla khosla@mit.edu
Jianxiong Xiao jxiao@csail.mit.edu
Antonio Torralba torralba@csail.mit.edu
Aude Oliva oliva@csail.mit.edu
Massachusetts Institute of Technology

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

T89 Bayesian Hierarchical Reinforcement Learning

Feng Cao fxc100@case.edu
Soumya Ray sray@case.edu
Case Western Reserve University

We describe an approach to incorporating Bayesian priors in the maxq framework for hierarchical reinforcement learning (HRL). We define priors on the primitive environment model and on task pseudo-rewards. Since models for composite tasks can be complex, we use a mixed model-based/model-free learning approach to find an optimal hierarchical policy. We show empirically that (i) our approach results in improved convergence over non-Bayesian baselines, given sensible priors, (ii) task hierarchies and Bayesian priors can be complementary sources of information, and using both sources is better than either alone, (iii) taking advantage of the structural decomposition induced by the task hierarchy significantly reduces the computational cost of Bayesian reinforcement learning and (iv) in this framework, task pseudo-rewards can be learned instead of being manually specified, leading to automatic learning of hierarchically optimal rather than recursively optimal policies.

T90 Spectral learning of linear dynamics from generalised-linear observations with application to neural population data

Lars Buesing lars@gatsby.ucl.ac.uk
Maneesh Sahani maneesh@gatsby.ucl.ac.uk
University College London
Jakob Macke Jakob.Macke@gmail.com
Max Planck Institute and Bernstein Center Tübingen

Latent linear dynamical systems with generalised-linear observation models arise in a variety of applications, for example when modelling the spiking activity of populations of neurons. Here, we show how spectral learning methods for linear systems with Gaussian observations (usually called subspace identification in this context) can be extended to estimate the parameters of dynamical system models observed through non-Gaussian noise models. We use this approach to obtain estimates of parameters for a dynamical model of neural population data, where the observed spike-counts are Poisson-distributed with log-rates determined by the latent dynamical process, possibly driven by external inputs. We show that the extended system identification algorithm is consistent and accurately recovers the correct parameters on large simulated data sets with much smaller computational cost than approximate expectation-maximisation (EM) due to the non-iterative nature of subspace identification. Even on smaller data sets, it provides an effective initialization for EM, leading to more robust performance and faster convergence. These benefits are shown to extend to real neural data.

T91 A systematic approach to extracting semantic information from functional MRI data

Francisco Pereira francisco.pereira@gmail.com
Siemens Corporation
Matthew Botvinick matthewb@princeton.edu
Princeton University

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

T92 Fully Bayesian inference for neural models with negative-binomial spiking

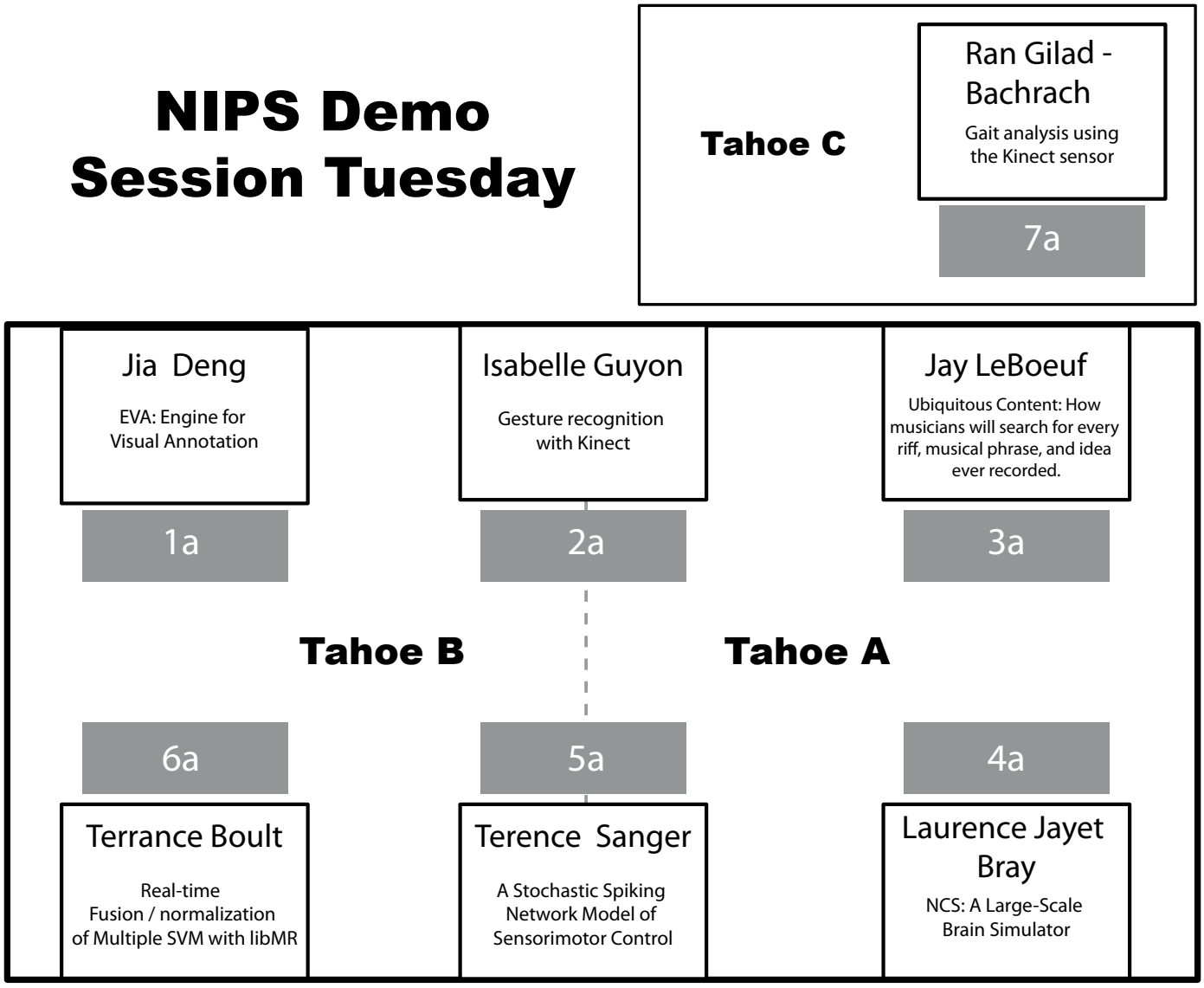
Jonathan Pillow pillow@mail.utexas.edu
James Scott James.Scott@mccombs.utexas.edu
University of Texas at Austin

Characterizing the information carried by neural populations in the brain requires accurate statistical models of neural spike responses. The negative-binomial distribution provides a convenient model for over-dispersed spike counts, that is, responses with greater-than-Poisson variability. Here we describe a powerful data-augmentation framework for fully Bayesian inference in neural models with negative-binomial spiking. Our approach relies on a recently described latent-variable representation of the negative-binomial distribution, which equates it to a Polya-gamma mixture of normals. This framework provides a tractable, conditionally Gaussian representation of the posterior that can be used to design efficient EM and Gibbs sampling based algorithms for inference in regression and dynamic factor models. We apply the model to neural data from primate retina and show that it substantially outperforms Poisson regression on held-out data, and reveals latent structure underlying spike count correlations in simultaneously recorded spike trains.

T93 Bayesian active learning with localized priors for fast receptive field characterization

Mijung Park mjpark@mail.utexas.edu
The University of Texas at Austin
Jonathan Pillow pillow@mail.utexas.edu
University of Texas at Austin

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



1A A Stochastic Spiking Network Model of Sensorimotor Control

A. Ghoreyshi, T. Sanger, J. Rocamora

We demonstrate the implementation of a new spike-based control scheme on an Omni robot arm. Our robotic demonstrations replicate a variety of complex human motor behavior, such as reaching and tracking, reflexes in task-relevant directions, and reward/penalty trade-off responses. Our scheme can provide a neurophysiological explanation of specific human sensorimotor functions under uncertainty.

Users can see and try out spike-based tracking ability of the arm using spheres representing reward (wanted) regions. They can also try and feel the robot's response to perturbations. They can try the robot's response to the simultaneous presence of reward (wanted) and penalty (feared) regions.

2A EVA: Engine for Visual Annotation

J. Deng, J. Krause, Z. Huang, A. Berg, F. Li

The EVA system, powered by ImageNet, recognizes over 20K visual classes. Using the DARTS algorithm (Deng et al. 2012), EVA is able to name objects in an image as informatively as possible while ensuring an arbitrarily high accuracy. This live demo showcases EVA on images supplied by a user.

3A Gait analysis using the Kinect sensor

M. Gabel, E. Renshaw, A. Schuster, R. Gilad - Bachrach

The human gait is an important health indicator. Gradual changes to gate are indicators to diseases such as Parkinson's disease, diabetes, Alzheimer and many others. Yet most technologies existing today for tracking changes in gate are either very expensive or intrusive. In this work we demonstrate how the Kinect sensor can be used to track gait. We build a layer of machine learned models that predict rich set of gait features from the skeleton model extracted by the Kinect SDK. These predictions were calibrated against wearable sensors and were shown to provide accurate predictions.

This approach has several advantages over existing techniques: it is affordable, accurate and not intrusive. Therefore, it allows for continuous monitoring of gait on large populations.

4A Gesture recognition with Kinect

I. Guyon

The Microsoft Kinect camera has revolutionized computer vision by providing an affordable 3D camera (RGB+depth). The applications, initially driven by the game industry, are rapidly diversifying and include video surveillance, computer interfaces, robot vision and control, and education. We recently organized two demonstration competitions of gesture recognition with Kinect(TM) in the context of the ChaLearn gesture challenge (http://gesture.chalearn.org). Videos and live demonstrations of the winning entries will be shown, including a hand gesture recognizer for American Sign Language (ASL), games using hand and arm gestures, and a head position recognizer.

5A NCS: A Large-Scale Brain Simulator

L. Jayet Bray, D. Tanna, F. Harris, Jr

Many different scales of experiments in neuroscience research attempt to clarify the functions of the brain. From the genetics of single molecules to the behavioral research of cognitive neuroscience, studies lead to a better understanding of nervous systems. When in vivo and in vitro experiments are impossible due to the complexity of neuronal structures or the lack of equipment, computational neuroscience provides new opportunities. Its unique access to any brain regions as well as its different levels of abstraction allows biologically-realistic neural simulations, and thus faster neuroscience advances.

However, neural simulations have always involved a trade-off between execution time and biophysical realism. Even as neuron models are simplified and approximated, the neural regions of interest may require an unreasonable amount of running time. To further drive computational neuroscience research, computer scientists and engineers have created optimized simulation programs and advanced hardware architecture, respectively.

We present a novel CPU/GPU simulation environment for large-scale neural modeling, called the Neocortical Simulator (NCS) version 6. At the cellular level NCS implements several built-in neuron models (e.g. Izhikevich, leaky integrate-and-fire). Computationally, shared-memory multiprocessor architectures and recent experiments with clustered GPUs indicate that we are close to simulate a million cells in real time without sacrificing biological detail. Our demonstration will be based on the design of large-scale brain models using NCS, and real-time simulations on a single or multiple machine(s).

6A Real-time Fusion/normalization of Multiple SVM with libMR

T. Boult

Learned visual attributes have become an important feature for recognition and search. Visual Attributes can be computed using SVMs, which then require normalization and fusion for effective searching. In this demo, we present a MugHunt, a highly scalable Google-like face search system based around Meta-Recognition fusion and implemented using of Cassandra, allowing searching over millions of face images using attribute descriptions (gender, race, hair color, expression, etc.). This demo will also include a look under the covers at the Meta-Recognition based normalization and implementation of similarity searching.

7A Ubiquitous Content: How musicians will search for every riff, musical phrase, and idea ever recorded.

J. LeBoeuf

This demonstration will showcase iZotope's MediaMined Discover - a search engine for sound. MediaMined was developed by Imagine Research (acquired by iZotope in 2012) and awarded multiple National Science Foundation Small Business Innovation Research grants (Phase I, Phase II, and Phase IIB). MediaMined uses state of the art digital signal processing and machine learning to analyze and index content.

This search engine enables a simple, intuitive workflow: "find me examples that sound like this." Musicians and engineers repeat this process - layering up musical phrases, loops, or samples that are rhythmically or timbrally what they hear in their mind - a process that is natural, fast, enjoyable, and intuitive.

WEDNESDAY

ORAL SESSION

SESSION 5 - 9:00 – 10:10 AM

Session Chair: Fei Sha

POSNER LECTURE: Challenges for Machine Learning in Computational Sustainability

Thomas Dietterich tgd@cs.orst.edu
Oregon State University

Research in computational sustainability seeks to develop and apply methods from computer science to the many challenges of managing the earth's ecosystems sustainably. Viewed as a control problem, ecosystem management is challenging for two reasons. First, we lack good models of the function and structure of the earth's ecosystems. Second, it is difficult to compute optimal management policies because ecosystems exhibit complex spatio-temporal interactions at multiple scales. This talk will discuss some of the many challenges and opportunities for machine learning research in computational sustainability. These include sensor placement, data interpretation, model fitting, computing robust optimal policies, and finally executing those policies successfully. Examples will be discussed on current work and open problems in each of these problems. All of these sustainability problems involve spatial modeling and optimization, and all of them can be usefully conceived in terms of facilitating or preventing flows along edges in spatial networks. For example, encouraging the recovery of endangered species involves creating a network of suitable habitat and encouraging spread along the edges of the network. Conversely, preventing the spread of diseases, invasive species, and pollutants involves preventing flow along edges of networks. Addressing these problems will require advances in several areas of machine learning and optimization.

Tom Dietterich *(AB Oberlin College 1977; MS University of Illinois 1979; PhD Stanford University 1984) is Professor and Director of Intelligent Systems Research at Oregon State University. Among his contributions to machine learning research are (a) the formalization of the multiple-instance problem, (b) the development of the error-correcting output coding method for multi-class prediction, (c) methods for ensemble learning, (d) the development of the MAXQ framework for hierarchical reinforcement learning, and (e) the application of gradient tree boosting to problems of structured prediction and latent variable models. Dietterich has pursued application-driven fundamental research in many areas including drug discovery, computer vision, computational sustainability, and intelligent user interfaces. Dietterich has served the machine learning community in a variety of roles including Executive Editor of the Machine Learning journal, co-founder of the Journal of Machine Learning Research, editor of the MIT Press Book Series on Adaptive Computation and Machine Learning, and editor of the Morgan-Claypool Synthesis series on Artificial Intelligence and Machine Learning. He was Program Co-Chair of AAAI-1990, Program Chair of NIPS-2000, and General Chair of NIPS-2001. He was first President of the International Machine Learning Society (the parent organization of ICML) and served a term on the NIPS Board of Trustees and the Council of AAAI.*

Augmented-SVM: Automatic space partitioning for combining multiple non-linear dynamics

Ashwini Shukla ashwini.shukla@epfl.ch
Aude Billard karin.elsea@epfl.ch
Ecole Polytechnique Fédérale de Lausanne

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

SPOTLIGHT SESSION

SESSION 5 - 10:10 – 10:30 AM

- **Majorization for CRFs and Latent Likelihoods**
  T. Jebara, A. Choromanska, Columbia University
  See abstract W63, page 96

- **Kernel Hyperalignment**
  A. Lorbert, P. Ramadge, Princeton University
  See abstract W62, page 96

- **Learning Networks of Heterogeneous Influence**
  N. DU, Georgia Institute of Technology; L. Song, Georgia Tech; A. Smola, Google; M. Yuan,
  See abstract W61, page 96

- **Learning from Distributions via Support Measure Machines**
  K. Muandet, F. Dinuzzo, B. Schölkopf, Max Planck Institute for Intelligent Systems; K. Fukumizu, Institute of Statistical Mathematics
  See abstract Th54, page 94

- **Multiclass Learning Approaches: A Theoretical Comparison with Implications**
  A. Daniely, Hebrew university; S. Sabato, Microsoft Research; S. Shalev-Shwartz, Hebrew University
  See abstract W58, page

ORAL SESSION

SESSION 6 - 11:00 - 11:40 AM

Session Chair: Doina Precup

On the Use of Non-Stationary Policies for Stationary Infinite-Horizon Markov Decision Processes

Bruno Scherrer scherrer@loria.fr
INRIA Loria
Boris Lesner boris.lesner@inria.fr
INRIA

We consider infinite-horizon stationary γ-discounted Markov Decision Processes, for which it is known that there exists a stationary optimal policy. Using Value and Policy Iteration with some error ϵ at each iteration, it is well-known that one can compute stationary policies that are \frac{2\gamma{(1-\ gamma)^2}\epsilon-optimal. After arguing that this guarantee is tight, we develop variations of Value and Policy Iteration for computing non-stationary policies that can be up to 2γ1−γϵ-optimal, which constitutes a significant improvement in the usual situation when γ is close to 1. Surprisingly, this shows that the problem of ``computing near-optimal non-stationary policies'' is much simpler than that of ``computing near-optimal stationary policies''.

A Unifying Perspective of Parametric Policy Search Methods for Markov Decision Processes

Thomas Furmston T.Furmston@cs.ucl.ac.uk
David Barber davidobarber@gmail.com
University College London

Parametric policy search algorithms are one of the methods of choice for the optimisation of Markov Decision Processes, with Expectation Maximisation and natural gradient ascent being considered the current state of the art in the field. In this article we provide a unifying perspective of these two algorithms by showing that their step-directions in the parameter space are closely related to the search direction of an approximate Newton method. This analysis leads naturally to the consideration of this approximate Newton method as an alternative gradient-based method for Markov Decision Processes. We are able show that the algorithm has numerous desirable properties, absent in the naive application of Newton's method, that make it a viable alternative to either Expectation Maximisation or natural gradient ascent. Empirical results suggest that the algorithm has excellent convergence and robustness properties, performing strongly in comparison to both Expectation Maximisation and natural gradient ascent.

SPOTLIGHT SESSION

SESSION 6 - 11:40 AM - 12:05 PM

- **Near Optimal Chernoff Bounds for Markov Decision Processes**
  T. Moldovan, UC Berkeley; P. Abbeel, Berkrley
  See abstract W11, page 84

- **Regularized Off-Policy TD-Learning**
  B. Liu, University of Massachusetts; S. Mahadevan, University of Massachusetts Amherst; J. Liu, University Wisconsin-Madison
  See abstract W88, page 102

- **Human memory search as a random walk in a semantic network**
  J. Abbott, T. Griffiths, University of California, Berkeley; J. Austerweil, UC Berkeley
  See abstract W86, page 102

- **Locating Changes in Highly Dependent Data with Unknown Number of Change Points**
  A. Khaleghi, INRIA Lille - Nord Europe; D. Ryabko, INRIA
  See abstract W68, page 97

- **Classification Calibration Dimension for General Multiclass Losses**
  H. Guruprasad, S. Agarwal, Indian Institute of Science
  See abstract W70, page 98

- **Multi-Stage Multi-Task Feature Learning**
  P. Gong, C. Zhang, Tsinghua University; J. Ye, Arizona State University
  See abstract Th29, page 118

ORAL SESSION

SESSION 7 - 2:00 - 3:30 PM

Session Chair: Katherine Heller

INVITED TALK: Signatures of Conscious Processing in the Human Brain

Stanislas Dehaene stanislas.dehaene@cea.fr
Collège de France & CEA

Understanding how brain activity leads to a conscious experience remains a major experimental challenge. I will describe a series of experiments that probe the signatures of conscious processing. In these experiments, my colleagues and I ask whether a specific type of brain activity can be detected when a person suddenly becomes aware of a piece of information. We create minimal contrasts whereby the very same visual stimulus is sometimes undetected, and sometimes consciously seen. We then use time-resolved methods of electro- and magnetoencephalography to follow the time course of brain activity. The results show that conscious access relates to a global burst of late synchronized activity (a cortical "ignition"), distributed through many cortical areas. We propose a theory of a global

neuronal workspace, according to which what we experience as a consciousness is the global availability of information in a large-scale network of pyramidal neurons with long-distance axons. This knowledge is now being applied to the monitoring of conscious states in non-communicating patients. Using real-time signal processing techniques, we believe that a few minutes of testing with simple sounds and two recording electrodes might suffice to determine whether a person is conscious.

Stanislas Dehaene is professor at the Collège de France, where he holds the chair of Experimental Cognitive Psychology. He directs the INSERM-CEA Cognitive Neuroimaging Unit at NeuroSpin in Saclay, south of Paris -- France's advanced neuroimaging research center. His research investigates the neural bases of human cognitive functions such as reading, calculation and language, with a particular interest for the differences between conscious and non-conscious processing. His main research findings include the discovery of automatic links between numbers and space, and of the role of the intraparietal sulcus in number sense; the operation of the "visual word form area", a left occipito-temporal region which acquires the visual component of reading; and the identification of physiological responses unique to conscious processing, supporting the theory of a "global neuronal workspace" for consciousness. He is the author of "The number sense" (1997/2010) and "Reading in the brain" (2009), and the editor of "The cognitive neuroscience of consciousness" (2001) and "From monkey brain to human brain" (2007). He is a member of the French and US Academies of Sciences and the Pontifical Academy of Sciences, and the recipient of the Dr. A.H. Heineken Prize in Cognitive Science (2009) and the McDonnell Centennial Award (1999).

Privacy Aware Learning

John Duchi jduchi@cs.berkeley.edu
Michael Jordan jordan@cs.berkeley.edu
Martin Wainwright wainwrig@eecs.berkeley.edu
UC Berkeley

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

On the (Non-)existence of Convex, Calibrated Surrogate Losses for Ranking

Clément Calauzènes clement.calauzenes@lip6.fr
Nicolas Usunier nicolas.usunier@lip6.fr
Université Pierre et Marie Curie
Patrick Gallinari patrick.gallinari@lip6.fr
University Paris 6

We study surrogate losses for learning to rank, in a framework where the rankings are induced by scores and the task is to learn the scoring function. We focus on the calibration of surrogate losses with respect to a ranking evaluation metric, where the calibration is equivalent to the guarantee that near-optimal values of the surrogate risk imply near-optimal values of the risk defined by the evaluation metric. We prove that if a surrogate loss is a convex function of the scores, then it is not calibrated with respect to two evaluation metrics widely used for search engine evaluation, namely the Average Precision and the Expected Reciprocal Rank. We also show that such convex surrogate losses cannot be calibrated with respect to the Pairwise Disagreement, an evaluation metric used when learning from pairwise preferences. Our results cast lights on the intrinsic difficulty of some ranking problems, as well as on the limitations of learning-to-rank algorithms based on the minimization of a convex surrogate risk.

SPOTLIGHT SESSION

SESSION 7 - 3:30 - 3:50 PM

- **Statistical Consistency of Ranking Methods in A Rank-Differentiable Probability Space**
  Y. Lan, X. Cheng, ICT; J. Guo, ; T. Liu, Microsoft
  See abstract W65, page 97

- **Iterative ranking from pair-wise comparisons**
  S. Negahban, University of California, Berkeley; S. Oh, University of Illinois at Urbana Champaign; D. Shah, Massachusetts Institute of Technology
  See abstract Th7, page 113

- **Hierarchical spike coding of sound**
  y. karklin, New York University; C. Ekanadham, Courant Institute, New York University; E. Simoncelli, HHMI / New York University
  See abstract W9, page 83

- **Patient Risk Stratification for Hospital-Associated C. Diff as a Time-Series Classification Task**
  J. Wiens, J. Guttag, Massachusetts Institute of Technology; E. Horvitz, Microsoft Research
  See abstract Th2, page 112

- **On Multilabel Classification and Ranking with Partial Feedback**
  C. Gentile, Universita' dell'Insubria; F. Orabona, Toyota Technological Institute at Chicago
  See abstract Th65, page 126

ORAL SESSION

SESSION 8 - 4:20 - 5:40 PM

Session Chair: Tiberio Caetano

Graphical Models via Generalized Linear Models

Eunho Yang eunho@cs.utexas.edu
Pradeep Ravikumar pradeepr@cs.utexas.edu
University of Texas, Austin
Genevera Allen giallen@stanford.edu
Rice University
zhandong Liu zhandong@mail.med.upenn.edu
University of Pennsylvania

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

No voodoo here! Learning discrete graphical models via inverse covariance estimation

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

We investigate the relationship between the support of the inverses of generalized covariance matrices and the structure of a discrete graphical model. We show that for certain graph structures, the support of the inverse covariance matrix of indicator variables on the vertices of a graph reflects the conditional independence structure of the graph. Our work extends results which were previously established only for multivariate Gaussian distributions, and partially answers an open question about the meaning of the inverse covariance matrix of a non-Gaussian distribution. We propose graph selection methods for a general discrete graphical model with bounded degree based on possibly corrupted observations, and verify our theoretical results via simulations. Along the way, we also establish new results for support recovery in the setting of sparse high-dimensional linear regression based on corrupted and missing observations.

