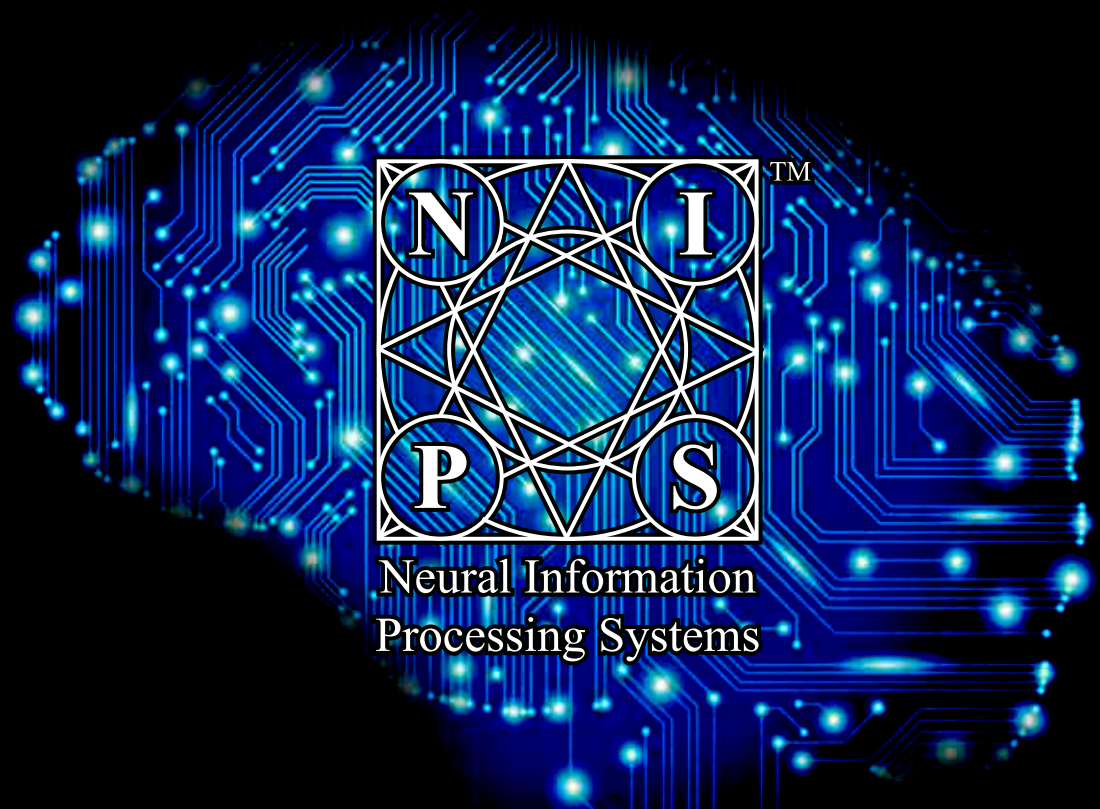
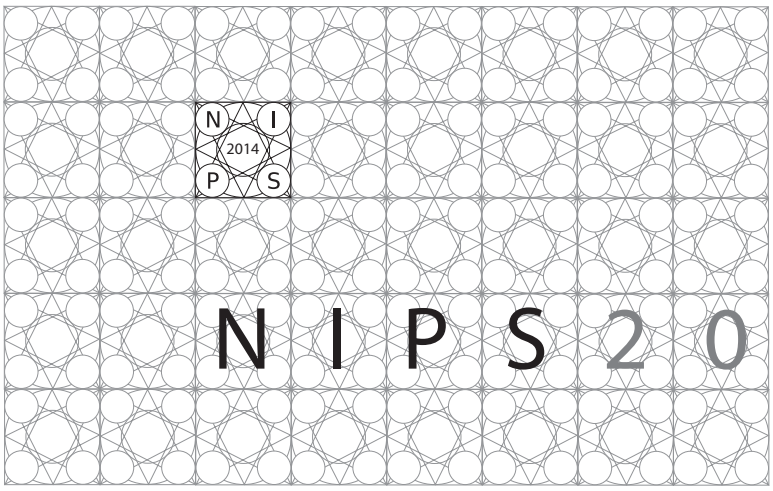
2014 NIPS Conference

MONTREAL 2014

CONFERENCE BOOK





NIPS 2014

Abstracts of Papers

TUTORIALS
December 8, 2014
Palais des Congrès de Montréal
Convention and Exhibition Center,
Montreal, Quebec, Canada

CONFERENCE SESSIONS
December 9 - 11, 2014
Palais des Congrès de Montréal
Convention and Exhibition Center,
Montreal, Quebec, Canada

WORKSHOPS
December 12-13, 2014
Palais des Congrès de Montréal
Convention and Exhibition Center,
Montreal, Quebec, Canada

Sponsored by the Neural Information Processing System Foundation, Inc

The technical program includes 6 invited talks and 414 accepted papers, selected from a total of 1678 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 27," edited by Corinna Cortes, Neil Lawrence, Max Welling, Zoubin Ghahramani and Kilian Weinberger

Neural Information Processing Systems Foundation

TABLE OF CONTENTS

Core Logistics Team 2

Organizing Committee 3

Program Committee 3

NIPS Foundation Offices and Board Members 3

Sponsors 4

PROGRAM HIGHLIGHTS 8

Exhibitors 9

MONDAY

Tutorials

Abstracts 11

Poster Sessions

Location of Presentations 13

Sessions M1 - M100 14

Monday Abstracts 17

TUESDAY

Oral Sessions

Sessions 1 - 4, Abstracts 40

Spotlights Sessions

Sessions 1 - 4, Abstracts 40

Poster Sessions

Sessions T1 - T104 44

Location of Presentations 48

Abstracts 49

Demonstrations 72

WEDNESDAY

Oral Sessions

Sessions 5 - 8, Abstracts 75

Spotlights Sessions

Sessions 5 - 8, Abstracts 75

Poster Sessions

Sessions W1 - W104 79

Location of Presentations 82

Abstracts 83

Demonstrations 105

THURSDAY

Oral Sessions

Sessions 9 - 10, Abstracts 108

Spotlights Sessions

Sessions 9 - 10, Abstracts 108

Poster Sessions

Sessions Thu1 - Thu106 109

Location of Presentations 113

Abstracts 114

Reviewers 138

Author Index 140

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

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

ORGANIZING COMMITTEE

General Chairs: **Zoubin Ghahramani** (University of Cambridge)
Max Welling (University of Amsterdam)
Program Chairs: **Corinna Cortes** (Google Research)
Neil Lawrence (University of Sheffield)
Tutorials Chair: **Jeff A. Blimes** (University of Washington)
Demonstration Chair: **Hugo Larochelle** (Univ. de Sherbrooke)
Workshop Chairs: **Francis Bach** ((INRIA & ENS),
Amir Globerson (Hebrew University)
Publications Chair and Electronic
Proceedings Chair: **Kilian Weinberger** (Washington Univ)
Program Manager: **Alan Saul** (University of Sheffield)

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

EXECUTIVE BOARD
Peter Bartlett, Queensland University and UC Berkeley
Léon Bottou, Microsoft Research
Chris J.C. Burges, Microsoft Research
Fernando Pereira, Google Research
John Shawe-Taylor, University College London
Rich Zemel, University of Toronto
Chris Williams, University of Edinburgh

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

ADVISORY BOARD
Sue Becker, McMaster University, Ontario, Canada
Yoshua Bengio, University of Montreal, Canada
Jack Cowan, University of Chicago
Thomas G. Dietterich, Oregon State University
Stephen Hanson, Rutgers University
Michael I. Jordan, University of California, Berkeley
Michael Kearns, University of Pennsylvania
Scott Kirkpatrick, Hebrew University, Jerusalem
Daphne Koller, Stanford University
John Lafferty, University of Chicago
Todd K. Leen, Oregon Health & Sciences University
Richard Lippmann, Massachusetts Institute of Technology
Bartlett Mel, University of Southern California
John Moody, Intl. Computer Science Inst., Berkeley and Portland
John C. Platt, Microsoft Research
Gerald Tesauro, IBM Watson Labs
Sebastian Thrun, Stanford University
Dave Touretzky, Carnegie Mellon University
Lawrence Saul, University of California, San Diego
Bernhard Schölkopf, Max Planck Institute for Intelligent Systems,
Dale Schuurmans, University of Alberta, Canada
Sara A. Solla, Northwestern University Medical School

PROGRAM COMMITTEE

Jacob Abernethy (University of Michigan)
Ryan Adams (Harvard University)
Deepak Agarwal (LinkedIn)
Edo Airoldi (Harvard University)
Anima Anandkumar (UC Irvine)
Paolo Avesani (FBK - NiLab, Trento, Italy)
Borja Balle (McGill University)
Peter Bartlett (Queensland University/UC Berkeley)
Mikhail Belkin (Ohio State University)
Samy Bengio (Google)
Kristin Bennett (Rensselaer Polytechnic Institute)
Jeff Bilmes (University of Washington)
Matthew Blaschko (Inria Saclay)
David Blei (Princeton)
Karsten Borgwardt (MPI for Intelligent Systems)
Leon Bottou (Microsoft Research)
Guillaume Bouchard (Xerox Research Centre)
Michael Bowling (Alberta)
Sebastien Bubeck (Princeton)
Lawrence Carin (Duke)
Miguel Carreira-Perpinan (UC Merced)
Trevor Cohn (University of Melbourne)
Marco Cuturi (Kyoto University)
Florence d'Alche-Buc (Université d'Evry-Val d'Essonne)
Viriginia De Sa (UCSD)
Arnaud Doucet (Oxford)
François Fleuret (Idiap Research Institute)
Claudio Gentile (DiSTA Universita dell'Insubria)
Mohammad Ghavamzadeh (INRIA & Adobe Research)
Geoff Gordon (Carnegie Mellon)
Dilan Gorur (Yahoo! Labs)
Philipp Hennig (MPI Tuebingen)
Tom Heskes (Radboud)
Bert Huang (University of Maryland)
Christian Igel (University of Copenhagen)
Thorsten Joachims (Cornell)
Brian Kingsbury (IBM Research)
Jyrki Kivinen (University of Helsinki)
Andreas Krause (ETH Zurich)
Brian Kulis (Ohio State University)
Sanjiv Kumar (Google Research)
Gert Lanckriet (UCSD)
Hugo Larochelle (Université de Sherbrooke)
Guy Lebanon (Amazon)
Honglak Lee (University of Michigan)
Robert Legenstein (Graz University of Technology)
Lihong Li (Microsoft Research)
Phil Long (Microsoft)
Aurelie Lozano (IBM Research)
Lester Mackey (Stanford University)
Michael Mahoney (UC Berkeley)
Benjamin Marlin (UMass Amherst)
Mehryar Mohri (Courant Institute and Google Research)
Iain Murray (University of Edinburgh)
Hartmut Neven (Google)
Jennifer Neville (Purdue University)
Sebastian Nowozin (MSR Cambridge)
Guillaume Obozinski (Ecole des Ponts - ParisTech)
Mike Osborne (University of Oxford)
Sinno Pan (National Univ. of Singapore)
Jan Peters (TU Darmstadt)
Joaquin Quinonero-Candela (Facebook)
Marc'Aurelio Ranzato (Facebook)
Steve Renals (University of Edinburgh)
Afshin Rostamizadeh (Google Research)
Dale Schuurmans (University of Alberta)
D Sculley (Google Research)
Ricardo Silva (UCL)
Alex Slivkins (Microsoft Research)
Suvrit Sra (MPI for Intelligent Systems & CMU)
Oliver Stegle (EMBL-EBI)
Amos Storkey (University of Edinburgh)
Charles Sutton (University of Edinburgh)
Umar Syed (Google Research)
Bill Triggs (CNRS)
Koji Tsuda (Advanced Industrial Science and Technology)
Nicolas Usunier (Université de Technologie de Compiègne)
Eleni Vasilaki (University of Sheffield)
Sergei Vassilvitskii (Google)
Aki Vehtari (Aalto University)
Jean-Philippe Vert (Mines ParisTech)
Kilian Weinberger (Washington University)
Yair Weiss (Hebrew University)
Rich Wilkinson (University of Nottingham)
Robert Williamson (NICTA)
David Wipf (Microsoft Research)
Jennifer Wortman Vaughan (Microsoft Research)
John Wright (Columbia University)
Byron Yu (CMU)
Alan Yuille (UCLA)
Richard Zemel (University of Toronto)
Tong Zhang (Baidu/Rutgers)

SPONSORS

NIPS gratefully acknowledges the generosity of those individuals and organizations who have provided financial support for the NIPS 2014 conference. Their financial support enables us to sponsor student travel and participation, the outstanding paper awards, and the volunteers who assist during NIPS.

facebook

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.

amazon.com

Amazon.com (jobs: http://www.amazon.jobs/) 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 fut re development of existing machine learning applications across Amazon, and help develop novel and infinitely-scalable applications.

Google™

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.

