2013

CONFERENCE BOOK

NIPS 2013

Abstracts of Papers

TUTORIALS
December 5, 2013
Harrah’s & Harveys
Lake Tahoe, Nevada

CONFERENCE SESSIONS
December 6 - 8, 2013
Harrah’s & Harveys
Lake Tahoe, Nevada

WORKSHOPS
December 9 - 10, 2013
Harrah’s & Harveys
Lake Tahoe, Nevada

Sponsored by the Neural Information Processing
System Foundation, Inc

The technical program includes 6 invited talks and 360 accepted papers, selected from a total of 1420 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 26,” edited by Léon Bottou, Chris Burges, Max Welling, Zoubin Ghahramani and Kilian Weinberger

Neural Information
Processing Systems
Foundation

TABLE OF CONTENTS

THURSDAY

Tutorials

Poster Sessions

FRIDAY

Oral Sessions

Spotlights Sessions

Poster Sessions

SATURDAY

Oral Sessions

Spotlights Sessions

Poster Sessions

SUNDAY

Oral Sessions

Spotlights Sessions

Poster Sessions

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

ORGANIZING COMMITTEE

General Chairs:	**Léon Bottou** (Microsoft Research),
	Chris J. C. Burges (Microsoft Research)
Program Chairs:	**Max Welling** (University of Amsterdam)
	Zoubin Ghahramani (Univ. of Cambridge)
Tutorials Chair:	**Neil Lawrence** (University of Sheffield)
Workshop Chairs:	**Rich Caruana** (Microsoft Research),
	Gunnar Rätsch (Memorial Sloan-Kettering Cancer Center)

Demonstration Chair:	**Russ Salakhutdinov** (Univ. of Toronto)
Publications Chair and Electronic Proceedings Chair:	**Kilian Weinberger** (Washington Univ)
Program Manager:	**Hong Ge** (University of Cambridge)

PROGRAM COMMITTEE

Jacob Abernethy (University of Pennsylvania)
Ryan Adams (Harvard University)
Alekh Agarwal (Microsoft Research)
Cedric Archambeau (Xerox)
Francis Bach (INRIA-ENS)
Serge Belongie (UCSD)
Matthias Bethge (University of Tübingen)
Jeff Bilmes (University of Washington)
Karsten Borgwardt (MPI for Intelligent Systems)
Miguel Carreira-Perpiñan (UC Merced)
Gal Chechik (Bar Ilan University)
Corinna Cortes (Google)
Koby Crammer (Technion)
John Cunningham (Washington University in St Louis)
Ofer Dekel (Microsoft Research)
Emily Fox (University of Washington)
Mohammad Ghavamzadeh (INRIA)
Amir Globerson (Hebrew University of Jerusalem)
Dilan Görür (Yahoo Research)
Stefan Harmeling (MPI for Intelligent Systems)
Elad Hazan (Technion)
Tamir Hazan (TTI Chicago)
Matt Hoffman (Adobe)
Daniel Hsu (Microsoft Research)
Alex Ihler (UC Irvine)
Prateek Jain (Microsoft Research)
Dominik Janzing (MPI for Intelligent Systems)
Rong Jin (Michigan State University)
Andreas Krause (ETH Zürich)
Gert Lanckriet (UCSD)
Hugo Larochelle (Sherbrooke University)
Neil Lawrence (University of Sheffield)
Jure Leskovec (Stanford University)
Christina Leslie (Memorial Sloan-Kettering Cancer Center)
Fei Fei Li (Stanford University)
Han Liu (Princeton)
Mehryar Mohri (NYU and Google)
Claire Monteleoni (George Washington University)
Iain Murray (University of Edinburgh)
Guillaume Obozinski (Ecole des Ponts-ParisTech)
Jan Peters (TU Darmstadt)
Jonathan Pillow (University of Texas at Austin)
Massimiliano Pontil (UCL)
Gunnar Rätsch (Memorial Sloan-Kettering Cancer Center)
Maxim Raginsky (UIUC)
Deva Ramanan (UC Irvine)
Marc'Aurelio Ranzato (Google)
Lorenzo Rosasco (University of Genova)
Daniel Roy (University of Cambridge)
Cynthia Rudin (MIT)
Odelia Schwartz (Albert Einstein College of Medicine)
Aarti Singh (CMU)
Satinder Singh (University of Michigan)
Le Song (Georgia Tech)
David Sontag (NYU)
Suvrit Sra (MPI for Intelligent Systems)
Bharath Sriperumbudur (University of Cambridge)
Ambuj Tewari (University of Michigan)
Richard Turner (University of Cambridge)
Nuno Vasconcelos (UCSD)
S.V.N. Vishwanathan (Purdue University)
Frank Wood (University of Oxford)
Jennifer Wortman Vaughan (Microsoft Research)
Jieping Ye (Arizona State University)
Angela Yu (UCSD)
Jun Zhu (Tsinghua University)
Andrew Zisserman (University of Oxford)

CORE LOGISTICS TEAM

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

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

2013 EXHIBITORS

Cambridge University Press
Robert Bosch LLC
CRC\Taylor & Francis Group
Now Publisher
The MIT Press
Springer
Pivotal

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
Peter Bartlett, Queensland University and University California, Berkeley; Yoshua Bengio, University of Montreal; John Lafferty, Carnegie Mellon University; Fernando Pereira, Google Research; John Shawe-Taylor, University College London; Rich Zemel, University of Toronto; Chris Williams, University of Edinburg

EMERITUS MEMBERS
Gary Blasdel, Harvard Medical School; T. L. Fine, Cornell University; Eve Marder, Brandeis University

ADVISORY BOARD
Sue Becker, McMaster Univ, Ontario, Canada; Jack Cowan, University of Chicago; Thomas G. Dietterich, Oregon State Univ. Stephen Hanson, Rutgers University; Michael I. Jordan, UC, Berkeley; Michael Kearns, University of Pennsylvania; Scott Kirkpatrick, Hebrew University; Daphne Koller, Stanford University; Todd K. Leen, Oregon Graduate Institute; Richard Lippmann, MIT; Bartlett Mel, University of Southern California; John Moody, ICSI, Berkeley and Portland; John C. Platt, Microsoft Research; Gerald Tesauro, IBM Watson Labs; Sebastian Thrun, Stanford University; Dave Touretzky, Carnegie Mellon University; Lawrence Saul, UC San Diego; Bernhard Schölkopf, Max Planck Institute; Dale Schuurmans, Univ. of Alberta, Canada; Sara A. Solla, Northwestern University; Yair Weiss, Hebrew University of Jerusalem

IN MEMORIAM

BEN TASKAR

NIPS mourns the passing of Ben Taskar, who died on November 17, 2013 at the age of 36. Ben moved in the spring of 2013 from the University of Pennsylvania to the University of Washington as Boeing Associate Professor. He worked in machine learning, computational linguistics and computer vision and his research on structured prediction won best paper awards at NIPS and EMNLP conferences. He received his bachelor's and doctoral degrees in Computer Science from Stanford University. After a postdoc at the University of California at Berkeley, he joined the faculty at the University of Pennsylvania's Computer and Information Science Department. He had a short but brilliant career, receiving a Sloan Research Fellowship, the NSF CAREER Award, and was selected for the Young Investigator Program by the Office of Naval Research and the DARPA Computer Science Study Group.

Ben was fun-loving, kind, and generous with others; Ben will be dearly missed by his friends and the whole NIPS community.

MESSAGE FROM THE PRESIDENT

PHIL SOTEL

The NIPS community has lost a valued friend and a link to its founding fathers. Phil Sotel, who served as the pro bono general counsel of the NIPS Foundation since it was formed by Ed Posner at Caltech in 1992, died on October 6, 2013 after a brief and previously undiagnosed illness at the age of 77. He guided NIPS through its early years and had been a steady hand in advising the board as NIPS grew into a major international meeting. He attended annual NIPS Board meetings and advised the Foundation on major decisions. He put in place corporate procedures that have served the Foundation well. The road that we have traveled over the last 26 years would have had a lot more bumps without Phil's advice and the next 26 years will be more challenging without him. NIPS owes Ed Posner a debt of gratitude not only for founding NIPS, but also for convincing his friend Phil Sotel to oversee its growth and health.

Phil was involved with petroleum exploration in Indonesian and the Philippines and although he was based in Pasadena, he had a ranch in Colorado. He had close ties with Caltech and was on the wine committee at the Caltech Athenium. He was a member of the Hoover Institution at Stanford and the Pacific Council on International Policy, the West Coast affiliate of the Council on Foreign Relations. He had a broad perspective on the world, and many friends in commerce and academia, which helped NIPS grow into an organization that has had a significant impact on the development of the knowledge economy.

The first NIPS conference and workshop was held in Denver in 1987. It attracted a broad range of scientists and engineers with diverse backgrounds and interests eager to solve intractable problems using massively-parallel architectures and learning algorithms to deal with the high dimensionality of the parameters spaces and data sets. Brains were an inspiration and an existence proof that these problems could in fact be solved. NIPS served as an incubator for a field now widely known as machine learning, with deep roots in artificial intelligence, statistics, computational neuroscience, cognitive science, computer vision, control theory, speech recognition, neuromorphic engineering and many other disciplines that rely on the computational tools that have been developed by the NIPS community over the past 26 years. The annual NIPS conference and workshop has grown from 600 participants in 1987 to over 1,600 in 2012.

The Board of Trustees of the NIPS Foundation includes recent general chairs and also relies on an Advisory Board for continuity with past conferences. The Foundation oversees the annual meeting and handles its infrastructure, with the Board making decisions on where the meetings are held and who is asked to serve on organizing committees. The program committee chairs have the important task of choosing the best submissions and invited lecturers. Workshops are the crucibles for future advances that keep NIPS healthy. The NIPS Foundation thanks the members of all the organizing committees who have given their valuable time and service over the years.

Terry Sejnowski
La Jolla, CA, November 4, 2013

SPONSORS

NIPS gratefully acknowledges the generosity of those individuals and organizations who have provided financial support for the NIPS 2013 conference. The financial support enabled us to sponsor student travel and participation, the outstanding paper awards, the demonstration track and the opening buffet.

AFOSR continues to expand the horizon of scientific knowledge through its leadership and management of the Air Force's basic research program. As a vital component of the Air Force Research Laboratory (AFRL), AFOSR's mission is to support Air Force goals of control and maximum utilization of air, space, and cyberspace.

AFOSR accomplishes its mission by investing in basic research efforts for the Air Force in relevant scientific areas. Central to AFOSR's strategy is the transfer of the fruits of basic research to industry, the supplier of Air Force acquisitions; to the academic community which can lead the way to still more accomplishment; and to the other directorates of Air Force Research Laboratory (AFRL) that carry the responsibility for applied and development research leading to acquisition.

Microsoft Research is dedicated to pursuing innovation through basic and applied research in computer science and software engineering. Basic long-term research, unconstrained by the demands of product cycles, leads to new discoveries and lays the foundation for future technology breakthroughs that can define new paradigms, such as the current move toward cloud computing and software-plus-services. Applied research focuses on the near-term goal of improving products by transferring research findings and innovative technology to development teams. By balancing basic and applied research, and by maintaining an effective bridge between the two, Microsoft Research continually advances the state of the art in computer science and redefines the computing experience for millions of people worldwide. Microsoft Research has more than 1,100 scientists and engineers specializing in over 60 disciplines and includes some of the world's finest computer scientists, sociologists, psychologists, mathematicians, physicists, and engineers, working in our worldwide locations.

Amazon.com strives to be Earth's most customer-centric company where people can find and discover virtually anything they want to buy online. Amazon's evolution from Web site to e-commerce partner to development platform is driven by the spirit of innovation that is part of the company's DNA. The world's brightest technology minds come to Amazon.com to research and develop technology that improves the lives of shoppers, sellers and developers around the world. At Amazon, our Machine Learning team is comprised of technical leaders who develop planet-scale platforms for machine learning on the cloud, assist in the benchmarking and future development of existing machine learning applications across Amazon, and help develop novel and infinitely-scalable applications.

Google's mission is to organize the world's information and make it universally accessible and useful. Perhaps as remarkable as two Stanford research students having the ambition to found a company with such a lofty objective is the progress the company has made to that end. Ten years ago, Larry Page and Sergey Brin applied their research to an interesting problem and invented the world's most popular search engine. The same spirit holds true at Google today. The mission of research at Google is to deliver cutting-edge innovation that improves Google products and enriches the lives of all who use them. We publish innovation through industry standards, and our researchers are often helping to define not just today's products but also tomorrow's.

Helping over a billion people share and connect around the globe requires constant innovation. At Facebook, research permeates everything we do. Here, research is more than a lab—it's a way of doing things.

At Facebook, we believe that the most interesting academic problems are derived from real-world problems. Our researchers work on cutting edge research problems with a practical focus and push product boundaries every day. At the same time, they are publishing papers, giving talks, attending and hosting conferences and collaborating with the academic community. Our research teams are an integral part of the engineering organization and work with real user data to solve real-world problems that impact millions of people.

Skytree® – The Machine Learning Company® is disrupting the Advanced Analytics market with a Machine Learning platform that gives organizations the power to discover deep analytic insights, predict future trends, make recommendations and reveal untapped markets and customers. Predictive Analytics is quickly becoming a must-have technology in the age of Big Data, and Skytree is at the forefront with enterprise-grade Machine Learning. Skytree's flagship product – Skytree Server – is the only general purpose scalable Machine Learning system on the market, built for the highest accuracy at unprecedented speed and scale.

United Technologies Research Center delivers the world's most advanced technologies, innovative thinking and disciplined research to the businesses of United Technologies -- industry leaders in aerospace propulsion, building infrastructure and services, heating and air conditioning, fire and security systems and power generation. Founded in 1929, UTRC is located in East Hartford, Connecticut (U.S.), with an office in Berkeley, California, and research and development centers in Shanghai, China, and Cork, Ireland. UTRC currently has several open roles for people with strong machine learning and distributed analytics skills to support service technologies across a wide array of applied industrial applications. If you're strong technically and enjoy working across a broad array of technical domains, UTRC may be the place for you.

IBM Research is a research and development organization consisting of twelve laboratories, worldwide. Major undertakings at IBM Research have included the invention of innovative materials and structures, high-performance microprocessors and computers, analytical methods and tools, algorithms, software architectures, methods for managing, searching and deriving meaning from data and in turning IBM's advanced services methodologies into reusable assets. IBM Research's numerous contributions to physical and computer sciences include the Scanning Tunneling Microscope and high temperature superconductivity, both of which were awarded the Nobel Prize. IBM Research was behind the inventions of the SABRE travel reservation system, the technology of laser eye surgery, magnetic storage, the relational database, UPC barcodes and Watson, the question-answering computing system that won a match against human champions on the Jeopardy! television quiz show. The Watson technology is now being commercialized as part of a project with healthcare company WellPoint. IBM Research is home to 5 Nobel Laureates, 9 US National Medals of Technology, 5 US National Medals of Science, 6 Turing Awards, and 13 Inductees in the National Inventors Hall of Fame.

We are a technology company that applies a rigorous, scientific method-based approach to investment management. Since our founding in 2001, Two Sigma's vision has been to develop technological innovations that intelligently analyze the world's data to consistently deliver value for our clients. Our technology – inspired by a diverse set of fields including artificial intelligence and distributed computing – and our commitment to Research & Development aim to ensure that our methods are constantly improving and advancing.

Headquartered in New York City, the D. E. Shaw group is a global investment and technology development firm with offices in North America, Europe, and Asia. Since its organization in 1988 by a former Columbia University computer science professor, David E. Shaw, the firm has earned an international reputation for successful investing based on financial innovation, careful risk management, and the quality and depth of our staff. Our investment activities are based on both mathematical models and our staff's expertise, and our multi-disciplinary approach combines insights from quantitative fields, software development, sector expertise, and finance. We offer the benefits of being one of the world's largest, most established alternative investment managers, with a world-class technology infrastructure, deep research capabilities, and programs that facilitate the ongoing growth and internal mobility of staff. We have a long history of looking for candidates who aren't conventional "financial types," and our culture doesn't fit the typical corporate mold.

DRW Trading Group (DRW) is a principal trading organization. This means that all of our trading is for our own account and risk, and all of our methods, systems and applications are solely for our own use. Unlike hedge funds, brokerage firms and banks, DRW has no customers, clients, investors or third party funds. Our trading spans a wide range of asset classes, instruments, geographies and trading venues, with a focus on trading listed, centrally-cleared instruments.

Founded in 1992, our mission is to empower a team of exceptional individuals to identify and capture trading opportunities in the global markets by leveraging and integrating technology, risk management and quantitative research. With that spirit, DRW has embraced the integration of trading and technology by devoting extensive time, capital and resources to develop fast, precise and reliable infrastructure and applications. DRW has a flexible and entrepreneurial culture that cultivates creativity and practicality.

YAHOO! LABS

Yahoo Labs is the scientific engine powering one of the most trafficked Internet destinations worldwide. From idea to product innovation, Yahoo Labs is responsible for the algorithms behind the quality of the Web experience for hundreds of millions of users. We impact more than 800 million people in 60 countries who use Yahoo, and we do it from some of the most interesting, diverse, creative and inspiring locations on the planet. Our scientists collaborate with each other and with scientists outside Yahoo, pioneering innovations that improve the Yahoo experience in both evolutionary and revolutionary ways. Yahoo Labs scientists invent the technologies of the future, and then make them a reality today.

PDT PARTNERS

PDT Partners is a top quantitative hedge fund where world class researchers analyze rich data to develop and deploy model-driven algorithmic trading strategies. We offer a strong track record of hiring, challenging and retaining scientists interested in conducting research where the lab is the financial markets. Our researchers come from a variety of disciplines and backgrounds, having published in the top conferences and journals in machine learning, statistics, information theory, computational biology, pure and applied mathematics, theoretical and experimental physics, and operations research.

Composed of a tight-knit community of researchers, technologists, and business professionals, we strive to build one of the best quantitative trading firms in the world. Our investment success is driven by rigorous research, state-of-the-art technology, and keen focus on risk management and trade execution. We accomplish our goals by valuing depth and expertise, encouraging intellectual curiosity, and seeking constant innovation.

TOYOTA

Toyota Research Institute of North America (TRI-NA) was established in 2008 as a division of Toyota Technical Center (TTC) in Ann Arbor, MI. Toyota has been pursuing Sustainable Mobility, which addresses four key priorities: advanced technologies, urban environment, energy, and partnerships with government and academia.

Recently Toyota Motor Corporation (TMC) and its Lexus Division unveiled its advanced active safety research vehicle for the first time at the International CES to demonstrate ongoing efforts around autonomous vehicle safety technologies and explain Toyota's approach to reducing global traffic fatalities and injuries. The vehicle, based on a Lexus LS, advances the industry toward a new era of integrated safety management technologies (see 1).

The Lexus advanced active safety research vehicle is equipped with an array of sensors and automated control systems to observe, process and respond to the vehicle's surroundings. These include GPS, stereo cameras, radar and Light Detection and Ranging (LIDAR) laser tracking.

xerox®

Xerox Research Centre Europe research covers a broad spectrum of activities linked to information, data, documents and processes. The centre is internationally reputed for its expertise in computer vision, data analytics, natural language processing, machine learning, ethnography and process modelling.

The Machine Learning for Services group conducts fundamental and applied research in machine learning, computational statistics, and algorithmic mechanism design. Our research results are used in a wide range of applications, including relational learning, personalised content creation, large-scale recommender systems, and dynamic pricing.

The evidence-driven solutions we develop are part of Xerox services offerings. Xerox is the world leader in document management and business process outsourcing and research in Europe ensures that Xerox maintains that position.

Xerox Research Centre Europe is part of the global Xerox Innovation Group made up of 650 researchers and engineers in five world-class research centres. The Grenoble site is set in a park in the heart of the French Alps in a stunning location only a few kilometres from the city centre.

criteo.

Criteo enables companies to engage and convert their customers online whether they are on a desktop,laptop, tablet or smartphone. Through its proprietary predictive algorithms, Criteo delivers performance- based online display advertising on real-time consumer data. Founded in 2005 in Paris, Criteo now employs more than 700 people across its 15 offices throughout the United States, Europe, Asia and Australia, serving more than 4,000 leading e-commerce companies across +35 countries globally. Our R&D team of 200+ engineers worldwide is building the next generation of digital advertising technologies that allow us to manage billions of ad impressions every month. For more information, please visit http://www.criteo.com

Springer
Machine Learning Journal

Springer Science+Business Media, LLC is an international e-publishing company for books and scholarly peer-reviewed journals in Science, Technology and Medicine (STM). E-content is distributed via numerous scientific databases, including SpringerLink, Springer Protocols, and SpringerImages. Book publications include major reference works, textbooks, monographs and various book series. Over 88,000 titles are available as e-books in 18 subject collections. The Machine Learning journal, published by Springer, is pleased to have supported the NIPS Best Student Awards.

PROGRAM HIGHLIGHTS

THURSDAY DEC 5TH

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
Harveys Convention Center Floor, CC

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

FRIDAY DEC 6TH

7:30 am – 9:00 am
Breakfast sponsored by Skytree
See map page 10

7:30 am – 5:30 pm
Registration Desk Open
Harveys Convention Center Floor, CC

9:00 – 10:10 am
Oral Session 1
INVITED TALK: Small, n=me, Data
Deborah Estrin

Scalable Influence Estimation in Continuous-Time Diffusion Networks
N. Du, L. Song, H. Zha, M. Gomez-Rodriguez

10:10 –10:30 am
Spotlights Session 1

10:30 – 10:55 am - Coffee Break

10:55 –11:40 am
NIPS award Session 1
Oral Session 2
On Decomposing the Proximal Map
Y. Yu

Non-Uniform Camera Shake Removal Using a Spatially-Adaptive Sparse Penalty
H. Zhang, D. Wipf

11:40 – 12:05 pm
Spotlights Session 2

FRIDAY DEC 6TH

12:05 – 2:00 pm - Lunch Break

2:00 – 3:30 pm
Oral Session 3
INVITED TALK: Memory Reactivation in Awake and Sleep States
Matthew Wilson

Correlations Strike Back (Again): The Case Of Associative Memory Retrieval
C. Savin, M. Lengyel, P. Dayan

A Memory Frontier for Complex Synapses
S. Lahiri, S. Ganguli

3:30 – 3:50 pm
Spotlights Session 3

3:50 – 4:15 pm - Coffee Break

4:15 – 5:40 pm
NIPS award Session 2
Oral Session 4:
Understanding Dropout
P. Baldi, P. Sadowski

Annealing Between Distributions By Averaging Moments
R. Grosse, C. Maddison, R. Salakhutdinov

A Simple Example of Dirichlet Process Mixture Inconsistency for the Number of Components
J. Miller, M. Harrison

Approximate Bayesian Image Interpretation using Generative Probabilistic Graphics Programs
V. Mansinghka, T. Kulkarni, Y. Perov, J. Tenenbaum

5:40 – 6:00 pm
Spotlight Session 4

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

PROGRAM HIGHLIGHTS

SATURDAY DEC 7TH

7:30 am – 9:00 am
Breakfast, See map page 10

7:30 am – 5:30 pm
Registration Desk Open
Harveys Convention Center Floor, CC

9:00 – 10:10 am
Oral Session 5
POSNER LECTURE: The Online Revolution: Learning without Limits
Daphne Koller

Optimizing Instructional Policies
R. Lindsey, M. Mozer, W. Huggins, H. Pashler

10:10 –10:30 am
Spotlights Session 5

10:30 – 10:55 am - Coffee Break

10:55 –11:40 am
NIPS award Session 3
Oral Session 6
A Kernel Test for Three-Variable Interactions
D. Sejdinovic, A. Gretton, W. Bergsma

More Effective Distributed ML via a Stale Synchronous Parallel Parameter Server
Q. Ho, J. Cipar, H. Cui, S. Lee, J. Kim, G. Gibson, G. Ganger, E. Xing, P.I Gibbons

11:40 – 12:05 pm
Spotlights Session 6

12:05 – 2:00 pm - Lunch Break

2:00 – 3:30 pm
Oral Session 7
INVITED TALK: Belief Propagation Algorithms: From Matching Problems to Network Discovery in Cancer Genomics
Jennifer Chayes

Message Passing Inference with Chemical Reaction Networks
N. Napp, R. Adams

Information-Theoretic Lower Bounds for Distributed Statistical Estimation With Communication Constraints
Y. Zhang, J. Duchi, M. Jordan, M. Wainwright

3:30 – 3:50 pm
Spotlights Session 7

3:50 – 4:20 pm - Coffee Break

4:20 – 5:40 pm
Oral Session 8
From Bandits to Experts: A Tale of Domination and Independence
N. Alon, Y. Mansour, N. Cesa-Bianchi, C. Gentile

Eluder Dimension and the Sample Complexity of Optimistic Exploration
D. Russo, B. Van Roy

Adaptive Market Making via Online Learning
J. Abernethy, S. Kale

Submodular Optimization with Submodular Cover and Submodular Knapsack Constraints
R. Iyer, J. Bilmes

5:40 – 6:00 pm
Spotlight Session 8

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

SUNDAY DEC 8TH

7:30 am – 9:00 am
Breakfast, See map page 10

7:30 am – 3:00 pm
Registration Desk Open
Harveys Convention Center Floor, CC

9:00 – 10:10 am
Oral Session 9
POSNER LECTURE: Neural Reinforcement Learning
Peter Dayan

Actor-Critic Algorithms for Risk-Sensitive MDPs
P. L.A., M. Ghavamzadeh

10:10 –10:30 am
Spotlights Session 9

10:30 – 10:50 am - Coffee Break

10:50 am –12:00 pm
Oral Session 10
INVITED TALK: New Methods for the Analysis of Genome Variation Datan
Richard Durbin

BIG & QUIC: Sparse Inverse Covariance Estimation for a Million Variables
C. Hsieh, P. Ravikumar, M. Sustik, I. Dhillon, R. Poldrack

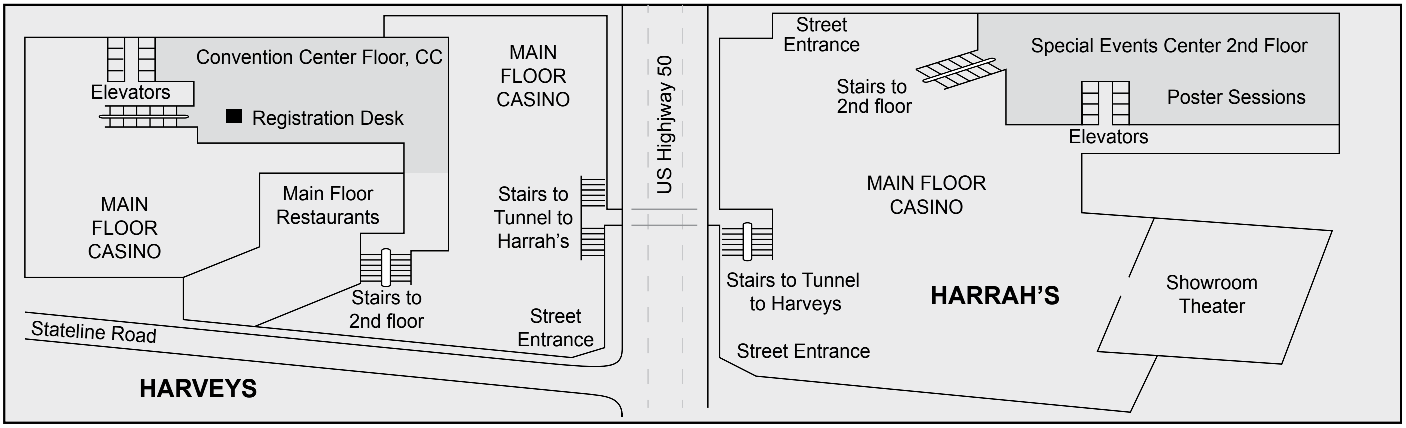
12:00 – 12:20 pm
Spotlights Session 10

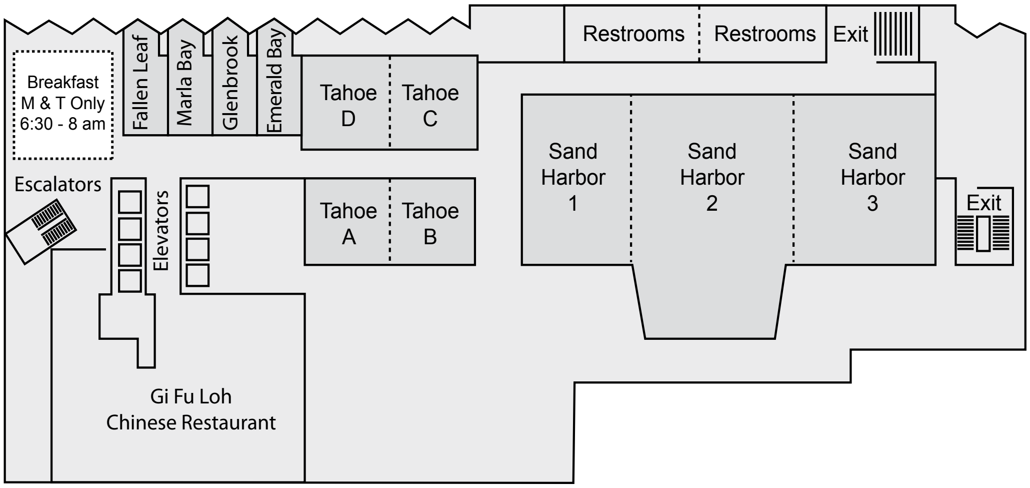
12:20 – 12:30 pm - Closing remarks

12:30 – 2:00 pm - Lunch Break

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

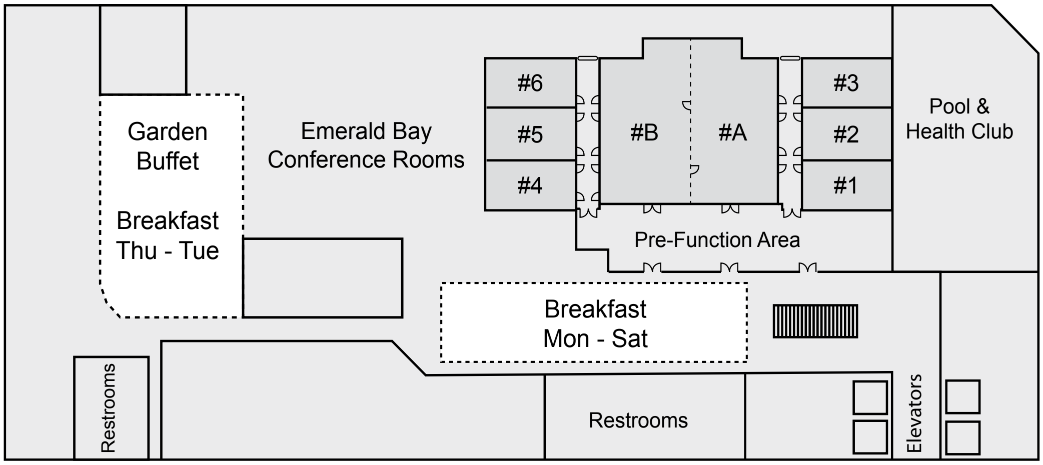
HARRAH'S & HARVEYS LOCATION MAPS





HARRAH'S SPECIAL EVENTS CENTER 2ND FLOOR

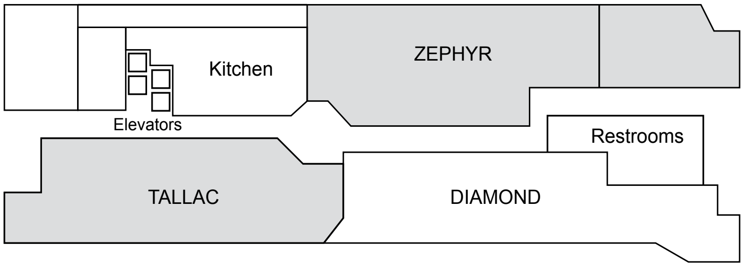
3 Breakfast Areas (dotted areas)
1 Breakfast area at Top Of The Wheel Restaurant in Harvey's 12th Floor



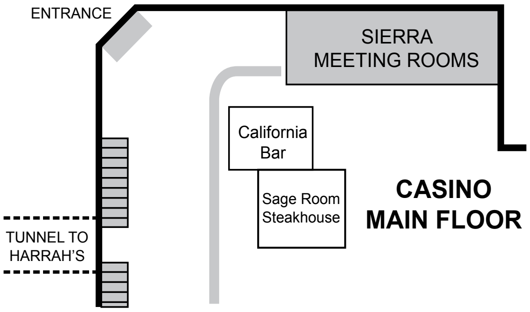
HARVEYS CONVENTION CENTER FLOOR, CC

HARVEYS

12TH FLOOR "TOP OF THE WHEEL"



HARVEYS



THURSDAY

THURSDAY TUTORIALS

Tutorial Session A, 9:30 – 11:30 AM
Session Chair: Christopher Bishop

Deep Learning for Computer Vision
Rob Fergus, NYU
Location: Emerald Bay A

Tutorial Session B, 9:30 – 11:30 AM
Session Chair: Joaquin Quiñonero Candela

Causes and Counterfactuals: Concepts, Principles and Tools.
Judea Pearl, UCLA
Elias Bareinboim, UCLA
Location: Emerald Bay B

Tutorial Session A, 1:00 – 3:00 PM
Session Chair: James Hensman

Deep Mathematical Properties of Submodularity with Applications to Machine Learning
Jeff Bilmes, University of Washington
Location: Emerald Bay A

Tutorial Session B, 1:00 – 3:00 PM
Session Chair: Dan Roy

Approximate Bayesian Computation (ABC)
Richard Wilkinson, The University of Nottingham
Location: Emerald Bay B

Tutorial Session A, 3:30 – 5:30 PM

Mechanisms Underlying Visual Object Recognition: Humans vs. Neurons vs. Machines
James DiCarlo, MIT
Location: Emerald Bay A

Tutorial Session B, 3:30 – 5:30 PM
Session Chair: Francis Bach

Learning to Interact
John Langford, Microsoft
Location: Emerald Bay B

ABSTRACTS OF TUTORIALS

Tutorial Session A, 9:30 – 11:30 AM
Session Chair: Christopher Bishop

Deep Learning for Computer Vision

Rob Fergus fergus@cs.nyu.edu
NYU

This tutorial will look at how deep learning methods can be applied to problems in computer vision, most notably object recognition. It will start by motivating the need to learn features, rather than hand-craft them. It will then introduce several basic architectures, explaining how they learn features, and showing how they can be "stacked" into hierarchies that can extract multiple layers of representation. Throughout, links will be drawn between these methods and existing approaches to recognition, particularly those involving hierarchical representations. The final part of the lecture will examine the current performances obtained by feature learning approaches on a range of standard vision benchmarks, highlighting their strengths and weaknesses. The tutorial will conclude with a discussion of vision problems that have yet to be successfully addressed by deep learning.

Tutorial Session B, 9:30 – 11:30 AM
Session Chair: Joaquin Quiñonero Candela

Causes and Counterfactuals: Concepts, Principles and Tools.

Judea Pearl judea@cs.ucla.edu
Elias Bareinboim eb@cs.ucla.edu
UCLA

The traditional aim of machine learning methods is to infer meaningful features of an underlying probability distribution from samples drawn of that distribution. With the help of such features, one can infer associations of interest and predict or classify yet unobserved samples. Causal analysis goes one step further; it aims at inferring features of the data-generating process, that is, of the invariant strategy by which Nature assigns values to the variables in the distribution. Process features enable us to predict, not merely relationships governed by the underlying distribution, but also how that distribution would CHANGE when conditions are altered, say, by deliberate interventions or by spontaneous transformations. We will review concepts, principles, and mathematical tools that were found useful in reasoning about causal and counterfactual relations, and will demonstrate their applications in several data-intensive sciences. These include questions of confounding control, policy analysis, misspecification tests, mediation, heterogeneity, selection bias, missing data, and the integration of findings from diverse studies. The following topics will be emphasized: 1. The 3-layer causal hierarchy: association, intervention and counterfactuals. http://ftp.cs.ucla.edu/pub/stat_ser/r350.pdf 2. What mathematics can tell us about "transfer learning" or "generalizing across domains" http://ftp.cs.ucla.edu/pub/stat_ser/r372.pdf http://ftp.cs.ucla.edu/pub/stat_ser/r387.pdf 3. What causal analysis tells us about recovery from selection bias and missing data. http://ftp.cs.ucla.edu/pub/stat_ser/r381.pdf http://ftp.cs.ucla.edu/pub/stat_ser/r410.pdf 4. The Mediation Formula, and what it tells us about "How nature works" http://ftp.cs.ucla.edu/pub/stat_ser/r379.pdf

Tutorial Session A, 1:00 – 3:00 PM

Session Chair: James Hensman

Deep Mathematical Properties of Submodularity with Applications to Machine Learning

Jeff Bilmes bilmes@ee.washington.edu
University of Washington

Submodular functions have received significant attention in the mathematics community owing to their natural and wide ranging applicability. Submodularity has a very simple definition which belies a treasure trove of consequent mathematical richness. This tutorial will attempt to convey some of this richness. We will start by defining submodularity and polymatroidality --- we will survey a surprisingly diverse set of functions that are submodular and operations that (sometimes remarkably) preserve submodularity. Next, we'll define the submodular polytope, and its relationship to the greedy algorithm and its exact and efficient solution to certain linear programs with an exponential number of constraints. We will see how submodularity shares certain properties with convexity (efficient minimization, discrete separation, subdifferentials, lattices and sub-lattices, and the convexity of the Lovasz extension), concavity (via its definition, submodularity via concave functions, superdifferentials), and neither (simultaneous sub- and super-differentials, efficient approximate maximization). The Lovasz extension will be given particular attention due to its growing use for structured convex norms and surrogates in relaxation methods. We will survey both constrained and unconstrained submodular optimization (including the minimum norm point algorithm), discussing what is currently known about hardness (both upper and lower bounds), and also when algorithms or instances are practical. As to applications, it is interesting that a submodular function itself can often be seen as a parameter to instantiate a machine-learning instance --- this includes active/semi-supervised learning, structured sparsity inducing norms, combinatorial independence and generalized entropy, and rank-order based divergences. Other examples include feature selection, data subset (or core set) selection, inference in graphical models with high tree-width and global potentials in computer vision, and influence determination in social networks.

Tutorial Session B, 1:00 – 3:00 PM

Session Chair: Dan Roy

Approximate Bayesian Computation (ABC)

Richard Wilkinson r.d.wilkinson@nottingham.ac.uk
The University of Nottingham

Approximate Bayesian computation (ABC) algorithms are a class of Monte Carlo methods for doing inference when the likelihood function can be simulated from, but not explicitly evaluated. This situation commonly occurs when using even relatively simple stochastic models. The algorithms can be viewed as methods for combining the scientific knowledge encoded in a computer model, with the empirical information contained in the data. The methods have become popular in the biological sciences, particularly in fields such as genetics and systematic biology, as they are simple to apply, and can be used on nearly any problem. However, there are several problems with ABC algorithms: they can be inefficient if applied naively; they only give approximate answers with the quality of the approximation unknown; they rely on a vector of summary statistics that is difficult to choose. In the first part of this tutorial, I will introduce the basic ideas behind ABC algorithms and illustrate their use on a problem from climate science. In the second part, I will describe some of the recent advances in ABC research, including regression adjustment methods, automatic summary selection, and the use of generalized acceptance kernels.

Tutorial Session A, 3:30 – 5:30 PM

Mechanisms Underlying Visual Object Recognition: Humans vs. Neurons vs. Machines

James DiCarlo dicarlo@mit.edu
Massachusetts Institute of Technology

Visual object recognition (OR) is a central problem in systems neuroscience, human psychophysics, and computer vision. A recognition system must be robust to image variation produced by different "views" of each object-- the so-called "invariance problem." My laboratory aims to understand and emulate the primate brain's solution to this problem. We have previously shown that a part of the non-human primate ventral visual stream (inferior temporal cortex, IT) rapidly and automatically conveys neuronal population rate codes that qualitatively solve the invariance problem for vision. But are such codes quantitatively sufficient to explain behavioral OR performance? Our results show that these codes are a powerful object representation, in that low complexity decoding tools can be applied to them to perfectly predict human performance over a large range of OR tasks. But how does the brain build this powerful representation? High-throughput computational methods can be used to explore a large family of biologically-constrained neural network architectures. Using this approach, we have recently discovered that functional optimization of this large family leads to specific algorithms that predict the response properties of IT dramatically better than all previous models. This suggests that these networks have captured key encoding mechanisms of human OR, and that today's computer vision algorithms are very close to emulating the power of the primate OR system.

Tutorial Session A, 3:30 – 5:30 PM

Session Chair: Francis Bach

Learning to Interact

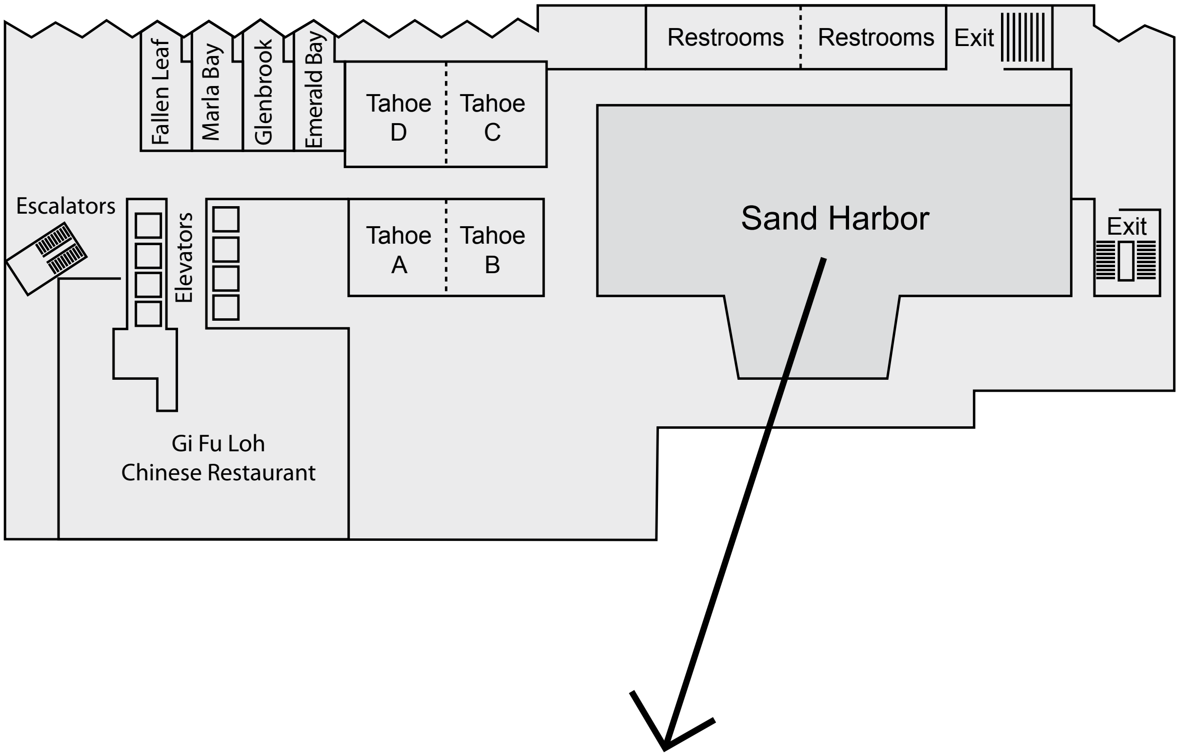
John Langford jl@hunch.net
Microsoft

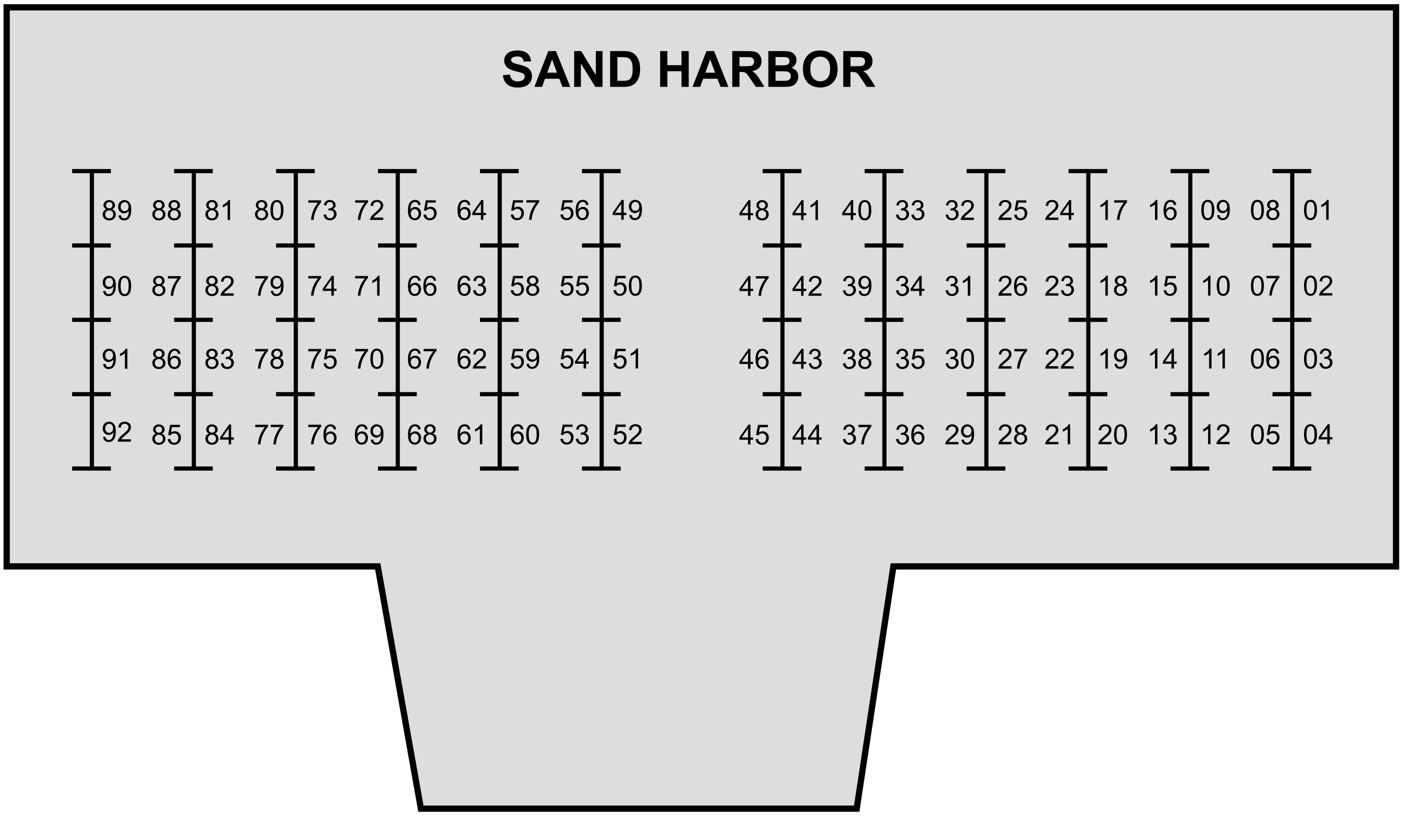
Machine Learning does magical things when it starts interacting with and changing the world, yet most algorithms are _not_ designed to do this. Systematically gathering the right data is the first order problem for learning with interaction. One simplistic example of this is ad recommendation where a high recommendation implies high placement which implies high click-through which implies high recommendation.... creating a self-fulfilling prophecy. This talk is about how to systematically avoid these problems by effectively (re)using randomization to engage in controlled exploration for learning algorithms. With these techniques, we can exponentially reduce the amount of exploration required, test many policies offline, and repurpose our existing learning algorithms to directly solve for optimal policies.

THURSDAY POSTER FLOORPLAN

HARRAH'S

2ND FLOOR SPECIAL EVENTS CENTER





89	88	81	80	73	72	65	64	57	56	49	48	41	40	33	32	25	24	17	16	09	08	01
90	87	82	79	74	71	66	63	58	55	50	47	42	39	34	31	26	23	18	15	10	07	02
91	86	83	78	75	70	67	62	59	54	51	46	43	38	35	30	27	22	19	14	11	06	03
92	85	84	77	76	69	68	61	60	53	52	45	44	37	36	29	28	21	20	13	12	05	04

THURSDAY, DECEMBER 5TH

6:30 – 6:40PM - OPENING REMARKS
Harveys Convention Center Floor, CC

POSTER SESSION

AND RECEPTION - 7:00 – 11:59 PM

T1 **Learning and using language via recursive pragmatic reasoning about other agents**
N. Smith, N. Goodman, M. Frank

T2 **Model Selection for High-Dimensional Regression under the Generalized Irrepresentability Condition**
A. Javanmard, A. Montanari

T3 **Confidence Intervals and Hypothesis Testing for High-Dimensional Statistical Models**
A. Javanmard, A. Montanari

T4 **Compressive Feature Learning**
H. Paskov, R. West, J. Mitchell, T. Hastie

T5 **Pass-efficient unsupervised feature selection**
C. Maung, H. Schweitzer

T6 **Better Approximation and Faster Algorithm Using the Proximal Average**
Y. Yu

T7 **Polar Operators for Structured Sparse Estimation**
X. Zhang, Y. Yu, D. Schuurmans

T8 **On the Linear Convergence of the Proximal Gradient Method for Trace Norm Regularization**
K. Hou, Z. Zhou, A. So, Z. Luo

T9 **Accelerating Stochastic Gradient Descent using Predictive Variance Reduction**
R. Johnson, T. Zhang

T10 **Accelerated Mini-Batch Stochastic Dual Coordinate Ascent**
S. Shalev-Shwartz, T. Zhang

T11 **Estimation, Optimization, and Parallelism when Data is Sparse**
J. Duchi, M. Jordan, B. McMahan

T12 **Linear Convergence with Condition Number Independent Access of Full Gradients**
L. Zhang, M. Mahdavi, R. Jin

T13 **Mixed Optimization for Smooth Functions**
M. Mahdavi, L. Zhang, R. Jin

T14 **Stochastic Convex Optimization with Multiple Objectives**
M. Mahdavi, T. Yang, R. Jin

T15 **Data-driven Distributionally Robust Polynomial Optimization**
M. Mevissen, E. Ragnoli, J. Yu

T16 **Multiscale Dictionary Learning for Estimating Conditional Distributions**
F. Petralia, J. Vogelstein, D. Dunson

T17 **On the Sample Complexity of Subspace Learning**
A. Rudi, G. Canas, L. Rosasco

T18 **Least Informative Dimensions**
F. Sinz, A. Stockl, J. Grewe, J. Benda

T19 **Blind Calibration in Compressed Sensing using Message Passing Algorithms**
C. Schulke, F. Caltagirone, F. Krzakala, L. Zdeborova

T20 **Estimating LASSO Risk and Noise Level**
M. Bayati, M. Erdogdu, A. Montanari

T21 **A Graphical Transformation for Belief Propagation: Maximum Weight Matchings and Odd-Sized Cycles**
J. Shin, A. Gelfand, M. Chertkov

T22 **Sensor Selection in High-Dimensional Gaussian Trees with Nuisances**
D. Levine, J. How

T23 **Σ-Optimality for Active Learning on Gaussian Random Fields**
Y. Ma, R. Garnett, J. Schneider

T24 **Bayesian optimization explains human active search**
A. Borji, L. Itti

T25 **Latent Structured Active Learning**
W. Luo, A. Schwing, R. Urtasun

T26 **Low-Rank Matrix and Tensor Completion via Adaptive Sampling**
A. Krishnamurthy, A. Singh

T27 **Adaptive Submodular Maximization in Bandit Setting**
V. Gabillon, B. Kveton, Z. Wen, B. Eriksson, S. Muthukrishnan

T28 **Auditing: Active Learning with Outcome-Dependent Query Costs**
S. Sabato, A. Sarwate, N. Srebro

T29 **Buy-in-Bulk Active Learning**
L. Yang, J. Carbonell

T30 **Active Learning for Probabilistic Hypotheses Using the Maximum Gibbs Error Criterion**
N. Cuong, W. Lee, N. Ye, K. Chai, H. Chieu

T31 **Marginals-to-Models Reducibility**
T. Roughgarden, M. Kearns

T32 Learning Chordal Markov Networks by Constraint Satisfaction
J. Corander, T. Janhunen, J. Rintanen, H. Nyman, J. Pensar

T33 Bayesian Estimation of Latently-grouped Parameters in Undirected Graphical Models
J. Liu, D. Page

T34 On Sampling from the Gibbs Distribution with Random Maximum A-Posteriori Perturbations
T. Hazan, S. Maji, T. Jaakkola

T35 EDML for Learning Parameters in Directed and Undirected Graphical Models
K. Refaat, A. Choi, A. Darwiche

T36 Projecting Ising Model Parameters for Fast Mixing
J. Domke, X. Liu

T37 Embed and Project: Discrete Sampling with Universal Hashing
S. Ermon, C. Gomes, A. Sabharwal, B. Selman

T38 Learning Stochastic Inverses
A. Stuhlmüller, J. Taylor, N. Goodman

T39 Approximate Gaussian process inference for the drift function in stochastic differential equations
A. Ruttor, P. Batz, M. Opper

T40 Online Learning of Nonparametric Mixture Models via Sequential Variational Approximation
D. Lin

T41 Memoized Online Variational Inference for Dirichlet Process Mixture Models
M. Hughes, E. Sudderth

T42 Regret based Robust Solutions for Uncertain Markov Decision Processes
A. Ahmed, P. Varakantham, Y. Adulyasak, P. Jaillet

T43 Improved and Generalized Upper Bounds on the Complexity of Policy Iteration
B. Scherrer

T44 Efficient Exploration and Value Function Generalization in Deterministic Systems
Z. Wen, B. Van Roy

T45 Aggregating Optimistic Planning Trees for Solving Markov Decision Processes
G. Kedenburg, R. Fonteneau, R. Munos

T46 Online learning in episodic Markovian decision processes by relative entropy policy search
A. Zimin, G. Neu

T47 Online Learning in Markov Decision Processes with Adversarially Chosen Transition Probability Distributions
Y. Abbasi, P. Bartlett, V. Kanade, Y. Seldin, C. Szepesvari

T48 Online Learning of Dynamic Parameters in Social Networks
S. Shahrampour, S. Rakhlin, A. Jadbabaie

T49 Modeling Overlapping Communities with Node Popularities
P. Gopalan, C. Wang, D. Blei

T50 A Scalable Approach to Probabilistic Latent Space Inference of Large-Scale Networks
J. Yin, Q. Ho, E. Xing

T51 Relevance Topic Model for Unstructured Social Group Activity Recognition
F. Zhao, Y. Huang, L. Wang, T. Tan

T52 Streaming Variational Bayes
T. Broderick, N. Boyd, A. Wibisono, A. Wilson, M. Jordan

T53 Scalable Inference for Logistic-Normal Topic Models
J. Chen, J. Zhu, Z. Wang, X. Zheng, B. Zhang

T54 When are Overcomplete Topic Models Identifiable? Uniqueness of Tensor Tucker Decompositions with Structured Sparsity
A. Anandkumar, D. Hsu, M. Janzamin, S. Kakade

T55 Minimax Theory for High-dimensional Gaussian Mixtures with Sparse Mean Separation
M. Azizyan, A. Singh, L. Wasserman

T56 Cluster Trees on Manifolds
S. Balakrishnan, S. Narayanan, A. Rinaldo, A. Singh, L. Wasserman

T57 Convex Tensor Decomposition via Structured Schatten Norm Regularization
R. Tomioka, T. Suzuki

T58 Convex Relaxations for Permutation Problems
F. Fogel, R. Jenatton, F. Bach, A. D’Aspremont

T59 Solving the multi-way matching problem by permutation synchronization
D. Pachauri, R. Kondor, V. Singh

T60 Reflection methods for user-friendly submodular optimization
S. Jegelka, F. Bach, S. Sra

T61 Curvature and Optimal Algorithms for Learning and Minimizing Submodular Functions
R. Iyer, S. Jegelka, J. Bilmes

T62 An Approximate, Efficient LP Solver for LP Rounding
S. Sridhar, S. Wright, C. Re, J. Liu, V. Bittorf, C. Zhang

T63 Hierarchical Modular Optimization of Convolutional Networks Achieves Representations Similar to Macaque IT and Human Ventral Stream
D. Yamins, H. Hong, C. Cadieu, J. DiCarlo

T64 Bayesian inference for low rank spatiotemporal neural receptive fields
M. Park, J. Pillow

T65 Spectral methods for neural characterization using generalized quadratic models
I. Park, E. Archer, N. Priebe, J. Pillow

T66 Optimal Neural Population Codes for High-dimensional Stimulus Variables
Z. Wang, A. Stocker, D. Lee

T67 Robust learning of low-dimensional dynamics from large neural ensembles
D. Pfau, E. Pnevmatikakis, L. Paninski

T68 Sparse nonnegative deconvolution for compressive calcium imaging: algorithms and phase transitions
E. Pnevmatikakis, L. Paninski

T69 Generalized Method-of-Moments for Rank Aggregation
H. Azari Soufiani, W. Chen, D. Parkes, L. Xia

T70 Generalized Random Utility Models with Multiple Types
H. Azari Soufiani, H. Diao, Z. Lai, D. Parkes

T71 Speedup Matrix Completion with Side Information: Application to Multi-Label Learning
M. Xu, R. Jin, Z. Zhou

T72 Correlated random features for fast semi-supervised learning
B. McWilliams, D. Balduzzi, J. Buhmann

T73 Manifold-based Similarity Adaptation for Label Propagation
M. Karasuyama, H. Mamitsuka

T74 Supervised Sparse Analysis and Synthesis Operators
P. Sprechmann, R. Litman, T. Ben Yakar, A. Bronstein, G. Sapiro

T75 When in Doubt, SWAP: High-Dimensional Sparse Recovery from Correlated Measurements
D. Vats, R. Baraniuk

T76 Deep content-based music recommendation
A. van den Oord, S. Dieleman, B. Schrauwen

T77 Probabilistic Low-Rank Matrix Completion with Adaptive Spectral Regularization Algorithms
A. Todeschini, F. Caron, M. Chavent

T78 A Gang of Bandits
N. Cesa-Bianchi, C. Gentile, G. Zappella

T79 Contrastive Learning Using Spectral Methods
J. Zou, D. Hsu, D. Parkes, R. Adams

T80 Fast Determinantal Point Process Sampling with Application to Clustering
B. Kang

T81 Computing the Stationary Distribution Locally
C. Lee, A. Ozdaglar, D. Shah

T82 Learning Prices for Repeated Auctions with Strategic Buyers
K. Amin, A. Rostamizadeh, U. Syed

T83 Efficient Algorithm for Privately Releasing Smooth Queries
Z. Wang, K. Fan, J. Zhang, L. Wang

T84 (Nearly) Optimal Algorithms for Private Online Learning in Full-information and Bandit Settings
A. Guha Thakurta, A. Smith

T85 Local Privacy and Minimax Bounds: Sharp Rates for Probability Estimation
J. Duchi, M. Wainwright, M. Jordan

T86 A Stability-based Validation Procedure for Differentially Private Machine Learning
K. Chaudhuri, S. Vinterbo

T87 Similarity Component Analysis
S. Changpinyo, K. Liu, F. Sha

T88 A message-passing algorithm for multi-agent trajectory planning
J. Bento, N. Derbinsky, J. Alonso-Mora, J. Yedidia

T89 The Power of Asymmetry in Binary Hashing
B. Neyshabur, N. Srebro, R. Salakhutdinov, Y. Makarychev, P. Yadollahpour

T90 Learning to Prune in Metric and Non-Metric Spaces
L. Boytsov, B. Naidan

T1 Learning and using language via recursive pragmatic reasoning about other agents

Nathaniel Smith nathaniel.smith@ed.ac.uk
University of Edinburgh
Noah Goodman ngoodman@stanford.edu
Michael Frank mcfrank@stanford.edu
Stanford University

Language users are remarkably good at making inferences about speakers' intentions in context, and children learning their native language also display substantial skill in acquiring the meanings of unknown words. These two cases are deeply related: Language users invent new terms in conversation, and language learners learn the literal meanings of words based on their pragmatic inferences about how those words are used. While pragmatic inference and word learning have both been independently characterized in probabilistic terms, no current work unifies these two. We describe a model in which language learners assume that they jointly approximate a shared, external lexicon and reason recursively about the goals of others in using this lexicon. This model captures phenomena in word learning and pragmatic inference; it additionally leads to insights about the emergence of communicative systems in conversation and the mechanisms by which pragmatic inferences become incorporated into word meanings.

T2 Model Selection for High-Dimensional Regression under the Generalized Irrepresentability Condition

Adel Javanmard adelj@stanford.edu
Andrea Montanari montanari@stanford.edu
Stanford University

In the high-dimensional regression model a response variable is linearly related to $p$covariates, but the sample size n is smaller than p. We assume that only a small subset of covariates is `active' (i.e., the corresponding coefficients are non-zero), and consider the model-selection problem of identifying the active covariates. A popular approach is to estimate the regression coefficients through the Lasso ($\ell 1$-regularized least squares). This is known to correctly identify the active set only if the irrelevant covariates are roughly orthogonal to the relevant ones, as quantified through the so called `irrepresentability' condition. In this paper we study the `Gauss-Lasso' selector, a simple two-stage method that first solves the Lasso, and then performs ordinary least squares restricted to the Lasso active set. We formulate `generalized irrepresentability condition' (GIC), an assumption that is substantially weaker than irrepresentability. We prove that, under GIC, the Gauss-Lasso correctly recovers the active set.

T3 Confidence Intervals and Hypothesis Testing for High-Dimensional Statistical Models

Adel Javanmard adelj@stanford.edu
Andrea Montanari montanari@stanford.edu
Stanford University

Fitting high-dimensional statistical models often requires the use of non-linear parameter estimation procedures. As a consequence, it is generally impossible to obtain an exact characterization of the probability distribution of the parameter estimates. This in turn implies that it is extremely challenging to quantify the `uncertainty' associated with a certain parameter estimate. Concretely, no commonly accepted procedure exists for computing classical measures of uncertainty and statistical significance as confidence intervals or p-values. We consider here a broad class of regression problems, and propose an efficient algorithm for constructing confidence intervals and p-values. The resulting confidence intervals have nearly optimal size. When testing for the null hypothesis that a certain parameter is vanishing, our method has nearly optimal power. Our approach is based on constructing a `de-biased' version of regularized M-estimators. The new construction improves over recent work in the field in that it does not assume a special structure on the design matrix. Furthermore, proofs are remarkably simple. We test our method on a diabetes prediction problem.

T4 Compressive Feature Learning

Hristo Paskov hpaskov@stanford.edu
Bob West west@cs.stanford.edu
John Mitchell mitchell@cs.stanford.edu
Trevor Hastie hastie@stanford.edu
Stanford University

This paper addresses the problem of unsupervised feature learning for text data. Our method is grounded in the principle of minimum description length and uses a dictionary-based compression scheme to extract a succinct feature set. Specifically, our method finds a set of word k-grams that minimizes the cost of reconstructing the text losslessly. We formulate document compression as a binary optimization task and show how to solve it approximately via a sequence of reweighted linear programs that are efficient to solve and parallelizable. As our method is unsupervised, features may be extracted once and subsequently used in a variety of tasks. We demonstrate the performance of these features over a range of scenarios including unsupervised exploratory analysis and supervised text categorization. Our compressed feature space is two orders of magnitude smaller than the full k-gram space and matches the text categorization accuracy achieved in the full feature space. This dimensionality reduction not only results in faster training times, but it can also help elucidate structure in unsupervised learning tasks and reduce the amount of training data necessary for supervised learning.

T5 Pass-efficient unsupervised feature selection

Crystal Maung Crystal.Maung@gmail.com
Haim Schweitzer HSchweitzer@utdallas.edu
UT Dallas

The goal of unsupervised feature selection is to identify a small number of important features that can represent the data. We propose a new algorithm, a modification of the classical pivoted QR algorithm of Businger and Golub, that requires a small number of passes over the data. The improvements are based on two ideas: keeping track of multiple features in each pass, and skipping calculations that can be shown not to affect the final selection. Our algorithm selects the exact same features as the classical pivoted QR algorithm, and has the same favorable numerical stability. We describe experiments on real-world datasets which sometimes show improvements of several orders of magnitude} over the classical algorithm. These results appear to be competitive with recently proposed randomized algorithms in terms of pass efficiency and run time. On the other hand, the randomized algorithms may produce better features, at the cost of small probability of failure.

T6 Better Approximation and Faster Algorithm Using the Proximal Average

Yao-Liang Yu yaoliang@cs.ualberta.ca
University of Alberta

It is a common practice to approximate “complicated” functions with more friendly ones. In large-scale machine learning applications, nonsmooth losses/regularizers that entail great computational challenges are usually approximated by smooth functions. We re-examine this powerful methodology and point out a nonsmooth approximation which simply pretends the linearity of the proximal map. The new approximation is justified using a recent convex analysis tool---proximal average, and yields a novel proximal gradient algorithm that is strictly better than the one based on smoothing, without incurring any extra overhead. Numerical experiments conducted on two important applications, overlapping group lasso and graph-guided fused lasso, corroborate the theoretical claims.

T7 Polar Operators for Structured Sparse Estimation

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

Structured sparse estimation has become an important technique in many areas of data analysis. Unfortunately, these estimators normally create computational difficulties that entail sophisticated algorithms. Our first contribution is to uncover a rich class of structured sparse regularizers whose polar operator can be evaluated efficiently. With such an operator, a simple conditional gradient method can then be developed that, when combined with smoothing and local optimization, significantly reduces training time vs. the state of the art. We also demonstrate a new reduction of polar to proximal maps that enables more efficient latent fused lasso.

T8 On the Linear Convergence of the Proximal Gradient Method for Trace Norm Regularization

Ke Hou khou@se.cuhk.edu.hk
Zirui Zhou zrzhou@se.cuhk.edu.hk
Anthony Man-Cho So manchoso@se.cuhk.edu.hk
CUHK
Zhi-Quan Luo luozq@ece.umn.edu
University of Minnesota, Twin Cites

Motivated by various applications in machine learning, the problem of minimizing a convex smooth loss function with trace norm regularization has received much attention lately. Currently, a popular method for solving such problem is the proximal gradient method (PGM), which is known to have a sublinear rate of convergence. In this paper, we show that for a large class of loss functions, the convergence rate of the PGM is in fact linear. Our result is established without any strong convexity assumption on the loss function. A key ingredient in our proof is a new Lipschitzian error bound for the aforementioned trace norm-regularized problem, which may be of independent interest.

T9 Accelerating Stochastic Gradient Descent using Predictive Variance Reduction

Rie Johnson riejohnson@gmail.com
RJ Research Consulting
Tong Zhang tzhang@stat.rutgers.edu
Baidu & Rutgers

Stochastic gradient descent is popular for large scale optimization but has slow convergence asymptotically due to the inherent variance. To remedy this problem, we introduce an explicit variance reduction method for stochastic gradient descent which we call stochastic variance reduced gradient (SVRG). For smooth and strongly convex functions, we prove that this method enjoys the same fast convergence rate as those of stochastic dual coordinate ascent (SDCA) and Stochastic Average Gradient (SAG). However, our analysis is significantly simpler and more intuitive. Moreover, unlike SDCA or SAG, our method does not require the storage of gradients, and Ts is more easily applicable to complex problems such as some structured prediction problems and neural network learning.

T10 Accelerated Mini-Batch Stochastic Dual Coordinate Ascent

Shai Shalev-Shwartz shai.shwartz@gmail.com
The Hebrew University
Tong Zhang tzhang@stat.rutgers.edu
Baidu & Rutgers

Stochastic dual coordinate ascent (SDCA) is an effective technique for solving regularized loss minimization problems in machine learning. This paper considers an extension of SDCA under the mini-batch setting that is often used in practice. Our main contribution is to introduce an accelerated mini-batch version of SDCA and prove a fast convergence rate for this method. We discuss an implementation of our method over a parallel computing system, and compare the results to both the vanilla stochastic dual coordinate ascent and to the accelerated deterministic gradient descent method of Nesterov [2007].

T11 Estimation, Optimization, and Parallelism when Data is Sparse

John Duchi jduchi@eecs.berkeley.edu
Michael Jordan jordan@cs.berkeley.edu
UC Berkeley
Brendan McMahan mcmahan@google.com
Google Research

We study stochastic optimization problems when the *data* is sparse, which is in a sense dual to the current understanding of high-dimensional statistical learning and optimization. We highlight both the difficulties---in terms of increased sample complexity that sparse data necessitates---and the potential benefits, in terms of allowing parallelism and asynchrony in the design of algorithms. Concretely, we derive matching upper and lower bounds on the minimax rate for optimization and learning with sparse data, and we exhibit algorithms achieving these rates. Our algorithms are adaptive: they achieve the best possible rate for the data observed. We also show how leveraging sparsity leads to (still minimax optimal) parallel and asynchronous algorithms, providing experimental evidence complementing our theoretical results on medium to large-scale learning tasks.