Near-Optimal MAP Inference for Determinantal Point Processes

Alex Kulesza kulesza@umich.edu
University of Michigan
Jennifer Gillenwater jengi@cis.upenn.edu
Ben Taskar taskar@cis.upenn.edu
University of Pennsylvania

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

Bayesian nonparametric models for bipartite graphs

Francois Caron Francois.Caron@inria.fr
INRIA Bordeaux

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

SPOTLIGHT SESSION

SESSION 8 - 5:40 - 6:00 PM

- **Scalable Inference of Overlapping Communities**
  P. Gopalan, D. Mimno, S. Gerrish, M. Freedman, D. Blei, Princeton University
  See abstract W29, page 88

- **Bayesian Nonparametric Modeling of Suicide Attempts**
  F. Ruiz, F. Perez-Cruz, University Carlos III in Madrid; I. Valera, University Carlos III at Madrid; C. Blanco, Columbia University College of Physicians and Surgeons
  See abstract Th41, page 121

- **Augment-and-Conquer Negative Binomial Processes**
  M. Zhou, Duke University; L. Carin, Duke
  See abstract W31, page 89

- **Symmetric Correspondence Topic Models for Multilingual Text Analysis**
  K. Fukumasu, K. Eguchi, Kobe University; E. Xing, Carnegie Mellon University
  See abstract W82, page 101

- **A Spectral Algorithm for Latent Dirichlet Allocation**
  A. Anandkumar, U.C.Irvine; D. Foster, University of Pennsylvania; D. Hsu, S. Kakade, Microsoft Research; Y. Liu, National Institute of Standards and Technology
  See abstract W66, page 97

POSTER SESSION

AND RECEPTION - 5:45 – 11:59 PM

W1 **Super-Bit Locality-Sensitive Hashing**
J. Ji, J. Li, s. yan, B. Zhang, Q. Tian

W2 **Learning Invariant Representations of Molecules for Atomization Energy Prediction**
G. Montavon, K. Hansen, S. Fazli, M. Rupp, F. Biegler, A. Ziehe, A. Tkatchenko, A. von Lilienfeld, K. Müller

W3 **Adaptive Learning of Smoothing Functions: Application to Electricity Load Forecasting**
A. Ba, M. Sinn, Y. Goude, P. Pompey

W4 **Automatic Feature Induction for Stagewise Collaborative Filtering**
J. Lee, M. Sun, S. Kim, G. Lebanon

W5 **Co-Regularized Hashing for Multimodal Data**
Y. Zhen, D. Yeung

W6 **Label Ranking with Partial Abstention based on Thresholded Probabilistic Models**
W. Cheng, E. Huellermeier, W. Waegeman, V. Welker

W7 **Random Utility Theory for Social Choice: Theory and Algorithms**
H. Azari, D. Parkes, L. Xia

W8 **Unsupervised Structure Discovery for Semantic Analysis of Audio**
S. Chaudhuri, B. Raj

W9 **Hierarchical spike coding of sound**
y. karklin, C. Ekanadham, E. Simoncelli

W10 **Nonparametric Bayesian Inverse Reinforcement Learning for Multiple Reward Functions**
J. Choi, K. Kim

W11 **Near Optimal Chernoff Bounds for Markov Decision Processes**
T. Moldovan, P. Abbeel

W12 **Sketch-Based Linear Value Function Approximation**
M. Bellemare, J. Veness, M. Bowling

W13 **A Unifying Perspective of Parametric Policy Search Methods for Markov Decision Processes**
T. Furmston, D. Barber

W14 **Large Scale Distributed Deep Networks**
J. Dean, G. Corrado, R. Monga, K. Chen, M. Devin, Q. Le, M. Mao, M. Ranzato, A. Senior, P. Tucker, K. Yang, A. Ng

W15 **Discriminative Learning of Sum-Product Networks**
R. Gens, P. Domingos

W16 **A Neural Autoregressive Topic Model**
H. Larochelle, S. Lauly

W17 **Feature-aware Label Space Dimension Reduction for Multi-label Classification**
Y. Chen, H. Lin

W18 **Multi-task Vector Field Learning**
B. Lin, S. Yang, C. Zhang, J. Ye, X. He

W19 **Minimax Multi-Task Learning and a Generalized Loss-Compositional Paradigm for MTL**
N. Mehta, D. Lee, A. Gray

W20 **Communication-Efficient Algorithms for Statistical Optimization**
Y. Zhang, J. Duchi, M. Wainwright

W21 **Multilabel Classification using Bayesian Compressed Sensing**
A. Kapoor, R. Viswanathan, P. Jain

W22 **MCMC for continuous-time discrete-state systems**
V. Rao, Y. Teh

W23 **Ancestor Sampling for Particle Gibbs**
F. Lindsten, M. Jordan, T. Schön

W24 **Probabilistic Event Cascades for Alzheimer's disease**
J. Huang, D. Alexander

W25 **Bayesian Pedigree Analysis using Measure Factorization**
B. Kirkpatrick, A. Bouchard-Côté

W26 **Fiedler Random Fields: A Large-Scale Spectral Approach to Statistical Network Modeling**
A. Freno, M. Keller, M. Tommasi

W27 Distributed Probabilistic Learning for Camera Networks with Missing Data
S. Yoon, V. Pavlovic

W28 No voodoo here! Learning discrete graphical models via inverse covariance estimation
P. Loh, M. Wainwright

W29 Scalable Inference of Overlapping Communities
P. Gopalan, D. Mimno, S. Gerrish, M. Freedman, D. Blei

W30 Probabilistic Topic Coding for Superset Label Learning
L. Liu, T. Dietterich

W31 Augment-and-Conquer Negative Binomial Processes
M. Zhou, L. Carin

W32 Priors for Diversity in Generative Latent Variable Models
J. Zou, R. Adams

W33 Dynamic Pruning of Factor Graphs for Maximum Marginal Prediction
C. Lampert

W34 Variational Inference for Crowdsourcing
Q. Liu, J. Peng, A. Ihler

W35 Near-Optimal MAP Inference for Determinantal Point Processes
A. Kulesza, J. Gillenwater, B. Taskar

W36 Affine Independent Variational Inference
E. Challis, D. Barber

W37 Truly Nonparametric Online Variational Inference for Hierarchical Dirichlet Processes
M. Bryant, E. Sudderth

W38 Effective Split-Merge Monte Carlo Methods for Nonparametric Models of Sequential Data
M. Hughes, E. Fox, E. Sudderth

W39 Truncation-free Online Variational Inference for Bayesian Nonparametric Models
C. Wang, D. Blei

W40 Modelling Reciprocating Relationships
C. Blundell, K. Heller, J. Beck

W41 A nonparametric variable clustering model
D. Knowles, K. Palla, Z. Ghahramani

W42 Bayesian nonparametric models for ranked data
F. Caron, Y. Teh

W43 A Marginalized Particle Gaussian Process Regression
Y. Wang, B. Chaib-draa

W44 Structured Sparse Learning of Multiple Gaussian Graphical Models
K. Mohan, M. Chung, S. Han, D. Witten, S. Lee, M. Fazel

W45 Topology Constraints in Graphical Models
M. Fiori, P. Musé, G. Sapiro

W46 Learning Mixtures of Tree Graphical Models
A. Anandkumar, D. Hsu, F. Huang, S. Kakade

W47 A Convex Formulation for Learning Scale-Free Networks via Submodular Relaxation
A. Defazio, T. Caetano

W48 Graphical Models via Generalized Linear Models
E. Yang, P. Ravikumar, G. Allen, z. Liu

W49 A latent factor model for highly multi-relational data
R. Jenatton, N. Le Roux, A. Bordes, G. Obozinski

W50 On Triangular versus Edge Representations --- Towards Scalable Modeling of Networks
Q. Ho, J. Yin, E. Xing

W51 Efficient high dimensional maximum entropy modeling via symmetric partition functions
P. Vernaza, D. Bagnell

W52 Simultaneously Leveraging Output and Task Structures for Multiple-Output Regression
P. Rai, A. Kumar, H. Daume III

W53 Augmented-SVM: Automatic space partitioning for combining multiple non-linear dynamics
A. Shukla, A. Billard

W54 Local Supervised Learning through Space Partitioning
J. Wang, V. Saligrama

W55 Mandatory Leaf Node Prediction in Hierarchical Multilabel Classification
W. Bi, J. Kwok

W56 Proper losses for learning from partial labels
J. Cid-Sueiro

W57 Multiclass Learning with Simplex Coding
Y. Mroueh, T. Poggio, L. Rosasco, J. Slotine

W58 Multiclass Learning Approaches: A Theoretical Comparison with Implications
A. Daniely, S. Sabato, S. Shalev-Shwartz

W59 Learning as MAP Inference in Discrete Graphical Models
T. Caetano, X. Liu, J. Petterson

W60 A new metric on the manifold of kernel matrices with application to matrix geometric means
S. Sra

W61 Learning Networks of Heterogeneous Influence
N. DU, L. Song, A. Smola, M. Yuan

W62 Kernel Hyperalignment
A. Lorbert, P. Ramadge

W63 Majorization for CRFs and Latent Likelihoods
T. Jebara, A. Choromanska

W64 Privacy Aware Learning
J. Duchi, M. Jordan, M. Wainwright

W65 Statistical Consistency of Ranking Methods in A Rank-Differentiable Probability Space
Y. Lan, J. Guo, X. Cheng, T. Liu

W66 A Spectral Algorithm for Latent Dirichlet Allocation
A. Anandkumar, D. Foster, D. Hsu, S. Kakade, Y. Liu

W67 Distributed Non-Stochastic Experts
V. Kanade, Z. Liu, B. Radunovic

W68 Locating Changes in Highly Dependent Data with Unknown Number of Change Points
A. Khaleghi, D. Ryabko

W69 On the (Non-)existence of Convex, Calibrated Surrogate Losses for Ranking
C. Calauzènes, N. Usunier, P. Gallinari

W70 Classification Calibration Dimension for General Multiclass Losses
H. Guruprasad, S. Agarwal

W71 A Linear Time Active Learning Algorithm for Link Classification
N. Cesa-Bianchi, C. Gentile, F. Vitale, G. Zappella

W72 Relax and Randomize : From Value to Algorithms
A. Rakhlin, O. Shamir, K. Sridharan

W73 Mirror Descent Meets Fixed Share (and feels no regret)
N. Cesa-Bianchi, P. Gaillard, G. Lugosi, G. Stoltz

W74 Best Arm Identification: A Unified Approach to Fixed Budget and Fixed Confidence
V. Gabillon, M. Ghavamzadeh, A. Lazaric

W75 Mixability in Statistical Learning
T. van Erven, P. Grunwald, M. Reid, R. Williamson

W76 No-Regret Algorithms for Unconstrained Online Convex Optimization
M. Streeter, B. McMahan

W77 Confusion-Based Online Learning and a Passive-Aggressive Scheme
L. Ralaivola

W78 Dimensionality Dependent PAC-Bayes Margin Bound
C. Jin, L. Wang

W79 The variational hierarchical EM algorithm for clustering hidden Markov models.
E. Coviello, A. Chan, G. Lanckriet

W80 FastEx: Fast Clustering with Exponential Families
A. Ahmed, S. Ravi, S. Narayanamurthy, A. Smola

W81 Clustering Sparse Graphs
Y. Chen, S. Sanghavi, H. Xu

W82 Symmetric Correspondence Topic Models for Multilingual Text Analysis
K. Fukumasu, K. Eguchi, E. Xing

W83 Factorial LDA: Sparse Multi-Dimensional Text Models
M. Paul, M. Dredze

W84 Visual Recognition using Embedded Feature Selection for Curvature Self-Similarity
A. Eigenstetter, B. Ommer

W85 Context-Sensitive Decision Forests for Object Detection
P. Kontschieder, S. Bulò, A. Criminisi, P. Kohli, M. Pelillo, H. Bischof

W86 Human memory search as a random walk in a semantic network
J. Abbott, J. Austerweil, T. Griffiths

W87 Transferring Expectations in Model-based Reinforcement Learning
T. Nguyen, T. Silander, T. Leong

W88 Regularized Off-Policy TD-Learning
B. Liu, S. Mahadevan, J. Liu

W89 Weighted Likelihood Policy Search with Model Selection
T. Ueno, Y. Kawahara, K. Hayashi, T. Washio

W90 A mechanistic model of early sensory processing based on subtracting sparse representations
S. Druckmann, T. Hu, D. Chklovskii

W91 High-Order Multi-Task Feature Learning to Identify Longitudinal Phenotypic Markers for Alzheimer Disease Progression Prediction
H. Wang, F. Nie, H. Huang, J. Yan, S. Kim, S. Risacher, A. Saykin, L. Shen

W92 Optimal Neural Tuning Curves for Arbitrary Stimulus Distributions: Discrimax, Infomax and Minimum Lp Loss
Z. Wang, A. Stocker, D. Lee

W93 Identifiability and Unmixing of Latent Parse Trees
P. Liang, S. Kakade, D. Hsu

DEMONSTRATIONS

7:00 - 11:59 PM

1B A Fast Accurate Training-less P300 Speller: Unsupervised Learning Uncovers new Possibilities
P. Kindermans, H. Verschore, D. Verstraeten, B. Schrauwen

2B Cynomix: A Machine Learning Aided Workbench for Rapid Comprehension of Large Malware Corpora
J. Saxe, D. Mentis, C. Greamo

3B DIRTBIS - Distributed Real-Time Bayesian Inference Service
R. Herbrich

4B GraphLab: A Framework For Machine Learning in the Cloud
Y. Low, H. Gu, C. Guestrin

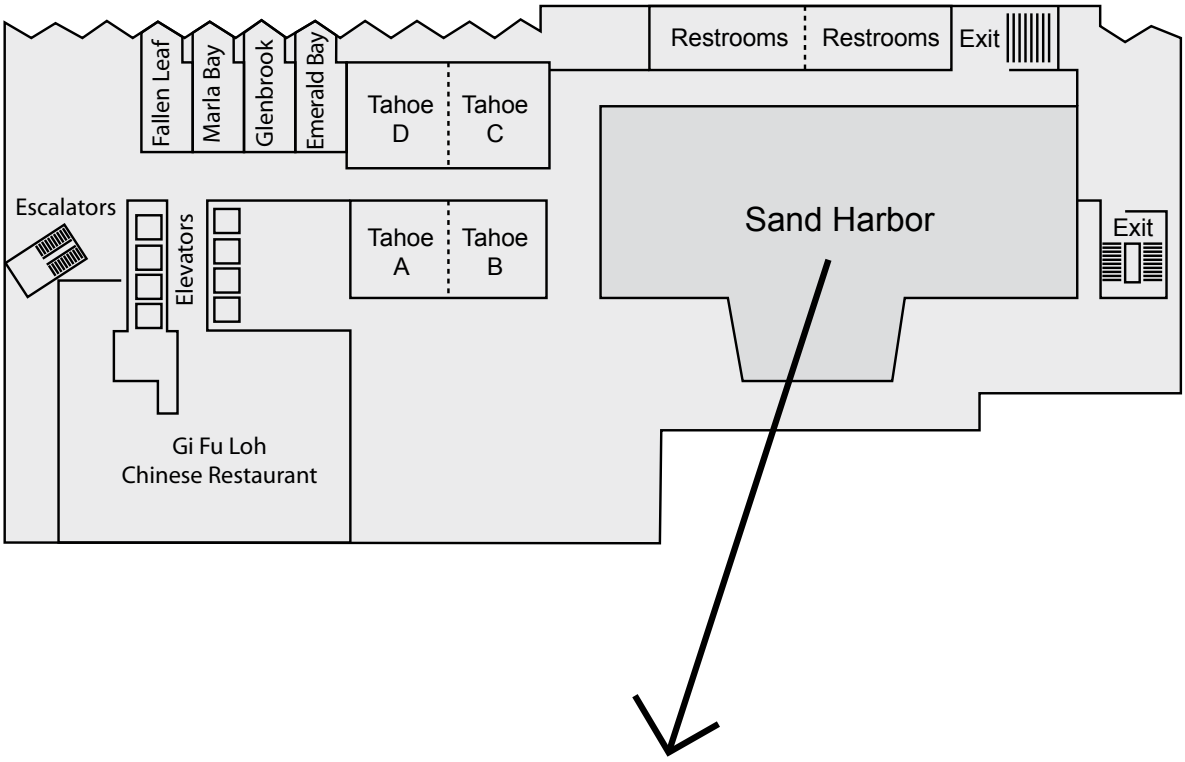
5B Hardware Accelerated Belief Propagation
S. Hershey, B. Vigoda

6B Protocols and Structures for Inference: A RESTful API for Machine Learning Services
J. Montgomery, M. Reid

7B The BUDS POMDP Spoken Dialogue System
M. Szummer, M. Henderson, C. Breslin, M. Gasic, D. Kim, B. Thomson, P. Tsiakoulis, S. Young

HARRAH'S

2ND FLOOR SPECIAL EVENTS CENTER



SAND HARBOR

	89	88	81	80	73	72	65	64	57	56	49
	90	87	82	79	74	71	66	63	58	55	50
	91	86	83	78	75	70	67	62	59	54	51
93	92	85	84	77	76	69	68	61	60	53	52

48	41	40	33	32	25	24	17	16	09	08	01
47	42	39	34	31	26	23	18	15	10	07	02
46	43	38	35	30	27	22	19	14	11	06	03
45	44	37	36	29	28	21	20	13	12	05	04

W1 Super-Bit Locality-Sensitive Hashing

Jianqiu Ji jijq10@mails.tsinghua.edu.cn
Jianmin Li lijianmin@mail.tsinghua.edu.cn
Tsinghua University
shuicheng yan scyan@math.pku.edu.cn
National University of Singapore
Bo Zhang bozhang@fairisaac.com
Fair Isaac Corp.
Qi Tian qitian@cs.utsa.edu
University of Texas at San Antonio

Sign-random-projection locality-sensitive hashing (SRP-LSH) is a probabilistic dimension reduction method which provides an unbiased estimate of angular similarity, yet suffers from the large variance of its estimation. In this work, we propose the Super-Bit locality-sensitive hashing (SBLSH). It is easy to implement, which orthogonalizes the random projection vectors in batches, and it is theoretically guaranteed that SBLSH also provides an unbiased estimate of angular similarity, yet with a smaller variance when the angle to estimate is within (0,π/2]. The extensive experiments on real data well validate that given the same length of binary code, SBLSH may achieve significant mean squared error reduction in estimating pairwise angular similarity. Moreover, SBLSH shows the superiority over SRP-LSH in approximate nearest neighbor (ANN) retrieval experiments.

W2 Learning Invariant Representations of Molecules for Atomization Energy Prediction

Grégoire Montavon g.montavon@gmail.com
Siamac Fazli fazli@tu-berlin.de
Franziska Biegler franziska.biegler@tu-berlin.de
Klaus-Robert Müller Klaus-Robert.Mueller@tu-berlin.de
TU Berlin
Katja Hansen hansen@fhi-berlin.mpg.de
Fritz-Haber-Institut
Matthias Rupp matthias.rupp@pharma.ethz.ch
ETH Zurich
Andreas Ziehe ziehe@first.fhg.de
Fraunhofer FIRST
Alexandre Tkatchenko atkatchenko@gmail.com
MPG Fritz Haber Institute
Anatole von Lilienfeld anatole@alcf.anl.gov
Argonne National Laboratory

The accurate prediction of molecular energetics in chemical compound space is a crucial ingredient for rational compound design. The inherently graph-like, non-vectorial nature of molecular data gives rise to a unique and difficult machine learning problem. In this paper, we adopt a learning-from-scratch approach where quantum-mechanical molecular energies are predicted directly from the raw molecular geometry. The study suggests a benefit from setting flexible priors and enforcing invariance stochastically rather than structurally. Our results improve the state-of-the-art by a factor of almost three, bringing statistical methods one step closer to the holy grail of ''chemical accuracy''.

W3 Adaptive Learning of Smoothing Functions: Application to Electricity Load Forecasting

Amadou Ba amadouba@ie.ibm.com
Pascal Pompey PAPOMPEY@ie.ibm.com
Mathieu Sinn mathsinn@ie.ibm.com
IBM Research - Ireland
Yannig Goude yannig.goude@edf.fr
EDF

This paper proposes an efficient online learning algorithm to track the smoothing functions of Additive Models. The key idea is to combine the linear representation of Additive Models with a Recursive Least Squares (RLS) filter. In order to quickly track changes in the model and put more weight on recent data, the RLS filter uses a forgetting factor which exponentially weights down observations by the order of their arrival. The tracking behaviour is further enhanced by using an adaptive forgetting factor which is updated based on the gradient of the a priori errors. Using results from Lyapunov stability theory, upper bounds for the learning rate are analyzed. The proposed algorithm is applied to 5 years of electricity load data provided by the French utility company Electricite de France (EDF). Compared to state-of-the-art methods, it achieves a superior performance in terms of model tracking and prediction accuracy.

W4 Automatic Feature Induction for Stagewise Collaborative Filtering

Joonseok Lee joonseok2010@gmail.com
Mingxuan Sun msun3@gatech.edu
Seungyeon Kim seungyeon.kim@gatech.edu
Guy Lebanon lebanon@cc.gatech.edu
Georgia Institute of Technology

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

W5 Co-Regularized Hashing for Multimodal Data

Linus Y. Zhen linuszhen@gmail.com
Dit-Yan Yeung dyyeung@cse.ust.hk
Hong Kong University of Science and Technology

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

W6 Label Ranking with Partial Abstention based on Thresholded Probabilistic Models

Weiwei Cheng cheng@mathematik.uni-marburg.de
Volkmar Welker welker@mathematik.uni-marburg.de
Eyke Huellermeier eyke@Mathematik.Uni-Marburg.de
Universitaet Marburg
Willem Waegeman willem.waegeman@ugent.be
Ghent University

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

W7 Random Utility Theory for Social Choice: Theory and Algorithms

Hossein Azari azari@fas.harvard.edu
David Parkes parkes@seas.harvard.edu
Lirong Xia xialirong@gmail.com
Harvard University

Random utility theory models an agent's preferences on alternatives by drawing a real-valued score on each alternative (typically independently) from a parameterized distribution, and then ranking according to scores. A special case that has received significant attention is the Plackett-Luce model, for which fast inference methods for maximum likelihood estimators are available. This paper develops conditions on general, random utility models that enable fast inference within a Bayesian framework through MC-EM, providing unimodal log-likelihood functions. Results on both real-world and simulated data provide support for the scalability of the approach, despite its flexibility.

W8 Unsupervised Structure Discovery for Semantic Analysis of Audio

Sourish Chaudhuri sourishc@cs.cmu.edu
Bhiksha Raj bhiksha@cs.cmu.edu
Carnegie Mellon University

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

W9 Hierarchical spike coding of sound

yan karklin yan.karklin@nyu.edu
Eero Simoncelli eero.simoncelli@nyu.edu
New York University
chaitu Ekanadham chaitu@math.nyu.edu
Courant Institute, New York University

We develop a probabilistic generative model for representing acoustic event structure at multiple scales via a two-stage hierarchy. The first stage consists of a spiking representation which encodes a sound with a sparse set of kernels at different frequencies positioned precisely in time. The coarse time and frequency statistical structure of the first-stage spikes is encoded by a second stage spiking representation, while fine-scale statistical regularities are encoded by recurrent interactions within the first-stage. When fitted to speech data, the model encodes acoustic features such as harmonic stacks, sweeps, and frequency modulations, that can be composed to represent complex acoustic events. The model is also able to synthesize sounds from the higher-level representation and provides significant improvement over wavelet thresholding techniques on a denoising task.

W10 Nonparametric Bayesian Inverse Reinforcement Learning for Multiple Reward Functions

Jaedeug Choi jdchoi@ai.kaist.ac.kr
Kee-Eung Kim kekim@cs.kaist.ac.kr
KAIST

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

W11 Near Optimal Chernoff Bounds for Markov Decision Processes

Teodor Mihai Moldovan moldovan@cs.berkeley.edu
Pieter Abbeel pabbeel@cs.berkeley.edu
Berkrley

The expected return is a widely used objective in decision making under uncertainty. Many algorithms, such as value iteration, have been proposed to optimize it. In risk-aware settings, however, the expected return is often not an appropriate objective to optimize. We propose a new optimization objective for risk-aware planning and show that it has desirable theoretical properties. We also draw connections to previously proposed objectives for risk-aware planing: minmax, exponential utility, percentile and mean minus variance. Our method applies to an extended class of Markov decision processes: we allow costs to be stochastic as long as they are bounded. Additionally, we present an efficient algorithm for optimizing the proposed objective. Synthetic and real-world experiments illustrate the effectiveness of our method, at scale.

W12 Sketch-Based Linear Value Function Approximation

Marc Bellemare mg17@ualberta.ca
Joel Veness jveness@gmail.com
Michael Bowling bowling@cs.ualberta.ca
University of Alberta

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

W13 A Unifying Perspective of Parametric Policy Search Methods for Markov Decision Processes

Thomas Furmston T.Furmston@cs.ucl.ac.uk
David Barber davidobarber@gmail.com
University College London

Parametric policy search algorithms are one of the methods of choice for the optimisation of Markov Decision Processes, with Expectation Maximisation and natural gradient ascent being considered the current state of the art in the field. In this article we provide a unifying perspective of these two algorithms by showing that their step-directions in the parameter space are closely related to the search direction of an approximate Newton method. This analysis leads naturally to the consideration of this approximate Newton method as an alternative gradient-based method for Markov Decision Processes. We are able show that the algorithm has numerous desirable properties, absent in the naive application of Newton's method, that make it a viable alternative to either Expectation Maximisation or natural gradient ascent. Empirical results suggest that the algorithm has excellent convergence and robustness properties, performing strongly in comparison to both Expectation Maximisation and natural gradient ascent.

W14 Large Scale Distributed Deep Networks

Jeff Dean jeff@google.com
Greg Corrado gcorrado@google.com
Rajat Monga rajatmonga@google.com
Andrew Senior andrewsenior@google.com
Kai Chen kaichen@google.com
Ke Yang yangke@google.com
Matthieu Devin mdevin@google.com
Mark Mao markmao@google.com
Marc'Aurelio Ranzato ranzato@google.com
Paul Tucker tucker@google.com
Google Inc.
Quoc Le quoc.le@stanford.edu
Andrew Ng ang@cs.stanford.edu
Stanford University

Recent work in unsupervised feature learning and deep learning has shown that being able to train large models can dramatically improve performance. In this paper, we consider the problem of training a deep network with billions of parameters using tens of thousands of CPU cores. We have developed a software framework called DistBelief that can utilize computing clusters with thousands of machines to train large models. Within this framework, we have developed two algorithms for large-scale distributed training: (i) Downpour SGD, an asynchronous stochastic gradient descent procedure supporting a large number of model replicas, and (ii) Sandblaster, a framework that supports for a variety of distributed batch optimization procedures, including a distributed implementation of L-BFGS. Downpour SGD and Sandblaster L-BFGS both increase the scale and speed of deep network training. We have successfully used our system to train a deep network 100x larger than previously reported in the literature, and achieves state-of-the-art performance on ImageNet, a visual object recognition task with 16 million images and 21k categories. We show that these same techniques dramatically accelerate the training of a more modestly sized deep network for a commercial speech recognition service. Although we focus on and report performance of these methods as applied to training large neural networks, the underlying algorithms are applicable to any gradient-based machine learning algorithm.

W15 Discriminative Learning of Sum-Product Networks

Robert Gens rcg@cs.washington.edu
Pedro Domingos pedrod@cs.washington.edu
University of Washington

Sum-product networks are a new deep architecture that can perform fast, exact inference on high-treewidth models. Only generative methods for training SPNs have been proposed to date. In this paper, we present the first discriminative training algorithms for SPNs, combining the high accuracy of the former with the representational power and tractability of the latter. We show that the class of tractable discriminative SPNs is broader than the class of tractable generative ones, and propose an efficient backpropagation-style algorithm for computing the gradient of the conditional log likelihood. Standard gradient descent suffers from the diffusion problem, but networks with many layers can be learned reliably using ''hard'' gradient descent, where marginal inference is replaced by MPE inference (i.e., inferring the most probable state of the non-evidence variables). The resulting updates have a simple and intuitive form. We test discriminative SPNs on standard image classification tasks. We obtain the best results to date on the CIFAR-10 dataset, using fewer features than prior methods with an SPN architecture that learns local image structure discriminatively. We also report the highest published test accuracy on STL-10 even though we only use the labeled portion of the dataset.

W16 A Neural Autoregressive Topic Model

Hugo Larochelle hugo.larochelle@usherbrooke.ca
Stanislas Lauly Stanislas.Lauly@USherbrooke.ca
Université de Sherbrooke

We describe a new model for learning meaningful representations of text documents from an unlabeled collection of documents. This model is inspired by the recently proposed Replicated Softmax, an undirected graphical model of word counts that was shown to learn a better generative model and more meaningful document representations. Specifically, we take inspiration from the conditional mean-field recursive equations of the Replicated Softmax in order to define a neural network architecture that estimates the probability of observing a new word in a given document given the previously observed words. This paradigm also allows us to replace the expensive softmax distribution over words with a hierarchical distribution over paths in a binary tree of words. The end result is a model whose training complexity scales logarithmically with the vocabulary size instead of linearly as in the Replicated Softmax. Our experiments show that our model is competitive both as a generative model of documents and as a document representation learning algorithm.

W17 Feature-aware Label Space Dimension Reduction for Multi-label Classification

Yao-Nan Chen r99922008@csie.ntu.edu.tw
Hsuan-Tien Lin htlin@csie.ntu.edu.tw
National Taiwan University

Label space dimension reduction (LSDR) is an efficient and effective paradigm for multi-label classification with many classes. Existing approaches to LSDR, such as compressive sensing and principal label space transformation, exploit only the label part of the dataset, but not the feature part. In this paper, we propose a novel approach to LSDR that considers both the label and the feature parts. The approach, called conditional principal label space transformation, is based on minimizing an upper bound of the popular Hamming loss. The minimization step of the approach can be carried out efficiently by a simple use of singular value decomposition. In addition, the approach can be extended to a kernelized version that allows the use of sophisticated feature combinations to assist LSDR. The experimental results verify that the proposed approach is more effective than existing ones to LSDR across many real-world datasets.

W18 Multi-task Vector Field Learning

Binbin Lin binbinlin@zju.edu.cn
Chiyuan Zhang pluskid@gmail.com
Xiaofei He xiaofeihe@cad.zju.edu.cn
Zhejiang University
Sen Yang senyang@asu.edu
Jieping Ye jieping.ye@asu.edu
Arizona State University

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

W19 Minimax Multi-Task Learning and a Generalized Loss-Compositional Paradigm for MTL

Nishant Mehta niche@cc.gatech.edu
Alexander Gray agray@cc.gatech.edu
Georgia Institute of Technology
Dongryeol Lee drselee@gmail.com
GE Global Research

Since its inception, the modus operandi of multi-task learning (MTL) has been to minimize the task-wise mean of the empirical risks. We introduce a generalized loss-compositional paradigm for MTL that includes a spectrum of formulations as a subfamily. One endpoint of this spectrum is minimax MTL: a new MTL formulation that minimizes the maximum of the tasks' empirical risks. Via a certain relaxation of minimax MTL, we obtain a continuum of MTL formulations spanning minimax MTL and classical MTL. The full paradigm itself is loss-compositional, operating on the vector of empirical risks. It incorporates minimax MTL, its relaxations, and many new MTL formulations as special cases. We show theoretically that minimax MTL tends to avoid worst case outcomes on newly drawn test tasks in the learning to learn (LTL) test setting. The results of several MTL formulations on synthetic and real problems in the MTL and LTL test settings are encouraging.

W20 Communication-Efficient Algorithms for Statistical Optimization

Yuchen Zhang yuczhang@eecs.berkeley.edu
Martin Wainwright wainwrig@eecs.berkeley.edu
John Duchi jduchi@cs.berkeley.edu
University of California Berkeley

We study two communication-efficient algorithms for distributed statistical optimization on large-scale data. The first algorithm is an averaging method that distributes the N data samples evenly to m machines, performs separate minimization on each subset, and then averages the estimates. We provide a sharp analysis of this average mixture algorithm, showing that under a reasonable set of conditions, the combined parameter achieves mean-squared error that decays as \order(N−1+(N/m)−2). Whenever m≤N, this guarantee matches the best possible rate achievable by a centralized algorithm having access to all N samples. The second algorithm is a novel method, based on an appropriate form of the bootstrap. Requiring only a single round of communication, it has mean-squared error that decays as \order(N−1+(N/m)−3), and so is more robust to the amount of parallelization. We complement our theoretical results with experiments on large-scale problems from the Microsoft Learning to Rank dataset.

W21 Multilabel Classification using Bayesian Compressed Sensing

Ashish Kapoor akapoor@microsoft.com
Raajay Viswanathan t-rviswa@microsoft.com
Prateek Jain prajain@microsoft.com
Microsoft Research Lab

In this paper, we present a Bayesian framework for multilabel classification using compressed sensing. The key idea in compressed sensing for multilabel classification is to first project the label vector to a lower dimensional space using a random transformation and then learn regression functions over these projections. Our approach considers both of these components in a single probabilistic model, thereby jointly optimizing over compression as well as learning tasks. We then derive an efficient variational inference scheme that provides joint posterior distribution over all the unobserved labels. The two key benefits of the model are that a) it can naturally handle datasets that have missing labels and b) it can also measure uncertainty in prediction. The uncertainty estimate provided by the model naturally allows for active learning paradigms where an oracle provides information about labels that promise to be maximally informative for the prediction task. Our experiments show significant boost over prior methods in terms of prediction performance over benchmark datasets, both in the fully labeled and the missing labels case. Finally, we also highlight various useful active learning scenarios that are enabled by the probabilistic model.