KETCHUM TRADING, LLC

Ketchum Trading, LLC is a privately held, proprietary trading firm in Chicago. Driven by algorithms and utilizing the most advanced software and technology in the industry, our multi-strategy firm makes markets and trades in futures, options, cash equities, and exchange traded funds. The fundamentals of Ketchum Trading are simple: manage the risks inherent in modern markets, leverage state-of-the-art technology and proprietary models, and gain a mathematical advantage. The rest takes care of itself.

Microsoft Research

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.

Artificial Intelligence
www.elsevier.com/locate/artint

Artificial Intelligence Journal (AIJ) which commenced publication in 1970, is now the generally accepted premier international forum for the publication of results of current research in this field. The journal welcomes foundational and applied papers describing mature work involving computational accounts of aspects of intelligence. Specifically, it welcomes papers on: AI and Philosophy, automated reasoning and inference, case-based reasoning, cognitive aspects of AI, commonsense reasoning, constraint processing, heuristic search, high-level computer vision, intelligent interfaces, intelligent robotics, knowledge representation, machine learning, multi-agent systems, natural language processing, planning and theories of action, reasoning under uncertainty or imprecision. The journal reports results achieved; proposals for new ways of looking at AI problems must include demonstrations of effectiveness. Papers describing systems or architectures integrating multiple technologies are welcomed. AIJ also invites papers on applications, which should describe a principled solution, emphasize its novelty, and present an in-depth evaluation of the AI techniques being exploited. The journal publishes an annual issue devoted to survey articles and also hosts a "competition section" devoted to reporting results from AI competitions. From time to time, there are special issues devoted to a particular topic; such special issues always have open calls.

Qualcomm Incorporated is the world leader in 3G, 4G and next-generation wireless technologies. Qualcomm innovations are enabling ultra-personal mobile devices; shaping relevant, next-generation mobile experiences; and inspiring transformative new business models and services. Dedicated to accelerating mobile growth and progress worldwide, Qualcomm is transforming the way people live, learn, work and play. Headquartered in San Diego, California, Qualcomm is included in the S&P 500 Index and is a FORTUNE 500(r) company traded on the NASDAQ Stock Market(r) under the ticker symbol QCOM. Qualcomm Research and Development, a division of Qualcomm Technologies, Inc., is where many of the industry's most talented engineers and scientists come to collaborate, share innovative ideas, and create the wireless technologies and solutions that will transform the future of wireless.

Baidu Research, based in Silicon Valley and Beijing, is led by Dr. Andrew Ng, Chief Scientist. Baidu Research comprises three interrelated labs: the Silicon Valley AI Lab, the Institute of Deep Learning and the Big Data Lab, led by Dr. Adam Coates, Dr. Kai Yu and Dr. Tong Zhang, respectively. The organization brings together global research talent to work on fundamental technologies in areas such as image recognition and image-based search, voice recognition, natural language processing and semantic intelligence.

United Technologies Research Center (UTRC) 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 open roles for people with strong machine learning, cloud analytics and HCI skills. If you're technically strong and enjoy continual learning of new application domains, UTRC is the place for you.

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.

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.

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.

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

NVIDIA is a world leader in visual computing. Our technologies are transforming a world of displays into a world of interactive discovery for everyone from gamers and scientists, to consumers and enterprises. We invest in our people, our technologies, and our research and development efforts to deliver the highest quality products to customers across the globe. NVIDIA's culture inspires our team of world-class engineers and developers to be at the top of their game. Data scientists in both industry and academia use GPUs for machine learning to make groundbreaking improvements across a variety of applications including image classification, video analytics, speech recognition and natural language processing. With thousands of computational cores GPUs have become the processor of choice for processing big data for data scientists.

Disney Research

As part of The Walt Disney Company, Disney Research draws on a legacy of innovation and technology leadership that continues to this day. In 1923, Walt Disney sold his animated/live-action series the Alice Comedies, founded his eponymous company, and launched a succession of firsts: The first cartoon with fully synchronized sound (1928). The first full-color cartoon (1932). The first animated feature film (1937). The first modern theme park (1955). The Walt Disney Company was also an early leader in entertainment technology development with inventions like the multiplane camera, Audio-Animatronics, Circle-Vision 360°, and Fantasound. In 2006, The Walt Disney Company acquired Pixar Animation Studios, a move that brought a host of valuable creative and technology assets, including a strong culture of excellence in research. Pixar is a major generator and publisher of world-class research in computer graphics. Its scientists contribute directly to Pixar's critically acclaimed films, consistently winning multiple technical Academy Awards®. The Pixar acquisition was a source of inspiration for the formation of Disney Research, and continues to influence the way we're organized and run. Disney Research was launched in 2008 as an informal network of research labs that collaborate closely with academic institutions such as Carnegie Mellon University and the Swiss Federal Institute of Technology Zürich (ETH). We're able to combine the best of academia and industry: we work on a broad range of commercially important challenges, we view publication as a principal mechanism for quality control, we encourage engagement with the global research community, and our research has applications that are experienced by millions of people. We're honoring Walt Disney's legacy of innovation by researching novel technologies and deploying them on a global scale.

Booking.com

Booking.com was established in 1996 in Amsterdam, The Netherlands and has grown into the world's leading online hotel and accommodation reservations company . We are part of the Priceline Group (NASDAQ: PCLN). Booking.com is still operated from Amsterdam and supported by over 130 offices in over 50 countries. The team 7.000+ people right across the world are dedicated to serving Booking.com's customers and property-owning partners. Our IT teams take care of creating a perfect shopping experience for our customers. Over 100 nationalities are employed at Booking.com. We are a melting pot of people from all over the world working together. Employees are encouraged to develop their own talents. We have a variety of trainings and other methods within our Booking.com University program to help to do just that. Our mission is to help leisure and business travellers, whatever their budgets, easily discover, book, and enjoy the world's best places to stay.Our values describe the "How" of Booking.com's operations: how we behave, how we deal with colleagues, customers, and partners. It basically describes who we are and what drives our behavior.

WINTON

Winton Capital Management is a world leading quantitative investment manager specialising in applying advanced scientific techniques to the analysis of financial markets. The company was founded in 1997 and now employs over 300 professionals across offices in London, Oxford, Zurich, Hong Kong, New York, Tokyo and Sydney.

Oracle Labs

The Mission of Oracle Labs is straightforward: Identify, explore, and transfer new technologies that have the potential to substantially improve Oracle's business. Oracle's commitment to R&D is a driving factor in the development of technologies that have kept Oracle at the forefront of the computer industry. Although many of Oracle's leading-edge technologies originate in its product development organizations, Oracle Labs is the sole organization at Oracle that is devoted exclusively to research. The acquisition of Sun Microsystems, along with dozens of other acquired companies, brought a wide array of technologies to Oracle's portfolio. Oracle executives recognized that in Sun Microsystems Laboratories, Sun brought the combined company the benefits of an independent research organization - now renamed Oracle Labs.

AdRoll

AdRoll's Data Science team is responsible for continuously and automatically leveraging our platform's data to provide actionable insights and models for our business. Our critical mission is to decide the value of displaying our clients' ads to maximize their return on investment. Our proprietary large-scale machine learning infrastructure ingests billions of events with billions of sparse features every few hours to produce models queried thousands of times per second. This scale pushes us to engineer supervised and unsupervised solutions that are frequently beyond the scope of the literature

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

Superfish is a developer and provider of an innovative, next generation visual discovery platform that's out to transform the way images are searched, seen, utilized and shared over the web. Through its cutting-edge, patented technology and pioneering vision, Superfish has developed a cunningly intuitive visual search engine that analyzes images algorithmically, without the need of text tags or human intervention. The company is reinventing image search and focusing on empowering consumers and businesses to experience new ways to engage with and search visual content. As the social web has exploded and photos play such a large role in people's lives online, Superfish sees a huge opportunity in solving new visual search problems. Founded in 2006 by Adi Pinhas and Michael Chertok, Superfish is headquartered in Palo Alto, California and is backed by Draper Fisher Jurvetson (DFJ) and Vintage Investment Partners

BOSCH

Intellectual Inspiration. Entrepreneurial Freedom. Global Opportunities. It's all yours with a career at Bosch. "Made by Bosch" stands for the first-class quality of a global player. The Bosch Research and Technology Center (RTC), with labs in Palo Alto, CA, Pittsburgh, PA, and Cambridge, MA focus on innovative research and development. The Bosch Research and Technology Center focuses on topics such as algorithms for Robotics, Autonomous Systems and Data Mining. Our team of research scientists and software engineers in Palo Alto will grow rapidly in the next couple of years. Therefore, now is the perfect time for you to join and make an impact with your passion to innovate! Please check our website for our current openings: www.boschresearch.com

IBM Research

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.

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 the Xerox services offerings. Today 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.

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.

Symantec

Symantec Corporation (NASDAQ: SYMC) is an information protection expert that helps people, businesses and governments seeking the freedom to unlock the opportunities technology brings – anytime, anywhere. Founded in April 1982, Symantec, a Fortune 500 company, operating one of the largest global data-intelligence networks, has provided leading security, backup and availability solutions for where vital information is stored, accessed and shared. The company's more than 21,500 employees reside in more than 50 countries. Ninety-nine percent of Fortune 500 companies are Symantec customers. In fiscal 2013, it recorded revenues of $6.9 billion. To learn more go to www.symantec.com or connect with Symantec at: go.symantec.com/socialmedia

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 support the NIPS Best Student Awards.

Frontiers in
Distributed Information
Systems Foundation

The focus is on the rapidly changing possibilities in the backbone and transport layers of the world's digital networks, and on the new services, new clients, and new communities that are coming into being as a result of their continuing growth. Opportunities and challenges are presented by high-performance computing and the explosion of digital storage.

PROGRAM HIGHLIGHTS

SUNDAY DEC 7TH

4:00 pm – 8:00 pm
Registration Desk Open, Level 2, room 210

MONDAY DEC 8TH

7:30 am – 6:30 pm
Registration Desk Open, Level 2, room 210

8:00 am – 9:30 am
Breakfast, Level 7, 710

9:30 am – 11:30 am
Climate Change: Challenges for Machine Learning
Arindam Banerjee, Claire Monteleoni

Recent Progress in the Structure of Large-Treewidth Graphs and Some Applications
Chandra Chekuri

10:45 – 11:15 am - Coffee Break

12:05 – 1:00 pm - Lunch Break

1:00 – 3:00 pm
IIntegrated Information Theory of Consciousness - Conceptual Formulation & Adventures in Simulated Evolution
Christof Koch

Non-Parametric Causal Models
Thomas Richardson

3:00 – 3:30 am - Coffee Break

3:30 – 5:30 pm
Computing Game-Theoretic Solutions
Vincent Conitzer

Differential Privacy and Learning: The Tools, The Results, and The Frontier
Katrina Ligett

6:30 – 6:55 pm
Opening Remarks, Awards and Reception

7:00 – 11:59 pm
Poster Session

TUESDAY DEC 9TH

7:30 am – 9:00 am
Breakfast, Level 7, 710

7:30 am – 5:30 pm
Registration Desk Open, Level 2, room 210

9:00 – 10:10 am
Oral Session 1
INVITED TALK: Subgradient Methods for Huge-Scale Optimization Problems
Yurii Nesterov

Provable Submodular Minimization using Wolfe's Algorithm
D. Chakrabarty, P. Jain, P. Kothari

10:10 –10:30 am
Spotlight Session 1: Optimization spotlights

10:30 – 10:55 am - Coffee Break

10:55 –11:40 am
NIPS award Session
Oral Session 2
Median selection subset aggregation for parallel inference
X. Wang, P. Peng, D. Dunson

Asymmetric LSH (ALSH) for Sublinear Time Maximum Inner Product Search (MIPS)
A. Shrivastava, P. Li

11:40 – 12:05 pm
Spotlight Session 2: Large Scale Learning Spotlights

12:05 – 2:00 pm - Lunch Break

2:00 – 3:30 pm
Oral Session 3
INVITED TALK: Role of Coupled Networks in 21st Century Energy Infrastructure
Arunava Majumdar

Analog Memories in a Balanced Rate-Based Network of E-I Neurons
D. Festa, G. Hennequin, M. Lengyel

Feedforward Learning of Mixture Models
M. Lawlor, S. Zucker

3:30 – 3:50 pm
Spotlight Session 3: Neuroscience and Neural Coding

3:50 – 4:15 pm - Coffee Break

4:15 – 5:40 pm
NIPS award Session
Oral Session 4
Conditional Random Field Autoencoders for Unsupervised Structured Predictione
W. Ammar, C. Dyer, N. Smith

Learning Generative Models with Visual Attention
Y. Tang, N. Srivastava, R. Salakhutdinov

Sequence to Sequence Learning with Neural Networks
I. Sutskever, O. Vinyals, Q. Le

Quantifying the Transferability of Features in Deep Neural Networks
J. Yosinski, J. Clune, Y. Bengio, H. Lipson