T12 Linear Convergence with Condition Number Independent Access of Full Gradients

Lijun Zhang zljzju@gmail.com
Mehrdad Mahdavi mahdavim@msu.edu
Rong Jin rong+@cs.cmu.edu
Michigan State University (MSU)

For smooth and strongly convex optimization, the optimal iteration complexity of the gradient-based algorithm is $O(\sqrt{}\kappa$ x $\log 1/\epsilon)$, where κ is the conditional number. In the case that the optimization problem is ill-conditioned, we need to evaluate a larger number of full gradients, which could be computationally expensive. In this paper, we propose to reduce the number of full gradient required by allowing the algorithm to access the stochastic gradients of the objective function. To this end, we present a novel algorithm named Epoch Mixed Gradient Descent (EMGD) that is able to utilize two kinds of gradients. A distinctive step in EMGD is the mixed gradient descent, where we use an combination of the gradient and the stochastic gradient to update the intermediate solutions. By performing a fixed number of mixed gradient descents, we are able to improve the sub-optimality of the solution by a constant factor, and Ts achieve a linear convergence rate. Theoretical analysis shows that EMGD is able to find an ϵ-optimal solution by computing $O(\log 1/\epsilon)$ full gradients and $O(\kappa_2 \log 1/\epsilon)$ stochastic gradients.

T13 Mixed Optimization for Smooth Functions

Mehrdad Mahdavi mahdavim@msu.edu
Lijun Zhang zljzju@gmail.com
Rong Jin rong+@cs.cmu.edu
Michigan State University (MSU)

It is well known that the optimal convergence rate for stochastic optimization of smooth functions is $[O(1/\sqrt{T})]$, which is same as stochastic optimization of Lipschitz continuous convex functions. This is in contrast to optimizing smooth functions using full gradients, which yields a convergence rate of $[O(1/T^2)]$. In this work, we consider a new setup for optimizing smooth functions, termed as **Mixed Optimization**, which allows to access both a stochastic oracle and a full gradient oracle. Our goal is to significantly improve the convergence rate of stochastic optimization of smooth functions by having an additional small number of accesses to the full gradient oracle. We show that, with an $[O(\ln T)]$ calls to the full gradient oracle and an $O(T)$ calls to the stochastic oracle, the proposed mixed optimization algorithm is able to achieve an optimization error of $[O(1/T)]$.

T14 Stochastic Convex Optimization with Multiple Objectives

Mehrdad Mahdavi mahdavim@msu.edu
Rong Jin rong+@cs.cmu.edu
Michigan State University (MSU)
Tianbao Yang tyang@ge.com
NEC Labs America

In this paper, we are interested in the development of efficient algorithms for convex optimization problems in the simultaneous presence of multiple objectives and stochasticity in the first-order information. We cast the stochastic multiple objective optimization problem into a constrained optimization problem by choosing one function as the objective and try to bound other objectives by appropriate thresholds. We first examine a two stages exploration-exploitation based algorithm which first approximates the stochastic objectives by sampling and then solves a constrained stochastic optimization problem by projected gradient method. This method attains a suboptimal convergence rate even under strong assumption on the objectives. Our second approach is an efficient primal-dual stochastic algorithm. It leverages on the theory of Lagrangian method in constrained optimization and attains the optimal convergence rate of $[O(1/\sqrt{T})]$ in high probability for general Lipschitz continuous objectives.

T15 Data-driven Distributionally Robust Polynomial Optimization

Martin Mevissen martmevi@ie.ibm.com
Emanuele Ragnoli eragnoli@ie.ibm.com
Jia Yuan Yu jiayuanyu@ie.ibm.com
IBM Research

We consider robust optimization for polynomial optimization problems where the uncertainty set is a set of candidate probability density functions. This set is a ball around a density function estimated from data samples, i.e., it is data-driven and random. Polynomial optimization problems are inherently hard due to nonconvex objectives and constraints. However, we show that by employing polynomial and histogram density estimates, we can introduce robustness with respect to distributional uncertainty sets without making the problem harder. We show that the solution to the distributionally robust problem is the limit of a sequence of tractable semidefinite programming relaxations. We also give finite-sample consistency guarantees for the data-driven uncertainty sets. Finally, we apply our model and solution method in a water network problem.

T16 Multiscale Dictionary Learning for Estimating Conditional Distributions

Francesca Petralia fp12@duke.edu
Mt Sinai School of Medicine
jovo Vogelstein jv.work@jhu.edu
David Dunson dunson@stat.duke.edu
Duke University

Nonparametric estimation of the conditional distribution of a response given high-dimensional features is a challenging problem. It is important to allow not only the mean but also the variance and shape of the response density to change flexibly with features, which are massive-dimensional. We propose a multiscale dictionary learning model, which expresses the conditional response density as a convex combination of dictionary densities, with the densities used and their weights dependent on the path through a tree decomposition of the feature space. A fast graph partitioning algorithm is applied to obtain the tree decomposition, with Bayesian methods then used to adaptively prune and average over different sub-trees in a soft probabilistic manner. The algorithm scales efficiently to approximately one million features. State of the art predictive performance is demonstrated for toy examples and two neuroscience applications including up to a million features.

T17 On the Sample Complexity of Subspace Learning

Alessandro Rudi alessandro.rudi@iit.it
Istituto Italiano di Tecnologia
Guille Canas guilledc@MIT.EDU
Lorenzo Rosasco lrosasco@mit.edu
Massachusetts Institute of Technology

A large number of algorithms in machine learning, from principal component analysis (PCA), and its non-linear (kernel) extensions, to more recent spectral embedding and support estimation methods, rely on estimating a linear subspace from samples. In this paper we introduce a general formulation of this problem and derive novel learning error estimates. Our results rely on natural assumptions on the spectral properties of the covariance operator associated to the data distribution, and hold for a wide class of metrics between subspaces. As special cases, we discuss sharp error estimates for the reconstruction properties of PCA and spectral support estimation. Key to our analysis is an operator theoretic approach that has broad applicability to spectral learning methods.

T18 Least Informative Dimensions

Fabian Sinz fabee@epagoge.de
January Grewe jan.grewe@uni-tuebingen.de
January Benda jan.benda@uni-tuebingen.de
Universität Tübingen
Anna Stockl Anna.Stockl@biol.lu.se
Lund University, Sweden

We present a novel non-parametric method for finding a subspace of stimulus features that contains all information about the response of a system. Our method generalizes similar approaches to this problem such as spike triggered average, spike triggered covariance, or maximally informative dimensions. Instead of maximizing the mutual information between features and responses directly, we use integral probability metrics in kernel Hilbert spaces to minimize the information between uninformative features and the combination of informative features and responses. Since estimators of these metrics access the data via kernels, are easy to compute, and exhibit good theoretical convergence properties, our method can easily be generalized to populations of neurons or spike patterns. By using a particular expansion of the mutual information, we can show that the informative features must contain all information if we can make the uninformative features independent of the rest.

T19 Blind Calibration in Compressed Sensing using Message Passing Algorithms

Christophe Schulke christophe.schulke@espci.fr
ESPCI ParisTech
Francesco Caltagirone f.calta@gmail.com
IPhT, CEA Saclay
Florent Krzakala florent.krzakala@gmail.com
École Normale Supérieure
Lenka Zdeborova lenka.zdeborova@cea.fr
CEA Saclay and CNRS URA 2306

Compressed sensing (CS) is a concept that allows to acquire compressible signals with a small number of measurements. As such, it is very attractive for hardware implementations. Therefore, correct calibration of the hardware is a central issue. In this paper we study the so-called blind calibration, i.e. when the training signals that are available to perform the calibration are sparse but unknown. We extend the approximate message passing (AMP) algorithm used in CS to the case of blind calibration. In the calibration-AMP, both the gains on the sensors and the elements of the signals are treated as unknowns. Our algorithm is also applicable to settings in which the sensors distort the measurements in other ways than multiplication by a gain, unlike previously suggested blind calibration algorithms based on convex relaxations. We study numerically the phase diagram of the blind calibration problem, and show that even in cases where convex relaxation is possible, our algorithm requires a smaller number of measurements and/or signals in order to perform well.

T20 Estimating LASSO Risk and Noise Level

Mohsen Bayati bayati@stanford.edu
Murat A. Erdogdu erdogdu@stanford.edu
Andrea Montanari montanari@stanford.edu
Stanford University

We study the fundamental problems of variance and risk estimation in high dimensional statistical modeling. In particular, we consider the problem of learning a coefficient vector $\theta_0 \in R^p$ from noisy linear observation $y = X\theta_0 + w \in R^n$ and the popular estimation procedure of solving an ℓ_1-penalized least squares objective known as the LASSO or Basis Pursuit DeNoising (BPDN). In this context, we develop new estimators for the ℓ_2 estimation risk $\|\theta\hat{} - \theta 0\|_2$ and the variance of the noise. These can be used to select the regularization parameter optimally. Our approach combines Stein unbiased risk estimate (Stein'81) and recent results of (Bayati and Montanari'11-12) on the analysis of approximate message passing and risk of LASSO. We establish high-dimensional consistency of our estimators for sequences of matrices X of increasing dimensions, with independent Gaussian entries. We establish validity for a broader class of Gaussian designs, conditional on the validity of a certain conjecture from statistical physics. Our approach is the first that provides an asymptotically consistent risk estimator. In addition, we demonstrate through simulation that our variance estimation outperforms several existing methods in the literature.

T21 A Graphical Transformation for Belief Propagation: Maximum Weight Matchings and Odd-Sized Cycles

Jinwoo Shin jinwoos@kaist.ac.kr
KAIST
Andrew Gelfand agelfand@ics.uci.edu
UC Irvine
Misha Chertkov m.chertkov@gmail.com
Los Alamos National Laboratory

Max-product 'belief propagation' (BP) is a popular distributed heuristic for finding the Maximum A Posteriori (MAP) assignment in a joint probability distribution represented by a Graphical Model (GM). It was recently shown that BP converges to the correct MAP assignment for a class of loopy GMs with the following common feature: the Linear Programming (LP) relaxation to the MAP problem is tight (has no integrality gap). Unfortunately, tightness of the LP relaxation does not, in general, guarantee convergence and correctness of the BP algorithm. The failure of BP in such cases motivates reverse engineering a solution – namely, given a tight LP, can we design a 'good' BP algorithm. In this paper, we design a BP algorithm for the Maximum Weight Matching (MWM) problem over general graphs. We prove that the algorithm converges to the correct optimum if the respective LP relaxation, which may include inequalities associated with non-intersecting odd-sized cycles, is tight. The most significant part of our approach is the introduction of a novel graph transformation designed to force convergence of BP. Our theoretical result suggests an efficient BP-based heuristic for the MWM problem, which consists of making sequential, "cutting plane", modifications to the underlying GM. Our experiments show that this heuristic performs as well as traditional cutting-plane algorithms using LP solvers on MWM problems.

T22 Sensor Selection in High-Dimensional Gaussian Trees with Nuisances

Daniel Levine denu@mit.edu
Jonathan How jhow@mit.edu
Massachusetts Institute of Technology

We consider the sensor selection problem on multivariate Gaussian distributions where only a subset of latent variables is of inferential interest. For pairs of vertices connected by a unique path in the graph, we show that there exist decompositions of nonlocal mutual information into local information measures that can be computed efficiently from the output of message passing algorithms. We integrate these decompositions into a computationally efficient greedy selector where the computational expense of quantification can be distributed across nodes in the network. Experimental results demonstrate the comparative efficiency of our algorithms for sensor selection in high-dimensional distributions. We additionally derive an online-computable performance bound based on augmentations of the relevant latent variable set that, when such a valid augmentation exists, is applicable for any distribution with nuisances.

T23 Σ-Optimality for Active Learning on Gaussian Random Fields

Yifei Ma yifeim@cs.cmu.edu
Jeff Schneider schneide@cs.cmu.edu
CMU
Roman Garnett rgarnett@uni-bonn.de
University of Bonn

A common classifier for unlabeled nodes on undirected graphs uses label propagation from the labeled nodes, equivalent to the harmonic predictor on Gaussian random fields (GRFs). For active learning on GRFs, the commonly used V-optimality criterion queries nodes that reduce the L2 (regression) loss. V-optimality satisfies a submodularity property showing that greedy reduction produces a (1 − 1/e) globally optimal solution. However, L2 loss may not characterise the true nature of 0/1 loss in classification problems and Ts may not be the best choice for active learning. We consider a new criterion we call Σ-optimality, which queries the node that minimizes the sum of the elements in the predictive covariance. Σ-optimality directly optimizes the risk of the surveying problem, which is to determine the proportion of nodes belonging to one class. In this paper we extend submodularity guarantees from V-optimality to Σ-optimality using properties specific to GRFs. We further show that GRFs satisfy the suppressor-free condition in addition to the conditional independence inherited from Markov random fields. We test Σ-optimality on real-world graphs with both synthetic and real data and show that it outperforms V-optimality and other related methods on classification.

T24 Bayesian optimization explains human active search

Ali Borji borji@usc.edu
Laurent Itti itti@usc.edu
University of Southern California (USC)

Many real-world problems have complicated objective functions. To optimize such functions, humans utilize sophisticated sequential decision-making strategies. Many optimization algorithms have also been developed for this same purpose, but how do they compare to humans in terms of both performance and behavior? We try to unravel the general underlying algorithm people may be using while searching for the maximum of an invisible 1D function. Subjects click on a blank screen and are shown the ordinate of the function at each clicked abscissa location. Their task is to find the function's maximum in as few clicks as possible. Subjects win if they get close enough to the maximum location. Analysis over 23 non-maths undergraduates, optimizing 25 functions from different families, shows that humans outperform 24 well-known optimization algorithms. Bayesian Optimization based on Gaussian Processes, which exploit all the x values tried and all the f(x) values obtained so far to pick the next x, predicts human performance and searched locations better. In 6 follow-up controlled experiments over 76 subjects, covering interpolation, extrapolation, and optimization tasks, we further confirm that Gaussian Processes provide a general and unified theoretical account to explain passive and active function learning and search in humans.

T25 Latent Structured Active Learning

Wenjie Luo wenjie.luo@ttic.edu
Raquel Urtasun rurtasun@ttic.edu
TTI Chicago
Alex Schwing aschwing@inf.ethz.ch
ETH Zurich

In this paper we present active learning algorithms in the context of structured prediction problems. To reduce the amount of labeling necessary to learn good models, our algorithms only label subsets of the output. To this end, we query examples using entropies of local marginals, which are a good surrogate for uncertainty. We demonstrate the effectiveness of our approach in the task of 3D layout prediction from single images, and show that good models are learned when labeling only a handful of random variables. In particular, the same performance as using the full training set can be obtained while only labeling 10\% of the random variables.

T26 Low-Rank Matrix and Tensor Completion via Adaptive Sampling

Akshay Krishnamurthy akshaykr@cs.cmu.edu
Aarti Singh aartisingh@cmu.edu
CMU

We study low rank matrix and tensor completion and propose novel algorithms that employ adaptive sampling schemes to obtain strong performance guarantees for these problems. Our algorithms exploit adaptivity to identify entries that are highly informative for identifying the column space of the matrix (tensor) and consequently, our results hold even when the row space is highly coherent, in contrast with previous analysis of matrix completion. In the absence of noise, we show that one can exactly recover a $n \times n$ matrix of rank r using $O(r^2 n \log(r))$ observations, which is better than the best known bound under random sampling. We also show that one can recover an order T tensor using $O(r^2(T-1)T^2 n \log(r))$. For noisy recovery, we show that one can consistently estimate a low rank matrix corrupted with noise using $O(nr \mathrm{polylog}(n))$ observations. We complement our study with simulations that verify our theoretical guarantees and demonstrate the scalability of our algorithms.

T27 Adaptive Submodular Maximization in Bandit Setting

Victor Gabillon victor.gabillon@inria.fr
INRIA
Branislav Kveton branislav.kveton@technicolor.com
Brian Eriksson brian.eriksson@technicolor.com
Technicolor Labs
Zheng Wen zhengwen@stanford.edu
Stanford University
S. MuTkrishnan muT@cs.rutgers.edu
Rutgers University

Maximization of submodular functions has wide applications in machine learning and artificial intelligence. Adaptive submodular maximization has been traditionally studied under the assumption that the model of the world, the expected gain of choosing an item given previously selected items and their states, is known. In this paper, we study the scenario where the expected gain is initially unknown and it is learned by interacting repeatedly with the optimized function. We propose an efficient algorithm for solving our problem and prove that its expected cumulative regret increases logarithmically with time. Our regret bound captures the inherent property of submodular maximization, earlier mistakes are more costly than later ones. We refer to our approach as Optimistic Adaptive Submodular Maximization (OASM) because it trades off exploration and exploitation based on the optimism in the face of uncertainty principle. We evaluate our method on a preference elicitation problem and show that non-trivial K-step policies can be learned from just a few hundred interactions with the problem.

T28 Auditing: Active Learning with Outcome-Dependent Query Costs

Sivan Sabato sivan.sabato@gmail.com
Microsoft Research
Anand Sarwate asarwate@ttic.edu
Nati Srebro nati@ttic.edu
TTI Chicago

We propose a learning setting in which unlabeled data is free, and the cost of a label depends on its value, which is not known in advance. We study binary classification in an extreme case, where the algorithm only pays for negative labels. Our motivation are applications such as fraud detection, in which investigating an honest transaction should be avoided if possible. We term the setting auditing, and consider the auditing complexity of an algorithm: The number of negative points it labels to learn a hypothesis with low relative error. We design auditing algorithms for thresholds on the line and axis-aligned rectangles, and show that with these algorithms, the auditing complexity can be significantly lower than the active label complexity. We discuss a general approach for auditing for a general hypothesis class, and describe several interesting directions for future work.

T29 Buy-in-Bulk Active Learning

Liu Yang liuy@cs.cmu.edu
Jaime Carbonell jgc@cs.cmu.edu
CMU

In many practical applications of active learning, it is more cost-effective to request labels in large batches, rather than one-at-a-time. This is because the cost of labeling a large batch of examples at once is often sublinear in the number of examples in the batch. In this work, we study the label complexity of active learning algorithms that request labels in a given number of batches, as well as the tradeoff between the total number of queries and the number of rounds allowed. We additionally study the total cost sufficient for learning, for an abstract notion of the cost of requesting the labels of a given number of examples at once. In particular, we find that for sublinear cost functions, it is often desirable to request labels in large batches (i.e., buying in bulk); although this may increase the total number of labels requested, it reduces the total cost required for learning.

T30 Active Learning for Probabilistic Hypotheses Using the Maximum Gibbs Error Criterion

Nguyen Viet Cuong cuongnv@nus.edu.sg
Wee Sun Lee leews@comp.nus.edu.sg
Nan Ye yenan@comp.nus.edu.sg
National University of Singapore
Kian Ming Chai ckianmin@dso.org.sg
Hai Leong Chieu chaileon@dso.org.sg
DSO National Laboratories

We introduce a new objective function for pool-based Bayesian active learning with probabilistic hypotheses. This objective function, called the policy Gibbs error, is the expected error rate of a random classifier drawn from the prior distribution on the examples adaptively selected by the active learning policy. Exact maximization of the policy Gibbs error is hard, so we propose a greedy strategy that maximizes the Gibbs error at each iteration, where the Gibbs error on an instance is the expected error of a random classifier selected from the posterior label distribution on that instance. We apply this maximum Gibbs error criterion to three active learning scenarios: non-adaptive, adaptive, and batch active learning. In each scenario, we prove that the criterion achieves near-maximal policy Gibbs error when constrained to a fixed budget. For practical implementations, we provide approximations to the maximum Gibbs error criterion for Bayesian conditional random fields and transductive Naive Bayes. Our experimental results on a named entity recognition task and a text classification task show that the maximum Gibbs error criterion is an effective active learning criterion for noisy models.

T31 Marginals-to-Models Reducibility

Tim Roughgarden tim@cs.stanford.edu
Stanford University
Michael Kearns mkearns@cis.upenn.edu
University of Pennsylvania

We consider a number of classical and new computational problems regarding marginal distributions, and inference in models specifying a full joint distribution. We prove general and efficient reductions between a number of these problems, which demonstrate that algorithmic progress in inference automatically yields progress for “pure data” problems. Our main technique involves formulating the problems as linear programs, and proving that the dual separation oracle for the Ellipsoid Method is provided by the target problem. This technique may be of independent interest in probabilistic inference.

T32 Learning Chordal Markov Networks by Constraint Satisfaction

Jukka Corander jukka.corander@helsinki.fi
University of Helsinki
Tomi Janhunen tomi.janhunen@aalto.fi
Jussi Rintanen jussi.rintanen@aalto.fi
Aalto University
Henrik Nyman henrik.j.nyman@abo.fi
Johan Pensar jopensar@abo.fi
Åbo Akademi

We investigate the problem of learning the structure of a Markov network from data. It is shown that the structure of such networks can be described in terms of constraints which enables the use of existing solver technology with optimization capabilities to compute optimal networks starting from initial scores computed from the data. To achieve efficient encodings, we develop a novel characterization of Markov network structure using a balancing condition on the separators between cliques forming the network. The resulting translations into propositional satisfiability and its extensions such as maximum satisfiability, satisfiability modulo theories, and answer set programming, enable us to prove the optimality of networks which have been previously found by stochastic search.

T33 Bayesian Estimation of Latently-grouped Parameters in Undirected Graphical Models

Jie Liu jieliu@cs.wisc.edu
David Page page@biostat.wisc.edu
UW-Madison

In large-scale applications of undirected graphical models, such as social networks and biological networks, similar patterns occur frequently and give rise to similar parameters. In this situation, it is beneficial to group the parameters for more efficient learning. We show that even when the grouping is unknown, we can infer these parameter groups during learning via a Bayesian approach. We impose a Dirichlet process prior on the parameters. Posterior inference usually involves calculating intractable terms, and we propose two approximation algorithms, namely a Metropolis-Hastings algorithm with auxiliary variables and a Gibbs sampling algorithm with stripped Beta approximation (Gibbs_SBA). Simulations show that both algorithms outperform conventional maximum likelihood estimation (MLE). Gibbs_SBA's performance is close to Gibbs sampling with exact likelihood calculation. Models learned with Gibbs_SBA also generalize better than the models learned by MLE on real-world Senate voting data.

T34 On Sampling from the Gibbs Distribution with Random Maximum A-Posteriori Perturbations

Tamir Hazan tamir@cs.haifa.ac.il
University of Haifa
Subhransu Maji smaji@ttic.edu
TTI Chicago
Tommi Jaakkola tommi@csail.mit.edu
Massachusetts Institute of Technology

In this paper we describe how MAP inference can be used to sample efficiently from Gibbs distributions. Specifically, we provide means for drawing either approximate or unbiased samples from Gibbs' distributions by introducing low dimensional perturbations and solving the corresponding MAP assignments. Our approach also leads to new ways to derive lower bounds on partition functions. We demonstrate empirically that our method excels in the typical "high signal - high coupling" regime. The setting results in ragged energy landscapes that are challenging for alternative approaches to sampling and/or lower bounds.

T35 EDML for Learning Parameters in Directed and Undirected Graphical Models

Khaled Refaat krefaat@cs.ucla.edu
ArTr Choi aychoi@cs.ucla.edu
Adnan Darwiche darwiche@cs.ucla.edu
UCLA

EDML is a recently proposed algorithm for learning parameters in Bayesian networks. It was originally derived in terms of approximate inference on a meta-network, which underlies the Bayesian approach to parameter estimation. While this initial derivation helped discover EDML in the first place and provided a concrete context for identifying some of its properties (e.g., in contrast to EM), the formal setting was somewhat tedious in the number of concepts it drew on. In this paper, we propose a greatly simplified perspective on EDML, which casts it as a general approach to continuous optimization. The new perspective has several advantages. First, it makes immediate some results that were non-trivial to prove initially. Second, it facilitates the design of EDML algorithms for new graphical models, leading to a new algorithm for learning parameters in Markov networks. We derive this algorithm in this paper, and show, empirically, that it can sometimes learn better estimates from complete data, several times faster than commonly used optimization methods, such as conjugate gradient and L-BFGS.

T36 Projecting Ising Model Parameters for Fast Mixing

Justin Domke justin.domke@nicta.com.au
NICTA
Xianghang Liu xianghang.liu@nicta.com.au
NICTA/UNSW

Inference in general Ising models is difficult, due to high treewidth making tree-based algorithms intractable. Moreover, when interactions are strong, Gibbs sampling may take exponential time to converge to the stationary distribution. We present an algorithm to project Ising model parameters onto a parameter set that is guaranteed to be fast mixing, under several divergences. We find that Gibbs sampling using the projected parameters is more accurate than with the original parameters when interaction strengths are strong and when limited time is available for sampling.

T37 Embed and Project: Discrete Sampling with Universal Hashing

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

We consider the problem of sampling from a probability distribution defined over a high-dimensional discrete set, specified for instance by a graphical model. We propose a sampling algorithm, called PAWS, based on embedding the set into a higher-dimensional space which is then randomly projected using universal hash functions to a lower-dimensional subspace and explored using combinatorial search methods. Our scheme can leverage fast combinatorial optimization tools as a blackbox and, unlike MCMC methods, samples produced are guaranteed to be within an (arbitrarily small) constant factor of the true probability distribution. We demonstrate that by using state-of-the-art combinatorial search tools, PAWS can efficiently sample from Ising grids with strong interactions and from software verification instances, while MCMC and variational methods fail in both cases.

T38 Learning Stochastic Inverses

Andreas Stuhlmüller ast@mit.edu
Massachusetts Institute of Technology
Jacob Taylor jacobt@stanford.edu
Noah Goodman ngoodman@stanford.edu
Stanford University

We describe a class of algorithms for amortized inference in Bayesian networks. In this setting, we invest computation upfront to support rapid online inference for a wide range of queries. Our approach is based on learning an inverse factorization of a model's joint distribution: a factorization that turns observations into root nodes. Our algorithms accumulate information to estimate the local conditional distributions that constitute such a factorization. These stochastic inverses can be used to invert each of the computation steps leading to an observation, sampling backwards in order to quickly find a likely explanation. We show that estimated inverses converge asymptotically in number of (prior or posterior) training samples. To make use of inverses before convergence, we describe the Inverse MCMC algorithm, which uses stochastic inverses to make block proposals for a Metropolis-Hastings sampler. We explore the efficiency of this sampler for a variety of parameter regimes and Bayes nets.

T39 Approximate Gaussian process inference for the drift function in stochastic differential equations

Andreas Ruttor andreas.ruttor@tu-berlin.de
Philipp Batz philipp.batz@tu-berlin.de
Manfred Opper manfred.opper@tu-berlin.de
TU Berlin

We introduce a nonparametric approach for estimating drift functions in systems of stochastic differential equations from incomplete observations of the state vector. Using a Gaussian process prior over the drift as a function of the state vector, we develop an approximate EM algorithm to deal with the unobserved, latent dynamics between observations. The posterior over states is approximated by a piecewise linearized process and the MAP estimation of the drift is facilitated by a sparse Gaussian process regression.

T40 Online Learning of Nonparametric Mixture Models via Sequential Variational Approximation

Dahua Lin dhlin@ttic.edu
TTI Chicago

Reliance on computationally expensive algorithms for inference has been limiting the use of Bayesian nonparametric models in large scale applications. To tackle this problem, we propose a Bayesian learning algorithm for DP mixture models. Instead of following the conventional paradigm -- random initialization plus iterative update, we take an progressive approach. Starting with a given prior, our method recursively transforms it into an approximate posterior through sequential variational approximation. In this process, new components will be incorporated on the fly when needed. The algorithm can reliably estimate a DP mixture model in one pass, making it particularly suited for applications with massive data. Experiments on both synthetic data and real datasets demonstrate remarkable improvement on efficiency -- orders of magnitude speed-up compared to the state-of-the-art.

T41 Memoized Online Variational Inference for Dirichlet Process Mixture Models

Mike Hughes mhughes@cs.brown.edu
Erik Sudderth sudderth@cs.brown.edu
Brown University

Variational inference algorithms provide the most effective framework for large-scale training of Bayesian nonparametric models. Stochastic online approaches are promising, but are sensitive to the chosen learning rate and often converge to poor local optima. We present a new algorithm, memoized online variational inference, which scales to very large (yet finite) datasets while avoiding the complexities of stochastic gradient. Our algorithm maintains finite-dimensional sufficient statistics from batches of the full dataset, requiring some additional memory but still scaling to millions of examples. Exploiting nested families of variational bounds for infinite nonparametric models, we develop principled birth and merge moves allowing non-local optimization. Births adaptively add components to the model to escape local optima, while merges remove redundancy and improve speed. Using Dirichlet process mixture models for image clustering and denoising, we demonstrate major improvements in robustness and accuracy.

T42 Regret based Robust Solutions for Uncertain Markov Decision Processes

Asrar Ahmed masrara@smu.edu.sg
Pradeep Varakantham pradeepv@smu.edu.sg
Singapore Management University
Yossiri Adulyasak yossiri@smart.mit.edu
Singapore-MIT Alliance for Research and Technology
Patrick Jaillet jaillet@mit.edu
Massachusetts Institute of Technology

In this paper, we seek robust policies for uncertain Markov Decision Processes (MDPs). Most robust optimization approaches for these problems have focussed on the computation of *maximin* policies which maximize the value corresponding to the worst realization of the uncertainty. Recent work has proposed *minimax* regret as a suitable alternative to the *maximin* objective for robust optimization. However, existing algorithms for handling minimax} regret are restricted to models with uncertainty over rewards only. We provide algorithms that employ sampling to improve across multiple dimensions: (a) Handle uncertainties over both transition and reward models; (b) Dependence of model uncertainties across state, action pairs and decision epochs; (c) Scalability and quality bounds. Finally, to demonstrate the empirical effectiveness of our sampling approaches, we provide comparisons against benchmark algorithms on two domains from literature. We also provide a Sample Average Approximation (SAA) analysis to compute a posteriori error bounds.

T43 Improved and Generalized Upper Bounds on the Complexity of Policy Iteration

Bruno Scherrer scherrer@loria.fr
INRIA

Given a Markov Decision Process (MDP) with n states and m actions per state, we study the number of iterations needed by Policy Iteration (PI) algorithms to converge to the optimal γ-discounted optimal policy. We consider two variations of PI: Howard's PI that changes the actions in all states with a positive advantage, and Simplex-PI that only changes the action in the state with maximal advantage. We show that Howard's PI terminates after at most $O\left(nm/1-\gamma \log\left(1/1-\gamma\right)\right)$ iterations, improving by a factor $O(\log n)$a result by Hansen et al. (2013), while Simplex-PI terminates after at most $O\left(n^2m/1-\gamma \quad \log\left(1/1-\gamma\right)\right)$ iterations, improving by a factor $O(\log n)$ a result by Ye (2011). Under some structural assumptions of the MDP, we then consider bounds that are independent of the discount factor~γ: given a measure of the maximal transient time τ_t and the maximal time τr to revisit states in recurrent classes under all policies, we show that Simplex-PI terminates after at most $O\text{~}\left(n^3m^2\tau_t\tau_r\right)$ iterations. This generalizes a recent result for deterministic MDPs by Post & Ye (2012), in which $\tau_t \leq n$ and $\tau_r \leq n$. We explain why similar results seem hard to derive for Howard's PI. Finally, under the additional (restrictive) assumption that the state space is partitioned in two sets, respectively states that are transient and recurrent for all policies, we show that Simplex-PI and Howard's PI terminate after at most $O\text{~}(nm(\tau_t+\tau_r))$ iterations.

T44 Efficient Exploration and Value Function Generalization in Deterministic Systems

Zheng Wen zhengwen@stanford.edu
Benjamin Van Roy bvr@stanford.edu
Stanford University

We consider the problem of reinforcement learning over episodes of a finite-horizon deterministic system and as a solution propose optimistic constraint propagation (OCP), an algorithm designed to synthesize efficient exploration and value function generalization. We establish that when the true value function lies within the given hypothesis class, OCP selects optimal actions over all but at most K episodes, where K is the eluder dimension of the given hypothesis class. We establish further efficiency and asymptotic performance guarantees that apply even if the true value function does not lie in the given hypothesis space, for the special case where the hypothesis space is the span of pre-specified indicator functions over disjoint sets.

T45 Aggregating Optimistic Planning Trees for Solving Markov Decision Processes

Gunnar Kedenburg gunnar.kedenburg@inria.fr
Remi Munos remi.munos@inria.fr
INRIA
Raphael Fonteneau raphael.fonteneau@ulg.ac.be
Université de Liège

This paper addresses the problem of online planning in Markov Decision Processes using only a generative model. We propose a new algorithm which is based on the construction of a forest of single successor state planning trees. For every explored state-action, such a tree contains exactly one successor state, drawn from the generative model. The trees are built using a planning algorithm which follows the optimism in the face of uncertainty principle, in assuming the most favorable outcome in the absence of further information. In the decision making step of the algorithm, the individual trees are combined. We discuss the approach, prove that our proposed algorithm is consistent, and empirically show that it performs better than a related algorithm which additionally assumes the knowledge of all transition distributions.

T46 Online learning in episodic Markovian decision processes by relative entropy policy search

Alexander Zimin ialexzimin@gmail.com
Institute of Science and Technology Austria
Gergo Neu neu.gergely@gmail.com
INRIA

We study the problem of online learning in finite episodic Markov decision processes where the loss function is allowed to change between episodes. The natural performance measure in this learning problem is the regret defined as the difference between the total loss of the best stationary policy and the total loss suffered by the learner. We assume that the learner is given access to a finite action space $\backslash\mathrm{A}$ and the state space $\backslash\mathrm{X}$ has a layered structure with L layers, so that state transitions are only possible between consecutive layers. We describe a variant of the recently proposed Relative Entropy Policy Search algorithm and show that its regret after T episodes is $2\sqrt{L}\backslash\mathrm{nX}\backslash\mathrm{nA}T \log(\backslash\mathrm{nX}\backslash\mathrm{nA}/L)$ in the bandit setting and $2L\sqrt{T}\log(\backslash\mathrm{nX}\backslash\mathrm{nA}/L)$ in the full information setting. These guarantees largely improve previously known results under much milder assumptions and cannot be significantly improved under general assumptions.

T47 Online Learning in Markov Decision Processes with Adversarially Chosen Transition Probability Distributions

Yasin Abbasi yasin.abbasi@gmail.com
Queensland University of Technology
Peter Bartlett bartlett@cs.berkeley.edu
Varun Kanade vkanade@eecs.berkeley.edu
UC Berkeley
Yevgeny Seldin yevgeny.seldin@gmail.com
Queensland Univ. of Technology & UC Berkeley
Csaba Szepesvari szepesva@cs.ualberta.ca
University of Alberta

We study the problem of online learning Markov Decision Processes (MDPs) when both the transition distributions and loss functions are chosen by an adversary. We present an algorithm that, under a mixing assumption, achieves $O\sqrt{(T\log|\Pi|+\log|\Pi|)}$regret with respect to a comparison set of policies Π. The regret is independent of the size of the state and action spaces. When expectations over sample paths can be computed efficiently and the comparison set Π has polynomial size, this algorithm is efficient. We also consider the episodic adversarial online shortest path problem. Here, in each episode an adversary may choose a weighted directed acyclic graph with an identified start and finish node. The goal of the learning algorithm is to choose a path that minimizes the loss while traversing from the start to finish node. At the end of each episode the loss function (given by weights on the edges) is revealed to the learning algorithm. The goal is to minimize regret with respect to a fixed policy for selecting paths. This problem is a special case of the online MDP problem. For randomly chosen graphs and adversarial losses, this problem can be efficiently solved. We show that it also can be efficiently solved for adversarial graphs and randomly chosen losses. When both graphs and losses are adversarially chosen, we present an efficient algorithm whose regret scales linearly with the number of distinct graphs. Finally, we show that designing efficient algorithms for the adversarial online shortest path problem (and hence for the adversarial MDP problem) is as hard as learning parity with noise, a notoriously difficult problem that has been used to design efficient cryptographic schemes.

T48 Online Learning of Dynamic Parameters in Social Networks

Shahin Shahrampour shahin@seas.upenn.edu
Sasha Rakhlin rakhlin@gmail.com
Ali Jadbabaie jadbabai@seas.upenn.edu
University of Pennsylvania

This paper addresses the problem of online learning in a dynamic setting. We consider a social network in which each individual observes a private signal about the underlying state of the world and communicates with her neighbors at each time period. Unlike many existing approaches, the underlying state is dynamic, and evolves according to a geometric random walk. We view the scenario as an optimization problem where agents aim to learn the true state while suffering the smallest possible loss. Based on the decomposition of the global loss function, we introduce two update mechanisms, each of which generates an estimate of the true state. We establish a tight bound on the rate of change of the underlying state, under which individuals can track the parameter with a bounded variance. Then, we characterize explicit expressions for the steady state mean-square deviation(MSD) of the estimates from the truth, per individual. We observe that only one of the estimators recovers the optimal MSD, which underscores the impact of the objective function decomposition on the learning quality. Finally, we provide an upper bound on the regret of the proposed methods, measured as an average of errors in estimating the parameter in a finite time.

T49 Modeling Overlapping Communities with Node Popularities

Prem Gopalan pgopalan@cs.princeton.edu
David Blei blei@cs.princeton.edu
Princeton University
Chong Wang chongw@cs.cmu.edu
CMU

We develop a probabilistic approach for accurate network modeling using node popularities within the framework of the mixed-membership stochastic blockmodel (MMSB). Our model integrates some of the basic properties of nodes in social networks: homophily and preferential connection to popular nodes. We develop a scalable algorithm for posterior inference, based on a novel nonconjugate variant of stochastic variational inference. We evaluate the link prediction accuracy of our algorithm on eight real-world networks with up to 60,000 nodes, and 24 benchmark networks. We demonstrate that our algorithm predicts better than the MMSB. Further, using benchmark networks we show that node popularities are essential to achieving high accuracy in the presence of skewed degree distribution and noisy links---both characteristics of real networks.

T50 A Scalable Approach to Probabilistic Latent Space Inference of Large-Scale Networks

Junming Yin junmingy@cs.cmu.edu
Qirong Ho qho@cs.cmu.edu
Eric Xing epxing@cs.cmu.edu
CMU

We propose a scalable approach for making inference about latent spaces of large networks. With a succinct representation of networks as a bag of triangular motifs, a parsimonious statistical model, and an efficient stochastic variational inference algorithm, we are able to analyze real networks with over a million vertices and hundreds of latent roles on a single machine in a matter of hours, a setting that is out of reach for many existing methods. When compared to the state-of-the-art probabilistic approaches, our method is several orders of magnitude faster, with competitive or improved accuracy for latent space recovery and link prediction.

T51 Relevance Topic Model for Unstructured Social Group Activity Recognition

Fang Zhao fang.zhao@nlpr.ia.ac.cn
Yongzhen Huang yzhuang@nlpr.ia.ac.cn
Liang Wang wangliang@nlpr.ia.ac.cn
Tieniu Tan tnt@nlpr.ia.ac.cn
Chinese Academy of Sciences

Unstructured social group activity recognition in web videos is a challenging task due to 1) the semantic gap between class labels and low-level visual features and 2) the lack of labeled training data. To tackle this problem, we propose a “relevance topic model” for jointly learning meaningful mid-level representations upon bag-of-words (BoW) video representations and a classifier with sparse weights. In our approach, sparse Bayesian learning is incorporated into an undirected topic model (i.e., Replicated Softmax) to discover topics which are relevant to video classes and suitable for prediction. Rectified linear units are utilized to increase the expressive power of topics so as to explain better video data containing complex contents and make variational inference tractable for the proposed model. An efficient variational EM algorithm is presented for model parameter estimation and inference. Experimental results on the Unstructured Social Activity Attribute dataset show that our model achieves state of the art performance and outperforms other supervised topic model in terms of classification accuracy, particularly in the case of a very small number of labeled training videos.

T52 Streaming Variational Bayes

Tamara Broderick tab@stat.berkeley.edu
Nicholas Boyd nickboyd@eecs.berkeley.edu
Andre Wibisono wibisono@eecs.berkeley.edu
Ashia Wilson ashia@stat.berkeley.edu
Michael Jordan jordan@cs.berkeley.edu
UC Berkeley

We present SDA-Bayes, a framework for (S)treaming, (D)istributed, (A)synchronous computation of a Bayesian posterior. The framework makes streaming updates to the estimated posterior according to a user-specified approximation primitive function. We demonstrate the usefulness of our framework, with variational Bayes (VB) as the primitive, by fitting the latent Dirichlet allocation model to two large-scale document collections. We demonstrate the advantages of our algorithm over stochastic variational inference (SVI), both in the single-pass setting SVI was designed for and in the streaming setting, to which SVI does not apply.

T53 Scalable Inference for Logistic-Normal Topic Models

Jianfei Chen chenjf10@mails.tsinghua.edu.cn
Jun Zhu dcszj@mail.tsinghua.edu.cn
Zi Wang wangzi10@mails.tsinghua.edu.cn
Xun Zheng xunzheng90@gmail.com
Bo Zhang dcszb@mail.tsinghua.edu.cn
Tsinghua University

Logistic-normal topic models can effectively discover correlation structures among latent topics. However, their inference remains a challenge because of the non-conjugacy between the logistic-normal prior and multinomial topic mixing proportions. Existing algorithms either make restricting mean-field assumptions or are not scalable to large-scale applications. This paper presents a partially collapsed Gibbs sampling algorithm that approaches the provably correct distribution by exploring the ideas of data augmentation. To improve time efficiency, we further present a parallel implementation that can deal with large-scale applications and learn the correlation structures of thousands of topics from millions of documents. Extensive empirical results demonstrate the promise.

T54 When are Overcomplete Topic Models Identifiable? Uniqueness of Tensor Tucker Decompositions with Structured Sparsity

Anima Anandkumar a.anandkumar@uci.edu
Majid Janzamin mjanzami@uci.edu
UC Irvine
Daniel Hsu danielhsu@gmail.com
Columbia University
Sham Kakade skakade@microsoft.com
Microsoft Research

Overcomplete latent representations have been very popular for unsupervised feature learning in recent years. In this paper, we specify which overcomplete models can be identified given observable moments of a certain order. We consider probabilistic admixture or topic models in the overcomplete regime, where the number of latent topics can greatly exceed the size of the observed word vocabulary. While general overcomplete topic models are not identifiable, we establish generic identifiability under a constraint, referred to as topic persistence. Our sufficient conditions for identifiability involve a novel set of “higher order’’ expansion conditions on the topic-word matrix or the population structure of the model. This set of higher-order expansion conditions allow for overcomplete models, and require the existence of a perfect matching from latent topics to higher order observed words. We establish that random structured topic models are identifiable w.h.p. in the overcomplete regime. Our identifiability results allow for general (non-degenerate) distributions for modeling the topic proportions, and thus, we can handle arbitrarily correlated topics in our framework. Our identifiability results imply uniqueness of a class of tensor decompositions with structured sparsity which is contained in the class of Tucker decompositions, but is more general than the Candecomp Parafac (CP) decomposition.

T55 Minimax Theory for High-dimensional Gaussian Mixtures with Sparse Mean Separation

Martin Azizyan mazizyan@cs.cmu.edu
Aarti Singh aartisingh@cmu.edu
Larry Wasserman larrywasserman.cool@gmail.com
CMU

While several papers have investigated computationally and statistically efficient methods for learning Gaussian mixtures, precise minimax bounds for their statistical performance as well as fundamental limits in high-dimensional settings are not well-understood. In this paper, we provide precise information theoretic bounds on the clustering accuracy and sample complexity of learning a mixture of two isotropic Gaussians in high dimensions under small mean separation. If there is a sparse subset of relevant dimensions that determine the mean separation, then the sample complexity only depends on the number of relevant dimensions and mean separation, and can be achieved by a simple computationally efficient procedure. Our results provide the first step of a theoretical basis for recent methods that combine feature selection and clustering.

T56 Cluster Trees on Manifolds

Sivaraman Balakrishnan the.seeing.stone@gmail.com
Srivatsan Narayanan srivatsa@cs.cmu.edu
Alessandro Rinaldo arinaldo@cmu.edu
Aarti Singh aartisingh@cmu.edu
Larry Wasserman larry@stat.cmu.edu
CMU

We investigate the problem of estimating the cluster tree for a density f supported on or near a smooth d-dimensional manifold M isometrically embedded in $\mathbb{R}^D$. We study a k-nearest neighbor based algorithm recently proposed by Chaudhuri and Dasgupta. Under mild assumptions on f and M, we obtain rates of convergence that depend on $d$only but not on the ambient dimension D. We also provide a sample complexity lower bound for a natural class of clustering algorithms that use D-dimensional neighborhoods.

T57 Convex Tensor Decomposition via Structured Schatten Norm Regularization

Ryota Tomioka tomioka@ttic.edu
TTI Chicago
Taiji Suzuki s-taiji@is.titech.ac.jp
Tokyo Institute of Technology

We propose a new class of structured Schatten norms for tensors that includes two recently proposed norms (“overlapped” and “latent”) for convex-optimization-based tensor decomposition. Based on the properties of the structured Schatten norms, we mathematically analyze the performance of “latent” approach for tensor decomposition, which was empirically found to perform better than the “overlapped” approach in some settings. We show theoretically that this is indeed the case. In particular, when the unknown true tensor is low-rank in a specific mode, this approach performs as well as knowing the mode with the smallest rank. Along the way, we show a novel duality result for structures Schatten norms, which is also interesting in the general context of structured sparsity. We confirm through numerical simulations that our theory can precisely predict the scaling behaviour of the mean squared error.

T58 Convex Relaxations for Permutation Problems

Fajwel Fogel fajwel.fogel@inria.fr
École Polytechnique
Rodolphe Jenatton rodolphe.jenatton@gmail.com
CMAP
Francis Bach francis.bach@mines.org
INRIA & ENS
Alexandre D’Aspremont alexandre.daspremont@m4x.org
CNRS – ENS

Seriation seeks to reconstruct a linear order between variables using unsorted similarity information. It has direct applications in archeology and shotgun gene sequencing for example. We prove the equivalence between the seriation and the combinatorial 2-sum problem (a quadratic minimization problem over permutations) over a class of similarity matrices. The seriation problem can be solved exactly by a spectral algorithm in the noiseless case and we produce a convex relaxation for the 2-sum problem to improve the robustness of solutions in a noisy setting. This relaxation also allows us to impose additional structural constraints on the solution, to solve semi-supervised seriation problems. We present numerical experiments on archeological data, Markov chains and gene sequences.

T59 Solving the multi-way matching problem by permutation synchronization

Deepti Pachauri pachauri@cs.wisc.edu
Vikas Singh vsingh@biostat.wisc.edu
UW-Madison
Risi Kondor risi@uchicago.edu
University of Chicago

The problem of matching not just two, but m different sets of objects to each other arises in a variety of contexts, including finding the correspondence between feature points across multiple images in computer vision. At present it is usually solved by matching the sets pairwise, in series. In contrast, we propose a new method, permutation synchronization, which finds all the matchings jointly, in one shot, via a relaxation to eigenvector decomposition. The resulting algorithm is both computationally efficient, and, as we demonstrate with theoretical arguments as well as experimental results, much more stable to noise than previous methods.

T60 Reflection methods for user-friendly submodular optimization

Stefanie Jegelka stefje@eecs.berkeley.edu
UC Berkeley
Francis Bach francis.bach@mines.org
INRIA & ENS
Suvrit Sra suvrit@gmail.com
MPI for Intelligent Systems & CMU

Recently, it has become evident that submodularity naturally captures widely occurring concepts in machine learning, signal processing and computer vision. In consequence, there is need for efficient optimization procedures for submodular functions, in particular for minimization problems. While general submodular minimization is challenging, we propose a new approach that exploits existing decomposability of submodular functions. In contrast to previous approaches, our method is neither approximate, nor impractical, nor does it need any cumbersome parameter tuning. Moreover, it is easy to implement and parallelize. A key component of our approach is a formulation of the discrete submodular minimization problem as a continuous best approximation problem. It is solved through a sequence of reflections and its solution can be automatically thresholded to obtain an optimal discrete solution. Our method solves both the continuous and discrete formulations of the problem, and therefore has applications in learning, inference, and reconstruction. In our experiments, we show the benefits of our new algorithms for two image segmentation tasks.

T61 Curvature and Optimal Algorithms for Learning and Minimizing Submodular Functions

Rishabh Iyer rkiyer@u.washington.edu
Jeff Bilmes bilmes@ee.washington.edu
University of Washington
Stefanie Jegelka stefje@eecs.berkeley.edu
UC Berkeley

We investigate three related and important problems connected to machine learning, namely approximating a submodular function everywhere, learning a submodular function (in a PAC like setting [26]), and constrained minimization of submodular functions. In all three problems, we provide improved bounds which depend on the “curvature” of a submodular function and improve on the previously known best results for these problems [9, 3, 7, 25] when the function is not too curved – a property which is true of many real-world submodular functions. In the former two problems, we obtain these bounds through a generic black-box transformation (which can potentially work for any algorithm), while in the case of submodular minimization, we propose a framework of algorithms which depend on choosing an appropriate surrogate for the submodular function. In all these cases, we provide almost matching lower bounds. While improved curvature-dependent bounds were shown for monotone submodular maximization [4, 27], the existence of similar improved bounds for the aforementioned problems has been open. We resolve this question in this paper by showing that the same notion of curvature provides these improved results. Empirical experiments add further support to our claims.

T62 An Approximate, Efficient LP Solver for LP Rounding

Srikrishna Sridhar srikris@cs.wisc.edu
Stephen Wright swright@cs.wisc.edu
Christopher Re chrisre@cs.wisc.edu
Ji Liu ji.liu.uwisc@gmail.com
Victor Bittorf bittorf@cs.wisc.edu
Ce Zhang czhang@cs.wisc.edu
UW-Madison

Many problems in machine learning can be solved by rounding the solution of an appropriate linear program. We propose a scheme that is based on a quadratic program relaxation which allows us to use parallel stochastic-coordinate-descent to approximately solve large linear programs efficiently. Our software is an order of magnitude faster than Cplex (a commercial linear programming solver) and yields similar solution quality. Our results include a novel perturbation analysis of a quadratic-penalty formulation of linear programming and a convergence result, which we use to derive running time and quality guarantees.

T63 Hierarchical Modular Optimization of Convolutional Networks Achieves Representations Similar to Macaque IT and Human Ventral Stream

Daniel Yamins yamins@mit.edu
Ha Hong hahong@mit.edu
Charles Cadieu c.cadieu@gmail.com
James DiCarlo dicarlo@mit.edu
Massachusetts Institute of Technology

Humans recognize visually-presented objects rapidly and accurately. To understand this ability, we seek to construct models of the ventral stream, the series of cortical areas thought to subserve object recognition. One tool to assess the quality of a model of the ventral stream is the Representation Dissimilarity Matrix (RDM), which uses a set of visual stimuli and measures the distances produced in either the brain (i.e. fMRI voxel responses, neural firing rates) or in models (features). Previous work has shown that all known models of the ventral stream fail to capture the RDM pattern observed in either IT cortex, the highest ventral area, or in the human ventral stream. In this work, we construct models of the ventral stream using a novel optimization procedure for category-level object recognition problems, and produce RDMs resembling both macaque IT and human ventral stream. The model, while novel in the optimization procedure, further develops a long-standing functional hypothesis that the ventral visual stream is a hierarchically arranged series of processing stages optimized for visual object recognition.

T64 Bayesian inference for low rank spatiotemporal neural receptive fields

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

The receptive field (RF) of a sensory neuron describes how the neuron integrates sensory stimuli over time and space. In typical experiments with naturalistic or flickering spatiotemporal stimuli, RFs are very high-dimensional, due to the large number of coefficients needed to specify an integration profile across time and space. Estimating these coefficients from small amounts of data poses a variety of challenging statistical and computational problems. Here we address these challenges by developing Bayesian reduced rank regression methods for RF estimation. This corresponds to modeling the RF as a sum of several space-time separable (i.e., rank-1) filters, which proves accurate even for neurons with strongly oriented space-time RFs. This approach substantially reduces the number of parameters needed to specify the RF, from 1K-100K down to mere 100s in the examples we consider, and confers substantial benefits in statistical power and computational efficiency. In particular, we introduce a novel prior over low-rank RFs using the restriction of a matrix normal prior to the manifold of low-rank matrices. We then use a "localized'' prior over row and column covariances to obtain sparse, smooth, localized estimates of the spatial and temporal RF components. We develop two methods for inference in the resulting hierarchical model: (1) a fully Bayesian method using blocked-Gibbs sampling; and (2) a fast, approximate method that employs alternating coordinate ascent of the conditional marginal likelihood. We develop these methods under Gaussian and Poisson noise models, and show that low-rank estimates substantially outperform full rank estimates in accuracy and speed using neural data from retina and V1.

T65 Spectral methods for neural characterization using generalized quadratic models

Il Park memming@austin.utexas.edu
Evan Archer earcher@utexas.edu
Nicholas Priebe nicholas@utexas.edu
Jonathan Pillow pillow@mail.utexas.edu
UT Austin

We describe a set of fast, tractable methods for characterizing neural responses to high-dimensional sensory stimuli using a model we refer to as the generalized quadratic model (GQM). The GQM consists of a low-rank quadratic form followed by a point nonlinearity and exponential-family noise. The quadratic form characterizes the neuron's stimulus selectivity in terms of a set linear receptive fields followed by a quadratic combination rule, and the invertible nonlinearity maps this output to the desired response range. Special cases of the GQM include the 2nd-order Volterra model (Marmarelis and Marmarelis 1978, Koh and Powers 1985) and the elliptical Linear-Nonlinear-Poisson model (Park and Pillow 2011). Here we show that for "canonical form" GQMs, spectral decomposition of the first two response-weighted moments yields approximate maximum-likelihood estimators via a quantity called the expected log-likelihood. The resulting theory generalizes moment-based estimators such as the spike-triggered covariance, and, in the Gaussian noise case, provides closed-form estimators under a large class of non-Gaussian stimulus distributions. We show that these estimators are fast and provide highly accurate estimates with far lower computational cost than full maximum likelihood. Moreover, the GQM provides a natural framework for combining multi-dimensional stimulus sensitivity and spike-history dependencies within a single model. We show applications to both analog and spiking data using intracellular recordings of V1 membrane potential and extracellular recordings of retinal spike trains.

T66 Optimal Neural Population Codes for High-dimensional Stimulus Variables

Jimmy Wang wangzhuo@sas.upenn.edu
Alan Stocker astocker@sas.upenn.edu
Daniel Lee ddlee@seas.upenn.edu
University of Pennsylvania

How does neural population process sensory information? Optimal coding theories assume that neural tuning curves are adapted to the prior distribution of the stimulus variable. Most of the previous work has discussed optimal solutions for only one-dimensional stimulus variables. Here, we expand some of these ideas and present new solutions that define optimal tuning curves for high-dimensional stimulus variables. We consider solutions for a minimal case where the number of neurons in the population is equal to the number of stimulus dimensions (diffeomorphic). In the case of two-dimensional stimulus variables, we analytically derive optimal solutions for different optimal criteria such as minimal L2 reconstruction error or maximal mutual information. For higher dimensional case, the learning rule to improve the population code is provided.

T67 Robust learning of low-dimensional dynamics from large neural ensembles

David Pfau pfau@neurotheory.columbia.edu
Eftychios Pnevmatikakis eftychios@stat.columbia.edu
Liam Paninski liam@stat.columbia.edu
Columbia University

Recordings from large populations of neurons make it possible to search for hypothesized low-dimensional dynamics. Finding these dynamics requires models that take into account biophysical constraints and can be fit efficiently and robustly. Here, we present an approach to dimensionality reduction for neural data that is convex, does not make strong assumptions about dynamics, does not require averaging over many trials and is extensible to more complex statistical models that combine local and global influences. The results can be combined with spectral methods to learn dynamical systems models. The basic method can be seen as an extension of PCA to the exponential family using nuclear norm minimization. We evaluate the effectiveness of this method using an exact decomposition of the Bregman divergence that is analogous to variance explained for PCA. We show on model data that the parameters of latent linear dynamical systems can be recovered, and that even if the dynamics are not stationary we can still recover the true latent subspace. We also demonstrate an extension of nuclear norm minimization that can separate sparse local connections from global latent dynamics. Finally, we demonstrate improved prediction on real neural data from monkey motor cortex compared to fitting linear dynamical models without nuclear norm smoothing.

T68 Sparse nonnegative deconvolution for compressive calcium imaging: algorithms and phase transitions

Eftychios Pnevmatikakis eftychios@stat.columbia.edu
Liam Paninski liam@stat.columbia.edu
Columbia University

We propose a compressed sensing (CS) calcium imaging framework for monitoring large neuronal populations, where we image randomized projections of the spatial calcium concentration at each timestep, instead of measuring the concentration at individual locations. We develop scalable nonnegative deconvolution methods for extracting the neuronal spike time series from such observations. We also address the problem of demixing the spatial locations of the neurons using rank-penalized matrix factorization methods. By exploiting the sparsity of neural spiking we demonstrate that the number of measurements needed per timestep is significantly smaller than the total number of neurons, a result that can potentially enable imaging of larger populations at considerably faster rates compared to traditional raster-scanning techniques. Unlike traditional CS setups, our problem involves a block-diagonal sensing matrix and a non-orthogonal sparse basis that spans multiple timesteps. We study the effect of these distinctive features in a noiseless setup using recent results relating conic geometry to CS. We provide tight approximations to the number of measurements needed for perfect deconvolution for certain classes of spiking processes, and show that this number displays a "phase transition," similar to phenomena observed in more standard CS settings; however, in this case the required measurement rate depends not just on the mean sparsity level but also on other details of the underlying spiking process.

T69 Generalized Method-of-Moments for Rank Aggregation

Hossein Azari Soufiani azari@fas.harvard.edu
William Chen williamzc@gmail.com
David Parkes parkes@eecs.harvard.edu
Lirong Xia lxia@seas.harvard.edu
Harvard University

In this paper we propose a class of efficient Generalized Method-of-Moments(GMM) algorithms for computing parameters of the Plackett-Luce model, where the data consists of full rankings over alternatives. Our technique is based on breaking the full rankings into pairwise comparisons, and then computing parameters that satisfy a set of generalized moment conditions. We identify conditions for the output of GMM to be unique, and identify a general class of consistent and inconsistent breakings. We then show by theory and experiments that our algorithms run significantly faster than the classical Minorize-Maximization (MM) algorithm, while achieving competitive statistical efficiency.

T70 Generalized Random Utility Models with Multiple Types

Hossein Azari Soufiani azari@fas.harvard.edu
Hansheng Diao diao@fas.harvard.edu
Zhenyu Lai zlai@fas.harvard.edu
David Parkes parkes@eecs.harvard.edu
Harvard University

We propose a model for demand estimation in multi-agent, differentiated product settings and present an estimation algorithm that uses reversible jump MCMC techniques to classify agents' types. Our model extends the popular setup in Berry, Levinsohn and Pakes (1995) to allow for the data-driven classification of agents' types using agent-level data. We focus on applications involving data on agents' ranking over alternatives, and present theoretical conditions that establish the identifiability of the model and uni-modality of the likelihood/posterior. Results on both real and simulated data provide support for the scalability of our approach.

T71 Speedup Matrix Completion with Side Information: Application to Multi-Label Learning

Miao Xu xum@lamda.nju.edu.cn
Zhi-Hua Zhou zhouzh@lamda.nju.edu.cn
Nanjing University
Rong Jin rong+@cs.cmu.edu
Michigan State University (MSU)

In standard matrix completion theory, it is required to have at least $O(n\ln^2 n)$observed entries to perfectly recover a low-rank matrix M of size $n\times n$, leading to a large number of observations when n is large. In many real tasks, side information in addition to the observed entries is often available. In this work, we develop a novel theory of matrix completion that explicitly explore the side information to reduce the requirement on the number of observed entries. We show that, under appropriate conditions, with the assistance of side information matrices, the number of observed entries needed for a perfect recovery of matrix M can be dramatically reduced to $O(\ln n)$. We demonstrate the effectiveness of the proposed approach for matrix completion in transductive incomplete multi-label learning.

T72 Correlated random features for fast semi-supervised learning

Brian McWilliams brian.mcwilliams@inf.ethz.ch
David Balduzzi david.balduzzi@inf.ethz.ch
Joachim Buhmann jbuhmann@inf.ethz.ch
ETH Zurich

This paper presents Correlated Nystrom Views (XNV), a fast semi-supervised algorithm for regression and classification. The algorithm draws on two main ideas. First, it generates two views consisting of computationally inexpensive random features. Second, multiview regression, using Canonical Correlation Analysis (CCA) on unlabeled data, biases the regression towards useful features. It has been shown that CCA regression can substantially reduce variance with a minimal increase in bias if the views contains accurate estimators. Recent theoretical and empirical work shows that regression with random features closely approximates kernel regression, implying that the accuracy requirement holds for random views. We show that XNV consistently outperforms a state-of-the-art algorithm for semi-supervised learning: substantially improving predictive performance and reducing the variability of performance on a wide variety of real-world datasets, whilst also reducing runtime by orders of magnitude.

T73 Manifold-based Similarity Adaptation for Label Propagation

Masayuki Karasuyama karasuyama@kuicr.kyoto-u.ac.jp
Hiroshi Mamitsuka mami@kuicr.kyoto-u.ac.jp
Kyoto University

Label propagation is one of the state-of-the-art methods for semi-supervised learning, which estimates labels by propagating label information through a graph. Label propagation assumes that data points (nodes) connected in a graph should have similar labels. Consequently, the label estimation heavily depends on edge weights in a graph which represent similarity of each node pair. We propose a method for a graph to capture the manifold structure of input features using edge weights parameterized by a similarity function. In this approach, edge weights represent both similarity and local reconstruction weight simultaneously, both being reasonable for label propagation. For further justification, we provide analytical considerations including an interpretation as a cross-validation of a propagation model in the feature space, and an error analysis based on a low dimensional manifold model. Experimental results demonstrated the effectiveness of our approach both in synthetic and real datasets.

T74 Supervised Sparse Analysis and Synthesis Operators

Pablo Sprechmann pablo.sprechmann@duke.edu
Guillermo Sapiro guillermo.sapiro@duke.edu
Duke University
Roy Litman roeelitm@post.tau.ac.il
Tal Ben Yakar talby10@gmail.com
Alexander Bronstein bron@eng.tau.ac.il
Tel Aviv University

In this paper, we propose a new and computationally efficient framework for learning sparse models. We formulate a unified approach that contains as particular cases models promoting sparse synthesis and analysis type of priors, and mixtures thereof. The supervised training of the proposed model is formulated as a bilevel optimization problem, in which the operators are optimized to achieve the best possible performance on a specific task, e.g., reconstruction or classification. By restricting the operators to be shift invariant, our approach can be thought as a way of learning analysis+synthesis sparsity-promoting convolutional operators. Leveraging recent ideas on fast trainable regressors designed to approximate exact sparse codes, we propose a way of constructing feed-forward neural networks capable of approximating the learned models at a fraction of the computational cost of exact solvers. In the shift-invariant case, this leads to a principled way of constructing task-specific convolutional networks. We illustrate the proposed models on several experiments in music analysis and image processing applications.

T75 When in Doubt, SWAP: High-Dimensional Sparse Recovery from Correlated Measurements

Divyanshu Vats dvats@rice.edu
Richard Baraniuk richb@rice.edu
Rice University

We consider the problem of accurately estimating a high-dimensional sparse vector using a small number of linear measurements that are contaminated by noise. It is well known that standard computationally tractable sparse recovery algorithms, such as the Lasso, OMP, and their various extensions, perform poorly when the measurement matrix contains highly correlated columns. We develop a simple greedy algorithm, called SWAP, that iteratively swaps variables until a desired loss function cannot be decreased any further. SWAP is surprisingly effective in handling measurement matrices with high correlations. We prove that SWAP can be easily used as a wrapper around standard sparse recovery algorithms for improved performance. We theoretically quantify the statistical guarantees of SWAP and complement our analysis with numerical results on synthetic and real data.

T76 Deep content-based music recommendation

Aaron van den Oord aaron.vandenoord@ugent.be
Sander Dielemans ander.dieleman@ugent.be
Benjamin Schrauwen benjamin.Schrauwen@ugent.be
Ghent University

Automatic music recommendation has become an increasingly relevant problem in recent years, since a lot of music is now sold and consumed digitally. Most recommender systems rely on collaborative filtering. However, this approach suffers from the cold start problem: it fails when no usage data is available, so it is not effective for recommending new and unpopular songs. In this paper, we propose to use a latent factor model for recommendation, and predict the latent factors from music audio when they cannot be obtained from usage data. We compare a traditional approach using a bag-of-words representation of the audio signals with deep convolutional neural networks, and evaluate the predictions quantitatively and qualitatively on the Million Song Dataset. We show that using predicted latent factors produces sensible recommendations, despite

the fact that there is a large semantic gap between the characteristics of a song that affect user preference and the corresponding audio signal. We also show that recent advances in deep learning translate very well to the music recommendation setting, with deep convolutional neural networks significantly outperforming the traditional approach.

T77 Probabilistic Low-Rank Matrix Completion with Adaptive Spectral Regularization Algorithms

Adrien Todeschini adrien.todeschini@inria.fr
INRIA
Francois Caron francois.caron@stats.ox.ac.uk
University of Oxford
Marie Chavent Marie.Chavent@u-bordeaux2.fr
Université de Bordeaux II & INRIA

We propose a novel class of algorithms for low rank matrix completion. Our approach builds on novel penalty functions on the singular values of the low rank matrix. By exploiting a mixture model representation of this penalty, we show that a suitably chosen set of latent variables enables to derive an Expectation-Maximization algorithm to obtain a Maximum A Posteriori estimate of the completed low rank matrix. The resulting algorithm is an iterative soft-thresholded algorithm which iteratively adapts the shrinkage coefficients associated to the singular values. The algorithm is simple to implement and can scale to large matrices. We provide numerical comparisons between our approach and recent alternatives showing the interest of the proposed approach for low rank matrix completion.

T78 A Gang of Bandits

Nicolò Cesa-Bianchi nicolo.cesa-bianchi@unimi.it
Giovanni Zappella giovanni.zappella@unimi.it
University of Milan
Claudio Gentile claudio.gentile@uninsubria.it
University of Insubria

Multi-armed bandit problems are receiving a great deal of attention because they adequately formalize the exploration-exploitation trade-offs arising in several industrially relevant applications, such as online advertisement and, more generally, recommendation systems. In many cases, however, these applications have a strong social component, whose integration in the bandit algorithm could lead to a dramatic performance increase. For instance, we may want to serve content to a group of users by taking advantage of an underlying network of social relationships among them. In this paper, we introduce novel algorithmic approaches to the solution of such networked bandit problems. More specifically, we design and analyze a global strategy which allocates a bandit algorithm to each network node (user) and allows it to “share” signals (contexts and payoffs) with the neghboring nodes. We then derive two more scalable variants of this strategy based on different ways of clustering the graph nodes. We experimentally compare the algorithm and its variants to state-of-the-art methods for contextual bandits that do not use the relational information. Our experiments, carried out on synthetic and real-world datasets, show a marked increase in prediction performance obtained by exploiting the network structure.

T79 Contrastive Learning Using Spectral Methods

James Zou jzou@fas.harvard.edu
David Parkes parkes@eecs.harvard.edu
Ryan Adams rpa@seas.harvard.edu
Harvard University
Daniel Hsu danielhsu@gmail.com
Columbia University

In many natural settings, the analysis goal is not to characterize a single data set in isolation, but rather to understand the difference between one set of observations and another. For example, given a background corpus of news articles together with writings of a particular author, one may want a topic model that explains word patterns and themes specific to the author. Another example comes from genomics, in which biological signals may be collected from different regions of a genome, and one wants a model that captures the differential statistics observed in these regions. This paper formalizes this notion of contrastive learning for mixture models, and develops spectral algorithms for inferring mixture components specific to a foreground data set when contrasted with a background data set. The method builds on recent moment-based estimators and tensor decompositions for latent variable models, and has the intuitive feature of using background data statistics to appropriately modify moments estimated from foreground data. A key advantage of the method is that the background data need only be coarsely modeled, which is important when the background is too complex, noisy, or not of interest. The method is demonstrated on applications in contrastive topic modeling and genomic sequence analysis.

T80 Fast Determinantal Point Process Sampling with Application to Clustering

Byungkon Kang
byungkon@kaist.ac.kr
Samsung Electronics

Determinantal Point Process (DPP) has gained much popularity for modeling sets of diverse items. The gist of DPP is that the probability of choosing a particular set of items is proportional to the determinant of a positive definite matrix that defines the similarity of those items. However, computing the determinant requires time cubic in the number of items, and is hence impractical for large sets. In this paper, we address this problem by constructing a rapidly mixing Markov chain, from which we can acquire a sample from the given DPP in sub-cubic time. In addition, we show that this framework can be extended to sampling from cardinality-constrained DPPs. As an application, we show how our sampling algorithm can be used to provide a fast heuristic for determining the number of clusters, resulting in better clustering.