W22 MCMC for continuous-time discrete-state systems

Vinayak Rao vrao@gatsby.ucl.ac.uk
Gatsby Unit, UCL
Yee Whye Teh teh@stats.ox.ac.uk
University of Oxford

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

W23 Ancestor Sampling for Particle Gibbs

Fredrik Lindsten lindsten@isy.liu.se
Thomas Schön schon@isy.liu.se
Linköping University
Michael Jordan jordan@cs.berkeley.edu
University of California

We present a novel method in the family of particle MCMC methods that we refer to as particle Gibbs with ancestor sampling (PG-AS). Similarly to the existing PG with backward simulation (PG-BS) procedure, we use backward sampling to (considerably) improve the mixing of the PG kernel. Instead of using separate forward and backward sweeps as in PG-BS, however, we achieve the same effect in a single forward sweep. We apply the PG-AS framework to the challenging class of non-Markovian state-space models. We develop a truncation strategy of these models that is applicable in principle to any backward-simulation-based method, but which is particularly well suited to the PG-AS framework. In particular, as we show in a simulation study, PG-AS can yield an order-of-magnitude improved accuracy relative to PG-BS due to its robustness to the truncation error. Several application examples are discussed, including Rao-Blackwellized particle smoothing and inference in degenerate state-space models.

W24 Probabilistic Event Cascades for Alzheimer's disease

Jonathan Huang jhuang11@stanford.edu
Stanford University
Daniel Alexander D.Alexander@cs.ucl.ac.uk
UCL

Accurate and detailed models of the progression of neurodegenerative diseases such as Alzheimer's (AD) are crucially important for reliable early diagnosis and the determination and deployment of effective treatments. In this paper, we introduce the ALPACA (Alzheimer's disease Probabilistic Cascades) model, a generative model linking latent Alzheimer's progression dynamics to observable biomarker data. In contrast with previous works which model disease progression as a fixed ordering of events, we explicitly model the variability over such orderings among patients which is more realistic, particularly for highly detailed disease progression models. We describe efficient learning algorithms for ALPACA and discuss promising experimental results on a real cohort of Alzheimer's patients from the Alzheimer's Disease Neuroimaging Initiative.

W25 Bayesian Pedigree Analysis using Measure Factorization

Bonnie Kirkpatrick bbkirk@cs.ubc.ca
Alexandre Bouchard-Côté bouchard@stat.ubc.ca
University of British Columbia

Pedigrees, or family trees, are directed graphs used to identify sites of the genome that are correlated with the presence or absence of a disease. With the advent of genotyping and sequencing technologies, there has been an explosion in the amount of data available, both in the number of individuals and in the number of sites. Some pedigrees number in the thousands of individuals. Meanwhile, analysis methods have remained limited to pedigrees of <100 individuals which limits analyses to many small independent pedigrees. Disease models, such those used for the linkage analysis log-odds (LOD) estimator, have similarly been limited. This is because linkage anlysis was originally designed with a different task in mind, that of ordering the sites in the genome, before there were technologies that could reveal the order. LODs are difficult to interpret and nontrivial to extend to consider interactions among sites. These developments and difficulties call for the creation of modern methods of pedigree analysis. Drawing from recent advances in graphical model inference and transducer theory, we introduce a simple yet powerful formalism for expressing genetic disease models. We show that these disease models can be turned into accurate and efficient estimators. The technique we use for constructing the variational approximation has potential applications to inference in other large-scale graphical models. This method allows inference on larger pedigrees than previously analyzed in the literature, which improves disease site prediction.

W26 Fiedler Random Fields: A Large-Scale Spectral Approach to Statistical Network Modeling

Antonino Freno antonino.freno@inria.fr
Marc Tommasi marc.tommasi@inria.fr
INRIA
Mikaela Keller mikaela.keller@univ-lille3.fr
Université Lille 3

Statistical models for networks have been typically committed to strong prior assumptions concerning the form of the modeled distributions. Moreover, the vast majority of currently available models are explicitly designed for capturing some specific graph properties (such as power-law degree distributions), which makes them unsuitable for application to domains where the behavior of the target quantities is not known a priori. The key contribution of this paper is twofold. First, we introduce the Fiedler delta statistic, based on the Laplacian spectrum of graphs, which allows to dispense with any parametric assumption concerning the modeled network properties. Second, we use the defined statistic to develop the Fiedler random field model, which allows for efficient estimation of edge distributions over large-scale random networks. After analyzing the dependence structure involved in Fiedler random fields, we estimate them over several real-world networks, showing that they achieve a much higher modeling accuracy than other well-known statistical approaches.

W27 Distributed Probabilistic Learning for Camera Networks with Missing Data

Sejong Yoon sjyoon@cs.rutgers.edu
Vladimir Pavlovic vladimir@cs.rutgers.edu
Rutgers University

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

W28 No voodoo here! Learning discrete graphical models via inverse covariance estimation

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

We investigate the relationship between the support of the inverses of generalized covariance matrices and the structure of a discrete graphical model. We show that for certain graph structures, the support of the inverse covariance matrix of indicator variables on the vertices of a graph reflects the conditional independence structure of the graph. Our work extends results which were previously established only for multivariate Gaussian distributions, and partially answers an open question about the meaning of the inverse covariance matrix of a non-Gaussian distribution. We propose graph selection methods for a general discrete graphical model with bounded degree based on possibly corrupted observations, and verify our theoretical results via simulations. Along the way, we also establish new results for support recovery in the setting of sparse high-dimensional linear regression based on corrupted and missing observations.

W29 Scalable Inference of Overlapping Communities

Prem Gopalan pgopalan@cs.princeton.edu
David Mimno mimno@cs.princeton.edu
Sean Gerrish sean.gerrish@gmail.com
Michael Freedman mfreed@cs.princeton.edu
David Blei blei@cs.princeton.edu
Princeton University

We develop a scalable algorithm for posterior inference of overlapping communities in large networks. Our algorithm is based on stochastic variational inference in the mixed-membership stochastic blockmodel. It naturally interleaves subsampling the network with estimating its community structure. We apply our algorithm on ten large, real-world networks with up to 60,000 nodes. It converges several orders of magnitude faster than the state-of-the-art algorithm for MMSB, finds hundreds of communities in large real-world networks, and detects the true communities in 280 benchmark networks with equal or better accuracy compared to other scalable algorithms.

W30 Probabilistic Topic Coding for Superset Label Learning

Liping Liu liping.liulp@gmail.com
Tom Dietterich tgd@oregonstate.edu
Oregon State University

In the superset label learning problem, each training instance provides a set of candidate labels of which one is the true label of the instance. Most approaches learn a discriminative classifier that tries to minimize an upper bound of the unobserved 0/1 loss. In this work, we propose a probabilistic model, Probabilistic Topic Coding (PTC), for the superset label learning problem. The PTC model is derived from logistic stick breaking process. It first maps the data to ``topics'', and then assigns to each topic a label drawn from a multinomial distribution. The layer of topics can capture underlying structure in the data, which is very useful when the model is weakly supervised. This advantage comes at little cost, since the model introduces few additional parameters. Experimental tests on several real-world problems with superset labels show results that are competitive or superior to the state of the art. The discovered underlying structures also provide improved explanations of the classification predictions.

W31 Augment-and-Conquer Negative Binomial Processes

Mingyuan Zhou mz1@ee.duke.edu
Lawrence Carin lcarin@ee.duke.edu
Duke University

By developing data augmentation methods unique to the negative binomial (NB) distribution, we unite seemingly disjoint count and mixture models under the NB process framework. We develop fundamental properties of the models and derive efficient Gibbs sampling inference. We show that the gamma-NB process can be reduced to the hierarchical Dirichlet process with normalization, highlighting its unique theoretical, structural and computational advantages. A variety of NB processes with distinct sharing mechanisms are constructed and applied to topic modeling, with connections to existing algorithms, showing the importance of inferring both the NB dispersion and probability parameters.

W32 Priors for Diversity in Generative Latent Variable Models

James Zou jzou@fas.harvard.edu
Ryan Adams rpa@seas.harvard.edu
Harvard University

Probabilistic latent variable models are one of the cornerstones of machine learning. They offer a convenient and coherent way to specify prior distributions over unobserved structure in data, so that these unknown properties can be inferred via posterior inference. Such models are useful for exploratory analysis and visualization, for building density models of data, and for providing features that can be used for later discriminative tasks. A significant limitation of these models, however, is that draws from the prior are often highly redundant due to i.i.d. assumptions on internal parameters. For example, there is no preference in the prior of a mixture model to make components non-overlapping, or in topic model to ensure that co-ocurring words only appear in a small number of topics. In this work, we revisit these independence assumptions for probabilistic latent variable models, replacing the underlying i.i.d.\ prior with a determinantal point process (DPP). The DPP allows us to specify a preference for diversity in our latent variables using a positive definite kernel function. Using a kernel between probability distributions, we are able to define a DPP on probability measures. We show how to perform MAP inference with DPP priors in latent Dirichlet allocation and in mixture models, leading to better intuition for the latent variable representation and quantitatively improved unsupervised feature extraction, without compromising the generative aspects of the model.

W33 Dynamic Pruning of Factor Graphs for Maximum Marginal Prediction

Christoph Lampert chl@ist.ac.at
IST Austria

We study the problem of maximum marginal prediction (MMP) in probabilistic graphical models, a task that occurs, for example, as the Bayes optimal decision rule under a Hamming loss. MMP is typically performed as a two-stage procedure: one estimates each variable's marginal probability and then forms a prediction from the states of maximal probability. In this work we propose a simple yet effective technique for accelerating MMP when inference is sampling-based: instead of the above two-stage procedure we directly estimate the posterior probability of each decision variable. This allows us to identify the point of time when we are sufficiently certain about any individual decision. Whenever this is the case, we dynamically prune the variable we are confident about from the underlying factor graph. Consequently, at any time only samples of variable whose decision is still uncertain need to be created. Experiments in two prototypical scenarios, multi-label classification and image inpainting, shows that adaptive sampling can drastically accelerate MMP without sacrificing prediction accuracy.

W34 Variational Inference for Crowdsourcing

Qiang Liu qliu1@uci.edu
Alexander Ihler ihler@ics.uci.edu
University of California, Irvine
Jian Peng jpengwhu@gmail.com
TTI Chicago

Crowdsourcing has become a popular paradigm for labeling large datasets. However, it has given rise to the computational task of aggregating the crowdsourced labels provided by a collection of unreliable annotators. We approach this problem by transforming it into a standard inference problem in graphical models, and applying approximate variational methods, including belief propagation (BP) and mean field (MF). We show that our BP algorithm generalizes both majority voting and a recent algorithm by Karger et al, while our MF method is closely related to a commonly used EM algorithm. In both cases, we find that the performance of the algorithms critically depends on the choice of a prior distribution on the workers' reliability; by choosing the prior properly, both BP and MF (and EM) perform surprisingly well on both simulated and real-world datasets, competitive with state-of-the-art algorithms based on more complicated modeling assumptions.

W35 Near-Optimal MAP Inference for Determinantal Point Processes

Alex Kulesza kulesza@umich.edu
University of Michigan
Jennifer Gillenwater jengi@cis.upenn.edu
Ben Taskar taskar@cis.upenn.edu
University of Pennsylvania

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

W36 Affine Independent Variational Inference

Edward Challis edward.challis.09@ucl.ac.uk
David Barber davidobarber@gmail.com
University College London

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

W37 Truly Nonparametric Online Variational Inference for Hierarchical Dirichlet Processes

Michael Bryant mbryantj@gmail.com
Erik Sudderth sudderth@cs.brown.edu
Brown University

Variational methods provide a computationally scalable alternative to Monte Carlo methods for large-scale, Bayesian nonparametric learning. In practice, however, conventional batch and online variational methods quickly become trapped in local optima. In this paper, we consider a nonparametric topic model based on the hierarchical Dirichlet process (HDP), and develop a novel online variational inference algorithm based on split-merge topic updates. We derive a simpler and faster variational approximation of the HDP, and show that by intelligently splitting and merging components of the variational posterior, we can achieve substantially better predictions of test data than conventional online and batch variational algorithms. For streaming analysis of large datasets where batch analysis is infeasible, we show that our split-merge updates better capture the nonparametric properties of the underlying model, allowing continual learning of new topics.

W38 Effective Split-Merge Monte Carlo Methods for Nonparametric Models of Sequential Data

Mike Hughes mhughes@cs.brown.edu
Erik Sudderth sudderth@cs.brown.edu
Brown University
Emily Fox ebfox@uw.edu
University of Washington

Applications of Bayesian nonparametric methods require learning and inference algorithms which efficiently explore models of unbounded complexity. We develop new Markov chain Monte Carlo methods for the beta process hidden Markov model (BP-HMM), enabling discovery of shared activity patterns in large video and motion capture databases. By introducing split-merge moves based on sequential allocation, we allow large global changes in the shared feature structure. We also develop data-driven reversible jump moves which more reliably discover rare or unique behaviors. Our proposals apply to any choice of conjugate likelihood for observed data, and we show success with multinomial, Gaussian, and autoregressive emission models. Together, these innovations allow tractable analysis of hundreds of time series, where previous inference required clever initialization and at least ten thousand burn-in iterations for just six sequences.

W39 Truncation-free Online Variational Inference for Bayesian Nonparametric Models

Chong Wang chongw@cs.princeton.edu
Carnegie Mellon University
David Blei blei@cs.princeton.edu
Princeton University

We present a truncation-free online variational inference algorithm for Bayesian nonparametric models. Unlike traditional (online) variational inference algorithms that require truncations for the model or the variational distribution, our method adapts model complexity on the fly. Our experiments for Dirichlet process mixture models and hierarchical Dirichlet process topic models on two large-scale data sets show better performance than previous online variational inference algorithms.

W40 Modelling Reciprocating Relationships

Charles Blundell c.blundell@gatsby.ucl.ac.uk
University College of London
Katherine Heller kheller@gmail.com
Duke University
Jeff Beck jeffbeck@gatsby.ucl.ac.uk
University of Rochester

We present a Bayesian nonparametric model that discovers implicit social structure from interaction time-series data. Social groups are often formed implicitly, through actions among members of groups. Yet many models of social networks use explicitly declared relationships to infer social structure. We consider a particular class of Hawkes processes, a doubly stochastic point process, that is able to model reciprocity between groups of individuals. We then extend the Infinite Relational Model by using these reciprocating Hawkes processes to parameterise its edges, making events associated with edges co-dependent through time. Our model outperforms general, unstructured Hawkes processes as well as structured Poisson process-based models at predicting verbal and email turn-taking, and military conflicts among nations.

W41 A nonparametric variable clustering model

David Knowles knowles84@gmail.com
Stanford University
Konstantina Palla kp376@cam.ac.uk
Zoubin Ghahramani zoubin@eng.cam.ac.uk
University of Cambridge

Factor analysis models effectively summarise the covariance structure of high dimensional data, but the solutions are typically hard to interpret. This motivates attempting to find a disjoint partition, i.e. a clustering, of observed variables so that variables in a cluster are highly correlated. We introduce a Bayesian non-parametric approach to this problem, and demonstrate advantages over heuristic methods proposed to date.

W42 Bayesian nonparametric models for ranked data

Francois Caron Francois.Caron@inria.fr
INRIA Bordeaux
Yee Whye Teh teh@stats.ox.ac.uk
University of Oxford

We develop a Bayesian nonparametric extension of the popular Plackett-Luce choice model that can handle an infinite number of choice items. Our framework is based on the theory of random atomic measures, with the prior specified by a gamma process. We derive a posterior characterization and a simple and effective Gibbs sampler for posterior simulation. We then develop a time-varying extension of our model, and apply our model to the New York Times lists of weekly bestselling books.

W43 A Marginalized Particle Gaussian Process Regression

Yali Wang wang@damas.ift.ulaval.ca
Brahim Chaib-draa chaib@ift.ulaval.ca
Laval University

We present a novel marginalized particle Gaussian process (MPGP) regression, which provides a fast, accurate online Bayesian filtering framework to model the latent function. Using a state space model established by the data construction procedure, our MPGP recursively filters out the estimation of hidden function values by a Gaussian mixture. Meanwhile, it provides a new online method for training hyperparameters with a number of weighted particles. We demonstrate the estimated performance of our MPGP on both simulated and real large data sets. The results show that our MPGP is a robust estimation algorithm with high computational efficiency, which outperforms other state-of-art sparse GP methods.

W44 Structured Sparse Learning of Multiple Gaussian Graphical Models

Karthik Mohan karna@uw.edu
Michael Chung mjyc@cs.washington.edu
Seungyeop Han syhan@cs.washington.edu
Daniela Witten dwitten@u.washington.edu
Maryam Fazel mfazel@uw.edu
University of Washington
Su-In Lee silee@cs.stanford.edu
Stanford University

We consider estimation of multiple high-dimensional Gaussian graphical models corresponding to a single set of nodes under several distinct conditions. We assume that most aspects of the networks are shared, but that there are some structured differences between them. Specifically, the network differences are generated from node perturbations: a few nodes are perturbed across networks, and most or all edges stemming from such nodes differ between networks. This corresponds to a simple model for the mechanism underlying many cancers, in which the gene regulatory network is disrupted due to the aberrant activity of a few specific genes. We propose to solve this problem using the structured joint graphical lasso, a convex optimization problem that is based upon the use of a novel symmetric overlap norm penalty, which we solve using an alternating directions method of multipliers algorithm. Our proposal is illustrated on synthetic data and on an application to brain cancer gene expression data.

W45 Topology Constraints in Graphical Models

Marcelo Fiori mfiori@fing.edu.uy
Facultad de Ingeniería, UdelaR
Pablo Musé pmuse@fing.edu.uy
Universidad de la República
Guillermo Sapiro guillermo.sapiro@duke.edu
Duke University

Graphical models are a very useful tool to describe and understand natural phenomena, from gene expression to climate change and social interactions. The topological structure of these graphs/networks is a fundamental part of the analysis, and in many cases the main goal of the study. However, little work has been done on incorporating prior topological knowledge onto the estimation of the underlying graphical models from sample data. In this work we propose extensions to the basic joint regression model for network estimation, which explicitly incorporate graph-topological constraints into the corresponding optimization approach. The first proposed extension includes an eigenvector centrality constraint, thereby promoting this important prior topological property. The second developed extension promotes the formation of certain motifs, triangle-shaped ones in particular, which are known to exist for example in genetic regulatory networks. The presentation of the underlying formulations, which serve as examples of the introduction of topological constraints in network estimation, is complemented with examples in diverse datasets demonstrating the importance of incorporating such critical prior knowledge.

W46 Learning Mixtures of Tree Graphical Models

Anima Anandkumar a.anandkumar@uci.edu
Furong Huang furongh@uci.edu
U.C.Irvine
Daniel Hsu danielhsu@gmail.com
Sham Kakade skakade@microsoft.com
Microsoft Research

We consider unsupervised estimation of mixtures of discrete graphical models, where the class variable is hidden and each mixture component can have a potentially different Markov graph structure and parameters over the observed variables. We propose a novel method for estimating the mixture components with provable guarantees. Our output is a tree-mixture model which serves as a good approximation to the underlying graphical model mixture. The sample and computational requirements for our method scale as \poly(p,r), for an r-component mixture of p-variate graphical models, for a wide class of models which includes tree mixtures and mixtures over bounded degree graphs.

W47 A Convex Formulation for Learning Scale-Free Networks via Submodular Relaxation

Aaron Defazio aaron.defazio@anu.edu.au
ANU
Tiberio Caetano tiberio.caetano@nicta.com.au
NICTA/ANU

A key problem in statistics and machine learning is the determination of network structure from data. We consider the case where the structure of the graph to be reconstructed is known to be scale-free. We show that in such cases it is natural to formulate structured sparsity inducing priors using submodular functions, and we use their Lovasz extension to obtain a convex relaxation. For tractable classes such as Gaussian graphical models, this leads to a convex optimization problem that can be efficiently solved. We show that our method results in an improvement in the accuracy of reconstructed networks for synthetic data. We also show how our prior encourages scale-free reconstructions on a bioinfomatics dataset.

W48 Graphical Models via Generalized Linear Models

Eunho Yang eunho@cs.utexas.edu
Pradeep Ravikumar pradeepr@cs.utexas.edu
University of Texas, Austin
Genevera Allen giallen@stanford.edu
Rice University
zhandong Liu zhandong@mail.med.upenn.edu
University of Pennsylvania

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

W49 A latent factor model for highly multi-relational data

Rodolphe Jenatton rodolphe.jenatton@inria.fr
INRIA
Nicolas Le Roux nicolas@le-roux.name
Criteo
Antoine Bordes antoine.bordes@hds.utc.fr
CNRS / UTC
Guillaume Obozinski Guillaume.Obozinski@ens.fr
INRIA / ENS

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

W50 On Triangular versus Edge Representations --- Towards Scalable Modeling of Networks

Qirong Ho qho@cs.cmu.edu
Eric Xing epxing@cs.cmu.edu
Junming Yin junmingy@cs.cmu.edu
Carnegie Mellon University

In this paper, we argue for representing networks as a bag of {\it triangular motifs}, particularly for important network problems that current model-based approaches handle poorly due to computational bottlenecks incurred by using edge representations. Such approaches require both 1-edges and 0-edges (missing edges) to be provided as input, and as a consequence, approximate inference algorithms for these models usually require Ω(N2) time per iteration, precluding their application to larger real-world networks. In contrast, triangular modeling requires less computation, while providing equivalent or better inference quality. A triangular motif is a vertex triple containing 2 or 3 edges, and the number of such motifs is Θ(∑iDi2) (where Di is the degree of vertex i), which is much smaller than N2 for low-maximum-degree networks. Using this representation, we develop a novel mixed-membership network model and approximate inference algorithm suitable for large networks with low max-degree. For networks with high maximum degree, the triangular motifs can be naturally subsampled in a {\ it node-centric} fashion, allowing for much faster inference at a small cost in accuracy. Empirically, we demonstrate that our approach, when compared to that of an edge-based model, has faster runtime and improved accuracy for mixed-membership community detection. We conclude with a large-scale demonstration on an N≈280,000-node network, which is infeasible for network models with Ω(N2) inference cost.

W51 Efficient high dimensional maximum entropy modeling via symmetric partition functions

Paul Vernaza paul.vernaza@gmail.com
Drew Bagnell dbagnell+nips@ri.cmu.edu
Carnegie Mellon University

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

W52 Simultaneously Leveraging Output and Task Structures for Multiple-Output Regression

Piyush Rai piyush@cs.utah.edu
University of Texas at Austin
Abhishek Kumar abhishek@cs.umd.edu
Hal Daume III me@hal3.name
University of Maryland

Multiple-output regression models require estimating multiple functions, one for each output. To improve parameter estimation in such models, methods based on structural regularization of the model parameters are usually needed. In this paper, we present a multiple-output regression model that leverages the covariance structure of the functions (i.e., how the multiple functions are related with each other) as well as the conditional covariance structure of the outputs. This is in contrast with existing methods that usually take into account only one of these structures. More importantly, unlike most of the other existing methods, none of these structures need be known a priori in our model, and are learned from the data. Several previously proposed structural regularization based multiple-output regression models turn out to be special cases of our model. Moreover, in addition to being a rich model for multiple-output regression, our model can also be used in estimating the graphical model structure of a set of variables (multivariate outputs) conditioned on another set of variables (inputs). Experimental results on both synthetic and real datasets demonstrate the effectiveness of our method.

W53 Augmented-SVM: Automatic space partitioning for combining multiple non-linear dynamics

Ashwini Shukla ashwini.shukla@epfl.ch
Aude Billard karin.elsea@epfl.ch
Ecole Polytechnique Fédérale de Lausanne

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

W54 Local Supervised Learning through Space Partitioning

Joseph Wang joewang@bu.edu
Venkatesh Saligrama srv@bu.edu
Boston University

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

W55 Mandatory Leaf Node Prediction in Hierarchical Multilabel Classification

Wei Bi weibi@cse.ust.hk
James Kwok jamesk@cse.ust.hk
Hong Kong University of Science and Technology

In hierarchical classification, the prediction paths may be required to always end at leaf nodes. This is called mandatory leaf node prediction (MLNP) and is particularly useful when the leaf nodes have much stronger semantic meaning than the internal nodes. However, while there have been a lot of MLNP methods in hierarchical multiclass classification, performing MLNP in hierarchical multilabel classification is much more difficult. In this paper, we propose a novel MLNP algorithm that (i) considers the global hierarchy structure; and (ii) can be used on hierarchies of both trees and DAGs. We show that one can efficiently maximize the joint posterior probability of all the node labels by a simple greedy algorithm. Moreover, this can be further extended to the minimization of the expected symmetric loss. Experiments are performed on a number of real-world data sets with tree- and DAG-structured label hierarchies. The proposed method consistently outperforms other hierarchical and flat multilabel classification methods.

W56 Proper losses for learning from partial labels

Jesus Cid-Sueiro jcid@tsc.uc3m.es
Univ. Carlos III de Madrid

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

W57 Multiclass Learning with Simplex Coding

Youssef Mroueh ymroueh@mit.edu
Tomaso Poggio tp@ai.mit.edu
Jean-Jacques Slotine jjs@mit.edu
Lorenzo Rosasco lrosasco@mit.edu
Massachusetts Institute of Technology

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

W58 Multiclass Learning Approaches: A Theoretical Comparison with Implications

Amit Daniely amit.daniely@mail.huji.ac.il
Shai Shalev-Shwartz shai.shwartz@gmail.com
Hebrew university
Sivan Sabato sivan.sabato@gmail.com
Microsoft Research

We theoretically analyze and compare the following five popular multiclass classification methods: One vs. All, All Pairs, Tree-based classifiers, Error Correcting Output Codes (ECOC) with randomly generated code matrices, and Multiclass SVM. In the first four methods, the classification is based on a reduction to binary classification. We consider the case where the binary classifier comes from a class of VC dimension d, and in particular from the class of halfspaces over \realsd. We analyze both the estimation error and the approximation error of these methods. Our analysis reveals interesting conclusions of practical relevance, regarding the success of the different approaches under various conditions. Our proof technique employs tools from VC theory to analyze the \emph{approximation error} of hypothesis classes. This is in sharp contrast to most, if not all, previous uses of VC theory, which only deal with estimation error.

W59 Learning as MAP Inference in Discrete Graphical Models

Tiberio Caetano tiberio.caetano@nicta.com.au
James Petterson james.petterson@nicta.com.au
NICTA/ANU
Xianghang Liu xianghang.liu@nicta.com.au
University of New South Wales

We present a new formulation for attacking binary classification problems. Instead of relying on convex losses and regularisers such as in SVMs, logistic regression and boosting, or instead non-convex but continuous formulations such as those encountered in neural networks and deep belief networks, our framework entails a non-convex but \emph{discrete} formulation, where estimation amounts to finding a MAP configuration in a graphical model whose potential functions are low-dimensional discrete surrogates for the misclassification loss. We argue that such a discrete formulation can naturally account for a number of issues that are typically encountered in either the convex or the continuous non-convex paradigms, or both. By reducing the learning problem to a MAP inference problem, we can immediately translate the guarantees available for many inference settings to the learning problem itself. We empirically demonstrate in a number of experiments that this approach is promising in dealing with issues such as severe label noise, while still having global optimality guarantees. Due to the discrete nature of the formulation, it also allows for \emph{direct} regularisation through cardinality-based penalties, such as the $\ell 0$ pseudo-norm, thus providing the ability to perform feature selection and trade-off interpretability and predictability in a principled manner. We also outline a number of open problems arising from the formulation.

W60 A new metric on the manifold of kernel matrices with application to matrix geometric means

Suvrit Sra suvrit@gmail.com
Max Planck Institute for Intelligent Systems

Symmetric positive definite (spd) matrices are remarkably pervasive in a multitude of scientific disciplines, including machine learning and optimization. We consider the fundamental task of measuring distances between two spd matrices; a task that is often nontrivial whenever an application demands the distance function to respect the non-Euclidean geometry of spd matrices. Unfortunately, typical non-Euclidean distance measures such as the Riemannian metric \riem(X,Y)=\froblog(X\invY), are computationally demanding and also complicated to use. To allay some of these difficulties, we introduce a new metric on spd matrices: this metric not only respects non-Euclidean geometry, it also offers faster computation than \riem while being less complicated to use. We support our claims theoretically via a series of theorems that relate our metric to \riem(X,Y), and experimentally by studying the nonconvex problem of computing matrix geometric means based on squared distances.

W61 Learning Networks of Heterogeneous Influence

NAN DU dunan@gatech.edu
Georgia Institute of Technology
Le Song lsong@cc.gatech.edu
Georgia Tech

Alexander Smola alex@smola.org
Google
Ming Yuan myuan@isye.gatech.edu

Information, disease, and influence diffuse over networks of entities in both natural systems and human society. Analyzing these transmission networks plays an important role in understanding the diffusion processes and predicting events in the future. However, the underlying transmission networks are often hidden and incomplete, and we observe only the time stamps when cascades of events happen. In this paper, we attempt to address the challenging problem of uncovering the hidden network only from the cascades. The structure discovery problem is complicated by the fact that the influence among different entities in a network are heterogeneous, which can not be described by a simple parametric model. Therefore, we propose a kernel-based method which can capture a diverse range of different types of influence without any prior assumption. In both synthetic and real cascade data, we show that our model can better recover the underlying diffusion network and drastically improve the estimation of the influence functions between networked entities.