5:40 – 6:00 pm
Spotlight Session 4: Deep spotlights

7:00 – 11:59 pm
Demonstrations
Poster Session

PROGRAM HIGHLIGHTS

WEDNESDAY DEC 10TH

7:30 am – 9:00 am
Breakfast, Level 7, 710

7:30 am – 5:30 pm
Registration Desk Open, Level 2, room 210

9:00 – 10:10 am
Oral Session 5
INVITED TALK: Networks in Climate Science
John Carlos Baez

Sparse Polynomial Learning and Graph Sketching
M. Kocaoglu, K. Shanmugam, A. Dimakis, A. Klivans

10:10 –10:30 am
Spotlight Session 5: Regression and time series spotlights

10:30 – 10:55 am - Coffee Break

10:55 –11:40 am
Oral Session 6
Combinatorial Pure Exploration of Multi-Armed Bandits
S. Chen, T. Lin, M. Lyu, I. King, W. Chen

From Stochastic Mixability to Fast Rates
N. Mehta, R. Williamson

11:40 – 12:05 pm
Spotlight Session 6: Learning theory spotlights

12:05 – 2:00 pm - Lunch Break

2:00 – 3:30 pm
Oral Session 7
POSNER LECTURE: Using the Emergent Dynamics of Attractor Networks for Computation
John J. Hopfield

"How hard is my MDP?" The distribution-norm to the rescue
T. Mann, O. Maillard, S. Mannor

On Communication Cost of Distributed Statistical Estimation and Dimensionality
A. Garg, T. Ma, H. Nguyen

3:30 – 3:50 pm
Spotlight Session 7: Reinforcement learning spotlights

3:50 – 4:20 pm - Coffee Break

4:15 – 5:40 pm
Oral Session 8
A* Sampling
C. Maddison, D. Tarlow, T. Minka

Asynchronous Anytime Sequential Monte Carlo
B. Paige, F. Wood, A. Doucet, Y. Teh

Probabilistic ODE Solvers with Runge-Kutta Means
M. Schober, D. Duvenaud, P. Hennig

A Wild Bootstrap for Degenerate Kernel Tests
K. Chwialkowski, D. Sejdinovic, A. Gretton

5:40 – 6:00 pm
Spotlight Session 8: GP spotlights, kernel spotlights, sampling spotlights, classification spotlights

7:00 – 11:59 pm
Demonstrations
Poster Session

THURSDAY DEC 11TH

7:30 am – 9:00 am
Breakfast, Level 7, 710

7:30 am – 11:59 am
Registration Desk Open, Level 2, room 210

9:00 – 10:10 am
Oral Session 9
INVITED TALK: Privacy in the Land of Plenty
Cynthia Dwork

Clamping Variables and Approximate Inference
A. Weller, T. Jebara

10:10 –10:30 am
Spotlight Session 9: Graphical models spotlights, model selection

10:30 – 10:50 am - Coffee Break

10:50 am –12:00 pm
Oral Session 10
POSNER LECTURE: Games, Networks, and People
Michael Kearns

A Statistical Decision-Theoretic Framework for Social Choice
H. Azari Soufiani, D. Parkes, L. Xia

12:00 – 12:20 pm
Spotlight Session 10: Cognitive science spotlights

12:20 – 12:30 pm - Closing remarks

12:30 – 2:00 pm - Lunch Break

2:00 – 6:00 pm
Poster Session

2014 EXHIBITORS

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

MONDAY SESSIONS

MONDAY TUTORIALS

Tutorial Session A, 9:30 – 11:30 AM
Session Chair: Kevin P. Murphy

Climate Change: Challenges for Machine Learning
Arindam Banerjee,University of Minnesota
Claire Monteleoni, George Washington Univ.
Location: Level 2, Room 210 a, b

Tutorial Session B, 9:30 – 11:30 AM
Session Chair: Jeff A. Blimes

Recent Progress in the Structure of Large-Treewidth Graphs and Some Applications
Chandra Chekuri, University of Illinois
Location: Level 2, Room 210 e, f

Tutorial Session A, 1:00 – 3:00 PM
Session Chair: Kevin P. Murphy

Integrated Information Theory of Consciousness - Conceptual Formulation & Adventures in Simulated Evolution
Christof Koch, Allen Institute for Brain Science
Location: Level 2, Room 210 a, b

Tutorial Session B, 1:00 – 3:00 PM
Session Chair: Jeff A. Blimes

Non-Parametric Causal Models
Thomas Richardson, University of Washington
Location: Level 2, Room 210 e, f

Tutorial Session A, 3:30 – 5:30 PM
Session Chair: Kevin P. Murphy

Differential Privacy and Learning: The Tools, The Results, and The Frontier
Katrina Ligett, Caltech
Location: Level 2, Room 210 e, f

Tutorial Session B, 3:30 – 5:30 PM
Session Chair: Jeff A. Blimes

Computing Game-Theoretic SolutionsJames
Vincent Conitzer, Duke University
Location: Level 2, Room 210 a, b

ABSTRACTS OF TUTORIALS

Tutorial Session A, 9:30 – 11:30 AM
Session Chair: Kevin P. Murphy

Climate Change: Challenges for Machine Learning

Arindam Banerjee banerjee@cs.umn.edu
University of Minnesota, Twin Cites
Claire Monteleoni cmontel@gwu.edu
George Washington University

Despite the scientific consensus on climate change, drastic uncertainties remain. Crucial questions about changes in regional climate, trends of extreme events such as heat waves, heavy precipitation, and mega-storms, and understanding how climate varied in the distant past, must be answered in order to improve predictions, assess impacts and vulnerability, and aid mitigation and adaptation efforts. Machine learning can help answer such questions and shed light on climate change. Similar to the case of bioinformatics, the study of climate change provides a data-rich scientific domain in which cutting-edge tools from machine learning can make a major impact. Further, such questions give rise to new challenges for the design of machine learning algorithms. This tutorial will give an overview of impactful open questions about climate change, highlight recent successes of machine learning in this domain, and outline significant remaining challenges. Machine learning problems in climate change include prediction, reconstruction, causal attribution, analysis of quantiles and extremes, and exploratory data analysis. Challenges arise because the climate system is extremely complex, comprised of physical processes and their interactions, and the data is massive, high-dimensional, and spatiotemporal, with non-stationarity and potential long-range dependencies over time and space.

Tutorial Session B, 9:30 – 11:30 AM
Session Chair: Jeff A. Blimes

Recent Progress in the Structure of Large-Treewidth Graphs and Some Applications

Chandra Chekuri chekuri@cs.illinois.edu
University of Illinois

Tree decompositions and treewidth play a key role in the seminal work of Robertson and Seymour on graph minors. One of their fundamental results is the Excluded-Grid Theorem which states that every graph G with treewidth at least k contains a f(k) x f(k) grid-minor for some (slowly growing) function f. Treewidth-based ideas have, over the years, yielded many algorithmic and structural results on graphs and related structures. There have been some recent advances in our understanding of the structure of graphs with large treewidth. These advances were partly motivated by the study of polynomial-time approximation algorithms for the maximum disjoint paths problem, culminating in a breakthrough by Chuzhoy in 2011. Subsequent work, building upon some of her ideas and other tools, has led to several new results including a polynomial relationship between treewidth of a graph and the size of its largest grid-minor. The tutorial will provide relevant background on treewidth, describe the highlights of recent developments, and a few applications.

ABSTRACTS OF TUTORIALS

Tutorial Session A, 1:00 – 3:00 PM

Session Chair: Kevin P. Murphy

Integrated Information Theory of Consciousness - Conceptual Formulation & Adventures in Simulated Evolution

Christof Koch koch.christof@gmail.com
Allen Institute for Brain Science

The science of consciousness has made great strides by focusing on the behavioral and neuronal correlates of experience. However, correlates are not enough if we are to understand even basic facts, for example, why the cerebral cortex gives rise to consciousness but the cerebellum does not, though it has even more neurons and appears to be just as complicated. Moreover, correlates are of little help in many instances where we would like to know if consciousness is present: patients with a few remaining islands of functioning cortex, pre-term infants, non-mammalian species, and machines that are rapidly outperforming people at driving, recognizing faces and objects, and answering difficult questions. To address these issues, we need not only more data, but also a theory of consciousness – one that says what experience is and what type of physical systems can have it. Integrated Information Theory (IIT) does so by starting from conscious experience itself via five phenomenological axioms of existence, composition, information, integration, and exclusion. From these it derives five postulates about the properties required of physical mechanisms to support consciousness. The theory provides a principled account of both the quantity and the quality of an individual experience (a quale), and a calculus to evaluate whether or not a particular system of mechanisms is conscious and of what. Moreover, IIT can explain a range of clinical and laboratory findings, makes a number of testable predictions, and extrapolates to a number of unusual conditions. The theory vindicates some intuitions often associated with panpsychism - that consciousness is an intrinsic, fundamental property, is graded, is common among biological organisms, and even some very simple systems may have some of it. However, unlike panpsychism, IIT implies that not everything is conscious, for example aggregates such as heaps of sand, a group of individuals or feed-forward networks, such as deep learning networks. Also, in sharp contrast with widespread functionalist beliefs, IIT implies that digital computers, even if their behavior were to be functionally equivalent to ours, and even if they were to run faithful simulations of the human brain, would experience next to nothing.

Tutorial Session B, 1:00 – 3:00 PM

Session Chair: Jeff A. Blimes

Non-Parametric Causal Models

Thomas Richardson thomasr@u.washington.edu
University of Washington

This tutorial offers a unified introduction to the modern theory of causality based on counterfactuals (aka potential outcomes), directed acyclic graphs (DAGs) and non-parametric structural equation models (NPSEMs). There are very large literatures associated with each of these frameworks, but the connections, which will be highlighted in this tutorial, are often obscure. In the first part of the tutorial we will introduce key concepts and distinctions by considering models that involve only a few variables, this will then provide the foundation for the consideration of more complex models based on graphs in the second. We begin by introducing counterfactuals in the simplest causal model involving a single binary 'treatment' and binary 'outcome'. This then leads naturally to the instrumental variable model, that applies in contexts where, for ethical or practical reasons, we are unable to 'compel' subjects to take a given treatment level (e.g. to take part in an exercise program, or to click on a link). This 'non-compliance' presents a challenge for causal inference since subjects who do participate are often not representative of the population of interest. However, if it possible to find another variable (e.g. payments for participation) that can be randomized and that increases the probability that a subject will take treatment then it is possible to make informative (non-parametric) inferences regarding causal effects. We will see that if the instrument does not directly affect the final outcome then the resulting set of distributions corresponds to a particular convex polytope, more specifically a rhombic dodecahedron, for a discrete instrument with finitely many levels. In the second part we will describe Single-World Intervention Graphs (SWIGs) that provide a simple bridge between potential outcomes and causal graphs. The application of Pearl's d-separation to the SWIG allows us to determine how and when simple adjustment for variables in an observational study may provide a basis for drawing causal conclusions by controlling for confounding. We will then present a simple reformulation of the ID algorithm (due to Tian) that provides a complete answer to the question of whether certain causal queries are non-parametrically identified in the context of a given DAG (that may include hidden variables). We will then discuss the problem of structure identification (aka causal discovery) describing the state of the art in (non-parametric) structure learning algorithms that allow for the presence of hidden variables. Lastly we will given an overview of current research on nested Markov models that encode non-parametric constraints that generalize conditional independence (aka Verma constraints). web-site: http://www.stat.washington.edu/tsr/.

Tutorial Session A, 3:30 – 5:30 PM

Session Chair: Kevin P. Murphy

Differential Privacy and Learning: The Tools, The Results, and The Frontier

Katrina Ligett katrina@caltech.edu
Caltech

When is working with private data safe, and when is it risky? Are the risks inherent to the computation? Widespread availability of detailed personal data makes understanding privacy necessary—an exciting yet daunting challenge. Differential privacy provides a framework for understanding the tradeoff between the loss of privacy for those whose data are input to a computation and the accuracy of that computation's output. This tutorial will not assume familiarity with differential privacy. We will cover the necessary definitions, help build intuition, and introduce the basic differential privacy toolkit. We will then highlight some connections to learning in the existing differential privacy literature, and challenges and open problems for differentially private learning tasks.