T81 Computing the Stationary Distribution Locally

Christina Lee celee@mit.edu
Asuman Ozdaglar asuman@mit.edu
Devavrat Shah devavrat@mit.edu
Massachusetts Institute of Technology

Computing the stationary distribution of a large finite or countably infinite state space Markov Chain (MC) has become central in many problems such as statistical inference and network analysis. Standard methods involve large matrix multiplications as in power iteration, or simulations of long random walks to sample states from the stationary distribution, as in Markov Chain Monte Carlo (MCMC). However these methods are computationally costly; either they involve operations at every state or they scale (in computation time) at least linearly in the size of the state space. In this paper, we provide a novel algorithm that answers whether a chosen state in a MC has stationary probability larger than some $\Delta\in(0,1)$. If so, it estimates the stationary probability. Our algorithm uses information from a local neighborhood of the state on the graph induced by the MC, which has constant size relative to the state space. We provide correctness and convergence guarantees that depend on the algorithm parameters and mixing properties of the MC. Simulation results show MCs for which this method gives tight estimates.

T82 Learning Prices for Repeated Auctions with Strategic Buyers

Kareem Amin akareem@seas.upenn.edu
University of Pennsylvania
Afshin Rostamizadeh rostami@google.com
Umar Syed usyed@google.com
Google Research

Inspired by real-time ad exchanges for online display advertising, we consider the problem of inferring a buyer's value distribution for a good when the buyer is repeatedly interacting with a seller through a posted-price mechanism. We model the buyer as a strategic agent, whose goal is to maximize her long-term surplus, and we are interested in mechanisms that maximize the seller's long-term revenue. We present seller algorithms that are no-regret when the buyer discounts her future surplus --- i.e. the buyer prefers showing advertisements to users sooner rather than later. We also give a lower bound on regret that increases as the buyer's discounting weakens and shows, in particular, that any seller algorithm will suffer linear regret if there is no discounting.

T83 Efficient Algorithm for Privately Releasing Smooth Queries

Ziteng Wang wangzt2012@gmail.com
Kai Fan interfk@hotmail.com
Jiaqi Zhang grzhang.jq@gmail.com
Liwei Wang wanglw@cis.pku.edu.cn
Peking University

We study differentially private mechanisms for answering smooth} queries on databases consisting of data points in $\mathbb{R}^d$. A K-smooth query is specified by a function whose partial derivatives up to order K are all bounded. We develop an ϵ-differentially private mechanism which for the class of K-smooth queries has accuracy $O(\left(1n\right)^{K/2d+K}/\epsilon)$. The mechanism first outputs a summary of the database. To obtain an answer of a query, the user runs a public evaluation algorithm which contains no information of the database. Outputting the summary runs in time $O(n^1{+}^{d/2d+K})$, and the evaluation algorithm for answering a query runs in time $O(n^{(d+2+2d/K)2d+K})$. Our mechanism is based on $L\infty$-approximation of (transformed) smooth functions by low degree even trigonometric polynomials with small and efficiently computable coefficients.

T84 (Nearly) Optimal Algorithms for Private Online Learning in Full-information and Bandit Settings

Abhradeep Guha Thakurta b-abhrag@microsoft.com
Stanford University & Microsoft
Adam Smith asmith@cse.psu.edu
Penn State University

We provide a general technique for making online learning algorithms differentially private, in both the full information and bandit settings. Our technique applies to algorithms that aim to minimize a convex loss function which is a sum of smaller convex loss terms, one for each data point. We modify the popular mirror descent approach, or rather a variant called follow the approximate leader. The technique leads to the first nonprivate algorithms for private online learning in the bandit setting. In the full information setting, our algorithms improve over the regret bounds of previous work. In many cases, our algorithms (in both settings) matching the dependence on the input length, T, of the optimal nonprivate} regret bounds up to logarithmic factors in T. Our algorithms require logarithmic space and update time.

T85 Local Privacy and Minimax Bounds: Sharp Rates for Probability Estimation

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

We provide a detailed study of the estimation of probability distributions---discrete and continuous---in a stringent setting in which data is kept private even from the statistician. We give sharp minimax rates of convergence for estimation in these locally private settings, exhibiting fundamental tradeoffs between privacy and convergence rate, as well as providing tools to allow movement along the privacy-statistical efficiency continuum. One of the consequences of our results is that Warner's classical work on randomized response is an optimal way to perform survey sampling while maintaining privacy of the respondents.

T86 A Stability-based Validation Procedure for Differentially Private Machine Learning

Kamalika Chaudhuri kamalika@cs.ucsd.edu
Staal Vinterbo sav@ucsd.edu
UC San Diego

Differential privacy is a cryptographically motivated definition of privacy which has gained considerable attention in the algorithms, machine-learning and data-mining communities. While there has been an explosion of work on differentially private machine learning algorithms, a major barrier to achieving end-to-end differential privacy in practical machine learning applications is the lack of an effective procedure for differentially private parameter tuning, or, determining the parameter value, such as a bin size in a histogram, or a regularization parameter, that is suitable for a particular application. In this paper, we introduce a generic validation procedure for differentially private machine learning algorithms that apply when a certain stability condition holds on the training algorithm and the validation performance metric. The training data size and the privacy budget used for training in our procedure is independent of the number of parameter values searched over. We apply our generic procedure to two fundamental tasks in statistics and machine-learning -- training a regularized linear classifier and building a histogram density estimator that result in end-to-end differentially private solutions for these problems.

T87 Similarity Component Analysis

Soravit Changpinyo martbeerina@gmail.com
Kuan Liu liukuan0518@gmail.com
Fei Sha feisha@usc.edu
University of Southern California (USC)

Measuring similarity is crucial to many learning tasks. It is also a richer and broader notion than what most metric learning algorithms can model. For example, similarity can arise from the process of aggregating the decisions of multiple latent components, where each latent component compares data in its own way by focusing on a different subset of features. In this paper, we propose Similarity Component Analysis (SCA), a probabilistic graphical model that discovers those latent components from data. In SCA, a latent component generates a local similarity value, computed with its own metric, independently of other components. The final similarity measure is then obtained by combining the local similarity values with a (noisy-)OR gate. We derive an EM-based algorithm for fitting the model parameters with similarity-annotated data from pairwise comparisons. We validate the SCA model on synthetic datasets where SCA discovers the ground-truth about the latent components. We also apply SCA to a multiway classification task and a link prediction task. For both tasks, SCA attains significantly better prediction accuracies than competing methods. Moreover, we show how SCA can be instrumental in exploratory analysis of data, where we gain insights about the data by examining patterns hidden in its latent components' local similarity values.

T88 A message-passing algorithm for multi-agent trajectory planning

Jose Bento jbento@disneyresearch.com
Nate Derbinsky nate.derbinsky@disneyresearch.com
Javier Alonso-Mora jalonso@disneyresearch.com
Jonathan Yedidia yedidia@disneyresearch.com
Disney Research

We describe a novel approach for computing collision-free global} trajectories for p agents with specified initial and final configurations, based on an improved version of the alternating direction method of multipliers (ADMM) algorithm. Compared with existing methods, our approach is naturally parallelizable and allows for incorporating different cost functionals with only minor adjustments. We apply our method to classical challenging instances and observe that its computational requirements scale well with p for several cost functionals. We also show that a specialization of our algorithm can be used for *local* motion planning by solving the problem of joint optimization in velocity space.

T89 The Power of Asymmetry in Binary Hashing

Behnam Neyshabur btavakoli@ttic.edu
Nati Srebro nati@ttic.edu
Yury Makarychev yury@ttic.edu
Payman Yadollahpour pyadolla@ttic.edu
TTI Chicago
Russ Salakhutdinov rsalakhu@cs.toronto.edu
University of Toronto

When approximating binary similarity using the hamming distance between short binary hashes, we shown that even if the similarity is symmetric, we can have shorter and more accurate hashes by using two distinct code maps. I.e.~by approximating the similarity between x and x' as the hamming distance between $f(x)$ and $g(x')$, for two distinct binary codes f,g, rather than as the hamming distance between $f(x)$ and $f(x')$.

T90 Learning to Prune in Metric and Non-Metric Spaces

Leonid Boytsov leo@boytsov.info
CMU
Bilegsaikhan Naidan bileg@idi.ntnu.no
Norwegian Univ. of Science and Technology (NTNU)

To the best of our knowledge, this work is the first successful attempt to employ a VP-tree with the learned pruning algorithm in non-metric spaces. We focus on approximate nearest neighbor (NN) retrieval methods and experiment with two simple yet effective learning-to-prune approaches: density estimation through sampling and “stretching” of the triangle inequality. Both methods are evaluated using data sets with a metric (the Euclidean) and a non-metric (the KL-divergence) distance functions. The VP-tree with a learned pruner is compared against the recently proposed state-of-the art approaches: the bbtree, the multi-probe locality sensitive hashing (LSH), and permutation methods. Our method was competitive against state-of-the art methods and outperformed them in most cases by a wide margin. Our experiments also showed that the bbtree (an exact search method) was typically slower than exhaustive searching, if the KL-divergence was evaluated efficiently (through precomputing logarithms at index time). Conditions on spaces where the VP-tree is applicable are discussed.

FRIDAY

ORAL SESSION

SESSION 1 - 9:00 – 10:10 AM

Session Chair: Neil Lawrence

INVITED TALK: Small, n=me, Data

Deborah Estrin destrin@cs.cornell.edu
Cornell NYC Tech

Consider a new kind of cloud-based app that would create a picture of an individual's behavior over time by continuously, securely, and privately analyzing the digital traces they generate 24x7 by virtue of the fact that they mediate, or at least accompany, their lives with mobile and other digital technologies. The social networks, search engines, mobile operators, online games, and e-commerce sites that they access every hour of most every day extensively use these digital traces to tailor service offerings and to improve system performance and in some cases to target advertisements. Most of these services do not make these individual traces available to the person who generated them; but they might begin to do so if we identify the market, technical, and social mechanisms that would derive value from these traces. Our premise is that this broad but highly personalized, data set can be analyzed to draw powerful inferences about an individual, and for that individual. Use of these traces could enhance, and even transform, our experiences as consumers, patients, passengers, customers, family members, as well as users of online media. These traces might fuel apps that offer individuals personalized, data-driven, insights into their habits and habitats. But for this to be realized the raw data sources will require extensive processing in order to generate an actionable representation of someone's relevant behaviors, e.g.., a personalized "behavioral pulse".

Deborah Estrin is a Professor of Computer Science at Cornell Tech in New York City (http://tech.cornell.edu/deborah-estrin) and a Professor of Public Health at Weill Cornell Medical College. She is co-founder of the non-profit startup, Open mHealth (http://openmhealth.org/). She was previously on faculty at UCLA and Founding Director of the NSF Center for Embedded Networked Sensing (CENS). Estrin is a pioneer in networked sensing, which uses mobile and wireless systems to collect and analyze real time data about the physical world and the people who occupy it. Estrin's current focus is on mobile health (mhealth), leveraging the programmability, proximity, and pervasiveness of mobile devices and the cloud for health management. She is an elected member of the American Academy of Arts and Sciences and the National Academy of Engineering. She recently presented at TEDMED about small data: https://smalldata.tech.cornell.edu/

Scalable Influence Estimation in Continuous-Time Diffusion Networks

Nan Du dunan@gatech.edu
Le Song lsong@cc.gatech.edu
Hongyuan Zha zha@cc.gatech.edu
Georgia Tech
Manuel Gomez-Rodriguez
manuelgr@tuebingen.mpg.de
MPI for Intelligent Systems

If a piece of information is released from a media site, can it spread, in 1 month, to a million web pages? This influence estimation problem is very challenging since both the time-sensitive nature of the problem and the issue of scalability need to be addressed simultaneously. In this paper, we propose a randomized algorithm for influence estimation in continuous-time diffusion networks. Our algorithm can estimate the influence of every node in a network with $|\backslash\mathrm{Vcal}|$nodes and $|\backslash\mathrm{Ecal}|$ edges to an accuracy of ϵ using $n{=}O(1/\epsilon 2)$randomizations and up to logarithmic factors $O(n|\backslash\mathrm{Ecal}|{+}n|\backslash\mathrm{Vcal}|)$computations. When used as a subroutine in a greedy influence maximization algorithm, our proposed method is guaranteed to find a set of nodes with an influence of at least $(1{-}1/e)\mathrm{OPT}{-}2\epsilon$, where OPT is the optimal value. Experiments on both synthetic and real-world data show that the proposed method can easily scale up to networks of millions of nodes while significantly improves over previous state-of-the-arts in terms of the accuracy of the estimated influence and the quality of the selected nodes in maximizing the influence.

SPOTLIGHT SESSION

Session 1, 10:10 – 10:30 AM

- **Adaptive Anonymity via b-Matching**
  K. Choromanski, Google Research; T. Jebara, K. Tang, Columbia University
  See abstract F89, page 68

- **Exact and Stable Recovery of Pairwise Interaction Tensors**
  S. Chen, M. Lyu, CUHK; I. King, Chinese University of Hong Kong; Z. Xu, University of Purdue
  See abstract F86, page 68

- **Matrix factorization with binary components**
  M. Slawski, M. Hein, P. Lutsik, Saarland University
  See abstract F84, page 67

- **On the Complexity and Approximation of Binary Evidence in Lifted Inference**
  G. Van den Broeck, A. Darwiche, UCLA
  See abstract F77, page 65

- **Unsupervised Spectral Learning of Finite State Transducers**
  R. Bailly, X. Carreras, A. Quattoni, Universitat Politècnica de Catalunya
  See abstract F67, page 63

- **Graphical Models for Inference with Missing Data**
  K. Mohan, J. Pearl, UCLA; J. Tian, Iowa State University
  See abstract F70, page 64

ORAL SESSION

Session 2, 10:55 – 11:40 AM

Session Chair: Francis Bach

NIPS Award 1 - 10:55 - 11:00

On Decomposing the Proximal Map

Yao-Liang Yu yaoliang@cs.ualberta.ca
University of Alberta

The proximal map is the key step in gradient-type algorithms, which have become prevalent in large-scale high-dimensional problems. For simple functions this proximal map is available in closed-form while for more complicated functions it can become highly nontrivial. Motivated by the need of combining regularizers to simultaneously induce different types of structures, this paper initiates a systematic investigation of when the proximal map of a sum of functions decomposes into the composition of the proximal maps of the individual summands. We not only unify a few known results scattered in the literature but also discover several new decompositions obtained almost effortlessly from our theory.

Non-Uniform Camera Shake Removal Using a Spatially-Adaptive Sparse Penalty

Haichao Zhang hczhang1@gmail.com
Duke University
David Wipf davidwip@microsoft.com
Microsoft Research

Typical blur from camera shake often deviates from the standard uniform convolutional assumption, in part because of problematic rotations which create greater blurring away from some unknown center point. Consequently, successful blind deconvolution for removing shake artifacts requires the estimation of a spatially-varying or non-uniform blur operator. Using ideas from Bayesian inference and convex analysis, this paper derives a non-uniform blind deblurring algorithm with several desirable, yet previously-unexplored attributes. The underlying objective function includes a spatially-adaptive penalty that couples the latent sharp image, non-uniform blur operator, and noise level together. This coupling allows the penalty to automatically adjust its shape based on the estimated degree of local blur and image structure such that regions with large blur or few prominent edges are discounted. Remaining regions with modest blur and revealing edges therefore dominate the overall estimation process without explicitly incorporating structure-selection heuristics. The algorithm can be implemented using an optimization strategy that is virtually parameter free and simpler than existing methods. Detailed theoretical analysis and empirical validation on real images serve to validate the proposed method.

SPOTLIGHT SESSION

Session 2, 11:40 AM – 12:05 PM

- **Provable Subspace Clustering: When LRR meets SSC**
  Y. Wang, National University of Singapore; H. Xu, NUS; C. Leng, University of Warwick
  See abstract F23, page 52

- **Matrix Completion From any Given Set of Observations**
  T. Lee, Centre for Quantum Technologies; A. Shraibman, Weizmann Institute of Science
  See abstract F27, page 53

- **Convex Two-Layer Modeling**
  Ö. Aslan, H. Cheng, D. Schuurmans, University of Alberta; X. Zhang, NICTA
  See abstract F81, page 66

- **Reconciling “priors” & “priors” without prejudice?**
  R. Gribonval, P. Machart, INRIA
  See abstract F34, page 55

- **Robust Sparse Principal Component Regression under the High Dimensional Elliptical Model**
  F. Han, Johns Hopkins University; H. Liu, Princeton University
  See abstract F36, page 55

- **Structured Learning via Logistic Regression**
  J. Domke, NICTA
  See abstract F69, page 64

ORAL SESSION

Session 3, 2:00 – 3:30 PM

Session Chair: Terry Sejnowski

INVITED TALK: Memory Reactivation in Awake and Sleep States

Matthew Wilson mwilson@mit.edu
MIT

By introducing arrays of microelectrodes into hippocampal, thalamic, and neocortical areas of freely behaving rodents, we have characterized the detailed structure and content of memory patterns across ensembles of individual neurons as they are formed during spatial behavior, and reactivated during quiet wakefulness, and sleep. I will describe the contributions of these brain systems to the expression and coordination of memory reactivation, including recent results demonstrating the ability to influence reactivated memory content during sleep.

Matthew Wilson received his bachelor’s degree in Electrical Engineering from Rensselaer Polytechnic Institute in 1983, his master’s degree in Electrical Engineering from the University of Wisconsin, Madison in 1986, and his Ph.D. in Computation and Neural Systems from the California Institute of Technology in 1990. In 1991 he began his work *studying the formation of memory in the hippocampus using large-scale multiple electrode recording of neuronal ensembles in the hippocampus of freely behaving rats at the University of Arizona, Tucson. He continues to study the mechanisms of memory formation in the rodent at the Massachusetts Institute of Technology as a member of the faculty of the Departments of Brain and Cognitive Sciences, and Biology, the Picower Center for Learning and Memory, and the RIKEN-MIT Neuroscience Research Center.*

Correlations strike back (again): the case of associative memory retrieval

Cristina Savin cs664@cam.ac.uk
Mate Lengyel m.lengyel@eng.cam.ac.uk
University of Cambridge
Peter Dayan dayan@gatsby.ucl.ac.uk
Gatsby Unit, UCL

It has long been recognised that statistical dependencies in neuronal activity need to be taken into account when decoding stimuli encoded in a neural population. Less studied, though equally pernicious, is the need to take account of dependencies between synaptic weights when decoding patterns previously encoded in an auto-associative memory. We show that activity-dependent learning generically produces such correlations, and failing to take them into account in the dynamics of memory retrieval leads to catastrophically poor recall. We derive optimal network dynamics for recall in the face of synaptic correlations caused by a range of synaptic plasticity rules. These dynamics involve well-studied circuit motifs, such as forms of feedback inhibition and experimentally observed dendritic nonlinearities. We therefore show how addressing the problem of synaptic correlations leads to a novel functional account of key biophysical features of the neural substrate.

A Memory Frontier for Complex Synapses

Subhaneil Lahiri sulahiri@stanford.edu
Surya Ganguli sganguli@stanford.edu
Stanford University

An incredible gulf separates theoretical models of synapses, often described solely by a single scalar value denoting the size of a postsynaptic potential, from the immense complexity of molecular signaling pathways underlying real synapses. To understand the functional contribution of such molecular complexity to learning and memory, it is essential to expand our theoretical conception of a synapse from a single scalar to an entire dynamical system with many internal molecular functional states. Moreover, theoretical considerations alone demand such an expansion; network models with scalar synapses assuming finite numbers of distinguishable synaptic strengths have strikingly limited memory capacity. This raises the fundamental question, how does synaptic complexity give rise to memory? To address this, we develop new mathematical theorems elucidating the relationship between the structural organization and memory properties of complex synapses that are themselves molecular networks. Moreover, in proving such theorems, we uncover a framework, based on first passage time theory, to impose an order on the internal states of complex synaptic models, thereby simplifying the relationship between synaptic structure and function.

SPOTLIGHT SESSION

Session 3, 3:30 – 3:50 PM

- **Bayesian entropy estimation for binary spike train data using parametric prior knowledge**
  E. Archer, I. Park, J. Pillow, UT Austin
  See abstract F43, page 57

- **Inferring neural population dynamics from multiple partial recordings of the same neural circuit**
  S. Turaga, L. Buesing, Gatsby Unit, UCL; A. Packer, H. Dalgleish, N. Pettit, M. Hausser, UCL; J. Macke, MPI for Biological Cybernetics
  See abstract F46, page 58

- **Noise-Enhanced Associative Memories**
  A. Karbasi, ETH Zurich; A. Salavati, A. Shokrollahi, EPFL; L. Varshney, IBM Watson Research Center
  See abstract F49, page 59

- **Demixing odors - fast inference in olfaction**
  A. Grabska-Barwinska, J. Beck, P. Latham, Gatsby Unit, UCL; A. Pouget, University of Geneva
  See abstract F54, page 60

- **Recurrent linear models of simultaneously-recorded neural populations**
  M. Pachitariu, M. Sahani, Gatsby Unit, UCL; B. Petreska, UCL
  See abstract F53, page 60

ORAL SESSION

Session 4, 4:15 – 5:40 PM

Session Chair: Emily Fox

NIPS Award 2 - 4:15 – 4:20 PM

Understanding Dropout

Pierre Baldi pfbaldi@ics.uci.edu
Peter Sadowski peterjsadowski@gmail.com
UC Irvine

Dropout is a relatively new algorithm for training neural networks which relies on stochastically “dropping out” neurons during training in order to avoid the co-adaptation of feature detectors. We introduce a general formalism for studying dropout on either units or connections, with arbitrary probability values, and use it to analyze the averaging and regularizing properties of dropout in both linear and non-linear networks. For deep neural networks, the averaging properties of dropout are characterized by three recursive equations, including the approximation of expectations by normalized weighted geometric means. We provide estimates and bounds for these approximations and corroborate the results with simulations. We also show in simple cases how dropout performs stochastic gradient descent on a regularized error function.

Annealing between distributions by averaging moments

Roger Grosse rgrosse@mit.edu
Massachusetts Institute of Technology
Chris J Maddison cmaddis@cs.toronto.edu
Russ Salakhutdinov rsalakhu@cs.toronto.edu
University of Toronto

Many powerful Monte Carlo techniques for estimating partition functions, such as annealed importance sampling (AIS), are based on sampling from a sequence of intermediate distributions which interpolate between a tractable initial distribution and an intractable target distribution. The near-universal practice is to use geometric averages of the initial and target distributions, but alternative paths can perform substantially better. We present a novel sequence of intermediate distributions for exponential families: averaging the moments of the initial and target distributions. We derive an asymptotically optimal piecewise linear schedule for the moments path and show that it performs at least as well as geometric averages with a linear schedule. Moment averaging performs well empirically at estimating partition functions of restricted Boltzmann machines (RBMs), which form the building blocks of many deep learning models, including Deep Belief Networks and Deep Boltzmann Machines.

A simple example of Dirichlet process mixture inconsistency for the number of components

Jeff Miller jeffrey_miller@brown.edu
Matthew Harrison Matthew_Harrison@Brown.edu
Brown University

For data assumed to come from a finite mixture with an unknown number of components, it has become common to use Dirichlet process mixtures (DPMs) not only for density estimation, but also for inferences about the number of components. The typical approach is to use the posterior distribution on the number of components occurring so far --- that is, the posterior on the number of clusters in the observed data. However, it turns out that this posterior is not consistent --- it does not converge to the true number of components. In this note, we give an elementary demonstration of this inconsistency in what is perhaps the simplest possible setting: a DPM with normal components of unit variance, applied to data from a “mixture” with one standard normal component. Further, we find that this example exhibits severe inconsistency: instead of going to 1, the posterior probability that there is one cluster goes to 0.

Approximate Bayesian Image Interpretation using Generative Probabilistic Graphics Programs

Vikash Mansinghka vkm@mit.edu
Tejas Kulkarni tejask@mit.edu
Yura Perov perov@mit.edu
Josh Tenenbaum jbt@mit.edu
Massachusetts Institute of Technology

The idea of computer vision as the Bayesian inverse problem to computer graphics has a long history and an appealing elegance, but it has proved difficult to directly implement. Instead, most vision tasks are approached via complex bottom-up processing pipelines. Here we show that it is possible to write short, simple probabilistic graphics programs that define flexible generative models and to automatically invert them to interpret real-world images. Generative probabilistic graphics programs consist of a stochastic scene generator, a renderer based on graphics software, a stochastic likelihood model linking the renderer’s output and the data, and latent variables that adjust the fidelity of the renderer and the tolerance of the likelihood model. Representations and algorithms from computer graphics, originally designed to produce high-quality images, are instead used as the deterministic backbone for highly approximate and stochastic generative models. This formulation combines probabilistic programming, computer graphics, and approximate Bayesian computation, and depends only on general-purpose, automatic inference techniques. We describe two applications: reading sequences of degraded and adversarially obscured alphanumeric characters, and inferring 3D road models from vehicle-mounted camera images. Each of the probabilistic graphics programs we present relies on under 20 lines of probabilistic code, and supports accurate, approximately Bayesian inferences about ambiguous real-world images.

SPOTLIGHT SESSION

Session 4, 5:40 – 6:00 PM

- **Dropout Training as Adaptive Regularization**
  S. Wager, S. Wang, P. Liang, Stanford University
  See abstract F80, page 66

- **Stochastic Gradient Riemannian Langevin Dynamics on the Probability Simplex**
  S. Patterson, Gatsby Unit, UCL; Y. Teh, University of Oxford
  See abstract F32, page 54

- **Restricting exchangeable nonparametric distributions**
  S. Williamson, UT Austin; S. MacEachern, Ohio State University; E. Xing, CMU
  See abstract F38, page 56

- **Approximate inference in latent Gaussian-Markov models from continuous time observations**
  B. Cseke, G. Sanguinetti, University of Edinburgh; M. Opper, TU Berlin
  See abstract F71, page 64

- **Bayesian inference as iterated random functions with applications to sequential inference in graphical models**
  A. Amini, X. Nguyen, University of Michigan
  See abstract F75, page 65

POSTER SESSION

7:00 – 11:59 PM

F1 **A Deep Architecture for Matching Short Texts**
Z. Lu, H. Li

F2 **On the Expressive Power of Restricted Boltzmann Machines**
J. Martens, A. Chattopadhya, T. Pitassi, R. Zemel

F3 **Distributed Representations of Words and Phrases and their Compositionality**
T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean

F4 **Stochastic Ratio Matching of RBMs for Sparse High-Dimensional Inputs**
Y. Dauphin, Y. Bengio

F5 **Generalized Denoising Auto-Encoders as Generative Models**
Y. Bengio, L. Yao, G. Alain, P. Vincent

F6 **Multi-Prediction Deep Boltzmann Machines**
I. Goodfellow, M. Mirza, A. Courville, Y. Bengio

F7 **Predicting Parameters in Deep Learning**
M. Denil, B. Shakibi, L. Dinh, M. Ranzato, N. de Freitas

F8 **Learning Stochastic Feedforward Neural Networks**
Y. Tang, R. Salakhutdinov

F9 **Zero-Shot Learning Through Cross-Modal Transfer**
R. Socher, M. Ganjoo, C. Manning, A. Ng

F10 **Reasoning With Neural Tensor Networks for Knowledge Base Completion**
R. Socher, D. Chen, C. Manning, A. Ng

F11 **Discriminative Transfer Learning with Tree-based Priors**
N. Srivastava, R. Salakhutdinov

F12 **Robust Image Denoising with Multi-Column Deep Neural Networks**
F. Agostinelli, M. Anderson, H. Lee

F13 **Annealing between distributions by averaging moments**
R. Grosse, C. Maddison, R. Salakhutdinov

F14 **Top-Down Regularization of Deep Belief Networks**
H. Goh, N. Thome, M. Cord, J. Lim

F15 **Adaptive dropout for training deep neural networks**
J. Ba, B. Frey

F16 **Stochastic Optimization of PCA with Capped MSG**
R. Arora, A. Cotter, N. Srebro

F17 **Variance Reduction for Stochastic Gradient Optimization**
C. Wang, X. Chen, A. Smola, E. Xing

F18 **Memory Limited, Streaming PCA**
I. Mitliagkas, C. Caramanis, P. Jain

F19 **Near-Optimal Entrywise Sampling for Data Matrices**
D. Achlioptas, Z. Karnin, E. Liberty

F20 **Large Scale Distributed Sparse Precision Estimation**
H. Wang, A. Banerjee, C. Hsieh, P. Ravikumar, I. Dhillon

F21 **Optimistic Concurrency Control for Distributed Unsupervised Learning**
X. Pan, J. Gonzalez, S. Jegelka, T. Broderick, M. Jordan

F22 **Distributed Submodular Maximization: Identifying Representative Elements in Massive Data**
B. Mirzasoleiman, A. Karbasi, R. Sarkar, A. Krause

F23 **Provable Subspace Clustering: When LRR meets SSC**
Y. Wang, H. Xu, C. Leng

F24 **Simultaneous Rectification and Alignment via Robust Recovery of Low-rank Tensors**
X. Zhang, D. Wang, Z. Zhou, Y. Ma

F25 **Phase Retrieval using Alternating Minimization**
P. Netrapalli, P. Jain, S. Sanghavi

F26 **Non-Uniform Camera Shake Removal Using a Spatially-Adaptive Sparse Penalty**
H. Zhang, D. Wipf

F27 **Matrix Completion From any Given Set of Observations**
T. Lee, A. Shraibman

F28 **Machine Teaching for Bayesian Learners in the Exponential Family**
X. Zhu

F29 **Approximate Bayesian Image Interpretation using Generative Probabilistic Graphics Programs**
V. Mansinghka, T. Kulkarni, Y. Perov, J. Tenenbaum

F30 **Analyzing Hogwild Parallel Gaussian Gibbs Sampling**
M. Johnson, J. Saunderson, A. Willsky

F31**Flexible sampling of discrete data correlations without the marginal distributions**
A. Kalaitzis, R. Silva

F32 **Stochastic Gradient Riemannian Langevin Dynamics on the Probability Simplex**
S. Patterson, Y. Teh

F33 **Auxiliary-variable Exact Hamiltonian Monte Carlo Samplers for Binary Distributions**
A. Pakman, L. Paninski

F34 **Reconciling “priors” & “priors” without prejudice?**
R. Gribonval, P. Machart

F35 **Wavelets on Graphs via Deep Learning**
R. Rustamov, L. Guibas

F36 **Robust Sparse Principal Component Regression under the High Dimensional Elliptical Model**
F. Han, H. Liu

F37 **A simple example of Dirichlet process mixture inconsistency for the number of components**
J. Miller, M. Harrison

F38 **Restricting exchangeable nonparametric distributions**
S. Williamson, S. MacEachern, E. Xing

F39 **Stochastic blockmodel approximation of a graphon: Theory and consistent estimation**
E. Airoldi, T. Costa, S. Chan

F40 **Bayesian Hierarchical Community Discovery**
C. Blundell, Y. Teh

F41 **Scalable Influence Estimation in Continuous-Time Diffusion Networks**
N. Du, L. Song, M. Gomez-Rodriguez, H. Zha

F42 **Nonparametric Multi-group Membership Model for Dynamic Networks**
M. Kim, J. Leskovec

F43 **Bayesian entropy estimation for binary spike train data using parametric prior knowledge**
E. Archer, I. Park, J. Pillow

F44 **Universal models for binary spike patterns using centered Dirichlet processes**
I. Park, E. Archer, K. Latimer, J. Pillow

F45 **A Determinantal Point Process Latent Variable Model for Inhibition in Neural Spiking Data**
J. Snoek, R. Zemel, R. Adams

F46 **Inferring neural population dynamics from multiple partial recordings of the same neural circuit**
S. Turaga, L. Buesing, A. Packer, H. Dalgleish, N. Pettit, M. Hausser, J. Macke

F47 **Neural representation of action sequences: how far can a simple snippet-matching model take us?**
C. Tan, J. Singer, T. Serre, D. Sheinberg, T. Poggio

F48**Firing rate predictions in optimal balanced networks**
D. Barrett, S. Denève, C. Machens

F49 **Noise-Enhanced Associative Memories**
A. Karbasi, A. Salavati, A. Shokrollahi, L. Varshney

F50 **A memory frontier for complex synapses**
S. Lahiri, S. Ganguli

F51 **Perfect Associative Learning with Spike-Timing-Dependent Plasticity**
C. Albers, M. Westkott, K. Pawelzik

F52 **Reciprocally Coupled Local Estimators Implement Bayesian Information Integration Distributively**
W. Zhang, S. Wu

F53 **Recurrent linear models of simultaneously-recorded neural populations**
M. Pachitariu, B. Petreska, M. Sahani

F54 **Demixing odors - fast inference in olfaction**
A. Grabska-Barwinska, J. Beck, A. Pouget, P. Latham

F55 **Multisensory Encoding, Decoding, and Identification**
A. Lazar, Y. Slutskiy

F56 **Recurrent networks of coupled Winner-Take-All oscillators for solving constraint satisfaction problems**
H. Mostafa, L. Muller, G. Indiveri

F57 **Capacity of strong attractor patterns to model behavioural and cognitive prototypes**
A. Edalat

F58 **Compete to Compute**
R. Srivastava, J. Masci, S. Kazerounian, F. Gomez, J. Schmidhuber

F59 **Understanding Dropout**
P. Baldi, P. Sadowski

F60 **RNADE: The real-valued neural autoregressive density-estimator**
B. Uria, I. Murray, H. Larochelle

F61 **Correlations strike back (again): the case of associative memory retrieval**
C. Savin, P. Dayan, M. Lengyel

F62 **Real-Time Inference for a Gamma Process Model of Neural Spiking**
D. Carlson, V. Rao, J. Vogelstein, L. Carin

F63 **Transportability from Multiple Environments with Limited Experiments**
E. Bareinboim, S. Lee, V. Honavar, J. Pearl

F64 **Causal Inference on Time Series using Restricted Structural Equation Models**
J. Peters, D. Janzing, B. Schölkopf

F65 **Discovering Hidden Variables in Noisy-Or Networks using Quartet Tests**
Y. Jernite, Y. Halpern, D. Sontag

F66 **Learning Hidden Markov Models from Non-sequence Data via Tensor Decomposition**
T. Huang, J. Schneider

F67 **Unsupervised Spectral Learning of Finite State Transducers**
R. Bailly, X. Carreras, A. Quattoni

F68 **Learning Efficient Random Maximum A-Posteriori Predictors with Non-Decomposable Loss Functions**
T. Hazan, S. Maji, J. Keshet, T. Jaakkola

F69 **Structured Learning via Logistic Regression**
J. Domke

F70 **Graphical Models for Inference with Missing Data**
K. Mohan, J. Pearl, J. Tian

F71 **Approximate inference in latent Gaussian-Markov models from continuous time observations**
B. Cseke, M. Opper, G. Sanguinetti

F72 **Variational Planning for Graph-based MDPs**
Q. Cheng, Q. Liu, F. Chen, A. Ihler

F73 **Integrated Non-Factorized Variational Inference**
S. Han, X. Liao, L. Carin

F74 **Global Solver and Its Efficient Approximation for Variational Bayesian Low-rank Subspace Clustering**
S. Nakajima, A. Takeda, S. Babacan, M. Sugiyama, I. Takeuchi

F75 **Bayesian inference as iterated random functions with applications to sequential inference in graphical models**
A. Amini, X. Nguyen

F76 **Learning to Pass Expectation Propagation Messages**
N. Heess, D. Tarlow, J. Winn

F77 **On the Complexity and Approximation of Binary Evidence in Lifted Inference**
G. Van den Broeck, A. Darwiche

F78 **Translating Embeddings for Modeling Multi-relational Data**
A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, O. Yakhnenko

F79 **Efficient Online Inference for Bayesian Nonparametric Relational Models**
D. Kim, P. Gopalan, D. Blei, E. Sudderth

F80 **Dropout Training as Adaptive Regularization**
S. Wager, S. Wang, P. Liang

F81 **Convex Two-Layer Modeling**
Ö. Aslan, H. Cheng, X. Zhang, D. Schuurmans

F82 **Learning with Noisy Labels**
N. Natarajan, I. Dhillon, P. Ravikumar, A. Tewari

F83 **Low-rank matrix reconstruction and clustering via approximate message passing**
R. Matsushita, T. Tanaka

F84 **Matrix factorization with binary components**
M. Slawski, M. Hein, P. Lutsik

F85 **Learning Multi-level Sparse Representations**
F. Diego Andilla, F. Hamprecht

F86 **Exact and Stable Recovery of Pairwise Interaction Tensors**
S. Chen, M. Lyu, I. King, Z. Xu

F87 **A New Convex Relaxation for Tensor Completion**
B. Romera-Paredes, M. Pontil

F88 **On Decomposing the Proximal Map**
Y. Yu

F89 **Adaptive Anonymity via b-Matching**
K. Choromanski, T. Jebara, K. Tang

DEMONSTRATIONS

7:00 – 11:59 PM

Accelerating Deep Neural Networks on Mobile Processor with Embedded Programmable Logic, E. Culurciello, A. Dundar, J. Jin, V. Gokhale, B. Martini

A Mobile Development Platform for Adaptive Machine Learning and Neuromorphic Computing in Robotics, J. Hunt, P. O'Connor

Codewebs: a Pedagogical Search Engine for Code Submissions to a MOOC, J. Huang, C. Piech, A. Nguyen, L. Guibas

Controlling Robot Dynamics With Spiking Neurons, S. Menon, S. Fok, K. Boahen

Di-BOSS™: Digital Building Operating System Solution, J. Forde, V. Rathod, H. Shookri, V. Bandari, A. Rajan, J. Min, A. Fan, L. Wu, A. Gagneja, D. Riecken, D. Solomon, L. Hannah, A. Boulanger, R. Anderson

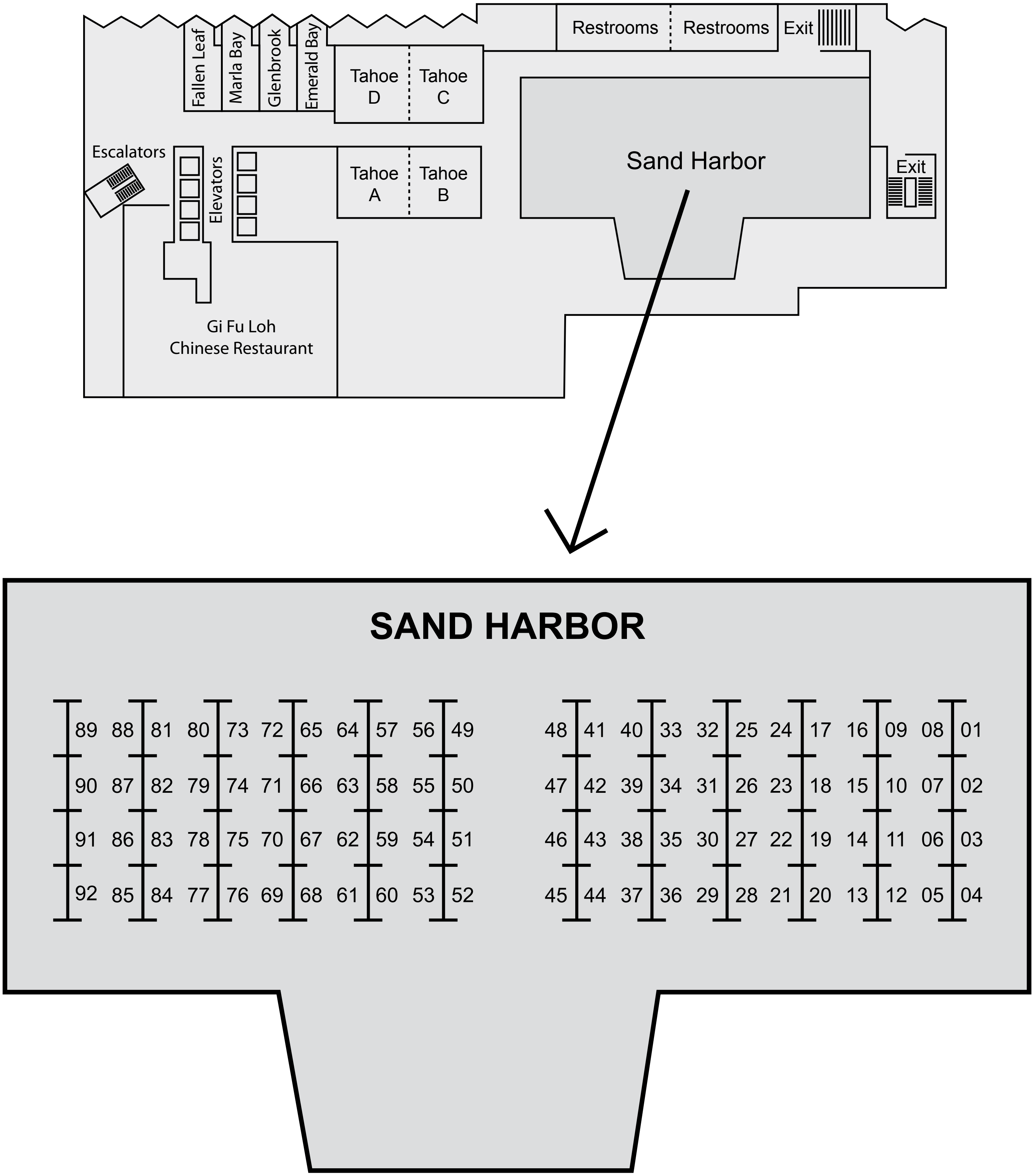
NCS: A Novel CPU/GPU Simulation Environment for Large-Scale Biologically-Realistic Neural Modeling, R. Hoang, D. Tanna, L. Jayet Bray, S. Dascalu, F. Harris, Jr

The Three-Weight Algorithm: Enhancing ADMM for Large-Scale Distributed Optimization, N. Derbinsky, J. Bento, J. Yedidia

FRIDAY POSTER FLOORPLAN

HARRAH'S

2ND FLOOR SPECIAL EVENTS CENTER



89	88	81	80	73	72	65	64	57	56	49	48	41	40	33	32	25	24	17	16	09	08	01
90	87	82	79	74	71	66	63	58	55	50	47	42	39	34	31	26	23	18	15	10	07	02
91	86	83	78	75	70	67	62	59	54	51	46	43	38	35	30	27	22	19	14	11	06	03
92	85	84	77	76	69	68	61	60	53	52	45	44	37	36	29	28	21	20	13	12	05	04

F1 A Deep Architecture for Matching Short Texts

Zhengdong Lu lu.zhengdong@huawei.com
Hang Li hangli.hl@huawei.com
Noah's Ark Lab, Huawei Technologies

Many machine learning problems can be interpreted as learning for matching two types of objects (e.g., images and captions, users and products, queries and documents). The matching level of two objects is usually measured as the inner product in a certain feature space, while the modeling effort focuses on mapping of objects from the original space to the feature space. This schema, although proven successful on a range of matching tasks, is insufficient for capturing the rich structure in the matching process of more complicated objects. In this paper, we propose a new deep architecture to more effectively model the complicated matching relations between two objects from heterogeneous domains. More specifically, we apply this model to matching tasks in natural language, e.g., finding sensible responses for a tweet, or relevant answers to a given question. This new architecture naturally combines the localness and hierarchy intrinsic to the natural language problems, and therefore greatly improves upon the state-of-the-art models.

F2 On the Expressive Power of Restricted Boltzmann Machines

James Martens jmartens@cs.toronto.edu
Toni Pitassi toni@cs.toronto.edu
Richard Zemel zemel@cs.toronto.edu
University of Toronto
Arkadev Chattopadhya arkadev.c@mailhost.tifr.res.in
Tata Institute of Fundamental Research

This paper examines the question: What kinds of distributions can be efficiently represented by Restricted Boltzmann Machines (RBMs)? We characterize the RBM's unnormalized log-likelihood function as a type of neural network (called an RBM network), and through a series of simulation results relate these networks to types that are better understood. We show the surprising result that RBM networks can efficiently compute any function that depends on the number of 1's in the input, such as parity. We also provide the first known example of a particular type of distribution which provably cannot be efficiently represented by an RBM (or equivalently, cannot be efficiently computed by an RBM network), assuming a realistic exponential upper bound on the size of the weights. By formally demonstrating that a relatively simple distribution cannot be represented efficiently by an RBM our results provide a new rigorous justification for the use of potentially more expressive generative models, such as deeper ones.

F3 Distributed Representations of Words and Phrases and their Compositionality

Tomas Mikolov tmikolov@google.com
Ilya Sutskever ilyasu@google.com
Kai Chen kaichen@google.com
Greg Corrado gcorrado@google.com
Jeff Dean jeff@google.com
Google Research

The recently introduced continuous Skip-gram model is an efficient method for learning high-quality distributed vector representations that capture a large number of precise syntactic and semantic word relationships. In this paper we present several improvements that make the Skip-gram model more expressive and enable it to learn higher quality vectors more rapidly. We show that by subsampling frequent words we obtain significant speedup, and also learn higher quality representations as measured by our tasks. We also introduce Negative Sampling, a simplified variant of Noise Contrastive Estimation (NCE) that learns more accurate vectors for frequent words compared to the hierarchical softmax. An inherent limitation of word representations is their indifference to word order and their inability to represent idiomatic phrases. For example, the meanings of "Canada'' and "Air'' cannot be easily combined to obtain "Air Canada''. Motivated by this example, we present a simple and efficient method for finding phrases, and show that their vector representations can be accurately learned by the Skip-gram model.

F4 Stochastic Ratio Matching of RBMs for Sparse High-Dimensional Inputs

Yann Dauphin yann-nicolas.dauphin@umontreal.ca
Yoshua Bengio bengioy@iro.umontreal.ca
University of Montreal

Sparse high-dimensional data vectors are common in many application domains where a very large number of rarely non-zero features can be devised. Unfortunately, this creates a computational bottleneck for unsupervised feature learning algorithms such as those based on auto-encoders and RBMs, because they involve a reconstruction step where the whole input vector is predicted from the current feature values. An algorithm was recently developed to successfully handle the case of auto-encoders, based on an importance sampling scheme stochastically selecting which input elements to actually reconstruct during training for each particular example. To generalize this idea to RBMs, we propose a stochastic ratio-matching algorithm that inherits all the computational advantages and unbiasedness of the importance sampling scheme. We show that stochastic ratio matching is a good estimator, allowing the approach to beat the state-of-the-art on two bag-of-word text classification benchmarks (20 Newsgroups and RCV1), while keeping computational cost linear in the number of non-zeros.

F5 Generalized Denoising Auto-Encoders as Generative Models

Yoshua Bengio bengioy@iro.umontreal.ca
Li Yao yaoli.email@gmail.com
Guillaume Alain alaingui@iro.umontreal.ca
Pascal Vincent vincentp@iro.umontreal.ca
University of Montreal

Recent work has shown how denoising and contractive autoencoders implicitly capture the structure of the data generating density, in the case where the corruption noise is Gaussian, the reconstruction error is the squared error, and the data is continuous-valued. This has led to various proposals for sampling from this implicitly learned density function, using Langevin and Metropolis-Hastings MCMC. However, it remained unclear how to connect the training procedure of regularized auto-encoders to the implicit estimation of the underlying data generating distribution when the data are discrete, or using other forms of corruption process and reconstruction errors. Another issue is the mathematical justification which is only valid in the limit of small corruption noise. We propose here a different attack on the problem, which deals with all these issues: arbitrary (but noisy enough) corruption, arbitrary reconstruction loss (seen as a log-likelihood), handling both discrete and continuous-valued variables, and removing the bias due to non-infinitesimal corruption noise (or non-infinitesimal contractive penalty).

F6 Multi-Prediction Deep Boltzmann Machines

Ian Goodfellow goodfeli@iro.umontreal.ca
Mehdi Mirza mirzamom@iro.umontreal.ca
Aaron Courville aaron.courville@umontreal.ca
Yoshua Bengio bengioy@iro.umontreal.ca
University of Montreal

We introduce the Multi-Prediction Deep Boltzmann Machine (MP-DBM). The MP-DBM can be seen as a single probabilistic model trained to maximize a variational approximation to the generalized pseudolikelihood, or as a family of recurrent nets that share parameters and approximately solve different inference problems. Prior methods of training DBMs either do not perform well on classification tasks or require an initial learning pass that trains the DBM greedily, one layer at a time. The MP-DBM does not require greedy layerwise pretraining, and outperforms the standard DBM at classification, classification with missing inputs, and mean field prediction tasks.

F7 Predicting Parameters in Deep Learning

Misha Denil mdenil@cs.ubc.ca
Babak Shakibi bshakibi@cs.ubc.ca
Nando de Freitas nando@cs.ubc.ca
UBC
Laurent Dinh laurent.dinh@centraliens.net
École Centrale Paris
Marc'Aurelio Ranzato ranzato@google.com
Google Research

We demonstrate that there is significant redundancy in the parameterization of several deep learning models. Given only a few weight values for each feature it is possible to accurately predict the remaining values. Moreover, we show that not only can the parameter values be predicted, but many of them need not be learned at all. We train several different architectures by learning only a small number of weights and predicting the rest. In the best case we are able to predict more than 95% of the weights of a network without any drop in accuracy.

F8 Learning Stochastic Feedforward Neural Networks

Yichuan Tang clarecorp@gmail.com
Russ Salakhutdinov rsalakhu@cs.toronto.edu
University of Toronto

Multilayer perceptrons (MLPs) or neural networks are popular models used for nonlinear regression and classification tasks. As regressors, MLPs model the conditional distribution of the predictor variables Y given the input variables X. However, this predictive distribution is assumed to be unimodal (e.g. Gaussian). For tasks such as structured prediction problems, the conditional distribution should be multimodal, forming one-to-many mappings. By using stochastic hidden variables rather than deterministic ones, Sigmoid Belief Nets (SBNs) can induce a rich multimodal distribution in the output space. However, previously proposed learning algorithms for SBNs are very slow and do not work well for real-valued data. In this paper, we propose a stochastic feedforward network with hidden layers having \emph{both deterministic and stochastic} variables. A new Generalized EM training procedure using importance sampling allows us to efficiently learn complicated conditional distributions. We demonstrate the superiority of our model to conditional Restricted Boltzmann Machines and Mixture Density Networks on synthetic datasets and on modeling facial expressions. Moreover, we show that latent features of our model improves classification and provide additional qualitative results on color images.

F9 Zero-Shot Learning Through Cross-Modal Transfer

Richard Socher	richard@socher.org
Milind Ganjoo	mganjoo@stanford.edu
Christopher Manning	manning@stanford.edu
Andrew Ng	ang@cs.stanford.edu
Stanford University	

This work introduces a model that can recognize objects in images even if no training data is available for the object class. The only necessary knowledge about unseen categories comes from unsupervised text corpora. Unlike previous zero-shot learning models, which can only differentiate between unseen classes, our model can operate on a mixture of objects, simultaneously obtaining state of the art performance on classes with thousands of training images and reasonable performance on unseen classes. This is achieved by seeing the distributions of words in texts as a semantic space for understanding what objects look like. Our deep learning model does not require any manually defined semantic or visual features for either words or images. Images are mapped to be close to semantic word vectors corresponding to their classes, and the resulting image embeddings can be used to distinguish whether an image is of a seen or unseen class. Then, a separate recognition model can be employed for each type. We demonstrate two strategies, the first gives high accuracy on unseen classes, while the second is conservative in its prediction of novelty and keeps the seen classes' accuracy high.

F10 Reasoning With Neural Tensor Networks for Knowledge Base Completion

Richard Socher	richard@socher.org
Christopher Manning	manning@stanford.edu
Andrew Ng	ang@cs.stanford.edu
Stanford University	
Danqi Chen	danqi@stanford.edu
Saarland University	

A common problem in knowledge representation and related fields is reasoning over a large joint knowledge graph, represented as triples of a relation between two entities. The goal of this paper is to develop a more powerful neural network model suitable for inference over these relationships. Previous models suffer from weak interaction between entities or simple linear projection of the vector space. We address these problems by introducing a neural tensor network (NTN) model which allow the entities and relations to interact multiplicatively. Additionally, we observe that such knowledge base models can be further improved by representing each entity as the average of vectors for the words in the entity name, giving an additional dimension of similarity by which entities can share statistical strength. We assess the model by considering the problem of predicting additional true relations between entities given a partial knowledge base. Our model outperforms previous models and can classify unseen relationships in WordNet and FreeBase with an accuracy of 86.2% and 90.0%, respectively.

F11 Discriminative Transfer Learning with Tree-based Priors

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

This paper proposes a way of improving classification performance for classes which have very few training examples. The key idea is to discover classes which are similar and transfer knowledge among them. Our method organizes the classes into a tree hierarchy. The tree structure can be used to impose a generative prior over classification parameters. We show that these priors can be combined with discriminative models such as deep neural networks. Our method benefits from the power of discriminative training of deep neural networks, at the same time using tree-based generative priors over classification parameters. We also propose an algorithm for learning the underlying tree structure. This gives the model some flexibility to tune the tree so that the tree is pertinent to task being solved. We show that the model can transfer knowledge across related classes using fixed semantic trees. Moreover, it can learn new meaningful trees usually leading to improved performance. Our method achieves state-of-the-art classification results on the CIFAR-100 image data set and the MIR Flickr multimodal data set.

F12 Robust Image Denoising with Multi-Column Deep Neural Networks

Forest Agostinelli	agostifo@umich.edu
Michael Anderson	mrander@umich.edu
Honglak Lee	honglak@eecs.umich.edu
University of Michigan	

Stacked sparse denoising auto-encoders (SSDAs) have recently been shown to be successful at removing noise from corrupted images. However, like most denoising techniques, the SSDA is not robust to variation in noise types beyond what it has seen during training. We present the multi-column stacked sparse denoising autoencoder, a novel technique of combining multiple SSDAs into a multi-column SSDA (MC-SSDA) by combining the outputs of each SSDA. We eliminate the need to determine the type of noise, let alone its statistics, at test time. We show that good denoising performance can be achieved with a single system on a variety of different noise types, including ones not seen in the training set. Additionally, we experimentally demonstrate the efficacy of MC-SSDA denoising by achieving MNIST digit error rates on denoised images at close to that of the uncorrupted images.

F13 Annealing between distributions by averaging moments

Roger Grosse rgrosse@mit.edu
Massachusetts Institute of Technology
Chris J Maddison cmaddis@cs.toronto.edu
Russ Salakhutdinov rsalakhu@cs.toronto.edu
University of Toronto

Many powerful Monte Carlo techniques for estimating partition functions, such as annealed importance sampling (AIS), are based on sampling from a sequence of intermediate distributions which interpolate between a tractable initial distribution and an intractable target distribution. The near-universal practice is to use geometric averages of the initial and target distributions, but alternative paths can perform substantially better. We present a novel sequence of intermediate distributions for exponential families: averaging the moments of the initial and target distributions. We derive an asymptotically optimal piecewise linear schedule for the moments path and show that it performs at least as well as geometric averages with a linear schedule. Moment averaging performs well empirically at estimating partition functions of restricted Boltzmann machines (RBMs), which form the building blocks of many deep learning models, including Deep Belief Networks and Deep Boltzmann Machines.

F14 Top-Down Regularization of Deep Belief Networks

Hanlin Goh hlgoh@i2r.a-star.edu.sg
LIP6/UPMC
Nicolas Thome nicolas.thome@lip6.fr
Matthieu Cord matthieu.cord@lip6.fr
University Pierre & Marie Curie and CNRS
Joo-Hwee Lim joohwee@i2r.a-star.edu.sg
Institute for Infocomm Research, Singapore

Designing a principled and effective algorithm for learning deep architectures is a challenging problem. The current approach involves two training phases: a fully unsupervised learning followed by a strongly discriminative optimization. We suggest a deep learning strategy that bridges the gap between the two phases, resulting in a three-phase learning procedure. We propose to implement the scheme using a method to regularize deep belief networks with top-down information. The network is constructed from building blocks of restricted Boltzmann machines learned by combining bottom-up and top-down sampled signals. A global optimization procedure that merges samples from a forward bottom-up pass and a top-down pass is used. Experiments on the MNIST dataset show improvements over the existing algorithms for deep belief networks. Object recognition results on the Caltech-101 dataset also yield competitive results.

F15 Adaptive dropout for training deep neural networks

Jimmy Ba jimmy@psi.utoronto.ca
Brendan Frey frey@psi.utoronto.ca
University of Toronto

Recently, it was shown that by dropping out hidden activities with a probability of 0.5, deep neural networks can perform very well. We describe a model in which a binary belief network is overlaid on a neural network and is used to decrease the information content of its hidden units by selectively setting activities to zero. This ‘’dropout network” can be trained jointly with the neural network by approximately computing local expectations of binary dropout variables, computing derivatives using back-propagation, and using stochastic gradient descent. Interestingly, experiments show that the learnt dropout network parameters recapitulate the neural network parameters, suggesting that a good dropout network regularizes activities according to magnitude. When evaluated on the MNIST and NORB datasets, we found our method can be used to achieve lower classification error rates than other feather learning methods, including standard dropout, denoising auto-encoders, and restricted Boltzmann machines. For example, our model achieves 5.8% error on the NORB test set, which is better than state-of-the-art results obtained using convolutional architectures.

F16 Stochastic Optimization of PCA with Capped MSG

Raman Arora rmnarora@gmail.com
University of Washington
Andy Cotter cotter@ttic.edu
Nati Srebro nati@ttic.edu
TTI Chicago

We study PCA as a stochastic optimization problem and propose a novel stochastic approximation algorithm which we refer to as “Matrix Stochastic Gradient’’ (MSG), as well as a practical variant, Capped MSG. We study the method both theoretically and empirically.

F17 Variance Reduction for Stochastic Gradient Optimization

Chong Wang chongw@cs.cmu.edu
Xi Chen xichen@cs.cmu.edu
Eric Xing epxing@cs.cmu.edu
CMU
Alex Smola alex@smola.org
Yahoo! Research

Stochastic gradient optimization is a class of widely used algorithms for training machine learning models. To optimize an objective, it uses the noisy gradient computed from the random data samples instead of the true gradient computed from the entire dataset. However, when the variance of the noisy gradient is large, the algorithm might spend much time bouncing around, leading to slower convergence and worse performance. In this paper, we develop a general approach of using control variate for variance reduction in stochastic gradient. Data statistics such as low-order moments (pre-computed or estimated online) is used to form the control variate. We demonstrate how to construct the control variate for two practical problems using stochastic gradient optimization. One is convex---the MAP estimation for logistic regression, and the other is non-convex---stochastic variational inference for latent Dirichlet allocation. On both problems, our approach shows faster convergence and better performance than the classical approach.

F18 Memory Limited, Streaming PCA

Ioannis Mitliagkas ioannis@utexas.edu
Constantine Caramanis constantine@utexas.edu
UT Austin
Prateek Jain prajain@microsoft.com
Microsoft Research

We consider streaming, one-pass principal component analysis (PCA), in the high-dimensional regime, with limited memory. Here, p-dimensional samples are presented sequentially, and the goal is to produce the k-dimensional subspace that best approximates these points. Standard algorithms require $O(p^2)$ memory; meanwhile no algorithm can do better than $O(kp)$ memory, since this is what the output itself requires. Memory (or storage) complexity is most meaningful when understood in the context of computational and sample complexity. Sample complexity for high-dimensional PCA is typically studied in the setting of the *spiked covariance model*, where p-dimensional points are generated from a population covariance equal to the identity (white noise) plus a low-dimensional perturbation (the spike) which is the signal to be recovered. It is now well-understood that the spike can be recovered when the number of samples, n, scales proportionally with the dimension, p. Yet, all algorithms that provably achieve this, have memory complexity $O(p^2)$. Meanwhile, algorithms with memory-complexity $O(kp)$ do not have provable bounds on sample complexity comparable to p. We present an algorithm that achieves both: it uses $O(kp)$ memory (meaning storage of any kind) and is able to compute the k-dimensional spike with $O(p \log p)$ sample-complexity -- the first algorithm of its kind. While our theoretical analysis focuses on the spiked covariance model, our simulations show that our algorithm is successful on much more general models for the data.

F19 Near-Optimal Entrywise Sampling for Data Matrices

Dimitris Achlioptas optas@cs.ucsc.edu
UC Santa Cruz
Zohar Karnin zkarnin@yahoo-inc.com
Yahoo! Labs
Edo Liberty edo@yahoo-inc.com
Yahoo! Research

We consider the problem of independently sampling s non-zero entries of a matrix A in order to produce a sparse sketch of it, B, that minimizes $\|A-B\|_2$. For large $m \times n$ matrices, such that $n \gg m$ (for example, representing n observations over m attributes) we give distributions exhibiting four important properties. First, they have closed forms for the probability of sampling each item which are computable from minimal information regarding A. Second, they allow sketching of matrices whose non-zeros are presented to the algorithm in arbitrary order as a stream, with $O(1)$computation per non-zero. Third, the resulting sketch matrices are not only sparse, but their non-zero entries are highly compressible. Lastly, and most importantly, under mild assumptions, our distributions are provably competitive with the optimal offline distribution. Note that the probabilities in the optimal offline distribution may be complex functions of all the entries in the matrix. Therefore, regardless of computational complexity, the optimal distribution might be impossible to compute in the streaming model.

F20 Large Scale Distributed Sparse Precision Estimation

Huahua Wang huwang@cs.umn.edu
Arindam Banerjee banerjee@cs.umn.edu
University of Minnesota, Twin Cites
Cho-Jui Hsieh cjhsieh@cs.utexas.edu
Pradeep Ravikumar pradeepr@cs.utexas.edu
UT Austin
Inderjit Dhillon inderjit@cs.utexas.edu
University of Texas

We consider the problem of sparse precision matrix estimation in high dimensions using the CLIME estimator, which has several desirable theoretical properties. We present an inexact alternating direction method of multiplier (ADMM) algorithm for CLIME, and establish rates of convergence for both the objective and optimality conditions. Further, we develop a large scale distributed framework for the computations, which scales to millions of dimensions and trillions of parameters, using hundreds of cores. The proposed framework solves CLIME in column-blocks and only involves elementwise operations and parallel matrix multiplications. We evaluate our algorithm on both shared-memory and distributed-memory architectures, which can use block cyclic distribution of data and parameters to achieve load balance and improve the efficiency in the use of memory hierarchies. Experimental results show that our algorithm is substantially more scalable than state-of-the-art methods and scales almost linearly with the number of cores.

F21 Optimistic Concurrency Control for Distributed Unsupervised Learning

Xinghao Pan xinghao@eecs.berkeley.edu
Joseph Gonzalez jegonzal@eecs.berkeley.edu
Stefanie Jegelka stefje@eecs.berkeley.edu
Tamara Broderick tab@stat.berkeley.edu
Michael Jordan jordan@cs.berkeley.edu
UC Berkeley

Research on distributed machine learning algorithms has focused primarily on one of two extremes---algorithms that obey strict concurrency constraints or algorithms that obey few or no such constraints. We consider an intermediate alternative in which algorithms optimistically assume that conflicts are unlikely and if conflicts do arise a conflict-resolution protocol is invoked. We view this “optimistic concurrency control'' paradigm as particularly appropriate for large-scale machine learning algorithms, particularly in the unsupervised setting. We demonstrate our approach in three problem areas: clustering, feature learning and online facility location. We evaluate our methods via large-scale experiments in a cluster computing environment.

F22 Distributed Submodular Maximization: Identifying Representative Elements in Massive Data

Baharan Mirzasoleiman baharanm@student.ethz.ch
Amin Karbasi amin.karbasi@gmail.com
Andreas Krause krausea@ethz.ch
ETH Zurich
Rik Sarkar rsarkar@inf.ed.ac.uk
University of Edinburgh

Many large-scale machine learning problems (such as clustering, non-parametric learning, kernel machines, etc.) require selecting, out of a massive data set, a manageable, representative subset. Such problems can often be reduced to maximizing a submodular set function subject to cardinality constraints. Classical approaches require centralized access to the full data set; but for truly large-scale problems, rendering the data centrally is often impractical. In this paper, we consider the problem of submodular function maximization in a distributed fashion. We develop a simple, two-stage protocol GreeDI, that is easily implemented using MapReduce style computations. We theoretically analyze our approach, and show, that under certain natural conditions, performance close to the (impractical) centralized approach can be achieved. In our extensive experiments, we demonstrate the effectiveness of our approach on several applications, including sparse Gaussian process inference on tens of millions of examples using Hadoop.

F23 Provable Subspace Clustering: When LRR meets SSC

Yu-Xiang Wang luke.yxwang@gmail.com
National University of Singapore
Huan Xu mpexuh@nus.edu.sg
NUS
Chenlei Leng c.leng@warwick.ac.uk
University of Warwick

Sparse Subspace Clustering (SSC) and Low-Rank Representation (LRR) are both considered as the state-of-the-art methods for *subspace clustering.* The two methods are fundamentally similar in that both are convex optimizations exploiting the intuition of “Self-Expressiveness”. The main difference is that SSC minimizes the vector ℓ_1 norm of the representation matrix to induce sparsity while LRR minimizes nuclear norm (aka trace norm) to promote a low-rank structure. Because the representation matrix is often simultaneously sparse and low-rank, we propose a new algorithm, termed Low-Rank Sparse Subspace Clustering (LRSSC), by combining SSC and LRR, and develops theoretical guarantees of when the algorithm succeeds. The results reveal interesting insights into the strength and weakness of SSC and LRR and demonstrate how LRSSC can take the advantages of both methods in preserving the “Self-Expressiveness Property” and “Graph Connectivity” at the same time.

F24 Simultaneous Rectification and Alignment via Robust Recovery of Low-rank Tensors

Xiaoqin Zhang zhangxiaoqinnan@gmail.com
Di Wang wangdi@wzu.edu.cn
Wenzhou University
Zhengyuan Zhou zhengyuanzhou24@gmail.com
Stanford University
Yi Ma mayi@microsoft.com
Microsoft Research

In this work, we propose a general method for recovering low-rank three-order tensors, in which the data can be deformed by some unknown transformation and corrupted by arbitrary sparse errors. Since the unfolding matrices of a tensor are interdependent, we introduce auxiliary variables and relax the hard equality constraints by the augmented Lagrange multiplier method. To improve the computational efficiency, we introduce a proximal gradient step to the alternating direction minimization method. We have provided proof for the convergence of the linearized version of the problem which is the inner loop of the overall algorithm. Both simulations and experiments show that our methods are more efficient and effective than previous work. The proposed method can be easily applied to simultaneously rectify and align multiple images or videos frames. In this context, the state-of-the-art algorithms “RASL” and “TILT” can be viewed as two special cases of our work, and yet each only performs part of the function of our method.

F25 Phase Retrieval using Alternating Minimization

Praneeth Netrapalli praneethn@gmail.com
Sujay Sanghavi sanghavi@mail.utexas.edu
UT Austin
Prateek Jain prajain@microsoft.com
Microsoft Research

Phase retrieval problems involve solving linear equations, but with missing sign (or phase, for complex numbers) information. Over the last two decades, a popular generic empirical approach to the many variants of this problem has been one of alternating minimization; i.e. alternating between estimating the missing phase information, and the candidate solution. In this paper, we show that a simple alternating minimization algorithm geometrically converges to the solution of one such problem -- finding a vector x from y,A, where $y=|A'x|$ and $|z|$ denotes a vector of element-wise magnitudes of z -- under the assumption that A is Gaussian. Empirically, our algorithm performs similar to recently proposed convex techniques for this variant (which are based on "lifting" to a convex matrix problem) in sample complexity and robustness to noise. However, our algorithm is much more efficient and can scale to large problems. Analytically, we show geometric convergence to the solution, and sample complexity that is off by log factors from obvious lower bounds. We also establish close to optimal scaling for the case when the unknown vector is sparse. Our work represents the only known proof of alternating minimization for any variant of phase retrieval problems in the non-convex setting.

F26 Non-Uniform Camera Shake Removal Using a Spatially-Adaptive Sparse Penalty

Haichao Zhang hczhang1@gmail.com
Duke University
David Wipf davidwip@microsoft.com
Microsoft Research

Typical blur from camera shake often deviates from the standard uniform convolutional assumption, in part because of problematic rotations which create greater blurring away from some unknown center point. Consequently, successful blind deconvolution for removing shake artifacts requires the estimation of a spatially-varying or non-uniform blur operator. Using ideas from Bayesian inference and convex analysis, this paper derives a non-uniform blind deblurring algorithm with several desirable, yet previously-unexplored attributes. The underlying objective function includes a spatially-adaptive penalty that couples the latent sharp image, non-uniform blur operator, and noise level together. This coupling allows the penalty to automatically adjust its shape based on the estimated degree of local blur and image structure such that regions with large blur or few prominent edges are discounted. Remaining regions with modest blur and revealing edges therefore dominate the overall estimation process without explicitly incorporating structure-selection heuristics. The algorithm can be implemented using an optimization strategy that is virtually parameter free and simpler than existing methods. Detailed theoretical analysis and empirical validation on real images serve to validate the proposed method.