W62 Kernel Hyperalignment

Alexander Lorbert alorbert@princeton.edu
Peter Ramadge ramadge@princeton.edu
Princeton University

We offer a regularized, kernel extension of the multi-set, orthogonal Procrustes problem, or hyperalignment. Our new method, called Kernel Hyperalignment, expands the scope of hyperalignment to include nonlinear measures of similarity and enables the alignment of multiple datasets with a large number of base features. With direct application to fMRI data analysis, kernel hyperalignment is well-suited for multi-subject alignment of large ROIs, including the entire cortex. We conducted experiments using real-world, multi-subject fMRI data.

W63 Majorization for CRFs and Latent Likelihoods

Tony Jebara jebara@cs.columbia.edu
Anna Choromanska aec2163@columbia.edu
Columbia University

The partition function plays a key role in probabilistic modeling including conditional random fields, graphical models, and maximum likelihood estimation. To optimize partition functions, this article introduces a quadratic variational upper bound. This inequality facilitates majorization methods: optimization of complicated functions through the iterative solution of simpler sub-problems. Such bounds remain efficient to compute even when the partition function involves a graphical model (with small tree-width) or in latent likelihood settings. For large-scale problems, low-rank versions of the bound are provided and outperform LBFGS as well as first-order methods. Several learning applications are shown and reduce to fast and convergent update rules. Experimental results show advantages over state-of-the-art optimization methods.

W64 Privacy Aware Learning

John Duchi jduchi@cs.berkeley.edu
University of California Berkeley
Michael Jordan jordan@cs.berkeley.edu
University of California
Martin Wainwright wainwrig@eecs.berkeley.edu
UC Berkeley

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

W65 Statistical Consistency of Ranking Methods in A Rank-Differentiable Probability Space

Yanyan Lan lanyanyan@ict.ac.cn
Xueqi Cheng cxq@ict.ac.cn
Jiafeng Guo guojiafeng@ict.ac.cn
ICT
Tie-Yan Liu tyliu@microsoft.com
Microsoft

This paper is concerned with the statistical consistency of ranking methods. Recently, it was proven that many commonly used pairwise ranking methods are inconsistent with the weighted pairwise disagreement loss (WPDL), which can be viewed as the true loss of ranking, even in a low-noise setting. This result is interesting but also surprising, given that the pairwise ranking methods have been shown very effective in practice. In this paper, we argue that the aforementioned result might not be conclusive, depending on what kind of assumptions are used. We give a new assumption that the labels of objects to rank lie in a rank-differentiable probability space (RDPS), and prove that the pairwise ranking methods become consistent with WPDL under this assumption. What is especially inspiring is that RDPS is actually not stronger than but similar to the low-noise setting. Our studies provide theoretical justifications of some empirical findings on pairwise ranking methods that are unexplained before, which bridge the gap between theory and applications.

W66 A Spectral Algorithm for Latent Dirichlet Allocation

Anima Anandkumar a.anandkumar@uci.edu
U.C.Irvine
Dean Foster foster@wharton.upenn.edu
University of Pennsylvania
Daniel Hsu danielhsu@gmail.com
Sham Kakade skakade@microsoft.com
Microsoft Research
Yi-Kai Liu yikailiu00@gmail.com
National Institute of Standards and Technology

Topic modeling is a generalization of clustering that posits that observations (words in a document) are generated by \emph{multiple} latent factors (topics), as opposed to just one. This increased representational power comes at the cost of a more challenging unsupervised learning problem of estimating the topic-word distributions when only words are observed, and the topics are hidden. This work provides a simple and efficient learning procedure that is guaranteed to recover the parameters for a wide class of topic models, including Latent Dirichlet Allocation (LDA). For LDA, the procedure correctly recovers both the topic-word distributions and the parameters of the Dirichlet prior over the topic mixtures, using only trigram statistics (\ emph{i.e.}, third order moments, which may be estimated with documents containing just three words). The method, called Excess Correlation Analysis, is based on a spectral decomposition of low-order moments via two singular value decompositions (SVDs). Moreover, the algorithm is scalable, since the SVDs are carried out only on k×k matrices, where k is the number of latent factors (topics) and is typically much smaller than the dimension of the observation (word) space.

W67 Distributed Non-Stochastic Experts

Varun Kanade vkanade@eecs.berkeley.edu
University of California, Berkeley
Zhenming Liu zliu@fas.harvard.edu
Harvard University
Bozidar Radunovic bozidar@microsoft.com
Microsoft Research

We consider the online distributed non-stochastic experts problem, where the distributed system consists of one coordinator node that is connected to k sites, and the sites are required to communicate with each other via the coordinator. At each time-step t, one of the k site nodes has to pick an expert from the set {1, . . . , n}, and the same site receives information about payoffs of all experts for that round. The goal of the distributed system is to minimize regret at time horizon T, while simultaneously keeping communication to a minimum. The two extreme solutions to this problem are: (i) Full communication: This essentially simulates the non-distributed setting to obtain the optimal O(\sqrt{log(n)T}) regret bound at the cost of T communication. (ii) No communication: Each site runs an independent copy – the regret is O(\sqrt{log(n)kT}) and the communication is 0. This paper shows the difficulty of simultaneously achieving regret asymptotically better than \sqrt{kT} and communication better than T. We give a novel algorithm that for an oblivious adversary achieves a non-trivial trade-off: regret O(\sqrt{k^{5(1+\epsilon)/6} T}) and communication O(T/k^\epsilon), for any value of \ epsilon in (0, 1/5). We also consider a variant of the model, where the coordinator picks the expert. In this model, we show that the label-efficient forecaster of Cesa-Bianchi et al. (2005) already gives us strategy that is near optimal in regret vs communication trade-off.

W68 Locating Changes in Highly Dependent Data with Unknown Number of Change Points

Azadeh Khaleghi azadeh.khaleghi@inria.fr
Daniil Ryabko daniil.ryabko@inria.fr
INRIA

The problem of multiple change point estimation is considered for sequences with unknown number of change points. A consistency framework is suggested that is suitable for highly dependent time-series, and an asymptotically consistent algorithm is proposed. In order for the consistency to be established the only assumption required is that the data is generated by stationary ergodic time-series distributions. No modeling, independence or parametric assumptions are made; the data are allowed to be dependent and the dependence can be of arbitrary form. The theoretical results are complemented with experimental evaluations.

W69 On the (Non-)existence of Convex, Calibrated Surrogate Losses for Ranking

Clément Calauzènes clement.calauzenes@lip6.fr
Nicolas Usunier nicolas.usunier@lip6.fr
Université Pierre et Marie Curie
Patrick Gallinari patrick.gallinari@lip6.fr
University Paris 6

We study surrogate losses for learning to rank, in a framework where the rankings are induced by scores and the task is to learn the scoring function. We focus on the calibration of surrogate losses with respect to a ranking evaluation metric, where the calibration is equivalent to the guarantee that near-optimal values of the surrogate risk imply near-optimal values of the risk defined by the evaluation metric. We prove that if a surrogate loss is a convex function of the scores, then it is not calibrated with respect to two evaluation metrics widely used for search engine evaluation, namely the Average Precision and the Expected Reciprocal Rank. We also show that such convex surrogate losses cannot be calibrated with respect to the Pairwise Disagreement, an evaluation metric used when learning from pairwise preferences. Our results cast lights on the intrinsic difficulty of some ranking problems, as well as on the limitations of learning-to-rank algorithms based on the minimization of a convex surrogate risk.

W70 Classification Calibration Dimension for General Multiclass Losses

Harish Guruprasad harish_gurup@csa.iisc.ernet.in
Shivani Agarwal shivani@csa.iisc.ernet.in
Indian Institute of Science

We study consistency properties of surrogate loss functions for general multiclass classification problems, defined by a general loss matrix. We extend the notion of classification calibration, which has been studied for binary and multiclass 0-1 classification problems (and for certain other specific learning problems), to the general multiclass setting, and derive necessary and sufficient conditions for a surrogate loss to be classification calibrated with respect to a loss matrix in this setting. We then introduce the notion of \emph{classification calibration dimension} of a multiclass loss matrix, which measures the smallest `size' of a prediction space for which it is possible to design a convex surrogate that is classification calibrated with respect to the loss matrix. We derive both upper and lower bounds on this quantity, and use these results to analyze various loss matrices. In particular, as one application, we provide a different route from the recent result of Duchi et al.\ (2010) for analyzing the difficulty of designing `low-dimensional' convex surrogates that are consistent with respect to pairwise subset ranking losses. We anticipate the classification calibration dimension may prove to be a useful tool in the study and design of surrogate losses for general multiclass learning problems.

W71 A Linear Time Active Learning Algorithm for Link Classification

Nicolò Cesa-Bianchi nicolo.cesa-bianchi@unimi.it
Fabio Vitale fabio.vitale@unimi.it
Giovanni Zappella giovanni.zappella@unimi.it
Università degli Studi di Milano
Claudio Gentile claudio.gentile@uninsubria.it
Universita' dell'Insubria

We present very efficient active learning algorithms for link classification in signed networks. Our algorithms are motivated by a stochastic model in which edge labels are obtained through perturbations of a initial sign assignment consistent with a two-clustering of the nodes. We provide a theoretical analysis within this model, showing that we can achieve an optimal (to whithin a constant factor) number of mistakes on any graph G=(V,E) such that |E| is at least order of |V|3/2 by querying at most order of |V|3/2 edge labels. More generally, we show an algorithm that achieves optimality to within a factor of order k by querying at most order of |V|+(|V|/k)3/2 edge labels. The running time of this algorithm is at most of order |E|+|V|log|V|.

W72 Relax and Randomize : From Value to Algorithms

Sasha Rakhlin rakhlin@gmail.com
Karthik Sridharan karthik.sridharan@gmail.com
University of Pennsylvania
Ohad Shamir ohadshamir@gmail.com
Microsoft Research

We show a principled way of deriving online learning algorithms from a minimax analysis. Various upper bounds on the minimax value, previously thought to be non-constructive, are shown to yield algorithms. This allows us to seamlessly recover known methods and to derive new ones, also capturing such ''unorthodox'' methods as Follow the Perturbed Leader and the R^2 forecaster. Understanding the inherent complexity of the learning problem thus leads to the development of algorithms. To illustrate our approach, we present several new algorithms, including a family of randomized methods that use the idea of a ''random play out''. New versions of the Follow-the-Perturbed-Leader algorithms are presented, as well as methods based on the Littlestone's dimension, efficient methods for matrix completion with trace norm, and algorithms for the problems of transductive learning and prediction with static experts.

W73 Mirror Descent Meets Fixed Share (and feels no regret)

Nicolò Cesa-Bianchi nicolo.cesa-bianchi@unimi.it
Università degli Studi di Milano
Pierre Gaillard pierre.gaillard@ens.fr
Gilles Stoltz gilles.stoltz@ens.fr
École Normale Supérieure
Gabor Lugosi gabor.lugosi@gmail.com
Pompeu Fabra University

Mirror descent with an entropic regularizer is known to achieve shifting regret bounds that are logarithmic in the dimension. This is done using either a carefully designed projection or by a weight sharing technique. Via a novel unified analysis, we show that these two approaches deliver essentially equivalent bounds on a notion of regret generalizing shifting, adaptive, discounted, and other related regrets. Our analysis also captures and extends the generalized weight sharing technique of Bousquet and Warmuth, and can be refined in several ways, including improvements for small losses and adaptive tuning of parameters.

W74 Best Arm Identification: A Unified Approach to Fixed Budget and Fixed Confidence

Victor Gabillon victor.gabillon@inria.fr
Mohammad Ghavamzadeh
mohammad.ghavamzadeh@inria.fr
Alessandro Lazaric alessandro.lazaric@inria.fr
INRIA Lille-Nord Europe

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

W75 Mixability in Statistical Learning

Tim van Erven tim@timvanerven.nl
Université Paris-Sud
Peter Grunwald pdg@cwi.nl
CWI
Mark Reid mark.reid@anu.edu.au
Robert Williamson Bob.Williamson@anu.edu.au
Australian National University

Statistical learning and sequential prediction are two different but related formalisms to study the quality of predictions. Mapping out their relations and transferring ideas is an active area of investigation. We provide another piece of the puzzle by showing that an important concept in sequential prediction, the mixability of a loss, has a natural counterpart in the statistical setting, which we call stochastic mixability. Just as ordinary mixability characterizes fast rates for the worst-case regret in sequential prediction, stochastic mixability characterizes fast rates in statistical learning. We show that, in the special case of log-loss, stochastic mixability reduces to a well-known (but usually unnamed) martingale condition, which is used in existing convergence theorems for minimum description length and Bayesian inference. In the case of 0/1-loss, it reduces to the margin condition of Mammen and Tsybakov, and in the case that the model under consideration contains all possible predictors, it is equivalent to ordinary mixability.

W76 No-Regret Algorithms for Unconstrained Online Convex Optimization

Matt Streeter matt@duolingo.com
Duolingo
Brendan McMahan mcmahan@google.com
Google

Some of the most compelling applications of online convex optimization, including online prediction and classification, are unconstrained: the natural feasible set is R^n. Existing algorithms fail to achieve sub-linear regret in this setting unless constraints on the comparator point x* are known in advance. We present an algorithm that, without such prior knowledge, offers near-optimal regret bounds with respect to _any_ choice of x*. In particular, regret with respect to x* = 0 is _constant_. We then prove lower bounds showing that our algorithm's guarantees are optimal in this setting up to constant factors.

W77 Confusion-Based Online Learning and a Passive-Aggressive Scheme

Liva Ralaivola liva.ralaivola@lif.univ-mrs.fr
Aix-Marseille University

This paper provides the first ---to the best of our knowledge--- analysis of online learning algorithms for multiclass problems when the {\em confusion} matrix is taken as a performance measure. The work builds upon recent and elegant results on noncommutative concentration inequalities, i.e. concentration inequalities that apply to matrices, and more precisely to matrix martingales. We do establish generalization bounds for online learning algorithm and show how the theoretical study motivate the proposition of a new confusion-friendly learning procedure. This learning algorithm, called \ copa (for COnfusion Passive-Aggressive) is a passive-aggressive learning algorithm; it is shown that the update equations for \copa can be computed analytically, thus allowing the user from having to recours to any optimization package to implement it.

W78 Dimensionality Dependent PAC-Bayes Margin Bound

Chi Jin chijin06@gmail.com
Liwei Wang wanglw@cis.pku.edu.cn
Peking University

Margin is one of the most important concepts in machine learning. Previous margin bounds, both for SVM and for boosting, are dimensionality independent. A major advantage of this dimensionality independency is that it can explain the excellent performance of SVM whose feature spaces are often of high or infinite dimension. In this paper we address the problem whether such dimensionality independency is intrinsic for the margin bounds. We prove a dimensionality dependent PAC-Bayes margin bound. The bound is monotone increasing with respect to the dimension when keeping all other factors fixed. We sho that our bound is strictly sharper than a previously well-known PAC-Bayes margin bound if the feature space is of finite dimension; and the two bounds tend to be equivalent as the dimension goes to infinity. In addition, we show that the VC bound for linear classifiers can be recovered from our bound under mild conditions. We conduct extensive experiments on benchmark datasets and find that the new bound is useful for model selection and is significantly sharper than the dimensionality independent PAC-Bayes margin bound as well as the VC bound for linear classifiers.

W79 The variational hierarchical EM algorithm for clustering hidden Markov models.

Emanuele Coviello ecoviell@ucsd.edu
Gert Lanckriet gert@ece.ucsd.edu
University of California, San Diego
Antoni Chan abchan@cityu.edu.hk
City University of Hong Kong

In this paper, we derive a novel algorithm to cluster hidden Markov models (HMMs) according to their probability distributions. We propose a variational hierarchical EM algorithm that i) clusters a given collection of HMMs into groups of HMMs that are similar, in terms of the distributions they represent, and ii) characterizes each group by a ``cluster center'', i.e., a novel HMM that is representative for the group. We illustrate the benefits of the proposed algorithm on hierarchical clustering of motion capture sequences as well as on automatic music tagging.

W80 FastEx: Fast Clustering with Exponential Families

Amr Ahmed amrahmed@yahoo-inc.com
Shravan Narayanamurthy shravanm@yahoo-inc.com
Yahoo! Research
Sujith Ravi ravi.sujith@gmail.com
Alexander Smola alex@smola.org
Google Inc.

Clustering is a key component in data analysis toolbox. Despite its importance, scalable algorithms often eschew rich statistical models in favor of simpler descriptions such as k-means clustering. In this paper we present a sampler, capable of estimating mixtures of exponential families. At its heart lies a novel proposal distribution using random projections to achieve high throughput in generating proposals, which is crucial for clustering models with large numbers of clusters.

W81 Clustering Sparse Graphs

Yudong Chen ydchen@utexas.edu
Sujay Sanghavi sanghavi@mail.utexas.edu
University of Texas, Austin
Huan Xu xuhuan@cim.mcgill.ca
NUS

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

W82 Symmetric Correspondence Topic Models for Multilingual Text Analysis

Kosuke Fukumasu fukumasu@cs25.scitec.kobe-u.ac.jp
Koji Eguchi eguchi@port.kobe-u.ac.jp
Kobe University
Eric Xing epxing@cs.cmu.edu
Carnegie Mellon University

Topic modeling is a widely used approach to analyzing large text collections. A small number of multilingual topic models have recently been explored to discover latent topics among parallel or comparable documents, such as in Wikipedia. Other topic models that were originally proposed for structured data are also applicable to multilingual documents. Correspondence Latent Dirichlet Allocation (CorrLDA) is one such model; however, it requires a pivot language to be specified in advance. We propose a new topic model, Symmetric Correspondence LDA (SymCorrLDA), that incorporates a hidden variable to control a pivot language, in an extension of CorrLDA. We experimented with two multilingual comparable datasets extracted from Wikipedia and demonstrate that SymCorrLDA is more effective than some other existing multilingual topic models.

W83 Factorial LDA: Sparse Multi-Dimensional Text Models

Michael Paul mpaul39@gmail.com
Mark Dredze mdredze@cs.jhu.edu
Johns Hopkins University

Multi-dimensional latent variable models can capture the many latent factors in a text corpus, such as topic, author perspective and sentiment. We introduce factorial LDA, a multi-dimensional latent variable model in which a document is influenced by K different factors, and each word token depends on a K-dimensional vector of latent variables. Our model incorporates structured word priors and learns a sparse product of factors. Experiments on research abstracts show that our model can learn latent factors such as research topic, scientific discipline, and focus (e.g. methods vs. applications.) Our modeling improvements reduce test perplexity and improve human interpretability of the discovered factors.

W84 Visual Recognition using Embedded Feature Selection for Curvature Self-Similarity

Angela Eigenstetter
angela.eigenstetter@iwr.uni-heidelberg.de
Bjorn Ommer ommer@uni-heidelberg.de
University of Heidelberg

Category-level object detection has a crucial need for informative object representations. This demand has led to feature descriptors of ever increasing dimensionality like co-occurrence statistics and self-similarity. In this paper we propose a new object representation based on curvature self-similarity that goes beyond the currently popular approximation of objects using straight lines. However, like all descriptors using second order statistics, ours also exhibits a high dimensionality. Although improving discriminability, the high dimensionality becomes a critical issue due to lack of generalization ability and curse of dimensionality. Given only a limited amount of training data, even sophisticated learning algorithms such as the popular kernel methods are not able to suppress noisy or superfluous dimensions of such high-dimensional data. Consequently, there is a natural need for feature selection when using present-day informative features and, particularly, curvature self-similarity. We therefore suggest an embedded feature selection method for SVMs that reduces complexity and improves generalization capability of object models. By successfully integrating the proposed curvature self-similarity representation together with the embedded feature selection in a widely used state-of-the-art object detection framework we show the general pertinence of the approach.

W85 Context-Sensitive Decision Forests for Object Detection

Peter Kontschieder kontschieder@icg.tugraz.at
Graz University of Technology
Samuel Rota Bulò srotabul@dais.unive.it
Marcello Pelillo marcello.pelillo@gmail.com
Università Ca' Foscari Venezia, Italy
Antonio Criminisi antcrim@microsoft.com
Pushmeet Kohli pkohli@microsoft.com
Microsoft Research
Horst Bischof bischof@icg.tugraz.at
TU Graz

In this paper we introduce Context-Sensitive Decision Forests - A new perspective to exploit contextual information in the popular decision forest framework for the object detection problem. They are tree-structured classifiers with the ability to access intermediate prediction (here: classification and regression) information during training and inference time. This intermediate prediction is available to each sample, which allows us to develop context-based decision criteria, used for refining the prediction process. In addition, we introduce a novel split criterion which in combination with a priority based way of constructing the trees, allows more accurate regression mode selection and hence improves the current context information. In our experiments, we demonstrate improved results for the task of pedestrian detection on the challenging TUD data set when compared to state-of-the-art methods.

W86 Human memory search as a random walk in a semantic network

Joshua Abbott joshua.abbott@berkeley.edu
Tom Griffiths tom_griffiths@berkeley.edu
University of California, Berkeley
Joseph Austerweil joseph.austerweil@gmail.com
UC Berkeley

The human mind has a remarkable ability to store a vast amount of information in memory, and an even more remarkable ability to retrieve these experiences when needed. Understanding the representations and algorithms that underlie human memory search could potentially be useful in other information retrieval settings, including internet search. Psychological studies have revealed clear regularities in how people search their memory, with clusters of semantically related items tending to be retrieved together. These findings have recently been taken as evidence that human memory search is similar to animals foraging for food in patchy environments, with people making a rational decision to switch away from a cluster of related information as it becomes depleted. We demonstrate that the results that were taken as evidence for this account also emerge from a random walk on a semantic network, much like the random web surfer model used in internet search engines. This offers a simpler and more unified account of how people search their memory, postulating a single process rather than one process for exploring a cluster and one process for switching between clusters.

W87 Transferring Expectations in Model-based Reinforcement Learning

Trung Nguyen nttrung@comp.nus.edu.sg
Tomi Silander silander@comp.nus.edu.sg
Tze Yun Leong leongty@comp.nus.edu.sg
National University of Singapore

We study how to automatically select and adapt multiple abstractions or representations of the world to support model-based reinforcement learning. We address the challenges of transfer learning in heterogeneous environments with varying tasks. We present an efficient, online framework that, through a sequence of tasks, learns a set of relevant representations to be used in future tasks. Without pre-defined mapping strategies, we introduce a general approach to support transfer learning across different state spaces. We demonstrate the potential impact of our system through improved jumpstart and faster convergence to near optimum policy in two benchmark domains.

W88 Regularized Off-Policy TD-Learning

Bo Liu boliu@cs.umass.edu
Sridhar Mahadevan mahadeva@cs.umass.edu
University of Massachusetts Amherst
Ji Liu ji.liu@asu.edu
University Wisconsin-Madison

We present a novel l1 regularized off-policy convergent TD-learning method (termed RO-TD), which is able to learn sparse representations of value functions with low computational complexity. The algorithmic framework underlying RO-TD integrates two key ideas: off-policy convergent gradient TD methods, such as TDC, and a convex-concave saddle-point formulation of non-smooth convex optimization, which enables first-order solvers and feature selection using online convex regularization. A detailed theoretical and experimental analysis of RO-TD is presented. A variety of experiments are presented to illustrate the off-policy convergence, sparse feature selection capability and low computational cost of the RO-TD algorithm.

W89 Weighted Likelihood Policy Search with Model Selection

Tsuyoshi Ueno ueno@ar.sanken.osaka-u.ac.jp
Japan Science and Technology
Yoshinobu Kawahara kawahara@ar.sanken.osaka-u.ac.jp
Takashi Washio washio@ar.sanken.osaka-u.ac.jp
Osaka University
Kohei Hayashi hayashi.kohei@gmail.com
The University of Tokyo

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

W90 A mechanistic model of early sensory processing based on subtracting sparse representations

Shaul Druckmann druckmanns@janelia.hhmi.org
Janelia Farm Research Campus
Tao Hu hut@janelia.hhmi.org
JFRC, HHMI
Dmitri Chklovskii chklovskiid@janelia.hhmi.org
HHMI

Early stages of sensory systems face the challenge of compressing information from numerous receptors onto a much smaller number of projection neurons, a so called communication bottleneck. To make more efficient use of limited bandwidth, compression may be achieved using predictive coding, whereby predictable, or redundant, components of the stimulus are removed. In the case of the retina, Srinivasan et al. (1982) suggested that feedforward inhibitory connections subtracting a linear prediction generated from nearby receptors implement such compression, resulting in biphasic center-surround receptive fields. However, feedback inhibitory circuits are common in early sensory circuits and furthermore their dynamics may be nonlinear. Can such circuits implement predictive coding as well? Here, solving the transient dynamics of nonlinear reciprocal feedback circuits through analogy to a signal-processing algorithm called linearized Bregman iteration we show that nonlinear predictive coding can be implemented in an inhibitory feedback circuit. In response to a step stimulus, interneuron activity in time constructs progressively less sparse but more accurate representations of the stimulus, a temporally evolving prediction. This analysis provides a powerful theoretical framework to interpret and understand the dynamics of early sensory processing in a variety of physiological experiments and yields novel predictions regarding the relation between activity and stimulus statistics.

W91 High-Order Multi-Task Feature Learning to Identify Longitudinal Phenotypic Markers for Alzheimer Disease Progression Prediction

Hua Wang huawangcs@gmail.com
Feiping Nie feipingnie@gmail.com
Heng Huang heng@uta.edu
University of Texas Arlington
Jingwen Yan jingyan@iupui.edu
Sungeun Kim sk31@iupui.edu
Shannon Risacher srisache@iupui.edu
Andrew Saykin asaykin@iupui.edu
IUPUI
Li Shen shenli@iupui.edu
Indiana University School of Medicine

Alzheimer disease (AD) is a neurodegenerative disorder characterized by progressive impairment of memory and other cognitive functions. Regression analysis has been studied to relate neuroimaging measures to cognitive status. However, whether these measures have further predictive power to infer a trajectory of cognitive performance over time is still an under-explored but important topic in AD research. We propose a novel high-order multi-task learning model to address this issue. The proposed model explores the temporal correlations existing in data features and regression tasks by the structured sparsity-inducing norms. In addition, the sparsity of the model enables the selection of a small number of MRI measures while maintaining high prediction accuracy. The empirical studies, using the baseline MRI and serial cognitive data of the ADNI cohort, have yielded promising results.

W92 Optimal Neural Tuning Curves for Arbitrary Stimulus Distributions: Discrimax, Infomax and Minimum Lp Loss

Jimmy Wang wangzhuo@sas.upenn.edu
Daniel Lee ddlee@seas.upenn.edu
University of Pennsylvania
Alan Stocker astocker@sas.upenn.edu
U Penn

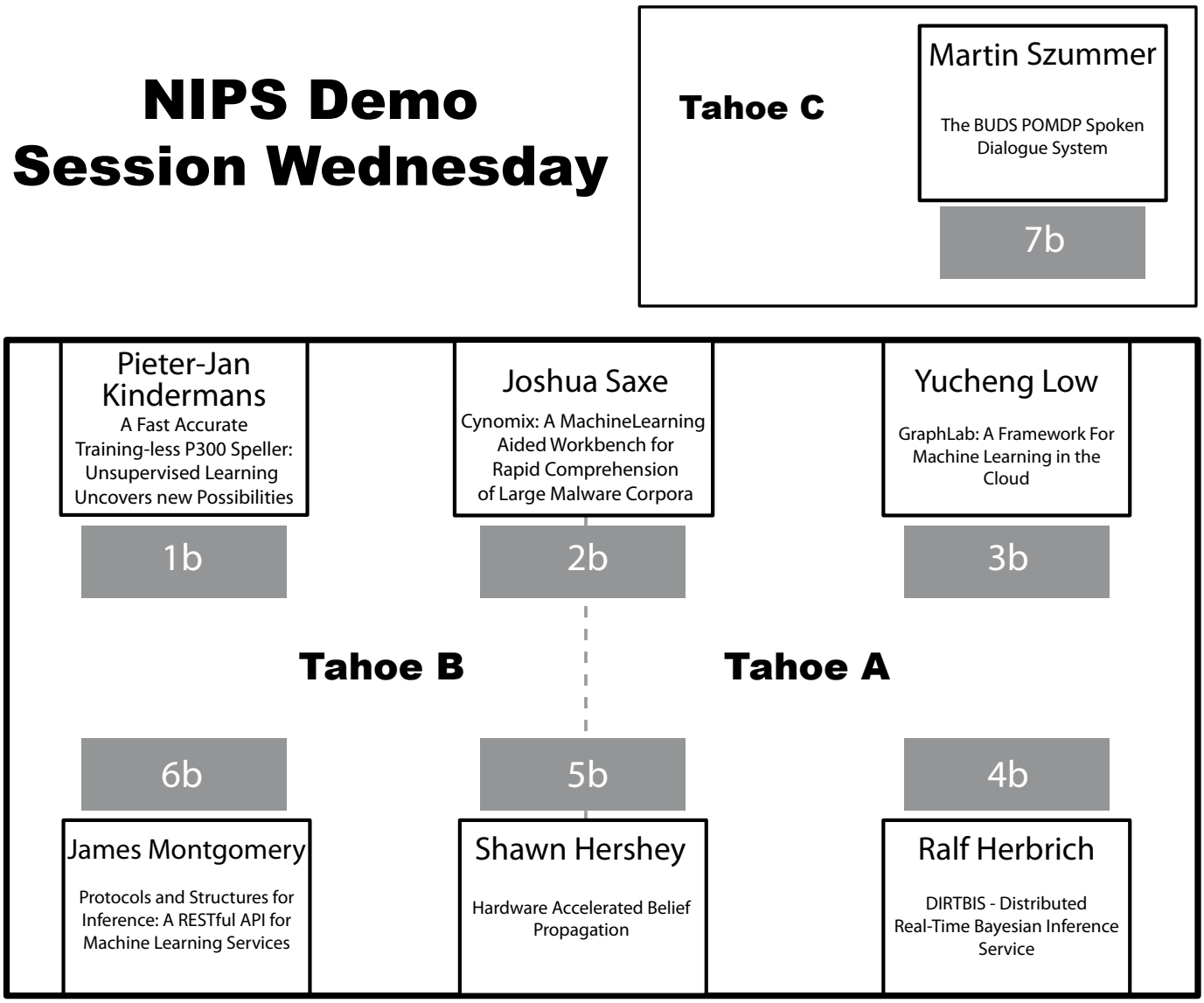
In this work we study how the stimulus distribution influences the optimal coding of an individual neuron. Closed-form solutions to the optimal sigmoidal tuning curve are provided for a neuron obeying Poisson statistics under a given stimulus distribution. We consider a variety of optimality criteria, including maximizing discriminability, maximizing mutual information and minimizing estimation error under a general Lp norm. We generalize the Cramer-Rao lower bound and show how the Lp loss can be written as a functional of the Fisher Information in the asymptotic limit, by proving the moment convergence of certain functions of Poisson random variables. In this manner, we show how the optimal tuning curve depends upon the loss function, and the equivalence of maximizing mutual information with minimizing Lp loss in the limit as p goes to zero.