Tutorial Session B, 3:30 – 5:30 PM

Session Chair: Jeff A. Blimes

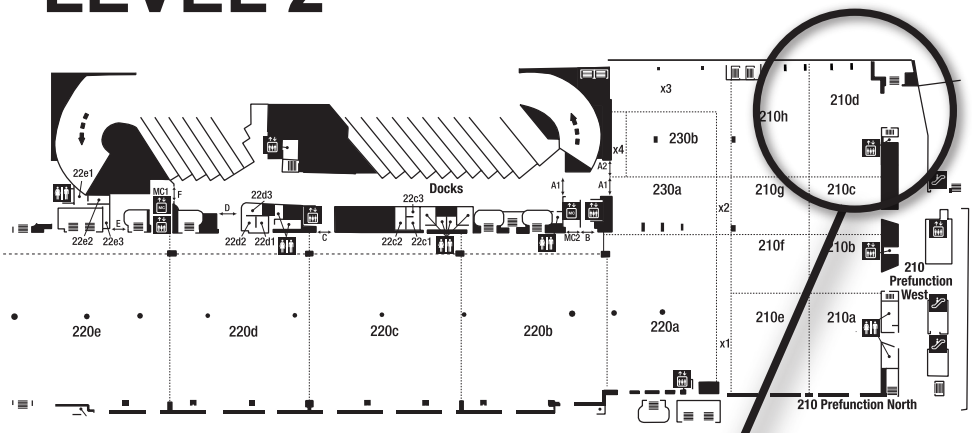
Computing Game-Theoretic Solutions

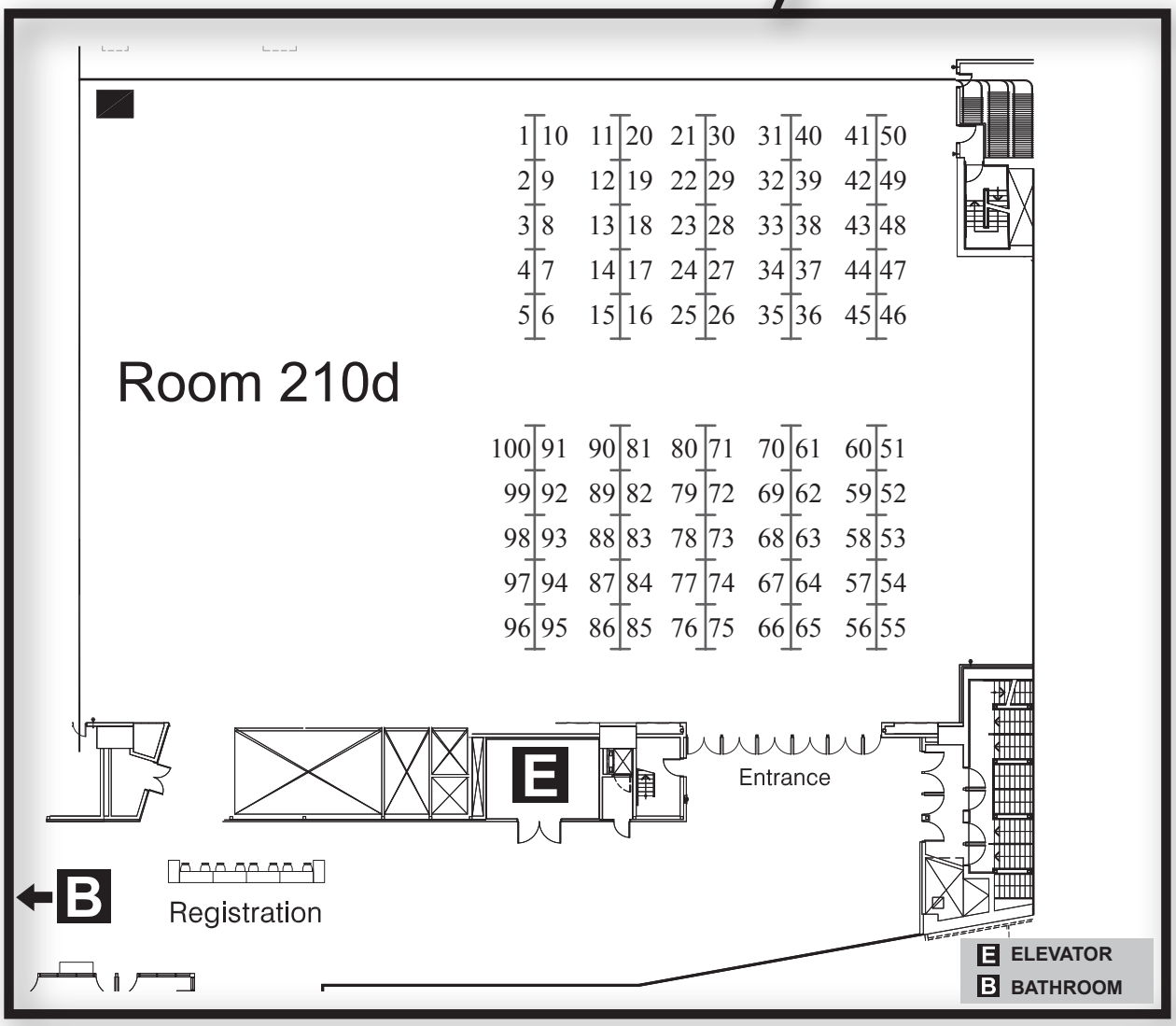
Vincent Conitzer conitzer@cs.duke.edu
Duke University

Game theory concerns how to form beliefs and act in settings with multiple self-interested agents. The best-known solution concept in game theory is that of Nash equilibrium. But there are also others, such as correlated equilibrium and Stackelberg mixed strategies. The latter concept offers a number of conceptual and computational advantages and has recently been deployed in a number of real-world security domains. I will review these concepts as well as algorithms for computing their solutions. I will also discuss learning in games. No previous background in game theory will be required.

MONDAY POSTER FLOORPLAN

LEVEL 2





MONDAY - CONFERENCE

MONDAY, DECEMBER 8TH

6:30 – 6:40PM - OPENING REMARKS
Level 2, Room 210

POSTER SESSION

AND RECEPTION - 7:00 – 11:59 PM

M**1** **Top Rank Optimization in Linear Time**
N. Li, R. Jin, Z. Zhou

M**2** **SerialRank: Spectral Ranking using Seriation**
F. Fogel, A. d'Aspremont, M. Vojnovic

M**3** **Magnitude-sensitive preference formation`**
N. Srivastava, E. Vul, P. Schrater[BR]

M**4** **Learning Mixed Multinomial Logit Model from Ordinal Data**
S. Oh, D. Shah

M**5** **Near-Optimal Density Estimation in Near-Linear Time Using Variable-Width Histograms**
S. Chan, I. Diakonikolas, R. Servedio, X. Sun

M**6** **Factoring Variations in Natural Images with Deep Gaussian Mixture Models**
A. van den Oord, B. Schrauwen

M**7** **Robust Kernel Density Estimation by Scaling and Projection in Hilbert Space**
R. Vandermeulen, C. Scott

M**8** **Distributed Estimation, Information Loss and Exponential Families**
Q. Liu, A. Ihler

M**9** **Sensory Integration and Density Estimation**
J. Makin, P. Sabes

M**10** **General Table Completion using a Bayesian Nonparametric Model**
I. Valera, Z. Ghahramani

M**11** **Dependent nonparametric trees for dynamic hierarchical clustering**
K. Dubey, Q. Ho, S. Williamson, E. Xing

M**12** **Sparse Bayesian Structure Learning with Dependent Relevance Determination Priors**
A. Wu, M. Park, O. Koyejo, J. Pillow

M**13** **Mondrian Forests: Efficient Online Random Forests**
B. Lakshminarayanan, D. Roy, Y. Teh

M**14** **Parallel Sampling of HDPs using Sub-Cluster Splits**
J. Chang, J. Fisher III

M**15** **Localized Data Fusion for Kernel k-Means Clustering with Application to Cancer Biology**
M. Gönen, A. Margolin

M**16** **Learning with Fredholm Kernels**
Q. Que, M. Belkin, Y. Wang

M**17** **Scalable Kernel Methods via Doubly Stochastic Gradients**
B. Dai, B. Xie, N. He, Y. Liang, A. Raj, M. Balcan, L. Song

M**18** **Kernel Mean Estimation via Spectral Filtering**
K. Muandet, B. Sriperumbudur, B. Schölkopf

M**19** **Subspace Embeddings for the Polynomial Kernel**
H. Avron, H. Nguyen, D. Woodruff

M**20** **Learning the Learning Rate for Prediction with Expert Advice**
W. Koolen, T. van Erven, P. Grünwald

M**21** **Delay-Tolerant Algorithms for Asynchronous Distributed Online Learning**
B. McMahan, M. Streeter

M**22** **Efficient Minimax Strategies for Square Loss Games**
W. Koolen, A. Malek, P. Bartlett

M**24** **Online Decision-Making in General Combinatorial Spaces**
A. Rajkumar, S. Agarwal

M**25** **Model-based Reinforcement Learning and the Eluder Dimension**
I. Osband, B. Van Roy

M**26** **Algorithms for CVaR Optimization in MDPs**
Y. Chow, M. Ghavamzadeh

M**27** **Sparse Multi-Task Reinforcement Learning**
D. Calandriello, A. Lazaric, M. Restelli

M**28** **Probabilistic Differential Dynamic Programming**
Y. Pan, E. Theodorou

M**29** **Weighted importance sampling for off-policy learning with linear function approximation**
A. Mahmood, H. van Hasselt, R. Sutton

M**30** **A Representation Theory for Ranking Functions**
H. Pareek, P. Ravikumar

M**31** **Near-Optimal-Sample Estimators for Spherical Gaussian Mixtures**
A. Suresh, A. Orlitsky, J. Acharya, A. Jafarpour

M**32** **Tighten after Relax: Minimax-Optimal Sparse PCA in Polynomial Time**
Z. Wang, H. Lu, H. Liu

M**33** **Consistency of weighted majority votes**
D. Berend, A. Kontorovitch

M**34** **Quantized Estimation of Gaussian Sequence Models in Euclidean Balls**
Y. Zhu, J. Lafferty

M**35** **On the Convergence Rate of Decomposable Submodular Function Minimization**
R. Nishihara, S. Jegelka, M. Jordan

M**36** **Parallel Direction Method of Multipliers**
H. Wang, A. Banerjee, Z. Luo

M37 **Constant Nullspace Strong Convexity and Fast Convergence of Proximal Methods under High-Dimensional Settings**
E. Yen, C. Hsieh, P. Ravikumar, I. Dhillon

M38 **SAGA: A Fast Incremental Gradient Method With Support for Non-Strongly Convex Composite Objectives**
A. Defazio, F. Bach, S. Lacoste-Julien

M39 **Time--Data Tradeoffs by Aggressive Smoothing**
J. Bruer, J. Tropp, V. Cevher, S. Becker

M40 **Local Linear Convergence of Forward--Backward under Partial Smoothness**
J. Liang, J. Fadili, G. Peyré

M41 **Simple MAP Inference via Low-Rank Relaxations**
R. Frostig, S. Wang, P. Liang, C. Manning

M42 **Inferring synaptic conductances from spike trains with a biophysically inspired point process model**
K. Latimer, E. Chichilnisky, F. Rieke, J. Pillow

M43 **Low-dimensional models of neural population activity in sensory cortical circuits**
E. Archer, U. Koster, J. Pillow, J. Macke

M44 **Inferring sparse representations of continuous signals with continuous orthogonal matching pursuit**
K. Knudson, J. Yates, A. Huk, J. Pillow

M45 **Fast Sampling-Based Inference in Balanced Neuronal Networks**
G. Hennequin, L. Aitchison, M. Lengyel

M46 **Information-based learning by agents in unbounded state spaces**
S. Mobin, J. Arnemann, F. Sommer

M47 **On the Computational Efficiency of Training Neural Networks**
R. Livni, S. Shalev-Shwartz, O. Shamir

M48 **Attentional Neural Network: Feature Selection Using Cognitive Feedback**
Q. Wang, J. Zhang, S. Song, Z. Zhang

M49 **Expectation Backpropagation: Parameter-Free Training of Multilayer Neural Networks with Continuous or Discrete Weights**
D. Soudry, I. Hubara, R. Meir

M50 **An Autoencoder Approach to Learning Bilingual Word Representations**
S. Chandar A P, S. Lauly, H. Larochelle, M. Khapra, B. Ravindran, V. Raykar, A. Saha

M51 **Pre-training of Recurrent Neural Networks via Linear Autoencoders**
L. Pasa, A. Sperduti

M52 **Using Convolutional Neural Networks to Recognize Rhythm Stimuli from Electroencephalography Recordings**
S. Stober, D. Cameron, J. Grahn

M53 **Neurons as Monte Carlo Samplers: Bayesian Inference and Learning in Spiking Networks**
Y. Huang, R. Rao

M54 **A framework for studying synaptic plasticity with neural spike train data**
S. Linderman, C. Stock, R. Adams

M55 **Global Belief Recursive Neural Networks**
R. Paulus, R. Socher, C. Manning

M56 **Deep Networks with Internal Selective Attention through Feedback Connections**
M. Stollenga, J. Masci, F. Gomez, J. Schmidhuber

M57 **Iterative Neural Autoregressive Distribution Estimator NADE-k**
T. Raiko, Y. Li, K. Cho, Y. Bengio

M58 **General Stochastic Networks for Classification**
M. Zöhrer, F. Pernkopf

M59 **Improved Multimodal Deep Learning with Variation of Information**
K. Sohn, W. Shang, H. Lee