F27 Matrix Completion From any Given Set of Observations

Troy Lee troyjlee@gmail.com
Centre for Quantum Technologies
Adi Shraibman adi.shribman@gmail.com
Weizmann Institute of Science

In the matrix completion problem the aim is to recover an unknown real matrix from a subset of its entries. This problem comes up in many application areas, and has received a great deal of attention in the context of the netflix prize. A central approach to this problem is to output a matrix of lowest possible complexity (e.g. rank or trace norm) that agrees with the partially specified matrix. The performance of this approach under the assumption that the revealed entries are sampled randomly has received considerable attention. In practice, often the set of revealed entries is not chosen at random and these results do not apply. We are therefore left with no guarantees on the performance of the algorithm we are using. We present a means to obtain performance guarantees with respect to any set of initial observations. The first step remains the same: find a matrix of lowest possible complexity that agrees with the partially specified matrix. We give a new way to interpret the output of this algorithm by next finding a probability distribution over the non-revealed entries with respect to which a bound on the generalization error can be proven. The more complex the set of revealed entries according to a certain measure, the better the bound on the generalization error.

F28 Machine Teaching for Bayesian Learners in the Exponential Family

Xiaojin Zhu jerryzhu@cs.wisc.edu
UW-Madison

What if there is a teacher who knows the learning goal and wants to design good training data for a machine learner? We propose an optimal teaching framework aimed at learners who employ Bayesian models. Our framework is expressed as an optimization problem over teaching examples that balance the future loss of the learner and the effort of the teacher. This optimization problem is in general hard. In the case where the learner employs conjugate exponential family models, we present an approximate algorithm for finding the optimal teaching set. Our algorithm optimizes the aggregate sufficient statistics, then unpacks them into actual teaching examples. We give several examples to illustrate our framework.

F29 Approximate Bayesian Image Interpretation using Generative Probabilistic Graphics Programs

Vikash Mansinghka vkm@mit.edu
Tejas Kulkarni tejask@mit.edu
Yura Perov perov@mit.edu
Josh Tenenbaum jbt@mit.edu
Massachusetts Institute of Technology

The idea of computer vision as the Bayesian inverse problem to computer graphics has a long history and an appealing elegance, but it has proved difficult to directly implement. Instead, most vision tasks are approached via complex bottom-up processing pipelines. Here we show that it is possible to write short, simple probabilistic graphics programs that define flexible generative models and to automatically invert them to interpret real-world images. Generative probabilistic graphics programs consist of a stochastic scene generator, a renderer based on graphics software, a stochastic likelihood model linking the renderer's output and the data, and latent variables that adjust the fidelity of the renderer and the tolerance of the likelihood model. Representations and algorithms from computer graphics, originally designed to produce high-quality images, are instead used as the deterministic backbone for highly approximate and stochastic generative models. This formulation combines probabilistic programming, computer graphics, and approximate Bayesian computation, and depends only on general-purpose, automatic inference techniques. We describe two applications: reading sequences of degraded and adversarially obscured alphanumeric characters, and inferring 3D road models from vehicle-mounted camera images. Each of the probabilistic graphics programs we present relies on under 20 lines of probabilistic code, and supports accurate, approximately Bayesian inferences about ambiguous real-world images.

F30 Analyzing Hogwild Parallel Gaussian Gibbs Sampling

Matthew Johnson mattjj@csail.mit.edu
James Saunderson jamess@mit.edu
Alan Willsky willsky@mit.edu
Massachusetts Institute of Technology

Sampling inference methods are computationally difficult to scale for many models in part because global dependencies can reduce opportunities for parallel computation. Without strict conditional independence structure among variables, standard Gibbs sampling theory requires sample updates to be performed sequentially, even if dependence between most variables is not strong. Empirical work has shown that some models can be sampled effectively by going "Hogwild" and simply running Gibbs updates in parallel with only periodic global communication, but the successes and limitations of such a strategy are not well understood. As a step towards such an understanding, we study the Hogwild Gibbs sampling strategy in the context of Gaussian distributions. We develop a framework which provides convergence conditions and error bounds along with simple proofs and connections to methods in numerical linear algebra. In particular, we show that if the Gaussian precision matrix is generalized diagonally dominant, then any Hogwild Gibbs sampler, with any update schedule or allocation of variables to processors, yields a stable sampling process with the correct sample mean.

F31 Flexible sampling of discrete data correlations without the marginal distributions

Freddie Kalaitzis a.kalaitzis@ucl.ac.uk
Ricardo Silva ricardo@stats.ucl.ac.uk
UCL

Learning the joint dependence of discrete variables is a fundamental problem in machine learning, with many applications including prediction, clustering and dimensionality reduction. More recently, the framework of copula modeling has gained popularity due to its modular parametrization of joint distributions. Among other properties, copulas provide a recipe for combining flexible models for univariate marginal distributions with parametric families suitable for potentially high dimensional dependence structures. More radically, the extended rank likelihood approach of Hoff (2007) bypasses learning marginal models completely when such information is ancillary to the learning task at hand as in, e.g., standard dimensionality reduction problems or copula parameter estimation. The main idea is to represent data by their observable rank statistics, ignoring any other information from the marginals. Inference is typically done in a Bayesian framework with Gaussian copulas, and it is complicated by the fact this implies sampling within a space where the number of constraints increase quadratically with the number of data points. The result is slow mixing when using off-the-shelf Gibbs sampling. We present an efficient algorithm based on recent advances on constrained Hamiltonian Markov chain Monte Carlo that is simple to implement and does not require paying for a quadratic cost in sample size.

F32 Stochastic Gradient Riemannian Langevin Dynamics on the Probability Simplex

Sam Patterson spatterson@gatsby.ucl.ac.uk
Gatsby Unit, UCL
Yee Whye Teh y.w.teh@stats.ox.ac.uk
University of Oxford

In this paper we investigate the use of Langevin Monte Carlo methods on the probability simplex and propose a new method, Stochastic gradient Riemannian Langevin dynamics, which is simple to implement and can be applied online. We apply this method to latent Dirichlet allocation in an online setting, and demonstrate that it achieves substantial performance improvements to the state of the art online variational Bayesian methods.

F33 Auxiliary-variable Exact Hamiltonian Monte Carlo Samplers for Binary Distributions

Ari Pakman aripakman@gmail.com
Liam Paninski liam@stat.columbia.edu
Columbia University

We present a new approach to sample from generic binary distributions, based on an exact Hamiltonian Monte Carlo algorithm applied to a piecewise continuous augmentation of the binary distribution of interest. An extension of this idea to distributions over mixtures of binary and continuous variables allows us to sample from posteriors of linear and probit regression models with spike-and-slab priors and truncated parameters. We illustrate the advantages of these algorithms in several examples in which they outperform the Metropolis or Gibbs samplers.

F34 Reconciling “priors” & “priors” without prejudice?

Remi Gribonval remi.gribonval@inria.fr
Pierre Machart pierre.machart@inria.fr
INRIA

There are two major routes to address linear inverse problems. Whereas regularization-based approaches build estimators as solutions of penalized regression optimization problems, Bayesian estimators rely on the posterior distribution of the unknown, given some assumed family of priors. While these may seem radically different approaches, recent results have shown that, in the context of additive white Gaussian denoising, the Bayesian conditional mean estimator is always the solution of a penalized regression problem. The contribution of this paper is twofold. First, we extend the additive white Gaussian denoising results to general linear inverse problems with colored Gaussian noise. Second, we characterize conditions under which the penalty function associated to the conditional mean estimator can satisfy certain popular properties such as convexity, separability, and smoothness. This sheds light on some tradeoff between computational efficiency and estimation accuracy in sparse regularization, and draws some connections between Bayesian estimation and proximal optimization.

F35 Wavelets on Graphs via Deep Learning

Raif Rustamov raifrustamov@gmail.com
Leonidas Guibas guibas@cs.stanford.edu
Stanford University

An increasing number of applications require processing of signals defined on weighted graphs. While wavelets provide a flexible tool for signal processing in the classical setting of regular domains, the existing graph wavelet constructions are less flexible in they are guided solely by the structure of the underlying graph and do not take directly into consideration the particular class of signals to be processed. This paper introduces a machine learning framework for constructing graph wavelets that can sparsely represent a given class of signals. Our construction uses the lifting scheme, and is based on the observation that the recurrent nature of the lifting scheme gives rise to a structure resembling a deep auto-encoder network. Particular properties that the resulting wavelets must satisfy determine the training objective and the structure of the involved neural networks. The training is unsupervised, and is conducted similarly to the greedy pre-training of a stack of auto-encoders. After training is completed, we obtain a linear wavelet transform that can be applied to any graph signal in time and memory linear in the size of the graph. Improved sparsity of our wavelet transform for the test signals is confirmed via experiments both on synthetic and real data.

F36 Robust Sparse Principal Component Regression under the High Dimensional Elliptical Model

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

In this paper we focus on the principal component regression and its application to high dimension non-Gaussian data. The major contributions are in two folds. First, in low dimensions and under a double asymptotic framework where both the dimension d and sample size n can increase, by borrowing the strength from recent development in minimax optimal principal component estimation, we first time sharply characterize the potential advantage of classical principal component regression over least square estimation under the Gaussian model. Secondly, we propose and analyze a new robust sparse principal component regression on high dimensional elliptically distributed data. The elliptical distribution is a semiparametric generalization of the Gaussian, including many well known distributions such as multivariate Gaussian, rank-deficient Gaussian, t, Cauchy, and logistic. It allows the random vector to be heavy tailed and have tail dependence. These extra flexibilities make it very suitable for modeling finance and biomedical imaging data. Under the elliptical model, we prove that our method can estimate the regression coefficients in the optimal parametric rate and therefore is a good alternative to the Gaussian based methods. Experiments on synthetic and real world data are conducted to illustrate the empirical usefulness of the proposed method.

F37 A simple example of Dirichlet process mixture inconsistency for the number of components

Jeff Miller jeffrey_miller@brown.edu
Matthew Harrison Matthew_Harrison@Brown.edu
Brown University

For data assumed to come from a finite mixture with an unknown number of components, it has become common to use Dirichlet process mixtures (DPMs) not only for density estimation, but also for inferences about the number of components. The typical approach is to use the posterior distribution on the number of components occurring so far --- that is, the posterior on the number of clusters in the observed data. However, it turns out that this posterior is not consistent --- it does not converge to the true number of components. In this note, we give an elementary demonstration of this inconsistency in what is perhaps the simplest possible setting: a DPM with normal components of unit variance, applied to data from a "mixture" with one standard normal component. Further, we find that this example exhibits severe inconsistency: instead of going to 1, the posterior probability that there is one cluster goes to 0.

F38 Restricting exchangeable nonparametric distributions

Sinead Williamson sineadannewilliamson@gmail.com
UT Austin
Steve MacEachern snm@stat.osu.edu
Ohio State University
Eric Xing epxing@cs.cmu.edu
CMU

Distributions over exchangeable matrices with infinitely many columns are useful in constructing nonparametric latent variable models. However, the distribution implied by such models over the number of features exhibited by each data point may be poorly-suited for many modeling tasks. In this paper, we propose a class of exchangeable nonparametric priors obtained by restricting the domain of existing models. Such models allow us to specify the distribution over the number of features per data point, and can achieve better performance on data sets where the number of features is not well-modeled by the original distribution.

F39 Stochastic blockmodel approximation of a graphon: Theory and consistent estimation

Edo Airoldi airoldi@fas.harvard.edu
Thiago Costa tcosta@fas.harvard.edu
Stanley Chan schan@seas.harvard.edu
Harvard University

Given a convergent sequence of graphs, there exists a limit object called the graphon from which random graphs are generated. This nonparametric perspective of random graphs opens the door to study graphs beyond the traditional parametric models, but at the same time also poses the challenging question of how to estimate the graphon underlying observed graphs. In this paper, we propose a computationally efficient algorithm to estimate a graphon from a set of observed graphs generated from it. We show that, by approximating the graphon with stochastic block models, the graphon can be consistently estimated, that is, the estimation error vanishes as the size of the graph approaches infinity.

F40 Bayesian Hierarchical Community Discovery

Charles Blundell charles@deepmind.com
Gatsby Unit, UCL
Yee Whye Teh y.w.teh@stats.ox.ac.uk
University of Oxford

We propose an efficient Bayesian nonparametric model for discovering hierarchical community structure in social networks. Our model is a tree-structured mixture of potentially exponentially many stochastic blockmodels. We describe a family of greedy agglomerative model selection algorithms whose worst case scales quadratically in the number of vertices of the network, but independent of the number of communities. Our algorithms are two orders of magnitude faster than the infinite relational model, achieving comparable or better accuracy.

F41 Scalable Influence Estimation in Continuous-Time Diffusion Networks

Nan Du dunan@gatech.edu
Le Song lsong@cc.gatech.edu
Hongyuan Zha zha@cc.gatech.edu
Georgia Tech
Manuel Gomez-Rodriguez
manuelgr@tuebingen.mpg.de
MPI for Intelligent Systems

If a piece of information is released from a media site, can it spread, in 1 month, to a million web pages? This influence estimation problem is very challenging since both the time-sensitive nature of the problem and the issue of scalability need to be addressed simultaneously. In this paper, we propose a randomized algorithm for influence estimation in continuous-time diffusion networks. Our algorithm can estimate the influence of every node in a network with $|\backslash Vcal|$ nodes and $|\backslash Ecal|$ edges to an accuracy of ϵ using $n{=}O(1/\epsilon^2)$ randomizations and up to logarithmic factors $O(n|\backslash Ecal|{+}n|\backslash Vcal|)$ computations. When used as a subroutine in a greedy influence maximization algorithm, our proposed method is guaranteed to find a set of nodes with an influence of at least $(1{-}1/e)\mathrm{OPT}{-}2\epsilon$, where OPT is the optimal value. Experiments on both synthetic and real-world data show that the proposed method can easily scale up to networks of millions of nodes while significantly improves over previous state-of-the-arts in terms of the accuracy of the estimated influence and the quality of the selected nodes in maximizing the influence.

F42 Nonparametric Multi-group Membership Model for Dynamic Networks

Myunghwan Kim mykim@stanford.edu
Jure Leskovec jure@cs.stanford.edu
Stanford University

Relational data—like graphs, networks, and matrices—is often dynamic, where the relational structure evolves over time. A fundamental problem in the analysis of time-varying network data is to extract a summary of the common structure and the dynamics of underlying relations between entities. Here we build on the intuition that changes in the network structure are driven by the dynamics at the level of groups of nodes. We propose a nonparametric multi-group membership model for dynamic networks. Our model contains three main components. We model the birth and death of groups with respect to the dynamics of the network structure via a distance dependent Indian Buffet Process. We capture the evolution of individual node group memberships via a Factorial Hidden Markov model. And, we explain the dynamics of the network structure by explicitly modeling the connectivity structure. We demonstrate our model's capability of identifying the dynamics of latent groups in a number of different types of network data. Experimental results show our model achieves higher predictive performance on the future network forecasting and missing link prediction.

F43 Bayesian entropy estimation for binary spike train data using parametric prior knowledge

Evan Archer earcher@utexas.edu
Il Park memming@austin.utexas.edu
Jonathan Pillow pillow@mail.utexas.edu
UT Austin

Shannon's entropy is a basic quantity in information theory, and a fundamental building block for the analysis of neural codes. Estimating the entropy of a discrete distribution from samples is an important and difficult problem that has received considerable attention in statistics and theoretical neuroscience. However, neural responses have characteristic statistical structure that generic entropy estimators fail to exploit. For example, existing Bayesian entropy estimators make the naive assumption that all spike words are equally likely a priori, which makes for an inefficient allocation of prior probability mass in cases where spikes are sparse. Here we develop Bayesian estimators for the entropy of binary spike trains using priors designed to flexibly exploit the statistical structure of simultaneously-recorded spike responses. We define two prior distributions over spike words using mixtures of Dirichlet distributions centered on simple parametric models. The parametric model captures high-level statistical features of the data, such as the average spike count in a spike word, which allows the posterior over entropy to concentrate more rapidly than with standard estimators (e.g., in cases where the probability of spiking differs strongly from 0.5). Conversely, the Dirichlet distributions assign prior mass to distributions far from the parametric model, ensuring consistent estimates for arbitrary distributions. We devise a compact representation of the data and prior that allow for computationally efficient implementations of Bayesian least squares and empirical Bayes entropy estimators with large numbers of neurons. We apply these estimators to simulated and real neural data and show that they substantially outperform traditional methods.

F44 Universal models for binary spike patterns using centered Dirichlet processes

Il Park memming@austin.utexas.edu
Evan Archer earcher@utexas.edu
Kenneth Latimer latimerk@utexas.edu
Jonathan Pillow pillow@mail.utexas.edu
UT Austin

Probabilistic models for binary spike patterns provide a powerful tool for understanding the statistical dependencies in large-scale neural recordings. Maximum entropy (or "maxent'') models, which seek to explain dependencies in terms of low-order interactions between neurons, have enjoyed remarkable success in modeling such patterns, particularly for small groups of neurons. However, these models are computationally intractable for large populations, and low-order maxent models have been shown to be inadequate for some datasets. To overcome these limitations, we propose a family of "universal'' models for binary spike patterns, where universality refers to the ability to model arbitrary distributions over all $2m$ binary patterns. We construct universal models using a Dirichlet process centered on a well-behaved parametric base measure, which naturally combines the flexibility of a histogram and the parsimony of a parametric model. We derive computationally efficient inference methods using Bernoulli and cascade-logistic base measures, which scale tractably to large populations. We also establish a condition for equivalence between the cascade-logistic and the 2nd-order maxent or "Ising" model, making cascade-logistic a reasonable choice for base measure in a universal model. We illustrate the performance of these models using neural data.

F45 A Determinantal Point Process Latent Variable Model for Inhibition in Neural Spiking Data

Jasper Snoek jasper@cs.toronto.edu
Richard Zemel zemel@cs.toronto.edu
University of Toronto
Ryan Adams rpa@seas.harvard.edu
Harvard University

Point processes are popular models of neural spiking behavior as they provide a statistical distribution over temporal sequences of spikes and help to reveal the complexities underlying a series of recorded action potentials. However, the most common neural point process models, the Poisson process and the gamma renewal process, do not capture interactions and correlations that are critical to modeling populations of neurons. We develop a novel model based on a determinantal point process over latent embeddings of neurons that effectively captures and helps visualize complex inhibitory and competitive interaction. We show that this model is a natural extension of the popular generalized linear model to sets of interacting neurons. The model is extended to incorporate gain control or divisive normalization, and the modulation of neural spiking based on periodic phenomena. Applied to neural spike recordings from the rat hippocampus, we see that the model captures inhibitory relationships, a dichotomy of classes of neurons, and a periodic modulation by the theta rhythm known to be present in the data.

F46 Inferring neural population dynamics from multiple partial recordings of the same neural circuit

Srini Turaga sturaga@gatsby.ucl.ac.uk
Lars Buesing lars@gatsby.ucl.ac.uk
Gatsby Unit, UCL
Adam Packer a.packer@ucl.ac.uk
Henry Dalgleish henry.dalgleish.09@ucl.ac.uk
Noah Pettit noah.pettit.10@ucl.ac.uk
Michael Hausser m.hausser@ucl.ac.uk
UCL
Jakob Macke Jakob.Macke@gmail.com
MPI for Biological Cybernetics

Simultaneous recordings of the activity of large neural populations are extremely valuable as they can be used to infer the dynamics and interactions of neurons in a local circuit, shedding light on the computations performed. It is now possible to measure the activity of hundreds of neurons using 2-photon calcium imaging. However, many computations are thought to involve circuits consisting of thousands of neurons, such as cortical barrels in rodent somatosensory cortex. Here we contribute a statistical method for “stitching” together sequentially imaged sets of neurons into one model by phrasing the problem as fitting a latent dynamical system with missing observations. This method allows us to substantially expand the population-sizes for which population dynamics can be characterized---beyond the number of simultaneously imaged neurons. In particular, we demonstrate using recordings in mouse somatosensory cortex that this method makes it possible to predict noise correlations between non-simultaneously recorded neuron pairs.

F47 Neural representation of action sequences: how far can a simple snippet-matching model take us?

Cheston Tan cheston-tan@i2r.a-star.edu.sg
Institute for Infocomm Research, Singapore
Jedediah Singer jedediah.singer@childrens.harvard.edu
Boston Children’s Hospital
Thomas Serre thomas_serre@brown.edu
David Sheinberg David_Sheinberg@brown.edu
Brown University
Tomaso Poggio tp@ai.mit.edu
Massachusetts Institute of Technology

The macaque Superior Temporal Sulcus (STS) is a brain area that receives and integrates inputs from both the ventral and dorsal visual processing streams (thought to specialize in form and motion processing respectively). For the processing of articulated actions, prior work has shown that even a small population of STS neurons contains sufficient information for the decoding of actor invariant to action, action invariant to actor, as well as the specific conjunction of actor and action. This paper addresses two questions. First, what are the invariance properties of individual neural representations (rather than the population representation) in STS? Second, what are the neural encoding mechanisms that can produce such individual neural representations from streams of pixel images? We find that a baseline model, one that simply computes a linear weighted sum of ventral and dorsal responses to short action “snippets”, produces surprisingly good fits to the neural data. Interestingly, even using inputs from a single stream, both actor-invariance and action-invariance can be produced simply by having different linear weights.

F48 Firing rate predictions in optimal balanced networks

David Barrett barrett@gatsby.ucl.ac.uk
Sophie Denève sophie.deneve@ens.fr
École Normale Supérieure
Christian Machens
christian.machens@neuro.fchampalimaud.org
Champalimaud Centre for the Unknown

How are firing rates in a spiking network related to neural input, connectivity and network function? This is an important problem because firing rates are one of the most important measures of network activity, in both the study of neural computation and neural network dynamics. However, it is a difficult problem, because the spiking mechanism of individual neurons is highly non-linear, and these individual neurons interact strongly through connectivity. We develop a new technique for calculating firing rates in optimal balanced networks. These are particularly interesting networks because they provide an optimal spike-based signal representation while producing cortex-like spiking activity through a dynamic balance of excitation and inhibition. We can calculate firing rates by treating balanced network dynamics as an algorithm for optimizing signal representation. We identify this algorithm and then calculate firing rates by finding the solution to the algorithm. Our firing rate calculation relates network firing rates directly to network input, connectivity and function. This allows us to explain the function and underlying mechanism of tuning curves in a variety of systems.

F49 Noise-Enhanced Associative Memories

Amin Karbasi amin.karbasi@gmail.com
ETH Zurich
Amir Hesam Salavati hesam.salavati@epfl.ch
Amin Shokrollahi amin.shokrollahi@epfl.ch
EPFL
Lav Varshney varshney@alum.mit.edu
IBM Watson Research Center

Recent advances in associative memory design through structured pattern sets and graph-based inference algorithms have allowed reliable learning and recall of an exponential number of patterns. Although these designs correct external errors in recall, they assume neurons that compute noiselessly, in contrast to the highly variable neurons in hippocampus and olfactory cortex. Here we consider associative memories with noisy internal computations and analytically characterize performance. As long as the internal noise level is below a specified threshold, the error probability in the recall phase can be made exceedingly small. More surprisingly, we show that internal noise actually improves the performance of the recall phase. Computational experiments lend additional support to our theoretical analysis. This work suggests a functional benefit to noisy neurons in biological neuronal networks.

F50 A memory frontier for complex synapses

Subhaneil Lahiri sulahiri@stanford.edu
Surya Ganguli sganguli@stanford.edu
Stanford University

An incredible gulf separates theoretical models of synapses, often described solely by a single scalar value denoting the size of a postsynaptic potential, from the immense complexity of molecular signaling pathways underlying real synapses. To understand the functional contribution of such molecular complexity to learning and memory, it is essential to expand our theoretical conception of a synapse from a single scalar to an entire dynamical system with many internal molecular functional states. Moreover, theoretical considerations alone demand such an expansion; network models with scalar synapses assuming finite numbers of distinguishable synaptic strengths have strikingly limited memory capacity. This raises the fundamental question, how does synaptic complexity give rise to memory? To address this, we develop new mathematical theorems elucidating the relationship between the structural organization and memory properties of complex synapses that are themselves molecular networks. Moreover, in proving such theorems, we uncover a framework, based on first passage time theory, to impose an order on the internal states of complex synaptic models, thereby simplifying the relationship between synaptic structure and function.

F51 Perfect Associative Learning with Spike-Timing-Dependent Plasticity

Christian Albers calbers@neuro.uni-bremen.de
Maren Westkott maren@neuro.uni-bremen.de
University of Bremen
Klaus Pawelzik pawelzik@neuro.uni-bremen.de
Universität Bremen

Recent extensions of the Perceptron, as e.g. the Tempotron, suggest that this theoretical concept is highly relevant also for understanding networks of spiking neurons in the brain. It is not known, however, how the computational power of the Perceptron and of its variants might be accomplished by the plasticity mechanisms of real synapses. Here we prove that spike-timing-dependent plasticity having an anti-Hebbian form for excitatory synapses as well as a spike-timing-dependent plasticity of Hebbian shape for inhibitory synapses are sufficient for realizing the original Perceptron Learning Rule if the respective plasticity mechanisms act in concert with the hyperpolarisation of the post-synaptic neurons. We also show that with these simple yet biologically realistic dynamics Tempotrons are efficiently learned. The proposed mechanism might underly the acquisition of mappings of spatio-temporal activity patterns in one area of the brain onto other spatio-temporal spike patterns in another region and of long term memories in cortex. Our results underline that learning processes in realistic networks of spiking neurons depend crucially on the interactions of synaptic plasticity mechanisms with the dynamics of participating neurons.

F52 Reciprocally Coupled Local Estimators Implement Bayesian Information Integration Distributively

Wenhao Zhang whzhang@ion.ac.cn
Si Wu wusi@bnu.edu.cn
Beijing Normal University

Psychophysical experiments have demonstrated that the brain integrates information from multiple sensory cues in a near Bayesian optimal manner. The present study proposes a novel mechanism to achieve this. We consider two reciprocally connected networks, mimicking the integration of heading direction information between the dorsal medial superior temporal (MSTd) and the ventral intraparietal (VIP) areas. Each network serves as a local estimator and receives an independent cue, either the visual or the vestibular, as direct input for the external stimulus. We find that positive reciprocal interactions can improve the decoding accuracy of each individual network as if it implements Bayesian inference from two cues. Our model successfully explains the experimental finding that both MSTd and VIP achieve Bayesian multisensory integration, though each of them only receives a single cue as direct external input. Our result suggests that the brain may implement optimal information integration distributively at each local estimator through the reciprocal connections between cortical regions.

F53 Recurrent linear models of simultaneously-recorded neural populations

Marius Pachitariu marius@gatsby.ucl.ac.uk
Maneesh Sahani maneesh@gatsby.ucl.ac.uk
Gatsby Unit, UCL
Biljana Petreska biljana.petreska@gmail.com
UCL

Population neural recordings with long-range temporal structure are often best understood in terms of a shared underlying low-dimensional dynamical process. Advances in recording technology provide access to an ever larger fraction of the population, but the standard computational approaches available to identify the collective dynamics scale poorly with the size of the dataset. Here we describe a new, scalable approach to discovering the low-dimensional dynamics that underlie simultaneously recorded spike trains from a neural population. Our method is based on recurrent linear models (RLMs), and relates closely to timeseries models based on recurrent neural networks. We formulate RLMs for neural data by generalising the Kalman-filter-based likelihood calculation for latent linear dynamical systems (LDS) models to incorporate a generalised-linear observation process. We show that RLMs describe motor-cortical population data better than either directly-coupled generalised-linear models or latent linear dynamical system models with generalised-linear observations. We also introduce the cascaded linear model (CLM) to capture low-dimensional instantaneous correlations in neural populations. The CLM describes the cortical recordings better than either Ising or Gaussian models and, like the RLM, can be fit exactly and quickly. The CLM can also be seen as a generalization of a low-rank Gaussian model, in this case factor analysis. The computational tractability of the RLM and CLM allow both to scale to very high-dimensional neural data.

F54 Demixing odors - fast inference in olfaction

Agnieszka Grabska-Barwinska agnieszka@gatsby.ucl.ac.uk
Jeff Beck jeffbeck@gatsby.ucl.ac.uk
Peter Latham pel@gatsby.ucl.ac.uk
Gatsby Unit, UCL
Alexandre Pouget alexandre.pouget@unige.ch
University of Geneva

The olfactory system faces a difficult inference problem: it has to determine what odors are present based on the distributed activation of its receptor neurons. Here we derive neural implementations of two approximate inference algorithms that could be used by the brain. One is a variational algorithm (which builds on the work of Beck. et al., 2012), the other is based on sampling. Importantly, we use a more realistic prior distribution over odors than has been used in the past: we use a “spike and slab’’ prior, for which most odors have zero concentration. After mapping the two algorithms onto neural dynamics, we find that both can infer correct odors in less than 100 ms, although it takes ~500 ms to eliminate false positives. Thus, at the behavioral level, the two algorithms make very similar predictions. However, they make different assumptions about connectivity and neural computations, and make different predictions about neural activity. Thus, they should be distinguishable experimentally. If so, that would provide insight into the mechanisms employed by the olfactory system, and, because the two algorithms use very different coding strategies, that would also provide insight into how networks represent probabilities.

F55 Multisensory Encoding, Decoding, and Identification

Aurel Lazar aurel@ee.columbia.edu
Yevgeniy Slutskiy ys2146@columbia.edu
Columbia University

We investigate a spiking neuron model of multisensory integration. Multiple stimuli from different sensory modalities are encoded by a single neural circuit comprised of a multisensory bank of receptive fields in cascade with a population of biophysical spike generators. We demonstrate that stimuli of different dimensions can be faithfully multiplexed and encoded in the spike domain and derive tractable algorithms for decoding each stimulus from the common pool of spikes. We also show that the identification of multisensory processing in a single neuron is dual to the recovery of stimuli encoded with a population of multisensory neurons, and prove that only a projection of the circuit onto input stimuli can be identified. We provide an example of multisensory integration using natural audio and video and discuss the performance of the proposed decoding and identification algorithms.

F56 Recurrent networks of coupled Winner-Take-All oscillators for solving constraint satisfaction problems

Hesham Mostafa hesham@ini.phys.ethz.ch
Lorenz Muller lorenz@ini.phys.ethz.ch
Giacomo Indiveri giacomo@ini.phys.ethz.ch
ETH Zurich

We present a recurrent neuronal network, modeled as a continuous-time dynamical system, that can solve constraint satisfaction problems. Discrete variables are represented by coupled Winner-Take-All (WTA) networks, and their values are encoded in localized patterns of oscillations that are learned by the recurrent weights in these networks. Constraints over the variables are encoded in the network connectivity. Although there are no sources of noise, the network can escape from local optima in its search for solutions that satisfy all constraints by modifying the effective network connectivity through oscillations. If there is no solution that satisfies all constraints, the network state changes in a pseudo-random manner and its trajectory approximates a sampling procedure that selects a variable assignment with a probability that increases with the fraction of constraints satisfied by this assignment. External evidence, or input to the network, can force variables to specific values. When new inputs are applied, the network re-evaluates the entire set of variables in its search for the states that satisfy the maximum number of constraints, while being consistent

with the external input. Our results demonstrate that the proposed network architecture can perform a deterministic search for the optimal solution to problems with non-convex cost functions. The network is inspired by canonical microcircuit models of the cortex and suggests possible dynamical mechanisms to solve constraint satisfaction problems that can be present in biological networks, or implemented in neuromorphic electronic circuits.

F57 Capacity of strong attractor patterns to model behavioural and cognitive prototypes

Abbas Edalat ae@ic.ac.uk
Imperial College London

We solve the mean field equations for a stochastic Hopfield network with temperature (noise) in the presence of strong, i.e., multiply stored patterns, and use this solution to obtain the storage capacity of such a network. Our result provides for the first time a rigorous solution of the mean field equations for the standard Hopfield model and is in contrast to the mathematically unjustifiable replica technique that has been hitherto used for this derivation. We show that the critical temperature for stability of a strong pattern is equal to its degree or multiplicity, when sum of the cubes of degrees of all stored patterns is negligible compared to the network size. In the case of a single strong pattern in the presence of simple patterns, when the ratio of the number of all stored patterns and the network size is a positive constant, we obtain the distribution of the overlaps of the patterns with the mean field and deduce that the storage capacity for retrieving a strong pattern exceeds that for retrieving a simple pattern by a multiplicative factor equal to the square of the degree of the strong pattern. This square law property provides justification for using strong patterns to model attachment types and behavioural prototypes in psychology and psychotherapy.

F58 Compete to Compute

Rupesh Srivastava rupesh@idsia.ch
Jonathan Masci jonathan@idsia.ch
Sohrob Kazerounian sohrob@idsia.ch
Faustino Gomez tino@idsia.ch
Jürgen Schmidhuber juergen@idsia.ch
IDSIA

Local competition among neighboring neurons is common in biological neural networks (NNs). We apply the concept to gradient-based, backprop-trained artificial multilayer NNs. NNs with competing linear units tend to outperform those with non-competing nonlinear units, and avoid catastrophic forgetting when training sets change over time.

F59 Understanding Dropout

Pierre Baldi pfbaldi@ics.uci.edu
Peter Sadowski peterjsadowski@gmail.com
UC Irvine

Dropout is a relatively new algorithm for training neural networks which relies on stochastically “dropping out’’ neurons during training in order to avoid the co-adaptation of feature detectors. We introduce a general formalism for studying dropout on either units or connections, with arbitrary probability values, and use it to analyze the averaging and regularizing properties of dropout in both linear and non-linear networks. For deep neural networks, the averaging properties of dropout are characterized by three recursive equations, including the approximation of expectations by normalized weighted geometric means. We provide estimates and bounds for these approximations and corroborate the results with simulations. We also show in simple cases how dropout performs stochastic gradient descent on a regularized error function.

F60 RNADE: The real-valued neural autoregressive density-estimator

Benigno Uria b.uria@ed.ac.uk
Iain Murray i.murray@ed.ac.uk
University of Edinburgh
Hugo Larochelle hugo.larochelle@usherbrooke.ca
Université de Sherbrooke (Quebec)

We introduce RNADE, a new model for joint density estimation of real-valued vectors. Our model calculates the density of a datapoint as the product of one-dimensional conditionals modeled using mixture density networks with shared parameters. RNADE learns a distributed representation of the data, while having a tractable expression for the calculation of densities. A tractable likelihood allows direct comparison with other methods and training by standard gradient-based optimizers. We compare the performance of RNADE on several datasets of heterogeneous and perceptual data, finding it outperforms mixture models in all but one case.

F61 Correlations strike back (again): the case of associative memory retrieval

Cristina Savin cs664@cam.ac.uk
Mate Lengyel m.lengyel@eng.cam.ac.uk
University of Cambridge
Peter Dayan dayan@gatsby.ucl.ac.uk
Gatsby Unit, UCL

It has long been recognised that statistical dependencies in neuronal activity need to be taken into account when decoding stimuli encoded in a neural population. Less studied, though equally pernicious, is the need to take account of dependencies between synaptic weights when decoding patterns previously encoded in an auto-associative memory. We show that activity-dependent learning generically produces such correlations, and failing to take them into account in the dynamics of memory retrieval leads to catastrophically poor recall. We derive optimal network dynamics for recall in the face of synaptic correlations caused by a range of synaptic plasticity rules. These dynamics involve well-studied circuit motifs, such as forms of feedback inhibition and experimentally observed dendritic nonlinearities. We therefore show how addressing the problem of synaptic correlations leads to a novel functional account of key biophysical features of the neural substrate.

F62 Real-Time Inference for a Gamma Process Model of Neural Spiking

David Carlson david.carlson@duke.edu
jovo Vogelstein jv.work@jhu.edu
Lawrence Carin lcarin@duke.edu
Duke University
Vinayak Rao vrao@gatsby.ucl.ac.uk
Gatsby Unit, UCL

With simultaneous measurements from ever increasing populations of neurons, there is a growing need for sophisticated tools to recover signals from individual neurons. In electrophysiology experiments, this classically proceeds in a two-step process: (i) threshold the waveforms to detect putative spikes and (ii) cluster the waveforms into single units (neurons). We extend previous Bayesian nonparamet- ric models of neural spiking to jointly detect and cluster neurons using a Gamma process model. Importantly, we develop an online approximate inference scheme enabling real-time analysis, with performance exceeding the previous state-of-the- art. Via exploratory data analysis—using data with partial ground truth as well as two novel data sets—we find several features of our model collectively contribute to our improved performance including: (i) accounting for colored noise, (ii) de- tecting overlapping spikes, (iii) tracking waveform dynamics, and (iv) using mul- tiple channels. We hope to enable novel experiments simultaneously measuring many thousands of neurons and possibly adapting stimuli dynamically to probe ever deeper into the mysteries of the brain.

F63 Transportability from Multiple Environments with Limited Experiments

Elias Bareinboim eb@cs.ucla.edu
Judea Pearl judea@cs.ucla.edu
UCLA
Sanghack Lee shlee@iastate.edu
Vasant Honavar vhonavar@ist.psu.edu
Penn State University

This paper considers the problem of transferring experimental findings learned from multiple heterogeneous domains to a target environment, in which only limited experiments can be performed. We reduce questions of transportability from multiple domains and with limited scope to symbolic derivations in the do-calculus, thus extending the treatment of transportability from full experiments introduced in Pearl and Bareinboim (2011). We further provide different graphical and algorithmic conditions for computing the transport formula for this setting, that is, a way of fusing the observational and experimental information scattered throughout different domains to synthesize a consistent estimate of the desired effects.

F64 Causal Inference on Time Series using Restricted Structural Equation Models

Jonas Peters peters@stat.math.ethz.ch
ETH Zurich
Dominik Janzing dominik.janzing@tuebingen.mpg.de
Bernhard Schölkopf bs@tuebingen.mpg.de
MPI Tübingen

Causal inference uses observational data to infer the causal structure of the data generating system. We study a class of restricted Structural Equation Models for time series that we call Time Series Models with Independent Noise (TiMINo). These models require independent residual time series, whereas traditional methods like Granger causality exploit the variance of residuals. This work contains two main contributions: (1) Theoretical: By restricting the model class (e.g. to additive noise) we provide more general identifiability results than existing ones. The results cover lagged and instantaneous effects that can be nonlinear and unfaithful, and non-instantaneous feedbacks between the time series. (2) Practical: If there are no feedback loops between time series, we propose an algorithm based on non-linear independence tests of time series. When the data are causally insufficient, or the data generating process does not satisfy the model assumptions, this algorithm may still give partial results, but mostly avoids incorrect answers. The Structural Equation Model point of view allows us to extend both the theoretical and the algorithmic part to situations in which the time series have been measured with different time delays (as may happen for fMRI data, for example). TiMINo outperforms existing methods on artificial and real data. Code is provided.

F65 Discovering Hidden Variables in Noisy-Or Networks using Quartet Tests

Yacine Jernite yacine.jernite@nyu.edu
Courant Institute, NYU
Yonatan Halpern halpern@cs.nyu.edu
David Sontag dsontag@cs.nyu.edu
NYU

We give a polynomial-time algorithm for provably learning the structure and parameters of bipartite noisy-or Bayesian networks of binary variables where the top layer is completely hidden. Unsupervised learning of these models is a form of discrete factor analysis, enabling the discovery of hidden variables and their causal relationships with observed data. We obtain an efficient learning algorithm for a family of Bayesian networks that we call quartet-learnable, meaning that every latent variable has four children that do not have any other parents in common. We show that the existence of such a quartet allows us to uniquely identify each latent variable and to learn all parameters involving that latent variable. Underlying our algorithm are two new techniques for structure learning: a quartet test to determine whether a set of binary variables are singly coupled, and a conditional mutual information test that we use to learn parameters. We also show how to subtract already learned latent variables from the model to create new singly-coupled quartets, which substantially expands the class of structures that we can learn. Finally, we give a proof of the polynomial sample complexity of our learning algorithm, and experimentally compare it to variational EM.

F66 Learning Hidden Markov Models from Non-sequence Data via Tensor Decomposition

T.-K. Huang tzukuoh@cs.cmu.edu
Jeff Schneider schneide@cs.cmu.edu
CMU

Learning dynamic models from observed data has been a central issue in many scientific studies or engineering tasks. The usual setting is that data are collected sequentially from trajectories of some dynamical system operation. In quite a few modern scientific modeling tasks, however, it turns out that reliable sequential data are rather difficult to gather, whereas out-of-order snapshots are much easier to obtain. Examples include the modeling of galaxies, chronic diseases such Alzheimer's, or certain biological processes. Existing methods for learning dynamic model from non-sequence data are mostly based on Expectation-Maximization, which involves non-convex optimization and is thus hard to analyze. Inspired by recent advances in spectral learning methods, we propose to study this problem from a different perspective: moment matching and spectral decomposition. Under that framework, we identify reasonable assumptions on the generative process of non-sequence data, and propose learning algorithms based on the tensor decomposition method \cite{anandkumar2012tensor} to \textit{provably} recover first-order Markov models and hidden Markov models. To the best of our knowledge, this is the first formal guarantee on learning from non-sequence data. Preliminary simulation results confirm our theoretical findings.

F67 Unsupervised Spectral Learning of Finite State Transducers

Raphael Bailly rbailly@lsi.upc.edu
Xavier Carreras carreras@lsi.upc.edu
Ariadna Quattoni aquattoni@lsi.upc.edu
Universitat Politècnica de Catalunya

Finite-State Transducers (FST) are a standard tool for modeling paired input-output sequences and are used in numerous applications, ranging from computational biology to natural language processing. Recently Balle et al. presented a spectral algorithm for learning FST from samples of aligned input-output sequences. In this paper we address the more realistic, yet challenging setting where the alignments are unknown to the learning algorithm. We frame FST learning as finding a low rank Hankel matrix satisfying constraints derived from observable statistics. Under this formulation, we provide identifiability results for FST distributions. Then, following previous work on rank minimization, we propose a regularized convex relaxation of this objective which is based on minimizing a nuclear norm penalty subject to linear constraints and can be solved efficiently.

F68 Learning Efficient Random Maximum A-Posteriori Predictors with Non-Decomposable Loss Functions

Tamir Hazan tamir@cs.haifa.ac.il
University of Haifa
Subhransu Maji smaji@ttic.edu
TTI Chicago
Joseph Keshet joseph.keshet@biu.ac.il
Bar-Ilan University
Tommi Jaakkola tommi@csail.mit.edu
Massachusetts Institute of Technology

In this work we develop efficient methods for learning random MAP predictors for structured label problems. In particular, we construct posterior distributions over perturbations that can be adjusted via stochastic gradient methods. We show that every smooth posterior distribution would suffice to define a smooth PAC-Bayesian risk bound suitable for gradient methods. In addition, we relate the posterior distributions to computational properties of the MAP predictors. We suggest multiplicative posteriors to learn super-modular potential functions that accompany specialized MAP predictors such as graph-cuts. We also describe label-augmented posterior models that can use efficient MAP approximations, such as those arising from linear program relaxations.

F69 Structured Learning via Logistic Regression

Justin Domke justin.domke@nicta.com.au
NICTA

A successful approach to structured learning is to write the learning objective as a joint function of linear parameters and inference messages, and iterate between updates to each. This paper observes that if the inference problem is "smoothed" through the addition of entropy terms, for fixed messages, the learning objective reduces to a traditional (non-structured) logistic regression problem with respect to parameters. In these logistic regression problems, each training example has a bias term determined by the current set of messages. Based on this insight, the structured energy function can be extended from linear factors to any function class where an "oracle" exists to minimize a logistic loss.

F70 Graphical Models for Inference with Missing Data

Karthika Mohan karthika@cs.ucla.edu
Judea Pearl judea@cs.ucla.edu
UCLA
Jin Tian jtian@iastate.edu
Iowa State University

We address the problem of recoverability i.e. deciding whether there exists a consistent estimator of a given relation Q, when data are missing not at random. We employ a formal representation called `Missingness Graphs' to explicitly portray the causal mechanisms responsible for missingness and to encode dependencies between these mechanisms and the variables being measured. Using this representation, we derive conditions that the graph should satisfy to ensure recoverability and devise algorithms to detect the presence of these conditions in the graph.

F71 Approximate inference in latent Gaussian-Markov models from continuous time observations

Botond Cseke bcseke@inf.ed.ac.uk
Guido Sanguinetti gsanguin@inf.ed.ac.uk
University of Edinburgh
Manfred Opper manfred.opper@tu-berlin.de
TU Berlin

We propose an approximate inference algorithm for continuous time Gaussian-Markov process models with both discrete and continuous time likelihoods. We show that the continuous time limit of the expectation propagation algorithm exists and results in a hybrid fixed point iteration consisting of (1) expectation propagation updates for the discrete time terms and (2) variational updates for the continuous time term. We introduce corrections methods that improve on the marginals of the approximation. This approach extends the classical Kalman-Bucy smoothing procedure to non-Gaussian observations, enabling continuous-time inference in a variety of models, including spiking neuronal models (state-space models with point process observations) and box likelihood models. Experimental results on real and simulated data demonstrate high distributional accuracy and significant computational savings compared to discrete-time approaches in a neural application.

F72 Variational Planning for Graph-based MDPs

Qiang Cheng cheng-q09@mails.tsinghua.edu.cn
Feng Chen chenfeng@mail.tsinghua.edu.cn
Tsinghua University
Qiang Liu qliu1@uci.edu
Alex Ihler ihler@ics.uci.edu
UC Irvine

Markov Decision Processes (MDPs) are extremely useful for modeling and solving sequential decision making problems. Graph-based MDPs provide a compact representation for MDPs with large numbers of random variables. However, the complexity of exactly solving a graph-based MDP usually grows exponentially in the number of variables, which limits their application. We present a new variational framework to describe and solve the planning problem of MDPs, and derive both exact and approximate planning algorithms. In particular, by exploiting the graph structure of graph-based MDPs, we propose a factored variational value iteration algorithm in which the value function is first approximated by the multiplication of local-scope value functions, then solved by minimizing a Kullback-Leibler (KL) divergence. The KL divergence is optimized using the belief propagation algorithm, with complexity exponential in only the cluster size of the graph. Experimental comparison on different models shows that our algorithm outperforms existing approximation algorithms at finding good policies.

F73 Integrated Non-Factorized Variational Inference

Shaobo Han shaobo.han@duke.edu
Xuejun Liao xjliao.work@gmail.com
Lawrence Carin lcarin@duke.edu
Duke University

We present a non-factorized variational method for full posterior inference in Bayesian hierarchical models, with the goal of capturing the posterior variable dependencies via efficient and possibly parallel computation. Our approach unifies the integrated nested Laplace approximation (INLA) under the variational framework. The proposed method is applicable in more challenging scenarios than typically assumed by INLA, such as Bayesian Lasso, which is characterized by the non-differentiability of the $\ell 1$ norm arising from independent Laplace priors. We derive an upper bound for the Kullback-Leibler divergence, which yields a fast closed-form solution via decoupled optimization. Our method is a reliable analytic alternative to Markov chain Monte Carlo (MCMC), and it results in a tighter evidence lower bound than that of mean-field variational Bayes (VB) method.

F74 Global Solver and Its Efficient Approximation for Variational Bayesian Low-rank Subspace Clustering

Shinichi Nakajima shinnkj23@gmail.com
Nikon
Akiko Takeda takeda@mist.i.u-tokyo.ac.jp
University of Tokyo
S. Derin Babacan dbabacan@gmail.com
Google Research
Masashi Sugiyama sugi@cs.titech.ac.jp
Tokyo Institute of Technology
Ichiro Takeuchi takeuchi.ichiro@nitech.ac.jp
Nagoya Institute of Technology

When a probabilistic model and its prior are given, Bayesian learning offers inference with automatic parameter tuning. However, Bayesian learning is often obstructed by computational difficulty: the rigorous Bayesian learning is intractable in many models, and its variational Bayesian (VB) approximation is prone to suffer from local minima. In this paper, we overcome this difficulty for low-rank subspace clustering (LRSC) by providing an exact global solver and its efficient approximation. LRSC extracts a low-dimensional structure of data by embedding samples into the union of low-dimensional subspaces, and its variational Bayesian variant has shown good performance. We first prove a key property that the VB-LRSC model is highly redundant. Thanks to this property, the optimization problem of VB-LRSC can be separated into small subproblems, each of which has only a small number of unknown variables. Our exact global solver relies on another key property that the stationary condition of each subproblem is written as a set of polynomial equations, which is solvable with the homotopy method. For further computational efficiency, we also propose an efficient approximate variant, of which the stationary condition can be written as a polynomial equation with a single variable. Experimental results show the usefulness of our approach.

F75 Bayesian inference as iterated random functions with applications to sequential inference in graphical models

Arash Amini aaamini@umich.edu
Long Nguyen xuanlong@umich.edu
University of Michigan

We propose a general formalism of iterated random functions with semigroup property, under which exact and approximate Bayesian posterior updates can be viewed as specific instances. A convergence theory for iterated random functions is presented. As an application of the general theory we analyze convergence behaviors of exact and approximate message-passing algorithms that arise in a sequential change point detection problem formulated via a latent variable directed graphical model. The sequential inference algorithm and its supporting theory are illustrated by simulated examples.

F76 Learning to Pass Expectation Propagation Messages

Nicolas Heess nicolas.heess@ucl.ac.uk
Gatsby Unit, UCL
Daniel Tarlow dtarlow@cs.toronto.edu
John Winn jwinn@microsoft.com
Microsoft Research

Expectation Propagation (EP) is a popular approximate posterior inference algorithm that often provides a fast and accurate alternative to sampling-based methods. However, while the EP framework in theory allows for complex non-Gaussian factors, there is still a significant practical barrier to using them within EP, because doing so requires the implementation of message update operators, which can be difficult and require hand-crafted approximations. In this work, we study the question of whether it is possible to automatically derive fast and accurate EP updates by learning a discriminative model e.g., a neural network or random forest) to map EP message inputs to EP message outputs. We address the practical concerns that arise in the process, and we provide empirical analysis on several challenging and diverse factors, indicating that there is a space of factors where this approach appears promising.

F77 On the Complexity and Approximation of Binary Evidence in Lifted Inference

Guy Van den Broeck
guy.vandenbroeck@cs.kuleuven.be
Adnan Darwiche darwiche@cs.ucla.edu
UCLA

Lifted inference algorithms exploit symmetries in probabilistic models to speed up inference. They show impressive performance when calculating unconditional probabilities in relational models, but often resort to non-lifted inference when computing conditional probabilities. The reason is that conditioning on evidence breaks many of the model's symmetries, which preempts standard lifting techniques. Recent theoretical results show, for example, that conditioning on evidence which corresponds to binary relations is #P-hard, suggesting that no lifting is to be expected in the worst case. In this paper, we balance this grim result by identifying the Boolean rank of the evidence as a key parameter for characterizing the complexity of conditioning in lifted inference. In particular, we show that conditioning on binary evidence with bounded Boolean rank is efficient. This opens up the possibility of approximating evidence by a low-rank Boolean matrix factorization, which we investigate both theoretically and empirically.

F78 Translating Embeddings for Modeling Multi-relational Data

Antoine Bordes antoine.bordes@utc.fr
Nicolas Usunier nicolas.usunier@utc.fr
Alberto Garcia-Duran agarciad@hds.utc.fr
Université de Technologie de Compiègne (UTC)
Jason Weston jweston@google.com
Oksana Yakhnenko oksana@google.com
Google Research

We consider the problem of embedding entities and relationships of multi-relational data in low-dimensional vector spaces. Our objective is to propose a canonical model which is easy to train, contains a reduced number of parameters and can scale up to very large databases. Hence, we propose, TransE, a method which models relationships by interpreting them as translations operating on the low-dimensional embeddings of the entities. Despite its simplicity, this assumption proves to be powerful since extensive experiments show that TransE significantly outperforms state-of-the-art methods in link prediction on two knowledge bases. Besides, it can be successfully trained on a large scale data set with 1M entities, 25k relationships and more than 17M training samples.

F79 Efficient Online Inference for Bayesian Nonparametric Relational Models

Dae Il Kim daeil.kim@gm.slc.edu
Erik Sudderth sudderth@cs.brown.edu
Brown University
Prem Gopalan pgopalan@cs.princeton.edu
David Blei blei@cs.princeton.edu
Princeton University

Stochastic block models characterize observed network relationships via latent community memberships. In large social networks, we expect entities to participate in multiple communities, and the number of communities to grow with the network size. We introduce a new model for these phenomena, the hierarchical Dirichlet process relational model, which allows nodes to have mixed membership in an unbounded set of communities. To allow scalable learning, we derive an online stochastic variational inference algorithm. Focusing on assortative models of undirected networks, we also propose an efficient structured mean field variational bound, and online methods for automatically pruning unused communities. Compared to state-of-the-art online learning methods for parametric relational models, we show significantly improved perplexity and link prediction accuracy for sparse networks with tens of thousands of nodes. We also showcase an analysis of LittleSis, a large network of who-knows-who at the heights of business and government.

F80 Dropout Training as Adaptive Regularization

Stefan Wager swager@stanford.edu
Sida Wang sidaw@cs.stanford.edu
Percy Liang pliang@cs.stanford.edu
Stanford University

Dropout and other feature noising schemes control overfitting by artificially corrupting the training data. For generalized linear models, dropout performs a form of adaptive regularization. Using this viewpoint, we show that the dropout regularizer is first-order equivalent to an \ LII regularizer applied after scaling the features by an estimate of the inverse diagonal Fisher information matrix. We also establish a connection to AdaGrad, an online learner, and find that a close relative of AdaGrad operates by repeatedly solving linear dropout-regularized problems. By casting dropout as regularization, we develop a natural semi-supervised algorithm that uses unlabeled data to create a better adaptive regularizer. We apply this idea to document classification tasks, and show that it consistently boosts the performance of dropout training, improving on state-of-the-art results on the IMDB reviews dataset.

F81 Convex Two-Layer Modeling

Özlem Aslan ozlem@ualberta.ca
Hao Cheng hcheng2@ualberta.ca
Dale Schuurmans dale@cs.ualberta.ca
University of Alberta
Xinhua Zhang xinhua.zhang.cs@gmail.com
NICTA

Latent variable prediction models, such as multi-layer networks, impose auxiliary latent variables between inputs and outputs to allow automatic inference of implicit features useful for prediction. Unfortunately, such models are difficult to train because inference over latent variables must be performed concurrently with parameter optimization---creating a highly non-convex problem. Instead of proposing another local training method, we develop a convex relaxation of hidden-layer conditional models that admits global training. Our approach extends current convex modeling approaches to handle two nested nonlinearities separated by a non-trivial adaptive latent layer. The resulting methods are able to acquire two-layer models that cannot be represented by any single-layer model over the same features, while improving training quality over local heuristics.

F82 Learning with Noisy Labels

Nagarajan Natarajan naga86@cs.utexas.edu
Pradeep Ravikumar pradeepr@cs.utexas.edu
UT Austin
Inderjit Dhillon inderjit@cs.utexas.edu
University of Texas
Ambuj Tewari tewaria@umich.edu
University of Michigan

In this paper, we theoretically study the problem of binary classification in the presence of random classification noise --- the learner, instead of seeing the true labels, sees labels that have independently been flipped with some small probability. Moreover, random label noise is \emph{class-conditional} --- the flip probability depends on the class. We provide two approaches to suitably modify any given surrogate loss function. First, we provide a simple unbiased estimator of any loss, and obtain performance bounds for empirical risk minimization in the presence of iid data with noisy labels. If the loss function satisfies a simple symmetry condition, we show that the method leads to an efficient algorithm for empirical minimization. Second, by leveraging a reduction of risk minimization under noisy labels to classification with weighted 0-1 loss, we suggest the use of a simple weighted surrogate loss, for which we are able to obtain strong empirical risk bounds. This approach has a very remarkable consequence --- methods used in practice such as biased SVM and weighted logistic regression are provably noise-tolerant. On a synthetic non-separable dataset, our methods achieve over 88\% accuracy even when 40\% of the labels are corrupted, and are competitive with respect to recently proposed methods for dealing with label noise in several benchmark datasets.

F83 Low-rank matrix reconstruction and clustering via approximate message passing

Ryosuke Matsushita matsur8@gmail.com
NTT DATA Mathematical Systems Inc.
Toshiyuki Tanaka tt@i.kyoto-u.ac.jp
Kyoto University

We study the problem of reconstructing low-rank matrices from their noisy observations. We formulate the problem in the Bayesian framework, which allows us to exploit structural properties of matrices in addition to low-rankedness, such as sparsity. We propose an efficient approximate message passing algorithm, derived from the belief propagation algorithm, to perform the Bayesian inference for matrix reconstruction. We have also successfully applied the proposed algorithm to a clustering problem, by formulating the problem of clustering as a low-rank matrix reconstruction problem with an additional structural property. Numerical experiments show that the proposed algorithm outperforms Lloyd's K-means algorithm.

F84 Matrix factorization with binary components

Martin Slawski ms@cs.uni-sb.de
Matthias Hein hein@cs.uni-saarland.de
Pavlo Lutsik p.lutsik@mx.uni-saarland.de
Saarland University

Motivated by an application in computational biology, we consider constrained low-rank matrix factorization problems with $\{0,1\}$-constraints on one of the factors. In addition to the the non-convexity shared with more general matrix factorization schemes, our problem is further complicated by a combinatorial constraint set of size $2m{\cdot}r$, where m is the dimension of the data points and r the rank of the factorization. Despite apparent intractability, we provide $-$in the line of recent work on non-negative matrix factorization by Arora et al.~(2012)$-$ an algorithm that provably recovers the underlying factorization in the exact case with operations of the order $O(mr2r+mnr)$ in the worst case. To obtain that result, we invoke theory centered around a fundamental result in combinatorics, the Littlewood-Offord lemma.

F85 Learning Multi-level Sparse Representations

Ferran Diego Andilla
ferran.diego@iwr.uni-heidelberg.de
Fred Hamprecht
fred.hamprecht@iwr.uni-heidelberg.de
University of Heidelberg

Bilinear approximation of a matrix is a powerful paradigm of unsupervised learning. In some applications, however, there is a natural hierarchy of concepts that ought to be reflected in the unsupervised analysis. For example, in the neurosciences image sequence considered here, there are the semantic concepts of pixel → neuron → assembly that should find their counterpart in the unsupervised analysis. Driven by this concrete problem, we propose a decomposition of the matrix of observations into a product of more than two sparse matrices, with the rank decreasing from lower to higher levels. In contrast to prior work, we allow for both hierarchical and heterarchical relations of lower-level to higher-level concepts. In addition, we learn the nature of these relations rather than imposing them. Finally, we describe an optimization scheme that allows to optimize the decomposition over all levels jointly, rather than in a greedy level-by-level fashion. The proposed bilevel SHMF (sparse heterarchical matrix factorization) is the first formalism that allows to simultaneously interpret a calcium imaging sequence in terms of the constituent neurons, their membership in assemblies, and the time courses of both neurons and assemblies. Experiments show that the proposed model fully recovers the structure from difficult synthetic data designed to imitate the experimental data. More importantly, bilevel SHMF yields plausible interpretations of real-world Calcium imaging data.

F86 Exact and Stable Recovery of Pairwise Interaction Tensors

Shouyuan Chen chenshouyuan@gmail.com
Michael Lyu lyu@cse.cuhk.edu.hk
CUHK
Irwin King king@cse.cuhk.edu.hk
Chinese University of Hong Kong
Zenglin Xu xu218@purdue.edu
University of Purdue

Tensor completion from incomplete observations is a problem of significant practical interest. However, it is unlikely that there exists an efficient algorithm with provable guarantee to recover a general tensor from a limited number of observations. In this paper, we study the recovery algorithm for pairwise interaction tensors, which has recently gained considerable attention for modeling multiple attribute data due to its simplicity and effectiveness. Specifically, in the absence of noise, we show that one can exactly recover a pairwise interaction tensor by solving a constrained convex program which minimizes the weighted sum of nuclear norms of matrices from $O(nr\log 2(n))$ observations. For the noisy cases, we also prove error bounds for a constrained convex program for recovering the tensors. Our experiments on the synthetic dataset demonstrate that the recovery performance of our algorithm agrees well with the theory. In addition, we apply our algorithm on a temporal collaborative filtering task and obtain state-of-the-art results.

F87 A New Convex Relaxation for Tensor Completion

Bernardino Romera-Paredes
bernardino.paredes.09@ucl.ac.uk
Massimiliano Pontil m.pontil@cs.ucl.ac.uk
UCL

We study the problem of learning a tensor from a set of linear measurements. A prominent methodology for this problem is based on the extension of trace norm regularization, which has been used extensively for learning low rank matrices, to the tensor setting. In this paper, we highlight some limitations of this approach and propose an alternative convex relaxation on the Euclidean unit ball. We then describe a technique to solve the associated regularization problem, which builds upon the alternating direction method of multipliers. Experiments on one synthetic dataset and two real datasets indicate that the proposed method improves significantly over tensor trace norm regularization in terms of estimation error, while remaining computationally tractable.

F88 On Decomposing the Proximal Map

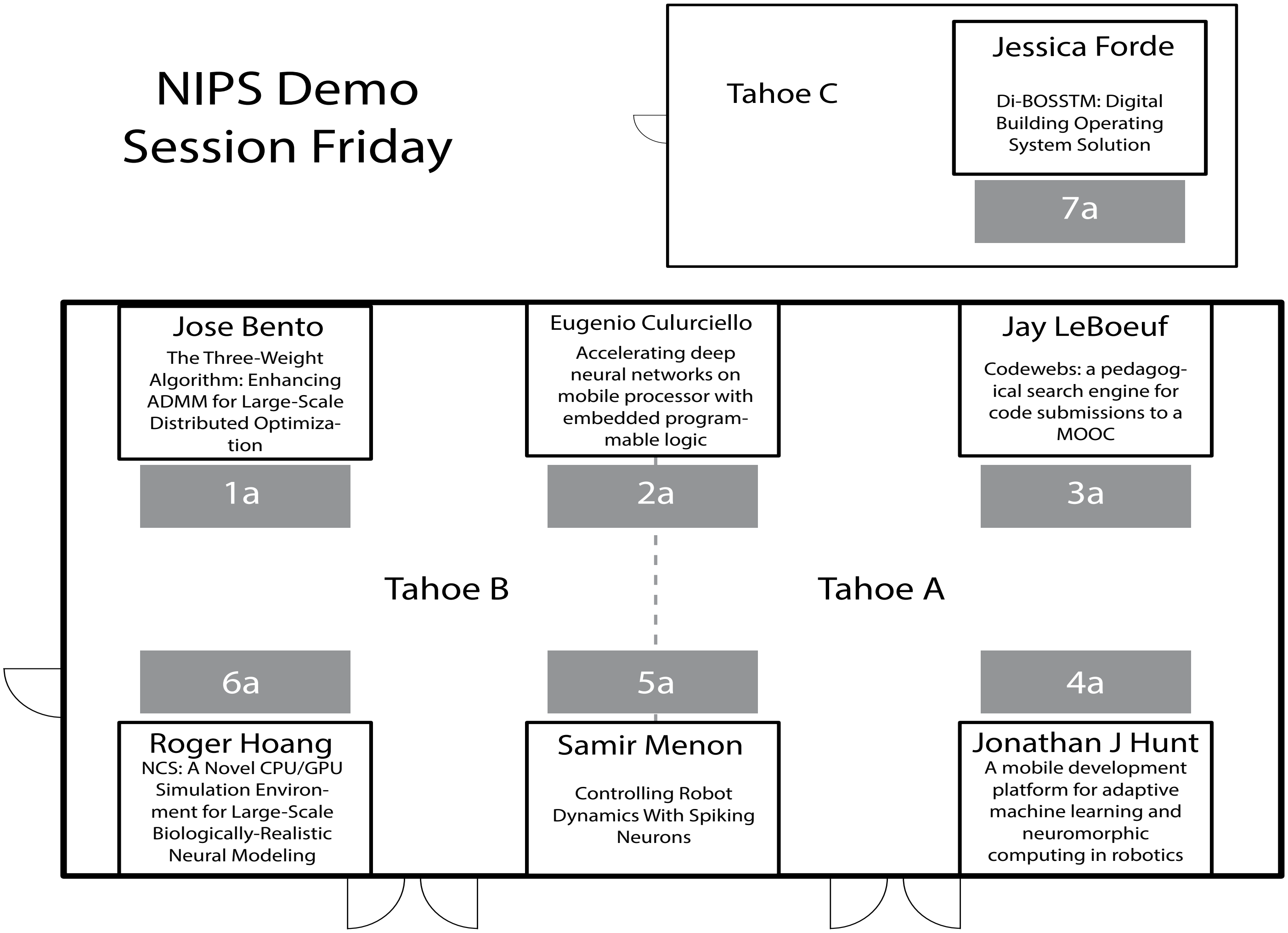
Yao-Liang Yu yaoliang@cs.ualberta.ca
University of Alberta

The proximal map is the key step in gradient-type algorithms, which have become prevalent in large-scale high-dimensional problems. For simple functions this proximal map is available in closed-form while for more complicated functions it can become highly nontrivial. Motivated by the need of combining regularizers to simultaneously induce different types of structures, this paper initiates a systematic investigation of when the proximal map of a sum of functions decomposes into the composition of the proximal maps of the individual summands. We not only unify a few known results scattered in the literature but also discover several new decompositions obtained almost effortlessly from our theory.

F89 Adaptive Anonymity via b-Matching

Krzysztof Choromanski choromanski1@gmail.com
Google Research
Tony Jebara jebara@cs.columbia.edu
Kui Tang kuitang@gmail.com
Columbia University

The adaptive anonymity problem is formalized where each individual shares their data along with an integer value to indicate their personal level of desired privacy. This problem leads to a generalization of k-anonymity to the b-matching setting. Novel algorithms and theory are provided to implement this type of anonymity. The relaxation achieves better utility, admits theoretical privacy guarantees that are as strong, and, most importantly, accommodates a variable level of anonymity for each individual. Empirical results confirm improved utility on benchmark and social data-sets.


NIPS Demo
Session Friday
Tahoe C
Jessica Forde
Di-BOSSTM: Digital Building Operating System Solution
7a
Jose Bento
The Three-Weight Algorithm: Enhancing ADMM for Large-Scale Distributed Optimization
1a
Eugenio Culurciello
Accelerating deep neural networks on mobile processor with embedded programmable logic
2a
Jay LeBoeuf
Codewebs: a pedagogical search engine for code submissions to a MOOC
3a
Tahoe B
Tahoe A
6a
5a
4a
Roger Hoang
NCS: A Novel CPU/GPU Simulation Environment for Large-Scale Biologically-Realistic Neural Modeling
Samir Menon
Controlling Robot Dynamics With Spiking Neurons
Jonathan J Hunt
A mobile development platform for adaptive machine learning and neuromorphic computing in robotics

1A The Three-Weight Algorithm: Enhancing ADMM for Large-Scale Distributed Optimization

Nate Derbinsky, Jose Bento, Jonathan Yedidia
Disney Research

We demonstrate the power of the Three Weight Algorithm (TWA), a new distributed optimization method inspired by the alternating direction method of multipliers (ADMM), a decomposition-coordination method dating back to the 70s. ADMM has received much attention in the convex-optimization community because (1) it is guaranteed to converge to the global optimum of convex problems and (2) it is natural to parallelize the algorithm, as recently emphasized in an extended review (Boyd et al. 2011). Our Three-Weight Algorithm (Derbinsky et al. 2013) is based upon a message-passing version of ADMM and while it maintains the convergence properties for convex problems, and naturally parallelizes, it also greatly improves scaling for non-convex problems. For example, TWA easily finds world-record solutions to large-scale packing problems on a laptop computer and can solve large instances of trajectory-planning problems for dozens of swarm robots in minutes (cf. Bento et al. 2013 at NIPS2013). Our demonstration exemplifies how to solve optimization problems by (a) decomposing the problem into many small, non-convex local cost functions, as well as (b) incorporating top-down sources of problem information, such as human guidance. TWA suffers no disadvantages compared to ADMM for convex problems, but can speed time-to-solution many orders of magnitude for many large-scale non-convex problems.

2A Accelerating Deep Neural Networks on Mobile Processor with Embedded Programmable Logic

Eugenio Culurciello, Aysegul Dundar, Jonghoon Jin
Vinayak Gokhale, Berin Martini
Purdue University

We present a live demonstration of a mobile platform aimed at accelerating deep convolutional neural networks (DCNNs). DCNNs is a powerful way to categorize images. They have achieved state of the art performance in many visual classification benchmarks and have won many competitions. However, their computational costs prevent them from being deployed for real-time applications. We implemented a hardware accelerator on the Xilinx Zynq SoC that can run DCNNs in real-time. The platform consists of a FPGA (PL) and two ARM Cortex-A9 cores (PS). The PL and PS share the same DDR3 memory which allows us to achieve a very high throughput when transferring data between software and co-processor. We will demonstrate live applications of DCNNs on our hardware.