W93 Identifiability and Unmixing of Latent Parse Trees

Percy Liang pliang@cs.stanford.edu
Stanford University
Sham Kakade skakade@microsoft.com
Daniel Hsu danielhsu@gmail.com
Microsoft Research

This paper explores unsupervised learning of parsing models along two directions. First, which models are identifiable from infinite data? We use a general technique for numerically checking identifiability based on the rank of a Jacobian matrix, and apply it to several standard constituency and dependency parsing models. Second, for identifiable models, how do we estimate the parameters efficiently? EM suffers from local optima, while recent work using spectral methods cannot be directly applied since the topology of the parse tree varies across sentences. We develop a strategy, unmixing, which deals with this additional complexity for restricted classes of parsing models.

NIPS Demo Session Wednesday



1B A Fast Accurate Training-less P300 Speller: Unsupervised Learning Uncovers new Possibilities

Pieter-Jan Kindermans, Hannes Verschore, David Verstraeten, Benjamin Schrauwen
Ghent University

The P300 speller paradigm is a Brain-Computer Interface, introduced in 1988 by Farwell and Donchin, which allows the users to type by simply focusing their attention on the desired symbol in a spelling matrix. This type of BCI has already been applied in a real world setting for patients, but is severely limited by the need for a tedious calibration session before each usage, whereby the system is trained supervisedly. In this demonstration we present our solution to this calibration session: an unsupervised P300 speller. The system is able to quickly learn online, entirely without labeled data, thus bypassing the need for calibration. The spelling accuracies are comparable to supervised spellers.

2B Cynomix: A Machine Learning Aided Workbench for Rapid Comprehension of Large Malware Corpora

Joshua Saxe, David Mentis, Chris Greamo,
Invincea Labs

Although the number of malware samples active on the Internet has risen above ten million and is growing at an exponential rate, in operational contexts today most analysis of malware is still done by hand, sample by sample, by expert reverse engineers. As a result, most malware samples have not been analyzed or understood. We will demonstrate a novel intelligent workbench for analysis of large malware corpora that sees beyond malware code obfuscation to identify code sharing relationships between malware samples. This allows our workbench tool to then propagate analyst annotations between samples when code is reused between samples. We believe our project can help to facilitate a paradigm shift in our approach to understanding the malware landscape, facilitating greater breadth and depth to the security community's understanding of the nature and evolution of malware. Our

system includes four components: a feature extraction component, a code-sharing estimation component, a malware behavioral trait identification component, and a visual interface which ties these components together. Our feature extraction component identifies semantically meaningful subsequences of malware system call behavior logs through a novel Markovian sequence extraction method which runs in linear time. Other features we extract are instruction n-grams from each sample's control flow graph, declared library and function imports, and printable strings information parsed from the binary sample files. To estimate code sharing we use a novel ensemble similarity function that incorporates sample control flow graph information, sample system call log subsequence features, and sample binary file metadata. To compute pairwise similarities we use a locally sensitive hashing technique that allows our system to scale up to tens of thousands of samples. Our code sharing detection approach was evaluated last month by a test team at MIT Lincoln Laboratory and scored extremely well both in absolute terms as well as relative to the other algorithms under test.

3B DIRTBIS - Distributed Real-Time Bayesian Inference Service

Ralf Herbrich
Facebook

No abstract

4B GraphLab: A Framework For Machine Learning in the Cloud

Yucheng Low, Haijie Gu, Carlos Guestrin
Carnegie Mellon University

GraphLab is a graph-based abstraction targeted at solving large scale Machine Learning and data-mining tasks. Our distributed GraphLab implementation scales graphs with billions of vertices and edges easily and outperforms other systems and abstractions by orders of magnitudes. We have implemented a number of toolkits on top of GraphLab ranging from classical graph analytics tasks to collaborative filtering, probabilistic inference and clustering algorithms. All code is open source and is available at http://graphlab.org/

5B Hardware Accelerated Belief Propagation

Shawn Hershey, Ben Vigoda
Analog Devices, Inc.

We introduce Dimple, a fully open-source API for probabilistic modeling. Dimple allows the user to specify probabilistic models in the form of graphical models, bayesian networks, or factor graphs, and performs inference (by automatically deriving an inference engine from a variety of algorithms) on the model. Dimple also serves as a compiler for GP5, a hardware accelerator for inference.

6B Protocols and Structures for Inference: A RESTful API for Machine Learning Services

James Montgomery, Mark Reid
The Australian National University

The last few years have seen rapid growth in the online delivery of machine learning services by a variety of both established and new companies. Each service offers its own, typically RESTful, API and a subset of machine learning techniques. Composing these services is difficult and extending their learning technique offerings impossible for anyone outside their respective development teams. We offer an alternative approach that is flexible and federated: data and learning algorithm providers are given considerable freedom in what they offer, while learning service consumers have the power to easily combine different services to produce new and powerful inference tools. We will demonstrate the features of our RESTful API for machine learning across a number of different example web services.

7B The BUDS POMDP Spoken Dialogue System

Martin Szummer, Matthew Henderson, Catherine Breslin, Milica Gasic, Dongho Kim, Blaise Thomson, Pirros Tsiakoulis, Steve Young
University of Cambridge

Bayesian update of dialogue state (BUDS) is a state-of-the art system for human-computer conversation in dialogues. Here, it is employed to build a speech-driven intelligent assistant. The system manages the conversation to help the user achieve their goal as quickly as possible. The main challenge is to converse in a way that overcomes mistakes made by the speech recognizer, or ambiguous utterances by the user. The system can ask for confirmations, pose choices, and ask for additional information, all in order to gain certainty while maximizing dialogue utility. The system contains a long machine learning pipeline. It preserves a large number of speech recognition hypotheses by representing them as a confusion network (a compact form of an HMM lattice), and applies a semantic decoder directly to this network. The dialogue state is tracked via a Dynamic Belief Network. The system chooses actions according to a policy that has been learned using a POMDP. The ability of the system to maintain uncertainty significantly improves dialogue utility compared to rule-based dialogue systems.

THURSDAY

ORAL SESSION

SESSION 9 - 9:00 - 10:10 AM

Session Chair: Jean-Philippe Vert

POSNER LECTURE: Suspicious Coincidences in the Brain

Terrence Sejnowski terry@salk.edu
Salk Institute

Brains need to make quick sense of massive amounts of ambiguous information with minimal energy costs and have evolved an intriguing mixture of analog and digital mechanisms to allow this efficiency. Analog electrical and biochemical signals inside neurons are used for integrating synaptic inputs from other neurons. The digital part is the all-or-none action potential, or spike, that lasts for a millisecond or less and is used to send messages over a long distance. Spike coincidences occur when neurons fire together at nearly the same time. In this lecture I will show how rare spike coincidences can be used efficiently to represent important visual events and how this architecture can be implemented with analog VLSI technology to simplify the early stages of visual processing.

Terrence Sejnowski *received his PhD in physics from Princeton University and was a postdoctoral fellow in the Department of Neurobiology at Harvard Medical School. He was on the faculty at the Johns Hopkins University and now holds the Francis Crick Chair at The Salk Institute for Biological Studies and is also a Professor of Biology at the University of California, San Diego, where he is co-director of the Institute for Neural Computation and co-director of the NSF Temporal Dynamics of Learning Center. He is the President of the Neural Information Processing Systems (NIPS) Foundation, which organizes an annual conference attended by over 1500 researchers in machine learning and neural computation and is the founding editor-in-chief of Neural Computation published by the MIT Press. An investigator with the Howard Hughes Medical Institute, he is also a Fellow of the American Association for the Advancement of Science, a Fellow of the Institute of Electrical and Electronics Engineers and a Fellow of the Cognitive Science Society. He has received many honors, including the NSF Young Investigators Award, the Wright Prize for interdisciplinary research from the Harvey Mudd College, the Neural Network Pioneer Award from the Institute of Electrical and Electronics Engineers and the Hebb Prize from the International Neural Network Society. He was elected to the Institute of Medicine in 2008, to the National Academy of Sciences in 2010, and to the National Academy of Engineering in 2011. He is one of only 10 living persons to be a member of all 3 national academies.*

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

Pradeep Shenoy pshenoy@ucsd.edu
Angela Yu ajyu@ucsd.edu
University of California, San Diego

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

SPOTLIGHT SESSION

SESSION 9 - 10:10 - 10:30 AM

- **Complex Inference in Neural Circuits with Probabilistic Population Codes and Topic Models**
  J. Beck, A. Pouget, University of Rochester; K. Heller, Duke University
  See abstract Th92, page 132

- **Identification of Recurrent Patterns in the Activation of Brain Networks**
  f. janoos, ExxonMobil Corp; W. Li, ExxonMobil Research; N. Subrahmanya, ExxonMobil Corporate Research; I. Morocz, W. Wells, Harvard Medical School
  See abstract Th84, page 130

- **Delay Compensation with Dynamical Synapses**
  C. Fung, The Hong Kong University of Science and Technology; K. Wong, Department of Physics, Hong Kong University of Science and Technology; S. Wu, Beijing Normal University
  See abstract Th90, page 132

- **Efficient Spike-Coding with Multiplicative Adaptation in a Spike Response Model**
  S. Bohte, Centrum Wiskunde Informatica
  See abstract Th87, page 131

- **Deep Spatio-Temporal Architectures and Learning for Protein Structure Prediction**
  P. Lena, UCI; P. Baldi, K. Nagata, UC Irvine
  See abstract Th23, page 117

ORAL SESSION

SESSION 10 - 10:50 AM - 12:00 PM

Session Chair: Alexander Smola

INVITED TALK: Fast Algorithms, Matrix Compression and Design by Simulation

Leslie Greengard greengar@cims.nyu.edu
New York University

During the last two decades, a variety of fast algorithms have been developed for large-scale problems in scientific computing, governed by the equations of electromagnetics, elasticity, and fluid mechanics. They are most easily understood, perhaps, in the case of particle simulations, where they reduce the cost of computing all pairwise interactions in a system of N particles from O(N^2) to O(N) or O(N log N) operations. Most recently, a number of researchers have been developing an infrastructure for such problems using a linear algebraic formulation that makes closer connections to some central ideas in machine learning. We will describe the computational foundations of these methods, as well as some of their applications to the problems of design in geometrically complicated environments.

Leslie Greengard *received a B.A. degree in Mathematics from Wesleyan University in 1979, and an M.D./Ph.D. degree from Yale University in 1987, with the Ph.D. in Computer Science. Since 1989, he has been at the Courant Institute of Mathematical Sciences, New York University, where he a Professor of Mathematics and Computer Science. He was the director of the Institute from 2006-2011. The research in Prof. Greengard's group focuses on fast, adaptive, and high-order accurate algorithms in computational physics and engineering, including electromagnetics, acoustics, heat flow and biomedical imaging. Among his honors are the Leroy P. Steele prize (with V. Rokhlin) from the American Mathematical Society, and plenary/invited talks at the International Congress on Industrial and Applied Mathematics, and the International Congress of Mathematicians. He is a member of both the National Academy of Sciences and the National Academy of Engineering.*

Gradient Weights help Nonparametric Regressors

Samory Kpotufe samory@ttic.edu
Toyota Technological Institute

Abdeslam Boularias boularias@tuebingen.mpg.de
Max Planck Institute for Intelligent Systems

In regression problems over \reald, the unknown function f often varies more in some coordinates than in others. We show that weighting each coordinate i with the estimated norm of the ith derivative of f is an efficient way to significantly improve the performance of distance-based regressors, e.g. kernel and k-NN regressors. We propose a simple estimator of these derivative norms and prove its consistency. Moreover, the proposed estimator is efficiently learned online.

SPOTLIGHT SESSION

SESSION 10 - 12:00 - 12:20 PM

- **Sparse Prediction with the k-Support Norm**
  A. Argyriou, Ecole Centrale de Paris; R. Foygel, Stanford University; N. Srebro, TTI-Chicago
  See abstract Th60, page 125

- **Compressive neural representation of sparse, high-dimensional probabilities**
  X. Pitkow, University of Rochester
  See abstract Th27, page 118

- **Exact and Stable Recovery of Sequences of Signals with Sparse Increments via Differential ℓ1-Minimization**
  D. Ba, B. Babadi, P. Purdon, E. Brown, MIT/Harvard
  See abstract Th26, page 117

- **Fused sparsity and robust estimation for linear models with unknown variance**
  A. Dalalyan, ENSAE - CREST; Y. Chen, Ecole des Ponts ParisTech
  See abstract Th61, page 125

- **Selecting Diverse Features via Spectral Regularization**
  A. Das, A. Dasgupta, R. Kumar, Yahoo!
  See abstract Th59, page 124

POSTER SESSION

SESSION - 2:00 - 6:00 PM

Th1 **Bandit Algorithms boost Brain Computer Interfaces for motor-task selection of a brain-controlled button**
J. Fruitet, A. Carpentier, R. Munos, M. Clerc

Th2 **Patient Risk Stratification for Hospital-Associated C. Diff as a Time-Series Classification Task**
J. Wiens, J. Guttag, E. Horvitz

Th3 **Learning Label Trees for Probabilistic Modelling of Implicit Feedback**
A. Mnih, Y. Teh

Th4 **Collaborative Ranking With 17 Parameters**
M. Volkovs, R. Zemel

Th5 **A dynamic excitatory-inhibitory network in a VLSI chip for spiking information reregistration**
J. Huo

Th6 **GenDeR: A Generic Diversified Ranking Algorithm**
J. He, H. Tong, Q. Mei, B. Szymanski

Th7 **Iterative ranking from pair-wise comparisons**
S. Negahban, S. Oh, D. Shah

Th8 **Accuracy at the Top**
S. Boyd, C. Cortes, M. Mohri, A. Radovanovic

Th9 Learning to Discover Social Circles in Ego Networks
J. McAuley, J. Leskovec

Th10 Cocktail Party Processing via Structured Prediction
Y. Wang, D. Wang

Th11 On the Use of Non-Stationary Policies for Stationary Infinite-Horizon Markov Decision Processes
B. Scherrer, B. Lesner

Th12 Non-parametric Approximate Dynamic Programming via the Kernel Method
N. Bhat, C. Moallemi, V. Farias

Th13 Imitation Learning by Coaching
H. He, H. Daume III, J. Eisner

Th14 Inverse Reinforcement Learning through Structured Classification
E. Klein, M. Geist, B. PIOT, O. Pietquin

Th15 Efficient Bayes-Adaptive Reinforcement Learning using Sample-Based Search
A. Guez, D. Silver, P. Dayan

Th16 A Bayesian Approach for Policy Learning from Trajectory Preference Queries
A. Wilson, A. Fern, P. Tadepalli

Th17 Value Pursuit Iteration
A. Farahmand, D. Precup

Th18 Online Regret Bounds for Undiscounted Continuous Reinforcement Learning
R. Ortner, D. Ryabko

Th19 Robustness and risk-sensitivity in Markov decision processes
T. Osogami

Th20 Trajectory-Based Short-Sighted Probabilistic Planning
F. Trevizan, M. Veloso

Th21 Cost-Sensitive Exploration in Bayesian Reinforcement Learning
D. Kim, K. Kim, P. Poupart

Th22 Recursive Deep Learning on 3D Point Clouds
R. Socher, B. Bath, B. Huval, C. Manning, A. Ng

Th23 Deep Spatio-Temporal Architectures and Learning for Protein Structure Prediction
P. Lena, P. Baldi, K. Nagata

Th24 Image Denoising and Inpainting with Deep Neural Networks
J. Xie, L. Xu, E. Chen

Th25 ImageNet Classification with Deep Convolutional Neural Networks
A. Krizhevsky, I. Sutskever, G. Hinton

Th26 Exact and Stable Recovery of Sequences of Signals with Sparse Increments via Differential ℓ1-Minimization
D. Ba, B. Babadi, P. Purdon, E. Brown

Th27 Compressive neural representation of sparse, high-dimensional probabilities
X. Pitkow

Th28 Shifting Weights: Adapting Object Detectors from Image to Video
K. Tang, V. Ramanathan, F. Li, D. Koller

Th29 Multi-Stage Multi-Task Feature Learning
P. Gong, J. Ye, C. Zhang

Th30 Factoring nonnegative matrices with linear programs
B. Recht, C. Re, J. Tropp, V. Bittorf

Th31 Proximal Newton-type Methods for Minimizing Convex Objective Functions in Composite Form
J. Lee, Y. Sun, M. Saunders

Th32 Communication/Computation Tradeoffs in Consensus-Based Distributed Optimization
K. Tsianos, S. Lawlor, M. Rabbat

Th33 Recovery of Sparse Probability Measures via Convex Programming
M. Pilanci, L. El Ghaoui, V. Chandrasekaran

Th34 Newton-Like Methods for Sparse Inverse Covariance Estimation
P. Olsen, F. Oztoprak, J. Nocedal, S. Rennie

Th35 A quasi-Newton proximal splitting method
S. Becker, J. Fadili

Th36 Query Complexity of Derivative-Free Optimization
K. Jamieson, R. Nowak, B. Recht

Th37 CPRL -- An Extension of Compressive Sensing to the Phase Retrieval Problem
H. Ohlsson, A. Yang, R. Dong, S. Sastry

Th38 Joint Modeling of a Matrix with Associated Text via Latent Binary Features
X. Zhang, L. Carin

Th39 Probabilistic Low-Rank Subspace Clustering
S. Babacan, S. Nakajima, M. Do

Th40 Bayesian n-Choose-k Models for Classification and Ranking
K. Swersky, D. Tarlow, R. Zemel, R. Adams, B. Frey

Th41 Bayesian Nonparametric Modeling of Suicide Attempts
F. Ruiz, I. Valera, C. Blanco, F. Perez-Cruz

Th42 Bayesian nonparametric models for bipartite graphs
F. Caron

Th43 Coupling Nonparametric Mixtures via Latent Dirichlet Processes
D. Lin, J. Fisher

Th44 Multiresolution Gaussian Processes
E. Fox, D. Dunson

Th45 Bayesian Warped Gaussian Processes
M. Lázaro-Gredilla

Th46 Collaborative Gaussian Processes for Preference Learning
N. Houlsby, J. Hernández-Lobato, F. Huszar, Z. Ghahramani

Th47 Nonparanormal Belief Propagation (NPBP)
G. Elidan, C. Cario

Th48 Latent Coincidence Analysis: A Hidden Variable Model for Distance Metric Learning
M. Der, L. Saul

Th49 Multiple Choice Learning: Learning to Produce Multiple Structured Outputs
A. Guzmán-Rivera, D. Batra, P. Kohli

Th50 Learning from the Wisdom of Crowds by Minimax Entropy
D. Zhou, S. Basu, Y. Mao, J. Platt

Th51 Bayesian models for Large-scale Hierarchical Classification
S. Gopal, Y. Yang, B. Bai, A. Niculescu-Mizil

Th52 Multiple Operator-valued Kernel Learning
H. Kadri, A. Rakotomamonjy, F. Bach, p. preux

Th53 Gradient-based kernel method for feature extraction and variable selection
K. Fukumizu, C. Leng

Th54 Learning from Distributions via Support Measure Machines
K. Muandet, K. Fukumizu, F. Dinuzzo, B. Schölkopf

Th55 Nonparametric Reduced Rank Regression
R. Foygel, M. Horrell, M. Drton, J. Lafferty

Th56 Pointwise Tracking the Optimal Regression Function
Y. Wiener, R. El-Yaniv

Th57 Link Prediction in Graphs with Autoregressive Features
E. Richard, S. Gaiffas, N. Vayatis

Th58 Gradient Weights help Nonparametric Regressors
S. Kpotufe, A. Boularias

Th59 Selecting Diverse Features via Spectral Regularization
A. Das, A. Dasgupta, R. Kumar

Th60 Sparse Prediction with the k-Support Norm
A. Argyriou, R. Foygel, N. Srebro

Th61 Fused sparsity and robust estimation for linear models with unknown variance
A. Dalalyan, Y. Chen

Th62 Dual-Space Analysis of the Sparse Linear Model
D. Wipf

Th63 Entropy Estimations Using Correlated Symmetric Stable Random Projections
P. Li, C. Zhang

Th64 Reducing statistical time-series problems to binary classification
D. Ryabko, J. Mary

Th65 On Multilabel Classification and Ranking with Partial Feedback
C. Gentile, F. Orabona

Th66 Nystr{ö}m Method vs Random Fourier Features: A Theoretical and Empirical Comparison
T. Yang, Y. Li, M. Mahdavi, R. Jin, Z. Zhou

Th67 Learning Manifolds with K-Means and K-Flats
G. Canas, T. Poggio, L. Rosasco

Th68 Selective Labeling via Error Bound Minimization
Q. Gu, T. Zhang, C. Ding, J. Han

Th69 Semi-Crowdsourced Clustering: Generalizing Crowd Labeling by Robust Distance Metric Learning
J. Yi, R. Jin, A. Jain, S. Jain

Th70 Forging The Graphs: A Low Rank and Positive Semidefinite Graph Learning Approach
D. Luo, C. Ding, H. Huang

Th71 Hamming Distance Metric Learning
M. Norouzi, R. Salakhutdinov, D. Fleet

Th72 Parametric Local Metric Learning for Nearest Neighbor Classification
J. Wang, A. Kalousis, A. Woznica

Th73 Non-linear Metric Learning
D. Kedem, S. Tyree, K. Weinberger, F. Sha, G. Lanckriet

Th74 Monte Carlo Methods for Maximum Margin Supervised Topic Models
Q. Jiang, J. Zhu, M. Sun, E. Xing

Th75 Topic-Partitioned Multinetwork Embeddings
P. Krafft, J. Moore, H. Wallach, B. Desmarais

Th76 Learning with Recursive Perceptual Representations
O. Vinyals, Y. Jia, L. Deng, T. Darrell

Th77 Natural Images, Gaussian Mixtures and Dead Leaves
D. Zoran, Y. Weiss

Th78 Deep Learning of invariant features via tracked video sequences
W. Zou, A. Ng, S. Zhu, K. Yu

Th79 Dynamical And-Or Graph Learning for Object Shape Modeling and Detection
x. wang, L. Lin

Th80 Searching for objects driven by context
B. Alexe, N. Heess, Y. Teh, V. Ferrari

Th81 Learning Image Descriptors with the Boosting-Trick
T. Trzcinski, M. Christoudias, V. Lepetit, P. Fua

Th82 Exploration in Model-based Reinforcement Learning by Empirically Estimating Learning Progress
M. Lopes, T. Lang, M. Toussaint, P. Oudeyer

Th83 Learning optimal spike-based representations
R. Bourdoukan, D. Barrett, C. Machens, S. Deneve

Th84 Identification of Recurrent Patterns in the Activation of Brain Networks
f. janoos, W. Li, N. Subrahmanya, I. Morocz, W. Wells

Th85 Efficient and direct estimation of a neural subunit model for sensory coding
B. Vintch, A. Zaharia, J. Movshon, E. Simoncelli

Th86 How Prior Probability Influences Decision Making: A Unifying Probabilistic Model
Y. Huang, A. Friesen, T. Hanks, M. Shadlen, R. Rao

Th87 Efficient Spike-Coding with Multiplicative Adaptation in a Spike Response Model
S. Bohte

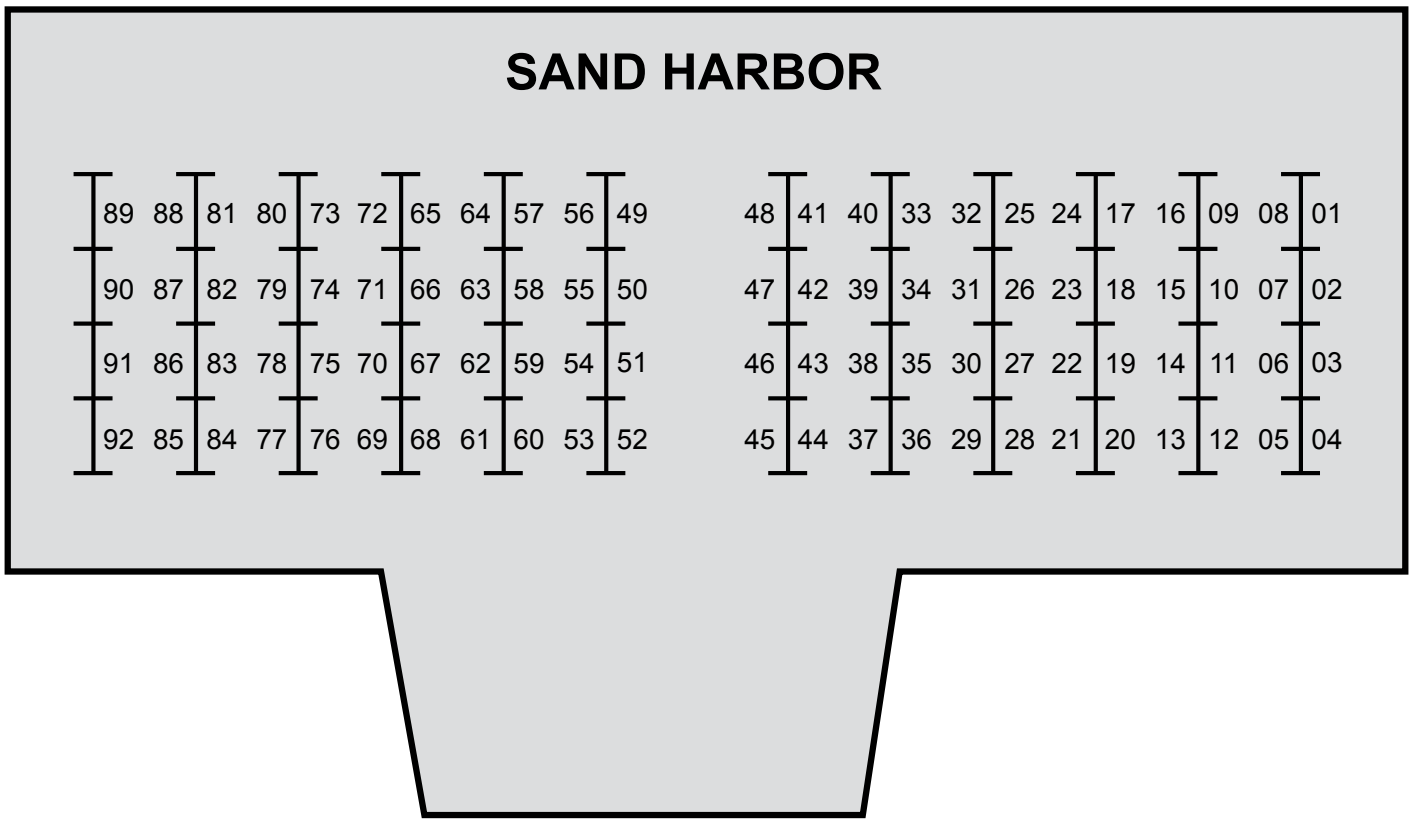
Th88 Te topographic unsupervised learning of natural sounds in the auditory cortex
H. Terashima, M. Okada

Th89 Strategic Impatience in Go/NoGo versus Forced-Choice Decision-Making
P. Shenoy, A. Yu

Th90 Delay Compensation with Dynamical Synapses
C. Fung, K. Wong, S. Wu

Th91 Neuronal spike generation mechanism as an oversampling, noise-shaping A-to-D converter
D. Chklovskii

Th92 Complex Inference in Neural Circuits with Probabilistic Population Codes and Topic Models
J. Beck, K. Heller, A. Pouget



Th1 Bandit Algorithms boost Brain Computer Interfaces for motor-task selection of a brain-controlled button

Joan Fruitet joan.fruitet@inria.fr
Maureen Clerc maureen.clerc@inria.fr
Remi Munos remi.munos@inria.fr
INRIA
Alexandra Carpentier a.carpentier@statslab.cam.ac.uk
Cambridge University

A brain-computer interface (BCI) allows users to “communicate” with a computer without using their muscles. BCI based on sensori-motor rhythms use imaginary motor tasks, such as moving the right or left hand to send control signals. The performances of a BCI can vary greatly across users but also depend on the tasks used, making the problem of appropriate task selection an important issue. This study presents a new procedure to automatically select as fast as possible a discriminant motor task for a brain-controlled button. We develop for this purpose an adaptive algorithm UCB-classif based on the stochastic bandit theory. This shortens the training stage, thereby allowing the exploration of a greater variety of tasks. By not wasting time on inefficient tasks, and focusing on the most promising ones, this algorithm results in a faster task selection and a more efficient use of the BCI training session. Comparing the proposed method to the standard practice in task selection, for a fixed time budget, UCB-classif leads to an improve classification rate, and for a fix classification rate, to a reduction of the time spent in training by 50%.