M60 **Restricted Boltzmann machines modeling human choice**
T. Osogami, M. Otsuka

M61 **Deep Fragment Embeddings for Bidirectional Image Sentence Mapping**
A. Karpathy, A. Joulin, F. Li

M62 **Recursive Context Propagation Network for Semantic Scene Labeling**
A. Sharma, O. Tuzel, M. Liu

M63 **Multiscale Fields of Patterns**
P. Felzenszwalb, J. Oberlin

M64 **Weakly-supervised Discovery of Visual Pattern Configurations**
H. Song, Y. Lee, S. Jegelka, T. Darrell

M65 **Encoding High Dimensional Local Features by Sparse Coding Based Fisher Vectors**
L. Liu, C. Shen, L. Wang, A. van den Hengel, C. Wang

M66 **Self-Adaptable Templates for Feature Coding**
X. Boix, G. Roig, S. Diether, L. Gool

M67 **Learning From Weakly Supervised Data by The Expectation Loss SVM (e-SVM) algorithm**
J. Zhu, J. Mao, A. Yuille

M68 **Predicting Useful Neighborhoods for Lazy Local Learning**
A. Yu, K. Grauman

M69 **A Unified Semantic Embedding: Relating Taxonomies and Attributes**
S. Hwang, L. Sigal

M70 **Zero-shot recognition with unreliable attributes**
D. Jayaraman, K. Grauman

M71 **Articulated Pose Estimation by a Graphical Model with Image Dependent Pairwise Relations**
X. Chen, A. Yuille

M72 **Exact Post Model Selection Inference for Marginal Screening**
J. Lee, J. Taylor

M73 **On Iterative Hard Thresholding Methods for High-dimensional M-Estimation**
P. Jain, A. Tewari, P. Kar

M74 **A Safe Screening Rule for Sparse Logistic Regression**
J. Wang, J. Zhou, J. Liu, P. Wonka, J. Ye

M75 **Generalized Dantzig Selector: Application to the k-support norm**
S. Chatterjee, S. Chen, A. Banerjee

M76 **Parallel Feature Selection Inspired by Group Testing**
Y. Zhou, U. Porwal, C. Zhang, H. Ngo, L. Nguyen, C. Ré, V. Govindaraju

M77 **Spectral k-Support Norm Regularization**
A. McDonald, M. Pontil, D. Stamos

M78 **Beta-Negative Binomial Process and Exchangeable Random Partitions for Mixed-Membership Modeling**
M. Zhou

M79 **The Infinite Mixture of Infinite Gaussian Mixtures**
H. Yerebakan, B. Rajwa, M. Dundar

M80 **Capturing Semantically Meaningful Word Dependencies with an Admixture of Poisson MRFs**
D. Inouye, P. Ravikumar, I. Dhillon

M81 **Dynamic Rank Factor Model for Text Streams**
S. Han, L. Du, E. Salazar, L. Carin

M82 **A provable SVD-based algorithm for learning topics in dominant admixture corpus**
T. Bansal, C. Bhattacharyya, R. Kannan

M83 **Learning a Concept Hierarchy from Multi-labeled Documents**
V. Nguyen, J. Boyd-Graber, P. Resnik, J. Chang

M84 **On a Theory of Nonparametric Pairwise Similarity for Clustering: Connecting Clustering to Classification**
Y. Yang, F. Liang, S. Yan, Z. Wang, T. Huang

M85 **Robust Bayesian Max-Margin Clustering**
C. Chen, J. Zhu, X. Zhang

M86 **On Integrated Clustering and Outlier Detection**
L. Ott, L. Pang, F. Ramos, S. Chawla

M87 **Convex Optimization Procedure for Clustering: Theoretical Revisit**
C. Zhu, H. Xu, C. Leng, S. Yan

M88 **Greedy Subspace Clustering**
D. Park, C. Caramanis, S. Sanghavi

M89 **Graph Clustering With Missing Data: Convex Algorithms and Analysis**
R. Korlakai Vinayak, S. Oymak, B. Hassibi

M90 **Dimensionality Reduction with Subspace Structure Preservation**
D. Arpit, I. Nwogu, V. Govindaraju

M91 **Distributed Balanced Clustering via Mapping Coresets**
M. Bateni, A. Bhaskara, S. Lattanzi, V. Mirrokni

M92 **Zeta Hull Pursuits: Learning Nonconvex Data Hulls**
Y. Xiong, W. Liu, D. Zhao, X. Tang

M93 **The Bayesian Case Model: A Generative Approach for Case-Based Reasoning and Prototype Classification**
B. Kim, C. Rudin, J. Shah

M94 **Consistency of Spectral Partitioning of Uniform Hypergraphs under Planted Partition Model**
D. Ghoshdastidar, A. Dukkipati

M95 **Projecting Markov Random Field Parameters for Fast Mixing**
X. Liu, J. Domke

M96 **Blossom Tree Graphical Models**
Z. Liu, J. Lafferty

M97 **Distributed Parameter Estimation in Probabilistic Graphical Models**
Y. Mizrahi, M. Denil, N. de Freitas

M98 **Elementary Estimators for Graphical Models**
E. Yang, A. Lozano, P. Ravikumar

M99 **Structure learning of antiferromagnetic Ising models**
G. Bresler, D. Gamarnik, D. Shah

M100 **On Sparse Gaussian Chain Graph Models**
C. McCarter, S. Kim

M1 Top Rank Optimization in Linear Time

Nan Li lin@lamda.nju.edu.cn
Zhi-Hua Zhou zhouzh@lamda.nju.edu.cn
Nanjing University, China
Rong Jin rong+@cs.cmu.edu
Michigan State University (MSU)

Bipartite ranking aims to learn a real-valued ranking function that orders positive instances before negative instances. Recent efforts of bipartite ranking are focused on optimizing ranking accuracy at the top of the ranked list. Most existing approaches are either to optimize task specific metrics or to extend the rank loss by emphasizing more on the error associated with the top ranked instances, leading to a high computational cost that is super-linear in the number of training instances. We propose a highly efficient approach, titled TopPush, for optimizing accuracy at the top that has computational complexity linear in the number of training instances. We present a novel analysis that bounds the generalization error for the top ranked instances for the proposed approach. Empirical study shows that the proposed approach is highly competitive to the state-of-the-art approaches and is 10-100 times faster.

M2 SerialRank: Spectral Ranking using Seriation

Fajwel Fogel fajwel.fogel@ens.fr
École Polytechnique
Alexandre d'Aspremont aspremon@princeton.edu
CNRS - ENS
Milan Vojnovic milanv@microsoft.com
Microsoft Research

We describe a seriation algorithm for ranking a set of n items given pairwise comparisons between these items. Intuitively, the algorithm assigns similar rankings to items that compare similarly with all others. It does so by constructing a similarity matrix from pairwise comparisons, using seriation methods to reorder this matrix and construct a ranking. We first show that this spectral seriation algorithm recovers the true ranking when all pairwise comparisons are observed and consistent with a total order. We then show that ranking reconstruction is still exact even when some pairwise comparisons are corrupted or missing, and that seriation based spectral ranking is more robust to noise than other scoring methods. An additional benefit of the seriation formulation is that it allows us to solve semi-supervised ranking problems. Experiments on both synthetic and real datasets demonstrate that seriation based spectral ranking achieves competitive and in some cases superior performance compared to classical ranking methods.

M3 Magnitude-sensitive preference formation`

Nisheeth Srivastava nisheeths@gmail.com
Ed Vul evul@ucsd.edu
University of California, San Diego
Paul Schrater schrater@umn.edu
University of Minnesota

Our understanding of the neural computations that underlie the ability of animals to choose among options has advanced through a synthesis of computational modeling, brain imaging and behavioral choice experiments. Yet, there remains a gulf between theories of preference learning and accounts of the real, economic choices that humans face in daily life, choices that are usually between some amount of money and an item. In this paper, we develop a theory of magnitude-sensitive preference learning that permits an agent to rationally infer its preferences for items compared with money options of different magnitudes. We show how this theory yields classical and anomalous supply-demand curves and predicts choices for a large panel of risky lotteries. Accurate replications of such phenomena without recourse to utility functions suggest that the theory proposed is both psychologically realistic and econometrically viable.

M4 Learning Mixed Multinomial Logit Model from Ordinal Data

Sewoong Oh swoh@illinois.edu
UIUC
Devavrat Shah devavrat@mit.edu
Massachusetts Institute of Technology

Motivated by generating personalized recommendations using ordinal (or preference) data, we study the question of learning a mixture of MultiNomial Logit (MNL) model, a parameterized class of distributions over permutations, from partial ordinal or preference data (e.g. pair-wise comparisons). Despite its long standing importance across disciplines including social choice, operations research and revenue management, little is known about this question. In case of single MNL models (no mixture), computationally and statistically tractable learning from pair-wise comparisons is feasible. However, even learning mixture of two MNL model is infeasible in general. Given this state of affairs, we seek conditions under which it is feasible to learn the mixture model in both computationally and statistically efficient manner. To that end, we present a sufficient condition as well as an efficient algorithm for learning mixed MNL models from partial preferences/comparisons data. In particular, a mixture of r MNL components over n objects can be learnt using samples whose size scales polynomially in n and r (concretely, $n^3 r^{3.5} \log^4 n$, with $r \ll n^{2/7}$ when the model parameters are sufficiently {\em incoherent}). The algorithm has two phases: first, learn the pair-wise marginals for each component using tensor decomposition; second, learn the model parameters for each component using RankCentrality introduced by Negahban et al. In the process of proving these results, we obtain a generalization of existing analysis for tensor decomposition to a more realistic regime where only partial information about each sample is available.

M5 Near-Optimal Density Estimation in Near-Linear Time Using Variable-Width Histograms

Siu On Chan sochan@gmail.com
Microsoft Research
Ilias Diakonikolas iliasdiakonikolas@gmail.com
University of Edinburgh
Rocco Servedio rocco@cs.columbia.edu
Xiaorui Sun xiaoruisun@cs.columbia.edu
Columbia University

Let p be an unknown and arbitrary probability distribution over $[0,1)$. We consider the problem of**density estimation**, in which a learning algorithm is given i.i.d. draws from p and must (with high probability) output a hypothesis distribution that is close to p. The main contribution of this paper is a highly efficient density estimation algorithm for learning using a variable-width histogram, i.e., a hypothesis distribution with a piecewise constant probability density function. In more detail, for any $k$and $\eps$, we give an algorithm that makes $O\tilde{}(k/\eps^2)$ draws from p, runs in $O\tilde{}(k/\eps^2)$ time, and outputs a hypothesis distribution h that is piecewise constant with $O(k\log^2(1/\eps))$ pieces. With high probability the hypothesis h satisfies $\dtv(p,h)\leq C\cdot\opt_k(p)+\eps$, where $\dtv$ denotes the total variation distance (statistical distance), C is a universal constant, and $\opt_k(p)$ is the smallest total variation distance between p and any k-piecewise constant distribution. The sample size and running time of our algorithm are both optimal up to logarithmic factors. The ``approximation factor'' C that is present in our result is inherent in the problem, as we prove that no algorithm with sample size bounded in terms of k and $\eps$ can achieve $C<2$ regardless of what kind of hypothesis distribution it uses.

M6 Factoring Variations in Natural Images with Deep Gaussian Mixture Models

Aaron van den Oord aaron.vandenoord@ugent.be
Benjamin Schrauwen Benjamin.Schrauwen@ugent.be
Ghent University

Generative models can be seen as the swiss army knives of machine learning, as many problems can be written probabilistically in terms of the distribution of the data, including prediction, reconstruction, imputation and simulation. One of the most promising directions for unsupervised learning may lie in Deep Learning methods, given their success in supervised learning. However, one of the current problems with deep unsupervised learning methods, is that they often are harder to scale. As a result there are some easier, more scalable shallow methods, such as the Gaussian Mixture Model and the Student-t Mixture Model, that remain surprisingly competitive. In this paper we propose a new scalable deep generative model for images, called the Deep Gaussian Mixture Model, that is a straightforward but powerful generalization of GMMs to multiple layers. The parametrization of a Deep GMM allows it to efficiently capture products of variations in natural images. We propose a new EM-based algorithm that scales well to large datasets, and we show that both the Expectation and the Maximization steps can easily be distributed over multiple machines. In our density estimation experiments we show that deeper GMM architectures generalize better than more shallow ones, with results in the same ballpark as the state of the art.

M7 Robust Kernel Density Estimation by Scaling and Projection in Hilbert Space

Robert Vandermeulen rvdm@umich.edu
Clay Scott cscott@eecs.umich.edu
University of Michigan

While robust parameter estimation has been well studied in parametric density estimation, there has been little investigation into robust density estimation in the nonparametric setting. We present a robust version of the popular kernel density estimator (KDE). As with other estimators, a robust version of the KDE is useful since sample contamination is a common issue with datasets. What ``robustness'' means for a nonparametric density estimate is not straightforward and is a topic we explore in this paper. To construct a robust KDE we scale the traditional KDE and project it to its nearest weighted KDE in the L_2 norm. Because the squared L_2 norm penalizes point-wise errors superlinearly this causes the weighted KDE to allocate more weight to high density regions. We demonstrate the robustness of the SPKDE with numerical experiments and a consistency result which shows that asymptotically the SPKDE recovers the uncontaminated density under sufficient conditions on the contamination.