3A Codewebs: a Pedagogical Search Engine for Code Submissions to a MOOC

Jonathan Huang, Leonidas Guibas, Chris Piech
Andy Nguyen
Stanford University

A knowledge of computer science is increasingly becoming an essential career skill in today's world. This demonstration showcases the Codewebs system, which we are developing for leveraging a massive database of code submissions to an online programming intensive course in order to deliver high quality feedback to students. For this demonstration, we will run the Codewebs system using a million code submissions to a machine learning course offered through Coursera. Under the hood, Codewebs can be viewed as a search engine for efficiently querying a massive collection of code submissions that all try to implement the same functionality. With so many submissions of the same assignment, we are able to obtain a dense sampling of the solution space, allowing for submissions to be meaningfully linked into a network connecting highly related solutions. The majority of erroneous solutions even in such a large dataset fall into a relatively small number of clusters that are made evident by the network. Human instructors can then evaluate one or a few assignments from each cluster, and their comments can be diffused along the network to provide specific feedback to a large number of student solutions. One of the novel features of our work is that we can compare code both by syntax and semantics. See also: http://www.stanford.edu/~jhuang11/research/pubs/moocshop13/codeweb.html

4A A Mobile Development Platform for Adaptive Machine Learning and Neuromorphic Computing in Robotics

Jonathan Hunt, Peter O'Connor
Brain Corpporation

One barrier to the incorporation of machine learning approaches in robotics is the lack of hardware platforms with the computational capacity and ease-of-use to facilitate testing new algorithms in standalone robotic devices. To address this problem we have developed bStem, a small, low-power mobile, computational platform running Linux. The board is powered by a state-of-the-art low-power Snapdragon processor with a GPU, FPGA and DSP available to accelerate algorithms. The platform also provides wireless connectivity and several adapter boards for controlling actuators, sensors and other robotic peripherals. In addition to the hardware platform, the bStem will ship with a complete SDK to facilitate rapid development and testing of algorithms in a standalone, embedded system. The bSTEM runs a full desktop version of Ubuntu including a full-featured development environment. bSTEM uses the Python to glue together the many subsystems on board, from reading camera images and sensors to sending motor commands and plotting data into one easy-to-use framework. It also comes with optimized numerical libraries and will run a full machine learning toolkit we are developing (with Python bindings). We will showcasing the potential of the bStem on a robotic platform. We will demonstrate learning algorithms which allow our robots to learn from a human teacher. Attendees will have the opportunity to try training the robots with complex behaviours and responses. Additionally, we will provide test machines where attendees can implement their own algorithms or tweaks using the bSTEM SDKs.

5A Controlling Robot Dynamics With Spiking Neurons

Samir Menon, Kwabena Boahen, Sam Fok
Stanford University

6A NCS: A Novel CPU/GPU Simulation Environment for Large-Scale Biologically-Realistic Neural Modeling

Roger Hoang, Devyani Tanna, Laurence Jayet Bray
Sergiu Dascalu, Frederick Harris, Jr
University of Nevada, Reno

We present a novel CPU/GPU simulation environment for large-scale biological networks, the NeoCortical Simulator version 6 (NCS6). NCS6 is a free, open-source, parallelizable, and scalable simulator, designed to run on clusters of multiple machines, potentially with high performance computing devices in each of them. It has built-in leaky-integrate-and-fire (LIF) and Izhikevich (IZH) neuron models, but users also have the capability to design their own plug-in interface for different neuron types as desired. NCS6 is currently able to simulate one million cells and 100 million synapses in quasi real time by distributing data across these heterogeneous clusters of CPUs and GPUs. A new python interface for NCS6 will also be presented.

7A Di-BOSS™: Digital Building Operating System Solution

Jessica Forde, Vivek Rathod, Hooshmand Shookri
Vaibhav Bandari, Ashwath Rajan, John Min, Ariel
Fan Leon Wu, Ashish Gagneja, Doug Riecken, Lauren
Hannah, Albert Boulanger, Roger Anderson
Columbia University
David Solomon
Selex ES

Our software, Di-BOSS™: Digital Building Operating System Solution, delivers recommendations that enable commercial building managers to make proactive, energy-efficient decisions. Di-BOSS continuously commissions data from multiple operating systems, floor-level occupancy, and ambient and forecast weather conditions. We rely on support vector regression and extremely randomized trees to model the building's thermodynamics and predict energy consumption. On top of our models, we use approximate dynamic programming to optimize operational decisions, such as when to turn on heat. Actionable data and analyses are presented in a real time control panel for building managers.

SATURDAY

ORAL SESSION

SESSION 5 - 9:00 – 10:10 AM

Session Chair: Zoubin Ghahramani

POSNER LECTURE: The Online Revolution: Learning without Limits

Daphne Koller koller@cs.stanford.edu
Stanford University

We are at the cusp of a major transformation in higher education. In the past year, we have seen the advent of MOOCs - massively open online classes (MOOCs) - top-quality courses from the best universities offered for free. These courses exploit technology to provide a real course experience to students, including video content, interactive exercises with meaningful feedback, using both auto-grading and peer-grading, and rich peer-to-peer interaction around the course materials. We now see MOOCs from dozens of top universities, offering courses to millions of students from every country in the world. The courses start from bridge/gateway courses all the way through graduate courses, and span a range of topics including computer science, business, medicine, science, humanities, social sciences, and more. In this talk, I'll discuss this far-reaching experiment in education, including some examples and preliminary analytics. I'll also discuss why we believe this model can support an improved learning experience for on-campus students, via blended learning, and provide unprecedented access to education to millions of students around the world.

Daphne Koller is the Rajeev Motwani Professor of Computer Science at Stanford University and the co-founder and co-CEO of Coursera, a social entrepreneurship company that works with the best universities to connect anyone around the world with the best education, for free. Coursera is the leading MOOC (Massive Open Online Course) platform, and has partnered with dozens of the world's top universities to offer hundreds of courses in a broad range of disciplines to millions of students, spanning every country in the world. In her research life, she works in the area of machine learning and probabilistic modeling, with applications to systems biology and personalized medicine. She is the author of over 200 refereed publications in venues that span a range of disciplines, and has given over 15 keynote talks at major conferences. She is the recipient of many awards, which include the Presidential Early Career Award for Scientists and Engineers (PECASE), the MacArthur Foundation Fellowship, the ACM/Infosys award, and membership in the US National Academy of Engineering. She is also an award winning teacher, who pioneered in her Stanford class many of the ideas that underlie the Coursera user experience. She received her BSc and MSc from the Hebrew University of Jerusalem, and her PhD from Stanford in 1994.

Optimizing Instructional Policies

Robert Lindsey robert.lindsey@colorado.edu
Michael Mozer mozer@colorado.edu
William Huggins w.j.huggins@gmail.com
University of Colorado
Harold Pashler hpashler@ucsd.edu
UC San Diego

Psychologists are interested in developing instructional policies that boost student learning. An instructional policy specifies the manner and content of instruction. For example, in the domain of concept learning, a policy might specify the nature of exemplars chosen over a training sequence. Traditional psychological studies compare several hand-selected policies, e.g., contrasting a policy that selects only difficult-to-classify exemplars with a policy that gradually progresses over the training sequence from easy exemplars to more difficult (known as {\em fading}). We propose an alternative to the traditional methodology in which we define a parameterized space of policies and search this space to identify the optimum policy. For example, in concept learning, policies might be described by a fading function that specifies exemplar difficulty over time. We propose an experimental technique for searching policy spaces using Gaussian process surrogate-based optimization and a generative model of student performance. Instead of evaluating a few experimental conditions each with many human subjects, as the traditional methodology does, our technique evaluates many experimental conditions each with a few subjects. Even though individual subjects provide only a noisy estimate of the population mean, the optimization method allows us to determine the shape of the policy space and identify the global optimum, and is as efficient in its subject budget as a traditional A-B comparison. We evaluate the method via two behavioral studies, and suggest that the method has broad applicability to optimization problems involving humans in domains beyond the educational arena.

SPOTLIGHT SESSION

Session 5, 10:10 – 10:30 AM

- **Linear decision rule as aspiration for simple decision heuristics**
  Ö. Şimşek, Max Planck Institute Berlin
  See abstract S78, page 96

- **Scoring Workers in Crowdsourcing: How Many Control Questions are Enough?**
  Q. Liu, A. Ihler, M. Steyvers, UC Irvine
  See abstract S72, page 95

- **Bayesian Inference and Online Experimental Design for Mapping Neural Microcircuits**
  B. Shababo, A. Pakman, L. Paninski, Columbia University; B. Paige, University of Oxford
  See abstract S18, page 83

- **Sparse Overlapping Sets Lasso for Multitask Learning and its Application to fMRI Analysis**
  N. Rao, C. Cox, R. Nowak, T. Rogers, UW-Madison
  See abstract S46, page 89

- **Lasso Screening Rules via Dual Polytope Projection**
  J. Wang, J. Zhou, P. Wonka, J. Ye, Arizona State University
  See abstract S81, page 97

ORAL SESSION

Session 6, 10:55 – 11:40 AM

Session Chair: Corinna Cortes

NIPS Award 3 - 10:55 – 11:00 AM

A Kernel Test for Three-Variable Interactions

Dino Sejdinovic	dino.sejdinovic@gmail.com
Gatsby Unit, UCL	
Arthur Gretton	arthur.gretton@gmail.com
UCL	
Wicher Bergsma	W.P.Bergsma@lse.ac.uk
LSE	

We introduce kernel nonparametric tests for Lancaster three-variable interaction and for total independence, using embeddings of signed measures into a reproducing kernel Hilbert space. The resulting test statistics are straightforward to compute, and are used in powerful three-variable interaction tests, which are consistent against all alternatives for a large family of reproducing kernels. We show the Lancaster test to be sensitive to cases where two independent causes individually have weak influence on a third dependent variable, but their combined effect has a strong influence. This makes the Lancaster test especially suited to finding structure in directed graphical models, where it outperforms competing nonparametric tests in detecting such V-structures.

More Effective Distributed ML via a Stale Synchronous Parallel Parameter Server

Qirong Ho	qho@cs.cmu.edu
James Cipar	jcipar@cmu.edu
Henggang Cui	hengganc@ece.cmu.edu
Seunghak Lee	seunghak@cs.cmu.edu
Jin Kyu Kim	jinkyuk@cs.cmu.edu
Garth Gibson	garth@cs.cmu.edu
Greg Ganger	ganger@ece.cmu.edu
Eric Xing	epxing@cs.cmu.edu
CMU	
Phil Gibbons	phillip.b.gibbons@intel.com
Intel Labs	

We propose a parameter server system for distributed ML, which follows a Stale Synchronous Parallel (SSP) model of computation that maximizes the time computational workers spend doing useful work on ML algorithms, while still providing correctness guarantees. The parameter server provides an easy-to-use shared interface for read/write access to an ML model's values (parameters and variables), and the SSP model allows distributed workers to read older, stale versions of these values from a local cache, instead of waiting to get them from a central storage. This significantly increases the proportion of time workers spend computing, as opposed to waiting. Furthermore, the SSP model ensures ML algorithm correctness by limiting the maximum age of the stale values. We provide a proof of correctness under SSP, as well as empirical results demonstrating that the SSP model achieves faster algorithm convergence on several different ML problems, compared to fully-synchronous and asynchronous schemes.

SPOTLIGHT SESSION

Session 6, 11:40 AM – 12:05 PM

- **Learning with Invariance via Linear Functionals on Reproducing Kernel Hilbert Space**
  X. Zhang, NICTA; W. Lee, National University of Singapore; Y. Teh, University of Oxford
  See abstract S07, page 81

- **Learning Kernels Using Local Rademacher Complexity**
  C. Cortes, Google Research; M. Kloft, Courant Institute, NYU & Sloan-Kettering Institute (MSKCC); M. Mohri, Courant Institute, NYU & Google
  See abstract S02, page 80

- **Inverse Density as an Inverse Problem: the Fredholm Equation Approach**
  Q. Que, M. Belkin, Ohio State University
  See abstract S55, page 91

- **Regression-tree Tuning in a Streaming Setting**
  S. Kpotufe, F. Orabona, TTI Chicago
  See abstract S40, page 88

- **Non-strongly-convex smooth stochastic approximation with convergence rate O(1/n)**
  F. Bach, INRIA & ENS; E. Moulines, Telecom ParisTech
  See abstract S59, page 92

ORAL SESSION

Session 7, 2:00 – 3:30 PM

Session Chair: Chris Burges

INVITED TALK: Belief Propagation Algorithms: From Matching Problems to Network Discovery in Cancer Genomics

Jennifer Chayes	jchayes@microsoft.com
Microsoft	

We review belief propagation algorithms inspired by the study of phase transitions in combinatorial optimization problems. In particular, we present rigorous results on convergence of such algorithms for matching and associated bargaining problems on networks. We also present a belief propagation algorithm for the prize- collecting Steiner tree problem, for which rigorous convergence results are not yet known. Finally, we show how this algorithm can be used to discover pathways in cancer genomics, and to suggest possible drug targets for cancer therapy. These methods give us the ability to share information across multiple patients to help reconstruct highly patient-specific networks.

Jennifer Tour Chayes is Distinguished Scientist and Managing Director of Microsoft Research New England in Cambridge, Massachusetts, which she co-founded in 2008, and Microsoft Research New York City, which she co-founded in 2012. Chayes was Research Area Manager for Mathematics, Theoretical Computer Science and Cryptography at Microsoft Research Redmond until 2008. Chayes joined Microsoft Research in 1997, when she co-

founded the Theory Group. Before that, she was for many years Professor of Mathematics at UCLA. Chayes is the author of about 125 academic papers and the inventor of over 25 patents. Her research areas include phase transitions in discrete mathematics and computer science, structural and dynamical properties of self-engineered networks, graph algorithms, and algorithmic game theory. Chayes received her B.A. in biology and physics at Wesleyan University, where she graduated first in her class, and her Ph.D. in mathematical physics at Princeton. She did her postdoctoral work in the Mathematics and Physics Departments at Harvard and Cornell. She is the recipient of the National Science Foundation Postdoctoral Fellowship, the Sloan Fellowship, and the UCLA Distinguished Teaching Award. Chayes has recently been the recipient of many leadership awards including the Leadership Award of Women Entrepreneurs in Science and Technology, the Women Who Lead Award, the Women to Watch Award of the Boston Business Journal, and the Women of Leadership Vision Award of the Anita Borg Institute. She has twice been a member of the Institute for Advanced Study in Princeton. Chayes is a Fellow of the American Association for the Advancement of Science, the Fields Institute, the Association for Computing Machinery, and the American Mathematical Society.

Message Passing Inference with Chemical Reaction Networks

Nils Napp nnapp@wyss.harvard.edu
Ryan Adams rpa@seas.harvard.edu
Harvard University

Recent work on molecular programming has explored new possibilities for computational abstractions with biomolecules, including logic gates, neural networks, and linear systems. In the future such abstractions might enable nanoscale devices that can sense and control the world at a molecular scale. Just as in macroscale robotics, it is critical that such devices can learn about their environment and reason under uncertainty. At this small scale, systems are typically modeled as chemical reaction networks. In this work, we develop a procedure that can take arbitrary probabilistic graphical models, represented as factor graphs over discrete random variables, and compile them into chemical reaction networks that implement inference. In particular, we show that marginalization based on sum-product message passing can be implemented in terms of reactions between chemical species whose concentrations represent probabilities. We show algebraically that the steady state concentration of these species correspond to the marginal distributions of the random variables in the graph and validate the results in simulations. As with standard sum-product inference, this procedure yields exact results for tree-structured graphs, and approximate solutions for loopy graphs.

Information-theoretic lower bounds for distributed statistical estimation with communication constraints

Yuchen Zhang yuczhang@eecs.berkeley.edu
John Duchi jduchi@eecs.berkeley.edu
Michael Jordan jordan@cs.berkeley.edu
Martin Wainwright wainwrig@stat.berkeley.edu
UC Berkeley

We establish minimax risk lower bounds for distributed statistical estimation given a budget B of the total number of bits that may be communicated. Such lower bounds in turn reveal the minimum amount of communication required by any procedure to achieve the classical optimal rate for statistical estimation. We study two classes of protocols in which machines send messages either independently or interactively. The lower bounds are established for a variety of problems, from estimating the mean of a population to estimating parameters in linear regression or binary classification.

SPOTLIGHT SESSION

Session 7, 3:30 – 3:50 PM

- **PAC-Bayes-Empirical-Bernstein Inequality**
  I. Tolstikhin, Russian Academy of Sciences; Y. Seldin, Queensland University of Technology & UC Berkeley
  See abstract S39, page 88

- **Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima**
  P. Loh, M. Wainwright, UC Berkeley
  See abstract S38, page 88

- **More data speeds up training time in learning halfspaces over sparse vectors**
  A. Daniely, Hebrew University; N. Linial, S. Shalev-Shwartz, The Hebrew University
  See abstract S34, page 87

- **Convex Calibrated Surrogates for Low-Rank Loss Matrices with Applications to Subset Ranking Losses**
  H. Ramaswamy, S. Agarwal, Indian Institute of Science; A. Tewari, University of Michigan
  See abstract S35, page 87

- **On the Relationship Between Binary Classification, Bipartite Ranking, and Binary Class Probability Estimation**
  H. Narasimhan, S. Agarwal, Indian Institute of Science
  See abstract S36, page 87

ORAL SESSION

Session 8, 4:20 – 5:40 PM

Session Chair: Ofer Dekel

From Bandits to Experts: A Tale of Domination and Independence

Noga Alon nogaa@tau.ac.il
Yishay Mansour mansour.yishay@gmail.com
Tel Aviv University
Nicolò Cesa-Bianchi nicolo.cesa-bianchi@unimi.it
University of Milan
Claudio Gentile claudio.gentile@uninsubria.it
University of Insubria

We consider the partial observability model for multi-armed bandits, introduced by Mannor and Shamir (2011). Our main result is a characterization of regret in the directed observability model in terms of the dominating and

independence numbers of the observability graph. We also show that in the undirected case, the learner can achieve optimal regret without even accessing the observability graph before selecting an action. Both results are shown using variants of the Exp3 algorithm operating on the observability graph in a time-efficient manner.

Eluder Dimension and the Sample Complexity of Optimistic Exploration

Dan Russo dan.joseph.russo@gmail.com
Benjamin Van Roy bvr@stanford.edu
Stanford University

This paper considers the sample complexity of the multi-armed bandit with dependencies among the arms. Some of the most successful algorithms for this problem use the principle of optimism in the face of uncertainty to guide exploration. The clearest example of this is the class of upper confidence bound (UCB) algorithms, but recent work has shown that a simple posterior sampling algorithm, sometimes called Thompson sampling, also shares a close theoretical connection with optimistic approaches. In this paper, we develop a regret bound that holds for both classes of algorithms. This bound applies broadly and can be specialized to many model classes. It depends on a new notion we refer to as the eluder dimension, which measures the degree of dependence among action rewards. Compared to UCB algorithm regret bounds for specific model classes, our general bound matches the best available for linear models and is stronger than the best available for generalized linear models.

Adaptive Market Making via Online Learning

Jacob Abernethy jabernet@umich.edu
University of Pennsylvania
Satyen Kale satyen.kale@gmail.com
IBM Research

We consider the design of strategies for \emph{market making} in a market like a stock, commodity, or currency exchange. In order to obtain profit guarantees for a market maker one typically requires very particular stochastic assumptions on the sequence of price fluctuations of the asset in question. We propose a class of spread-based market making strategies whose performance can be controlled even under worst-case (adversarial) settings. We prove structural properties of these strategies which allows us to design a master algorithm which obtains low regret relative to the best such strategy in hindsight. We run a set of experiments showing favorable performance on real-world price data.

Submodular Optimization with Submodular Cover and Submodular Knapsack Constraints

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

We investigate two new optimization problems — minimizing a submodular function subject to a submodular lower bound constraint (submodular cover) and maximizing a submodular function subject to a submodular upper bound constraint (submodular knapsack). We are motivated by a number of real-world applications in machine learning including sensor placement and data subset selection, which require maximizing a certain submodular function (like coverage or diversity) while simultaneously minimizing another (like cooperative cost). These problems are often posed as minimizing the difference between submodular functions [9, 23] which is in the worst case inapproximable. We show, however, that by phrasing these problems as constrained optimization, which is more natural for many applications, we achieve a number of bounded approximation guarantees. We also show that both these problems are closely related and, an approximation algorithm solving one can be used to obtain an approximation guarantee for the other. We provide hardness results for both problems thus showing that our approximation factors are tight up to log-factors. Finally, we empirically demonstrate the performance and good scalability properties of our algorithms.

SPOTLIGHT SESSION

Session 8, 5:40 – 6:00 PM

- **How to Hedge an Option Against an Adversary: Black-Scholes Pricing is Minimax Optimal**
  J. Abernethy, University of Pennsylvania; P. Bartlett, A. Wibisono, UC Berkeley; R. Frongillo, Microsoft Research
  See abstract S10, page 82

- **Small-Variance Asymptotics for Hidden Markov Models**
  A. Roychowdhury, K. Jiang, B. Kulis, Ohio State University
  See abstract S62, page 93

- **The Total Variation on Hypergraphs - Learning on Hypergraphs Revisited**
  M. Hein, S. Setzer, L. Jost, S. Rangapuram, Saarland University
  See abstract S25, page 85

- **Using multiple samples to learn mixture models**
  J. Lee, Stanford University; R. Gilad-Bachrach, R. Caruana, Microsoft Research
  See abstract S30, page 86

- **Approximate Inference in Continuous Determinantal Processes**
  R. Affandi, University of Pennsylvania; E. Fox, B. Taskar, University of Washington
  See abstract S54, page 91

POSTER SESSION

Session, 7:00 – 11:59 PM

S1 Latent Maximum Margin Clustering
G. Zhou, T. Lan, A. Vahdat, G. Mori

S2 Learning Kernels Using Local Rademacher Complexity
C. Cortes, M. Kloft, M. Mohri

S3 Statistical analysis of coupled time series with Kernel Cross-Spectral Density operators.
M. Besserve, N. Logothetis, B. Schölkopf

S4 Robust Low Rank Kernel Embeddings of Multivariate Distributions
L. Song, B. Dai

S5 B-test: A Non-parametric, Low Variance Kernel Two-sample Test
W. Zaremba, A. Gretton, M. Blaschko

S6 A Kernel Test for Three-Variable Interactions
D. Sejdinovic, A. Gretton, W. Bergsma

S7 Learning with Invariance via Linear Functionals on Reproducing Kernel Hilbert Space
X. Zhang, W. Lee, Y. Teh

S8 On Flat versus Hierarchical Classification in Large-Scale Taxonomies
R. Babbar, I. Partalas, E. Gaussier, M. Amini

S9 Robust Bloom Filters for Large MultiLabel Classification Tasks
M. Cisse, N. Usunier, T. Artières, P. Gallinari

S1 0How to Hedge an Option Against an Adversary: Black-Scholes Pricing is Minimax Optimal
J. Abernethy, P. Bartlett, R. Frongillo, A. Wibisono

S11 Adaptive Market Making via Online Learning
J. Abernethy, S. Kale

S12 Gaussian Process Conditional Copulas with Applications to Financial Time Series
J. Hernández-Lobato, J. Lloyd, D. Hernández-Lobato

S13 Bayesian Inference and Learning in Gaussian Process State-Space Models with Particle MCMC
R. Frigola, F. Lindsten, T. Schon, C. Rasmussen

S14 Multi-Task Bayesian Optimization
K. Swersky, J. Snoek, R. Adams

S15 Efficient Optimization for Sparse Gaussian Process Regression
Y. Cao, M. Brubaker, D. Fleet, A. Hertzmann

S16 Variational Inference for Mahalanobis Distance Metrics in Gaussian Process Regression
M. Titsias, M. Lazaro-Gredilla

S17 It is all in the noise: Efficient multi-task Gaussian process inference with structured residuals
B. Rakitsch, C. Lippert, K. Borgwardt, O. Stegle

S18 Bayesian Inference and Online Experimental Design for Mapping Neural Microcircuits
B. Shababo, B. Paige, A. Pakman, L. Paninski

S19 Spike train entropy-rate estimation using hierarchical Dirichlet process priors
K. Knudson, J. Pillow

S20 Information-theoretic lower bounds for distributed statistical estimation with communication constraints
Y. Zhang, J. Duchi, M. Jordan, M. Wainwright

S21 Designed Measurements for Vector Count Data
L. Wang, D. Carlson, M. Rodrigues, D. Wilcox, R. Calderbank, L. Carin

S22 Dirty Statistical Models
E. Yang, P. Ravikumar

S23 Summary Statistics for Partitionings and Feature Allocations
I. Fidaner, T. Cemgil

S24 Dynamic Clustering via Asymptotics of the Dependent Dirichlet Process Mixture
T. Campbell, M. Liu, B. Kulis, J. How, L. Carin

S25 The Total Variation on Hypergraphs - Learning on Hypergraphs Revisited
M. Hein, S. Setzer, L. Jost, S. Rangapuram

S26 k-Prototype Learning for 3D Rigid Structures
H. Ding, R. Berezney, J. Xu

S27 Distributed k-means and k-median clustering on general communication topologies
M. Balcan, S. Ehrlich, Y. Liang

S28 Multiclass Total Variation Clustering
X. Bresson, T. Laurent, D. Uminsky, J. von Brecht

S29 Learning Multiple Models via Regularized Weighting
D. Vainsencher, S. Mannor, H. Xu

S30 Using multiple samples to learn mixture models
J. Lee, R. Gilad-Bachrach, R. Caruana

S31 Regularized Spectral Clustering under the Degree-Corrected Stochastic Blockmodel
T. Qin, K. Rohe

S32 Moment-based Uniform Deviation Bounds for k-means and Friends
M. Telgarsky, S. Dasgupta

S33 Statistical Active Learning Algorithms
M. Balcan, V. Feldman

S34 More data speeds up training time in learning halfspaces over sparse vectors
A. Daniely, N. Linial, S. Shalev-Shwartz

S35 Convex Calibrated Surrogates for Low-Rank Loss Matrices with Applications to Subset Ranking Losses
H. Ramaswamy, S. Agarwal, A. Tewari

S36 On the Relationship Between Binary Classification, Bipartite Ranking, and Binary Class Probability Estimation
H. Narasimhan, S. Agarwal

S37 Predictive PAC Learning and Process Decompositions
C. Shalizi, A. Kontorovitch

S38 Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima
P. Loh, M. Wainwright

S39 PAC-Bayes-Empirical-Bernstein Inequality
I. Tolstikhin, Y. Seldin

S40 Regression-tree Tuning in a Streaming Setting
S. Kpotufe, F. Orabona

S41 Adaptivity to Local Smoothness and Dimension in Kernel Regression
S. Kpotufe, V. Garg

S42 A Comparative Framework for Preconditioned Lasso Algorithms
F. Wauthier, N. Jojic, M. Jordan

S43 New Subsampling Algorithms for Fast Least Squares Regression
P. Dhillon, Y. Lu, D. Foster, L. Ungar

S44 Faster Ridge Regression via the Subsampled Randomized Hadamard Transform
Y. Lu, P. Dhillon, D. Foster, L. Ungar

S45 Submodular Optimization with Submodular Cover and Submodular Knapsack Constraints
R. Iyer, J. Bilmes

S46 Sparse Overlapping Sets Lasso for Multitask Learning and its Application to fMRI Analysis
N. Rao, C. Cox, R. Nowak, T. Rogers

S47 Sequential Transfer in Multi-armed Bandit with Finite Set of Models
M. Gheshlaghi azar, A. Lazaric, E. Brunskill

S48 Eluder Dimension and the Sample Complexity of Optimistic Exploration
D. Russo, B. Van Roy

S49 Prior-free and prior-dependent regret bounds for Thompson Sampling
S. Bubeck, C. Liu

S50 From Bandits to Experts: A Tale of Domination and Independence
N. Alon, N. Cesa-Bianchi, C. Gentile, Y. Mansour

S51 Two-Target Algorithms for Infinite-Armed Bandits with Bernoulli Rewards
T. Bonald, A. Proutiere

S52 Thompson Sampling for 1-Dimensional Exponential Family Bandits
N. Korda, E. Kaufmann, R. Munos

S53 Bayesian Mixture Modelling and Inference based Thompson Sampling in Monte-Carlo Tree Search
A. Bai, F. Wu, X. Chen

S54 Approximate Inference in Continuous Determinantal Processes
R. Affandi, E. Fox, B. Taskar

S55 Inverse Density as an Inverse Problem: the Fredholm Equation Approach
Q. Que, M. Belkin

S56 Density estimation from unweighted k-nearest neighbor graphs: a roadmap
U. Von Luxburg, M. Alamgir

S57 Sketching Structured Matrices for Faster Nonlinear Regression
H. Avron, V. Sindhwani, D. Woodruff

S58 More Effective Distributed ML via a Stale Synchronous Parallel Parameter Server
Q. Ho, J. Cipar, H. Cui, S. Lee, J. Kim, P. Gibbons, G. Gibson, G. Ganger, E. Xing

S59 Non-strongly-convex smooth stochastic approximation with convergence rate O(1/n)
F. Bach, E. Moulines

S60 Trading Computation for Communication: Distributed Stochastic Dual Coordinate Ascent
T. Yang

S61 Locally Adaptive Bayesian Multivariate Time Series
D. Durante, B. Scarpa, D. Dunson

S62 Small-Variance Asymptotics for Hidden Markov Models
A. Roychowdhury, K. Jiang, B. Kulis

S63 A Latent Source Model for Nonparametric Time Series Classification
G. Chen, S. Nikolov, D. Shah

S64 Multilinear Dynamical Systems for Tensor Time Series
M. Rogers, L. Li, S. Russell

S65 What do row and column marginals reveal about your dataset?
B. Golshan, J. Byers, E. Terzi

S66 Error-Minimizing Estimates and Universal Entry-Wise Error Bounds for Low-Rank Matrix Completion
F. Kiraly, L. Theran

S67 Synthesizing Robust Plans under Incomplete Domain Models
T. Nguyen, S. Kambhampati, M. Do

S68 Message Passing Inference with Chemical Reaction Networks
N. Napp, R. Adams

S69 Which Space Partitioning Tree to Use for Search?
P. Ram, A. Gray

S70 Solving inverse problem of Markov chain with partial observations
T. Morimura, T. Osogami, T. Ide

S71 Robust Data-Driven Dynamic Programming
G. Hanasusanto, D. Kuhn

S72 Scoring Workers in Crowdsourcing: How Many Control Questions are Enough?
Q. Liu, A. Ihler, M. Steyvers

S73 Online Variational Approximations to non-Exponential Family Change Point Models: With Application to Radar Tracking
R. Turner, S. Bottone, C. Stanek

S74 q-OCSVM: A q-Quantile Estimator for High-Dimensional Distributions
A. Glazer, M. Lindenbaoum, S. Markovitch

S75 Unsupervised Structure Learning of Stochastic And-Or Grammars
K. Tu, M. Pavlovskaia, S. Zhu

S76 Rapid Distance-Based Outlier Detection via Sampling
M. Sugiyama, K. Borgwardt

S77 One-shot learning by inverting a compositional causal process
B. Lake, R. Salakhutdinov, J. Tenenbaum

S78 Linear decision rule as aspiration for simple decision heuristics
Ö. Şimşek

S79 Optimizing Instructional Policies
R. Lindsey, M. Mozer, W. Huggins, H. Pashler

S80 Stochastic Majorization-Minimization Algorithms for Large-Scale Optimization
J. Mairal

S81 Lasso Screening Rules via Dual Polytope Projection
J. Wang, J. Zhou, P. Wonka, J. Ye

S82 Robust Transfer Principal Component Analysis with Rank Constraints
Y. Guo

S83 Online Robust PCA via Stochastic Optimization
J. Feng, H. Xu, S. Yan

S84 The Fast Convergence of Incremental PCA
A. Balsubramani, S. Dasgupta, Y. Freund

S85 Probabilistic Principal Geodesic Analysis
M. Zhang, P. Fletcher

S86 Fast Algorithms for Gaussian Noise Invariant Independent Component Analysis
J. Voss, L. Rademacher, M. Belkin

S87 Online PCA for Contaminated Data
J. Feng, H. Xu, S. Mannor, S. Yan

S88 Fantope Projection and Selection: A near-optimal convex relaxation of sparse PCA
V. Vu, J. Cho, J. Lei, K. Rohe

S89 One-shot learning and big data with n=2
L. Dicker, D. Foster

S90 The Randomized Dependence Coefficient
D. Lopez-Paz, P. Hennig, B. Schölkopf

S**91 Sign Cauchy Projections and Chi-Square Kernel**
P. Li, G. Samorodnitsk, J. Hopcroft

DEMONSTRATIONS

7:00 – 11:59 PM

Cross-Lingual Technologies: Text to Logic Mapping, Search and Classification over 100 Languages,
J. Rupnik, A. Muhic, B. Fortuna, J. Starc, M. Grobelnik, M. Witbrock

Deep Content-Based Music Recommendation,
A. van den Oord, S. Dieleman, B. Schrauwen

Distributed Representations of Words and Phrases and their Compositionality,
T. Mikolov, K. Chen, G. Corrado

Easy Text Classification with Machine Learning,
R. Socher, R. Paulus, B. McCann, A. Ng

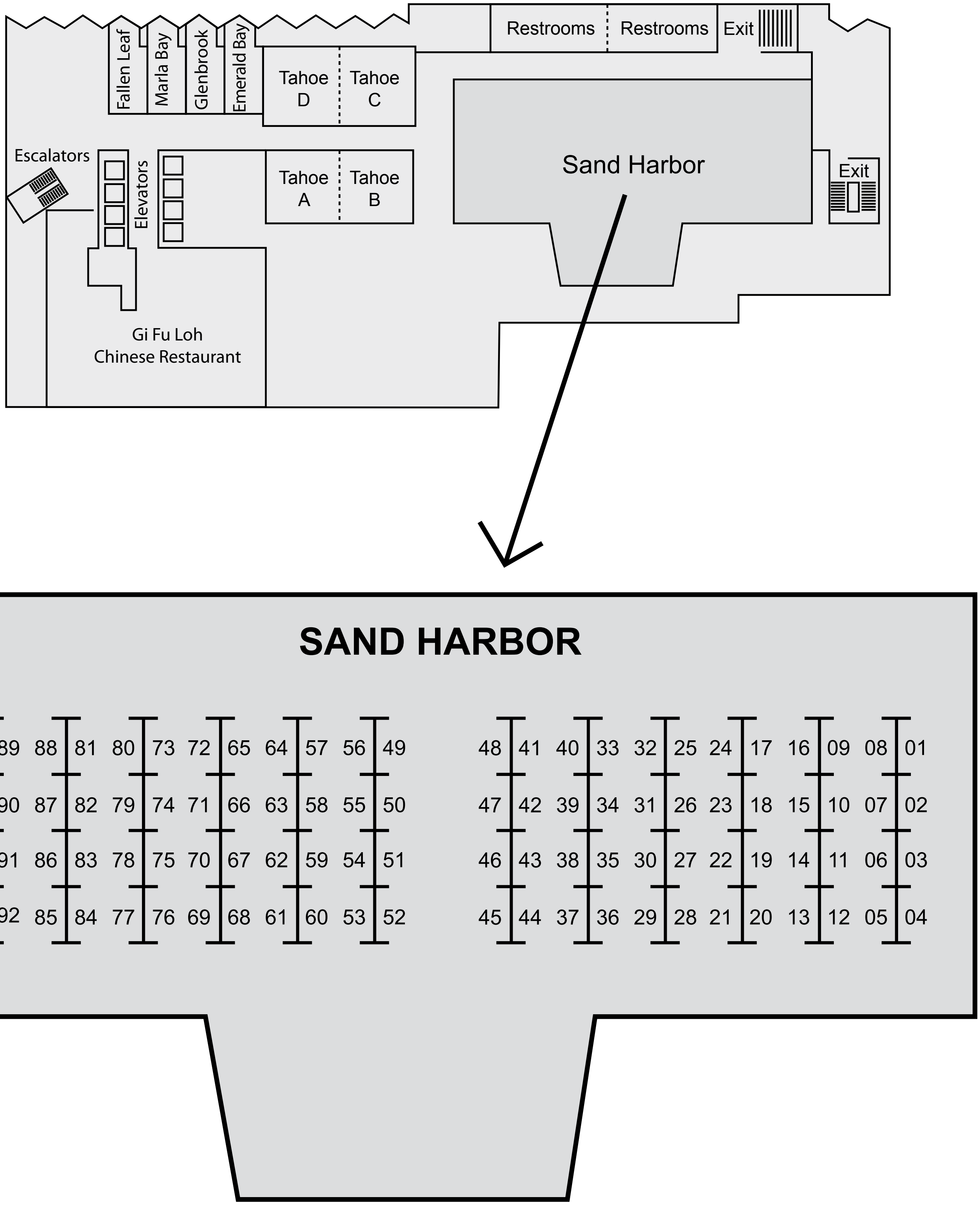
Making Smooth Topical Connections on Touch Devices,
N. Jojic, A. Perina, A. Truski

Topic Modeling for Robots,
Y. Girdhar, G. Dudek

SATURDAY POSTER FLOORPLAN

HARRAH'S

2ND FLOOR SPECIAL EVENTS CENTER



S01 Latent Maximum Margin Clustering

Guang-Tong Zhou gza11@sfu.ca
Tian Lan taran.lan1986@gmail.com
Arash Vahdat avahdat@sfu.ca
Greg Mori mori@cs.sfu.ca
Simon Fraser University

We present a maximum margin framework that clusters data using latent variables. Using latent representations enables our framework to model unobserved information embedded in the data. We implement our idea by large margin learning, and develop an alternating descent algorithm to effectively solve the resultant non-convex optimization problem. We instantiate our latent maximum margin clustering framework with tag-based video clustering tasks, where each video is represented by a latent tag model describing the presence or absence of video tags. Experimental results obtained on three standard datasets show that the proposed method outperforms non-latent maximum margin clustering as well as conventional clustering approaches.

S02 Learning Kernels Using Local Rademacher Complexity

Corinna Cortes corinna@google.com
Google Research
Marius Kloft mkloft@cs.nyu.edu
Courant Institute, NYU & Sloan-Kettering Institute
Mehryar Mohri mohri@google.com
Courant Institute, NYU & Google

We use the notion of local Rademacher complexity to design new algorithms for learning kernels. Our algorithms thereby benefit from the sharper learning bounds based on that notion which, under certain general conditions, guarantee a faster convergence rate. We devise two new learning kernel algorithms: one based on a convex optimization problem for which we give an efficient solution using existing learning kernel techniques, and another one that can be formulated as a DC-programming problem for which we describe a solution in detail. We also report the results of experiments with both algorithms in both binary and multi-class classification tasks.

S03 Statistical analysis of coupled time series with Kernel Cross-Spectral Density operators.

Michel Besserve
michel.besserve@tuebingen.mpg.de
MPI for Intelligent Systems
Nikos Logothetis
nikos.logothetis@tuebingen.mpg.de
MPI for Biological Cybernetics
Bernhard Schölkopf bs@tuebingen.mpg.de
MPI Tübingen

Many applications require the analysis of complex interactions between time series. These interactions can be non-linear and involve vector valued as well as complex data structures such as graphs or strings. Here we provide a general framework for the statistical analysis of these interactions when random variables are sampled from stationary time-series of arbitrary objects. To achieve this goal we analyze the properties of the kernel cross-spectral density operator induced by positive definite kernels on arbitrary input domains. This framework enables us to develop an independence test between time series as well as a similarity measure to compare different types of coupling. The performance of our test is compared to the HSIC test using i.i.d. assumptions, showing improvement in terms of detection errors as well as the suitability of this approach for testing dependency in complex dynamical systems. Finally, we use this approach to characterize complex interactions in electrophysiological neural time series.

S04 Robust Low Rank Kernel Embeddings of Multivariate Distributions

Le Song lsong@cc.gatech.edu
Bo Dai bodai@gatech.edu
Georgia Tech

Kernel embedding of distributions has led to many recent advances in machine learning. However, latent and low rank structures prevalent in real world distributions have rarely been taken into account in this setting. Furthermore, no prior work in kernel embedding literature has addressed the issue of robust embedding when the latent and low rank information are misspecified. In this paper, we propose a hierarchical low rank decomposition of kernels embeddings which can exploit such low rank structures in data while being robust to model misspecification. We also illustrate with empirical evidence that the estimated low rank embeddings lead to improved performance in density estimation.

S05 B-test: A Non-parametric, Low Variance Kernel Two-sample Test

Wojciech Zaremba woj.zaremba@gmail.com
Matthew Blaschko matthew.blaschko@inria.fr
École Centrale Paris
Arthur Gretton arthur.gretton@gmail.com
UCL

We propose a family of maximum mean discrepancy (MMD) kernel two-sample tests that have low sample complexity and are consistent. The test has a hyperparameter that allows one to control the tradeoff between sample complexity and computational time. Our family of tests, which we denote as B-tests, is both computationally and statistically efficient, combining favorable properties of previously proposed MMD two-sample tests. It does so by better leveraging samples to produce low variance estimates in the finite sample case, while avoiding a quadratic number of kernel evaluations and complex null-hypothesis approximation as would be required by tests relying on one sample U-statistics. The B-test uses a smaller than quadratic number of kernel evaluations and avoids completely the computational burden of complex null-hypothesis approximation while maintaining consistency and probabilistically conservative thresholds on Type I error. Finally, recent results of combining multiple kernels transfer seamlessly to our hypothesis test, allowing a further increase in discriminative power and decrease in sample complexity.

S06 A Kernel Test for Three-Variable Interactions

Dino Sejdinovic dino.sejdinovic@gmail.com
Gatsby Unit, UCL
Arthur Gretton arthur.gretton@gmail.com
UCL
Wicher Bergsma W.P.Bergsma@lse.ac.uk
LSE

We introduce kernel nonparametric tests for Lancaster three-variable interaction and for total independence, using embeddings of signed measures into a reproducing kernel Hilbert space. The resulting test statistics are straightforward to compute, and are used in powerful three-variable interaction tests, which are consistent against all alternatives for a large family of reproducing kernels. We show the Lancaster test to be sensitive to cases where two independent causes individually have weak influence on a third dependent variable, but their combined effect has a strong influence. This makes the Lancaster test especially suited to finding structure in directed graphical models, where it outperforms competing nonparametric tests in detecting such V-structures.

S07 Learning with Invariance via Linear Functionals on Reproducing Kernel Hilbert Space

Xinhua Zhang xinhua.zhang.cs@gmail.com
NICTA
Wee Sun Lee leews@comp.nus.edu.sg
National University of Singapore
Yee Whye Teh y.w.teh@stats.ox.ac.uk
University of Oxford

Incorporating invariance information is important for many learning problems. To exploit invariances, most existing methods resort to approximations that either lead to expensive optimization problems such as semi-definite programming, or rely on separation oracles to retain tractability. Some methods further limit the space of functions and settle for non-convex models. In this paper, we propose a framework for learning in reproducing kernel Hilbert spaces (RKHS) using local invariances that explicitly characterize the behavior of the target function around data instances. These invariances are \emph{compactly} encoded as linear functionals whose value are penalized by some loss function. Based on a representer theorem that we establish, our formulation can be efficiently optimized via a convex program. For the representer theorem to hold, the linear functionals are required to be bounded in the RKHS, and we show that this is true for a variety of commonly used RKHS and invariances. Experiments on learning with unlabeled data and transform invariances show that the proposed method yields better or similar results compared with the state of the art.

S08 On Flat versus Hierarchical Classification in Large-Scale Taxonomies

Rohit Babbar rohit.babbar@imag.fr
Eric Gaussier eric.gaussier@imag.fr
Massih-Reza Amini Massih-Reza.Amini@imag.fr
Université Joseph Fourier, Grenoble
Ioannis Partalas ioannis.partalas@imag.fr
UJF/LIG

We study in this paper flat and hierarchical classification strategies in the context of large-scale taxonomies. To this end, we first propose a multiclass, hierarchical data dependent bound on the generalization error of classifiers deployed in large-scale taxonomies. This bound provides an explanation to several empirical results reported in the literature, related to the performance of flat and hierarchical classifiers. We then introduce another type of bounds targeting the approximation error of a family of classifiers, and derive from it features used in a meta-classifier to decide which nodes to prune (or flatten) in a large-scale taxonomy. We finally illustrate the theoretical developments through several experiments conducted on two widely used taxonomies.

S09 Robust Bloom Filters for Large MultiLabel Classification Tasks

Moustapha Cisse cisse@poleia.lip6.fr
Thierry Artières thierry.artieres@lip6.fr
Patrick Gallinari patrick.gallinari@lip6.fr
LIP6/UPMC
Nicolas Usunier nicolas.usunier@utc.fr
Université de Technologie de Compiègne (UTC)

This paper presents an approach to multilabel classification (MLC) with a large number of labels. Our approach is a reduction to binary classification in which label sets are represented by low dimensional binary vectors. This representation follows the principle of Bloom filters, a space-efficient data structure originally designed for approximate membership testing. We show that a naive application of Bloom filters in MLC is not robust to individual binary classifiers' errors. We then present an approach that exploits a specific feature of real-world datasets when the number of labels is large: many labels (almost) never appear together. Our approch is provably robust, has sublinear training and inference complexity with respect to the number of labels, and compares favorably to state-of-the-art algorithms on two large scale multilabel datasets.

S10 How to Hedge an Option Against an Adversary: Black-Scholes Pricing is Minimax Optimal

Jacob Abernethy jabernet@umich.edu
University of Pennsylvania
Peter Bartlett bartlett@cs.berkeley.edu
Andre Wibisono wibisono@eecs.berkeley.edu
UC Berkeley
Rafael Frongillo raf@cs.berkeley.edu
Microsoft Research

We consider a popular problem in finance, option pricing, through the lens of an online learning game between Nature and an Investor. In the Black-Scholes option pricing model from 1973, the Investor can continuously hedge the risk of an option by trading the underlying asset, assuming that the asset's price fluctuates according to Geometric Brownian Motion (GBM). We consider a worst-case model, in which Nature chooses a sequence of price fluctuations under a cumulative quadratic volatility constraint, and the Investor can make a sequence of hedging decisions. Our main result is to show that the value of our proposed game, which is the “regret'' of hedging strategy, converges to the Black-Scholes option price. We use significantly weaker assumptions than previous work---for instance, we allow large jumps in the asset price---and show that the Black-Scholes hedging strategy is near-optimal for the Investor even in this non-stochastic framework.

S11 Adaptive Market Making via Online Learning

Jacob Abernethy jabernet@umich.edu
University of Pennsylvania
Satyen Kale satyen.kale@gmail.com
IBM Research

We consider the design of strategies for *market making* in a market like a stock, commodity, or currency exchange. In order to obtain profit guarantees for a market maker one typically requires very particular stochastic assumptions on the sequence of price fluctuations of the asset in question. We propose a class of spread-based market making strategies whose performance can be controlled even under worst-case (adversarial) settings. We prove structural properties of these strategies which allows us to design a master algorithm which obtains low regret relative to the best such strategy in hindsight. We run a set of experiments showing favorable performance on real-world price data.

S12 Gaussian Process Conditional Copulas with Applications to Financial Time Series

José Miguel Hernández-Lobato jmh233@cam.ac.uk
James Lloyd jrl44@cam.ac.uk
University of Cambridge
Daniel Hernández-Lobato daniel.hernandez@uam.es
Universidad Autónoma de Madrid

The estimation of dependencies between multiple variables is a central problem in the analysis of financial time series. A common approach is to express these dependencies in terms of a copula function. Typically the copula function is assumed to be constant but this may be innacurate when there are covariates that could have a large influence on the dependence structure of the data. To account for this, a Bayesian framework for the estimation of conditional copulas is proposed. In this framework the parameters of a copula are non-linearly related to some arbitrary conditioning variables. We evaluate the ability of our method to predict time-varying dependencies on several equities and currencies and observe consistent performance gains compared to static copula models and other time-varying copula methods.

S13 Bayesian Inference and Learning in Gaussian Process State-Space Models with Particle MCMC

Roger Frigola rf342@cam.ac.uk
Carl Rasmussen cer54@cam.ac.uk
University of Cambridge
Fredrik Lindsten lindsten@isy.liu.se
Linköping University
Thomas Schon thomas.schon@it.uu.se
Uppsala University

State-space models are successfully used in many areas of science, engineering and economics to model time series and dynamical systems. We present a fully Bayesian approach to inference and learning in nonlinear nonparametric state-space models. We place a Gaussian process prior over the transition dynamics, resulting in a flexible model able to capture complex dynamical phenomena. However, to enable efficient inference, we marginalize over the dynamics of the model and instead infer directly the joint smoothing distribution through the use of specially tailored Particle Markov Chain Monte Carlo samplers. Once an approximation of the smoothing distribution is computed, the state transition predictive distribution can be formulated analytically. We make use of sparse Gaussian process models to greatly reduce the computational complexity of the approach.

S14 Multi-Task Bayesian Optimization

Kevin Swersky kswersky@cs.toronto.edu
Jasper Snoek jasper@cs.toronto.edu
University of Toronto
Ryan Adams rpa@seas.harvard.edu
Harvard University

Bayesian optimization has recently been proposed as a framework for automatically tuning the hyperparameters of machine learning models and has been shown to yield state-of-the-art performance with impressive ease and efficiency. In this paper, we explore whether it is possible to transfer the knowledge gained from previous optimizations to new tasks in order to find optimal hyperparameter settings more efficiently. Our approach is based on extending multi-task Gaussian processes to the framework of Bayesian optimization. We show that this method significantly speeds up the optimization process when compared to the standard single-task approach. We further propose a straightforward extension of our algorithm in order to jointly minimize the average error across multiple tasks and demonstrate how this can be used to greatly speed up k-fold cross-validation. Lastly, our

most significant contribution is an adaptation of a recently proposed acquisition function, entropy search, to the cost-sensitive and multi-task settings. We demonstrate the utility of this new acquisition function by utilizing a small dataset in order to explore hyperparameter settings for a large dataset. Our algorithm dynamically chooses which dataset to query in order to yield the most information per unit cost.

S15 Efficient Optimization for Sparse Gaussian Process Regression

Yanshuai Cao caoy@cs.toronto.edu
David Fleet fleet@cs.toronto.edu
University of Toronto
Marcus Brubaker mbrubake@cs.toronto.edu
TTI Chicago
Aaron Hertzmann hertzman@adobe.com
Adobe Research

We propose an efficient discrete optimization algorithm for selecting a subset of training data to induce sparsity for Gaussian process regression. The algorithm estimates this inducing set and the hyperparameters using a single objective, either the marginal likelihood or a variational free energy. The space and time complexity are linear in the training set size, and the algorithm can be applied to large regression problems on discrete or continuous domains. Empirical evaluation shows state-of-art performance in the discrete case and competitive results in the continuous case.

S16 Variational Inference for Mahalanobis Distance Metrics in Gaussian Process Regression

Michalis Titsias mtitsias@aueb.gr
Athens University of Economics and Business
Miguel Lazaro-Gredilla miguel@tsc.uc3m.es
Universidad Carlos III de Madrid

We introduce a novel variational method that allows to approximately integrate out kernel hyperparameters, such as length-scales, in Gaussian process regression. This approach consists of a novel variant of the variational framework that has been recently developed for the Gaussian process latent variable model which additionally makes use of a standardised representation of the Gaussian process. We consider this technique for learning Mahalanobis distance metrics in a Gaussian process regression setting and provide experimental evaluations and comparisons with existing methods by considering datasets with high-dimensional inputs.

S17 It is all in the noise: Efficient multi-task Gaussian process inference with structured residuals

Barbara Rakitsch rakitsch@tuebingen.mpg.de
MPI Tübingen
Christoph Lippert lippert@microsoft.com
Microsoft Research
Karsten Borgwardt karsten.borgwardt@tuebingen.mpg.de
MPI Tübingen & University of Tübingen
Oliver Stegle oliver.stegle@ebi.ac.uk
EMBL-EBI

Multi-task prediction models are widely being used to couple regressors or classification models by sharing information across related tasks. A common pitfall of these models is that they assume that the output tasks are independent conditioned on the inputs. Here, we propose a multi-task Gaussian process approach to model both the relatedness between regressors as well as the task correlations in the residuals, in order to more accurately identify true sharing between regressors. The resulting Gaussian model has a covariance term that is the sum of Kronecker products, for which efficient parameter inference and out of sample prediction are feasible. On both synthetic examples and applications to phenotype prediction in genetics, we find substantial benefits of modeling structured noise compared to established alternatives.

S18 Bayesian Inference and Online Experimental Design for Mapping Neural Microcircuits

Ben Shababo bms2156@columbia.edu
Ari Pakman aripakman@gmail.com
Liam Paninski liam@stat.columbia.edu
Columbia University
Brooks Paige brooks@robots.ox.ac.uk
University of Oxford

We develop an inference and optimal design procedure for recovering synaptic weights in neural microcircuits. We base our procedure on data from an experiment in which populations of putative presynaptic neurons can be stimulated while a subthreshold recording is made from a single postsynaptic neuron. We present a realistic statistical model which accounts for the main sources of variability in this experiment and allows for large amounts of information about the biological system to be incorporated if available. We then present a simpler model to facilitate online experimental design which entails the use of efficient Bayesian inference. The optimized approach results in equal quality posterior estimates of the synaptic weights in roughly half the number of experimental trials under experimentally realistic conditions, tested on synthetic data generated from the full model.

S19 Spike train entropy-rate estimation using hierarchical Dirichlet process priors

Karin Knudson kknudson@math.utexas.edu
Jonathan Pillow pillow@mail.utexas.edu
UT Austin

Entropy rate quantifies the amount of disorder in a stochastic process. For spiking neurons, the entropy rate places an upper bound on the rate at which the spike train can convey stimulus information, and a large literature has focused on the problem of estimating entropy rate from spike train data. Here we present Bayes Least Squares and Empirical Bayesian entropy rate estimators for binary spike trains using Hierarchical Dirichlet Process (HDP) priors. Our estimator leverages the fact that the entropy rate of an ergodic Markov Chain with known transition probabilities can be calculated analytically, and many stochastic processes that are non-Markovian can still be well approximated by Markov processes of sufficient depth. Choosing an appropriate depth of Markov model presents challenges due to possibly long time dependencies and short data sequences: a deeper model can better account for long time-dependencies, but is more difficult to infer from limited data. Our approach mitigates this difficulty by using a hierarchical prior to share statistical power across Markov chains of different depths. We present both a fully Bayesian and empirical Bayes entropy rate estimator based on this model, and demonstrate their performance on simulated and real neural spike train data.

S20 Information-theoretic lower bounds for distributed statistical estimation with communication constraints

Yuchen Zhang yuczhang@eecs.berkeley.edu
John Duchi jduchi@eecs.berkeley.edu
Michael Jordan jordan@cs.berkeley.edu
Martin Wainwright wainwrig@stat.berkeley.edu
UC Berkeley

We establish minimax risk lower bounds for distributed statistical estimation given a budget B of the total number of bits that may be communicated. Such lower bounds in turn reveal the minimum amount of communication required by any procedure to achieve the classical optimal rate for statistical estimation. We study two classes of protocols in which machines send messages either independently or interactively. The lower bounds are established for a variety of problems, from estimating the mean of a population to estimating parameters in linear regression or binary classification.

S21 Designed Measurements for Vector Count Data

Liming Wang liming.w@duke.edu
David Carlson david.carlson@duke.edu
Robert Calderbank robert.calderbank@duke.edu
Lawrence Carin lcarin@duke.edu
Duke University
Miguel Rodrigues m.rodrigues@ucl.ac.uk
UCL
David Wilcox wilcoxds@purdue.edu
Purdue University

We consider design of linear projection measurements for a vector Poisson signal model. The projections are performed on the vector Poisson rate, $X \in \mathbb{R}^n_+$, and the observed data are a vector of counts, $Y \in \mathbb{Z}^m_+$. The projection matrix is designed by maximizing mutual information between Y and X, $I(Y;X)$. When there is a latent class label $C \in \{1,\ldots,L\}$ associated with X, we consider the mutual information with respect to Y and C, $I(Y;C)$. New analytic expressions for the gradient of $I(Y;X)$ and $I(Y;C)$ are presented, with gradient performed with respect to the measurement matrix. Connections are made to the more widely studied Gaussian measurement model. Example results are presented for compressive topic modeling of a document corpora (word counting), and hyperspectral compressive sensing for chemical classification (photon counting).

S22 Dirty Statistical Models

Eunho Yang eunho@cs.utexas.edu
Pradeep Ravikumar pradeepr@cs.utexas.edu
UT Austin

We provide a unified framework for the high-dimensional analysis of “superposition-structured” or “dirty” statistical models: where the model parameters are a “superposition” of structurally constrained parameters. We allow for any number and types of structures, and any statistical model. We consider the general class of M-estimators that minimize the sum of any loss function, and an instance of what we call a “hybrid” regularization, that is the infimal convolution of weighted regularization functions, one for each structural component. We provide corollaries showcasing our unified framework for varied statistical models such as linear regression, multiple regression and principal component analysis, over varied superposition structures.

S23 Summary Statistics for Partitionings and Feature Allocations

Isik Fidaner fidaner@gmail.com
A. Taylan Cemgil taylan.cemgil@boun.edu.tr
Boğaziçi University

Infinite mixture models are commonly used for clustering. One can sample from the posterior of mixture assignments by Monte Carlo methods or find its maximum a posteriori solution by optimization. However, in some problems the posterior is diffuse and it is hard to interpret the sampled partitionings. In this paper, we introduce novel statistics based on block sizes for representing sample sets of partitionings and feature allocations. We develop an element-based definition of entropy to

quantify segmentation among their elements. Then we propose a simple algorithm called entropy agglomeration (EA) to summarize and visualize this information. Experiments on various infinite mixture posteriors as well as a feature allocation dataset demonstrate that the proposed statistics are useful in practice.

S24 Dynamic Clustering via Asymptotics of the Dependent Dirichlet Process Mixture

Trevor Campbell tdjc@mit.edu
Jonathan How jhow@mit.edu
Massachusetts Institute of Technology
Miao Liu miao.liu@duke.edu
Lawrence Carin lcarin@duke.edu
Duke University
Brian Kulis brian.kulis@gmail.com
Ohio State University

This paper presents a novel algorithm, based upon the dependent Dirichlet process mixture model (DDPMM), for clustering batch-sequential data containing an unknown number of evolving clusters. The algorithm is derived via a low-variance asymptotic analysis of the Gibbs sampling algorithm for the DDPMM, and provides a hard clustering with convergence guarantees similar to those of the k-means algorithm. Empirical results from a synthetic test with moving Gaussian clusters and a test with real ADS-B aircraft trajectory data demonstrate that the algorithm requires orders of magnitude less computational time than contemporary probabilistic and hard clustering algorithms, while providing higher accuracy on the examined datasets.

S25 The Total Variation on Hypergraphs - Learning on Hypergraphs Revisited

Matthias Hein hein@cs.uni-saarland.de
Simon Setzer setzer@mia.uni-saarland.de
Leonardo Jost leo@santorin.cs.uni-saarland.de
Syama Sundar Rangapuram
srangapu@mpi-inf.mpg.de
Saarland University

Hypergraphs allow to encode higher-order relationships in data and are thus a very flexible modeling tool. Current learning methods are either based on approximations of the hypergraphs via graphs or on tensor methods which are only applicable under special conditions. In this paper we present a new learning framework on hypergraphs which fully uses the hypergraph structure. The key element is a family of regularization functionals based on the total variation on hypergraphs.

S26 k-Prototype Learning for 3D Rigid Structures

Hu Ding huding@buffalo.edu
Jinhui Xu jinhui@buffalo.edu
SUNY at Buffalo
Ronald Berezney berezney@buffalo.edu
University of Buffalo

In this paper, we study the following new variant of prototype learning, called *k-prototype learning problem for 3D rigid structures*: Given a set of 3D rigid structures, find a set of k rigid structures so that each of them is a prototype for a cluster of the given rigid structures and the total cost (or dissimilarity) is minimized. Prototype learning is a core problem in machine learning and has a wide range of applications in many areas. Existing results on this problem have mainly focused on the graph domain. In this paper, we present the first algorithm for learning multiple prototypes from 3D rigid structures. Our result is based on a number of new insights to rigid structures alignment, clustering, and prototype reconstruction, and is practically efficient with quality guarantee. We validate our approach using two type of data sets, random data and biological data of chromosome territories. Experiments suggest that our approach can effectively learn prototypes in both types of data.

S27 Distributed k-means and k-median clustering on general communication topologies

Maria-Florina Balcan ninamf@cc.gatech.edu
Steven Ehrlich sehrlich@cc.gatech.edu
Yingyu Liang yliang39@gatech.edu
Georgia Tech

This paper provides new algorithms for distributed clustering for two popular center-based objectives, k-median and k-means. These algorithms have provable guarantees and improve communication complexity over existing approaches. Following a classic approach in clustering by har2004coresets, we reduce the problem of finding a clustering with low cost to the problem of finding a `coreset' of small size. We provide a distributed method for constructing a global coreset which improves over the previous methods by reducing the communication complexity, and which works over general communication topologies. We provide experimental evidence for this approach on both synthetic and real data sets.

S28 Multiclass Total Variation Clustering

Xavier Bresson xavier.bresson@unil.ch
City University of Hong Kong
Thomas Laurent tlaurent@lmu.edu
Loyola Marymount University
David Uminsky duminsky@usfca.edu
University of San Francisco
James von Brecht jub@math.ucla.edu
UCLA

Ideas from the image processing literature have recently motivated a new set of clustering algorithms that rely on the concept of total variation. While these algorithms perform well for bi-partitioning tasks, their recursive extensions yield unimpressive results for multiclass clustering tasks. This paper presents a general framework for multiclass total variation clustering that does not rely on recursion. The results greatly outperform previous total variation algorithms and compare well with state-of-the-art NMF approaches.

S29 Learning Multiple Models via Regularized Weighting

Daniel Vainsencher daniel.vainsencher@gmail.com
Shie Mannor shie@ee.technion.ac.il
Technion
Huan Xu mpexuh@nus.edu.sg
NUS

We consider the general problem of Multiple Model Learning (MML) from data, from the statistical and algorithmic perspectives; this problem includes clustering, multiple regression and subspace clustering as special cases. A common approach to solving new MML problems is to generalize Lloyd's algorithm for clustering (or Expectation-Maximization for soft clustering). However this approach is unfortunately sensitive to outliers and large noise: a single exceptional point may take over one of the models. We propose a different general formulation that seeks for each model a distribution over data points; the weights are regularized to be sufficiently spread out. This enhances robustness by making assumptions on class balance. We further provide generalization bounds and explain how the new iterations may be computed efficiently. We demonstrate the robustness benefits of our approach with some experimental results and prove for the important case of clustering that our approach has a non-trivial breakdown point, i.e., is guaranteed to be robust to a fixed percentage of adversarial unbounded outliers.

S30 Using multiple samples to learn mixture models

Jason Lee jdl17@stanford.edu
Stanford University
Ran Gilad-Bachrach rang@microsoft.com
Rich Caruana rcaruana@microsoft.com
Microsoft Research

In the mixture models problem it is assumed that there are K distributions $\theta 1,\ldots,\theta K$ and one gets to observe a sample from a mixture of these distributions with unknown coefficients. The goal is to associate instances with their generating distributions, or to identify the parameters of the hidden distributions. In this work we make the assumption that we have access to several samples drawn from the same K underlying distributions, but with different mixing weights. As with topic modeling, having multiple samples is often a reasonable assumption. Instead of pooling the data into one sample, we prove that it is possible to use the differences between the samples to better recover the underlying structure. We present algorithms that recover the underlying structure under milder assumptions than the current state of art when either the dimensionality or the separation is high. The methods, when applied to topic modeling, allow generalization to words not present in the training data.

S31 Regularized Spectral Clustering under the Degree-Corrected Stochastic Blockmodel

Tai Qin tqin3@wisc.edu
Karl Rohe karlrohe@stat.wisc.edu
UW-Madison

Spectral clustering is a fast and popular algorithm for finding clusters in networks. Recently, Chaudhuri et al. and Amini et al. proposed variations on the algorithm that artificially inflate the node degrees for improved statistical performance. The current paper extends the previous theoretical results to the more canonical spectral clustering algorithm in a way that removes any assumption on the minimum degree and provides guidance on the choice of tuning parameter. Moreover, our results show how the "star shape" in the eigenvectors--which are consistently observed in empirical networks--can be explained by the Degree-Corrected Stochastic Blockmodel and the Extended Planted Partition model, two statistical model that allow for highly heterogeneous degrees. Throughout, the paper characterizes and justifies several of the variations of the spectral clustering algorithm in terms of these models.

S32 Moment-based Uniform Deviation Bounds for k-means and Friends

Matus Telgarsky mtelgars@cs.ucsd.edu
Sanjoy Dasgupta dasgupta@eng.ucsd.edu
UC San Diego

Suppose k centers are fit to m points by heuristically minimizing the k-means cost; what is the corresponding fit over the source distribution? This question is resolved here for distributions with $p\geq 4$ bounded moments; in particular, the difference between the sample cost and distribution cost decays with m and p as $m^{\min\{-1/4,-1/2+2/p\}}$. The essential technical contribution is a mechanism to uniformly control deviations in the face of unbounded parameter sets, cost functions, and source distributions. To further demonstrate this mechanism, a soft clustering variant of k-means cost is also considered, namely the log likelihood of a Gaussian mixture, subject to the constraint that all covariance matrices have bounded spectrum. Lastly, a rate with refined constants is provided for k-means instances possessing some cluster structure.

S33 Statistical Active Learning Algorithms

Maria-Florina Balcan ninamf@cc.gatech.edu
Georgia Tech
Vitaly Feldman vitaly.edu@gmail.com
IBM Research

We describe a framework for designing efficient active learning algorithms that are tolerant to random classification noise. The framework is based on active learning algorithms that are statistical in the sense that they rely on estimates of expectations of functions of filtered random examples. It builds on the powerful statistical query framework of Kearns (1993). We show that any efficient active statistical learning algorithm can be automatically converted to an efficient active learning algorithm which is tolerant to random classification noise as well as other forms of "uncorrelated" noise. The complexity of the resulting algorithms has information-theoretically optimal quadratic dependence on $1/(1-2\eta)$, where η is the noise rate. We demonstrate the power of our framework by showing that commonly studied concept classes including thresholds, rectangles, and linear separators can be efficiently actively learned in our framework. These results combined with our generic conversion lead to the first known computationally-efficient algorithms for actively learning some of these concept classes in the presence of random classification noise that provide exponential improvement in the dependence on the error ϵ over their passive counterparts. In addition, we show that our algorithms can be automatically converted to efficient active differentially-private algorithms. This leads to the first differentially-private active learning algorithms with exponential label savings over the passive case.

S34 More data speeds up training time in learning halfspaces over sparse vectors

Amit Daniely amit.daniely@mail.huji.ac.il
Hebrew University
Nati Linial nati@cs.huji.ac.il
Shai Shalev-Shwartz shai.shwartz@gmail.com
The Hebrew University

The increased availability of data in recent years led several authors to ask whether it is possible to use data as a *computational* resource. That is, if more data is available, beyond the sample complexity limit, is it possible to use the extra examples to speed up the computation time required to perform the learning task? We give the first positive answer to this question for a natural supervised learning problem. We consider agnostic PAC learning of halfspaces over 3-sparse vectors in $\{-1{,}1{,}0\}n$. This class is inefficiently learnable using $O\left(n/\epsilon_2\right)$ examples. Our main contribution is a novel, non-cryptographic, methodology for establishing computational-statistical gaps, which allows us to show that, under a widely believed assumption that refuting random $3\mathrm{CNF}$ formulas is hard, efficiently learning this class using $\mathrm{O\backslash left(n/\epsilon\hat{\ }2\backslash right)}$ examples is impossible.We further show that under stronger hardness assumptions, even $\mathrm{O[(n\hat{\ }\{1.499\}/\epsilon\hat{\ }2\backslash]}$ examples do not suffice. On the other hand, we show a new algorithm that learns this class efficiently using $\Omega\mathrm{[(n\hat{\ }2/\epsilon\hat{\ }2]}$examples. This formally establishes the trade off between sample and computational complexity for a natural supervised learning problem

S35 Convex Calibrated Surrogates for Low-Rank Loss Matrices with Applications to Subset Ranking Losses

Harish Ramaswamy harish_gurup@csa.iisc.ernet.in
Shivani Agarwal shivani@csa.iisc.ernet.in
Indian Institute of Science
Ambuj Tewari tewaria@umich.edu
University of Michigan

The design of convex, calibrated surrogate losses, whose minimization entails consistency with respect to a desired target loss, is an important concept to have emerged in the theory of machine learning in recent years. We give an explicit construction of a convex least-squares type surrogate loss that can be designed to be calibrated for any multiclass learning problem for which the target loss matrix has a low-rank structure; the surrogate loss operates on a surrogate target space of dimension at most the rank of the target loss. We use this result to design convex calibrated surrogates for a variety of subset ranking problems, with target losses including the precision@q, expected rank utility, mean average precision, and pairwise disagreement.