Th2 Patient Risk Stratification for Hospital-Associated C. Diff as a Time-Series Classification Task

Jenna Wiens jwiens@mit.edu
John Guttag guttag@csail.mit.edu
Massachusetts Institute of Technology
Eric Horvitz horvitz@microsoft.com
Microsoft Research

A patient’s risk for adverse events is affected by temporal processes including the nature and timing of diagnostic and therapeutic activities, and the overall evolution of the patient’s pathophysiology over time. Yet many investigators ignore this temporal aspect when modeling patient risk, considering only the patient’s current or aggregate state. We explore representing patient risk as a time series. In doing so, patient risk stratification becomes a time-series classification task. The task differs from most applications of time-series analysis, like speech processing, since the time series itself must first be extracted. Thus, we begin by defining and extracting approximate \textit{risk processes}, the evolving approximate daily risk of a patient. Once obtained, we use these signals to explore different approaches to time-series classification with the goal of identifying high-risk patterns. We apply the classification to the specific task of identifying patients at risk of testing positive for hospital acquired colonization with \textit{Clostridium Difficile}. We achieve an area under the receiver operating characteristic curve of 0.79 on a held-out set of several hundred patients. Our two-stage approach to risk stratification outperforms classifiers that consider only a patient’s current state ($p<0.05$).

Th3 Learning Label Trees for Probabilistic Modelling of Implicit Feedback

Andriy Mnih amnih@gatsby.ucl.ac.uk
Gatsby Unit, UCL
Yee Whye Teh teh@stats.ox.ac.uk
University of Oxford

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

Th4 Collaborative Ranking With 17 Parameters

Maksims Volkovs mvolkovs@cs.toronto.edu
Richard Zemel zemel@cs.toronto.edu
University of Toronto

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

Th5 A dynamic excitatory-inhibitory network in a VLSI chip for spiking information reregistration

Juan Huo juanhuo@126.com

Inhibitory synapse is an important component both in physiology and artificial neural network, which has been widely investigated and used. A typical inhibitory synapse in very large scale integrated (VLSI) circuit is simplified from related research and applied in a VLSI chip for spike train reregistration. The spike train reregistration network is derived from a neural network model for sensory map realignment for network adaptation. In this paper, we introduce the design of spike train registration in CMOS circuit and analyze the performance of the inhibitory network in it, which shows representative characters for the firing rate of inhibited neuron and information transmission in circuit compared to math model.

Th6 GenDeR: A Generic Diversified Ranking Algorithm

Jingrui He jingrui.he@gmail.com
Hanghang Tong hanghang.tong@gmail.com
IBM Research
Qiaozhu Mei qmei@umich.edu
University of Michigan
Bolek Szymanski boleslaw.szymanski@gmail.com
RPI

Diversified ranking is a fundamental task in machine learning. It is broadly applicable in many real world problems, e.g., information retrieval, team assembling, product search, etc. In this paper, we consider a generic setting where we aim to diversify the top-k ranking list based on an arbitrary relevance function and an arbitrary similarity function among all the examples. We formulate it as an optimization problem and show that in general it is NP-hard. Then, we show that for a large volume of the parameter space, the proposed objective function enjoys the diminishing returns property, which enables us to design a scalable, greedy algorithm to find the near-optimal solution. Experimental results on real data sets demonstrate the effectiveness of the proposed algorithm.

Th7 Iterative ranking from pair-wise comparisons

Sahand Negahban sahand_n@eecs.berkeley.edu
University of California, Berkeley
Sewoong Oh swoh@illinois.edu
University of Illinois at Urbana Champaign
Devavrat Shah devavrat@mit.edu
Massachusetts Institute of Technology

The question of aggregating pairwise comparisons to obtain a global ranking over a collection of objects has been of interest for a very long time: be it ranking of online gamers (e.g. MSR's TrueSkill system) and chess players, aggregating social opinions, or deciding which product to sell based on transactions. In most settings, in addition to obtaining ranking, finding 'scores' for each object (e.g. player's rating) is of interest to understanding the intensity of the preferences. In this paper, we propose a novel iterative rank aggregation algorithm for discovering scores for objects from pairwise comparisons. The algorithm has a natural random walk interpretation over the graph of objects with edges present between two objects if they are compared; the scores turn out to be the stationary probability of this random walk. The algorithm is model independent. To establish the efficacy of our method, however, we consider the popular Bradley-Terry-Luce (BTL) model in which each object has an associated score which determines the probabilistic outcomes of pairwise comparisons between objects. We bound the finite sample error rates between the scores assumed by the BTL model and those estimated by our algorithm. This, in essence, leads to order-optimal dependence on the number of samples required to learn the scores well by our algorithm. Indeed, the experimental evaluation shows that our (model independent) algorithm performs as well as the Maximum Likelihood Estimator of the BTL model and outperforms a recently proposed algorithm by Ammar and Shah [1].

Th8 Accuracy at the Top

Stephen Boyd boyd@stanford.edu
Stanford University
Corinna Cortes corinna@google.com
Ana Radovanovic anaradovanovic@google.com
Google, Inc
Mehryar Mohri mohri@google.com
Courant Institute & Google Research

We introduce a new notion of classification accuracy based on the top τ-quantile values of a scoring function, a relevant criterion in a number of problems arising for search engines. We define an algorithm optimizing a convex surrogate of the corresponding loss, and show how its solution can be obtained by solving several convex optimization problems. We also present margin-based guarantees for this algorithm based on the τ-quantile of the functions in the hypothesis set. Finally, we report the results of several experiments evaluating the performance of our algorithm. In a comparison in a bipartite setting with several algorithms seeking high precision at the top, our algorithm achieves a better performance in precision at the top.

Th9 Learning to Discover Social Circles in Ego Networks

Julian McAuley julian.mcauley@gmail.com
Jure Leskovec jure@cs.stanford.edu
Stanford University

Our personal social networks are big and cluttered, and currently there is no good way to organize them. Social networking sites allow users to manually categorize their friends into social circles (e.g. `circles' on Google+, and `lists' on Facebook and Twitter), however they are laborious to construct and must be updated whenever a user's network grows. We define a novel machine learning task of identifying users' social circles. We pose the problem as a node clustering problem on a user's ego-network, a network of connections between her friends. We develop a model for detecting circles that combines network structure as well as user profile information. For each circle we learn its members and the circle-specific user profile similarity metric. Modeling node membership to multiple circles allows us to detect overlapping as well as hierarchically nested circles. Experiments show that our model accurately identifies circles on a diverse set of data from Facebook, Google+, and Twitter for all of which we obtain hand-labeled ground-truth data.

Th10 Cocktail Party Processing via Structured Prediction

Yuxuan Wang wangyuxu@cse.ohio-state.edu
DeLiang Wang dwang@cis.ohio-state.edu
Ohio State University

While human listeners excel at selectively attending to a conversation in a cocktail party, machine performance is still far inferior by comparison. We show that the cocktail party problem, or the speech separation problem, can be effectively approached via structured prediction. To account for temporal dynamics in speech, we employ conditional random fields (CRFs) to classify speech dominance within each time-frequency unit for a sound mixture. To capture complex, nonlinear relationship between input and output, both state and transition feature functions in CRFs are learned by deep neural networks. The formulation of the problem as classification allows us to directly optimize a measure that is well correlated with human speech intelligibility. The proposed system substantially outperforms existing ones in a variety of noises.

Th11 On the Use of Non-Stationary Policies for Stationary Infinite-Horizon Markov Decision Processes

Bruno Scherrer scherrer@loria.fr
Boris Lesner boris.lesner@inria.fr
INRIA

We consider infinite-horizon stationary γ-discounted Markov Decision Processes, for which it is known that there exists a stationary optimal policy. Using Value and Policy Iteration with some error ϵ at each iteration, it is well-known that one can compute stationary policies that are \frac{2\gamma{(1-\gamma)^2}\epsilon-optimal. After arguing that this guarantee is tight, we develop variations of Value and Policy Iteration for computing non-stationary policies that can be up to 2γ1−γϵ-optimal, which constitutes a significant improvement in the usual situation when γ is close to 1. Surprisingly, this shows that the problem of ``computing near-optimal non-stationary policies'' is much simpler than that of ``computing near-optimal stationary policies''.

Th12 Non-parametric Approximate Dynamic Programming via the Kernel Method

Nikhil Bhat nikhil.p.b@gmail.com
Ciamac Moallemi ciamac@gsb.columbia.edu
Columbia University
Vivek Farias vivekf@mit.edu
Massachusetts Institute of Technology

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

Th13 Imitation Learning by Coaching

He He hhe@cs.umd.edu
Hal Daume III me@hal3.name
University of Maryland
Jason Eisner jason@cs.jhu.edu
Johns Hopkins University

Imitation Learning has been shown to be successful in solving many challenging real-world problems. Some recent approaches give strong performance guarantees by training the policy iteratively. However, it is important to note that these guarantees depend on how well the policy we found can imitate the oracle on the training data. When there is a substantial difference between the oracle's ability and the learner's policy space, we may fail to find a policy that has low error on the training set. In such cases, we propose to use a coach that demonstrates easy-to-learn actions for the learner and gradually approaches the oracle. By a reduction of learning by demonstration to online learning, we prove that coaching can yield a lower regret bound than using the oracle. We apply our algorithm to a novel cost-sensitive dynamic feature selection problem, a hard decision problem that considers a user-specified accuracy-cost trade-off. Experimental results on UCI datasets show that our method outperforms state-of-the-art imitation learning methods in dynamic features selection and two static feature selection methods.

Th14 Inverse Reinforcement Learning through Structured Classification

Edouard Klein edouard.klein@supelec.fr
Matthieu Geist matthieu.geist@supelec.fr
BILAL PIOT bilal.piot@supelec.fr
Olivier Pietquin olivier.pietquin@supelec.fr
SUPELEC

This paper adresses the inverse reinforcement learning (IRL) problem, that is inferring a reward for which a demonstrated expert behavior is optimal. We introduce a new algorithm, SCIRL, whose principle is to use the so-called feature expectation of the expert as the parameterization of the score function of a multi-class classifier. This approach produces a reward function for which the expert policy is provably near-optimal. Contrary to most of existing IRL algorithms, SCIRL does not require solving the direct RL problem. Moreover, with an appropriate heuristic, it can succeed with only trajectories sampled according to the expert behavior. This is illustrated on a car driving simulator.

Th15 Efficient Bayes-Adaptive Reinforcement Learning using Sample-Based Search

Arthur Guez aguez@gatsby.ucl.ac.uk
David Silver davidstarsilver@gmail.com
Peter Dayan dayan@gatsby.ucl.ac.uk
University College of London

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

Th16 A Bayesian Approach for Policy Learning from Trajectory Preference Queries

Aaron Wilson wilsonaa@eecs.oregonstate.edu
Prasad Tadepalli tadepall@eecs.oregonstate.edu
Alan Fern afern@eecs.oregonstate.edu
Oregon State University

We consider the problem of learning control policies via trajectory preference queries to an expert. In particular, the learning agent can present an expert with short runs of a pair of policies originating from the same state and the expert then indicates the preferred trajectory. The agent's goal is to elicit a latent target policy from the expert with as few queries as possible. To tackle this problem we propose a novel Bayesian model of the querying process and introduce two methods that exploit this model to actively select expert queries. Experimental results on four benchmark problems indicate that our model can effectively learn policies from trajectory preference queries and that active query selection can be substantially more efficient than random selection.

Th17 Value Pursuit Iteration

Amir-massoud Farahmand amirf@ualberta.ca
Doina Precup dprecup@cs.mcgill.ca
McGill University

Value Pursuit Iteration (VPI) is an approximate value iteration algorithm that finds a close to optimal policy for reinforcement learning and planning problems with large state spaces. VPI has two main features: First, it is a nonparametric algorithm that finds a good sparse approximation of the optimal value function given a dictionary of features. The algorithm is almost insensitive to the number of irrelevant features. Second, after each iteration of VPI, the algorithm adds a set of functions based on the currently learned value function to the dictionary. This increases the representation power of the dictionary in a way that is directly relevant to the goal of having a good approximation of the optimal value function. We theoretically study VPI and provide a finite-sample error upper bound for it.

Th18 Online Regret Bounds for Undiscounted Continuous Reinforcement Learning

Ronald Ortner ronald.ortner@unileoben.ac.at
Montanuniversitaet Leoben
Daniil Ryabko daniil.ryabko@inria.fr
INRIA

We derive sublinear regret bounds for undiscounted reinforcement learning in continuous state space. The proposed algorithm combines state aggregation with the use of upper confidence bounds for implementing optimism in the face of uncertainty. Beside the existence of an optimal policy which satisfies the Poisson equation, the only assumptions made are Hoelder continuity of rewards and transition probabilities.

Th19 Robustness and risk-sensitivity in Markov decision processes

Takayuki Osogami osogami@jp.ibm.com
IBM Research - Tokyo

We uncover relations between robust MDPs and risk-sensitive MDPs. The objective of a robust MDP is to minimize a function, such as the expectation of cumulative cost, for the worst case when the parameters have uncertainties. The objective of a risk-sensitive MDP is to minimize a risk measure of the cumulative cost when the parameters are known. We show that a risk-sensitive MDP of minimizing the expected exponential utility is equivalent to a robust MDP of minimizing the worst-case expectation with a penalty for the deviation of the uncertain parameters from their nominal values, which is measured with the Kullback-Leibler divergence. We also show that a risk-sensitive MDP of minimizing an iterated risk measure that is composed of certain coherent risk measures is equivalent to a robust MDP of minimizing the worst-case expectation when the possible deviations of uncertain parameters from their nominal values are characterized with a concave function.

Th20 Trajectory-Based Short-Sighted Probabilistic Planning

Felipe Trevizan fwt@cs.cmu.edu
Manuela Veloso mmv@cs.cmu.edu
Carnegie Mellon University

Probabilistic planning captures the uncertainty of plan execution by probabilistically modeling the effects of actions in the environment, and therefore the probability of reaching different states from a given state and action. In order to compute a solution for a probabilistic planning problem, planners need to manage the uncertainty associated with the different paths from the initial state to a goal state. Several approaches to manage uncertainty were proposed, e.g., consider all paths at once, perform determinization of actions, and sampling. In this paper, we introduce trajectory-based short-sighted Stochastic Shortest Path Problems (SSPs), a novel approach to manage uncertainty for probabilistic planning problems in which states reachable with low probability are substituted by artificial goals that heuristically estimate their cost to reach a goal state. We also extend the theoretical results of Short-Sighted Probabilistic Planner (SSiPP) [ref] by proving that SSiPP always finishes and is asymptotically optimal under sufficient conditions on the structure of short-sighted SSPs. We empirically compare SSiPP using trajectory-based short-sighted SSPs with the winners of the previous probabilistic planning competitions and other state-of-the-art planners in the triangle tireworld problems. Trajectory-based SSiPP outperforms all the competitors and is the only planner able to scale up to problem number 60, a problem in which the optimal solution contains approximately 1070 states.

Th21 Cost-Sensitive Exploration in Bayesian Reinforcement Learning

Dongho Kim dk449@cam.ac.uk
University of Cambridge
Kee-Eung Kim kekim@cs.kaist.ac.kr
KAIST
Pascal Poupart ppoupart@cs.uwaterloo.ca
University of Waterloo

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

Th22 Recursive Deep Learning on 3D Point Clouds

Richard Socher richard@socher.org
Bharath Bath bbhat@stanford.edu
Brody Huval brody38h@gmail.com
Andrew Ng ang@cs.stanford.edu
Christopher Manning manning@stanford.edu
Stanford University

Recent advances in 3D sensing technologies make it possible to easily record color and depth images which together can improve object recognition. Most current methods rely on very well-designed features for this new 3D modality. We introduce a novel model based on sparse and recursive autoencoders (RAE) for learning both features and object categories from raw 3D point clouds as well as standard images. The model differs from previous RAE models in that it fixes the tree structures and includes short-circuit connections from all tree nodes to the final classifier. This allows the model to take into consideration both low-level features as well as global features of the object. Using our fully learned architecture, we achieve state of the art performance on a standard RGB-D object recognition dataset, rivaling random forest classifiers on hand-designed features such as SIFT and spin images. Our method is very fast and can classify 71 images in 1 second on a standard desktop machine in Matlab. This is possible because the method only requires 16 matrix multiplications to classify each image into one of 51 household objects.

Th23 Deep Spatio-Temporal Architectures and Learning for Protein Structure Prediction

Pietro Di Lena pdilena@uci.edu
Pierre Baldi pfbaldi@ics.uci.edu
Ken Nagata knagata@uci.edu
University of California, Irvine

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

Th24 Image Denoising and Inpainting with Deep Neural Networks

Junyuan Xie piiswrong@gmail.com
Linli Xu linli@cs.ualberta.ca
Enhong Chen cheneh@ustc.edu.cn
University of Science and Technology of China

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

Th25 ImageNet Classification with Deep Convolutional Neural Networks

Alex Krizhevsky kriz@cs.toronto.edu
Ilya Sutskever ilya@cs.utoronto.ca
Geoffrey Hinton hinton@cs.toronto.edu
University of Toronto

We trained a large, deep convolutional neural network to classify the 1.3 million high-resolution images in the LSVRC-2010 ImageNet training set into the 1000 different classes. On the test data, we achieved top-1 and top-5 error rates of 39.7\% and 18.9\% which is considerably better than the previous state-of-the-art results. The neural network, which has 60 million parameters and 500,000 neurons, consists of five convolutional layers, some of which are followed by max-pooling layers, and two globally connected layers with a final 1000-way softmax. To make training faster, we used non-saturating neurons and a very efficient GPU implementation of convolutional nets. To reduce overfitting in the globally connected layers we employed a new regularization method that proved to be very effective.

Th26 Exact and Stable Recovery of Sequences of Signals with Sparse Increments via Differential ℓ1-Minimization

Demba Ba demba@mit.edu
Behtash Babadi behtash@nmr.mgh.harvard.edu
Patrick Purdon patrickp@nmr.mgh.harvard.edu
Emery Brown enb@neurostat.mit.edu
MIT/Harvard

We consider the problem of recovering a sequence of vectors, (xk)k=0K, for which the increments xk−xk−1 are Sk-sparse (with Sk typically smaller than S1), based on linear measurements (yk=Akxk+ek)k=1K, where Ak and ek denote the measurement matrix and noise, respectively. Assuming each Ak obeys the restricted isometry property (RIP) of a certain order---depending only on Sk---we show that in the absence of noise a convex program, which minimizes the weighted sum of the ℓ1-norm of successive differences subject to the linear measurement constraints, recovers the sequence (xk)k=1K \emph{exactly}. This is an interesting result because this convex program is equivalent to a standard compressive sensing problem with a highly-structured aggregate measurement matrix which does not satisfy the RIP requirements in the standard sense, and yet we can achieve exact recovery. In the presence of bounded noise, we propose a quadratically-constrained convex program for recovery and derive bounds on the reconstruction error of the sequence. We supplement our theoretical analysis with simulations and an application to real video data. These further support the validity of the proposed approach for acquisition and recovery of signals with time-varying sparsity.

Th27 Compressive neural representation of sparse, high-dimensional probabilities

Xaq Pitkow xaq@neurotheory.columbia.edu
University of Rochester

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

Th28 Shifting Weights: Adapting Object Detectors from Image to Video

Kevin Tang kdtang@cs.stanford.edu
Fei Fei Li feifeili@cs.stanford.edu
Daphne Koller koller@cs.stanford.edu
Vignesh Ramanathan vigneshr@stanford.edu
Stanford University

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

Th29 Multi-Stage Multi-Task Feature Learning

Pinghua Gong pinghuag@gmail.com
Changshui Zhang zcs@mail.tsinghua.edu.cn
Tsinghua University
Jieping Ye jieping.ye@asu.edu
Arizona State University

Multi-task sparse feature learning aims to improve the generalization performance by exploiting the shared features among tasks. It has been successfully applied to many applications including computer vision and biomedical informatics. Most of the existing multi-task sparse feature learning algorithms are formulated as a convex sparse regularization problem, which is usually suboptimal, due to its looseness for approximating an ℓ0-type regularizer. In this paper, we propose a non-convex formulation for multi-task sparse feature learning based on a novel regularizer. To solve the non-convex optimization problem, we propose a Multi-Stage Multi-Task Feature Learning (MSMTFL) algorithm. Moreover, we present a detailed theoretical analysis showing that MSMTFL achieves a better parameter estimation error bound than the convex formulation. Empirical studies on both synthetic and real-world data sets demonstrate the effectiveness of MSMTFL in comparison with the state of the art multi-task sparse feature learning algorithms.

Th30 Factoring nonnegative matrices with linear programs

Benjamin Recht brecht@cs.wisc.edu
Christopher Re chrisre@cs.wisc.edu
Victor Bittorf bittorf@cs.wisc.edu
University of Wisconsin
Joel Tropp jtropp@cms.caltech.edu
California Institute of Technology

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

Th31 Proximal Newton-type Methods for Minimizing Convex Objective Functions in Composite Form

Jason Lee jdl17@stanford.edu
Yuekai Sun yuekai@stanford.edu
Michael Saunders saunders@stanford.edu
Stanford University

We consider minimizing convex objective functions in \ emph{composite form} \minimizex∈\Rnf(x):=g(x)+h(x), where g is convex and twice-continuously differentiable and h:\Rn→\R is a convex but not necessarily differentiable function whose proximal mapping can be evaluated efficiently. We derive a generalization of Newton-type methods to handle such convex but nonsmooth objective functions. Many problems of relevance in high-dimensional statistics, machine learning, and signal processing can be formulated in composite form. We prove such methods are globally convergent to a minimizer and achieve quadratic rates of convergence in the vicinity of a unique minimizer. We also demonstrate the performance of such methods using problems of relevance in machine learning and high-dimensional statistics.

Th32 Communication/Computation Tradeoffs in Consensus-Based Distributed Optimization

Konstantinos Tsianos konstantinos.tsianos@gmail.com
McGill University
Sean Lawlor slawlor@slawlor.com
Michael Rabbat rabbat@cae.wisc.edu
University of Wisconsin-Madison

We study the scalability of consensus-based distributed optimization algorithms by considering two questions: How many processors should we use for a given problem, and how often should they communicate when communication is not free? Central to our analysis is a problem-specific value r which quantifies the communication/computation tradeoff. We show that organizing the communication among nodes as a k-regular expander graph~\ cite{kRegExpanders} yields speedups, while when all pairs of nodes communicate (as in a complete graph), there is an optimal number of processors that depends on r. Surprisingly, a speedup can be obtained, in terms of the time to reach a fixed level of accuracy, by communicating less and less frequently as the computation progresses. Experiments on a real cluster solving metric learning and non-smooth convex minimization tasks demonstrate strong agreement between theory and practice.

Th33 Recovery of Sparse Probability Measures via Convex Programming

Mert Pilanci mert@eecs.berkeley.edu
Laurent El Ghaoui elghaoui@eecs.berkeley.edu
UC Berkeley
Venkat Chandrasekaran venkatc@caltech.edu
California Institute of Technology

We consider the problem of cardinality penalized optimization of a convex function over the probability simplex with additional convex constraints. It's well-known that the classical L1 regularizer fails to promote sparsity on the probability simplex since L1 norm on the probability simplex is trivially constant. We propose a direct relaxation of the minimum cardinality problem and show that it can be efficiently solved using convex programming. As a first application we consider recovering a sparse probability measure given moment constraints, in which our formulation becomes linear programming, hence can be solved very efficiently. A sufficient condition for exact recovery of the minimum cardinality solution is derived for arbitrary affine constraints. We then develop a penalized version for the noisy setting which can be solved using second order cone programs. The proposed method outperforms known heuristics based on L1 norm. As a second application we consider convex clustering using a sparse Gaussian mixture and compare our results with the well known soft k-means algorithm.

Th34 Newton-Like Methods for Sparse Inverse Covariance Estimation

Peder Olsen pederao@us.ibm.com
Steven Rennie rennie@eecg.utoronto.ca
IBM research
Figen Oztoprak figen@eecs.northwestern.edu
Jorge Nocedal nocedal@eecs.northwestern.edu
Northwestern University

We propose two classes of second-order optimization methods for solving the sparse inverse covariance estimation problem. The first approach, which we call the Newton-LASSO method, minimizes a piecewise quadratic model of the objective function at every iteration to generate a step. We employ the fast iterative shrinkage thresholding method (FISTA) to solve this subproblem. The second approach, which we call the Orthant-Based Newton method, is a two-phase algorithm that first identifies an orthant face and then minimizes a smooth quadratic approximation of the objective function using the conjugate gradient method. These methods exploit the structure of the Hessian to efficiently compute the search direction and to avoid explicitly storing the Hessian. We show that quasi-Newton methods are also effective in this context, and describe a limited memory BFGS variant of the orthant-based Newton method. We present numerical results that suggest that all the techniques described in this paper have attractive properties and constitute useful tools for solving the sparse inverse covariance estimation problem. Comparisons with the method implemented in the QUIC software package are presented.

Th35 A quasi-Newton proximal splitting method

Stephen Becker stephen.becker@upmc.fr
Paris-6/CNRS
Jalal Fadili Jalal.Fadili@greyc.ensicaen.fr
CNRS-ENSICAEN-Univ. Caen

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

Th36 Query Complexity of Derivative-Free Optimization

Kevin Jamieson kgjamieson@wisc.edu
Benjamin Recht brecht@cs.wisc.edu
Rob Nowak nowak@ece.wisc.edu
University of Wisconsin-Madison

Derivative Free Optimization (DFO) is attractive when the objective function's derivatives are not available and evaluations are costly. Moreover, if the function evaluations are noisy, then approximating gradients by finite differences is difficult. This paper gives quantitative lower bounds on the performance of DFO with noisy function evaluations, exposing a fundamental and unavoidable gap between optimization performance based on noisy evaluations versus noisy gradients. This challenges the conventional wisdom that the method of finite differences is comparable to a stochastic gradient. However, there are situations in which DFO is unavoidable, and for such situations we propose a new DFO algorithm that is proved to be near optimal for the class of strongly convex objective functions. A distinctive feature of the algorithm is that it only uses Boolean-valued function comparisons, rather than evaluations. This makes the algorithm useful in an even wider range of applications, including optimization based on paired comparisons from human subjects, for example. Remarkably, we show that regardless of whether DFO is based on noisy function evaluations or Boolean-valued function comparisons, the convergence rate is the same.

T37 CPRL -- An Extension of Compressive Sensing to the Phase Retrieval Problem

Henrik Ohlsson ohlsson@isy.liu.se
Linköping university
Allen Yang yang@eecs.berkeley.edu
Shankar Sastry sastry@eecs.berkeley.edu
Roy Dong roydong@eecs.berkeley.edu
University of California - Berkeley

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

Th38 Joint Modeling of a Matrix with Associated Text via Latent Binary Features

XianXing Zhang xianxing.zhang@duke.edu
Lawrence Carin lcarin@ee.duke.edu
Duke University

A new methodology is developed for joint analysis of a matrix and accompanying documents, with the documents associated with the matrix rows/columns. The documents are modeled with a focused topic model, inferring latent binary features (topics) for each document. A new matrix decomposition is developed, with latent binary features associated with the rows/columns, and with imposition of a low-rank constraint. The matrix decomposition and topic model are coupled by sharing the latent binary feature vectors associated with each. The model is applied to roll-call data, with the associated documents defined by the legislation. State-of-the-art results are manifested for prediction of votes on a new piece of legislation, based only on the observed text legislation. The coupling of the text and legislation is also demonstrated to yield insight into the properties of the matrix decomposition for roll-call data.