M8 Distributed Estimation, Information Loss and Exponential Families

Qiang Liu qliu1@uci.edu
Alex Ihler ihler@ics.uci.edu
UC Irvine

Distributed learning of probabilistic models from multiple data repositories with minimum communication is increasingly important. We study a simple communication-efficient learning framework that first calculates the local maximum likelihood estimates (MLE) based on the data subsets, and then combines the local MLEs to achieve the best possible approximation to the global MLE, based on the whole dataset jointly. We study the statistical properties of this framework, showing that the loss of efficiency compared to the global setting relates to how much the underlying distribution families deviate from full exponential families, drawing connection to the theory of information loss by Fisher, Rao and Efron. We show that the “full-exponential-family-ness” represents the lower bound of the error rate of arbitrary combinations of local MLEs, and is achieved by a KL-divergence-based combination method but not by a more common linear combination method. We also study the empirical properties of the KL and linear combination methods, showing that the KL method significantly outperforms linear combination in practical settings with issues such as model misspecification, non-convexity, and heterogeneous data partitions.

M9 Sensory Integration and Density Estimation

Joseph Makin makin@phy.ucsf.edu
Philip Sabes sabes@phy.ucsf.edu
UCSF

The integration of partially redundant information from multiple sensors is a standard computational problem for agents interacting with the world. In man and other primates, integration has been shown psychophysically to be nearly optimal in the sense of error minimization.

An influential generalization of this notion of optimality is that populations of multisensory neurons should retain all the information from their unisensory afferents about the underlying, common stimulus [1]. More recently, it was shown empirically that a neural network trained to perform latent-variable density estimation, with the activities of the unisensory neurons as observed data, satisfies the information-preservation criterion, even though the model architecture was not designed to match the true generative process for the data [2]. We prove here an analytical connection between these seemingly different tasks, density estimation and sensory integration; that the former implies the latter for the model used in [2]; but that this does not appear to be true for all models.

M10 General Table Completion using a Bayesian Nonparametric Model

Isabel Valera ivalera@tsc.uc3m.es
UC3M
Zoubin Ghahramani zoubin@eng.cam.ac.uk
University of Cambridge

Even though heterogeneous databases can be found in a broad variety of applications, there exists a lack of tools for estimating missing data in such databases. In this paper, we provide an efficient and robust table completion tool, based on a Bayesian nonparametric latent feature model. In particular, we propose a general observation model for the Indian buffet process (IBP) adapted to mixed continuous (real-valued and positive real-valued) and discrete (categorical, ordinal and count) observations. Then, we propose an inference algorithm that scales linearly with the number of observations. Finally, our experiments over five real databases show that the proposed approach provides more robust and accurate estimates than the standard IBP and the Bayesian probabilistic matrix factorization with Gaussian observations.

M11 Dependent nonparametric trees for dynamic hierarchical clustering

Kumar Dubey avinava.dubey@gmail.com
Carnegie Mellon University
Qirong Ho hoqirong@gmail.com
Institute for Infocomm Research, A*STAR
Sinead Williamson
sinead.williamson@mccombs.utexas.edu
University of Texas at Austin
Eric Xing epxing@cs.cmu.edu
CMU

Hierarchical clustering methods offer an intuitive and powerful way to model a wide variety of data sets. However, the assumption of a fixed hierarchy is often overly restrictive when working with data generated over a period of time: We expect both the structure of our hierarchy, and the parameters of the clusters, to evolve with time. In this paper, we present a distribution over collections of time-dependent, infinite-dimensional trees that can be used to model evolving hierarchies, and present an efficient and scalable algorithm for performing approximate inference in such a model. We demonstrate the efficacy of our model and inference algorithm on both synthetic data and real-world document corpora.

M12 Sparse Bayesian Structure Learning with Dependent Relevance Determination Priors

Anqi Wu wuanqi1129@gmail.com
Jonathan Pillow pillow@mail.utexas.edu
UT Austin
Mijung Park mijung@gatsby.ucl.ac.uk
University College London
Sanmi Koyejo sanmi.k@gmail.com
Stanford University

In many problem settings, vectors of regression weights are not merely sparse, but are also dependent in such a way that non-zero weights tend to cluster together. We refer to this local form of dependency as “region sparsity”. Region sparsity differs from “group sparsity” in that it does not require a prior specification of groups, only a notion of distance between coefficients, such as one finds in spatial and time series regression problems. Classical methods for sparse regression, such as automatic relevance determination and the Lasso, do not exploit such dependencies, and effectively model regression weights as independent a priori. Here we introduce a hierarchical model for smooth, region-sparse weight tensors. Our approach represents a hierarchical extension of Tipping’s relevance determination framework, in which we use a transformed Gaussian process to describe the dependencies between the prior variances of nearby weights. To simultaneously impose smoothness, we employ a structured model of the prior variances of Fourier coefficients, which prunes high frequencies and encourages estimates to be sparse in two bases simultaneously. We develop efficient approximate inference methods and show substantial improvements over comparable methods (e.g., group Lasso and smooth RVM) on both simulated and real datasets from brain imaging.

M13 Mondrian Forests: Efficient Online Random Forests

Balaji Lakshminarayanan balaji@gatsby.ucl.ac.uk
Gatsby Unit, University College London
Dan Roy droy@utstat.toronto.edu
University of Toronto
Yee Whye Teh y.w.teh@stats.ox.ac.uk
University of Oxford

Ensembles of randomized decision trees, usually referred to as random forests, are widely used for classification and regression tasks in machine learning and statistics. Random forests achieve competitive predictive performance and are computationally efficient to train and test, making them excellent candidates for real-world prediction tasks. The most popular random forest variants (such as Breiman’s random forest and extremely randomized trees) operate on batches of training data. Online methods are now in greater demand. Existing online random forests, however, require more training data than their batch counterpart to achieve comparable predictive performance. In this work, we use Mondrian processes (Roy and Teh, 2009) to construct ensembles of random decision trees we call Mondrian forests. Mondrian forests can be grown in an incremental/online fashion and remarkably, the distribution of online Mondrian forests is the same as that of batch Mondrian forests. Mondrian forests achieve competitive predictive performance comparable with existing online random forests and periodically re-trained batch random forests, while being more than an order of magnitude faster, thus representing a better computation vs accuracy tradeoff.

M14 Parallel Sampling of HDPs using Sub-Cluster Splits

Jason Chang jchang7@csail.mit.edu
CSAIL, MIT
John Fisher III fisher@csail.mit.edu
Massachusetts Institute of Technology

We develop a sampling technique for Hierarchical Dirichlet process models. The parallel algorithm builds upon [Chang & Fisher 2013] by proposing large split and merge moves based on learned sub-clusters. The additional global split and merge moves drastically improve convergence in the experimental results. Furthermore, we discover that cross-validation techniques do not adequately determine convergence, and that previous sampling methods converge slower than were previously expected.

M15 Localized Data Fusion for Kernel k-Means Clustering with Application to Cancer Biology

Mehmet Gönen gonen@ohsu.edu
Adam Margolin margolin@ohsu.edu
Oregon Health & Science University

In many modern applications from, for example, bioinformatics and computer vision, samples have multiple feature representations coming from different data sources. Multiview learning algorithms try to exploit all these available information to obtain a better learner in such scenarios. In this paper, we propose a novel multiple kernel learning algorithm that extends kernel k-means clustering to the multiview setting, which combines kernels calculated on the views in a localized way to better capture sample-specific characteristics of the data. We demonstrate the better performance of our localized data fusion approach on a human colon and rectal cancer data set by clustering patients. Our method finds more relevant prognostic patient groups than global data fusion methods when we evaluate the results with respect to three commonly used clinical biomarkers.

M16 Learning with Fredholm Kernels

Qichao Que que@cse.ohio-state.edu
Mikhail Belkin mbelkin@cse.ohio-state.edu
Yusu Wang yusu@cse.ohio-state.edu
Ohio State University

In this paper we propose a framework for supervised and semi-supervised learning based on reformulating the learning problem as a regularized Fredholm integral equation. Our approach fits naturally into the kernel framework and can be interpreted as constructing new data-dependent kernels, which we call Fredholm kernels. We proceed to discuss the “noise assumption” for semi-supervised learning and provide evidence evidence both theoretical and experimental that Fredholm kernels can effectively utilize unlabeled data under the noise assumption. We demonstrate that methods based on Fredholm learning show very competitive performance in the standard semi-supervised learning setting.

M17 Scalable Kernel Methods via Doubly Stochastic Gradients

Bo Dai bodai@gatech.edu
Bo Xie zixu1986@gmail.com
Maria-Florina Balcan ninamf@cc.gatech.edu
Le Song l song@cc.gatech.edu
Georgia Tech
Niao He nhe6@isye.gatech.edu
Gatech
Yingyu Liang yingyul@cs.princeton.edu
Princeton University
Anant Raj araj34@mail.gatech.edu
IIT, Kanpur

The general perception is that kernel methods are not scalable, so neural nets become the choice for large-scale nonlinear learning problems. Have we tried hard enough for kernel methods? In this paper, we propose an approach that scales up kernel methods using a novel concept called ``doubly stochastic functional gradients''. Based on the fact that many kernel methods can be expressed as convex optimization problems, our approach solves the optimization problems by making two unbiased stochastic approximations to the functional gradient---one using random training points and another using random features associated with the kernel---and performing descent steps with this noisy functional gradient. Our algorithm is simple, need no commit to a preset number of random features, and allows the flexibility of the function class to grow as we see more incoming data in the streaming setting. We demonstrate that a function learned by this procedure after t iterations converges to the optimal function in the reproducing kernel Hilbert space in rate O(1/t), and achieves a generalization bound of O(1/\sqrt{t}). Our approach can readily scale kernel methods up to the regimes which are dominated by neural nets. We show competitive performances of our approach as compared to neural nets in datasets such as 2.3 million energy materials from MolecularSpace, 8 million handwritten digits from MNIST, and 1 million photos from ImageNet using convolution features.

M18 Kernel Mean Estimation via Spectral Filtering

Krikamol Muandet krikamol@gmail.com
Bernhard Schölkopf bs@tuebingen.mpg.de
Max Planck Institute for Intelligent Systems
Bharath Sriperumbudur bharathsv.ucsd@gmail.com
Pennsylvania State University

The problem of estimating the kernel mean in a reproducing kernel Hilbert space (RKHS) is central to kernel methods in that it is used by classical approaches (e.g., when centering a kernel PCA matrix), and it also forms the core inference step of modern kernel methods (e.g., kernel-based non-parametric tests) that rely on embedding probability distributions in RKHSs. Previous work [1] has shown that shrinkage can help in constructing “better” estimators of the kernel mean than the empirical estimator. The present paper studies the consistency and admissibility of the estimators in [1], and proposes a wider class of shrinkage estimators that improve upon the empirical estimator by considering appropriate basis functions. Using the kernel PCA basis, we show that some of these estimators can be constructed using spectral

filtering algorithms which are shown to be consistent under some technical assumptions. Our theoretical analysis also reveals a fundamental connection to the kernel-based supervised learning framework. The proposed estimators are simple to implement and perform well in practice.

M19 Subspace Embeddings for the Polynomial Kernel

Haim Avron haim.avron@gmail.com
David Woodruff dpwoodru@us.ibm.com
IBM Research
Huy Nguyen huylenguyen@gmail.com
University of California, Berkeley

Sketching is a powerful dimensionality reduction tool for accelerating statistical learning algorithms. However, its applicability has been limited to a certain extent since the crucial ingredient, the so-called oblivious subspace embedding, can only be applied to data spaces with an explicit representation as the column span or row span of a matrix, while in many settings learning is done in a high-dimensional space implicitly defined by the data matrix via a kernel transformation. We propose the first {\em fast} oblivious subspace embeddings that are able to embed a space induced by a non-linear kernel {\em without} explicitly mapping the data to the high-dimensional space. In particular, we propose an embedding for mappings induced by the polynomial kernel. Using the subspace embeddings, we obtain the fastest known algorithms for computing an implicit low rank approximation of the higher-dimension mapping of the data matrix, and for computing an approximate kernel PCA of the data, as well as doing approximate kernel principal component regression.