S36 On the Relationship Between Binary Classification, Bipartite Ranking, and Binary Class Probability Estimation

Harikrishna Narasimhan nhari88@gmail.com
Shivani Agarwal shivani@csa.iisc.ernet.in
Indian Institute of Science

We investigate the relationship between three fundamental problems in machine learning: binary classification, bipartite ranking, and binary class probability estimation (CPE). It is known that a good binary CPE model can be used to obtain a good binary classification model (by thresholding at 0.5), and also to obtain a good bipartite ranking model (by using the CPE model directly as a ranking model); it is also known that a binary classification model does not necessarily yield a CPE model. However, not much is known about other directions. Formally, these relationships involve regret transfer bounds. In this paper, we introduce the notion of weak regret transfer bounds, where the mapping needed to transform a model from one problem to another depends on the underlying probability distribution (and in practice, must be estimated from data). We then show that, in this weaker sense, a good bipartite ranking model can be used to construct a good classification model (by thresholding at a suitable point), and more surprisingly, also to construct a good binary CPE model (by calibrating the scores of the ranking model).

S37 Predictive PAC Learning and Process Decompositions

Cosma Shalizi cshalizi@cmu.edu
CMU
Aryeh Kontorovitch karyeh@cs.bgu.ac.il
Ben Gurion University

We informally call a stochastic process learnable if it admits a generalization error approaching zero in probability for any concept class with finite VC-dimension (IID processes are the simplest example). A mixture of learnable processes need not be learnable itself, and certainly its generalization error need not decay at the same rate. In this paper, we argue that it is natural in predictive PAC to condition not on the past observations but on the mixture component of the sample path. This definition not only matches what a realistic learner might demand, but also allows us to sidestep several otherwise grave problems in learning from dependent data. In particular, we give a novel PAC generalization bound for mixtures of learnable processes with a generalization error that is not worse than that of each mixture component. We also provide a characterization of mixtures of absolutely regular (β-mixing) processes, of independent interest.

S38 Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima

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

We establish theoretical results concerning all local optima of various regularized M-estimators, where both loss and penalty functions are allowed to be nonconvex. Our results show that as long as the loss function satisfies restricted strong convexity and the penalty function satisfies suitable regularity conditions, any local optimum of the composite objective function lies within statistical precision of the true parameter vector. Our theory covers a broad class of nonconvex objective functions, including corrected versions of the Lasso for errors-in-variables linear models; regression in generalized linear models using nonconvex regularizers such as SCAD and MCP; and graph and inverse covariance matrix estimation. On the optimization side, we show that a simple adaptation of composite gradient descent may be used to compute a global optimum up to the statistical precision epsilon in log(1/epsilon) iterations, which is the fastest possible rate of any first-order method. We provide a variety of simulations to illustrate the sharpness of our theoretical predictions.

S39 PAC-Bayes-Empirical-Bernstein Inequality

Ilya Tolstikhin iliya.tolstikhin@gmail.com
Russian Academy of Sciences
Yevgeny Seldin yevgeny.seldin@gmail.com
Queensland University of Technology & UC Berkeley

We present PAC-Bayes-Empirical-Bernstein inequality. The inequality is based on combination of PAC-Bayesian bounding technique with Empirical Bernstein bound. It allows to take advantage of small empirical variance and is especially useful in regression. We show that when the empirical variance is significantly smaller than the empirical loss PAC-Bayes-Empirical-Bernstein inequality is significantly tighter than PAC-Bayes-kl inequality of Seeger (2002) and otherwise it is comparable. PAC-Bayes-Empirical-Bernstein inequality is an interesting example of application of PAC-Bayesian bounding technique to self-bounding functions. We provide empirical comparison of PAC-Bayes-Empirical-Bernstein inequality with PAC-Bayes-kl inequality on a synthetic example and several UCI datasets.

S40 Regression-tree Tuning in a Streaming Setting

Samory Kpotufe samory@ttic.edu
Francesco Orabona orabona@ttic.edu
TTI Chicago

We consider the problem of maintaining the data-structures of a partition-based regression procedure in a setting where the training data arrives sequentially over time. We prove that it is possible to maintain such a structure in time $O(\log\ n)$ at any time step n while achieving a nearly-optimal regression rate of õ$\left(^{n-2/(2+d)}\right)$ in terms of the unknown metric dimension d. Finally we prove a new regression lower-bound which is independent of a given data size, and hence is more appropriate for the streaming setting.

S41 Adaptivity to Local Smoothness and Dimension in Kernel Regression

Samory Kpotufe samory@ttic.edu
Vikas Garg vkg@ttic.edu
TTI Chicago

We present the first result for kernel regression where the procedure adapts locally at a point $x$to both the unknown local dimension of the metric and the unknown H\"{o}lder-continuity of the regression function at x. The result holds with high probability simultaneously at all points x in a metric space of unknown structure.

S42 A Comparative Framework for Preconditioned Lasso Algorithms

Fabian Wauthier flw@berkeley.edu
Michael Jordan jordan@cs.berkeley.edu
UC Berkeley
Nebojsa Jojic jojic@microsoft.com
Microsoft Research

The Lasso is a cornerstone of modern multivariate data analysis, yet its performance suffers in the common situation in which covariates are correlated. This limitation has led to a growing number of \emph{Preconditioned Lasso} algorithms that pre-multiply X and y by matrices P_X, P_y prior to running the standard Lasso. A direct comparison of these and similar Lasso-style algorithms to the original Lasso is difficult because the performance of all of these methods depends critically on an auxiliary penalty parameter λ. In this paper we propose an agnostic, theoretical framework for comparing Preconditioned Lasso algorithms to the Lasso without having to choose λ. We apply our framework to three Preconditioned Lasso instances and highlight when they will outperform the Lasso. Additionally, our theory offers insights into the fragilities of these algorithms to which we provide partial solutions.

S43 New Subsampling Algorithms for Fast Least Squares Regression

Paramveer Dhillon dhillon@cis.upenn.edu
Yichao Lu luyichao1123@gmail.com
Dean Foster foster@wharton.upenn.edu
Lyle Ungar ungar@cis.upenn.edu
University of Pennsylvania

We address the problem of fast estimation of ordinary least squares (OLS) from large amounts of data ($n \gg p$). We propose three methods which solve the big data problem by subsampling the covariance matrix using either a single or two stage estimation. All three run in the order of size of input i.e. O(np) and our best method, {\it Uluru}, gives an error bound of $O(\sqrt{p/n})$ which is independent of the amount of subsampling as long as it is above a threshold. We provide theoretical bounds for our algorithms in the fixed design (with Randomized Hadamard preconditioning) as well as sub-Gaussian random design setting. We also compare the performance of our methods on synthetic and real-world datasets and show that if observations are i.i.d., sub-Gaussian then one can directly subsample without the expensive Randomized Hadamard preconditioning without loss of accuracy.

S44 Faster Ridge Regression via the Subsampled Randomized Hadamard Transform

Yichao Lu luyichao1123@gmail.com
Paramveer Dhillon dhillon@cis.upenn.edu
Dean Foster foster@wharton.upenn.edu
Lyle Ungar ungar@cis.upenn.edu
University of Pennsylvania

We propose a fast algorithm for ridge regression when the number of features is much larger than the number of observations ($p \gg n$). The standard way to solve ridge regression in this setting works in the dual space and gives a running time of $O(n^2p)$. Our algorithm (SRHT-DRR) runs in time $O(np\ \log(n))$ and works by preconditioning the design matrix by a Randomized Walsh-Hadamard Transform with a subsequent subsampling of features. We provide risk bounds for our SRHT-DRR algorithm in the fixed design setting and show experimental results on synthetic and real datasets.

S45 Submodular Optimization with Submodular Cover and Submodular Knapsack Constraints

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

We investigate two new optimization problems — minimizing a submodular function subject to a submodular lower bound constraint (submodular cover) and maximizing a submodular function subject to a submodular upper bound constraint (submodular knapsack). We are motivated by a number of real-world applications in machine learning including sensor placement and data subset selection, which require maximizing a certain submodular function (like coverage or diversity) while simultaneously minimizing another (like cooperative cost). These problems are often posed as minimizing the difference between submodular functions [9, 23] which is in the worst case inapproximable. We show, however, that by phrasing these problems as constrained optimization, which is more natural for many applications, we achieve a number of bounded approximation guarantees. We also show that both these problems are closely related and, an approximation algorithm solving one can be used to obtain an approximation guarantee for the other. We provide hardness results for both problems thus showing that our approximation factors are tight up to log-factors. Finally, we empirically demonstrate the performance and good scalability properties of our algorithms.

S46 Sparse Overlapping Sets Lasso for Multitask Learning and its Application to fMRI Analysis

Nikhil Rao nrao2@wisc.edu
Christopher Cox crcox@wisc.edu
Rob Nowak nowak@ece.wisc.edu
Timothy Rogers ttrogers@wisc.edu
UW-Madison

Multitask learning can be effective when features useful in one task are also useful for other tasks, and the group lasso is a standard method for selecting a common subset of features. In this paper, we are interested in a less restrictive form of multitask learning, wherein (1) the available features can be organized into subsets according to a notion of similarity and (2) features useful in one task are similar, but not necessarily identical, to the features best suited for other tasks. The main contribution of this paper is a new procedure called *Sparse Overlapping Sets (SOS) lasso*, a convex optimization that automatically selects similar features for related learning tasks. Error bounds are derived for SOSlasso and its consistency is established for squared error loss. In particular, SOSlasso is motivated by multi-subject fMRI studies in which functional activity is classified using brain voxels as features. Experiments with real and synthetic data demonstrate the advantages of SOSlasso compared to the lasso and group lasso.

S47 Sequential Transfer in Multi-armed Bandit with Finite Set of Models

Mohammad Gheshlaghi azar mazar@cs.cmu.edu
Emma Brunskill ebrun@cs.cmu.edu
CMU
Alessandro Lazaric alessandro.lazaric@gmail.com
INRIA

Learning from prior tasks and transferring that experience to improve future performance is critical for building lifelong learning agents. Although results in supervised and reinforcement learning show that transfer may significantly improve the learning performance, most of the literature on transfer is focused on batch learning tasks. In this paper we study the problem of sequential transfer in online learning, notably in the multi-arm bandit framework, where the objective is to minimize the cumulative regret over a sequence of tasks by incrementally transferring knowledge from prior tasks. We introduce a novel bandit algorithm based on a method-of-moments approach for the estimation of the possible tasks and derive regret bounds for it.

S48 Eluder Dimension and the Sample Complexity of Optimistic Exploration

Dan Russo dan.joseph.russo@gmail.com
Benjamin Van Roy bvr@stanford.edu
Stanford University

This paper considers the sample complexity of the multi-armed bandit with dependencies among the arms. Some of the most successful algorithms for this problem use the principle of optimism in the face of uncertainty to guide exploration. The clearest example of this is the class of upper confidence bound (UCB) algorithms, but recent work has shown that a simple posterior sampling algorithm, sometimes called Thompson sampling, also shares a close theoretical connection with optimistic approaches. In this paper, we develop a regret bound that holds for both classes of algorithms. This bound applies broadly and can be specialized to many model classes. It depends on a new notion we refer to as the eluder dimension, which measures the degree of dependence among action rewards. Compared to UCB algorithm regret bounds for specific model classes, our general bound matches the best available for linear models and is stronger than the best available for generalized linear models.

S49 Prior-free and prior-dependent regret bounds for Thompson Sampling

Sebastien Bubeck sbubeck@princeton.edu
Che-Yu Liu cheliu@princeton.edu
Princeton University

We consider the stochastic multi-armed bandit problem with a prior distribution on the reward distributions. We are interested in studying prior-free and prior-dependent regret bounds, very much in the same spirit than the usual distribution-free and distribution-dependent bounds for the non-Bayesian stochastic bandit. We first show that Thompson Sampling attains an optimal prior-free bound in the sense that for any prior distribution its Bayesian regret is bounded from above by $14\sqrt{nK}$. This result is unimprovable in the sense that there exists a prior distribution such that any algorithm has a Bayesian regret bounded from below by $1/20\sqrt{nK}$. We also study the case of priors for the setting of Bubeck et al. [2013] (where the optimal mean is known as well as a lower bound on the smallest gap) and we show that in this case the regret of Thompson Sampling is in fact uniformly bounded over time, thus showing that Thompson Sampling can greatly take advantage of the nice properties of these priors.

S50 From Bandits to Experts: A Tale of Domination and Independence

Noga Alon nogaa@tau.ac.il
Yishay Mansour mansour.yishay@gmail.com
Tel Aviv University
Nicolò Cesa-Bianchi nicolo.cesa-bianchi@unimi.it
University of Milan
Claudio Gentile claudio.gentile@uninsubria.it
University of Insubria

We consider the partial observability model for multi-armed bandits, introduced by Mannor and Shamir (2011). Our main result is a characterization of regret in the directed observability model in terms of the dominating and independence numbers of the observability graph. We also show that in the undirected case, the learner can achieve optimal regret without even accessing the observability graph before selecting an action. Both results are shown using variants of the Exp3 algorithm operating on the observability graph in a time-efficient manner.

S51 Two-Target Algorithms for Infinite-Armed Bandits with Bernoulli Rewards

Thomas Bonald thomas.bonald@telecom-paristech.fr
Telecom ParisTech
Alexandre Proutiere alepro@kth.se
KTH

We consider an infinite-armed bandit problem with Bernoulli rewards. The mean rewards are independent, uniformly distributed over $[0,1]$. Rewards 0 and 1 are referred to as a success and a failure, respectively. We propose a novel algorithm where the decision to exploit any arm is based on two successive targets, namely, the total number of successes until the first failure and the first m failures, respectively, where m is a fixed parameter. This two-target algorithm achieves a long-term average regret in $\sqrt{2n}$ for a large parameter m and a known time horizon n. This regret is optimal and strictly less than the regret achieved by the best known algorithms, which is in $2\sqrt{n}$. The results are extended to any mean-reward distribution whose support contains 1 and to unknown time horizons. Numerical experiments show the performance of the algorithm for finite time horizons.

S52 Thompson Sampling for 1-Dimensional Exponential Family Bandits

Nathaniel Korda nathaniel.korda@inria.fr
Remi Munos remi.munos@inria.fr
INRIA
Emilie Kaufmann kaufmann@telecom-paristech.fr
Telecom ParisTech

Thompson Sampling has been demonstrated in many complex bandit models, however the theoretical guarantees available for the parametric multi-armed bandit are still limited to the Bernoulli case. Here we extend them by proving asymptotic optimality of the algorithm using the Jeffreys prior for 1-dimensional exponential family bandits. Our proof builds on previous work, but also makes extensive use of closed forms for Kullback-Leibler divergence and Fisher information (and thus Jeffreys prior) available in an exponential family. This allow us to give a finite time exponential concentration inequality for posterior distributions on exponential families that may be of interest in its own right. Moreover our analysis covers some distributions for which no optimistic algorithm has yet been proposed, including heavy-tailed exponential families.

S53 Bayesian Mixture Modelling and Inference based Thompson Sampling in Monte-Carlo Tree Search

Aijun Bai baj@mail.ustc.edu.cn
Xiaoping Chen xpchen@ustc.edu.cn
University of Science and Technology of China (USTC)
Feng Wu fw6e11@ecs.soton.ac.uk
University of Southampton

Monte-Carlo tree search is drawing great interest in the domain of planning under uncertainty, particularly when little or no domain knowledge is available. One of the central problems is the trade-off between exploration and exploitation. In this paper we present a novel Bayesian mixture modelling and inference based Thompson sampling approach to addressing this dilemma. The proposed Dirichlet-NormalGamma MCTS (DNG-MCTS) algorithm represents the uncertainty of the accumulated reward for actions in the MCTS search tree as a mixture of Normal distributions and inferences on it in Bayesian settings by choosing conjugate priors in the form of combinations of Dirichlet and NormalGamma distributions. Thompson sampling is used to select the best action at each decision node. Experimental results show that our proposed algorithm has achieved the state-of-the-art comparing with popular UCT algorithm in the context of online planning for general Markov decision processes.

S54 Approximate Inference in Continuous Determinantal Processes

Raja Hafiz Affandi rajara@wharton.upenn.edu
University of Pennsylvania
Emily Fox ebfox@uw.edu
Ben Taskar taskar@cs.washington.edu
University of Washington

Determinantal point processes (DPPs) are random point processes well-suited for modeling repulsion. In machine learning, the focus of DPP-based models has been on diverse subset selection from a discrete and finite base set. This discrete setting admits an efficient algorithm for sampling based on the eigendecomposition of the defining kernel matrix. Recently, there has been growing interest in using DPPs defined on continuous spaces. While the discrete-DPP sampler extends formally to the continuous case, computationally, the steps required cannot be directly extended except in a few restricted cases. In this paper, we present efficient approximate DPP sampling schemes based on Nystrom and random Fourier feature approximations that apply to a wide range of kernel functions. We demonstrate the utility of continuous DPPs in repulsive mixture modeling applications and synthesizing human poses spanning activity spaces.

S55 Inverse Density as an Inverse Problem: the Fredholm Equation Approach

Qichao Que que@cse.ohio-state.edu
Mikhail Belkin mbelkin@cse.ohio-state.edu
Ohio State University

We address the problem of estimating the ratio qp where p is a density function and q is another density, or, more generally an arbitrary function. Knowing or approximating this ratio is needed in various problems of inference and integration, in particular, when one needs to average a function with respect to one probability distribution, given a sample from another. It is often referred as *importance sampling* in statistical inference and is also closely related to the problem of covariate shift in transfer learning as well as to various MCMC methods. Our approach is based on reformulating the problem of estimating the ratio as an inverse problem in terms of an integral operator corresponding to a kernel, and thus reducing it to an integral equation, known as the Fredholm problem of the first kind. This formulation, combined with the techniques of regularization and kernel methods, leads to a principled kernel-based framework for constructing algorithms and for analyzing them theoretically. The resulting family of algorithms (FIRE, for Fredholm Inverse Regularized Estimator) is flexible, simple and easy to implement. We provide detailed theoretical analysis including concentration bounds and convergence rates for the Gaussian kernel for densities defined on $\backslash \mathrm{R}^{d}$ and smooth d-dimensional sub-manifolds of the Euclidean space. Model selection for unsupervised or semi-supervised inference is generally a difficult problem. Interestingly, it turns out that in the density ratio estimation setting, when samples from both distributions are available, there are simple completely unsupervised methods for choosing parameters. We call this model selection mechanism CD-CV for Cross-Density Cross-Validation. Finally, we show encouraging experimental results including applications to classification within the covariate shift framework.

S56 Density estimation from unweighted k-nearest neighbor graphs: a roadmap

Ulrike Von Luxburg
luxburg@informatik.uni-hamburg.de
Morteza Alamgir
alamgir@informatik.uni-hamburg.de
University of Hamburg

Consider an unweighted k-nearest neighbor graph on n points that have been sampled i.i.d. from some unknown density p on R^d. We prove how one can estimate the density p just from the unweighted adjacency matrix of the graph, without knowing the points themselves or their distance or similarity scores. The key insights are that local differences in link numbers can be used to estimate some local function of p, and that integrating this function along shortest paths leads to an estimate of the underlying density.

S57 Sketching Structured Matrices for Faster Nonlinear Regression

Haim Avron haim.avron@gmail.com
Vikas Sindhwani vikas.sindhwani@gmail.com
David Woodruff dpwoodru@us.ibm.com
IBM Research

Motivated by the desire to extend fast randomized techniques to nonlinear lp regression, we consider a class of structured regression problems. These problems involve Vandermonde matrices which arise naturally in various statistical modeling settings, including classical polynomial fitting problems and recently developed randomized techniques for scalable kernel methods. We show that this structure can be exploited to further accelerate the solution of the regression problem, achieving running times that are faster than "input sparsity". We present empirical results confirming both the practical value of our modeling framework, as well as speedup benefits of randomized regression.

S58 More Effective Distributed ML via a Stale Synchronous Parallel Parameter Server

Qirong Ho qho@cs.cmu.edu
James Cipar jcipar@cmu.edu
Henggang Cui hengganc@ece.cmu.edu
Seunghak Lee seunghak@cs.cmu.edu
Jin Kyu Kim jinkyuk@cs.cmu.edu
Garth Gibson garth@cs.cmu.edu
Greg Ganger ganger@ece.cmu.edu
Eric Xing epxing@cs.cmu.edu
CMU
Phil Gibbons phillip.b.gibbons@intel.com
Intel Labs

We propose a parameter server system for distributed ML, which follows a Stale Synchronous Parallel (SSP) model of computation that maximizes the time computational workers spend doing useful work on ML algorithms, while still providing correctness guarantees. The parameter server provides an easy-to-use shared interface for read/write access to an ML model's values (parameters and variables), and the SSP model allows distributed workers to read older, stale versions of these values from a local cache, instead of waiting to get them from a central storage. This significantly increases the proportion of time workers spend computing, as opposed to waiting. Furthermore, the SSP model ensures ML algorithm correctness by limiting the maximum age of the stale values. We provide a proof of correctness under SSP, as well as empirical results demonstrating that the SSP model achieves faster algorithm convergence on several different ML problems, compared to fully-synchronous and asynchronous schemes.

S59 Non-strongly-convex smooth stochastic approximation with convergence rate O(1/n)

Francis Bach francis.bach@mines.org
INRIA & ENS
Eric Moulines eric.moulines@telecom-paristech.fr
Telecom ParisTech

We consider the stochastic approximation problem where a convex function has to be minimized, given only the knowledge of unbiased estimates of its gradients at certain points, a framework which includes machine learning methods based on the minimization of the empirical risk. We focus on problems without strong convexity, for which all previously known algorithms achieve a convergence rate for function values of $O(1/\sqrt{n})$. We consider and analyze two algorithms that achieve a rate of $O(1/n)$ for classical supervised learning problems. For least-squares regression, we show that averaged stochastic gradient descent with constant step-size achieves the desired rate. For logistic regression, this is achieved by a simple novel stochastic gradient algorithm that (a) constructs successive local quadratic approximations of the loss functions, while (b) preserving the same running time complexity as stochastic gradient descent. For these algorithms, we provide a non-asymptotic analysis of the generalization error (in expectation, and also in high probability for least-squares), and run extensive experiments showing that they often outperform existing approaches.

S60 Trading Computation for Communication: Distributed Stochastic Dual Coordinate Ascent

Tianbao Yang yangtia1@msu.edu
NEC Labs America

We present and study a distributed optimization algorithm by employing a stochastic dual coordinate ascent method. Stochastic dual coordinate ascent methods enjoy strong theoretical guarantees and often have better performances than stochastic gradient descent methods in optimizing regularized loss minimization problems. It still lacks of efforts in studying them in a distributed framework. We make a progress along the line by presenting a distributed stochastic dual coordinate ascent algorithm in a star network, with an analysis of the tradeoff between computation and communication. We verify our analysis by experiments on real data sets. Moreover, we compare the proposed algorithm with distributed stochastic gradient descent methods and distributed alternating direction methods of multipliers for optimizing SVMs in the same distributed framework, and observe competitive performances.

S61 Locally Adaptive Bayesian Multivariate Time Series

Daniele Durante durante@stat.unipd.it
Bruno Scarpa scarpa@stat.unipd.it
University of Padua
David Dunson dunson@stat.duke.edu
Duke University

In modeling multivariate time series, it is important to allow time-varying smoothness in the mean and covariance process. In particular, there may be certain time intervals exhibiting rapid changes and others in which changes are slow. If such locally adaptive smoothness is not accounted for, one can obtain misleading inferences and predictions, with over-smoothing across erratic time intervals and under-smoothing across times exhibiting slow variation. This can lead to miscalibration of predictive intervals, which can be substantially too narrow or wide depending on the time. We propose a continuous multivariate stochastic process for time series having locally varying smoothness in both the mean and covariance matrix. This process is constructed utilizing latent dictionary functions in time, which are given nested Gaussian process priors and linearly related to the observed data through a sparse mapping. Using a differential equation representation, we bypass usual computational bottlenecks in obtaining MCMC and online algorithms for approximate Bayesian inference. The performance is assessed in simulations and illustrated in a financial application.

S62 Small-Variance Asymptotics for Hidden Markov Models

Anirban Roychowdhury roychowdhury.7@osu.edu
Ke Jiang jiangk@cse.ohio-state.edu
Brian Kulis brian.kulis@gmail.com
Ohio State University

Small-variance asymptotics provide an emerging technique for obtaining scalable combinatorial algorithms from rich probabilistic models. We present a small-variance asymptotic analysis of the Hidden Markov Model and its infinite-state Bayesian nonparametric extension. Starting with the standard HMM, we first derive a “hard” inference algorithm analogous to k-means that arises when particular variances in the model tend to zero. This analysis is then extended to the Bayesian nonparametric case, yielding a simple, scalable, and flexible algorithm for discrete-state sequence data with a non-fixed number of states. We also derive the corresponding combinatorial objective functions arising from our analysis, which involve a k-means-like term along with penalties based on state transitions and the number of states. A key property of such algorithms is that — particularly in the nonparametric setting — standard probabilistic inference algorithms lack scalability and are heavily dependent on good initialization. A number of results on synthetic and real data sets demonstrate the advantages of the proposed framework.

S63 A Latent Source Model for Nonparametric Time Series Classification

George Chen georgehc@mit.edu
Devavrat Shah devavrat@mit.edu
Massachusetts Institute of Technology
Stan Nikolov snikolov@twitter.com
Twitter

For classifying time series, a nearest-neighbor approach is widely used in practice with performance often competitive with or better than more elaborate methods such as neural networks, decision trees, and support vector machines. We develop theoretical justification for the effectiveness of nearest-neighbor-like classification of time series. Our guiding hypothesis is that in many applications, such as forecasting which topics will become trends on Twitter, there aren’t actually that many prototypical time series to begin with, relative to the number of time series we have access to, e.g., topics become trends on Twitter only in a few distinct manners whereas we can collect massive amounts of Twitter data. To operationalize this hypothesis, we propose a latent source model for time series, which naturally leads to a “weighted majority voting” classification rule that can be approximated by a nearest-neighbor classifier. We establish nonasymptotic performance guarantees of both weighted majority voting and nearest-neighbor classification under our model accounting for how much of the time series we observe and the model complexity. Experimental results on synthetic data show weighted majority voting achieving the same misclassification rate as nearest-neighbor classification while observing less of the time series. We then use weighted majority to forecast which news topics on Twitter become trends, where we are able to detect such “trending topics” in advance of Twitter 79% of the time, with a mean early advantage of 1 hour and 26 minutes, a true positive rate of 95%, and a false positive rate of 4%.

S64 Multilinear Dynamical Systems for Tensor Time Series

Mark Rogers markrogersjr@gmail.com
Lei Li leili@cs.berkeley.edu
Stuart Russell russell@cs.berkeley.edu
UC Berkeley

Many scientific data occur as sequences of multidimensional arrays called tensors. How can hidden, evolving trends in such data be extracted while preserving the tensor structure? The model that is traditionally used is the linear dynamical system (LDS), which treats the observation at each time slice as a vector. In this paper, we propose the multilinear dynamical system (MLDS) for modeling tensor time series and an expectation-maximization (EM) algorithm to estimate the parameters. The MLDS models each time slice of the tensor time series as the multilinear projection of a corresponding member of a sequence of latent, low-dimensional tensors. Compared to the LDS with an equal number of parameters, the MLDS achieves higher prediction accuracy and marginal likelihood for both simulated and real datasets.

S65 What do row and column marginals reveal about your dataset?

Behzad Golshan behzad@cs.bu.edu
John Byers byers@cs.bu.edu
Evimaria Terzi evimaria@cs.bu.edu
Boston University

Numerous datasets ranging from group memberships within social networks to purchase histories on e-commerce sites are represented by binary matrices. While this data is often either proprietary or sensitive, aggregated data, notably row and column marginals, is often viewed as much less sensitive, and may be furnished for analysis. Here, we investigate how these data can be exploited to make inferences about the underlying matrix H. Instead of assuming a generative model for H, we view the input marginals as constraints on the dataspace of possible realizations of H and compute the probability density function of particular entries H(i,j) of interest. We do this, for all the cells of H simultaneously, without generating realizations but rather via implicitly sampling the datasets that satisfy the input marginals. The end result is an efficient algorithm with running time equal to the time required by standard sampling techniques to generate a single dataset from the same dataspace. Our experimental evaluation demonstrates the efficiency and the efficacy of our framework in multiple settings.

S66 Error-Minimizing Estimates and Universal Entry-Wise Error Bounds for Low-Rank Matrix Completion

Franz Kiraly f.kiraly@ucl.ac.uk
TU Berlin
Louis Theran theran@math.fu-berlin.de
Freie Universität Berlin

We propose a general framework for reconstructing and denoising single entries of incomplete and noisy entries. We describe: effective algorithms for deciding if and entry can be reconstructed and, if so, for reconstructing and denoising it; and a priori bounds on the error of each entry, individually. In the noiseless case our algorithm is exact. For rank-one matrices, the new algorithm is fast, admits a highly-parallel implementation, and produces an error minimizing estimate that is qualitatively close to our theoretical and the state-of-the-art Nuclear Norm and OptSpace methods.

S67 Synthesizing Robust Plans under Incomplete Domain Models

Tuan Nguyen natuan@asu.edu
Subbarao Kambhampati rao@asu.edu
Arizona State University
Minh Do minh.do@nasa.gov
NASA

Most current planners assume complete domain models and focus on generating correct plans. Unfortunately, domain modeling is a laborious and error-prone task, thus real world agents have to plan with incomplete domain models. While domain experts cannot guarantee completeness, often they are able to circumscribe the incompleteness of the model by providing annotations as to which parts of the domain model may be incomplete. In such cases, the goal should be to synthesize plans that are robust with respect to any known incompleteness of the domain. In this paper, we first introduce annotations expressing the knowledge of the domain incompleteness and formalize the notion of plan robustness with respect to an incomplete domain model. We then show an approach to compiling the problem of finding robust plans to the conformant probabilistic planning problem, and present experimental results with Probabilistic-FF planner.

S68 Message Passing Inference with Chemical Reaction Networks

Nils Napp nnapp@wyss.harvard.edu
Ryan Adams rpa@seas.harvard.edu
Harvard University

Recent work on molecular programming has explored new possibilities for computational abstractions with biomolecules, including logic gates, neural networks, and linear systems. In the future such abstractions might enable nanoscale devices that can sense and control the world at a molecular scale. Just as in macroscale robotics, it is critical that such devices can learn about their environment and reason under uncertainty. At this small scale, systems are typically modeled as chemical reaction networks. In this work, we develop a procedure that can take arbitrary probabilistic graphical models, represented as factor graphs over discrete random variables, and compile them into chemical reaction networks that implement inference. In particular, we show that marginalization based on sum-product message passing can be implemented in terms of reactions between chemical species whose concentrations represent probabilities. We show algebraically that the steady state concentration of these species correspond to the marginal distributions of the random variables in the graph and validate the results in simulations. As with standard sum-product inference, this procedure yields exact results for tree-structured graphs, and approximate solutions for loopy graphs.

S69 Which Space Partitioning Tree to Use for Search?

Pari Ram p.ram@gatech.edu
Alexander Gray agray@cc.gatech.edu
Georgia Tech

We consider the task of nearest-neighbor search with the class of binary-space-partitioning trees, which includes kd-trees, principal axis trees and random projection trees, and try to rigorously answer the question “which tree to use for nearest-neighbor search?’’ To this end, we present the theoretical results which imply that trees with better vector quantization performance have better search performance guarantees. We also explore another factor affecting the search performance -- margins of the partitions in these trees. We demonstrate, both theoretically and empirically, that large margin partitions can improve the search performance of a space-partitioning tree.

S70 Solving inverse problem of Markov chain with partial observations

Tetsuro Morimura tetsuro@jp.ibm.com
Takayuki Osogami osogami@jp.ibm.com
Tsuyoshi Ide tide@us.ibm.com
IBM Research

The Markov chain is a convenient tool to represent the dynamics of complex systems such as traffic and social systems, where probabilistic transition takes place between internal states. A Markov chain is characterized by initial-state probabilities and a state-transition probability matrix. In the traditional setting, a major goal is to figure out properties of a Markov chain when those probabilities are known. This paper tackles an inverse version of the problem: we find those probabilities from partial observations at a limited number of states. The observations include the frequency of visiting a state and the rate of reaching a state from another. Practical examples of this task include traffic monitoring systems in cities, where we need to infer the traffic volume on every single link on a road network from a very limited number of observation points. We formulate this task as a regularized optimization problem for probability functions, which is efficiently solved using the notion of natural gradient. Using synthetic and real-world data sets including city traffic monitoring data, we demonstrate the effectiveness of our method.

S71 Robust Data-Driven Dynamic Programming

Grani Adiwena Hanasusanto g.hanasusanto11@imperial.ac.uk
Imperial College London
Daniel Kuhn daniel.kuhn@epfl.ch
EPFL

In stochastic optimal control the distribution of the exogenous noise is typically unknown and must be inferred from limited data before dynamic programming (DP)-based solution schemes can be applied. If the conditional expectations in the DP recursions are estimated via kernel regression, however, the historical sample paths enter the solution procedure directly as they determine the evaluation points of the cost-to-go functions. The resulting data-driven DP scheme is asymptotically consistent and admits efficient computational solution when combined with parametric value function approximations. If training data is sparse, however, the estimated cost-to-go functions display a high variability and an optimistic bias, while the corresponding control policies perform poorly in out-of-sample tests. To mitigate these small sample effects, we propose a robust data-driven DP scheme, which replaces the expectations in the DP recursions with worst-case expectations over a set of distributions close to the best estimate. We show that the arising min-max problems in the DP recursions reduce to tractable conic programs. We also demonstrate that this robust algorithm dominates state-of-the-art benchmark algorithms in out-of-sample tests across several application domains.

S72 Scoring Workers in Crowdsourcing: How Many Control Questions are Enough?

Qiang Liu qliu1@uci.edu
Alex Ihler ihler@ics.uci.edu
Mark Steyvers mark.steyvers@uci.edu
UC Irvine

We study the problem of estimating continuous quantities, such as prices, probabilities, and point spreads, using a crowdsourcing approach. A challenging aspect of combining the crowd's answers is that workers' reliabilities and biases are usually unknown and highly diverse. Control items with known answers can be used to evaluate workers' performance, and hence improve the combined results on the target items with unknown answers. This raises the problem of how many control items to use when the total number of items each workers can answer is limited: more control items evaluates the workers better, but leaves fewer resources for the target items that are of direct interest, and vice versa. We give theoretical results for this problem under different scenarios, and provide a simple rule of thumb for crowdsourcing practitioners. As a byproduct, we also provide theoretical analysis of the accuracy of different consensus methods.

S73 Online Variational Approximations to non-Exponential Family Change Point Models: With Application to Radar Tracking

Ryan Turner Ryan.Turner@ngc.com
Steven Bottone steven.bottone@ngc.com
Clay Stanek clay.stanek@ngc.com
Northrop Grumman

The Bayesian online change point detection (BOCPD) algorithm provides an efficient way to do exact inference when the parameters of an underlying model may suddenly change over time. BOCPD requires computation of the underlying model's posterior predictives, which can only be computed online in $O(1)$ time and memory for exponential family models. We develop variational approximations to the posterior on change point times (formulated as run lengths) for efficient inference when the underlying model is not in the exponential family, and does not have tractable posterior predictive distributions. In doing so, we develop improvements to online variational inference. We apply our methodology to a tracking problem using radar data with a signal-to-noise feature that is Rice distributed. We also develop a variational method for inferring the parameters of the (non-exponential family) Rice distribution.

S74 q-OCSVM: A q-Quantile Estimator for High-Dimensional Distributions

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

In this paper we introduce a novel method that can efficiently estimate a family of hierarchical dense sets in high-dimensional distributions. Our method can be regarded as a natural extension of the one-class SVM (OCSVM) algorithm that finds multiple parallel separating hyperplanes in a reproducing kernel Hilbert space. We call our method q-OCSVM, as it can be used to estimate q quantiles of a high-dimensional distribution. For this purpose, we introduce a new global convex optimization program that finds all estimated sets at once and show that it can be solved efficiently. We prove the correctness of our method and present empirical results that demonstrate its superiority over existing methods.

S75 Unsupervised Structure Learning of Stochastic And-Or Grammars

Kewei Tu tukw@ucla.edu
Maria Pavlovskaia mariapavl@gmail.com
Song-Chun Zhu sczhu@stat.ucla.edu
UCLA

Stochastic And-Or grammars compactly represent both compositionality and reconfigurability and have been used to model different types of data such as images and events. We present a unified formalization of stochastic And-Or grammars that is agnostic to the type of the data being modeled, and propose an unsupervised approach to learning the structures as well as the parameters of such grammars. Starting from a trivial initial grammar, our approach iteratively induces compositions and reconfigurations in a unified manner and optimizes the posterior probability of the grammar. In our empirical evaluation, we applied our approach to learning event grammars and image grammars and achieved comparable or better performance than previous approaches.

S76 Rapid Distance-Based Outlier Detection via Sampling

Mahito Sugiyama mahito.sugiyama@tuebingen.mpg.de
MPI Tübingen
Karsten Borgwardt karsten.borgwardt@tuebingen.mpg.de
MPI Tübingen & University of Tübingen

Distance-based approaches to outlier detection are popular in data mining, as they do not require to model the underlying probability distribution, which is particularly challenging for high-dimensional data. We present an empirical comparison of various approaches to distance-based outlier detection across a large number of datasets. We report the surprising observation that a simple, sampling-based scheme outperforms state-of-the-art techniques in terms of both efficiency and effectiveness. To better understand this phenomenon, we provide a theoretical analysis why the sampling-based approach outperforms alternative methods based on k-nearest neighbor search.

S77 One-shot learning by inverting a compositional causal process

Brenden Lake brenden@mit.edu
Josh Tenenbaum jbt@mit.edu
Massachusetts Institute of Technology
Russ Salakhutdinov rsalakhu@cs.toronto.edu
University of Toronto

People can learn a new visual class from just one example, yet machine learning algorithms typically require hundreds or thousands of examples to tackle the same problems. Here we present a Hierarchical Bayesian model based on compositionality and causality that can learn a wide range of natural (although simple) visual concepts, generalizing in human-like ways from just one image. We evaluated performance on a challenging one-shot classification task, where our model achieved a human-level error rate while substantially outperforming two deep learning models. We also used a “visual Turing test” to show that our model produces human-like performance on other conceptual tasks, including generating new examples and parsing.

S78 Linear decision rule as aspiration for simple decision heuristics

Özgür Şimşek ozgur@mpib-berlin.mpg.de
Max Planck Institute Berlin

Many attempts to understand the success of simple decision heuristics have examined heuristics as an approximation to a linear decision rule. This research has identified three environmental structures that aid heuristics: dominance, cumulative dominance, and noncompensatoriness. Here, we further develop these ideas and examine their empirical relevance in 51 natural environments. We find that all three structures are prevalent, making it possible for some simple rules to reach the accuracy levels of the linear decision rule using less information.

S79 Optimizing Instructional Policies

Robert Lindsey robert.lindsey@colorado.edu
Michael Mozer mozer@colorado.edu
William Huggins w.j.huggins@gmail.com
University of Colorado
Harold Pashler hpashler@ucsd.edu
UC San Diego

Psychologists are interested in developing instructional policies that boost student learning. An instructional policy specifies the manner and content of instruction. For example, in the domain of concept learning, a policy might specify the nature of exemplars chosen over a training sequence. Traditional psychological studies compare several hand-selected policies, e.g., contrasting a policy that selects only difficult-to-classify exemplars with a policy that gradually progresses over the training sequence from easy exemplars to more difficult (known as {\em fading}). We propose an alternative to the traditional methodology in which we define a parameterized space of policies and search this space to identify the optimum policy. For example, in concept learning, policies might be described by a fading function that specifies exemplar difficulty over time. We propose an experimental technique for searching policy spaces using Gaussian process surrogate-based optimization and a generative model of student performance. Instead of evaluating a few experimental conditions each with many human subjects, as the traditional methodology does, our technique evaluates many experimental conditions each with a few subjects. Even though individual subjects provide only a noisy estimate of the population mean, the optimization method allows us to determine the shape of the policy space and identify the global optimum, and is as efficient in its subject budget as a traditional A-B comparison. We evaluate the method via two behavioral studies, and suggest that the method has broad applicability to optimization problems involving humans in domains beyond the educational arena.

S80 Stochastic Majorization-Minimization Algorithms for Large-Scale Optimization

Julien Mairal julien.mairal@m4x.org
INRIA

Majorization-minimization algorithms consist of iteratively minimizing a majorizing surrogate of an objective function. Because of its simplicity and its wide applicability, this principle has been very popular in statistics and in signal processing. In this paper, we intend to make this principle scalable. We introduce a stochastic majorization-minimization scheme which is able to deal with large-scale or possibly infinite data sets. When applied to convex optimization problems under suitable assumptions, we show that it achieves an expected convergence rate of $O(1/\sqrt{n})$ after~n iterations, and of $O(1/n)$ for strongly convex functions. Equally important, our scheme almost surely converges to stationary points for a large class of non-convex problems. We develop several efficient algorithms based on our framework. First, we propose a new stochastic proximal gradient method, which experimentally matches state-of-the-art solvers for large-scale ℓ_1-logistic regression. Second, we develop an online DC programming algorithm for non-convex sparse estimation. Finally, we demonstrate the effectiveness of our technique for solving large-scale structured matrix factorization problems.

S81 Lasso Screening Rules via Dual Polytope Projection

Jie Wang jie.wang.ustc@asu.edu
Jiayu Zhou jiayu.zhou@asu.edu
Peter Wonka peter.wonka@asu.edu
Jieping Ye jieping.ye@asu.edu
Arizona State University

Lasso is a widely used regression technique to find sparse representations. When the dimension of the feature space and the number of samples are extremely large, solving the Lasso problem remains challenging. To improve the efficiency of solving large-scale Lasso problems, El Ghaoui and his colleagues have proposed the SAFE rules which are able to quickly identify the inactive predictors, i.e., predictors that have 0 components in the solution vector. Then, the inactive predictors or features can be removed from the optimization problem to reduce its scale. By transforming the standard Lasso to its dual form, it can be shown that the inactive predictors include the set of inactive constraints on the optimal dual solution. In this paper, we propose an efficient and effective screening rule via Dual Polytope Projections (DPP), which is mainly based on the uniqueness and nonexpansiveness of the optimal dual solution due to the fact that the feasible set in the dual space is a convex and closed polytope. Moreover, we show that our screening rule can be extended to identify inactive groups in group Lasso. To the best of our knowledge, there is currently no "exact" screening rule for group Lasso. We have evaluated our screening rule using many real data sets. Results show that our rule is more effective to identify inactive predictors than existing state-of-the-art screening rules for Lasso.

S82 Robust Transfer Principal Component Analysis with Rank Constraints

Yuhong Guo yuhong@temple.edu
Temple University

Principal component analysis (PCA), a well-established technique for data analysis and processing, provides a convenient form of dimensionality reduction that is effective for cleaning small Gaussian noises presented in the data. However, the applicability of standard principal component analysis in real scenarios is limited by its sensitivity to large errors. In this paper, we tackle the challenge problem of recovering data corrupted with errors of high magnitude by developing a novel robust transfer principal component analysis method. Our method is based on the assumption that useful information for the recovery of a corrupted data matrix can be gained from an uncorrupted related data matrix. Specifically, we formulate the data recovery problem as a joint robust principal component analysis problem on the two data matrices, with shared common principal components across matrices and individual principal components specific to each data matrix. The formulated optimization problem is a minimization problem over a convex objective function but with non-convex rank constraints. We develop an efficient proximal projected gradient descent algorithm to solve the proposed optimization problem with convergence guarantees. Our empirical results over image denoising tasks show the proposed method can effectively recover images with random large errors, and significantly outperform both standard PCA and robust PCA.

S83 Online Robust PCA via Stochastic Optimization

Jiashi Feng a0066331@nus.edu.sg
Huan Xu mpexuh@nus.edu.sg
NUS
Shuicheng Yan eleyans@nus.edu.sg
National University of Singapore

Robust PCA methods are typically based on batch optimization and have to load all the samples into memory. This prevents them from efficiently processing big data. In this paper, we develop an Online Robust Principal Component Analysis (OR-PCA) that processes one sample per time instance and hence its memory cost is independent of the data size, significantly enhancing the computation and storage efficiency. The proposed method is based on stochastic optimization of an equivalent reformulation of the batch RPCA method. Indeed, we show that OR-PCA provides a sequence of subspace estimations converging to the optimum of its batch counterpart and hence is provably robust to sparse corruption. Moreover, OR-PCA can naturally be applied for tracking dynamic subspace. Comprehensive simulations on subspace recovering and tracking demonstrate the robustness and efficiency advantages of the OR-PCA over online PCA and batch RPCA methods.

S84 The Fast Convergence of Incremental PCA

Akshay Balsubramani abalsubr@cs.ucsd.edu
Sanjoy Dasgupta dasgupta@cs.ucsd.edu
Yoav Freund yfreund@cs.ucsd.edu
UC San Diego

We prove the first finite-sample convergence rates for any incremental PCA algorithm using sub-quadratic time and memory per iteration. The algorithm analyzed is Oja’s learning rule, an efficient and well-known scheme for estimating the top principal component. Our analysis of this non-convex problem yields expected and high-probability convergence rates of $O\tilde{}(1/n)$ through a novel technique. We relate our guarantees to existing rates for stochastic gradient descent on strongly convex functions, and extend those results. We also include experiments which demonstrate convergence behaviors predicted by our analysis.

S85 Probabilistic Principal Geodesic Analysis

Miaomiao Zhang miaomiao@sci.utah.edu
P.T. Fletcher fletcher@sci.utah.edu
University of Utah

Principal geodesic analysis (PGA) is a generalization of principal component analysis (PCA) for dimensionality reduction of data on a Riemannian manifold. Currently PGA is defined as a geometric fit to the data, rather than as a probabilistic model. Inspired by probabilistic PCA, we present a latent variable model for PGA that provides a probabilistic framework for factor analysis on manifolds. To compute maximum likelihood estimates of the parameters in our model, we develop a Monte Carlo Expectation Maximization algorithm, where the expectation is approximated by Hamiltonian Monte Carlo sampling of the latent variables. We demonstrate the ability of our method to recover the ground truth parameters in simulated sphere data, as well as its effectiveness in analyzing shape variability of a corpus callosum data set from human brain images.

S86 Fast Algorithms for Gaussian Noise Invariant Independent Component Analysis

James Voss vossj@cse.ohio-state.edu
Luis Rademacher lrademac@cse.ohio-state.edu
Mikhail Belkin mbelkin@cse.ohio-state.edu
Ohio State University

The performance of standard algorithms for Independent Component Analysis quickly deteriorates under the addition of Gaussian noise. This is partially due to a common first step that typically consists of whitening, i.e., applying Principal Component Analysis (PCA) and rescaling the components to have identity covariance, which is not invariant under Gaussian noise. In our paper we develop the first practical algorithm for Independent Component Analysis that is provably invariant under Gaussian noise. The two main contributions of this work are as follows: 1. We develop and implement a more efficient version of a Gaussian noise invariant decorrelation (quasi-orthogonalization) algorithm using Hessians of the cumulant functions. 2. We propose a very simple and efficient fixed-point GI-ICA (Gradient Iteration ICA) algorithm, which is compatible with quasi-orthogonalization, as well as with the usual PCA-based whitening in the noiseless case. The algorithm is based on a special form of gradient iteration (different from gradient descent). We provide an analysis of our algorithm demonstrating fast convergence following from the basic properties of cumulants. We also present a number of experimental comparisons with the existing methods, showing superior results on noisy data and very competitive performance in the noiseless case.

S87 Online PCA for Contaminated Data

Jiashi Feng a0066331@nus.edu.sg
Huan Xu mpexuh@nus.edu.sg
NUS
Shie Mannor shie@ee.technion.ac.il
Technion
Shuicheng Yan eleyans@nus.edu.sg
National University of Singapore

We consider the online Principal Component Analysis (PCA) for contaminated samples (containing outliers) which are revealed sequentially to the Principal Components (PCs) estimator. Due to their sensitiveness to outliers, previous online PCA algorithms fail in this case and their results can be arbitrarily bad. Here we propose the online robust PCA algorithm, which is able to improve the PCs estimation upon an initial one steadily, even when faced with a constant fraction of outliers. We show that the final result of the proposed online RPCA has an acceptable degradation from the optimum. Actually, under mild conditions, online RPCA achieves the maximal robustness with a 50% breakdown point. Moreover, online RPCA is shown to be efficient for both storage and computation, since it need not re-explore the previous samples as in traditional robust PCA algorithms. This endows online RPCA with scalability for large scale data.

S88 Fantope Projection and Selection: A near-optimal convex relaxation of sparse PCA

Vince Vu vqv@stat.osu.edu
Ohio State University
Juhee Cho chojuhee@stat.wisc.edu
Karl Rohe karlrohe@stat.wisc.edu
UW-Madison
Jing Lei leij09@gmail.com
CMU

We propose a novel convex relaxation of sparse principal subspace estimation based on the convex hull of rank-d projection matrices (the Fantope). The convex problem can be solved efficiently using alternating direction method of multipliers (ADMM). We establish a near-optimal convergence rate, in terms of the sparsity, ambient dimension, and sample size, for estimation of the principal subspace of a general covariance matrix without assuming the spiked covariance model. In the special case of $d=1$, our result implies the near- optimality of DSPCA even when the solution is not rank 1. We also provide a general theoretical framework for analyzing the statistical properties of the method for arbitrary input matrices that extends the applicability and provable guarantees to a wide array of settings. We demonstrate this with an application to Kendall's tau correlation matrices and transelliptical component analysis.

S89 One-shot learning and big data with n=2

Lee Dicker ldicker@stat.rutgers.edu
Rutgers University
Dean Foster foster@wharton.upenn.edu
University of Pennsylvania

We model a "one-shot learning" situation, where very few (scalar) observations $y_1,...,y_n$ are available. Associated with each observation y_i is a very high-dimensional vector x_i, which provides context for y_i and enables us to predict subsequent observations, given their own context. One of the salient features of our analysis is that the problems studied here are easier when the dimension of $x_i$is large; in other words, prediction becomes easier when more context is provided. The proposed methodology is a variant of principal component regression (PCR). Our rigorous analysis sheds new light on PCR. For instance, we show that classical PCR estimators may be inconsistent in the specified setting, unless they are multiplied by a scalar $c>1$; that is, unless the classical estimator is expanded. This expansion phenomenon appears to be somewhat novel and contrasts with shrinkage methods ($c<1$), which are far more common in big data analyses.

S90 The Randomized Dependence Coefficient

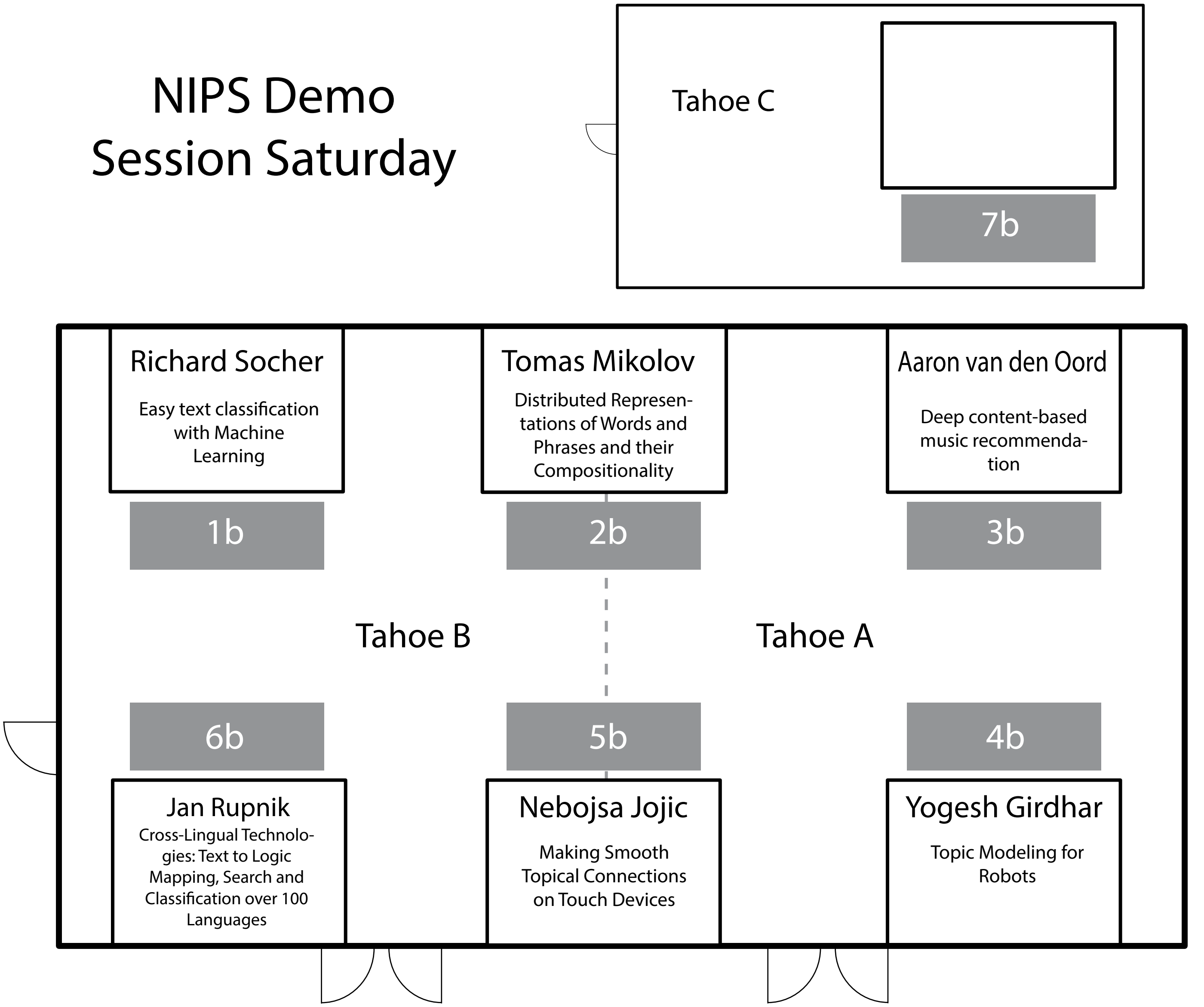
David Lopez-Paz david.lopez@tuebingen.mpg.de
MPI for Intelligent Systems & University of Cambridge
Philipp Hennig phennig@tuebingen.mpg.de
Bernhard Schölkopf bs@tuebingen.mpg.de
MPI Tübingen

We introduce the Randomized Dependence Coefficient (RDC), a measure of non-linear dependence between random variables of arbitrary dimension based on the Hirschfeld-Gebelein-Rényi Maximum Correlation Coefficient. RDC is defined in terms of correlation of random non-linear copula projections; it is invariant with respect to marginal distribution transformations, has low computational cost and is easy to implement: just five lines of R code, included at the end of the paper.

S91 Sign Cauchy Projections and Chi-Square Kernel

Ping Li pingli@cornell.edu
Gennady Samorodnitsk gs18@cornell.edu
John Hopcroft jeh@cs.cornell.edu
Cornell University

The method of Cauchy random projections is popular for computing the l_1 distance in high dimension. In this paper, we propose to use only the signs of the projected data and show that the probability of collision (i.e., when the two signs differ) can be accurately approximated as a function of the chi-square (χ_2) similarity, which is a popular measure for nonnegative data (e.g., when features are generated from histograms as common in text and vision applications). Our experiments confirm that this method of sign Cauchy random projections is promising for large-scale learning applications. Furthermore, we extend the idea to sign α-stable random projections and derive a bound of the collision probability.


NIPS Demo
Session Saturday
Tahoe C
7b
Richard Socher
Easy text classification with Machine Learning
Tomas Mikolov
Distributed Representations of Words and Phrases and their Compositionality
Aaron van den Oord
Deep content-based music recommendation
1b
2b
3b
Tahoe B
Tahoe A
6b
5b
4b
Jan Rupnik
Cross-Lingual Technologies: Text to Logic Mapping, Search and Classification over 100 Languages
Nebojsa Jojic
Making Smooth Topical Connections on Touch Devices
Yogesh Girdhar
Topic Modeling for Robots

1B Easy Text Classification with Machine Learning

Richard Socher, Bryan McCann, Andrew Ng
Stanford University
Romain Paulus
Institut supérieur d’électronique de Paris

The goal of this website is to make text classification with machine learning so easy to use that it becomes accessible to a wide audience. The website integrates with different data sources, such as Twitter. It allows people to upload their own unlabeled text via a simple drag and drop interface and have it tagged with our trained classifiers. It is also possible to upload a labeled dataset and train your own classifier. User supplied classifiers can then in turn be used by other users. Two of many possible exciting classifiers will classify sentiment on Twitter or whether a kickstarter proposal is likely to succeed.

2B Distributed Representations of Words and Phrases and their Compositionality

Tomas Mikolov, Kai Chen, Greg Corrado
Google Research

We will demonstrate quality of word and phrase representations derived from neural network models that were trained on about one hundred billion words. Several interactive applications will be available to the audience for exploration: - search for similar words, phrases, search queries and sentences - visualization of the word vectors - answering of analogy questions

3B Deep Content-Based Music Recommendation

Aaron van den Oord, Sander Dieleman, Benjamin Schrauwen
Ghent University

Automatic music recommendation has become an increasingly relevant problem in recent years, since a lot of music is now sold and consumed digitally. Most recommender systems rely on collaborative filtering. However, this approach suffers from the cold start problem: it fails when no usage data is available, so it is not effective for recommending new and unpopular songs. We propose to use a latent factor model for recommendation, and predict the latent factors from music audio using a deep convolutional neural network when they cannot be obtained from usage data. Our demo processes music clips obtained from YouTube, selected by the user, and finds other clips with similar (predicted) usage patterns from a large database of 600,000 songs (a subset of the Million Song Dataset).

4B Topic Modeling for Robots

Yogesh Girdhar, Gregory Dudek
McGill University

ROST is a realtime online spatiotemporal topic modeling framework for data such as streaming video and audio observed by a robot, where topics represent the latent causes that produce these observations. When new observations are made, we not only compute its topic labels, but also use it to update the global topic model and the labels of older observations, resulting in a consistent semantic description of the scene without the use of any prior knowledge. The proposed approximations have constant update time as new data is observed, which allows for the technique to work in real-time; a critical requirement for its use in the robotics.

5B Making Smooth Topical Connections on Touch Devices

Nebojsa Jojic, Alessandro Perina, Andrzej Truski
Microsoft Research

A strategy is proposed for mining, browsing, and searching through documents consisting of text, images, and other modalities: A collection of documents is represented as a grid of keywords with varying font sizes that indicate the words' weights. The grid is based on the counting-grid model so that each document matches in its word usage the word-weight distribution in some window in the grid. This strategy leads to denser packing and higher relatedness of nearby documents—documents that map to overlapping windows literally share the words found in the overlap. Smooth thematic shifts become evident in the grid, providing connections among distant topics and guiding the user's attention in search for the spot of interest.

6B Cross-Lingual Technologies: Text to Logic Mapping, Search and Classification over 100 Languages

Jan Rupnik, Andrej Muhic, Blaz Fortuna, Janez Starc, Marko Grobelnik
Jozef Stefan Institute
Michael Witbrock
Cycorp

We demonstrate two approaches that enable language independent document representations. The first approach is based on factorization techniques where documents are expressed in terms of multi-lingual topic vectors. The second approach is based on mapping documents to statements in first-order logic. Our demonstration is composed of two parts: the first part demonstrates a scalable solution to cross-lingual document retrieval and classification over 100 languages. The users will be able to input a document in any of the supported languages, find similar documents in other languages, and classify the document in the Open Directory Project taxonomy. The second part demonstrates a solution for the problem of text understanding aiming to enable machines to "understand" and reason about the semantics of a human composed text. The goal is to extract inferentially capable knowledge from textual documents of the given domain at economically viable cost. The audience will be able to participate in the process of making "text to logic" mappings and reasoning about the extracted knowledge.

SUNDAY

ORAL SESSION

SESSION 9 - 9:00 - 10:10 AM

Session Chair: Satinder Singh

POSNER LECTURE: Neural Reinforcement Learning

Peter Dayan dayan@gatsby.ucl.ac.uk
Gatsby Unit, UCL

Reinforcement learning has become a wide and deep conduit that links ideas and results in computer science, statistics, control theory and economics to psychological data on animal and human decision-making, and the neural basis of choice. There is a ready and free flow of ideas among these disciplines, providing a powerful foundation for exploring some of the complexities of both normal and abnormal behaviours. I will outline some of the happy circumstances that led us to this point; discuss current computational, algorithmic and implementational themes; and provide some pointers to the future.

I am Director of the Gatsby Computational Neuroscience Unit at University College London. I studied mathematics at the University of Cambridge and then did a PhD at the University of Edinburgh, specialising in associative memory and reinforcement learning. I did postdocs with Terry Sejnowski at the Salk Institute and Geoff Hinton at the University of Toronto, then became an Assistant Professor in Brain and Cognitive Science at the Massachusetts Institute of Technology before moving to UCL.

Actor-Critic Algorithms for Risk-Sensitive MDPs

Prashanth L.A. prashanth.la@inria.fr
INRIA
Mohammad Ghavamzadeh
mohammad.ghavamzadeh@inria.fr
INRIA & Adobe Research

In many sequential decision-making problems we may want to manage risk by minimizing some measure of variability in rewards in addition to maximizing a standard criterion. Variance related risk measures are among the most common risk-sensitive criteria in finance and operations research. However, optimizing many such criteria is known to be a hard problem. In this paper, we consider both discounted and average reward Markov decision processes. For each formulation, we first define a measure of variability for a policy, which in turn gives us a set of risk-sensitive criteria to optimize. For each of these criteria, we derive a formula for computing its gradient. We then devise actor-critic algorithms for estimating the gradient and updating the policy parameters in the ascent direction. We establish the convergence of our algorithms to locally risk-sensitive optimal policies. Finally, we demonstrate the usefulness of our algorithms in a traffic signal control application.

SPOTLIGHT SESSION

Session 9, 10:10 – 10:30 AM

- **Learning from Limited Demonstrations**
  B. Kim, A. Farahmand, J. Pineau, D. Precup, McGill University
  See abstract Su48, page 119

- **Distributed Exploration in Multi-Armed Bandits**
  E. Hillel, Z. Karnin, R. Lempel, O. Somekh, Yahoo! Labs; T. Koren, Technion
  See abstract Su15, page 111

- **Dimension-Free Exponentiated Gradient**
  F. Orabona, TTI Chicago
  See abstract Su13, page 110

- **Generalizing Analytic Shrinkage for Arbitrary Covariance Structures**
  D. Bartz, K. Müller, TU Berlin
  See abstract Su43, page 117

- **Robust Spatial Filtering with Beta Divergence**
  W. Samek, D. Blythe, K. Müller, TU Berlin; M. Kawanabe, ATR
  See abstract Su44, page 118

ORAL SESSION

Session 10, 10:50 AM – 12:00 PM

Session Chair: Michael Jordan

INVITED TALK: New Methods for the Analysis of Genome Variation Data

Richard Durbin rd@sanger.ac.uk
Wellcome Trust Sanger Institute

Genetic variation in genome sequences within a species such as humans underpins our biological diversity, is the basis for the genetic contribution to disease, provides information about our ancestry, and is the substrate for evolution. Genetic variation has a complex structure of shared inheritance from a common ancestor at each position in the genome, with the pattern of sharing changing along the genome as a consequence of genetic recombination. The scale of data sets that can be obtained from modern sequencing and genotyping methods, currently of the order of hundreds of terabytes, makes analysis computationally challenging. During the last few years, a number of tools such as BWA, Bowtie have been developed for sequence matching based on suffix array derived data structures, in particular the Burrows-Wheeler tranform (BWT) and Ferragina-Manzini (FM) index, which have the nice property that they not only give asymptotically optimal search, but also are highly compressed data structures (they underlie the bzip compression algorithms). I will discuss a number of approaches based on these data structures for primary data processing, sequence assembly, variation detection and large scale genetic analysis, with applications to very large scale human genetic variation data sets.

Richard Durbin is a Senior Group Leader and joint Head of Human Genetics at The Wellcome Trust Sanger Institute. He is currently co-leading the 1000 Genomes Project to produce a deep catalogue of human genetic variation by large scale sequencing, and the UK10K collaboration to extend sequence based genetics to samples with clinically relevant phenotypes. Previously Richard contributed to the human genome project, and development of the Pfam database of protein families and the Ensembl genome data resource. He has also made theoretical and algorithmic contributions to biological sequence analysis. Richard has a BA in Mathematics, and a PhD in Biology from Cambridge University, where he was also a Research Fellow, at King's College, from 1986 to 1988. He was a Fulbright Visiting Scholar in Biophysics at Harvard University from 1982 to 1983 and a Lucille P Markey visiting Fellow in the Department of Psychology, Stanford University from 1988 to 1990. He was a staff scientist at the MRC Laboratory of Molecular Biology from 1990 to 1996, and was Head of Informatics at the Sanger Institute from 1992-2006 and Deputy Director from 1997 to 2006. He was elected a Fellow of the Royal Society in 2004. Richard's home page can be found at http://www.sanger.ac.uk/research/faculty/rdurbin/

BIG & QUIC: Sparse Inverse Covariance Estimation for a Million Variables

Cho-Jui Hsieh cjhsieh@cs.utexas.edu
Pradeep Ravikumar pradeepr@cs.utexas.edu
UT Austin
Matyas Sustik msustik@gmail.com
Inderjit Dhillon inderjit@cs.utexas.edu
Russell Poldrack poldrack@utexas.edu
University of Texas

The l1-regularized Gaussian maximum likelihood estimator (MLE) has been shown to have strong statistical guarantees in recovering a sparse inverse covariance matrix even under high-dimensional settings. However, it requires solving a difficult non-smooth log-determinant program with number of parameters scaling quadratically with the number of Gaussian variables. State-of-the-art methods thus do not scale to problems with more than 20,000 variables. In this paper, we develop an algorithm BigQUIC, which can solve 1 million dimensional l1-regularized Gaussian MLE problems (which would thus have 1000 billion parameters) using a single machine, with bounded memory. In order to do so, we carefully exploit the underlying structure of the problem. Our innovations include a novel block-coordinate descent method with the blocks chosen via a clustering scheme to minimize repeated computations; and allowing for inexact computation of specific components. In spite of these modifications, we are able to theoretically analyze our procedure and show that BigQUIC can achieve super-linear or even quadratic convergence rates.