Th39 Probabilistic Low-Rank Subspace Clustering

S. Derin Babacan dbabacan@gmail.com
Minh Do minhdo@illinois.edu
University of Illinois
Shinichi Nakajima shinnkj23@gmail.com
Nikon Corporation

In this paper, we consider the problem of clustering data points into low-dimensional subspaces in the presence of outliers. We pose the problem using a density estimation formulation with an associated generative model. Based on this probability model, we first develop an iterative expectation-maximization (EM) algorithm and then derive its global solution. In addition, we develop two Bayesian methods based on variational Bayesian (VB) approximation, which are capable of automatic dimensionality selection. While the first method is based on an alternating optimization scheme for all unknowns, the second method makes use of recent results in VB matrix factorization leading to fast and effective estimation. Both methods are extended to handle sparse outliers for robustness and can handle missing values. Experimental results suggest that proposed methods are very effective in clustering and identifying outliers.

Th40 Bayesian n-Choose-k Models for Classification and Ranking

Kevin Swersky kswersky@cs.toronto.edu
Danny Tarlow dtarlow@cs.toronto.edu
Richard Zemel zemel@cs.toronto.edu
Brendan Frey frey@psi.toronto.edu
University of Toronto
Ryan Adams rpa@seas.harvard.edu
Harvard University

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

Th41 Bayesian Nonparametric Modeling of Suicide Attempts

Francisco J. R. Ruiz franrruiz@tsc.uc3m.es
Fernando Perez-Cruz fernandop@ieee.org
Isabel Valera ivalera@tsc.uc3m.es
University Carlos III at Madrid
Carlos Blanco Cblanco@nyspi.columbia.edu
Columbia University College of Physicians and Surgeons

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

Th42 Bayesian nonparametric models for bipartite graphs

Francois Caron Francois.Caron@inria.fr
INRIA Bordeaux

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

Th43 Coupling Nonparametric Mixtures via Latent Dirichlet Processes

Dahua Lin dhlin@mit.edu
John Fisher fisher@csail.mit.edu
Massachusetts Institute of Technology

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

Th44 Multiresolution Gaussian Processes

Emily Fox ebfox@uw.edu
University of Washington
David Dunson dunson@stat.duke.edu
Duke University

We propose a multiresolution Gaussian process to capture long-range, non-Markovian dependencies while allowing for abrupt changes. The multiresolution GP hierarchically couples a collection of smooth GPs, each defined over an element of a random nested partition. Long-range dependencies are captured by the top-level GP while the partition points define the abrupt changes. Due to the inherent conjugacy of the GPs, one can analytically marginalize the GPs and compute the conditional likelihood of the observations given the partition tree. This allows for efficient inference of the partition itself, for which we employ graph-theoretic techniques. We apply the multiresolution GP to the analysis of Magnetoencephalography (MEG) recordings of brain activity.

Th45 Bayesian Warped Gaussian Processes

Miguel Lázaro-Gredilla miguel@tsc.uc3m.es
Universidad Carlos III de Madrid

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

Th46 Collaborative Gaussian Processes for Preference Learning

Neil Houlsby nmT2@cam.ac.uk
Ferenc Huszar fh277@cam.ac.uk
Zoubin Ghahramani zoubin@eng.cam.ac.uk
Jose Miguel Hernández-Lobato jmh233@cam.ac.uk
Cambridge University

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

Th47 Nonparanormal Belief Propagation (NPBP)

Gal Elidan galel@huji.ac.il
Cobi Cario cobi.cario@mail.huji.ac.il
Hebrew University

The empirical success of the belief propagation approximate inference algorithm has inspired numerous theoretical and algorithmic advances. Yet, for continuous non-Gaussian domains performing belief propagation remains a challenging task: recent innovations such as nonparametric or kernel belief propagation, while useful, come with a substantial computational cost and offer little theoretical guarantees, even for tree structured models. In this work we present Nonparanormal BP for performing efficient inference on distributions parameterized by a Gaussian copulas network and any univariate marginals. For tree structured networks, our approach is guaranteed to be exact for this powerful class of non-Gaussian models. Importantly, the method is as efficient as standard Gaussian BP, and its convergence properties do not depend on the complexity of the univariate marginals, even when a nonparametric representation is used.

Th48 Latent Coincidence Analysis: A Hidden Variable Model for Distance Metric Learning

Matt Der mfder@cs.ucsd.edu
Lawrence Saul saul@cs.ucsd.edu
UC San Diego

We describe a latent variable model for supervised dimensionality reduction and distance metric learning. The model discovers linear projections of high dimensional data that shrink the distance between similarly labeled inputs and expand the distance between differently labeled ones. The model's continuous latent variables locate pairs of examples in a latent space of lower dimensionality. The model differs significantly from classical factor analysis in that the posterior distribution over these latent variables is not always multivariate Gaussian. Nevertheless we show that inference is completely tractable and derive an Expectation-Maximization (EM) algorithm for parameter estimation. We also compare the model to other approaches in distance metric learning. The model's main advantage is its simplicity: at each iteration of the EM algorithm, the distance metric is re-estimated by solving an unconstrained least-squares problem. Experiments show that these simple updates are highly effective.

Th49 Multiple Choice Learning: Learning to Produce Multiple Structured Outputs

Abner Guzmán-Rivera aguzman5@illinois.edu
University of Illinois
Dhruv Batra dbatra@ttic.edu
TTI-Chicago
Pushmeet Kohli pkohli@microsoft.com
Microsoft Research

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

Th50 Learning from the Wisdom of Crowds by Minimax Entropy

Dengyong Zhou dengyong.zhou@microsoft.com
Sumit Basu sumitb@microsoft.com
John Platt jplatt@microsoft.com
Yi Mao yimao@microsoft.com
Microsoft Research

We consider the multiclass crowd labeling issue. Each instance is labeled several times by different workers, while one instance might be labeled more times than another. We propose a minimax entropy principle to simultaneously estimate worker expertise, task ambiguities, and true labels. We also suggest an objectivity requirement for reasonably measuring worker expertise and task ambiguities, and show that the proposed method is unique in meeting the objectivity requirement. Experimental results are presented for both synthetic and real data.

Th51 Bayesian models for Large-scale Hierarchical Classification

Siddharth Gopal sgopal1@andrew.cmu.edu
Yiming Yang yiming@cs.cmu.edu
Carnegie Mellon University
Bing Bai bbai@nec-labs.com
Alexandru Niculescu-Mizil alexnic@gmail.com
NEC Laboratories America

A challenging problem in hierarchical classification is to leverage the hierarchical relations among classes for improving classification performance. An even greater challenge is to do so in a manner that is computationally feasible for the large scale problems usually encountered in practice. This paper proposes a set of Bayesian methods to model hierarchical dependencies among class labels using multivari- ate logistic regression. Specifically, the parent-child relationships are modeled by placing a hierarchical prior over the children nodes centered around the parameters of their parents; thereby encouraging classes nearby in the hierarchy to share similar model parameters. We present new, efficient variational algorithms for tractable posterior inference in these models, and provide a parallel implementa- tion that can comfortably handle large-scale problems with hundreds of thousands of dimensions and tens of thousands of classes. We run a comparative evaluation on multiple large-scale benchmark datasets that highlights the scalability of our approach, and shows a significant performance advantage over the other state-of- the-art hierarchical methods.

Th52 Multiple Operator-valued Kernel Learning

Hachem Kadri hachem.kadri@lif.univ-mrs.fr
Aix-Marseille University
Alain Rakotomamonjy alain.rakoto@insa-rouen.fr
University of Rouen
Francis Bach francis.bach@mines.org
philippe preux philippe.preux@univ-lille3.fr
INRIA - Ecole Normale Superieure

Positive definite operator-valued kernels generalize the well-known notion of reproducing kernels, and are naturally adapted to multi-output learning situations. This paper addresses the problem of learning a finite linear combination of infinite-dimensional operator-valued kernels which are suitable for extending functional data analysis methods to nonlinear contexts. We study this problem in the case of kernel ridge regression for functional responses with an lr-norm constraint on the combination coefficients. The resulting optimization problem is more involved than those of multiple scalar-valued kernel learning since operator-valued kernels pose more technical and theoretical issues. We propose a multiple operator-valued kernel learning algorithm based on solving a system of linear operator equations by using a block coordinate-descent procedure. We experimentally validate our approach on a functional regression task in the context of finger movement prediction in brain-computer interfaces.

Th53 Gradient-based kernel method for feature extraction and variable selection

Kenji Fukumizu fukumizu@ism.ac.jp
Institute of Statistical Mathematics
Chenlei Leng stalc@nus.edu.sg
National University of Singapore

We propose a novel kernel approach to dimension reduction for supervised learning: feature extraction and variable selection; the former constructs a small number of features from predictors, and the latter finds a subset of predictors. First, a method of linear feature extraction is proposed using the gradient of regression function, based on the recent development of the kernel method. In comparison with other existing methods, the proposed one has wide applicability without strong assumptions on the regressor or type of variables, and uses computationally simple eigendecomposition, thus applicable to large data sets. Second, in combination of a sparse penalty, the method is extended to variable selection, following the approach by Chen et al. (2010). Experimental results show that the proposed methods successfully find effective features and variables without parametric models.

Th54 Learning from Distributions via Support Measure Machines

Krikamol Muandet krikamol@gmail.com
Francesco Dinuzzo francesco.dinuzzo@gmail.com
Bernhard Schölkopf bs@tuebingen.mpg.de
Max Planck Institute for Intelligent Systems
Kenji Fukumizu fukumizu@ism.ac.jp
Institute of Statistical Mathematics

This paper presents a kernel-based discriminative learning framework on probability measures. Rather than relying on large collections of vectorial training examples, our framework learns using a collection of probability distributions that have been constructed to meaningfully represent training data. By representing these probability distributions as mean embeddings in the

reproducing kernel Hilbert space (RKHS), we are able to apply many standard kernel-based learning techniques in straightforward fashion. To accomplish this, we construct a generalization of the support vector machine (SVM) called a support measure machine (SMM). Our analyses of SMMs provides several insights into their relationship to traditional SVMs. Based on such insights, we propose a flexible SVM (Flex-SVM) that places different kernel functions on each training example. Experimental results on both synthetic and real-world data demonstrate the effectiveness of our proposed framework.

Th55 Nonparametric Reduced Rank Regression

Rina Foygel rinafb@stanford.edu
Stanford University
Michael Horrell horrell@galton.uchicago.edu
John Lafferty lafferty@gmail.com
University of Chicago
Mathias Drton md5@uw.edu
University of Washington

We propose an approach to multivariate nonparametric regression that generalizes reduced rank regression for linear models. An additive model is estimated for each dimension of a q-dimensional response, with a shared p-dimensional predictor variable. To control the complexity of the model, we employ a functional form of the Ky-Fan or nuclear norm, resulting in a set of function estimates that have low rank. Backfitting algorithms are derived and justified using a nonparametric form of the nuclear norm subdifferential. Oracle inequalities on excess risk are derived that exhibit the scaling behavior of the procedure in the high dimensional setting. The methods are illustrated on gene expression data.

Th56 Pointwise Tracking the Optimal Regression Function

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

This paper examines the possibility of a `reject option' in the context of least squares regression. It is shown that using rejection it is theoretically possible to learn `selective' regressors that can ϵ-pointwise track the best regressor in hindsight from the same hypothesis class, while rejecting only a bounded portion of the domain. Moreover, the rejected volume vanishes with the training set size, under certain conditions. We then develop efficient and exact implementation of these selective regressors for the case of linear regression. Empirical evaluation over a suite of real-world datasets corroborates the theoretical analysis and indicates that our selective regressors can provide substantial advantage by reducing estimation error.

Th57 Link Prediction in Graphs with Autoregressive Features

Emile Richard r.emile.richard@gmail.com
Nicolas Vayatis nicolas.vayatis@cmla.ens-cachan.fr
ENS Cachan
Stephane Gaiffas stephane.gaiffas@cmap.polytechnique.fr
Ecole Polytechnique and University Paris 6

In the paper, we consider the problem of link prediction in time-evolving graphs. We assume that certain graph features, such as the node degree, follow a vector autoregressive (VAR) model and we propose to use this information to improve the accuracy of prediction. Our strategy involves a joint optimization procedure over the space of adjacency matrices and VAR matrices which takes into account both sparsity and low rank properties of the matrices. Oracle inequalities are derived and illustrate the trade-offs in the choice of smoothing parameters when modeling the joint effect of sparsity and low rank property. The estimate is computed efficiently using proximal methods through a generalized forward-backward agorithm.

Th58 Gradient Weights help Nonparametric Regressors

Samory Kpotufe samory@ttic.edu
Toyota Technological Institute
Abdeslam Boularias boularias@tuebingen.mpg.de
Max Planck Institute for Intelligent Systems

In regression problems over \reald, the unknown function f often varies more in some coordinates than in others. We show that weighting each coordinate i with the estimated norm of the ith derivative of f is an efficient way to significantly improve the performance of distance-based regressors, e.g. kernel and k-NN regressors. We propose a simple estimator of these derivative norms and prove its consistency. Moreover, the proposed estimator is efficiently learned online.

Th59 Selecting Diverse Features via Spectral Regularization

Abhimanyu Das abhidas@yahoo-inc.com
Anirban Dasgupta anirban.dasgupta@gmail.com
Ravi Kumar ravikumar@yahoo-inc.com
Yahoo!

We study the problem of diverse feature selection in linear regression: selecting a small subset of diverse features that can predict a given objective. Diversity is useful for several reasons such as interpretability, robustness to noise, etc. We propose several spectral regularizers that capture a notion of diversity of features and show that these are all submodular set functions. These regularizers, when added to the objective function for linear regression, result in approximately submodular functions, which can then be maximized approximately by efficient greedy and local search algorithms, with provable guarantees. We compare our algorithms to traditional greedy and $\ell 1$-regularization schemes and show that we obtain a more diverse set of features that result in the regression problem being stable under perturbations.

Th60 Sparse Prediction with the k-Support Norm

Andreas Argyriou argyriou@ttic.edu
Ecole Centrale de Paris
Rina Foygel rinafb@stanford.edu
Stanford University
Nati Srebro nati@ttic.edu
TTI-Chicago

We derive a novel norm that corresponds to the tightest convex relaxation of sparsity combined with an ℓ2 penalty. We show that this new norm provides a tighter relaxation than the elastic net, and is thus a good replacement for the Lasso or the elastic net in sparse prediction problems. But through studying our new norm, we also bound the looseness of the elastic net, thus shedding new light on it and providing justification for its use.

Th61 Fused sparsity and robust estimation for linear models with unknown variance

Arnak Dalalyan dalalyan@imagine.enpc.fr
ENSAE - CREST
Yin Chen ychen.cy@gmail.com
Ecole des Ponts ParisTech

In this paper, we develop a novel approach to the problem of learning sparse representations in the context of fused sparsity and unknown noise level. We propose an algorithm, termed Scaled Fused Dantzig Selector (SFDS), that accomplishes the aforementioned learning task by means of a second-order cone program. A special emphasize is put on the particular instance of fused sparsity corresponding to the learning in presence of outliers. We establish finite sample risk bounds and carry out an experimental evaluation on both synthetic and real data.

Th62 Dual-Space Analysis of the Sparse Linear Model

David Wipf davidwipf@gmail.com
Microsoft Research Asia

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

Th63 Entropy Estimations Using Correlated Symmetric Stable Random Projections

Ping Li pingli@cornell.edu
Cornell
Cun-Hui Zhang czhang@stat.rutgers.edu
Rutgers University

Methods for efficiently estimating the Shannon entropy of data streams have important applications in learning, data mining, and network anomaly detections (e.g., the DDoS attacks). For nonnegative data streams, the method of Compressed Counting (CC) based on maximally-skewed stable random projections can provide accurate estimates of the Shannon entropy using small storage. However, CC is no longer applicable when entries of data streams can be below zero, which is a common scenario when comparing two streams. In this paper, we propose an algorithm for entropy estimation in general data streams which allow negative entries. In our method, the Shannon entropy is approximated by the finite difference of two correlated frequency moments estimated from correlated samples of symmetric stable random variables. Our experiments confirm that this method is able to substantially better approximate the Shannon entropy compared to the prior state-of-the-art.

Th64 Reducing statistical time-series problems to binary classification

Daniil Ryabko daniil.ryabko@inria.fr
Jeremie Mary jeremie.mary@inria.fr
INRIA / Univ. Lille

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

Th65 On Multilabel Classification and Ranking with Partial Feedback

Claudio Gentile claudio.gentile@uninsubria.it
Universita' dell'Insubria
Francesco Orabona francesco@orabona.com
Toyota Technological Institute at Chicago

We present a novel multilabel/ranking algorithm working in partial information settings. The algorithm is based on 2nd-order descent methods, and relies on upper-confidence bounds to trade-off exploration and exploitation. We analyze this algorithm in a partial adversarial setting, where covariates can be adversarial, but multilabel probabilities are ruled by (generalized) linear models. We show O(T1/2logT) regret bounds, which improve in several ways on the existing results. We test the effectiveness of our upper-confidence scheme by contrasting against full-information baselines on real-world multilabel datasets, often obtaining comparable performance.

Th66 Nystr{ö}m Method vs Random Fourier Features: A Theoretical and Empirical Comparison

Tianbao Yang yangtia1@msu.edu
GE Global Research
Yu-Feng Li liyf@lamda.nju.edu.cn
Zhi-hua Zhou zhouzh@nju.edu.cn
Nanjing University
Mehrdad Mahdavi mahdavim@msu.edu
Rong Jin rong+@cs.cmu.edu
Michigan State University

Both random Fourier features and the Nystr{ö}m method have been successfully applied to efficient kernel learning. In this work, we investigate the fundamental difference between these two approaches, and how the difference could affect their generalization performances. Unlike approaches based on random Fourier features where the basis functions (i.e., cosine and sine functions) are sampled from a distribution {\it independent} from the training data, basis functions used by the Nystr{ö}m method are randomly sampled from the training examples and are therefore {\it data dependent}. By exploring this difference, we show that when there is a large gap in the eigen-spectrum of the kernel matrix, approaches based the Nystr{ö}m method can yield impressively better generalization error bound than random Fourier features based approach. We empirically verify our theoretical findings on a wide range of large data sets.

Th67 Learning Manifolds with K-Means and K-Flats

Guille Canas guilledc@MIT.EDU
Tomaso Poggio tp@ai.mit.edu
Lorenzo Rosasco lrosasco@mit.edu
MIT and Italian Institute of Technology

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

Th68 Selective Labeling via Error Bound Minimization

Quanquan Gu qgu3@illinois.edu
University of Illinois at Urbana-Champaign
Tong Zhang tzhang@stat.rutgers.edu
Rutgers University
Chris Ding chqding@uta.edu
University of Texas at Arlington
Jiawei Han hanj@cs.uiuc.edu
UIUC

In many practical machine learning problems, the acquisition of labeled data is often expensive and/or time consuming. This motivates us to study a problem as follows: given a label budget, how to select data points to label such that the learning performance is optimized. We propose a selective labeling method by analyzing the generalization error of Laplacian regularized Least Squares (LapRLS). In particular, we derive a deterministic generalization error bound for LapRLS trained on subsampled data, and propose to select a subset of data points to label by minimizing this upper bound. Since the minimization is a combinational problem, we relax it into continuous domain and solve it by projected gradient descent. Experiments on benchmark datasets show that the proposed method outperforms the state-of-the-art methods.

Th69 Semi-Crowdsourced Clustering: Generalizing Crowd Labeling by Robust Distance Metric Learning

Jinfeng Yi jinfengyi.ustc@gmail.com
Anil Jain jain@cse.msu.edu
Rong Jin rong+@cs.cmu.edu
Michigan State University
Shaili Jain shailij@gmail.com
Yale University

One of the main challenges in data clustering is to define an appropriate similarity measure between two objects. Crowdclustering addresses this challenge by defining the pairwise similarity based on the manual annotations obtained through crowdsourcing. Despite its encouraging results, a key limitation of crowdclustering is that it can only cluster objects when their manual annotations are available. To address this limitation, we propose a new approach for clustering, called \textit{semi-crowdsourced clustering} that effectively combines the low-level features of objects with the manual annotations of a subset of the objects obtained via crowdsourcing. The key idea is to learn an appropriate similarity measure, based on the low-level features of objects, from the manual annotations of only a small portion of the data to be clustered. One

difficulty in learning the pairwise similarity measure is that there is a significant amount of noise and inter-worker variations in the manual annotations obtained via crowdsourcing. We address this difficulty by developing a metric learning algorithm based on the matrix completion method. Our empirical study with two real-world image data sets shows that the proposed algorithm outperforms state-of-the-art distance metric learning algorithms in both clustering accuracy and computational efficiency.

Th70 Forging The Graphs: A Low Rank and Positive Semidefinite Graph Learning Approach

Dijun Luo dijun.luo@gmail.com
WhaleShark Media
Chris Ding chqding@uta.edu
Heng Huang heng@uta.edu
University of Texas Arlington

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

Th71 Hamming Distance Metric Learning

Mohammad Norouzi mohammad.n@gmail.com
Russ Salakhutdinov rsalakhu@mit.edu
David Fleet fleet@cs.toronto.edu
University of Toronto

Motivated by large-scale multimedia applications we propose to learn mappings from high-dimensional data to binary codes that preserve semantic similarity. Binary codes are well suited to large-scale applications as they are storage efficient and permit exact sub-linear kNN search. The framework is applicable to broad families of mappings, and uses a flexible form of triplet ranking loss. We overcome discontinuous optimization of the discrete mappings by minimizing a piecewise-smooth upper bound on empirical loss, inspired by latent structural SVMs. We develop a new loss-augmented inference algorithm that is quadratic in the code length. We show strong retrieval performance on CIFAR-10 and MNIST, with promising classification results using no more than kNN on the binary codes.

Th72 Parametric Local Metric Learning for Nearest Neighbor Classification

Jun Wang Jun.Wang@unige.ch
Adam Woznica Adam.Woznica@unige.ch
University of Geneva
Alexandros Kalousis Alexandros.Kalousis@unige.ch
University of Applied Sciences, Western Switzerland

We study the problem of learning local metrics for nearest neighbor classification. Most previous works on local metric learning learn a number of local unrelated metrics. While this ''independence'' approach delivers an increased flexibility its downside is the considerable risk of overfitting. We present a new parametric local metric learning method in which we learn a smooth metric matrix function over the data manifold. Using an approximation error bound of the metric matrix function we learn local metrics as linear combinations of basis metrics defined on anchor points over different regions of the instance space. We constrain the metric matrix function by imposing on the linear combinations manifold regularization which makes the learned metric matrix function vary smoothly along the geodesics of the data manifold. Our metric learning method has excellent performance both in terms of predictive power and scalability. We experimented with several large-scale classification problems, tens of thousands of instances, and compared it with several state of the art metric learning methods, both global and local, as well as to SVM with automatic kernel selection, all of which it outperforms in a significant manner.

Th73 Non-linear Metric Learning

Dor Kedem kedem.dor@wustl.edu
Stephen Tyree swtyree@wustl.edu
Kilian Weinberger kilian@wustl.edu
Washington University

Fei Sha feisha@usc.edu
University of Southern California
Gert Lanckriet gert@ece.ucsd.edu
U.C. San Diego

In this paper, we introduce two novel metric learning algorithms, χ2-LMNN and GB-LMNN, which are explicitly designed to be non-linear and easy-to-use. The two approaches achieve this goal in fundamentally different ways: χ2-LMNN inherits the computational benefits of a linear mapping from linear metric learning, but uses a non-linear χ2-distance to explicitly capture similarities within histogram data sets; GB-LMNN applies gradient-boosting to learn non-linear mappings directly in function space and takes advantage of this approach's robustness, speed, parallelizability and insensitivity towards the single additional hyper-parameter. On various benchmark data sets, we demonstrate these methods not only match the current state-of-the-art in terms of kNN classification error, but in the case of χ2-LMNN, obtain best results in 19 out of 20 learning settings.

Th74 Monte Carlo Methods for Maximum Margin Supervised Topic Models

Qixia Jiang qixia.jiang@gmail.com
Jun Zhu jjzhunet9@hotmail.com
Maosong Sun sms@tsinghua.edu.cn
Tsinghua University
Eric Xing epxing@cs.cmu.edu
Carnegie Mellon University

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

Th75 Topic-Partitioned Multinetwork Embeddings

Peter Krafft pkrafft@mit.edu
Massachusetts Institute of Technology
Juston Moore jmoore@cs.umass.edu
Hanna Wallach wallach@cs.umass.edu
Bruce Desmarais desmarais@polsci.umass.edu
University of Massachusetts Amherst

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

Th76 Learning with Recursive Perceptual Representations

Oriol Vinyals vinyals@eecs.berkeley.edu
Yangqing Jia jiayq@eecs.berkeley.edu
Trevor Darrell trevor@eecs.berkeley.edu
UC Berkeley
Li Deng deng@microsoft.com
Microsoft Research

Linear Support Vector Machines (SVMs) have become very popular in vision as part of state-of-the-art object recognition and other classification tasks but require high dimensional feature spaces for good performance. Deep learning methods can find more compact representations but current methods employ multilayer perceptrons that require solving a difficult, non-convex optimization problem. We propose a deep non-linear classifier whose layers are SVMs and which incorporates random projection as its core stacking element. Our method learns layers of linear SVMs recursively transforming the original data manifold through a random projection of the weak prediction computed from each layer. Our method scales as linear SVMs, does not rely on any kernel computations or nonconvex optimization, and exhibits better generalization ability than kernel-based SVMs. This is especially true when the number of training samples is smaller than the dimensionality of data, a common scenario in many real-world applications. The use of random projections is key to our method, as we show in the experiments section, in which we observe a consistent improvement over previous --often more complicated-- methods on several vision and speech benchmarks.

Th77 Natural Images, Gaussian Mixtures and Dead Leaves

Daniel Zoran daniez@cs.huji.ac.il
Yair Weiss yweiss@cs.huji.ac.il
Hebrew University

Simple Gaussian Mixture Models (GMMs) learned from pixels of natural image patches have been recently shown to be surprisingly strong performers in modeling the statistics of natural images. Here we provide an in depth analysis of this simple yet rich model. We show that such a GMM model is able to compete with even the most successful models of natural images in log likelihood scores, denoising performance and sample quality. We provide an analysis of what such a model learns from natural images as a function of number of mixture components --- including covariance structure, contrast variation and intricate structures such as textures, boundaries and more. Finally, we show that the salient properties of the GMM learned from natural images can be derived from a simplified Dead Leaves model which explicitly models occlusion, explaining its surprising success relative to other models.

Th78 Deep Learning of invariant features via tracked video sequences

Will Zou wzou@stanford.edu
Andrew Ng ang@cs.stanford.edu
Stanford University
Shenghuo Zhu zsh@nec-labs.com
NEC Laboratories America
Kai Yu kai.yu.cool@gmail.com
Baidu

We use video sequences produced by tracking as training data to learn invariant features. These features are spatial instead of temporal, and well suited to extract from still images. With a temporal coherence objective, a multi-layer neural network encodes invariance that grow increasingly complex with layer hierarchy. Without fine-tuning with labels, we achieve competitive performance on five non-temporal image datasets and state-of-the-art classification accuracy 61% on STL-10 object recognition dataset.

Th79 Dynamical And-Or Graph Learning for Object Shape Modeling and Detection

xiaolong wang dragonwxl123@gmail.com
Sun Yat-Sen University
Liang Lin liang@stat.ucla.edu
University of California, Los Angeles

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

Th80 Searching for objects driven by context

Bogdan Alexe bogdan@vision.ee.ethz.ch
ETH ZURICH
Nicolas Heess nheess@gatsby.ucl.ac.uk
University College London
Yee Whye Teh teh@stats.ox.ac.uk
University of Oxford
Vittorio Ferrari vittoferrari@gmail.com
University of Edinburgh

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

Th81 Learning Image Descriptors with the Boosting-Trick

Tomasz Trzcinski tomasz.trzcinski@epfl.ch
Mario Christoudias mario.christoudias@epfl.ch
Vincent Lepetit vincent.lepetit@epfl.ch
Pascal Fua josiane.gisclon@epfl.ch
Ecole Polytechnique Federal de Lausanne

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

Th82 Exploration in Model-based Reinforcement Learning by Empirically Estimating Learning Progress

Manuel Lopes manuel.lopes@inria.fr
Pierre-Yves Oudeyer pierre-yves.oudeyer@inria.fr
INRIA
Tobias Lang tobias.lang@fu-berlin.de
Marc Toussaint mtoussai@cs.tu-berlin.de
TU Berlin

Formal exploration approaches in model-based reinforcement learning estimate the accuracy of the currently learned model without consideration of the empirical prediction error. For example, PAC-MDP approaches such as Rmax base their model certainty on the amount of collected data, while Bayesian approaches assume a prior over the transition dynamics. We propose extensions to such approaches which drive exploration solely based on empirical estimates of the learner's accuracy and learning progress. We provide a ``sanity check'' theoretical analysis, discussing the behavior of our extensions in the standard stationary finite state-action case. We then provide experimental studies demonstrating the robustness of these exploration measures in cases of non-stationary environments or where original approaches are misled by wrong domain assumptions.