M20 Learning the Learning Rate for Prediction with Expert Advice

Wouter Koolen wouter.koolen@qut.edu.au
Queensland University of Technology
Tim van Erven tim@timvanerven.nl
Leiden University
Peter Grünwald pdg@cwi.nl
CWI and Leiden University

Most standard algorithms for prediction with expert advice depend on a parameter called the learning rate. This learning rate needs to be large enough to fit the data well, but small enough to prevent overfitting. For the exponential weights algorithm, a sequence of prior work has established theoretical guarantees for higher and higher data-dependent tunings of the learning rate, which allow for increasingly aggressive learning. But in practice such theoretical tunings often still perform worse (as measured by their regret) than ad hoc tuning with an even higher learning rate. To close the gap between theory and practice we introduce an approach to learn the learning rate. Up to a factor that is at most (poly)logarithmic in the number of experts and the inverse of the learning rate, our method performs as well as if we would know the empirically best learning rate from a large range that includes both conservative small values and values that are much higher than those for which formal guarantees were previously available. Our method employs a grid of learning rates, yet runs in linear time regardless of the size of the grid.

M21 Delay-Tolerant Algorithms for Asynchronous Distributed Online Learning

Brendan McMahan mcmahan@google.com
Google
Matthew Streeter matts@cs.cmu.edu
Duolingo

We analyze new online gradient descent algorithms for distributed systems with large delays between gradient computations and the corresponding updates. Using insights from adaptive gradient methods, we develop algorithms that adapt not only to the sequence of gradients, but also to the precise update delays that occur. We first give an impractical algorithm that achieves a regret bound that precisely quantifies the impact of the delays. We then analyze AdaptiveRevision, an algorithm that is efficiently implementable and achieves comparable guarantees. The key algorithmic technique is appropriately and efficiently revising the learning rate used for previous gradient steps. Experimental results show when the delays grow large (1000 updates or more), our new algorithms perform significantly better than standard adaptive gradient methods.

M22 Efficient Minimax Strategies for Square Loss Games

Wouter Koolen wouter.koolen@qut.edu.au
Queensland University of Technology
Alan Malek malek@berkeley.edu
Peter Bartlett bartlett@cs.berkeley.edu
UC Berkeley

We consider online prediction problems where the loss between the prediction and the outcome is measured by the squared Euclidean distance and its generalization, the squared Mahalanobis distance. We derive the minimax solutions for the case where the prediction and action spaces are the simplex (this setup is sometimes called the Brier game) and the ℓ_2 ball (this setup is related to Gaussian density estimation). We show that in both cases the value of each sub-game is a quadratic function of a simple statistic of the state, with coefficients that can be efficiently computed using an explicit recurrence relation. The resulting deterministic minimax strategy and randomized maximin strategy are linear functions of the statistic.

M24 Online Decision-Making in General Combinatorial Spaces

Arun Rajkumar arun_r@csa.iisc.ernet.in
Shivani Agarwal shivani@csa.iisc.ernet.in
Indian Institute of Science

We study online combinatorial decision problems, where one must make sequential decisions in some combinatorial space without knowing in advance the cost of decisions on each trial; the goal is to minimize the total regret over some sequence of trials relative to the best fixed decision in hindsight. Such problems have been studied mostly in settings where decisions are represented by Boolean vectors and costs are linear in this representation. Here we study a general setting where costs may be linear in any suitable low-dimensional vector representation of elements of the decision space. We give a general algorithm for such problems that we call low-dimensional online mirror descent (LDOMD); the algorithm generalizes both the Component Hedge algorithm of Koolen et al. (2010), and a recent algorithm of Suehiro et al. (2012). Our study offers a unification and generalization of previous work, and emphasizes the role of the convex polytope arising from the vector representation of the decision space; while Boolean representations lead to 0-1 polytopes, more general vector representations lead to more general polytopes. We study several examples of both types of polytopes. Finally, we demonstrate the benefit of having a general framework for such problems via an application to an online transportation problem; the associated transportation polytopes generalize the Birkhoff polytope of doubly stochastic matrices, and the resulting algorithm generalizes the PermELearn algorithm of Helmbold and Warmuth (2009).

M25 Model-based Reinforcement Learning and the Eluder Dimension

Ian Osband iosband@stanford.edu
Stanford
Benjamin Van Roy bvr@stanford.edu
Stanford University

We consider the problem of learning to optimize an unknown Markov decision process (MDP). We show that, if the MDP can be parameterized within some known function class, we can obtain regret bounds that scale with the dimensionality, rather than cardinality, of the system. We characterize this dependence explicitly as $O\tilde{}(d_K d_E T\tilde{}\sqrt{})$ where T is time elapsed, d_K is the Kolmogorov dimension and d_E is the **eluder dimension**. These represent the first unified regret bounds for model-based reinforcement learning and provide state of the art guarantees in several important settings. Moreover, we present a simple and computationally efficient algorithm **posterior sampling for reinforcement learning** (PSRL) that satisfies these bounds.

M26 Algorithms for CVaR Optimization in MDPs

Yinlam Chow yldick.chow@gmail.com
Stanford
Mohammad Ghavamzadeh mohammad.ghavamzadeh@inria.fr
Adobe Research & INRIA

In many sequential decision-making problems we may want to manage risk by minimizing some measure of variability in costs in addition to minimizing a standard criterion. Conditional value-at-risk (CVaR) is a relatively new risk measure that addresses some of the shortcomings of the well-known variance-related risk measures, and because of its computational efficiencies has gained popularity in finance and operations research. In this paper, we consider the mean-CVaR optimization problem in MDPs. We first derive a formula for computing the gradient of this risk-sensitive objective function. We then devise policy gradient and actor-critic algorithms that each uses a specific method to estimate this gradient and updates the policy parameters in the descent direction. We establish the convergence of our algorithms to locally risk-sensitive optimal policies. Finally, we demonstrate the usefulness of our algorithms in an optimal stopping problem.

M27 Sparse Multi-Task Reinforcement Learning

Daniele Calandriello daniele.calandriello@mail.polimi.it
Alessandro Lazaric alessandro.lazaric@inria.fr
INRIA Lille-Nord Europe
Marcello Restelli restelli@elet.polimi.it
Politecnico di Milano

In multi-task reinforcement learning (MTRL), the objective is to simultaneously learn multiple tasks and exploit their similarity to improve the performance w.r.t.\ single-task learning. In this paper we investigate the case when all the tasks can be accurately represented in a linear approximation space using the same small subset of the original (large) set of features. This is equivalent to assuming that the weight vectors of the task value functions are \textit{jointly sparse}, i.e., the set of their non-zero components is small and it is shared across tasks. Building on existing results in multi-task regression, we develop two multi-task extensions of the fitted Q-iteration algorithm. While the first algorithm assumes that the tasks are jointly sparse in the given representation, the second one learns a transformation of the features in the attempt of finding a more sparse representation. For both algorithms we provide a sample complexity analysis and numerical simulations.

M28 Probabilistic Differential Dynamic Programming

Yunpeng Pan ypan37@gatech.edu
Evangelos Theodorou
evangelos.theodorou@ae.gatech.edu
Georgia Institute of Technology

We present a data-driven, probabilistic trajectory optimization framework for systems with unknown dynamics, called Probabilistic Differential Dynamic Programming (PDDP). PDDP takes into account uncertainty explicitly for dynamics models using Gaussian processes (GPs). Based on the second-order local approximation of the value function, PDDP performs Dynamic Programming around a nominal trajectory in Gaussian belief spaces. Different from typical gradient-

based policy search methods, PDDP does not require a policy parameterization and learns a locally optimal, time-varying control policy. We demonstrate the effectiveness and efficiency of the proposed algorithm using two nontrivial tasks. Compared with the classical DDP and a state-of-the-art GP-based policy search method, PDDP offers a superior combination of data-efficiency, learning speed, and applicability.

M29 Weighted importance sampling for off-policy learning with linear function approximation

A. Rupam Mahmood ashique@ualberta.ca
Rich Sutton rich@richsutton.com
University of Alberta
Hado van Hasselt h. p.van.hasselt@gmail.com
Centrum Wiskunde & Informatica (CWI)

Importance sampling is an essential component of off-policy model-free reinforcement learning algorithms. However, its most effective variant, **weighted** importance sampling, does not carry over easily to function approximation and, because of this, it is not utilized in existing off-policy learning algorithms. In this paper, we take two steps toward bridging this gap. First, we show that weighted importance sampling can be viewed as a special case of weighting the error of individual training samples, and that this weighting has theoretical and empirical benefits similar to those of weighted importance sampling. Second, we show that these benefits extend to a new weighted-importance-sampling version of off-policy LSTD(lambda). We show empirically that our new WIS-LSTD(lambda) algorithm can result in much more rapid and reliable convergence than conventional off-policy LSTD(lambda) (Yu 2010, Bertsekas & Yu 2009).

M30 A Representation Theory for Ranking Functions

Harsh Pareek harshp@cs.utexas.edu
Pradeep Ravikumar pradeepr@cs.utexas.edu
UT Austin

This paper presents a representation theory for permutation-valued functions, which in their general form can also be called listwise ranking functions. Pointwise ranking functions assign a score to each object independently, without taking into account the other objects under consideration; whereas listwise loss functions evaluate the set of scores assigned to all objects as a whole. In many supervised learning to rank tasks, it might be of interest to use listwise ranking functions instead; in particular, the Bayes Optimal ranking functions might themselves be listwise, especially if the loss function is listwise. A key caveat to using listwise ranking functions has been the lack of an appropriate representation theory for such functions. We show that a natural symmetricity assumption that we call exchangeability allows us to explicitly characterize the set of such exchangeable listwise ranking functions. Our analysis draws from the theories of tensor analysis, functional analysis and De Finetti theorems. We also present experiments using a novel reranking method motivated by our representation theory.

M31 Near-Optimal-Sample Estimators for Spherical Gaussian Mixtures

Ananda Theertha Suresh asuresh@ucsd.edu
Alon Orlitsky alon@ucsd.edu
Ashkan Jafarpour ashkan@ucsd.edu
University of California, San Diego
Jayadev Acharya jayadev@csail.mit.edu
Massachusetts Institute of Technology

Many important distributions are high dimensional, and often they can be modeled as Gaussian mixtures. We derive the first sample-efficient polynomial-time estimator for high-dimensional spherical Gaussian mixtures. Based on intuitive spectral reasoning, it approximates mixtures of k spherical Gaussians in d-dimensions to withinℓ_1 distance ϵ using $\mathcal{O}(dk^9(\log^2 d)/\epsilon^4)$ samples and $\mathcal{O}_{k,\epsilon}(d^3\log^5 d)$computation time. Conversely, we show that any estimator requires $\Omega\big(dk/\epsilon^2\big)$ samples, hence the algorithm's sample complexity is nearly optimal in the dimension. The implied time-complexity factor \ mathcal{O}_{k,\epsilon}is exponential ink, but much smaller than previously known. We also construct a simple estimatorforone−dimensional Gaussian mixtures that uses\tilde\mathcal{O}(k/\epsilon^2)samplesand\tilde\ mathcal{O}((k/\epsilon)^{3k+1})$ computation time.

M32 Tighten after Relax: Minimax-Optimal Sparse PCA in Polynomial Time

Zhaoran Wang zhaoran@princeton.edu
Huanran Lu huanranl@princeton.edu
Han Liu hanliu@princeton.edu
Princeton University

We provide statistical and computational analysis of sparse Principal Component Analysis (PCA) in high dimensions. The sparse PCA problem is highly nonconvex in nature. Consequently, though its global solution attains the optimal statistical rate of convergence, such solution is computationally intractable to obtain. Meanwhile, although its convex relaxations are tractable to compute, they yield estimators with suboptimal statistical rates of convergence. On the other hand, existing nonconvex optimization procedures, such as greedy methods, lack statistical guarantees. In this paper, we propose a two-stage sparse PCA procedure that attains the optimal principal subspace estimator in polynomial time. The main stage employs a novel algorithm named sparse orthogonal iteration pursuit, which iteratively solves the underlying nonconvex problem. However, our analysis shows that this algorithm only has desired computational and statistical guarantees within a restricted region, namely the basin of attraction. To obtain the desired initial estimator that falls into this region, we solve a convex formulation of sparse PCA with early stopping. Under an integrated analytic framework, we simultaneously characterize the computational and statistical performance of this two-stage procedure. Computationally, our procedure converges at the rate of $1/t\sqrt{}$ within the initialization stage, and at a geometric rate within the main stage. Statistically, the final principal subspace estimator achieves the minimax-optimal statistical rate of convergence with respect to the sparsity level $s*$, dimension d and sample size n. Our procedure motivates a general paradigm of tackling nonconvex statistical learning problems with provable statistical guarantees.

M33 Consistency of weighted majority votes

Daniel Berend berend@cs.bgu.ac.il
Aryeh Kontorovitch karyeh@cs.bgu.ac.il
Ben Gurion University

We revisit from a statistical learning perspective the classical decision-theoretic problem of weighted expert voting. In particular, we examine the consistency (both asymptotic and finitary) of the optimal Nitzan-Paroush weighted majority and related rules. In the case of known expert competence levels, we give sharp error estimates for the optimal rule. When the competence levels are unknown, they must be empirically estimated. We provide frequentist and Bayesian analyses for this situation. Some of our proof techniques are non-standard and may be of independent interest. The bounds we derive are nearly optimal, and several challenging open problems are posed. Experimental results are provided to illustrate the theory.