SPOTLIGHT SESSION

Session 10, 12:00 – 12:20 PM

- **Speeding up Permutation Testing in Neuroimaging**
  C. Hinrichs, V. Ithapu, Q. Sun, S. Johnson, V. Singh, UW-Madison
  See abstract Sun34, page 115

- **Robust Multimodal Graph Matching: Sparse Coding Meets Graph Matching**
  M. Fiori, P. Muse, Universidad de la República, Uruguay; P. Sprechmann, J. Vogelstein, G. Sapiro, Duke University
  See abstract Sun22, page 113

- **Deep Fisher Networks for Large-Scale Image Classification**
  K. Simonyan, A. Vedaldi, A. Zisserman, University of Oxford
  See abstract Sun79, page 127

Sinkhorn Distances: Lightspeed Computation of Optimal Transportation
M. Cuturi, Kyoto University
See abstract Sun89, page 129

- **Understanding variable importances in forests of randomized trees**
  G. Louppe, L. Wehenkel, A. Sutera, P. Geurts, Université de Liège
  See abstract Sun88, page 129

12:20 – 12:30 PM - Closing Remarks

POSTER SESSION

Session, 2:00 – 6:00 PM

Su1 Sparse Additive Text Models with Low Rank Background
L. Shi

Su2 Documents as multiple overlapping windows into grids of counts
A. Perina, N. Jojic, M. Bicego, A. Truski

Su3 On Algorithms for Sparse Multi-factor NMF
S. Lyu, X. Wang

Su4 Learning Adaptive Value of Information for Structured Prediction
D. Weiss, B. Taskar

Su5 Symbolic Opportunistic Policy Iteration for Factored-Action MDPs
A. Raghavan, R. Khardon, A. Fern, P. Tadepalli

Su6 Point Based Value Iteration with Optimal Belief Compression for Dec-POMDPs
L. MacDermed, C. Isbell

Su7 Convergence of Monte Carlo Tree Search in Simultaneous Move Games
V. Lisy, V. Kovarik, M. Lanctot, B. Bosansky

Su8 Estimation Bias in Multi-Armed Bandit Algorithms for Search Advertising
M. Xu, T. Qin, T. Liu

Su9 Optimization, Learning, and Games with Predictable Sequences
S. Rakhlin, K. Sridharan

Su10 Minimax Optimal Algorithms for Unconstrained Linear Optimization
B. McMahan, J. Abernethy

Su11 Online Learning with Costly Features and Labels
N. Zolghadr, G. Bartok, R. Greiner, A. György, C. Szepesvari

Su12 The Pareto Regret Frontier
W. Koolen

Su13 Dimension-Free Exponentiated Gradient
F. Orabona

Su14 Online Learning with Switching Costs and Other Adaptive Adversaries
N. Cesa-Bianchi, O. Dekel, O. Shamir

Su15 Distributed Exploration in Multi-Armed Bandits
E. Hillel, Z. Karnin, T. Koren, R. Lempel, O. Somekh

Su16 High-Dimensional Gaussian Process Bandits
J. Djolonga, A. Krause, V. Cevher

Su17 On Poisson Graphical Models
E. Yang, P. Ravikumar, G. Allen, Z. Liu

Su18 Conditional Random Fields via Univariate Exponential Families
E. Yang, P. Ravikumar, G. Allen, Z. Liu

Su19 Scalable kernels for graphs with continuous attributes
A. Feragen, N. Kasenburg, J. Petersen, M. de Bruijne, K. Borgwardt

Su20 Near-optimal Anomaly Detection in Graphs using Lovasz Extended Scan Statistic
J. Sharpnack, A. Krishnamurthy, A. Singh

Su21 Analyzing the Harmonic Structure in Graph-Based Learning
X. Wu, Z. Li, S. Chang

Su22 Robust Multimodal Graph Matching: Sparse Coding Meets Graph Matching
M. Fiori, P. Sprechmann, J. Vogelstein, P. Muse, G. Sapiro

Su23 Learning Gaussian Graphical Models with Observed or Latent FVSs
Y. Liu, A. Willsky

Su24 Global MAP-Optimality by Shrinking the Combinatorial Search Area with Convex Relaxation
B. Savchynskyy, J. Kappes, P. Swoboda, C. Schnörr

Su25 First-order Decomposition Trees
N. Taghipour, J. Davis, H. Blockeel

Su26 Binary to Bushy: Bayesian Hierarchical Clustering with the Beta Coalescent
Y. Hu, J. Boyd-Graber, H. Daume III, Z. Ying

Su27 Parallel Sampling of DP Mixture Models using Sub-Cluster Splits
J. Chang, J. Fisher III

Su28 Lexical and Hierarchical Topic Regression
V. Nguyen, J. Boyd-Graber, P. Resnik

Su29 A Novel Two-Step Method for Cross Language Representation Learning
M. Xiao, Y. Guo

Su30 Learning word embeddings efficiently with noise-contrastive estimation
A. Mnih, k. kavukcuoglu

Su31 Training and Analysing Deep Recurrent Neural Networks
M. Hermans, B. Schrauwen

Su32 Extracting regions of interest from biological images with convolutional sparse block coding
M. Pachitariu, A. Packer, N. Pettit, H. Dalgleish, M. Hausser, M. Sahani

Su33 Mapping paradigm ontologies to and from the brain
Y. Schwartz, B. Thirion, G. Varoquaux

Su34 Speeding up Permutation Testing in Neuroimaging
C. Hinrichs, V. Ithapu, Q. Sun, S. Johnson, V. Singh

Su35 BIG & QUIC: Sparse Inverse Covariance Estimation for a Million Variables
C. Hsieh, M. Sustik, I. Dhillon, P. Ravikumar, R. Poldrack

Su36 Geometric optimisation on positive definite matrices for elliptically contoured distributions
S. Sra, R. Hosseini

Su37 Estimating the Unseen: Improved Estimators for Entropy and other Properties
P. Valiant, G. Valiant

Su38 Factorized Asymptotic Bayesian Inference for Latent Feature Models
K. Hayashi, R. Fujimaki

Su39 Tracking Time-varying Graphical Structure
E. Kummerfeld, D. Danks

Su40 Sparse Inverse Covariance Estimation with Calibration
T. Zhao, H. Liu

Su41 A* Lasso for Learning a Sparse Bayesian Network Structure for Continuous Variables
J. Xiang, S. Kim

Su42 On model selection consistency of penalized M-estimators: a geometric theory
J. Lee, Y. Sun, J. Taylor

Su43 Generalizing Analytic Shrinkage for Arbitrary Covariance Structures
D. Bartz, K. Müller

Su44 Robust Spatial Filtering with Beta Divergence
W. Samek, D. Blythe, K. Müller, M. Kawanabe

Su45 A multi-agent control framework for co-adaptation in brain-computer interfaces
J. Merel, R. Fox, T. Jebara, L. Paninski

Su46 Probabilistic Movement Primitives
A. Paraschos, C. Daniel, J. Peters, G. Neumann

Su47 Variational Policy Search via Trajectory Optimization
S. Levine, V. Koltun

Su48 Learning from Limited Demonstrations
B. Kim, A. Farahmand, J. Pineau, D. Precup

Su49 Learning Trajectory Preferences for Manipulators via Iterative Improvement
A. Jain, B. Wojcik, T. Joachims, A. Saxena

Su50 Forgetful Bayes and myopic planning: Human learning and decision-making in a bandit setting
S. Zhang, A. Yu

Su51 Context-sensitive active sensing in humans
S. Ahmad, H. Huang, A. Yu

Su52 Bellman Error Based Feature Generation using Random Projections on Sparse Spaces
M. Milani Fard, Y. Grinberg, A. Farahmand, J. Pineau, D. Precup

Su53 Reinforcement Learning in Robust Markov Decision Processes
S. Lim, H. Xu, S. Mannor

Su54 Projected Natural Actor-Critic
P. Thomas, W. Dabney, S. Giguere, S. Mahadevan

Su55 (More) Efficient Reinforcement Learning via Posterior Sampling
I. Osband, D. Russo, B. Van Roy

Su56 Adaptive Step-Size for Policy Gradient Methods
M. Pirotta, M. Restelli, L. Bascetta

Su57 Policy Shaping: Integrating Human Feedback with Reinforcement Learning
S. Griffith, K. Subramanian, J. Scholz, C. Isbell, A. Thomaz

Su58 Optimistic policy iteration and natural actor-critic: A unifying view and a non-optimality result
P. Wagner

Su59 Actor-Critic Algorithms for Risk-Sensitive MDPs
P. L.A., M. Ghavamzadeh

Su60 DESPOT: Online POMDP Planning with Regularization
A. Somani, N. Ye, D. Hsu, W. Lee

Su61 Approximate Dynamic Programming Finally Performs Well in the Game of Tetris
V. Gabillon, M. Ghavamzadeh, B. Scherrer

Su62 Reward Mapping for Transfer in Long-Lived Agents
X. Guo, S. Singh, R. Lewis

Su63 Learning a Deep Compact Image Representation for Visual Tracking
N. Wang, D. Yeung

Su64 Learning the Local Statistics of Optical Flow
D. Rosenbaum, D. Zoran, Y. Weiss

Su65 Third-Order Edge Statistics: Contour Continuation, Curvature, and Cortical Connections
M. Lawlor, S. Zucker

Su66 What Are the Invariant Occlusive Components of Image Patches? A Probabilistic Generative Approach
Z. Dai, G. Exarchakis, J. Lücke

Su67 Action from Still Image Dataset and Inverse Optimal Control to Learn Task Specific Visual Scanpaths
S. Mathe, C. Sminchisescu

Su68 Action is in the Eye of the Beholder: Eye-gaze Driven Model for Spatio-Temporal Action Localization
N. Shapovalova, M. Raptis, L. Sigal, G. Mori

Su69 Higher Order Priors for Joint Intrinsic Image, Objects, and Attributes Estimation
V. Vineet, C. Rother, P. Torr

Su70 Decision Jungles: Compact and Rich Models for Classification
J. Shotton, T. Sharp, P. Kohli, S. Nowozin, J. Winn, A. Criminisi

Su71 Non-Linear Domain Adaptation with Boosting
C. Becker, C. Christoudias, P. Fua

Su72 Modeling Clutter Perception using Parametric Proto-object Partitioning
C. Yu, W. Hua, D. Samaras, G. Zelinsky

Su73 Mid-level Visual Element Discovery as Discriminative Mode Seeking
C. Doersch, A. Gupta, A. Efros

Su74 Optimal integration of visual speed across different spatiotemporal frequency channels
M. Jogan, A. Stocker

Su75 DeViSE: A Deep Visual-Semantic Embedding Model
A. Frome, G. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, T. Mikolov

Su76 Visual Concept Learning: Combining Machine Vision and Bayesian Generalization on Concept Hierarchies
Y. Jia, J. Abbott, J. Austerweil, T. Griffiths, T. Darrell

Su77 Learning invariant representations and applications to face verification
Q. Liao, J. Leibo, T. Poggio

Su78 Deep Neural Networks for Object Detection
C. Szegedy, A. Toshev, D. Erhan

Su79 Deep Fisher Networks for Large-Scale Image Classification
K. Simonyan, A. Vedaldi, A. Zisserman

Su80 Fast Template Evaluation with Vector Quantization
M. Sadeghi, D. Forsyth

Su81 Transfer Learning in a Transductive Setting
M. Rohrbach, S. Ebert, B. Schiele

Su82 Reshaping Visual Datasets for Domain Adaptation
B. Gong, K. Grauman, F. Sha

Su83 Heterogeneous-Neighborhood-based Multi-Task Local Learning Algorithms
Y. Zhang

Su84 Learning Feature Selection Dependencies in Multi-task Learning
D. Hernández-Lobato, J. Hernández-Lobato

Su85 Parametric Task Learning
I. Takeuchi, T. Hongo, M. Sugiyama, S. Nakajima

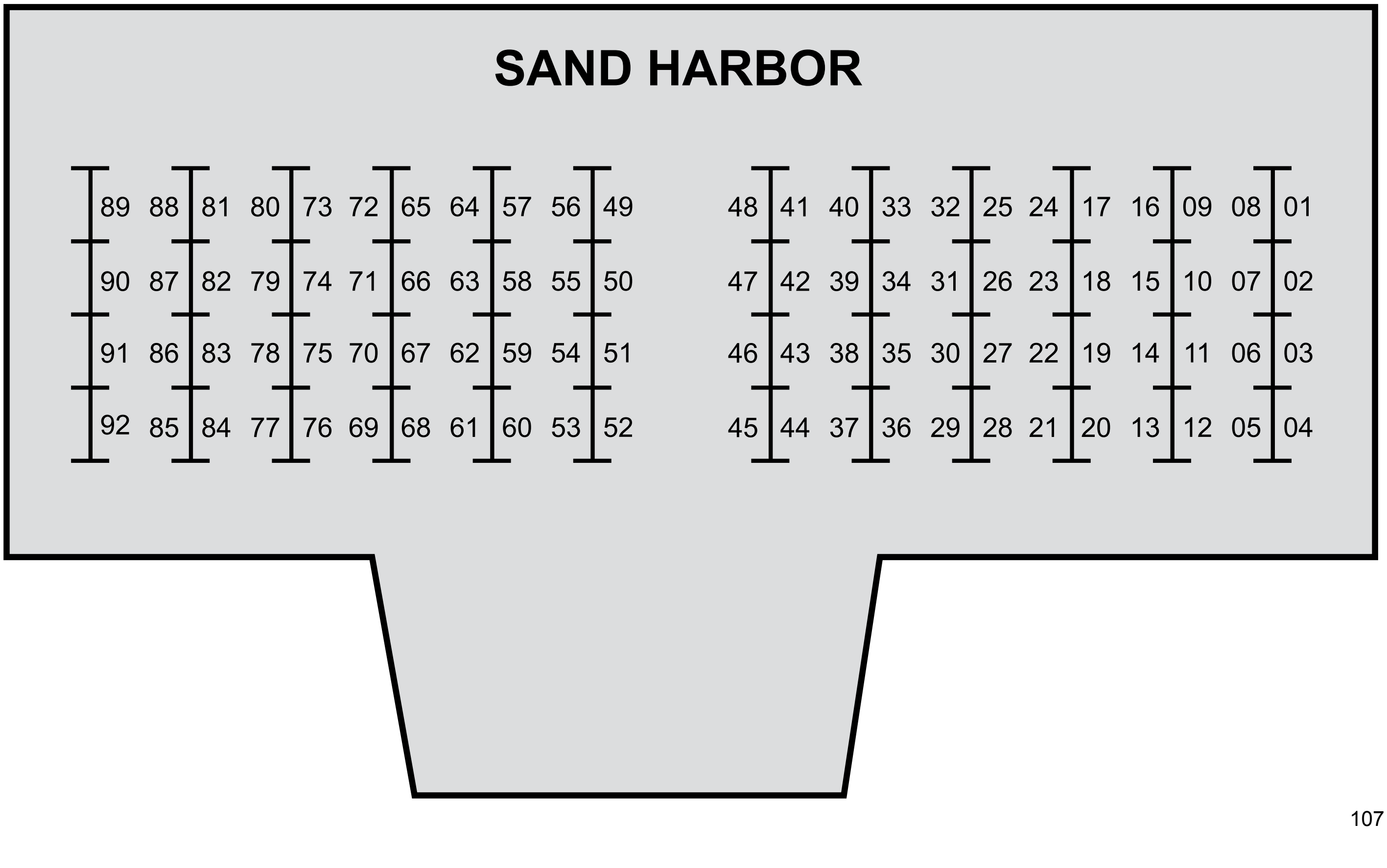
Su86 Direct 0-1 Loss Minimization and Margin Maximization with Boosting
S. Zhai, T. Xia, M. Tan, S. Wang

Su87 Reservoir Boosting : Between Online and Offline Ensemble Learning
L. Lefakis, F. Fleuret

Su88 Understanding variable importances in forests of randomized trees
G. Louppe, L. Wehenkel, A. Sutera, P. Geurts

Su89 Sinkhorn Distances: Lightspeed Computation of Optimal Transportation
M. Cuturi

Su90 Beyond Pairwise: Provably Fast Algorithms for Approximate k-Way Similarity Search
A. Shrivastava, P. Li[



Su01 Sparse Additive Text Models with Low Rank Background

Lei Shi shilei06@baidu.com
Baidu

The sparse additive model for text modeling involves the sum-of-exp computing, with consuming costs for large scales. Moreover, the assumption of equal background across all classes/topics may be too strong. This paper extends to propose sparse additive model with low rank background (SAM-LRB), and simple yet efficient estimation. Particularly, by employing a double majorization bound, we approximate the log-likelihood into a quadratic lower-bound with the sum-of-exp terms absent. The constraints of low rank and sparsity are then simply embodied by nuclear norm and ℓ_1-norm regularizers. Interestingly, we find that the optimization task in this manner can be transformed into the same form as that in Robust PCA. Consequently, parameters of supervised SAM-LRB can be efficiently learned using an existing algorithm for Robust PCA based on accelerated proximal gradient. Besides the supervised case, we extend SAM-LRB to also favor unsupervised and multifaceted scenarios. Experiments on real world data demonstrate the effectiveness and efficiency of SAM-LRB, showing state-of-the-art performances.

Su02 Documents as multiple overlapping windows into grids of counts

Alessandro Perina alessandro.perina@gmail.com
Nebojsa Jojic jojic@microsoft.com
Andrzej Truski andrzejt@microsoft.com
Microsoft Research
Manuele Bicego manuele.bicego@univr.it
University of Verona

In text analysis documents are represented as disorganized bags of words, models of count features are typically based on mixing a small number of topics \cite{lda,sam}. Recently, it has been observed that for many text corpora documents evolve into one another in a smooth way, with some features dropping and new ones being introduced. The counting grid cgUai models this spatial metaphor literally: it is multidimensional grid of word distributions learned in such a way that a document's own distribution of features can be modeled as the sum of the histograms found in a window into the grid. The major drawback of this method is that it is essentially a mixture and all the content much be generated by a single contiguous area on the grid. This may be problematic especially for lower dimensional grids. In this paper, we overcome to this issue with the *Componential Counting Grid* which brings the componential nature of topic models to the basic counting grid. We also introduce a generative kernel based on the document's grid usage and a visualization strategy useful for understanding large text corpora. We evaluate our approach on document classification and multimodal retrieval obtaining state of the art results on standard benchmarks.

Su03 On Algorithms for Sparse Multi-factor NMF

Siwei Lyu slyu@albany.edu
Xin Wang xwang26@albany.edu
SUNY at Albany

Nonnegative matrix factorization (NMF) is a popular data analysis method, the objective of which is to decompose a matrix with all nonnegative components into the product of two other nonnegative matrices. In this work, we describe a new simple and efficient algorithm for multi-factor nonnegative matrix factorization problem ({mfNMF}), which generalizes the original NMF problem to more than two factors. Furthermore, we extend the mfNMF algorithm to incorporate a regularizer based on Dirichlet distribution over normalized columns to encourage sparsity in the obtained factors. Our sparse NMF algorithm affords a closed form and an intuitive interpretation, and is more efficient in comparison with previous works that use fix point iterations. We demonstrate the effectiveness and efficiency of our algorithms on both synthetic and real data sets.

Su04 Learning Adaptive Value of Information for Structured Prediction

David Weiss djweiss@cis.upenn.edu
University of Pennsylvania
Ben Taskar taskar@cs.washington.edu
University of Washington

Discriminative methods for learning structured models have enabled wide-spread use of very rich feature representations. However, the computational cost of feature extraction is prohibitive for large-scale or time-sensitive applications, often dominating the cost of inference in the models. Significant efforts have been devoted to sparsity-based model selection to decrease this cost. Such feature selection methods control computation statically and miss the opportunity to fine-tune feature extraction to each input at run-time. We address the key challenge of learning to control fine-grained feature extraction adaptively, exploiting non-homogeneity of the data. We propose an architecture that uses a rich feedback loop between extraction and prediction. The run-time control policy is learned using efficient value-function approximation, which adaptively determines the value of information of features at the level of individual variables for each input. We demonstrate significant speedups over state-of-the-art methods on two challenging datasets. For articulated pose estimation in video, we achieve a more accurate state-of-the-art model that is simultaneously $4\times$ faster while using only a small fraction of possible features, with similar results on an OCR task.

Su05 Symbolic Opportunistic Policy Iteration for Factored-Action MDPs

Aswin Raghavan nadamuna@eecs.oregonstate.edu
Alan Fern afern@eecs.oregonstate.edu
Prasad Tadepalli tadepall@eecs.oregonstate.edu
Oregon State University
Roni Khardon roni@cs.tufts.edu
Tufts University

We address the scalability of symbolic planning under uncertainty with factored states and actions. Prior work has focused almost exclusively on factored states but not factored actions, and on value iteration (VI) compared to policy iteration (PI). Our first contribution is a novel method for symbolic policy backups via the application of constraints, which is used to yield a new efficient symbolic imple- mentation of modified PI (MPI) for factored action spaces. While this approach improves scalability in some cases, naive handling of policy constraints comes with its own scalability issues. This leads to our second and main contribution, symbolic Opportunistic Policy Iteration (OPI), which is a novel convergent algorithm lying between VI and MPI. The core idea is a symbolic procedure that applies policy constraints only when they reduce the space and time complexity of the update, and otherwise performs full Bellman backups, thus automatically adjusting the backup per state. We also give a memory bounded version of this algorithm allowing a space-time tradeoff. Empirical results show significantly improved scalability over the state-of-the-art.

Su06 Point Based Value Iteration with Optimal Belief Compression for Dec-POMDPs

Liam MacDermed liam@cc.gatech.edu
Charles Isbell isbell@cc.gatech.edu
Georgia Tech

This paper presents four major results towards solving decentralized partially observable Markov decision problems (DecPOMDPs) culminating in an algorithm that outperforms all existing algorithms on all but one standard infinite-horizon benchmark problems. (1) We give an integer program that solves collaborative Bayesian games (CBGs). The program is notable because its linear relaxation is very often integral. (2) We show that a DecPOMDP with bounded belief can be converted to a POMDP (albeit with actions exponential in the number of beliefs). These actions correspond to strategies of a CBG. (3) We present a method to transform any DecPOMDP into a DecPOMDP with bounded beliefs (the number of beliefs is a free parameter) using optimal (not lossless) belief compression. (4) We show that the combination of these results opens the door for new classes of DecPOMDP algorithms based on previous POMDP algorithms. We choose one such algorithm, point-based valued iteration, and modify it to produce the first tractable value iteration method for DecPOMDPs which outperforms existing algorithms.

Su07 Convergence of Monte Carlo Tree Search in Simultaneous Move Games

Viliam Lisy viliam.lisy@fel.cvut.cz
Vojta Kovarik vojta.kovarik@gmail.com
Branislav Bosansky
branislav.bosansky@agents.fel.cvut.cz
CTU in Prague
Marc Lanctot marc.lanctot@maastrichtuniversity.nl
Maastricht University

In this paper, we study Monte Carlo tree search (MCTS) in zero-sum extensive-form games with perfect information and simultaneous moves. We present a general template of MCTS algorithms for these games, which can be instantiated by various selection methods. We formally prove that if a selection method is ϵ-Hannan consistent in a matrix game and satisfies additional requirements on exploration, then the MCTS algorithm eventually converges to an approximate Nash equilibrium (NE) of the extensive-form game. We empirically evaluate this claim using regret matching and Exp3 as the selection methods on randomly generated and worst case games. We confirm the formal result and show that additional MCTS variants also converge to approximate NE on the evaluated games.

Su08 Estimation Bias in Multi-Armed Bandit Algorithms for Search Advertising

Min Xu minx@cs.cmu.edu
CMU
Tao Qin taoqin@microsoft.com
Tie-Yan Liu tie-yan.liu@microsoft.com
Microsoft Research

In search advertising, the search engine needs to select the most profitable advertisements to display, which can be formulated as an instance of online learning with partial feedback, also known as the stochastic multi-armed bandit (MAB) problem. In this paper, we show that the naive application of MAB algorithms to search advertising for advertisement selection will produce sample selection bias that harms the search engine by decreasing expected revenue and “estimation of the largest mean” (ELM) bias that harms the advertisers by increasing game-theoretic player-regret. We then propose simple bias-correction methods with benefits to both the search engine and the advertisers.

Su09 Optimization, Learning, and Games with Predictable Sequences

Sasha Rakhlin rakhlin@gmail.com
Karthik Sridharan karthik@ttic.edu
University of Pennsylvania

We provide several applications of Optimistic Mirror Descent, an online learning algorithm based on the idea of predictable sequences. First, we recover the Mirror-Prox algorithm, prove an extension to Holder-smooth functions, and apply the results to saddle-point type problems. Second, we prove that a version of Optimistic Mirror Descent (which has a close relation to the Exponential Weights algorithm) can be used by two strongly-uncoupled players in a finite zero-sum matrix game to converge to the minimax equilibrium at the rate of O(log T / T). This addresses a question of Daskalakis et al, 2011. Further, we consider a partial information version of the problem. We then apply the results to approximate convex programming and show a simple algorithm for the approximate Max-Flow problem.

Su10 Minimax Optimal Algorithms for Unconstrained Linear Optimization

Brendan McMahan mcmahan@google.com
Google Research
Jacob Abernethy jabernet@umich.edu
University of Pennsylvania

We design and analyze minimax-optimal algorithms for online linear optimization games where the player's choice is unconstrained. The player strives to minimize regret, the difference between his loss and the loss of a post-hoc benchmark strategy. The standard benchmark is the loss of the best strategy chosen from a bounded comparator set, whereas we consider a broad range of benchmark functions. We consider the problem as a sequential multi-stage zero-sum game, and we give a thorough analysis of the minimax behavior of the game, providing characterizations for the value of the game, as well as both the player's and the adversary's optimal strategy. We show how these objects can be computed efficiently under certain circumstances, and by selecting an appropriate benchmark, we construct a novel hedging strategy for an unconstrained betting game.

Su11 Online Learning with Costly Features and Labels

Navid Zolghadr zolghadr@ualberta.ca
Russell Greiner rgreiner@ualberta.ca
András György gyorgy@ualberta.ca
Csaba Szepesvari szepesva@cs.ualberta.ca
University of Alberta
Gabor Bartok bartok@inf.ethz.ch
ETH Zurich

This paper introduces the "online probing" problem: In each round, the learner is able to purchase the values of a subset of feature values. After the learner uses this information to come up with a prediction for the given round, he then has the option of paying for seeing the loss that he is evaluated against. Either way, the learner pays for the imperfections of his predictions and whatever he chooses to observe, including the cost of observing the loss function for the given round and the cost of the observed features. We consider two variations of this problem, depending on whether the learner can observe the label for free or not. We provide algorithms and upper and lower bounds on the regret for both variants. We show that a positive cost for observing the label significantly increases the regret of the problem.

Su12 The Pareto Regret Frontier

Wouter Koolen wouter.koolen@qut.edu.au
Queensland University of Technology

Performance guarantees for online learning algorithms typically take the form of regret bounds, which express that the cumulative loss overhead compared to the best expert in hindsight is small. In the common case of large but structured expert sets we typically wish to keep the regret especially small compared to simple experts, at the cost of modest additional overhead compared to more complex others. We study which such regret trade-offs can be achieved, and how. We analyse regret w.r.t. each individual expert as a multi-objective criterion in the simple but fundamental case of absolute loss. We characterise the achievable and Pareto optimal trade-offs, and the corresponding optimal strategies for each sample size both exactly for each finite horizon and asymptotically.

Su13 Dimension-Free Exponentiated Gradient

Francesco Orabona orabona@ttic.edu
TTI Chicago

We present a new online learning algorithm that extends the exponentiated gradient to infinite dimensional spaces. Our analysis shows that the algorithm is implicitly able to estimate the L_2 norm of the unknown competitor, U, achieving a regret bound of the order of $O(U \log(UT+1))\sqrt{T})$, instead of the standard $O((U^2+1)\sqrt{T})$, achievable without knowing U. For this analysis, we introduce novel tools for algorithms with time-varying regularizers, through the use of local smoothness. Through a lower bound, we also show that the algorithm is optimal up to $\log\sqrt{T}$ term for linear and Lipschitz losses.

Su14 Online Learning with Switching Costs and Other Adaptive Adversaries

Nicolò Cesa-Bianchi nicolo.cesa-bianchi@unimi.it
University of Milan
Ofer Dekel oferd@microsoft.com
Microsoft Research
Ohad Shamir ohad.shamir@weizmann.ac.il
The Weizmann Institute

We study the power of different types of adaptive (nonoblivious) adversaries in the setting of prediction with expert advice, under both full-information and bandit feedback. We measure the player's performance using a new notion of regret, also known as policy regret, which better captures the adversary's adaptiveness to the player's behavior. In a setting where losses are allowed to drift, we characterize ---in a nearly complete manner--- the power of adaptive adversaries with bounded memories and switching costs. In particular, we show that with switching costs, the attainable rate with bandit feedback is $T^{2/3}$. Interestingly, this rate is significantly worse than the $\sqrt{T}$ rate attainable with switching costs in the full-information case. Via a novel reduction from experts to bandits, we also show that a bounded memory adversary can force $T^{2/3}$ regret even in the full information case, proving that switching costs are easier to control than bounded memory adversaries. Our lower bounds rely on a new stochastic adversary strategy that generates loss processes with strong dependencies.

Su15 Distributed Exploration in Multi-Armed Bandits

Eshcar Hillel eshcar@yahoo-inc.com
Zohar Karnin zkarnin@yahoo-inc.com
Ronny Lempel rlempel@yahoo-inc.com
Oren Somekh orens@yahoo-inc.com
Yahoo! Labs
Tomer Koren tomerk@technion.ac.il
Technion

We study exploration in Multi-Armed Bandits (MAB) in a setting where~k players collaborate in order to identify an ϵ-optimal arm. Our motivation comes from recent employment of MAB algorithms in computationally intensive, large-scale applications. Our results demonstrate a non-trivial tradeoff between the number of arm pulls required by each of the players, and the amount of communication between them. In particular, our main result shows that by allowing the k players to communicate *only once*, they are able to learn $\sqrt{k}$ times faster than a single player. That is, distributing learning to k players gives rise to a factor ~$\sqrt{k}$ parallel speed-up. We complement this result with a lower bound showing this is in general the best possible. On the other extreme, we present an algorithm that achieves the ideal factor k speed-up in learning performance, with communication only logarithmic in~$1/\epsilon$.

Su16 High-Dimensional Gaussian Process Bandits

Josip Djolonga josipd@student.ethz.ch
Andreas Krause krausea@ethz.ch
ETH Zurich
Volkan Cevher volkan.cevher@epfl.ch
EPFL

Many applications in machine learning require optimizing unknown functions defined over a high-dimensional space from noisy samples that are expensive to obtain. We address this notoriously hard challenge, under the assumptions that the function varies only along some low-dimensional subspace and is smooth (i.e., it has a low norm in a Reproducible Kernel Hilbert Space). In particular, we present the SI-BO algorithm, which leverages recent low-rank matrix recovery techniques to learn the underlying subspace of the unknown function and applies Gaussian Process Upper Confidence sampling for optimization of the function. We carefully calibrate the exploration–exploitation tradeoff by allocating sampling budget to subspace estimation and function optimization, and obtain the first subexponential cumulative regret bounds and convergence rates for Bayesian optimization in high-dimensions under noisy observations. Numerical results demonstrate the effectiveness of our approach in difficult scenarios.

Su17 On Poisson Graphical Models

Eunho Yang eunho@cs.utexas.edu
Pradeep Ravikumar pradeepr@cs.utexas.edu
UT Austin
Genevera Allen gallen@rice.edu
Rice University
Zhandong Liu zhandonl@bcm.edu
Baylor College of Medicine

Undirected graphical models, such as Gaussian graphical models, Ising, and multinomial/categorical graphical models, are widely used in a variety of applications for modeling distributions over a large number of variables. These standard instances, however, are ill-suited to modeling count data, which are increasingly ubiquitous in big-data settings such as genomic sequencing data, user-ratings data, spatial incidence data, climate studies, and site visits. Existing classes of Poisson graphical models, which arise as the joint distributions that correspond to Poisson distributed node-conditional distributions, have a major drawback: they can only model negative conditional dependencies for reasons of normalizability given its infinite domain. In this paper, our objective is to modify the Poisson graphical model distribution so that it can capture a rich dependence structure between count-valued variables. We begin by discussing two strategies for truncating the Poisson distribution and show that only one of these leads to a valid joint distribution; even this model, however, has limitations on the types of variables and dependencies that may be modeled. To address this, we propose two novel variants of the Poisson distribution and their corresponding joint graphical model distributions. These models provide a class of Poisson graphical models that can capture both positive and negative conditional dependencies between count-valued variables. One can learn the graph structure of our model via penalized neighborhood selection, and we demonstrate the performance of our methods by learning simulated networks as well as a network from microRNA-Sequencing data.

Su18 Conditional Random Fields via Univariate Exponential Families

Eunho Yang eunho@cs.utexas.edu
Pradeep Ravikumar pradeepr@cs.utexas.edu
UT Austin
Genevera Allen gallen@rice.edu
Rice University
Zhandong Liu zhandonl@bcm.edu
Baylor College of Medicine

Conditional random fields, which model the distribution of a multivariate response conditioned on a set of covariates using undirected graphs, are widely used in a variety of multivariate prediction applications. Popular instances of this class of models such as categorical-discrete CRFs, Ising CRFs, and conditional Gaussian based CRFs, are not however best suited to the varied types of response variables in many applications, including count-valued responses. We thus introduce a “novel subclass of CRFs”, derived by imposing node-wise conditional distributions of response variables conditioned on the rest of the responses and the covariates as arising from univariate exponential families. This allows us to derive novel multivariate CRFs given any univariate exponential distribution, including the Poisson, negative binomial, and exponential distributions. Also in particular, it addresses the common CRF problem of specifying “feature'' functions determining the interactions between response variables and covariates. We develop a class of tractable penalized M-estimators to learn these CRF distributions from data, as well as a unified sparsistency analysis for this general class of CRFs showing exact structure recovery can be achieved with high probability.

Su19 Scalable kernels for graphs with continuous attributes

Aasa Feragen aasa.feragen@tuebingen.mpg.de
Niklas Kasenburg
niklas.kasenburg@tuebingen.mpg.de
MPI Tübingen & University of Copenhagen
Jens Petersen phup@diku.dk
University of Copenhagen
Marleen de Bruijne marleen@diku.dk
Erasmus MC
Karsten Borgwardt
karsten.borgwardt@tuebingen.mpg.de
MPI Tübingen & University of Tübingen

While graphs with continuous node attributes arise in many applications, state-of-the-art graph kernels for comparing continuous-attributed graphs suffer from a high runtime complexity; for instance, the popular shortest path kernel scales as $\mathcal{O}(n^4)$, where n is the number of nodes. In this paper, we present a class of path kernels with computational complexity $\mathcal{O}(n^2(m+\delta^2))$, where δ is the graph diameter and m the number of edges. Due to the sparsity and small diameter of real-world graphs, these kernels scale comfortably to large graphs. In our experiments, the presented kernels outperform state-of-the-art kernels in terms of speed and accuracy on classification benchmark datasets.

Su20 Near-optimal Anomaly Detection in Graphs using Lovasz Extended Scan Statistic

James Sharpnack jsharpna@cs.cmu.edu
Akshay Krishnamurthy akshaykr@cs.cmu.edu
Aarti Singh aartisingh@cmu.edu
CMU

The detection of anomalous activity in graphs is a statistical problem that arises in many applications, such as network surveillance, disease outbreak detection, and activity monitoring in social networks. Beyond its wide applicability, graph structured anomaly detection serves as a case study in the difficulty of balancing computational complexity with statistical power. In this work, we develop from first principles the generalized likelihood ratio test for determining if there is a well connected region of activation over the vertices in the graph in Gaussian noise. Because this test is computationally infeasible, we provide a relaxation, called the Lov\’asz extended scan statistic (LESS) that uses submodularity to approximate the intractable generalized likelihood ratio. We demonstrate a connection between LESS and maximum a-posteriori inference in Markov random fields, which provides us with a poly-time algorithm for LESS. Using electrical network theory, we are able to control type 1 error for LESS and prove conditions under which LESS is risk consistent. Finally, we consider specific graph models, the torus, k-nearest neighbor graphs, and ϵ-random graphs. We show that on these graphs our results provide near-optimal performance by matching our results to known lower bounds.

Su21 Analyzing the Harmonic Structure in Graph-Based Learning

Xiao-Ming Wu xmwu@ee.columbia.edu
Zhenguo Li zhenguol@gmail.com
Shih-Fu Chang sfchang@ee.columbia.edu
Columbia University

We show that either explicitly or implicitly, various well-known graph-based models exhibit a common significant *harmonic* structure in its target function -- the value of a vertex is approximately the weighted average of the values of its adjacent neighbors. Understanding of such structure and analysis of the loss defined over such structure help reveal important properties of the target function over a graph. In this paper, we show that the variation of the target function across a cut can be upper and lower bounded by the ratio of its harmonic loss and the cut cost. We use this to develop an analytical tool and analyze 5 popular models in graph-based learning: absorbing random walks, partially absorbing random walks, hitting times, pseudo-inverse of graph Laplacian, and eigenvectors of the Laplacian matrices. Our analysis well explains several open questions of these models reported in the literature. Furthermore, it provides theoretical justifications and guidelines for their practical use. Simulations on synthetic and real datasets support our analysis.

Su22 Robust Multimodal Graph Matching: Sparse Coding Meets Graph Matching

Marcelo Fiori mfiori@fing.edu.uy
Pablo Muse pmuse@fing.edu.uy
Universidad de la República, Uruguay
Pablo Sprechmann pablo.sprechmann@duke.edu
jovo Vogelstein jv.work@jhu.edu
Guillermo Sapiro guillermo.sapiro@duke.edu
Duke University

Graph matching is a challenging problem with very important applications in a wide range of fields, from image and video analysis to biological and biomedical problems. We propose a robust graph matching algorithm inspired in sparsity-related techniques. We cast the problem, resembling group or collaborative sparsity formulations, as a non-smooth convex optimization problem that can be efficiently solved using augmented Lagrangian techniques. The method can deal with weighted or unweighted graphs, as well as multimodal data, where different graphs represent different types of data. The proposed approach is also naturally integrated with collaborative graph inference techniques, solving general network inference problems where the observed variables, possibly coming from different modalities, are not in correspondence. The algorithm is tested and compared with state-of-the-art graph matching techniques in both synthetic and real graphs. We also present results on multimodal graphs and applications to collaborative inference of brain connectivity from alignment-free functional magnetic resonance imaging (fMRI) data.

Su23 Learning Gaussian Graphical Models with Observed or Latent FVSs

Ying Liu liu_ying@mit.edu
Alan Willsky willsky@mit.edu
Massachusetts Institute of Technology

Gaussian Graphical Models (GGMs) or Gauss Markov random fields are widely used in many applications, and the trade-off between the modeling capacity and the efficiency of learning and inference has been an important research problem. In this paper, we study the family of GGMs with small feedback vertex sets (FVSs), where an FVS is a set of nodes whose removal breaks all the cycles. Exact inference such as computing the marginal distributions and the partition function has complexity $O(k^2n)$ using message-passing algorithms, where k is the size of the FVS, and n is the total number of nodes. We propose efficient structure learning algorithms for two cases: 1) All nodes are observed, which is useful in modeling social or flight networks where the FVS nodes often correspond to a small number of high-degree nodes, or hubs, while the rest of the networks is modeled by a tree. Regardless of the maximum degree, without knowing the full graph structure, we can exactly compute the maximum likelihood estimate in $O(kn^2+n^2\log n)$ if the FVS is known or in polynomial time if the FVS is unknown but has bounded size. 2) The FVS nodes are latent variables, where structure learning is equivalent to decomposing a inverse covariance matrix (exactly or approximately) into the sum of a tree-structured matrix and a low-rank matrix. By incorporating efficient inference into the learning steps, we can obtain a learning algorithm using alternating low-rank correction with complexity $O(kn^2+n^2\log n)$ per iteration. We also perform experiments using both synthetic data as well as real data of flight delays to demonstrate the modeling capacity with FVSs of various sizes. We show that empirically the family of GGMs of size $O(\log n)$ strikes a good balance between the modeling capacity and the efficiency.

Su24 Global MAP-Optimality by Shrinking the Combinatorial Search Area with Convex Relaxation

Bogdan Savchynskyy
bogdan.savchynskyy@iwr.uni-heidelberg.de
Jörg Hendrik Kappes kappes@math.uni-heidelberg.de
Paul Swoboda swoboda@math.uni-heidelberg.de
Christoph Schnörr schnoerr@math.uni-heidelberg.de
University of Heidelberg

We consider energy minimization for undirected graphical models, also known as MAP-inference problem for Markov random fields. Although combinatorial methods, which return a provably optimal integral solution of the problem, made a big progress in the past decade, they are still typically unable to cope with large-scale datasets. On the other hand, large scale datasets are typically defined on sparse graphs, and convex relaxation methods, such as linear programming relaxations often provide good approximations to integral solutions. We propose a novel method of combining combinatorial and convex programming techniques to obtain a global solution of the initial combinatorial problem. Based on the information obtained from the solution of the convex relaxation, our method confines application of the combinatorial solver to a small fraction of the initial graphical model, which allows to optimally solve big problems. We demonstrate the power of our approach on a computer vision energy minimization benchmark.

Su25 First-order Decomposition Trees

Nima Taghipour nima.taghipour@cs.kuleuven.be
Jesse Davis jesse.davis@cs.kuleuven.be
Hendrik Blockeel hendrik.blockeel@cs.kuleuven.be
KU Leuven

Lifting attempts to speedup probabilistic inference by exploiting symmetries in the model. Exact lifted inference methods, like their propositional counterparts, work by recursively decomposing the model and the problem. In the propositional case, there exist formal structures, such as decomposition trees (dtrees), that represent such a decomposition and allow us to determine the complexity of inference a priori. However, there is currently no equivalent structure nor analogous complexity results for lifted inference. In this paper, we introduce FO-dtrees, which upgrade propositional dtrees to the first-order level. We show how these trees can characterize a lifted inference solution for a probabilistic logical model (in terms of a sequence of lifted operations), and make a theoretical analysis of the complexity of lifted inference in terms of the novel notion of lifted width for the tree.

Su26 Binary to Bushy: Bayesian Hierarchical Clustering with the Beta Coalescent

Yuening Hu ynhu@cs.umd.edu
Jordan Boyd-Graber jbg@umiacs.umd.edu
Hal Daume III hal@umiacs.umd.edu
University of Maryland
Z. Irene Ying zhu.ying@ars.usda.gov
US Department of Agriculture

Discovering hierarchical regularities in data is a key problem in interacting with large datasets, modeling cognition, and encoding knowledge. A previous Bayesian solution---Kingman's coalescent---provides a convenient probabilistic model for data represented as a binary tree. Unfortunately, this is inappropriate for data better described by bushier trees. We generalize an existing belief propagation framework of Kingman's coalescent to the beta coalescent, which models a wider range of tree structures. Because of the complex combinatorial search over possible structures, we develop new sampling schemes using sequential Monte Carlo and Dirichlet process mixture models, which render inference efficient and tractable. We present results on both synthetic and real data that show the beta coalescent outperforms Kingman's coalescent on real datasets and is qualitatively better at capturing data in bushy hierarchies.

Su27 Parallel Sampling of DP Mixture Models using Sub-Cluster Splits

Jason Chang jchang7@csail.mit.edu
John Fisher III fisher@csail.mit.edu
Massachusetts Institute of Technology

We present a novel MCMC sampler for Dirichlet process mixture models that can be used for conjugate or non-conjugate prior distributions. The proposed sampler can be massively parallelized to achieve significant computational gains. A non-ergodic restricted Gibbs iteration is mixed with split/merge proposals to produce a valid sampler. Each regular cluster is augmented with two sub-clusters to construct likely split moves. Unlike many previous parallel samplers, the proposed sampler accurately enforces the correct stationary distribution of the Markov chain without the need for approximate models. Empirical results illustrate that the new sampler exhibits better convergence properties than current methods.

Su28 Lexical and Hierarchical Topic Regression

Viet-An Nguyen vietan@cs.umd.edu
Jordan Boyd-Graber jbg@umiacs.umd.edu
Philip Resnik resnik@umd.edu
University of Maryland

Inspired by a two-level theory that unifies agenda setting and ideological framing, we propose supervised hierarchical latent Dirichlet allocation (SHLDA) which jointly captures documents' multi-level topic structure and their polar response variables. Our model extends the nested Chinese restaurant process to discover a tree-structured topic hierarchy and uses both per-topic hierarchical and per-word lexical regression parameters to model the response variables. Experiments in a political domain and on sentiment analysis tasks show that SHLDA improves predictive accuracy while adding a new dimension of insight into how topics under discussion are framed.

Su29 A Novel Two-Step Method for Cross Language Representation Learning

Min Xiao minxiao@temple.edu
Yuhong Guo yuhong@temple.edu
Temple University

Cross language text classification is an important learning task in natural language processing. A critical challenge of cross language learning lies in that words of different languages are in disjoint feature spaces. In this paper, we propose a two-step representation learning method to bridge the feature spaces of different languages by exploiting a set of parallel bilingual documents. Specifically, we first formulate a matrix completion problem to produce a complete parallel document-term matrix for all documents in two languages, and then induce a cross-lingual document representation by applying latent semantic indexing on the obtained matrix. We use a projected gradient descent algorithm to solve the formulated matrix completion problem with convergence guarantees. The proposed approach is evaluated by conducting a set of experiments with cross language sentiment classification tasks on Amazon product reviews. The experimental results demonstrate that the proposed learning approach outperforms a number of comparison cross language representation learning methods, especially when the number of parallel bilingual documents is small.

Su30 Learning word embeddings efficiently with noise-contrastive estimation

Andriy Mnih amnih@gatsby.ucl.ac.uk
Gatsby Unit, UCL
koray kavukcuoglu koray@kavukcuoglu.org
NEC Labs

Continuous-valued word embeddings learned by neural language models have recently been shown to capture semantic and syntactic information about words very well, setting performance records on several word similarity tasks. The best results are obtained by learning high-dimensional embeddings from very large quantities of data, which makes scalability of the training method a critical factor. We propose a simple and scalable new approach to learning word embeddings based on training log-bilinear models with noise-contrastive estimation. Our approach is simpler, faster, and produces better results than the current state-of-the art method of Mikolov et al. (2013a). We achieve results comparable to the best ones reported, which were obtained on a cluster, using four times less data and more than an order of magnitude less computing time. We also investigate several model types and find that the embeddings learned by the simpler models perform at least as well as those learned by the more complex ones.

Su31 Training and Analysing Deep Recurrent Neural Networks

Michiel Hermans michiel.hermans@ugent.be
Benjamin Schrauwen Benjamin.Schrauwen@ugent.be
Ghent University

Time series often have a temporal hierarchy, with information that is spread out over multiple time scales. Common recurrent neural networks, however, do not explicitly accommodate such a hierarchy, and most research on them has been focusing on training algorithms rather than on their basic architecture. In this pa- per we study the effect of a hierarchy of recurrent neural networks on processing time series. Here, each layer is a recurrent network which receives the hidden state of the previous layer as input. This architecture allows us to perform hierarchical processing on difficult temporal tasks, and more naturally capture the structure of time series. We show that they reach state-of-the-art performance for recurrent networks in character-level language modelling when trained with sim- ple stochastic gradient descent. We also offer an analysis of the different emergent time scales.

Su32 Extracting regions of interest from biological images with convolutional sparse block coding

Marius Pachitariu marius@gatsby.ucl.ac.uk
Maneesh Sahani maneesh@gatsby.ucl.ac.uk
Gatsby Unit, UCL
Adam Packer a.packer@ucl.ac.uk
Noah Pettit noah.pettit.10@ucl.ac.uk
Henry Dalgleish hwpdalgleish@gmail.com
Michael Hausser m.hausser@ucl.ac.uk
UCL

Biological tissue is often composed of cells with similar morphologies replicated throughout large volumes and many biological applications rely on the accurate identification of these cells and their locations from image data. Here we develop a generative model that captures the regularities present in images composed of repeating elements of a few different types. Formally, the model can be described as convolutional sparse block coding. For inference we use a variant of convolutional matching pursuit adapted to block-based representations. We extend the K-SVD learning algorithm to subspaces by retaining several principal vectors from the SVD decomposition instead of just one. Good models with little cross-talk between subspaces can be obtained by learning the blocks incrementally. We perform extensive experiments on simulated images and the inference algorithm consistently recovers a large proportion of the cells with a small number of false positives. We fit the convolutional model to noisy GCaMP6 two-photon images of spiking neurons and to Nissl-stained slices of cortical tissue and show that it recovers cell body locations without supervision. The flexibility of the block-based representation is reflected in the variability of the recovered cell shapes.

Su33 Mapping paradigm ontologies to and from the brain

Yannick Schwartz yannick.schwartz@inria.fr
Bertrand Thirion bertrand.thirion@inria.fr
Gael Varoquaux gael.varoquaux@inria.fr
INRIA

Imaging neuroscience links brain activation maps to behavior and cognition via correlational studies. Due to the nature of the individual experiments, based on eliciting neural response from a small number of stimuli, this link is incomplete, and unidirectional from the causal point of view. To come to conclusions on the function implied by the activation of brain regions, it is necessary to combine a wide exploration of the various brain functions and some inversion of the statistical inference. Here we introduce a methodology for accumulating knowledge towards a bidirectional link between observed brain activity and the corresponding function. We rely on a large corpus of imaging studies and a predictive engine. Technically, the challenges are to find commonality between the studies without denaturing the richness of the corpus. The key elements that we contribute are labeling the tasks performed with a cognitive ontology, and modeling the long tail of rare paradigms in the corpus. To our knowledge, our approach is the first demonstration of predicting the cognitive content of completely new brain images. To that end, we propose a method that predicts the experimental paradigms across different studies.

Su34 Speeding up Permutation Testing in Neuroimaging

Chris Hinrichs hinrichs@cs.wisc.edu
Vamsi Ithapu ithapu@wisc.edu
Qinyuan Sun qSu28@wisc.edu
Sterling Johnson scj@medicine.wisc.edu
Vikas Singh vsingh@biostat.wisc.edu
UW-Madison

Multiple hypothesis testing is a significant problem in nearly all neuroimaging studies. In order to correct for this phenomena, we require a reliable estimate of the Family-Wise Error Rate (FWER). The well known Bonferroni correction method, while being simple to implement, is quite conservative, and can substantially under-power a study because it ignores dependencies between test statistics. Permutation testing, on the other hand, is an exact, non parametric method of estimating the FWER for a given α threshold, but for acceptably low thresholds the computational burden can be prohibitive. In this paper, we observe that permutation testing in fact amounts to populating the columns of a very large matrix P. By analyzing the spectrum of this matrix, under certain conditions, we see that P has a low-rank plus a low-variance residual decomposition which makes it suitable for highly sub–sampled — on the order of 0.5% — matrix completion methods. Thus, we propose a novel permutation testing methodology which offers a large speedup, without sacrificing the fidelity of the estimated FWER. Our valuations on four different neuroimaging datasets show that a computational speedup factor of roughly 50× can be achieved while recovering the FWER distribution up to very high accuracy. Further, we show that the estimated α threshold is also recovered faithfully, and is stable.

Su35 BIG & QUIC: Sparse Inverse Covariance Estimation for a Million Variables

Cho-Jui Hsieh cjhsieh@cs.utexas.edu
Pradeep Ravikumar pradeepr@cs.utexas.edu
UT Austin
Matyas Sustik msustik@gmail.com
Inderjit Dhillon inderjit@cs.utexas.edu
Russell Poldrack poldrack@utexas.edu
University of Texas

The l1-regularized Gaussian maximum likelihood estimator (MLE) has been shown to have strong statistical guarantees in recovering a sparse inverse covariance matrix even under high-dimensional settings. However, it requires solving a difficult non-smooth log-determinant program with number of parameters scaling quadratically with the number of Gaussian variables. State-of-the-art methods thus do not scale to problems with more than 20,000 variables. In this paper, we develop an algorithm BigQUIC, which can solve 1 million dimensional l1-regularized Gaussian MLE problems (which would thus have 1000 billion parameters) using a single machine, with bounded memory. In order to do so, we carefully exploit the underlying structure of the problem. Our innovations include a novel block-coordinate descent method with the blocks chosen via a clustering scheme to minimize repeated computations; and allowing for inexact computation of specific components. In spite of these modifications, we are able to theoretically analyze our procedure and show that BigQUIC can achieve super-linear or even quadratic convergence rates.

Su36 Geometric optimisation on positive definite matrices for elliptically contoured distributions

Suvrit Sra suvrit@gmail.com
MPI for Intelligent Systems & CMU
Reshad Hosseini hosseini@tuebingen.mpg.de
MPI Tübingen

Hermitian positive definite matrices (HPD) recur throughout statistics and machine learning. In this paper we develop *geometric optimisation* for globally optimising certain nonconvex loss functions arising in the modelling of data via elliptically contoured distributions (ECDs). We exploit the remarkable structure of the convex cone of positive definite matrices which allows one to uncover hidden geodesic convexity of objective functions that are nonconvex in the ordinary Euclidean sense. Going even beyond manifold convexity we show how further metric properties of HPD matrices can be exploited to globally optimise several ECD log-likelihoods that are not even geodesic convex. We present key results that help recognise this geometric structure, as well as obtain efficient fixed-point algorithms to optimise the corresponding objective functions. To our knowledge, ours are the most general results on geometric optimisation of HPD matrices known so far. Experiments reveal the benefits of our approach---it avoids any eigenvalue computations which makes it very competitive.

Su37 Estimating the Unseen: Improved Estimators for Entropy and other Properties

Paul Valiant pvaliant@gmail.com
Brown University
Gregory Valiant gregory.valiant@gmail.com
Stanford University

Recently, [Valiant and Valiant] showed that a class of distributional properties, which includes such practically relevant properties as entropy, the number of distinct elements, and distance metrics between pairs of distributions, can be estimated given a SUBLINEAR sized sample. Specifically, given a sample consisting of independent draws from any distribution over at most n distinct elements, these properties can be estimated accurately using a sample of size $O(n/\log n)$. We propose a novel modification of this approach and show: 1) theoretically, our estimator is optimal (to constant factors, over worst-case instances), and 2) in practice, it performs exceptionally well for a variety of estimation tasks, on a variety of natural distributions, for a wide range of parameters. Perhaps unsurprisingly, the key step in this approach is to first use the sample to characterize the “unseen” portion of the distribution. This goes beyond such tools as the Good-Turing frequency estimation scheme, which estimates the total probability mass of the unobserved portion of the distribution: we seek to estimate the “shape” of the unobserved portion of the distribution. This approach is robust, general, and theoretically principled; we expect that it may be fruitfully used as a component within larger machine learning and data analysis systems.

Su38 Factorized Asymptotic Bayesian Inference for Latent Feature Models

Kohei Hayashi hayashi.kohei@gmail.com
NII
Ryohei Fujimaki rfujimaki@nec-labs.com
NEC Labs America

This paper extends factorized asymptotic Bayesian (FAB) inference for latent feature models (LFMs). FAB inference has not been applicable to models, including LFMs, without a specific condition on the Hesqsian matrix of a complete log-likelihood, which is required to derive a “factorized information criterion’’ (FIC). Our asymptotic analysis of the Hessian matrix of LFMs shows that FIC of LFMs has the same form as those of mixture models. FAB/LFMs have several desirable properties (e.g., automatic hidden states selection and parameter identifiability) and empirically perform better than state-of-the-art Indian Buffet processes in terms of model selection, prediction, and computational efficiency.

Su39 Tracking Time-varying Graphical Structure

Erich Kummerfelde kummerfeld@gmail.com
David Danks ddanks@cmu.edu
CMU

Structure learning algorithms for graphical models have focused almost exclusively on stable environments in which the underlying generative process does not change; that is, they assume that the generating model is globally stationary. In real-world environments, however, such changes often occur without warning or signal. Real-world data often come from generating models that are only locally stationary. In this paper, we present LoSST, a novel, heuristic structure learning algorithm that tracks changes in graphical model structure or parameters in a dynamic, real-time manner. We show by simulation that the algorithm performs comparably to batch-mode learning when the generating graphical structure is globally stationary, and significantly better when it is only locally stationary.

Su40 Sparse Inverse Covariance Estimation with Calibration

Tuo Zhao tzhao5@jhu.edu
Johns Hopkins University
Han Liu hanliu@princeton.edu
Princeton University

We propose a semiparametric procedure for estimating high dimensional sparse inverse covariance matrix. Our method, named ALICE, is applicable to the elliptical family. Computationally, we develop an efficient dual inexact iterative projection (D_2P) algorithm based on the alternating direction method of multipliers (ADMM). Theoretically, we prove that the ALICE estimator achieves the parametric rate of convergence in both parameter estimation and model selection. Moreover, ALICE calibrates regularizations when estimating each column of the inverse covariance matrix. So it not only is asymptotically tuning free, but also achieves an improved finite sample performance. We present numerical simulations to support our theory, and a real data example to illustrate the effectiveness of the proposed estimator.

Su41 A* Lasso for Learning a Sparse Bayesian Network Structure for Continuous Variables

Jing Xiang jingx@cs.cmu.edu
Seyoung Kim sssykim@cs.cmu.edu
CMU

We address the problem of learning a sparse Bayesian network structure for continuous variables in a high-dimensional space. The constraint that the estimated Bayesian network structure must be a directed acyclic graph (DAG) makes the problem challenging because of the huge search space of network structures. Most previous methods were based on a two-stage approach that prunes the search space in the first stage and then searches for a network structure that satisfies the DAG constraint in the second stage. Although this approach is effective in a low-dimensional setting, it is difficult to ensure that the correct network structure is not pruned in the first stage in a high-dimensional setting. In this paper, we propose a single-stage method, called A* lasso, that recovers the optimal sparse Bayesian network structure by solving a single optimization problem with A* search algorithm that uses lasso in its scoring system. Our approach substantially improves the computational efficiency of the well-known exact methods based on dynamic programming. We also present a heuristic scheme that further improves the efficiency of A* lasso without significantly compromising the quality of solutions and demonstrate this on benchmark Bayesian networks and real data.

Su42 On model selection consistency of penalized M-estimators: a geometric theory

Jason Lee jdl17@stanford.edu
Yuekai Sun yuekai@stanford.edu
Jonathan Taylor jonathan.taylor@stanford.edu
Stanford University

Penalized M-estimators are used in diverse areas of science and engineering to fit high-dimensional models with some low-dimensional structure. Often, the penalties are *geometrically decomposable,* \ie\ can be expressed as a sum of (convex) support functions. We generalize the notion of irrepresentable to geometrically decomposable penalties and develop a general framework for establishing consistency and model selection consistency of M-estimators with such penalties. We then use this framework to derive results for some special cases of interest in bioinformatics and statistical learning.

Su43 Generalizing Analytic Shrinkage for Arbitrary Covariance Structures

Daniel Bartz daniel.bartz@tu-berlin.de
Klaus-Robert Müller
Klaus-Robert.Mueller@tu-berlin.de
TU Berlin

Analytic shrinkage is a statistical technique that offers a fast alternative to cross-validation for the regularization of covariance matrices and has appealing consistency properties. We show that the proof of consistency implies bounds on the growth rates of eigenvalues and their dispersion, which are often violated in data. We prove consistency under assumptions which do not restrict the covariance structure and therefore better match real world data. In addition, we propose an extension of analytic shrinkage --orthogonal complement shrinkage-- which adapts to the covariance structure. Finally we demonstrate the superior performance of our novel approach on data from the domains of finance, spoken letter and optical character recognition, and neuroscience.

Su44 Robust Spatial Filtering with Beta Divergence

Wojciech Samek wojciech.samek@tu-berlin.de
Duncan Blythe duncan.blythe@bccn-berlin.de
Klaus-Robert Müller
Klaus-Robert.Mueller@tu-berlin.de
TU Berlin
Motoaki Kawanabe kawanabe@atr.jp
ATR

The efficiency of Brain-Computer Interfaces (BCI) largely depends upon a reliable extraction of informative features from the high-dimensional EEG signal. A crucial step in this protocol is the computation of spatial filters. The Common Spatial Patterns (CSP) algorithm computes filters that maximize the difference in band power between two conditions, thus it is tailored to extract the relevant information in motor imagery experiments. However, CSP is highly sensitive to artifacts in the EEG data, i.e. few outliers may alter the estimate drastically and decrease classification performance. Inspired by concepts from the field of information geometry we propose a novel approach for robustifying CSP. More precisely, we formulate CSP as a divergence maximization problem and utilize the property of a particular type of divergence, namely beta divergence, for robustifying the estimation of spatial filters in the presence of artifacts in the data. We demonstrate the usefulness of our method on toy data and on EEG recordings from 80 subjects.

Su45 A multi-agent control framework for co-adaptation in brain-computer interfaces

Josh Merel jsm2183@columbia.edu
Tony Jebara jebara@cs.columbia.edu
Liam Paninski liam@stat.columbia.edu
Columbia University
Roy Fox royf@cs.huji.ac.il
Hebrew University

In a closed-loop brain-computer interface (BCI), adaptive decoders are used to learn parameters suited to decoding the user's neural response. Feedback to the user provides information which permits the neural tuning to also adapt. We present an approach to model this process of co-adaptation between the encoding model of the neural signal and the decoding algorithm as a multi-agent formulation of the linear quadratic Gaussian (LQG) control problem. In simulation we characterize how decoding performance improves as the neural encoding and adaptive decoder optimize, qualitatively resembling experimentally demonstrated closed-loop improvement. We then propose a novel, modified decoder update rule which is aware of the fact that the encoder is also changing and show it can improve simulated co-adaptation dynamics. Our modeling approach offers promise for gaining insights into co-adaptation as well as improving user learning of BCI control in practical settings.

Su46 Probabilistic Movement Primitives

Alexandros Paraschos paraschos@ias.tu-darmstadt.de
Christian Daniel daniel@ias.tu-darmstadt.de
Jan Peters mail@jan-peters.net
Gerhard Neumann neumann@ias.tu-darmstadt.de
TU Darmstadt

Movement Primitives (MP) are a well-established approach for representing modular and re-usable robot movement generators. Many state-of-the-art robot learning successes are based MPs, due to their compact representation of the inherently continuous and high dimensional robot movements. A major goal in robot learning is to combine multiple MPs as building blocks in a modular control architecture to solve complex tasks. To this effect, a MP representation has to allow for blending between motions, adapting to altered task variables, and co-activating multiple MPs in parallel. We present a probabilistic formulation of the MP concept that maintains a distribution over trajectories. Our probabilistic approach allows for the derivation of new operations which are essential for implementing all aforementioned properties in one framework. In order to use such a trajectory distribution for robot movement control, we analytically derive a stochastic feedback controller which reproduces the given trajectory distribution. We evaluate and compare our approach to existing methods on several simulated as well as real robot scenarios.

Su47 Variational Policy Search via Trajectory Optimization

Sergey Levine svlevine@stanford.edu
Stanford University
Vladlen Koltun vladlen@stanford.edu
Adobe Research

In order to learn effective control policies for dynamical systems, policy search methods must be able to discover successful executions of the desired task. While random exploration can work well in simple domains, complex and high-dimensional tasks present a serious challenge, particularly when combined with high-dimensional policies that make parameter-space exploration infeasible. We present a method that uses trajectory optimization as a powerful exploration strategy that guides the policy search. A variational decomposition of a maximum likelihood policy objective allows us to use standard trajectory optimization algorithms such as differential dynamic programming, interleaved with standard supervised learning for the policy itself. We demonstrate that the resulting algorithm can outperform prior methods on two challenging locomotion tasks.

Su48 Learning from Limited Demonstrations

Beomjoon Kim beomjoon.kim0@gmail.com
Amir massoud Farahmand amirf@ualberta.ca
Joelle Pineau jpineau@cs.mcgill.ca
Doina Precup dprecup@cs.mcgill.ca
McGill University

We propose an approach to learning from demonstration (LfD) which leverages expert data, even if the expert examples are very few or inaccurate. We achieve this by integrating LfD in an approximate policy iteration algorithm. The key idea of our approach is that expert examples are used to generate linear constraints on the optimization, in a similar fashion to large-margin classification. We prove an upper bound on the true Bellman error of the approximation computed by the algorithm at each iteration. We show empirically that the algorithm outperforms both pure policy iteration, as well as DAgger (a state-of-art LfD algorithm) and supervised learning in a variety of scenarios, including when very few and/or imperfect demonstrations are available. Our experiments include simulations as well as a real robotic navigation task.

Su49 Learning Trajectory Preferences for Manipulators via Iterative Improvement

Ashesh Jain ashesh@cs.cornell.edu
Brian Wojcik bmw75@cornell.edu
Thorsten Joachims tj@cs.cornell.edu
Ashutosh Saxena asaxena@cs.cornell.edu
Cornell University

We consider the problem of learning good trajectories for manipulation tasks. This is challenging because the criterion defining a good trajectory varies with users, tasks and environments. In this paper, we propose a co-active online learning framework for teaching robots the preferences of its users for object manipulation tasks. The key novelty of our approach lies in the type of feedback expected from the user: the human user does not need to demonstrate optimal trajectories as training data, but merely needs to iteratively provide trajectories that slightly improve over the trajectory currently proposed by the system. We argue that this co-active preference feedback can be more easily elicited from the user than demonstrations of optimal trajectories, which are often challenging and non-intuitive to provide on high degrees of freedom manipulators. Nevertheless, theoretical regret bounds of our algorithm match the asymptotic rates of optimal trajectory algorithms. We also formulate a score function to capture the contextual information and demonstrate the generalizability of our algorithm on a variety of household tasks, for whom, the preferences were not only influenced by the object being manipulated but also by the surrounding environment.

Su50 Forgetful Bayes and myopic planning: Human learning and decision-making in a bandit setting

Shunan Zhang s6zhang@ucsd.edu
Angela Yu ajyu@ucsd.edu
UC San Diego

How humans achieve long-term goals in an uncertain environment, via repeated trials and noisy observations, is an important problem in cognitive science. We investigate this behavior in the context of a multi-armed bandit task. We compare human behavior to a variety of models that vary in their representational and computational complexity. Our result shows that subjects' choices, on a trial-to-trial basis, are best captured by a "forgetful" Bayesian iterative learning model in combination with a partially myopic decision policy known as Knowledge Gradient. This model accounts for subjects' trial-by-trial choice better than a number of other previously proposed models, including optimal Bayesian learning and risk minimization, epsilon-greedy and win-stay-lose-shift. It has the added benefit of being closest in performance to the optimal Bayesian model than all the other heuristic models that have the same computational complexity (all are significantly less complex than the optimal model). These results constitute an advancement in the theoretical understanding of how humans negotiate the tension between exploration and exploitation in a noisy, imperfectly known environment.

Su51 Context-sensitive active sensing in humans

Sheeraz Ahmad sahmad@cs.ucsd.edu
He Huang heh001@ucsd.edu
Angela Yu ajyu@ucsd.edu
UC San Diego

Humans and animals readily utilize active sensing, or the use of self-motion, to focus sensory and cognitive resources on the behaviorally most relevant stimuli and events in the environment. Understanding the computational basis of natural active sensing is important both for advancing brain sciences and for developing more powerful artificial systems. Recently, a goal-directed, context-sensitive, Bayesian control strategy for active sensing, termed C-DAC (Context-Dependent Active Controller), was proposed (Ahmad & Yu, 2013). In contrast to previously proposed algorithms for human active vision, which tend to optimize abstract statistical objectives and therefore cannot adapt to changing behavioral context or task goals, C-DAC directly minimizes behavioral costs and thus, automatically adapts itself to different task conditions. However, C-DAC is limited as a model of human active sensing, given its computational/representational requirements, especially for more complex, real-world situations. Here, we propose a myopic approximation to C-DAC, which also takes behavioral costs into account, but achieves a significant reduction in complexity by looking only one step ahead. We also present data from a human active visual search experiment, and compare the performance of the various models against human behavior. We find that C-DAC and its myopic variant both achieve better fit to human data than Infomax (Butko & Movellan, 2010), which maximizes expected cumulative future information gain. In summary, this work provides novel experimental results that differentiate theoretical models for human active sensing, as well as a novel active sensing algorithm that retains the context-sensitivity of the optimal controller while achieving significant computational savings.

Su52 Bellman Error Based Feature Generation using Random Projections on Sparse Spaces

Mahdi Milani Fard mahdi.milanifard@mail.mcgill.ca
Yuri Grinberg yuri.grinberg@mail.mcgill.ca
Amir massoud Farahmand amirf@ualberta.ca
Joelle Pineau jpineau@cs.mcgill.ca
Doina Precup dprecup@cs.mcgill.ca
McGill University

This paper addresses the problem of automatic generation of features for value function approximation in reinforcement learning. Bellman Error Basis Functions (BEBFs) have been shown to improve the error of policy evaluation with function approximation, with a convergence rate similar to that of value iteration. We propose a simple, fast and robust algorithm based on random projections, which generates BEBFs for sparse feature spaces. We provide a finite sample analysis of the proposed method, and prove that projections logarithmic in the dimension of the original space guarantee a contraction in the error. Empirical results demonstrate the strength of this method in domains in which choosing a good state representation is challenging.

Su53 Reinforcement Learning in Robust Markov Decision Processes

Shiau-Hong Lim shonglim@gmail.com
National University of Singapore
Huan Xu mpexuh@nus.edu.sg
NUS
Shie Mannor shie@ee.technion.ac.il
Technion

An important challenge in Markov decision processes is to ensure robustness with respect to unexpected or adversarial system behavior while taking advantage of well-behaving parts of the system. We consider a problem setting where some unknown parts of the state space can have arbitrary transitions while other parts are purely stochastic. We devise an algorithm that is adaptive to potentially adversarial behavior and show that it achieves similar regret bounds as the purely stochastic case.

Su54 Projected Natural Actor-Critic

Philip Thomas pthomas@cs.umass.edu
William Dabney wdabney@cs.umass.edu
Stephen Giguere sgiguere9@gmail.com
Sridhar Mahadevan mahadeva@cs.umass.edu
UMass Amherst

Natural actor-critics are a popular class of policy search algorithms for finding locally optimal policies for Markov decision processes. In this paper we address a drawback of natural actor-critics that limits their real-world applicability - their lack of safety guarantees. We present a principled algorithm for performing natural gradient descent over a constrained domain. In the context of reinforcement learning, this allows for natural actor-critic algorithms that are guaranteed to remain within a known safe region of policy space. While deriving our class of constrained natural actor-critic algorithms, which we call Projected Natural Actor-Critics (PNACs), we also elucidate the relationship between natural gradient descent and mirror descent.