Th83 Learning optimal spike-based representations

Ralph Bourdoukan ralph.bourdoukan@gmail.com
David Barrett david.barrett@ens.fr
Sophie Deneve sophie.deneve@ens.fr
École Normale Supérieure
Christian Machens
christian.machens@neuro.fchampalimaud.org
Champalimaud Institute

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

Th84 Identification of Recurrent Patterns in the Activation of Brain Networks

firdaus janoos firdaus.janoos@exxonmobil.com
Weichang Li lwc@alum.mit.edu
Niranjan Subrahmanya
niranjan.a.subrahmanya@exxonmobil.com
ExxonMobil Corporate Research
Istvan Morocz pisti@bwh.harvard.edu
William Wells sw@bwh.harvard.edu
Harvard Medical School

Identifying patterns from the neuroimaging recordings of brain activity related to the unobservable psychological or mental state of an individual can be treated as a unsupervised pattern recognition problem. The main challenges, however, for such an analysis of fMRI data are: a) defining a physiologically meaningful feature-space for representing the spatial patterns across time; b) dealing with the high-dimensionality of the data; and c) robustness to the various artifacts and confounds in the fMRI time-series. In this paper, we present a network-aware feature-space to represent the states of a general network, that enables comparing and clustering such states in a manner that is a) meaningful in terms of the network connectivity structure; b)computationally efficient; c) low-dimensional; and d) relatively robust to structured and random noise artifacts. This feature-space is obtained from a spherical relaxation of the transportation distance metric which measures the cost of transporting ``mass'' over the network to transform one function into another. Through theoretical and empirical assessments, we demonstrate the accuracy and efficiency of the approximation, especially for large problems. While the application presented here is for identifying distinct brain activity patterns from fMRI, this feature-space can be applied to the problem of identifying recurring patterns and detecting outliers in measurements on many different types of networks, including sensor, control and social networks.

Th85 Efficient and direct estimation of a neural subunit model for sensory coding

Brett Vintch brett.vintch@gmail.com
Andrew Zaharia zaharia@cns.nyu.edu
J Movshon movshon@nyu.edu
Eero Simoncelli eero.simoncelli@nyu.edu
HHMI / New York University

Many visual and auditory neurons have response properties that are well explained by pooling the rectified responses of a set of self-similar linear filters. These filters cannot be found using spike-triggered averaging (STA), which estimates only a single filter. Other methods, like spike-triggered covariance (STC), define a multi-dimensional response subspace, but require substantial amounts of data and do not produce unique estimates of the linear filters. Rather, they provide a linear basis for the subspace in which the filters reside. Here, we define a `subunit' model as an LN-LN cascade, in which the first linear stage is restricted to a set of shifted (``convolutional'') copies of a common filter, and the first nonlinear stage consists of rectifying nonlinearities that are identical for all filter outputs; we refer to these initial LN elements as the `subunits' of the receptive field. The

second linear stage then computes a weighted sum of the responses of the rectified subunits. We present a method for directly fitting this model to spike data. The method performs well for both simulated and real data (from primate V1), and the resulting model outperforms STA and STC in terms of both cross-validated accuracy and efficiency.

Th86 How Prior Probability Influences Decision Making: A Unifying Probabilistic Model

Yanping Huang huangyp@cs.washington.edu
Abe Friesen afriesen@cs.washington.edu
Michael Shadlen shadlen@u.washington.edu
Rajesh Rao rao@cs.washington.edu
University of Washington
Timothy Hanks thanks@princeton.edu
Princeton University

How does the brain combine prior knowledge with sensory evidence when making decisions under uncertainty? Two competing descriptive models have been proposed based on experimental data. The first posits an additive offset to a decision variable, implying a static effect of the prior. However, this model is inconsistent with recent data from a motion discrimination task involving temporal integration of uncertain sensory evidence. To explain this data, a second model has been proposed which assumes a time-varying influence of the prior. Here we present a normative model of decision making that incorporates prior knowledge in a principled way. We show that the additive offset model and the time-varying prior model emerge naturally when decision making is viewed within the framework of partially observable Markov decision processes (POMDPs). Decision making in the model reduces to (1) computing beliefs given observations and prior information in a Bayesian manner, and (2) selecting actions based on these beliefs to maximize the expected sum of future rewards. We show that the model can explain both data previously explained using the additive offset model as well as more recent data on the time-varying influence of prior knowledge on decision making.

Th87 Efficient Spike-Coding with Multiplicative Adaptation in a Spike Response Model

Sander Bohte nips@bohte.com
Centrum Wiskunde Informatica

Neural adaptation underlies the ability of neurons to maximize encoded information over a wide dynamic range of input stimuli. While adaptation is an intrinsic feature of neuronal models like the Hodgkin-Huxley model, the challenge is to integrate adaptation in models of neural computation. Recent computational models like the Adaptive Spike Response Model implement adaptation as spike-based addition of fixed-size fast spike-triggered threshold dynamics and slow spike-triggered currents. Such adaptation has been shown to accurately model neural spiking behavior over a limited dynamic range. Taking a cue from kinetic models of adaptation, we propose a multiplicative Adaptive Spike Response Model where the spike-triggered adaptation dynamics are scaled multiplicatively by the adaptation state at the time of spiking. We show that unlike the additive adaptation model, the firing rate in the multiplicative adaptation model saturates to a maximum spike-rate. When simulating variance switching experiments, the model also quantitatively fits the experimental data over a wide dynamic range. Furthermore, dynamic threshold models of adaptation suggest a straightforward interpretation of neural activity in terms of dynamic signal encoding with shifted and weighted exponential kernels. We show that when thus encoding rectified filtered stimulus signals, the multiplicative Adaptive Spike Response Model achieves a high coding efficiency and maintains this efficiency over changes in the dynamic signal range of several orders of magnitude, without changing model parameters.

Th88 The topographic unsupervised learning of natural sounds in the auditory cortex

Hiroki Terashima teratti@teratti.jp
Masato Okada okada@k.u-tokyo.ac.jp
The University of Tokyo

The computational modelling of the primary auditory cortex (A1) has been less fruitful than that of the primary visual cortex (V1) due to the less organized properties of A1. Greater disorder has recently been demonstrated for the tonotopy of A1 that has traditionally been considered to be as ordered as the retinotopy of V1. This disorder appears to be incongruous, given the uniformity of the neocortex; however, we hypothesized that both A1 and V1 would adopt an efficient coding strategy and that the disorder in A1 reflects natural sound statistics. To provide a computational model of the tonotopic disorder in A1, we used a model that was originally proposed for the smooth V1 map. In contrast to natural images, natural sounds exhibit distant correlations, which were learned and reflected in the disordered map. The auditory model predicted harmonic relationships among neighbouring A1 cells; furthermore, the same mechanism used to model V1 complex cells reproduced nonlinear responses similar to the pitch selectivity. These results contribute to the understanding of the sensory cortices of different modalities in a novel and integrated manner.

Th89 Strategic Impatience in Go/NoGo versus Forced-Choice Decision-Making

Pradeep Shenoy pshenoy@ucsd.edu
Angela Yu ajyu@ucsd.edu
University of California, San Diego

Two-alternative forced choice (2AFC) and Go/NoGo (GNG) tasks are behavioral choice paradigms commonly used to study sensory and cognitive processing in choice behavior. While GNG is thought to isolate the sensory/decisional component by removing the need for response selection, a consistent bias towards the Go response (higher hits and false alarm rates) in the GNG task suggests possible fundamental differences in the sensory or cognitive processes engaged in the two tasks. Existing mechanistic models of these choice tasks, mostly variants of the drift-diffusion model (DDM; [1,2]) and the related leaky competing accumulator models [3,4] capture various aspects of behavior but do not address the provenance of

the Go bias. We postulate that this ``impatience'' to go is a strategic adjustment in response to the implicit asymmetry in the cost structure of GNG: the NoGo response requires waiting until the response deadline, while a Go response immediately terminates the current trial. We show that a Bayes-risk minimizing decision policy that minimizes both error rate and average decision delay naturally exhibits the experimentally observed bias. The optimal decision policy is formally equivalent to a DDM with a time-varying threshold that initially rises after stimulus onset, and collapses again near the response deadline. The initial rise is due to the fading temporal advantage of choosing the Go response over the fixed-delay NoGo response. We show that fitting a simpler, fixed-threshold DDM to the optimal model reproduces the counterintuitive result of a higher threshold in GNG than 2AFC decision-making, previously observed in direct DDM fit to behavioral data [2], although such approximations cannot reproduce the Go bias. Thus, observed discrepancies between GNG and 2AFC decision-making may arise from rational strategic adjustments to the cost structure, and need not imply additional differences in the underlying sensory and cognitive processes.

Th90 Delay Compensation with Dynamical Synapses

C. C. Alan Fung alanfung@ust.hk
K. Y. Michael Wong phkywong@ust.hk
Hong Kong University of Science and Technology
Si Wu wusi@bnu.edu.cn
Beijing Normal University

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

Th91 Neuronal spike generation mechanism as an oversampling, noise-shaping A-to-D converter

Dmitri Chklovskii chklovskiid@janelia.hhmi.org
HHMI

We explore the hypothesis that the neuronal spike generation mechanism is an analog-to-digital converter, which rectifies low-pass filtered summed synaptic currents and encodes them into spike trains linearly decodable in post-synaptic neurons. To digitally encode an analog current waveform, the sampling rate of the spike generation mechanism must exceed its Nyquist rate. Such oversampling is consistent with the experimental observation that the precision of the spike-generation mechanism is an order of magnitude greater than the cut-off frequency of dendritic low-pass filtering. To achieve additional reduction in the error of analog-to-digital conversion, electrical engineers rely on noise-shaping. If noise-shaping were used in neurons, it would introduce correlations in spike timing to reduce low-frequency (up to Nyquist) transmission error at the cost of high-frequency one (from Nyquist to sampling rate). Using experimental data from three different classes of neurons, we demonstrate that biological neurons utilize noise-shaping. We also argue that rectification by the spike-generation mechanism may improve energy efficiency and carry out de-noising. Finally, the zoo of ion channels in neurons may be viewed as a set of predictors, various subsets of which are activated depending on the statistics of the input current.

Th92 Complex Inference in Neural Circuits with Probabilistic Population Codes and Topic Models

Jeff Beck jeffbeck@gatsby.ucl.ac.uk
Alexandre Pouget alex@bcs.rochester.edu
University of Rochester
Katherine Heller kheller@gmail.com
Duke University

Recent experiments have demonstrated that humans and animals typically reason probabilistically about their environment. This ability requires a neural code that represents probability distributions and neural circuits that are capable of implementing the operations of probabilistic inference. The proposed probabilistic population coding (PPC) framework provides a statistically efficient neural representation of probability distributions that is both broadly consistent with physiological measurements and capable of implementing some of the basic operations of probabilistic inference in a biologically plausible way. However, these experiments and the corresponding neural models have largely focused on simple (tractable) probabilistic computations such as cue combination, coordinate transformations, and decision making. As a result it remains unclear how to generalize this framework to more complex probabilistic computations. Here we address this short coming by showing that a very general approximate inference algorithm known as Variational Bayesian Expectation Maximization can be implemented within the linear PPC framework. We apply this approach to a generic problem faced by any given layer of cortex, namely the identification of latent causes of complex mixtures of spikes. We identify a formal equivalent between this spike pattern demixing problem and topic models used for document classification, in particular Latent Dirichlet Allocation (LDA). We then construct a neural network implementation of variational inference and learning for LDA that utilizes a linear PPC. This network relies critically on two non-linear operations: divisive normalization and super-linear facilitation, both of which are ubiquitously observed in neural circuits. We also demonstrate how online learning can be achieved using a variation of Hebb's rule and describe an extesion of this work which allows us to deal with time varying and correlated latent causes.

REVIEWERS

Abbasi Yasin
Abeel Thomas
Ackerman Margareta
Adams Ryan
Agarwal Alekh
Agarwal Alekh
Agarwal Shivani
Ahmed Amr
Ailon Nir
Airoldi Edo
Alahari Karteek
Alan Qi
Ali Karim
Altun Yasemin
Alvarez Mauricio
Alvarez Marco
Alzate Carlos
Amini Massih-Reza
Anandkumar Anima
Andres Bjoern
Archambeau Cedric
Argyriou Andreas
Arora Raman
Arras Kai
Asuncion Arthur
Atkeson Chris
Austerweil Joseph
Avitan Lilach
Bach Francis
Bai Bing
Balakrishnan Suhrid
Balakrishnan Sivaraman
Baldassarre Luca
Balle Borja
Balzano Laura
Baraniuk Richard
Barash Yoseph
Barreto Andre
Basu Sumit
Batra Dhruv
Belkin Mikhail
Bell Robert
Bellemare Marc
Ben Shitrit Horesh
Bengio Yoshua
Bennett Kristin
Berens Philipp
Beygelzimer Alina
Bhattacharyya Chiranjib
Bi Jinbo
Biessmann Felix
Bilenko Misha
Birnbaum Aharon
Blanchard Gilles
Blaschko Matthew
Blei David
Blitzer John
Blundell Charles
Bollegala Danushka
Bonilla Edwin
Boots Byron
Bordes Antoine
Borgwardt Karsten
Bosagh Zadeh Reza
Botvinick Matthew
Bouchard Guillaume
Boucheron Stephane
Boureau YLan
Boutsidis Christos
Bowling Michael
Boyd-Graber Jordan
Braun David
Braun Mikio
Brefeld Ulf
Brubaker Marcus
Buesing Lars
Buhmann Joachim
Buntine Wray
Burgard Wolfram
Burkitt Anthony
Busa-Fekete Róbert
Busoniu Lucian
Cadieu Charles
Cai Deng
Camerer Colin
Campbell Colin
Canini Kevin
Canu Stephane
Cappe Olivier
Caputo Barbara
Caramanis Constantine
Carpentier Alexandra
Carreira-Perpinan Miguel
Cauwenberghs Gert
Cawley Gavin
Celikyilmaz Asli
Cesa-Bianchi Nicolò
Cevher Volkan
Chan Hubert
Chang Jonathan
Chaudhuri Kamalika
Chechik Gal
Chen Minmin
Chen Jieqiu
Chen Minhua
Chen Shuo
Chiappa Silvia
Chiquet Julien
Chklovskii Dmitri
Choi Arthur
Choi Seungjin
Chopra Sumit
Clemencon Stephan
Coates Adam
Coates Mark
Coen-cagli Ruben
Cohen Mark
Cormode Graham
Corrado Greg
Cosatto Eric
Cottrell Garrison
Cour Timothee
Courville Aaron
Craven Mark
Cremers Daniel
Cussens James
Dalalyan Arnak
d'Alche-Buc Florence
Daniely Amit
Das Dipanjan
Das Sanmay
Dasgupta Sanjoy
Dauce Emmanuel
Daume III Hal
Davidson Ian
Davis Jesse
Dayan Peter
de Campos Cassio
De Sa Viriginia
Decoste Dennis
Deisenroth Marc
Dekel Ofer
Dembczynski Krzysztof
DeNero John
Deneve Sophie
Deng Li
Deng Jia
Dietterich Thomas
Dietz Laura
Dillon Joshua V.
Ding Chris
Dinuzzo Francesco
Doi Eizaburo
Dollar Piotr
Domke Justin
Doshi-Velez Finale
Doucet Arnaud
Dredze Mark
Drineas Petros
Dror Gideon
Duchenne Olivier
Duchi John
Dudik Miroslav
Dunson David
Dutta Haimonti
Dy Jennifer
El Ghaoui Laurent
Elder James
Elkan Charles
Elliott Lloyd
Emmert-Streib Frank
Erhan Dumitru
Ernst Damien
Farabet Clement
Farahmand Amir massoud
Farhadi Ali
Faugeras Olivier
Favaro Paolo
Ferrari Vittorio
Figueiredo Mario
Fiser Jozsef
Flach Boris
Fonteneau Raphael
Forsyth David
Fowlkes Charless
Fox Emily
Frank Andrew
Frank Jordan
Frank Michael
Frazier Peter
Freeman William
Friston Karl
Fuernkranz Johannes
Fujimoto Yu
Fukumizu Kenji
Gall Juergen
Galyardt April
Ganti Ravi
Gasso Gilles
Gasthaus Jan
Gaussier Eric
Gehler Peter
Geiger Andreas
Geist Matthieu
Gelfand Andrew E.
Gerrish Sean
Gershman Sam
Gerwinn Sebastian
Geurts Pierre
Ghavamzadeh Mohammad
Gheshlaghi azar Mohammad
Gilad-Bachrach Ran
Girolami Mark
Girshick Ross
Glasmachers Tobias
Gogate Vibhav
Goldberger Jacob
Goldenberg Anna
Goldman Mark
Gonen Alon
Gong Pinghua
Gopalan Raghuraman
Gordon Geoff
Goschin Sergiu
Gould Stephen
Graepel Thore
Grangier David
Gray Alexander
Gregor Karol
Gretton Arthur
Griffin Jim
Griffiths Thomas
Grosse-Wentrup Moritz
Gunawardana Asela
Gupta Abhinav
Gupta Maya
Gureckis Todd
Gutmann Michael
Habeck Michael
Haefner Ralf
Haffari Gholamreza
Hagai Attias
Hall David
Hamm Jihun
Hansen Katja
Hansen Katja
Hansen Lars-Kai
Harchaoui Zaid
Harmeling Stefan
Hartemink Alexander
Haruno Masahiko
Hasselmo Mike
Hatano Kohei
Haufe Stefan
Hays James
Hazan Elad

REVIEWERS

He Jingrui
He Xiaodong
Heess Nicolas
Hein Matthias
Hensman James
Herbrich Ralf
Herbster Mark
Hero Alfred
Heskes Tom
Hino Hideitsu
Hirsch Michael
Hlavac Vaclav
Hoey Jesse
Hoffman Matt
Hofmann Thomas
Hoiem Derek
Honkela Antti
Hsieh Cho-Jui
Hsu Chun-Nan
Huang Jonathan
Huellermeier Eyke
Hukushima Koji
Hunt Jonny
Hwang Sung Ju
Ide Tsuyoshi
Igel Christian
Ihler Alex
Ikeda Kazushi
Isbell Charles
Jaakkola Tommi
Jacob Laurent
Jacobs Robert
Jaeger Manfred
Jafarpour Sina
Jagarlamudi Jagadeesh
Jaggi Martin
Jancsary Jeremy
Janzing Dominik
Jenatton Rodolphe
Ji Shuiwang
Jia Zhaoyin
Jiang Yun
Jin Rong
Johansen Adam
Jojic Vladimir
Joulin Armand
Jurie Frederic
Kadri Hachem
Kale Satyen
Kameoka Hirokazu
Kanamori Takafumi
Kanan Christopher
Kapoor Ashish
Kar Purushottam
Karaletsos Theofanis
Karasuyama Masayuki
Karklin Yan
Kashima Hisashi
Kaski Samuel
Kavukcuoglu Koray
Kawahara Yoshinobu
Kazawa Hideto
Keerthi Sathiya
Kegl Balazs
Keller Yosi
Kemp Charles
Kersting Kristian
Kienzle Wolf
Kim Seyoung
Kimura Akisato
Kirkpatrick Bonnie
Kirshner Sergey
Kivinen Jyrki
Kloft Marius
Knowles David
Kolar Mladen
Kolmogorov Vladimir
Kolter Zico
Komachi Mamoru
Komori Osamu
Kondor Risi
Konidaris George
Koo Terry
Koppula Hema
Korc Filip
Korda Nathaniel
Kotlowski Wojciech
Kowalczyk Adam
Krause Andreas
Krawczyk Bartosz
Krishnan Dilip
Krishnapuram Balaji
Krizhevsky Alex
Kuang Rui
Kulesza Alex
Kulis Brian
Kumar M. Pawan
Kumar Sanjiv
Kundaje Anshul
Kurita Takio
Kveton Branislav
Kwok James
KyungHyun Cho
Laber Eric
Lacoste-Julien Simon
Lahaie Sebastien
Lai Kevin
Lampert Christoph
Lan Guanghui
Lanckriet Gert
Landwehr Niels
Laptev Ivan
Larochelle Hugo
Larsen Jan
Laskov Pavel
Lazaric Alessandro
Lazebnik Svetlana
Lazic Nevena
Le Quoc
Le Roux Nicolas
Lee Honglak
Lee Sangkyun
Lefakis Leonidas
Lempitsky Victor
Lenz Ian
Lepetit Vincent
Leslie Christina
Lewicki Mike
Li Hang
Li Lihong
Liao Li
Liao Xuejun
Ligett Katrina
Lin Yuanqing
Lippert Christoph
Liu Ce
Liu Han
Liu Jun
Liu Qiang
Liu Tie-Yan
Liu Wei
Lizotte Dan
Long Bo
Lowd Daniel
Lozano Aurelie
Ludvig Elliot
Ma Shiqian
Macke Jakob
Mackey Lester
Mahadevan Sridhar
Mahoney Michael
Maillard Odalric-Ambrym
Mairal Julien
Malisiewicz Tomasz
Mallat Stephane
Mamitsuka Hiroshi
Mann Gideon
Mansinghka Vikash
Margalit Oded
Marlin Ben
Martens James
Martin Cichy Radoslaw
Martins Andre
Mason Winter
Matsui Tomoko
Matsumoto Yuji
McAuley Julian
McAuliffe Jon
McCallum Andrew
McFee Brian
McMahan Brendan
Meek Chris
Meinecke Frank
Meir Ron
Melo Francisco
Memisevic Roland
Meshi Ofer
Mesterharm Chris
Mezuman Elad
Mimno David
Minka Tom
Mnih Andriy
Mochihashi Daichi
Modayil Joseph
Mohamed Shakir
Montavon Gregoire
Monteleoni Claire
Mooij Joris
Mori Greg
Morimura Tetsuro
Morris Quaid
Moschitti Alessandro
Mukherjee Sayan
Munoz Daniel
Murphy Robert
Murua Alejandro
Muthukumarana Saman
Nakajima Shinichi
Negahban Sahand
Nelson Blaine
Netrapalli Praneeth
Neu Gergely
Neumann Gerhard
Neville Jennifer
Newman David
Nguyen Patrick
Nguyen Huy
Niculescu-Mizil Alexandru
Niu Gang
Niv Yael
Obozinski Guillaume
Odobez Jean-Marc
Oh Sewoong
Olshausen Bruno
Onoda Takshi
Opper Manfred
Orabona Francesco
O'Reilly Una May
Ortner Ronald
Page David
Paiement Jean-Francois
Paisley John
Pal David
Papandreou George
Paquet Ulrich
Parikh Devi
Paris Sylvain
Parr Ronald
Passerini Andrea
Pavlovic Vladimir
Pawelzik Klaus
Pelillo Marcello
Peng Jian
Perez-Cruz Fernando
Perronnin Florent
Petreska Biljana
Petrik Marek
Petrov Slav
Pfeifer Nico
Pillow Jonathan
Pitkow Xaq
Platt John
Pless Robert
Pletscher Patrick
Poczos Barnabas
Poupart Pascal
Preux Philippe
Qi Yanjun
Quadrianto Novi
Quattoni Ariadna
Rabbat Michael
Raginsky Maxim
Rai Piyush
Rajashekar Umesh
Rakotomamonjy Alain
Ralaivola Liva

REVIEWERS

Ramadge Peter
Ramamoorthy Subramanian
Ramanan Deva
Rao Vinayak
Raskutti Garvesh
Ravikumar Pradeep
Ravindran Balaraman
Raykar Vikas
Reichart Roi
Reid Mark
Reisinger Joseph
Rieck Konrad
Robin St?phane
Rosasco Lorenzo
Rossi Fabrice
Roth Andrew
Roth Stefan
Roth Volker
Rothblum Guy
Rousu Juho
Rueckert Ulrich
Rush Alexander
Russell Bryan
Ryabko Daniil
Saberian Mohammad Ehsan
Sadrzadeh Mehrnoosh
Saenko Kate
Saha Ankan
Sahani Maneesh
Saigo Hiroto
Sakuma Jun
Salzmann Mathieu
Samaras Dimitris
Sanghavi Sujay
Sanguinetti Guido
Sankararaman Sriram
Sanner Scott
Sapp Ben
Sarawagi Sunita
Saria Suchi
Sarwate Anand
Sato Issei
Savchynskyy Bogdan
Savin Cristina
Schaal Stefan
Schapire Robert
Scheffer Tobias
Scheinberg Katya
Scherrer Bruno
Schiele Bernt
Schlesinger Dmitrij
Schliep Alexander
Schmidt Mark
Schneider Jeff
Schulz Hannes
Schuurmans Dale
Schwartz Odelia
Schweikert Gabriele
Schwing Alex
Scott Clayton
Scott James
Sebag Michele
Seeger Matthias
Senior Andrew
Seong-Whan Lee
Sermanet Pierre
Shafto Patrick
Shah Devavrat
Shakhnarovich Greg
Shalit Uri
Shalizi Cosma
Shamir Ohad
Sharma Mohit
Sharpee Tatyana
Sharpnack James
Shawe-Taylor John
Sheffet Or
Shen Xiaotong
Shenoy Pradeep
Shimizu Nobuyuki
Shimizu Shohei
Shin Helen
Shivaswamy Pannaga
Shpigelman Lavi
Silva Ricardo
Silver David
Simsek Ozgur
Sindhwani Vikas
Singer Yoram
Singh Ajit
Singh Sameer
Sinha Kaushik
Sinz Fabian
Sivic Josef
Small Kevin
Sminchisescu Cristian
Socher Richard
Sohl-Dickstein Jascha
Sollich Peter
Sommer Fritz
Song Le
Sontag David
Soudry Daniel
Sridharan Karthik
Sriperumbudur Bharath
Stegle Oliver
Stocker Alan
Stokes Jay
Streeter Matt
Sturm Juergen
Subramanya Amar
Sun Min
Sutskever Ilya
Sutton Richard
Suykens Johan
Suzuki Taiji
Syed Umar
Szafranski Marie
Sznitman Raphael
Takenouchi Takashi
Takeuchi Ichiro
Takimoto Eiji
Talukdar Partha
Talvitie Erik
Talwalkar Ameet
Tanaka Toshiyuki
Tang Yichuan
Tangermann Michael
Tao Dacheng
Tappen Marshall
Tarlow Danny
Taskar Ben
Tatsuno Masami
Tatti Nikolaj
Taylor Graham
Telgarsky Matus
Tenenbaum Josh
Teytaud Olivier
Thiesson Bo
Thijssen Sep
Timme Marc
Titov Ivan
Titsias Michalis
Todorovic Sinisa
Tomioka Ryota
Tong Hanghang
Torralba Antonio
Torresani Lorenzo
Toshev Alexander
Toussaint Marc
Tran-Thanh Long
Tresp Volker
Tsang Ivor
Tsochantaridis Ioannis
Tsuda Koji
Turaga Srini
Turner Richard
Ueda Naonori
Ujfalussy Balazs
Ungar Lyle
Urner Ruth
Usunier Nicolas
Uther William
Valko Michal
van der Maaten Laurens
Van Gael Jurgen
Vanhoucke Vincent
Vasconcelos Nuno
Vedaldi Andrea
Vembu Shankar
Verbeek Jakob
Verma Nakul
Verri Alessandro
Vijayanarasimhan Sudheendra
Vincent Pascal
Vishwanathan SVN
Vitale Fabio
Vogt Julia
von Luxburg Ulrike
Vul Ed
Wagstaff Kiri
Wallis Guy
Wang Chong
Wang Jack
Wang Yang
Wang Liwei
Wang Tong
Wasserman Larry
Watanabe Kazuho
Watkins Chris
Wauthier Fabian
Wei Chu
Weiss Yair
Wen Zaiwen
Werner Tomas
Whiteson Shimon
Willet Rebecca
Williamson Robert
Williamson Sinead
Winther Ole
Wolf Lior
Woznica Adam
Wright John
Wu Mingrui
Xiao Jianxiong
Xiaofeng Ren
Xie Lexing
Xu Min
Xu Min
Xu Huan
Yamada Makoto
Yamada Makoto
Yamakawa Hiroshi
Yamanishi Yoshihiro
Yanai Keiji
Yang Liu
Yao Bangpeng
Yao Angela
Yeung Dit-Yan
Yih Scott
Yin Junming
Yin Junming
Yin Wotao
Ying Yiming
Yoo Chang D.
Yoshii Kazuyoshi
Yoshimoto Junichiro
Yoshioka Taku
Yu Byron
Yu Chun-Nam
Yu Shipeng
Yu Yaoliang
Yuan Xiaoming
Yue Yisong
Zeiler Matt
Zhang Shunan
Zhang Xinhua
Zhou Dengyong
Zhou Mingyuan
Zhou Xueyuen
Zhou Zhi-Hua
Zhu Jun
Zhu Shenghuo
Zickler Todd
Ziehe Andreas
Zinkevich Martin
Zisserman Andrew
Zitnick Larry
Zoran Daniel

AUTHOR INDEX

AUTHOR INDEX

AUTHOR INDEX

AUTHOR INDEX

AUTHOR INDEX

NEXT CONFERENCE

2013 - 2014
LAKE TAHOE
NEVADA

TM

Lake Tahoe, Nevada