Su55 (More) Efficient Reinforcement Learning via Posterior Sampling

Ian Osband ian.osband@gmail.com
Dan Russo dan.joseph.russo@gmail.com
Benjamin Van Roy bvr@stanford.edu
Stanford University

Most provably efficient learning algorithms introduce optimism about poorly-understood states and actions to encourage exploration. We study an alternative approach for efficient exploration, posterior sampling for reinforcement learning (PSRL). This algorithm proceeds in repeated episodes of known duration. At the start of each episode, PSRL updates a prior distribution over Markov decision processes and takes one sample from this posterior. PSRL then follows the policy that is optimal for this sample during the episode. The algorithm is conceptually simple, computationally efficient and allows an agent to encode prior knowledge in a natural way. We establish an õ$(\tau S\sqrt{AT})$ bound on the expected regret, where T is time, τ is the episode

length and S and A are the cardinalities of the state and action spaces. This bound is one of the first for an algorithm not based on optimism and close to the state of the art for any reinforcement learning algorithm. We show through simulation that PSRL significantly outperforms existing algorithms with similar regret bounds.

Su56 Adaptive Step-Size for Policy Gradient Methods

Matteo Pirotta matteo.pirotta@polimi.it
Marcello Restelli restelli@elet.polimi.it
Luca Bascetta luca.bascetta@polimi.it
Politecnico di Milano

In the last decade, policy gradient methods have significantly grown in popularity in the reinforcement--learning field. In particular, they have been largely employed in motor control and robotic applications, thanks to their ability to cope with continuous state and action domains and partial observable problems. Policy gradient researches have been mainly focused on the identification of effective gradient directions and the proposal of efficient estimation algorithms. Nonetheless, the performance of policy gradient methods is determined not only by the gradient direction, since convergence properties are strongly influenced by the choice of the step size: small values imply slow convergence rate, while large values may lead to oscillations or even divergence of the policy parameters. Step--size value is usually chosen by hand tuning and still little attention has been paid to its automatic selection. In this paper, we propose to determine the learning rate by maximizing a lower bound to the expected performance gain. Focusing on Gaussian policies, we derive a lower bound that is second--order polynomial of the step size, and we show how a simplified version of such lower bound can be maximized when the gradient is estimated from trajectory samples. The properties of the proposed approach are empirically evaluated in a linear--quadratic regulator problem.

Su57 Policy Shaping: Integrating Human Feedback with Reinforcement Learning

Shane Griffith sgriffith7@gatech.edu
Kaushik Subramanian kausubbu@gatech.edu
Jonathan Scholz jkscholz@gatech.edu
Charles Isbell isbell@cc.gatech.edu
Andrea Thomaz athomaz@cc.gatech.edu
Georgia Tech

A long term goal of Interactive Reinforcement Learning is to incorporate non-expert human feedback to solve complex tasks. State-of-the-art methods have approached this problem by mapping human information to reward and value signals to indicate preferences and then iterating over them to compute the necessary control policy. In this paper we argue for an alternate, more effective characterization of human feedback: Policy Shaping. We introduce Advise, a Bayesian approach that attempts to maximize the information gained from human feedback by utilizing it as direct labels on the policy. We compare Advise to state-of-the-art approaches and highlight scenarios where it outperforms them and importantly is robust to infrequent and inconsistent human feedback.

Su58 Optimistic policy iteration and natural actor-critic: A unifying view and a non-optimality result

Paul Wagner pwagner@cis.hut.fi
Aalto University

Approximate dynamic programming approaches to the reinforcement learning problem are often categorized into greedy value function methods and value-based policy gradient methods. As our first main result, we show that an important subset of the latter methodology is, in fact, a limiting special case of a general formulation of the former methodology; optimistic policy iteration encompasses not only most of the greedy value function methods but also natural actor-critic methods, and permits one to directly interpolate between them. The resulting continuum adjusts the strength of the Markov assumption in policy improvement and, as such, can be seen as dual in spirit to the continuum in TD(λ)-style algorithms in policy evaluation. As our second main result, we show for a substantial subset of soft-greedy value function approaches that, while having the potential to avoid policy oscillation and policy chattering, this subset can never converge toward any optimal policy, except in a certain pathological case. Consequently, in the context of approximations, the majority of greedy value function methods seem to be deemed to suffer either from the risk of oscillation/chattering or from the presence of systematic sub-optimality.

Su59 Actor-Critic Algorithms for Risk-Sensitive MDPs

Prashanth L.A. prashanth.la@inria.fr
INRIA
Mohammad Ghavamzadeh
mohammad.ghavamzadeh@inria.fr
INRIA & Adobe Research

In many sequential decision-making problems we may want to manage risk by minimizing some measure of variability in rewards in addition to maximizing a standard criterion. Variance related risk measures are among the most common risk-sensitive criteria in finance and operations research. However, optimizing many such criteria is known to be a hard problem. In this paper, we consider both discounted and average reward Markov decision processes. For each formulation, we first define a measure of variability for a policy, which in turn gives us a set of risk-sensitive criteria to optimize. For each of these criteria, we derive a formula for computing its gradient. We then devise actor-critic algorithms for estimating the gradient and updating the policy parameters in the ascent direction. We establish the convergence of our algorithms to locally risk-sensitive optimal policies. Finally, we demonstrate the usefulness of our algorithms in a traffic signal control application.

Su60 DESPOT: Online POMDP Planning with Regularization

Adhiraj Somani adhirajsomani@gmail.com
NUS
Nan Ye yenan@comp.nus.edu.sg
David Hsu dyhsu@comp.nus.edu.sg
Wee Sun Lee leews@comp.nus.edu.sg
National University of Singapore

POMDPs provide a principled framework for planning under uncertainty, but are computationally intractable, due to the "curse of dimensionality" and the "curse of history". This paper presents an online lookahead search algorithm that alleviates these difficulties by limiting the search to a set of sampled scenarios. The execution of all policies on the sampled scenarios is summarized using a Determinized Sparse Partially Observable Tree (DESPOT), which is a sparsely sampled belief tree. Our algorithm, named Regularized DESPOT (R-DESPOT), searches the DESPOT for a policy that optimally balances the size of the policy and the accuracy on its value estimate obtained through sampling. We give an output-sensitive performance bound for all policies derived from the DESPOT, and show that R-DESPOT works well if a small optimal policy exists. We also give an anytime approximation to R-DESPOT. Experiments show strong results, compared with two of the fastest online POMDP algorithms.

Su61 Approximate Dynamic Programming Finally Performs Well in the Game of Tetris

Victor Gabillon victor.gabillon@inria.fr
Bruno Scherrer scherrer@loria.fr
INRIA
Mohammad Ghavamzadeh
mohammad.ghavamzadeh@inria.fr
INRIA & Adobe Research

Tetris is a popular video game that has been widely used as a benchmark for various optimization techniques including approximate dynamic programming (ADP) algorithms. A close look at the literature of this game shows that while ADP algorithms, that have been (almost) entirely based on approximating the value function (value function based), have performed poorly in Tetris, the methods that search directly in the space of policies by learning the policy parameters using an optimization black box, such as the cross entropy (CE) method, have achieved the best reported results. This makes us conjecture that Tetris is a game in which good policies are easier to represent, and thus, learn than their corresponding value functions. So, in order to obtain a good performance with ADP, we should use ADP algorithms that search in a policy space, instead of the more traditional ones that search in a value function space. In this paper, we put our conjecture to test by applying such an ADP algorithm, called classification-based modified policy iteration (CBMPI), to the game of Tetris. Our extensive experimental results show that for the first time an ADP algorithm, namely CBMPI, obtains the best results reported in the literature for Tetris in both small 10×10 and large 10×20 boards. Although the CBMPI's results are similar to those achieved by the CE method in the large board, CBMPI uses considerably fewer (almost 1/10) samples (call to the generative model of the game) than CE.

Su62 Reward Mapping for Transfer in Long-Lived Agents

Xiaoxiao Guo guoxiao@umich.edu
Satinder Singh baveja@umich.edu
Richard Lewis rickl@umich.edu
University of Michigan

We consider how to transfer knowledge from previous tasks to a current task in long-lived and bounded agents that must solve a sequence of MDPs over a finite lifetime. A novel aspect of our transfer approach is that we reuse reward functions. While this may seem counterintuitive, we build on the insight of recent work on the optimal rewards problem that guiding an agent's behavior with reward functions other than the task-specifying reward function can help overcome computational bounds of the agent. Specifically, we use good guidance reward functions learned on previous tasks in the sequence to incrementally train a reward mapping function that maps task-specifying reward functions into good initial guidance reward functions for subsequent tasks. We demonstrate that our approach can substantially improve the agent's performance relative to other approaches, including an approach that transfers policies.

Su63 Learning a Deep Compact Image Representation for Visual Tracking

Naiyan Wang winsty@gmail.com
Dit-Yan Yeung dyyeung@cse.ust.hk
Hong Kong University of Science and Technology

In this paper, we study the challenging problem of tracking the trajectory of a moving object in a video with possibly very complex background. In contrast to most existing trackers which only learn the appearance of the tracked object online, we take a different approach, inspired by recent advances in deep learning architectures, by putting more emphasis on the (unsupervised) feature learning problem. Specifically, by using auxiliary natural images, we train a stacked denoising autoencoder offline to learn generic image features that are more robust against variations. This is then followed by knowledge transfer from offline training to the online tracking process. Online tracking involves a classification neural network which is constructed from the encoder part of the trained autoencoder as a feature extractor and an additional classification layer. Both the feature extractor and the classifier can be further tuned to adapt to appearance changes of the moving object. Comparison with the state-of-the-art trackers on some challenging benchmark video sequences shows that our deep learning tracker is very efficient as well as more accurate.

Su64 Learning the Local Statistics of Optical Flow

Dan Rosenbaum danrsm@cs.huji.ac.il
Daniel Zoran danielzoran@gmail.com
Yair Weiss yweiss@cs.huji.ac.il
Hebrew University

Motivated by recent progress in natural image statistics, we use newly available datasets with ground truth optical flow to learn the local statistics of optical flow and rigorously compare the learned model to prior models assumed by computer vision optical flow algorithms. We find that a Gaussian mixture model with 64 components provides a significantly better model for local flow statistics when compared to commonly used models. We investigate the source of the GMMs success and show it is related to an explicit representation of flow boundaries. We also learn a model that jointly models the local intensity pattern and the local optical flow. In accordance with the assumptions often made in computer vision, the model learns that flow boundaries are more likely at intensity boundaries. However, when evaluated on a large dataset, this dependency is very weak and the benefit of conditioning flow estimation on the local intensity pattern is marginal.

Su65 Third-Order Edge Statistics: Contour Continuation, Curvature, and Cortical Connections

Matthew Lawlor matthew.lawlor@yale.edu
Steven Zucker steven.zucker@yale.edu
Yale University

Association field models have been used to explain human contour grouping performance and to explain the mean frequency of long-range horizontal connections across cortical columns in V1. However, association fields essentially depend on pairwise statistics of edges in natural scenes. We develop a spectral test of the sufficiency of pairwise statistics and show that there is significant higher-order structure. An analysis using a probabilistic spectral embedding reveals curvature-dependent components to the association field, and reveals a challenge for biological learning algorithms.

Su66 What Are the Invariant Occlusive Components of Image Patches? A Probabilistic Generative Approach

Zhenwen Dai dai@fias.uni-frankfurt.de
Goethe-University Frankfurt
Georgios Exarchakis exarchakis@berkeley.edu
UC Berkeley
Jörg Lücke luecke@tu-berlin.de
TU Berlin

We study optimal image encoding based on a generative approach with non-linear feature combinations and explicit position encoding. By far most approaches to unsupervised learning learning of visual features, such as sparse coding or ICA, account for translations by representing the same features at different positions. Some earlier models used a separate encoding of features and their positions to facilitate invariant data encoding and recognition. All probabilistic generative models with explicit position encoding have so far assumed a linear superposition of components to encode image patches. Here, we for the first time apply a model with non-linear feature superposition and explicit position encoding. By avoiding linear superpositions, the studied model represents a closer match to component occlusions which are ubiquitous in natural images. In order to account for occlusions, the non-linear model encodes patches qualitatively very different from linear models by using component representations separated into mask and feature parameters. We first investigated encodings learned by the model using artificial data with mutually occluding components. We find that the model extracts the components, and that it can correctly identify the occlusive components with the hidden variables of the model. On natural image patches, the model learns component masks and features for typical image components. By using reverse correlation, we estimate the receptive fields associated with the model's hidden units. We find many Gabor-like or globular receptive fields as well as fields sensitive to more complex structures. Our results show that probabilistic models that capture occlusions and invariances can be trained efficiently on image patches, and that the resulting encoding represents an alternative model for the neural encoding of images in the primary visual cortex.

Su67 Action from Still Image Dataset and Inverse Optimal Control to Learn Task Specific Visual Scanpaths

Stefan Mathe mstefan@cs.toronto.edu
University of Toronto
Cristian Sminchisescu
cristian.sminchisescu@math.lth.se
LTH

Human eye movements provide a rich source of information into the human visual processing. The complex interplay between the task and the visual stimulus is believed to determine human eye movements, yet it is not fully understood. This has precluded the development of reliable dynamic eye movement prediction systems. Our work makes three contributions towards addressing this problem. First, we complement one of the largest and most challenging static computer vision datasets, VOC 2012 Actions, with human eye movement annotations collected under the task constraints of action and context recognition. Our dataset is unique among eyetracking datasets for still images in terms of its large scale (over 1 million fixations, 9157 images), task control and action from a single image emphasis. Second, we introduce models to automatically discover areas of interest (AOI) and introduce novel dynamic consistency metrics, based on them. Our method can automatically determine the number and spatial support of the AOIs, in addition to their locations. Based on such encodings, we show that, on unconstrained read-world stimuli, task instructions have significant influence on visual behavior. Finally, we leverage our large scale dataset in conjunction with powerful machine learning techniques and computer vision features, to introduce novel dynamic eye movement prediction methods which learn task-sensitive reward functions from eye movement data and efficiently integrate these rewards to plan future saccades based on inverse optimal control. We show that the propose methodology achieves state of the art scanpath modeling results.

Su68 Action is in the Eye of the Beholder: Eye-gaze Driven Model for Spatio-Temporal Action Localization

Nataliya Shapovalova nshapova@sfu.ca
Greg Mori mori@cs.sfu.ca
Simon Fraser University
Michalis Raptis mraptis@disneyresearch.com
Leonid Sigal lsigal@disneyresearch.com
Disney Research

We propose a new weakly-supervised structured learning approach for recognition and spatio-temporal localization of actions in video. As part of the proposed approach we develop a generalization of the Max-Path search algorithm, which allows us to efficiently search over a structured space of multiple spatio-temporal paths, while also allowing to incorporate context information into the model. Instead of using spatial annotations, in the form of bounding boxes, to guide the latent model during training, we utilize human gaze data in the form of a weak supervisory signal. This is achieved by incorporating gaze, along with the classification, into the structured loss within the latent SVM learning framework. Experiments on a challenging benchmark dataset, UCF-Sports, show that our model is more accurate, in terms of classification, and achieves state-of-the-art results in localization. In addition, we show how our model can produce top-down saliency maps conditioned on the classification label and localized latent paths.

Su69 Higher Order Priors for Joint Intrinsic Image, Objects, and Attributes Estimation

Vibhav Vineet vibhav.vineet-2010@brookes.ac.uk
Oxford Brookes University
Carsten Rother carsten.rother@tu-dresden.de
TU Dresden
Philip Torr philiptorr@hotmail.com
University of Oxford

Many methods have been proposed to recover the intrinsic scene properties such as shape, reflectance and illumination from a single image. However, most of these models have been applied on laboratory datasets. In this work we explore the synergy effects between intrinsic scene properties recovered from an image, and the objects and attributes present in the scene. We cast the problem in a joint energy minimization framework; thus our model is able to encode the strong correlations between intrinsic properties (reflectance, shape, illumination), objects (table, tv-monitor), and materials (wooden, plastic) in a given scene. We tested our approach on the NYU and Pascal datasets, and observe both qualitative and quantitative improvements in the overall accuracy.

Su70 Decision Jungles: Compact and Rich Models for Classification

Jamie Shotton jamiesho@microsoft.com
Toby Sharp tsharp@microsoft.com
Pushmeet Kohli pkohli@microsoft.com
Sebastian Nowozin senowozi@microsoft.com
John Winn jwinn@microsoft.com
Antonio Criminisi antcrim@microsoft.com
Microsoft Research

Randomized decision trees and forests have a rich history in machine learning and have seen considerable success in application, perhaps particularly so for computer vision. However, they face a fundamental limitation: given enough data, the number of nodes in decision trees will grow exponentially with depth. For certain applications, for example on mobile or embedded processors, memory is a limited resource, and so the exponential growth of trees limits their depth, and thus their potential accuracy. This paper proposes decision jungles, revisiting the idea of ensembles of rooted decision directed acyclic graphs (DAGs), and shows these to be compact and powerful discriminative models for classification. Unlike conventional decision trees that only allow one path to every node, a DAG in a decision jungle allows multiple paths from the root to each leaf. We present and compare two new node merging algorithms that jointly optimize both the features and the structure of the DAGs efficiently. During training, node splitting and node merging are driven by the minimization of exactly the same objective function, here the weighted sum of entropies at the leaves. Results on varied datasets show that, compared to decision forests and several other baselines, decision jungles require dramatically less memory while considerably improving generalization.

Su71 Non-Linear Domain Adaptation with Boosting

Carlos Becker carlos.becker@epfl.ch
Christos Christoudias mario.christoudias@epfl.ch
Pascal Fua pascal.fua@epfl.ch
EPFL

A common assumption in machine vision is that the training and test samples are drawn from the same distribution. However, there are many problems when this assumption is grossly violated, as in bio-medical applications where different acquisitions can generate drastic variations in the appearance of the data due to changing experimental conditions. This problem is accentuated with 3D data, for which annotation is very time-consuming, limiting the amount of data that can be labeled in new acquisitions for training. In this paper we present a multi-task learning algorithm for domain adaptation based on boosting. Unlike previous approaches that learn task-specific decision boundaries, our method learns a single decision boundary in a shared feature space, common to all tasks. We use the boosting-trick to learn a non-linear mapping of the observations in each task, with no need for specific a-priori knowledge of its global analytical form. This yields a more parameter-free domain adaptation approach that successfully leverages learning on new tasks where labeled data is scarce. We evaluate our approach on two challenging bio-medical datasets and achieve a significant improvement over the state-of-the-art.

Su72 Modeling Clutter Perception using Parametric Proto-object Partitioning

Chen-Ping Yu cxy7452@gmail.com
Dimitris Samaras samaras@cs.stonybrook.edu
Greg Zelinsky gregory.zelinsky@stonybrook.edu
Stony Brook University
Wen-Yu Hua littlehanag@gmail.com
Penn State University

Visual clutter, the perception of an image as being crowded and disordered, affects aspects of our lives ranging from object detection to aesthetics, yet relatively little effort has been made to model this important and ubiquitous percept. Our approach models clutter as the number of proto-objects segmented from an image, with proto-objects defined as groupings of superpixels that are similar in intensity, color, and gradient orientation features. We introduce a novel parametric method of merging superpixels by modeling mixture of Weibull distributions on similarity distance statistics, then taking the normalized number of proto-objects following partitioning as our estimate of clutter perception. We validated this model using a new $90-$image dataset of realistic scenes rank ordered by human raters for clutter, and showed that our method not only predicted clutter extremely well (Spearman's $\rho=0.81$, $p<0.05$), but also outperformed all existing clutter perception models and even a behavioral object segmentation ground truth. We conclude that the number of proto-objects in an image affects clutter perception more than the number of objects or features.

Su73 Mid-level Visual Element Discovery as Discriminative Mode Seeking

Carl Doersch cdoersch@cs.cmu.edu
Abhinav Gupta abhinavg@cs.cmu.edu
CMU
Alexei Efros efros@cs.berkeley.edu
UC Berkeley

Recent work on mid-level visual representations aims to capture information at the level of complexity higher than typical "visual words", but lower than full-blown semantic objects. Several approaches have been proposed to discover mid-level visual elements, that are both 1) representative, i.e. frequently occurring within a visual dataset, and 2) visually discriminative. However, the current approaches are rather ad hoc and difficult to analyze and evaluate. In this work, we pose visual element discovery as discriminative mode seeking, drawing connections to the the well-known and well-studied mean-shift algorithm. Given a weakly-labeled image collection, our method discovers visually-coherent patch clusters that are maximally discriminative with respect to the labels. One advantage of our formulation is that it requires only a single pass through the data. We also propose the Purity-Coverage plot as a principled way of experimentally analyzing and evaluating different visual discovery approaches, and compare our method against prior work on the Paris Street View dataset. We also evaluate our method on the task of scene classification, demonstrating state-of-the-art performance on the MIT Scene-67 dataset.

Su74 Optimal integration of visual speed across different spatiotemporal frequency channels

Matjaz Jogan mjogan@sas.upenn.edu
Alan Stocker astocker@sas.upenn.edu
University of Pennsylvania

How does the human visual system compute the speed of a coherent motion stimulus that contains motion energy in different spatiotemporal frequency bands? Here we propose that perceived speed is the result of optimal integration of speed information from independent spatiotemporal frequency tuned channels. We formalize this hypothesis with a Bayesian observer model that treats the channel activity as independent cues, which are optimally combined with a prior expectation for slow speeds. We test the model against behavioral data from a 2AFC speed discrimination task with which we measured subjects' perceived speed of drifting sinusoidal gratings with different contrasts and spatial frequencies, and of various combinations of these single gratings. We find that perceived speed of the combined stimuli is independent of the relative phase of the underlying grating components, and that the perceptual biases and discrimination thresholds are always smaller for the combined stimuli, supporting the cue combination hypothesis. The proposed Bayesian model fits the data well, accounting for perceptual biases and thresholds of both simple and combined stimuli. Fits are improved if we assume that the channel responses are subject to divisive normalization, which is in line with physiological evidence. Our results provide an important step toward a more complete model of visual motion perception that can predict perceived speeds for stimuli of arbitrary spatial structure.

Su75 DeViSE: A Deep Visual-Semantic Embedding Model

Andrea Frome afrome@google.com
Greg Corrado gcorrado@google.com
Jon Shlens shlens@google.com
Samy Bengio bengio@google.com
Jeff Dean jeff@google.com
Marc'Aurelio Ranzato ranzato@google.com
Tomas Mikolov tmikolov@google.com
Google Research

Modern visual recognition systems are often limited in their ability to scale to large numbers of object categories. This limitation is in part due to the increasing difficulty of acquiring sufficient training data in the form of labeled images as the number of object categories grows. One remedy is to leverage data from other sources -- such as text data -- both to train visual models and to constrain their predictions. In this paper we present a new deep visual-semantic embedding model trained to identify visual objects using both labeled image data as well as semantic information gleaned from unannotated text. We demonstrate that this model matches state-of-the-art performance on the 1000-class ImageNet object recognition challenge while making more semantically reasonable errors, and also show that the semantic information can be exploited to make predictions about tens of thousands of image labels not observed during training. Semantic knowledge improves such zero-shot predictions by up to 65%, achieving hit rates of up to 10% across thousands of novel labels never seen by the visual model.

Su76 Visual Concept Learning: Combining Machine Vision and Bayesian Generalization on Concept Hierarchies

Yangqing Jia jiayq@eecs.berkeley.edu
Joshua Abbott joshua.abbott@berkeley.edu
Thomas Griffiths tom_griffiths@berkeley.edu
Trevor Darrell trevor@eecs.berkeley.edu
UC Berkeley
Joseph Austerweil joseph.austerweil@gmail.com
Brown University

Learning a visual concept from a small number of positive examples is a significant challenge for machine learning algorithms. Current methods typically fail to find the appropriate level of generalization in a concept hierarchy for a given set of visual examples. Recent work in cognitive science on Bayesian models of generalization addresses this challenge, but prior results assumed that objects were perfectly recognized. We present an algorithm for learning visual concepts directly from images, using probabilistic predictions generated by visual classifiers as the input to a Bayesian generalization model. As no existing challenge data tests this paradigm, we collect and make available a new, large-scale dataset for visual concept learning using the ImageNet hierarchy as the source of possible concepts, with human annotators to provide ground truth labels as to whether a new image is an instance of each concept using a paradigm similar to that used in experiments studying word learning in children. We compare the performance of our system to several baseline algorithms, and show a significant advantage results from combining visual classifiers with the ability to identify an appropriate level of abstraction using Bayesian generalization.

Su77 Learning invariant representations and applications to face verification

Qianli Liao lql@mit.edu
Joel Leibo jzleibo@mit.edu
Tomaso Poggio tp@csail.mit.edu
Massachusetts Institute of Technology

One approach to computer object recognition and modeling the brain's ventral stream involves unsupervised learning of representations that are invariant to common transformations. However, applications of these ideas have usually been limited to 2D affine transformations, e.g., translation and scaling, since they are easiest to solve via convolution. In accord with a recent theory of transformation-invariance, we propose a model that, while capturing other common convolutional networks as special cases, can also be used with arbitrary identity-preserving transformations. The model's wiring can be learned from videos of transforming objects---or any other grouping of images into sets by their depicted object. Through a series of successively more complex empirical tests, we study the invariance/discriminability properties of this model with respect to different transformations. First, we empirically confirm theoretical predictions for the case of 2D affine transformations. Next, we apply the model to non-affine transformations: as expected, it performs well on face verification tasks requiring invariance to the relatively smooth transformations of 3D rotation-in-depth and changes in illumination direction. Surprisingly, it can also tolerate clutter "transformations" which map an image of a face on one background to an image of the same face on a different background. Motivated by these empirical findings, we tested the same model on face verification benchmark tasks from the computer vision literature: Labeled Faces in the Wild, PubFig and a new dataset we gathered---achieving strong performance in these highly unconstrained cases as well.

Su78 Deep Neural Networks for Object Detection

Christian Szegedy szegedy@google.com
Alexander Toshev toshev@google.com
Dumitru Erhan dumitru@google.com
Google Research

Deep Neural Networks (DNNs) have recently shown outstanding performance on the task of whole image classification. In this paper we go one step further and address the problem of object detection -- not only classifying but also precisely localizing objects of various classes using DNNs. We present a simple and yet powerful formulation of object detection as a regression to object masks. We define a multi-scale inference procedure which is able to produce a high-resolution object detection at a low cost by a few network applications. The approach achieves state-of-the-art performance on Pascal 2007 VOC.

Su79 Deep Fisher Networks for Large-Scale Image Classification

Karen Simonyan karen@robots.ox.ac.uk
Andrea Vedaldi vedaldi@robots.ox.ac.uk
Andrew Zisserman az@robots.ox.ac.uk
University of Oxford

As massively parallel computations have become broadly available with modern GPUs, deep architectures trained on very large datasets have risen in popularity. Discriminatively trained convolutional neural networks, in particular, were recently shown to yield state-of-the-art performance in challenging image classification benchmarks such as ImageNet. However, elements of these architectures are similar to standard hand-crafted representations used in computer vision. In this paper, we explore the extent of this analogy, proposing a version of the state-of-the-art Fisher vector image encoding that can be stacked in multiple layers. This architecture significantly improves on standard Fisher vectors, and obtains competitive results with deep convolutional networks at a significantly smaller computational cost. Our hybrid architecture allows us to measure the performance improvement brought by a deeper image classification pipeline, while staying in the realms of conventional SIFT features and FV encodings.

Su80 Fast Template Evaluation with Vector Quantization

Mohammad Amin Sadeghi msadegh2@illinois.edu
David Forsyth daf@illinois.edu
University of Illinois at Urbana-Champaign

Applying linear templates is an integral part of many object detection systems and accounts for a significant portion of computation time. We describe a method that achieves a substantial end-to-end speedup over the best current methods, without loss of accuracy. Our method is a combination of approximating scores by vector quantizing feature windows and a number of speedup techniques including cascade. Our procedure allows speed and accuracy to be traded off in two ways: by choosing the number of Vector Quantization levels, and by choosing to rescore windows or not. Our method can be directly plugged into any recognition system that relies on linear templates. We demonstrate our method to speed up the original Exemplar SVM detector [1] by an order of magnitude and Deformable Part models [2] by two orders of magnitude with no loss of accuracy.

Su81 Transfer Learning in a Transductive Setting

Marcus Rohrbach rohrbach@mpi-inf.mpg.de
Sandra Ebert ebert@mpi-inf.mpg.de
Bernt Schiele schiele@mpi-inf.mpg.de
Max Planck Institute for Informatics

Category models for objects or activities typically rely on supervised learning requiring sufficiently large training sets. Transferring knowledge from known categories to novel classes with no or only a few labels however is far less researched even though it is a common scenario. In this work, we extend transfer learning with semi-supervised learning to exploit unlabeled instances of (novel) categories with no or only a few labeled instances. Our proposed approach Propagated Semantic Transfer combines three main ingredients. First, we transfer information from known to novel categories by incorporating external knowledge, such as linguistic or expert-specified information, e.g., by a mid-level layer of semantic attributes. Second, we exploit the manifold structure of novel classes. More specifically we adapt a graph-based learning algorithm - so far only used for semi-supervised learning - to zero-shot and few-shot learning. Third, we improve the local neighborhood in such graph structures by replacing the raw feature-based representation with a mid-level object- or attribute-based representation. We evaluate our approach on three challenging datasets in two different applications, namely on Animals with Attributes and ImageNet for image classification and on MPII Composites for activity recognition. Our approach consistently outperforms state-of-the-art transfer and semi-supervised approaches on all datasets.

Su82 Reshaping Visual Datasets for Domain Adaptation

Boqing Gong boqinggo@usc.edu
Fei Sha feisha@usc.edu
University of Southern California (USC)
Kristen Grauman grauman@cs.utexas.edu
UT Austin

In visual recognition problems, the common data distribution mismatches between training and testing make domain adaptation essential. However, image data is difficult to manually divide into the discrete domains required by adaptation algorithms, and the standard practice of equating datasets with domains is a weak proxy for all the real conditions that alter the statistics in complex ways (lighting, pose, background, resolution, etc.) We propose an approach to automatically discover latent domains in image or video datasets. Our formulation imposes two key properties on domains: maximum distinctiveness and maximum learnability. By maximum distinctiveness, we require the underlying distributions of the identified domains to be different from each other; by maximum learnability, we ensure that a strong discriminative model can be learned from the domain. We devise a nonparametric representation and efficient optimization procedure for distinctiveness, which, when coupled with our learnability constraint, can successfully discover domains among both training and test data. We extensively evaluate our approach on object recognition and human activity recognition tasks.

Su83 Heterogeneous-Neighborhood-based Multi-Task Local Learning Algorithms

Yu Zhang yuzhang@comp.hkbu.edu.hk
Hong Kong Baptist University

All the existing multi-task local learning methods are defined on homogeneous neighborhood which consists of all data points from only one task. In this paper, different from existing methods, we propose local learning methods for multi-task classification and regression problems based on heterogeneous neighborhood which is defined on data points from all tasks. Specifically, we extend the k-nearest-neighbor classifier by formulating the decision function for each data point as a weighted voting among the neighbors from all tasks where the weights are task-specific. By defining a regularizer to enforce the task-specific weight matrix to approach a symmetric one, a regularized objective function is proposed and an efficient coordinate descent method is developed to solve it. For regression problems, we extend the kernel regression to multi-task setting in a similar way to the classification case. Experiments on some toy data and real-world datasets demonstrate the effectiveness of our proposed methods.

Su84 Learning Feature Selection Dependencies in Multi-task Learning

Daniel Hernández-Lobato daniel.hernandez@uam.es
Universidad Autónoma de Madrid
José Miguel Hernández-Lobato jmh233@cam.ac.uk
University of Cambridge

A probabilistic model based on the horseshoe prior is proposed for learning dependencies in the process of identifying relevant features for prediction. Exact inference is intractable in this model. However, expectation propagation offers an approximate alternative. Because the process of estimating feature selection dependencies may suffer from over-fitting in the model proposed, additional data from a multi-task learning scenario are considered for induction. The same model can be used in this setting with few modifications. Furthermore, the assumptions made are less restrictive than in other multi-task methods: The different tasks must share feature selection dependencies, but can have different relevant features and model coefficients. Experiments with real and synthetic data show that this model performs better than other multi-task alternatives from the literature. The experiments also show that the model is able to induce suitable feature selection dependencies for the problems considered, only from the training data.

Su85 Parametric Task Learning

Ichiro Takeuchi takeuchi.ichiro@nitech.ac.jp
Tatsuya Hongo hongo.mllab.nit@gmail.com
Nagoya Institute of Technology
Masashi Sugiyama sugi@cs.titech.ac.jp
Tokyo Institute of Technology
Shinichi Nakajima shinnkj23@gmail.com
Nikon

We introduce a novel formulation of multi-task learning (MTL) called parametric task learning (PTL) that can systematically handle infinitely many tasks parameterized by a continuous parameter. Our key finding is that, for a certain class of PTL problems, the path of optimal task-wise solutions can be represented as piecewise-linear functions of the continuous task parameter. Based on this fact, we employ a parametric programming technique to obtain the common shared representation across all the continuously parameterized tasks efficiently. We show that our PTL formulation is useful in various scenarios such as learning under non-stationarity, cost-sensitive learning, and quantile regression, and demonstrate the usefulness of the proposed method experimentally in these scenarios.

Su86 Direct 0-1 Loss Minimization and Margin Maximization with Boosting

Shaodan Zhai zhai.6@wright.edu
Tian Xia xia.7@wright.edu
Ming Tan tan.6@wright.edu
Shaojun Wang shaojun.wang@wright.edu
Wright State University

We propose a boosting method, DirectBoost, a greedy coordinate descent algorithm that builds an ensemble classifier of weak classifiers through directly minimizing empirical classification error over labeled training examples; once the training classification error is reduced to a local coordinatewise minimum, DirectBoost runs a greedy coordinate ascent algorithm that continuously adds weak classifiers to maximize any targeted arbitrarily defined margins until reaching a local coordinatewise maximum of the margins in a certain sense. Experimental results on a collection of machine-learning benchmark datasets show that DirectBoost gives consistently better results than AdaBoost, LogitBoost, LPBoost with column generation and BrownBoost, and is noise tolerant when it maximizes an n'th order bottom sample margin.

Su87 Reservoir Boosting : Between Online and Offline Ensemble Learning

Leonidas Lefakis leonidas.lefakis@idiap.ch
François Fleuret francois.fleuret@idiap.ch
Idiap Research Institute

We propose to train an ensemble with the help of a reservoir in which the learning algorithm can store a limited number of samples. This novel approach lies in the area between offline and online ensemble approaches and can be seen either as a restriction of the former or an enhancement of the latter. We identify some basic strategies that can be used to populate this reservoir and present our main contribution, dubbed Greedy Edge Expectation Maximization (GEEM), that maintains the reservoir content in the case of Boosting by viewing the samples through their projections into the weak classifier response space. We propose an efficient algorithmic implementation which makes it tractable in practice, and demonstrate its efficiency experimentally on several compute-vision data-sets, on which it outperforms both online and offline methods in a memory constrained setting.

Su88 Understanding variable importances in forests of randomized trees

Gilles Louppe g.louppe@ulg.ac.be
Louis Wehenkel l.wehenkel@ulg.ac.be
Antonio Sutera a.sutera@ulg.ac.be
Pierre Geurts p.geurts@ulg.ac.be
Université de Liège

Despite growing interest and practical use in various scientific areas, variable importances derived from tree-based ensemble methods are not well understood from a theoretical point of view. In this work we characterize the Mean Decrease Impurity (MDI) variable importances as measured by an ensemble of totally randomized trees in asymptotic sample and ensemble size conditions. We derive a three-level decomposition of the information jointly provided by all input variables about the output in terms of i) the MDI importance of each input variable, ii) the degree of interaction of a given input variable with the other input variables, iii) the different interaction terms of a given degree. We then show that this MDI importance of a variable is equal to zero if and only if the variable is irrelevant and that the MDI importance of a relevant variable is invariant with respect to the removal or the addition of irrelevant variables. We illustrate these properties on a simple example and discuss how they may change in the case of non-totally randomized trees such as Random Forests and Extra-Trees.

Su89 Sinkhorn Distances: Lightspeed Computation of Optimal Transportation

Marco Cuturi mcuturi@i.kyoto-u.ac.jp
Kyoto University

Optimal transportation distances are a fundamental family of parameterized distances for histograms in the probability simplex. Despite their appealing theoretical properties, excellent performance and intuitive formulation, their computation involves the resolution of a linear program whose cost is prohibitive whenever the histograms' dimension exceeds a few hundreds. We propose in this work a new family of optimal transportation distances that look at transportation problems from a maximum-entropy perspective. We smooth the classical optimal transportation problem with an entropic regularization term, and show that the resulting optimum is also a distance which can be computed through Sinkhorn's matrix scaling algorithm at a speed that is several orders of magnitude faster than that of transportation solvers. We also report improved performance on the MNIST benchmark problem over competing distances.

Su90 Beyond Pairwise: Provably Fast Algorithms for Approximate k-Way Similarity Search

Anshumali Shrivastava anshu@cs.cornell.edu
Ping Li pingli@cornell.edu
Cornell University

We go beyond the notion of pairwise similarity and look into search problems with k-way similarity functions. In this paper, we focus on problems related to *3-way Jaccard* similarity: $\mathcal{R}^{3way} = {}^{|S1 \cap S2 \cap S3|}/_{|S1 \cup S2 \cup S3|}$, $S1, S2, S3 \in \mathcal{C}$, where $\mathcal{C}$ is a size n collection of sets (or binary vectors). We show that approximate $\mathcal{R}^{3way}$ similarity search problems admit fast algorithms with provable guarantees, analogous to the pairwise case. Our analysis and speedup guarantees naturally extend to$k$-way resemblance. In the process, we extend traditional framework of \ emph{locality sensitive hashing (LSH)} to handle higher order similarities, which could be of independent theoretical interest. The applicability of $\mathcal{R}^{3way}$ search is shown on the "Google sets" application. In addition, we demonstrate the advantage of $\mathcal{R}^{3way}$ resemblance over the pairwise case in improving retrieval quality.

REVIEWERS

Yasin Abbasi
Jacob Abernethy
Margareta Ackerman
Ryan Adams
Raja Hafiz Affandi
Alekh Agarwal
Shipra Agrawal
Yashar Ahmadian
Amr Ahmed
Nir Ailon
Edo Airoldi
Karteek Alahari
Morteza Alamgir
Qi Alan
Pierre Alquier
Mauricio Alvarez
Marco Alvarez
Carlos Alzate
Kareem Amin
Massih-Reza Amini
Anima Anandkumar
Oren Anava
Charles Anderson
Bjoern Andres
Ron Appel
Cedric Archambeau
Evan Archer
andreas argyriou
Raman Arora
Kai Arras
Arthur Asuncion
Chris Atkeson
Joseph Austerweil
Lilach Avitan
Yusuf Aytar
Chloé-Agathe Azencott
Martin Azizyan
S. Derin Babacan
Stephen Bach
Francis Bach
Bing Bai
Suhrid Balakrishnan
Sivaraman Balakrishnan
Krishnakumar Balasubramanian
Christopher Baldassano
Luca Baldassarre
Pierre Baldi
David Balduzzi
Borja Balle
Arindam Banerjee
Ying-Ze Bao
Yoseph Barash
Andre Barreto
Jon Barron
Peter Bartlett
Gabor Bartok
Sumit Basu
Dhruv Batra
Francoise Beaufays
Stephen Becker
Niko Beerenwinkel
Oscar Beijbom
Marc Bellemare
Serge Belongie
Horesh Ben Shitrit
Shai Ben-David
Samy Bengio
Yoshua Bengio
Jose Bento
Philipp Berens
Alex Berg
James Bergstra
Michel Besserve
Matthias Bethge
Chiranjib Bhattacharyya
Jinbo Bi
Jacob Bien
Felix Biessmann
Misha Bilenko
Aude Billard
Jeff Bilmes
Aharon Birnbaum
Matthew Blaschko
David Blei
Liefeng Bo
jeannette Bohg
Sander Bohte
Danushka Bollegala
Edwin Bonilla
Byron Boots
Antoine Bordes
Karsten Borgwardt
Jorg Bornschein
Reza Bosagh Zadeh
Leon Bottou
Guillaume Bouchard
Alexandre Bouchard-Côté
Abdeslam Boularias
YLan Boureau
Christos Boutsidis
Jake Bouvrie
Michael Bowling
Jordan Boyd-Graber
Levi Boyles
Steve Branson
David Braun
Mikio Braun
Tamara Broderick
Marcus Brubaker
Nicholas Bryan
Sebastien Bubeck
Lars Buesing
Wray Buntine
Wolfram Burgard
Robert Busa-Fekete
Lucian Busoniu
Tiberio Caetano
Deng Cai
Ben Calderhead
Colin Camerer
Colin Campbell
William Campbell
Stephane Canu
Liangliang Cao
Olivier Cappe
Barbara Caputo
Constantine Caramanis
Peter Carbonetto
Lawrence Carin
Alexandra Carpentier
Miguel Carreira-Perpinan
Gert Cauwenberghs
Gavin Cawley
Asli Celikyilmaz
Taylan Cemgil
Nicolò Cesa-Bianchi
Volkan Cevher
Brahim Chaib-draa
Jonathan Chang
Kai-Wei Chang
Denis Charles
Laurent Charlin
Duen Horng Chau
Sougata Chaudhuri
Gal Chechik
Minmin Chen
Jianhui Chen
Ning Chen
Shuo Chen
Xi Chen
Yutian Chen
Yuxin Chen
Silvia Chiappa
Max Chickering
Gillian Chin
Julien Chiquet
Dmitri Chklovskii
Arthur Choi
Jaedeug Choi
Seungjin Choi
Sumit Chopra
Andreas Christmann
Andrzej Cichocki
Adam Coates
Mark Coates
Ruben Coen-cagli
Shay Cohen
Trevor Cohn
Ronan Collobert
Greg Corrado
Corinna Cortes
Aaron Courville
Koby Crammer
Chris Cueva
John Cunningham
James Cussens
Marco Cuturi
George Dahl
Arnak Dalalyan
Andreas Damianou
Christian Daniel
Amit Daniely
Abhimanyu Das
Dipanjan Das
Emmanuel Dauce
Yann Dauphin
Ian Davidson
Jesse Davis
Nathaniel Daw
Peter Dayan
Cassio de Campos
Fernando de la Torre
Viriginia De Sa
Eyal Dechter
Dennis Decoste
Marc Deisenroth
Ofer Dekel
Krzysztof Dembczynski
Vasil Denchev
Sophie Deneve
Li Deng
Jia Deng
Brian Depasquale
Guillaume Desjardins
Paramveer Dhillon
Thomas Dietterich
Tom Diettrich
Laura Dietz
Chris Ding
Nan Ding
Carlos Diuk
Santosh Divvala
Chuong Do
Huyen Do
Eizaburo Doi
Justin Domke
Finale Doshi-Velez
Arnaud Doucet
Petros Drineas
Shaul Druckmann
Christopher DuBois
John Duchi
Miroslav Dudik
David Dunson
Nicolas Durrande
Haimonti Dutta
David Duvenaud
Jennifer Dy
Frederik Eaton
Elad Eban
Alexander Ecker
Jacob Eisenstein
Jason Eisner
Carl Henrik Ek
Chaitanya Ekanadham
Khalid El-Arini
James Elder
Olivier Elemento
Tina Eliassi-Rad
Gal Elidan
Charles Elkan
Lloyd Elliott
Frank Emmert-Streib
Dominik Endres
Peter Englert
Dumitru Erhan
Stefano Ermon
Tim van Erven
Jo Etzel
Clement Farabet
Alireza Fathi
Paolo Favaro
Aasa Feragen
Rob Fergus
Vittorio Ferrari
Sanja Fidler
Mario Figueiredo
Jozsef Fiser
John Fisher III
Boris Flach
David Fleet
Alyson Fletcher
François Fleuret
Raphael Fonteneau
Justin Foster
Nick Foti
James Foulds
Charless Fowlkes
Emily Fox
Rina Foygel
Vojtech Franc
Andrew Frank
Paolo Frasconi
Peter Frazier
William Freeman
Jeremy Freeman
Abe Friesen
Karl Friston
Mario Fritz
Johannes Fuernkranz
Kenji Fukumizu
C. C. Alan Fung
Nicolo Fusi
Alona Fyshe
Juergen Gall
April Galyardt
Surya Ganguli
Ravi Ganti
Jing Gao
Xin Gao
Dan Garber
Gilles Gasso
Jan Gasthaus
Eric Gaussier
Rong Ge
Peter Gehler
Andreas Geiger
Matthieu Geist
Andrew Gelfand
Sebastien Gerchinovitz
Sean Gerrish
Sam Gershman
Sebastian Gerwinn
Pierre Geurts
Mohammad Ghavamzadeh
Mohammad Gheshlaghi azar
Soumya Ghosh
Richard Gibson
Ran Gilad-Bachrach
Elad Gilboa
Jennifer Gillenwater
Kevin Gimpel
Mark Girolami
Tobias Glasmachers
Amir Globerson
Vibhav Gogate
Jacob Goldberger
Anna Goldenberg
Daniel Golovin
Ryan Gomes
Manuel Gomez-Rodriguez
Alon Gonen
Pinghua Gong
Joseph Gonzalez
Ian Goodfellow
Noah Goodman
Nakul Gopalan
Raghuraman Gopalan
Nico Görnitz
Dilan Gorur
Sergiu Goschin
Stephen Gould
Navin Goyal
Agnieszka Grabska-Barwinska
Thore Graepel
Hans-Peter Graf
Alexandre Gramfort
Yves Grandvalet
David Grangier
Edouard Grave
Mihajlo Grbovic
Karol Gregor
Arthur Gretton
Jim Griffin
Thomas Griffiths
Roger Grosse
Moritz Grosse-Wentrup
Steffen Grunewalder
Asela Gunawardana
Yuhong Guo
Abhinav Gupta
Maya Gupta
Todd Gureckis
Andras Gyorgy

REVIEWERS

Minh Ha Quang
Michael Habeck
Hirotaka Hachiya
Ralf Haefner
Gholamreza Haffari
Patrick Haffner
Attias Hagai
David Hall
Jihun Hamm
Lars-Kai Hansen
Zaid Harchaoui
Bharath Hariharan
Stefan Harmeling
Nicol Harper
Masahiko Haruno
Rob Haslinger
Kohei Hatano
Jarvis Haupt
Elad Hazan
Tamir Hazan
Jingrui He
Creighton Heaukulani
Nicolas Heess
Chinmay Hegde
Hoda Heidari
Matthias Hein
Uri Heinemann
Saied Hemati
James Henderson
Philipp Hennig
James Hensman
Mark Herbster
Jose Miguel Hernández Lobato
Aaron Hertzmann
Hideitsu Hino
Chris Hinrichs
Michael Hirsch
Vaclav Hlavac
Chien-Ju Ho
Qirong Ho
Jesse Hoey
Matt Hoffman
Thomas Hofmann
Steven Hoi
Chris Holmes
Eric Horvitz
Xiaodi Hou
Neil Houlsby
Cho-Jui Hsieh
Chun-Nan Hsu
Daniel Hsu
Tao Hu
Jonathan Huang
Junzhou Huang
Ling Huang
Tzu-Kuo Huang
Eyke Huellermeier
Jonathan Huggins
Michael Hughes
Koji Hukushima
Bui Hung
Jonny Hunt
Sung Ju Hwang
Tsuyoshi Ide
Christian Igel
Alex Ihler
Kazushi Ikeda
Giacomo Indiveri
Marius Cătălin Iordan
Charles Isbell
Mariya Ishteva

Tomoharu Iwata
Rishabh Iyer
Laurent Jacob
Robert Jacobs
Jagadeesh Jagarlamudi
Martin Jaggi
Ashesh Jain
Prateek Jain
Jeremy Jancsary
Dominik Janzing
Qiang Ji
Shuiwang Ji
Yangqing Jia
Zhaoyin Jia
Jiarong Jiang
Luo Jie
Rong Jin
Adam Johansen
Michael Johanson
Matthew Johnson
Nebojsa Jojic
Vladimir Jojic
Armand Joulin
Anatoli Juditsky
Frederic Jurie
Hachem Kadri
Alfredo Kalaitzis
Satyen Kale
Hirokazu Kameoka
Varun Kanade
Takafumi Kanamori
Christopher Kanan
Jonathan Kao
Ashish Kapoor
Purushottam Kar
Theofanis Karaletsos
Theofanis Karaletsos
Nikos Karampatziakis
Masayuki Karasuyama
Amin Karbasi
Yan Karklin
Hisashi Kashima
Samuel Kaski
Kentaro Katahira
Koray Kavukcuoglu
Yoshinobu Kawahara
Motoaki Kawanabe
Sathiya Keerthi
Balazs Kegl
Kristian Kersting
Hossein Keshavarz
Azadeh Khaleghi
Emtiyaz Khan
Aditya Khosla
Bryan Kian Hsiang Low
Martin Kiefel
Seyoung Kim
Akisato Kimura
Irwin King
Franz Kiraly
Sergey Kirshner
Jyrki Kivinen
Negar Kiyavash
Fatma Kilinc Karzan
Arto Klami
Marius Kloft
David Knowles
Ryota Kobayashi
Jens Kober
Pushmeet Kohli
Mikko Koivisto
Mladen Kolar

J. Zico Kolter
Vladlen Koltun
Mamoru Komachi
Nikos Komodakis
Risi Kondor
George Konidaris
Aryeh Kontorovitch
Wouter Koolen
Hema Koppula
Filip Korc
Nathaniel Korda
Petar Kormushev
Adriana Kovashka
Matthieu Kowalski
Andreas Krause
Bartosz Krawczyk
Akshay Krishnamurthy
Dilip Krishnan
Balaji Krishnapuram
Rui Kuang
Brian Kulis
Abhishek Kumar
Neeraj Kumar
M. Pawan Kumar
Ravi Kumar
Sanjiv Kumar
Anshul Kundaje
Branislav Kveton
James Kwok
Cho KyungHyun
Simon Lacoste-Julien
John Lafferty
John Lai
Kevin Lai
Balaji Lakshminarayanan
Christoph Lampert
Gert Lanckriet
Marc Lanctot
Niels Landwehr
John Langford
Ivan Laptev
Hugo Larochelle
Jan Larsen
Pavel Laskov
Neil Lawrence
Alessandro Lazaric
Miguel Lázaro-Gredilla
Svetlana Lazebnik
Hai-Son Le
Quoc Le
Erik Learned-Miller
Daniel Lee
Dongryeol Lee
Honglak Lee
Jason Lee
Wee Sun Lee
Sangkyun Lee
Tai Sing Lee
Leonidas Lefakis
Robert Legenstein
Victor Lempitsky
Ian Lenz
Vincent Lepetit
Jure Leskovec
Christina Leslie
Guy Lever
Fei Fei Li
Fuxin Li
Hang Li
Lihong Li
Limin Li
Ping Li

Dawen Liang
Percy Liang
Li Liao
Xuejun Liao
Katrina Ligett
Binbin Lin
Dahua Lin
Hsuan-Tien Lin
Yuanqing Lin
Scott Linderman
Christoph Lippert
Matthew Liptrot
Bo Liu
Ce Liu
Han Liu
Ji Liu
Jun Liu
Xiaoming Liu
Qiang Liu
Tie-Yan Liu
Wei Liu
Yi-Kai Liu
Karen Livescu
Roi Livni
Dan Lizotte
James Lloyd
Po-Ling Loh
Maria Lomeli
Bo Long
Phil Long
David Lopez-Paz
Yucheng Low
Daniel Lowd
Aurelie Lozano
HongJing Lu
Zhengdong Lu
Chris Lucas
Jorg Lucke
Elliot Ludvig
Gediminas Luksys
Rui Ma
Shiqian Ma
Christian Machens
Jakob Macke
Lester Mackey
Malik Magdon-Ismail
Sridhar Mahadevan
Vijay Mahadevan
Michael Mahoney
Odalric-Ambrym Maillard
Julien Mairal
Subhransu Maji
Arian Maleki
Tomasz Malisiewicz
Stephane Mallat
Jonathan Malmaud
Hiroshi Mamitsuka
Chris Manning
Vikash Mansinghka
Yi Mao
Oded Margalit
Shaul Markovitch
Benjamin Marlin
James Martens
Radoslaw Martin Cichy
Andre Martins
Winter Mason
Tomoko Matsui
Yuji Matsumoto
Julian McAuley
Jon McAuliffe
Andrew McCallum

Brian McFee
Andrew McHutchon
Brendan McMahan
Scott McQuade
Brian McWilliams
Ted Meeds
Ron Meir
Talya Meltzer
Roland Memisevic
Ofer Meshi
Timm Meyer
Elad Mezuman
Mahdi Milani Fard
David Mimno
Tom Minka
Andriy Mnih
Volodymyr Mnih
Daichi Mochihashi
Joseph Modayil
Shakir Mohamed
Karthik Mohan
Mehryar Mohri
Ankur Moitra
Gregoire Montavon
Claire Monteleoni
Greg Mori
Tetsuro Morimura
Edward Moroshko
Quaid Morris
Morten Morup
Alessandro Moschitti
Michael Mozer
Youssef Mroueh
Krikamol Muandet
Klaus-Robert Mueller
Sayan Mukherjee
Andres Munoz Medina
Noboru Murata
Robert Murphy
Iain Murray
Lawrence Murray
Alejandro Murua
Saman Muthukumarana
Gautham Mysore
Boaz Nadler
Sri Nagarajan
Vinod Nair
Shinichi Nakajima
Mukund Narasimhan
Karthik Narayan
Karthik Narayan
Hari Narayanan
Nagarajan Natarajan
Saketha Nath
Daniel Navarro
Sahand Negahban
Willie Neiswanger
Blaine Nelson
Praneeth Netrapalli
Gergely Neu
Gerhard Neumann
Tyler Neylon
Huy Nguyen
Minh Hoai Nguyen
Duy Nguyen-Tuong
Hannes Nickisch
Juan Carlos Niebles
Yang Ning
Gang Niu
William Noble
Yung-Kyun Noh
Rob Nowak

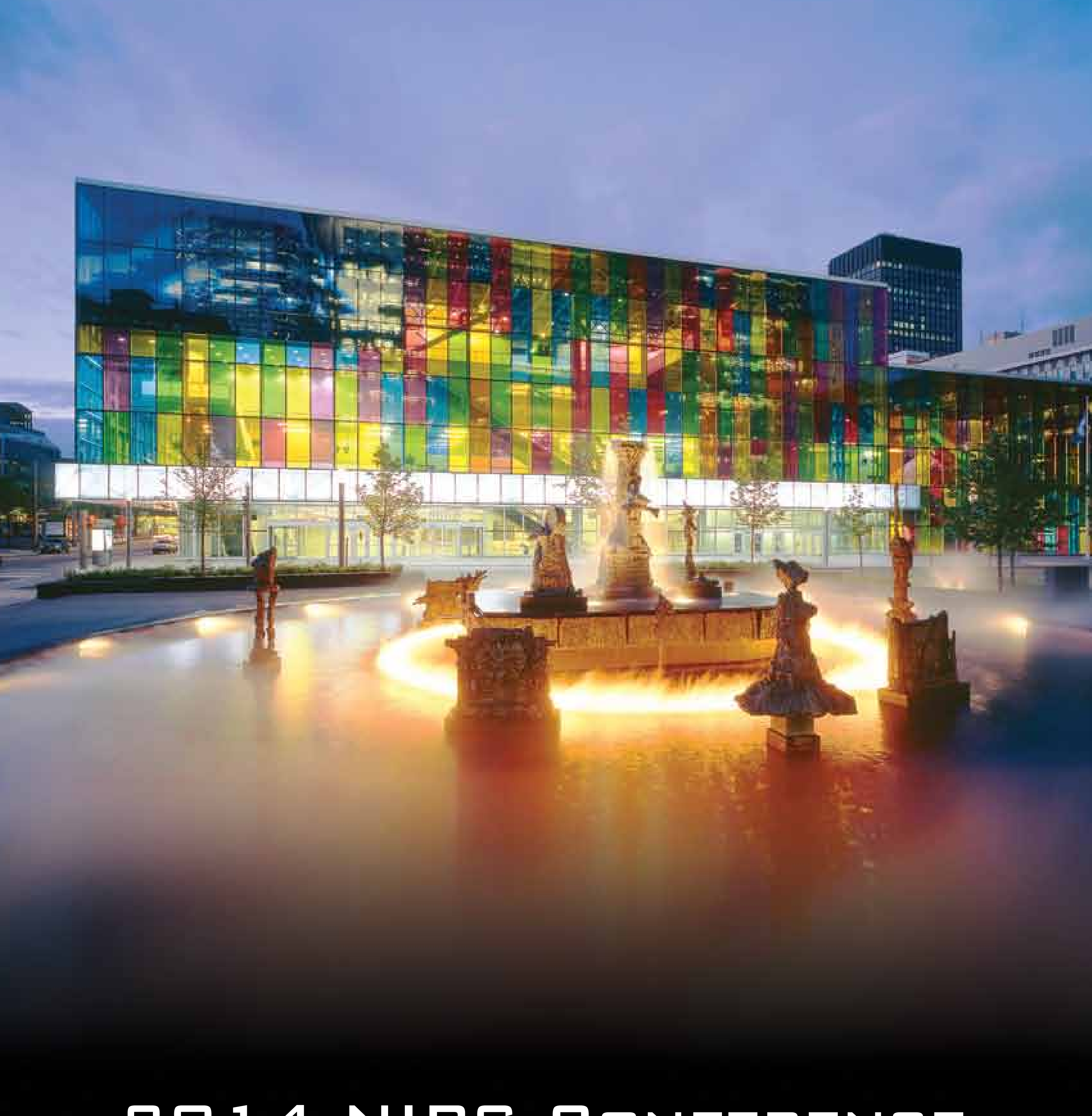
REVIEWERS

Sebastian Nowozin
Timothy O'Donnell
Una May O'Reilly
Guillaume Obozinski
Sewoong Oh
Masato Okada
Bruno Olshausen
Arno Onken
Takshi Onoda
Manfred Opper
Francesco Orabona
Randall Oreilly
Ronald Ortner
Michael Osborne
Sarah Osentoski
Hua Ouyang
Benjamin Packer
David Page
Jean-Francois Paiement
Brooks Paige
John Paisley
David Pal
Konstantina Palla
Liam Paninski
George Papandreou
Ulrich Paquet
Ankur Parikh
Neal Parikh
Devi Parikh
Dennis Park
Il Park
Mijung Park
Ronald Parr
Andrea Passerini
alexandre Passos
Ofer Paternak
Genevieve Patterson
Vladimir Pavlovic
Klaus Pawelzik
Barak Pearlmutter
Jian Peng
Daniel Percival
Fernando Pereira
Fernando Perez-Cruz
Alessandro Perina
Florent Perronnin
Jan Peters
Biljana Petreska
Marek Petrik
Slav Petrov
Nico Pfeifer
Jean-Pascal Pfister
Jonathan Pillow
Joelle Pineau
Hamed Pirsiavash
Xaq Pitkow
John Platt
Robert Pless
Patrick Pletscher
Barnabas Poczos
Jan Poland
Daniel Polani
Massimiliano Pontil
David Poole
Pascal Poupart
Doina Precup
Philippe Preux
Guido Pusiol
Yanjun Qi
Tao Qin
Novi Quadrianto
Ariadna Quattoni
Joaquin Quiñonero-Candela
Neil Rabinowitz
Gunnar Raetsch
Anna Rafferty
Maxim Raginsky
Piyush Rai
Sasha Rakhlin
Alain Rakotomamonjy
Liva Ralaivola
Parikshit Ram
Peter Ramadge
Subramanian Ramamoorthy
Deva Ramanan
Aaditya Ramdas
Marc'Aurelio Ranzato
Vinayak Rao
Carl Rasmussen
Magnus Rattray
Pradeep Ravikumar
Balaraman Ravindran
Debajyoti Ray
Vikas Raykar
Colorado Reed
Khaled Refaat
Roi Reichart
David Reichert
Mark Reid
Joseph Reisinger
Marcello Restelli
Emile Richard
Thomas Richardson
Salah Rifai
Ludovic Righetti
Stephen Roberts
Stéphane Robin
Abel Rodriguez
Jaldert Rombouts
Bernardino Romera-Paredes
Romer Rosales
Lorenzo Rosasco
David Rosenberg
Arun Ross
Fabrice Rossi
Afshin Rostamizadeh
Andrew Roth
Volker Roth
Constantin Rothkopf
Juho Rousu
Benjamin Van Roy
Daniel Roy
Cynthia Rudin
Ulrich Rueckert
Nicholas Ruozzi
Alexander Rush
Olga Russakovsky
Bryan Russell
Andreas Ruttor
Paul Ruvolo
Daniil Ryabko
Sivan Sabato
Regis Sabbadin
Regis Sabbadin
Mohammad Ehsan Saberian
Robin Sabhnani
Mehrnoosh Sadrzadeh
Ankan Saha
Hiroto Saigo
Jun Sakuma
Ruslan Salakhutdinov
Venkatesh Saligrama
Joseph Salmon
Mathieu Salzmann
Saverio Salzo
Adam Sanborn
Sujay Sanghavi
Guido Sanguinetti
Aswin Sankaranarayanan
Sriram Sankararaman
Scott Sanner
Suchi Saria
Anand Sarwate
Issei Sato
Richard Savage
Christoph Sawade
Andrew Saxe
Stefan Schaal
Robert Schapire
Tobias Scheffer
Katya Scheinberg
Bruno Scherrer
Bernt Schiele
Alexander Schliep
Mark Schmidt
Mikkel Schmidt
Jeff Schneider
Benjamin Schrauwen
Christian Schuler
Hannes Schulz
Dale Schuurmans
Odelia Schwartz
Alex Schwing
James Scott
D Sculley
Michele Sebag
Matthias Seeger
Jeff Seeley
Jeff Seeley
Frank Sehnke
Yevgeny Seldin
Sundararajan Sellamanickam
Bart Selman
Lee Seong-Whan
Pierre Sermanet
Ben Shababo
Patrick Shafto
Greg Shakhnarovich
Uri Shalit
Cosma Shalizi
Ohad Shamir
Tatyana Sharpee
James Sharpnack
Or Sheffet
Daniel Sheldon
Christian Shelton
Jacquelyn Shelton
Bin Shen
Xiaotong Shen
Pradeep Shenoy
Nino Shervashidze
Bertram Shi
Qinfeng Shi
Nobuyuki Shimizu
Shohei Shimizu
Helen Shin
Shigeru Shinomoto
Pannaga Shivaswamy
Lavi Shpigelman
Si Si
Olivier Sigaud
Marco Signoretto
Ricardo Silva
Karen Simonyan
Özgür Şimşek
Vikas Sindhwani
Aarti Singh
Ajit Singh
Satinder Singh
Sameer Singh
Vikas Singh
Kaushik Sinha
Mathieu Sinn
Josef Sivic
John Skilling
Kevin Small
Paris Smaragdis
Cristian Sminchisescu
Alex Smola
Jasper Snoek
Richard Socher
Jascha Sohl-Dickstein
Kyung-Ah Sohn
Peter Sollich
Fritz Sommer
Le Song
Sören Sonnenburg
David Sontag
Daniel Soudry
Matthijs Spaan
Henning Sprekeler
Suvrit Sra
Karthik Sridharan
Bharath Sriperumbudur
Michael Stark
Oliver Stegle
Jacob Steinhardt
Florian Steinke
Bastian Steudel
Ian Stevenson
Mark Steyvers
Alan Stocker
Jay Stokes
Karl Stratos
Andreas Stuhlmüller
Juergen Sturm
Erik Sudderth
Mahito Sugiyama
Masashi Sugiyama
Min Sun
Dennis Sun
Liang Sun
Ilya Sutskever
Charles Sutton
Richard Sutton
Johan Suykens
Taiji Suzuki
Kevin Swersky
Zeeshan Syed
Marie Szafranski
Csaba Szepesvari
Arthur Szlam
Raphael Sznitman
Yasuo Tabei
Takashi Takenouchi
Ichiro Takeuchi
Eiji Takimoto
Partha Talukdar
Erik Talvitie
Toshiyuki Tanaka
Yichuan Tang
Cheng Tang
Dacheng Tao
Daniel Tarlow
Masami Tatsuno
Nikolaj Tatti
Graham Taylor
Yee Whye Teh
Matus Telgarsky
Josh Tenenbaum
Choon-Hui Teo
Ambuj Tewari
Lucas Theis
Georgios Theocharous
Bertrand Thirion
Ryan Tibshirani
Robert Tillman
Ivan Titov
Michalis Titsias
Sinisa Todorovic
Ryota Tomioka
Hanghang Tong
Lorenzo Torresani
Ivana Tosic
Behrouz Touri
Long Tran-Thanh
Volker Tresp
Bill Triggs
Ivor Tsang
Ioannis Tsochantaridis
Koji Tsuda
Srini Turaga
Richard Turner
Naonori Ueda
Balazs Ujfalussy
Tomer Ullman
Lyle Ungar
Ruth Urner
Matthew Urry
Raquel Urtasun
Nicolas Usunier
Daniel Vainsencher
Gregory Valiant
Michal Valko
Jan Willem van de Meent
Laurens van der Maaten
Jurgen Van Gael
Herke Van Hoof
Vincent Vanhoucke
Nuno Vasconcelos
Eleni Vasilaki
Andrea Vedaldi
Shankar Vembu
Archana Venkataraman
Dan Ventura
Jakob Verbeek
Nakul Verma
Alessandro Verri
Jean-Philippe Vert
Rene Vidal
Sudheendra Vijayanarasimhan
Silvia Villa
Pascal Vincent
Brett Vinch
S. Vishwanathan
Max Vladymyrov
Joshua Vogelstein
Julia Vogt
Maksims Volkovs
Ed Vul
Sara Wade
Willem Waegeman
Mike Wakin
Guy Wallis

REVIEWERS

Tom Wallis
Tom Walters
Chong Wang
Huan Wang
Huayan Wang
Jack Wang
Jun Wang
Yang Wang
Lei Wang
Lie Wang
Shaojun Wang
Liwei Wang
Tong Wang
Wei Wang
Weiran Wang
Zhaoran Wang
Takashi Washio
Larry Wasserman
Takanori Watanabe
Kazuho Watanabe
Chris Watkins
Kevin Waugh
Chu Wei
Markus Weimer
Kilian Weinberger
Yair Weiss
Zaiwen Wen
Tomas Werner
Martha White
Shimon Whiteson
Andre Wibisono
Michael Wick
Jenna Wiens
Daan Wierstra
Ami Wiesel
Matthew Wilder
Chris Williams
Robert Williamson
Ross Williamson
Sinead Williamson
Andrew Wilson
Robert Wilson
David Wingate
Ole Winther
David Wipf
Frank Wood
Jennifer Wortman Vaughan
Mingrui Wu
Lei Wu
Xiao-Ming Wu
Yihong Wu
Jianxiong Xiao
Lexing Xie
Yu Xin
Linli Xu
Min Xu
Huan Xu
Zenglin Xu
Zhixiang (Eddie) Xu
Makoto Yamada
Yoshihiro Yamanishi
Feng Yan
Keiji Yanai
Pinar Yanardag
Zhi Yang
Eunho Yang
Jimei Yang
Ming-Hsuan Yang
Qiang Yang
Shuanghong Yang
Weilong Yang
Tianbao Yang
Yi Yang
Zhirong Yang
Bangpeng Yao
Angela Yao
Jieping Ye
Ainur Yessenalina
Dit-Yan Yeung
Scott Yih
Junming Yin
Wotao Yin
Yiming Ying
Chang D. Yoo
Junichiro Yoshimoto
Taku Yoshioka
Angela Yu
Byron Yu
Chun-Nam Yu
Kai Yu
Hsiang-Fu Yu
Shipeng Yu
Yao-Liang Yu
Hyokun Yun
Bianca Zadrozny
Thorsten Zander
Giovanni Zappella
Matt Zeiler
Richard Zemel
Kun Zhang
Lei Zhang
Shunan Zhang
Xinhua Zhang
Yichuan Zhang
Yi Zhang
Yuchen Zhang
Yu Zhang
Changshui Zhang
Zhihua Zhang
Lijun Zhang
Zheng Zhao
Alice Zheng
Chunxiao Zhou
Dengyong Zhou
Mingyuan Zhou
Shuheng Zhou
Xueyuen Zhou
Zhi-Hua Zhou
Hankui Zhou
Jun Zhu
Xiaojin Zhu
Xiangxin Zhu
Shenghuo Zhu
Brian Ziebart
Martin Zinkevich
Andrew Zisserman
Larry Zitnick
Onno Zoeter
Daniel Zoran
Or Zuk
Alon Zweig

AUTHOR INDEX



2014 NIPS Conference
Montreal • Canada
Palais des congrès de Montréal