M34 Quantized Estimation of Gaussian Sequence Models in Euclidean Balls

Yuancheng Zhu yuancheng@uchicago.edu
John Lafferty lafferty@gmail.com
University of Chicago

A central result in statistical theory is Pinsker's theorem, which characterizes the minimax rate in the normal means model of nonparametric estimation. In this paper, we present an extension to Pinsker's theorem where estimation is carried out under storage or communication constraints. In particular, we place limits on the number of bits used to encode an estimator, and analyze the excess risk in terms of this constraint, the signal size, and the noise level. We give sharp upper and lower bounds for the case of a Euclidean ball, which establishes the Pareto-optimal minimax tradeoff between storage and risk in this setting.

M35 On the Convergence Rate of Decomposable Submodular Function Minimization

Robert Nishihara robertnishihara@gmail.com
Stefanie Jegelka stefje@eecs.berkeley.edu
Michael Jordan jordan@cs.berkeley.edu
UC Berkeley

Submodular functions describe a variety of discrete problems in machine learning, signal processing, and computer vision. However, minimizing submodular functions poses a number of algorithmic challenges. Recent work introduced an easy-to-use, parallelizable algorithm for minimizing submodular functions that decompose as the sum of "simple" submodular functions. Empirically, this algorithm performs extremely well, but no theoretical analysis was given. In this paper, we show that the algorithm converges linearly, and we provide upper and lower bounds on the rate of convergence. Our proof relies on the geometry of submodular polyhedra and draws on results from spectral graph theory.

M36 Parallel Direction Method of Multipliers

Huahua Wang huwang@cs.umn.edu
Arindam Banerjee banerjee@cs.umn.edu
Zhi-Quan Luo luozq@ece.umn.edu
University of Minnesota, Twin Cites

We consider the problem of minimizing block-separable convex functions subject to linear constraints. While the Alternating Direction Method of Multipliers (ADMM) for two-block linear constraints has been intensively studied both theoretically and empirically, in spite of some preliminary work, effective generalizations of ADMM to multiple blocks is still unclear. In this paper, we propose a parallel randomized block coordinate method named Parallel Direction Method of Multipliers (PDMM) to solve the optimization problems with multi-block linear constraints. PDMM randomly updates some blocks in parallel, behaving like parallel randomized block coordinate descent. We establish the global convergence and the iteration complexity for PDMM with constant step size. We also show that PDMM can do randomized block coordinate descent on overlapping blocks. Experimental results show that PDMM performs better than state-of-the-arts methods in two applications, robust principal component analysis and overlapping group lasso.

M37 Constant Nullspace Strong Convexity and Fast Convergence of Proximal Methods under High-Dimensional Settings

En-Hsu Yen a061105@gmail.com
Cho-Jui Hsieh cjhsieh@cs.utexas.edu
Pradeep Ravikumar pradeepr@cs.utexas.edu
Inderjit Dhillon inderjit@cs.utexas.edu
University of Texas

State of the art statistical estimators for high-dimensional problems take the form of regularized, and hence non-smooth, convex programs. A key facet of thesestatistical estimation problems is that these are typically not strongly convex under a high-dimensional sampling regime when the Hessian matrix becomes rank-deficient. Under vanilla convexity however, proximal optimization methods attain only a sublinear rate. In this paper, we investigate a novel variant of strong convexity, which we call Constant Nullspace Strong Convexity (CNSC), where we require that the objective function be strongly convex only over a constant subspace. As we show, the CNSC condition is naturally satisfied by high-dimensional statistical estimators. We then analyze the behavior of proximal methods under this CNSC condition: we show global linear convergence of Proximal Gradient and local quadratic convergence of Proximal Newton Method, when the regularization function comprising the statistical estimator is decomposable. We corroborate our theory via numerical experiments, and show a qualitative difference in the convergence rates of the proximal algorithms when the loss function does satisfy the CNSC condition.

M38 SAGA: A Fast Incremental Gradient Method With Support for Non-Strongly Convex Composite Objectives

Aaron Defazio aaron.defazio@anu.edu.au
NICTA
Francis Bach francis.bach@mines.org
Simon Lacoste-Julien simon.lacoste-julien@inria.fr
INRIA / ENS

In this work we introduce a new fast incremental gradient method SAGA, in the spirit of SAG, SDCA, MISO and SVRG. SAGA improves on the theory behind SAG and SVRG, with better theoretical convergence rates, and support for composite objectives where a proximal operator is used on the regulariser. Unlike SDCA, SAGA supports non-strongly convex problems directly, and is adaptive to any inherent strong convexity of the problem. We give experimental results showing the effectiveness of our method.

M39 Time--Data Tradeoffs by Aggressive Smoothing

John Bruer jbruer@cms.caltech.edu
Joel Tropp jtropp@cms.caltech.edu
Caltech
Volkan Cevher volkan.cevher@epfl.ch
EPFL
Stephen Becker stephen.beckr@gmail.com
University of Colorado

This paper proposes a tradeoff between sample complexity and computation time that applies to statistical estimators based on convex optimization. As the amount of data increases, we can smooth optimization problems more and more aggressively to achieve accurate estimates more quickly. This work provides theoretical and experimental evidence of this tradeoff for a class of regularized linear inverse problems.

M40 Local Linear Convergence of Forward--Backward under Partial Smoothness

Jingwei Liang Jingwei.Liang@ensicaen.fr
Jalal Fadili Jalal.Fadili@greyc.ensicaen.fr
CNRS-ENSICAEN-Univ. Caen
Gabriel Peyré gabriel.peyre@ceremade.dauphine.fr
Université Paris Dauphine

In this paper, we consider the Forward--Backward proximal splitting algorithm to minimize the sum of two proper closed convex functions, one of which having a Lipschitz continuous gradient and the other being partly smooth relatively to an active manifold $\mathcal{M}$. We propose a generic framework in which we show that the Forward--Backward (i) correctly identifies the active manifold $\mathcal{M}$ in a finite number of iterations, and then (ii) enters a local linear convergence regime that we characterize precisely. This gives a grounded and unified explanation to the typical behaviour that has been observed numerically for many problems encompassed in our framework, including the Lasso, the group Lasso, the fused Lasso and the nuclear norm regularization to name a few. These results may have numerous applications including in signal/image processing processing, sparse recovery and machine learning.

M41 Simple MAP Inference via Low-Rank Relaxations

Roy Frostig rf@cs.stanford.edu
Sida Wang sidaw@cs.stanford.edu
Percy Liang pliang@cs.stanford.edu
Christopher Manning manning@stanford.edu
Stanford University

We focus on the problem of maximum a posteriori (MAP) inference in Markov random fields with binary variables and pairwise interactions. For this common subclass of inference tasks, we consider low-rank relaxations that interpolate between the discrete problem and its full-rank semidefinite relaxation, followed by randomized rounding. We develop new theoretical bounds studying the effect of rank, showing that as the rank grows, the relaxed objective increases but saturates, and that the fraction in objective value retained by the rounded discrete solution decreases. In practice, we show two algorithms for optimizing the low-rank objectives which are simple to implement, enjoy ties to the underlying theory, and outperform existing approaches on benchmark MAP inference tasks.

M42 Inferring synaptic conductances from spike trains with a biophysically inspired point process model

Kenneth Latimer latimerk@utexas.edu
Jonathan Pillow pillow@mail.utexas.edu
UT Austin
E. J. Chichilnisky ej@stanford.edu
Stanford University
Fred Rieke rieke@uw.edu
University of Washington, Seattle

A popular approach to neural characterization describes neural responses in terms of a cascade of linear and nonlinear stages: a linear filter to describe stimulus integration, followed by a nonlinear function to convert the filter output to spike rate. However, real neurons respond to stimuli in a manner that depends on the nonlinear integration of excitatory and inhibitory synaptic inputs. Here we introduce a biophysically inspired point process model that explicitly incorporates stimulus-induced changes in synaptic conductance in a dynamical model of neuronal membrane potential. Our work makes two important contributions. First, on a theoretical level, it offers a novel interpretation of the popular generalized linear model (GLM) for neural spike trains. We show that the classic GLM is a special case of our conductance-based model in which the stimulus linearly modulates excitatory and inhibitory conductances in an equal and opposite "push-pull" fashion. Our model can therefore be viewed as a direct extension of the GLM in which we relax these constraints; the resulting model can exhibit shunting as well as hyperpolarizing inhibition, and time-varying changes in both gain and membrane time constant. Second, on a practical level, we show that our model provides a tractable model of spike responses in early sensory neurons that is both more accurate and more interpretable than the GLM. Most importantly, we show that we can accurately infer intracellular synaptic conductances from extracellularly recorded spike trains. We validate these estimates using direct intracellular measurements of excitatory and inhibitory conductances in parasol retinal ganglion cells. We show that the model fit to extracellular spike trains can predict excitatory and

inhibitory conductances elicited by novel stimuli with nearly the same accuracy as a model trained directly with intracellular conductances.

M43 Low-dimensional models of neural population activity in sensory cortical circuits

Evan Archer earcher@utexas.edu
Jakob Macke Jakob.Macke@gmail.com
MPI for Biological Cybernetics
Urs Koster urs@berkeley.edu
University of California at Berkely
Jonathan Pillow pillow@mail.utexas.edu
UT Austin

Neural responses in visual cortex are influenced by visual stimuli and by ongoing spiking activity in local circuits. An important challenge in computational neuroscience is to develop models that can account for both of these features in large multi-neuron recordings and to reveal how stimulus representations interact with and depend on cortical dynamics. Here we introduce a statistical model of neural population activity that integrates a nonlinear receptive field model with a latent dynamical model of ongoing cortical activity. This model captures the temporal dynamics, effective network connectivity in large population recordings, and correlations due to shared stimulus drive as well as common noise. Moreover, because the nonlinear stimulus inputs are mixed by the ongoing dynamics, the model can account for a relatively large number of idiosyncratic receptive field shapes with a small number of nonlinear inputs to a low-dimensional latent dynamical model. We introduce a fast estimation method using online expectation maximization with Laplace approximations. Inference scales linearly in both population size and recording duration. We apply this model to multi-channel recordings from primary visual cortex and show that it accounts for a large number of individual neural receptive fields using a small number of nonlinear inputs and a low-dimensional dynamical model.

M44 Inferring sparse representations of continuous signals with continuous orthogonal matching pursuit

Karin Knudson kknudson@math.utexas.edu
Jonathan Pillow pillow@mail.utexas.edu
UT Austin
Jacob Yates jakelyates@gmail.com
Alexander Huk huk_@utexas.edu
University of Texas at Austin

Many signals, such as spike trains recorded in multi-channel electrophysiological recordings, may be represented as the sparse sum of translated and scaled copies of waveforms whose timing and amplitudes are of interest. From the aggregate signal, one may seek to estimate the identities, amplitudes, and translations of the waveforms that compose the signal. Here we present a fast method for recovering these identities, amplitudes, and translations. The method involves greedily selecting component waveforms and then refining estimates of their amplitudes and translations, moving iteratively between these steps in a process analogous to the well-known Orthogonal Matching Pursuit (OMP) algorithm. Our approach for modeling translations borrows from Continuous Basis Pursuit (CBP), which we extend in several ways: by selecting a subspace that optimally captures translated copies of the waveforms, replacing the convex optimization problem with a greedy approach, and moving to the Fourier domain to more precisely estimate time shifts. We test the resulting method, which we call Continuous Orthogonal Matching Pursuit (COMP), on simulated and neural data, where it shows gains over CBP in both speed and accuracy.

M45 Fast Sampling-Based Inference in Balanced Neuronal Networks

Guillaume Hennequin gjeh2@cam.ac.uk
Mate Lengyel m.lengyel@eng.cam.ac.uk
University of Cambridge
Laurence Aitchison laurence@gatsby.ucl.ac.uk
University College London

Multiple lines of evidence support the notion that the brain performs probabilistic inference in multiple cognitive domains, including perception and decision making. There is also evidence that probabilistic inference may be implemented in the brain through the (quasi-)stochastic activity of neural circuits, producing samples from the appropriate posterior distributions, effectively implementing a Markov chain Monte Carlo algorithm. However, time becomes a fundamental bottleneck in such sampling-based probabilistic representations: the quality of inferences depends on how fast the neural circuit generates new, uncorrelated samples from its stationary distribution (the posterior). We explore this bottleneck in a simple, linear-Gaussian latent variable model, in which posterior sampling can be achieved by stochastic neural networks with linear dynamics. The well-known Langevin sampling (LS) recipe, so far the only sampling algorithm for continuous variables of which a neural implementation has been suggested, naturally fits into this dynamical framework. However, we first show analytically and through simulations that the symmetry of the synaptic weight matrix implied by LS yields critically slow mixing when the posterior is high-dimensional. Next, using methods from control theory, we construct and inspect networks that are optimally fast, and hence orders of magnitude faster than LS, while being far more biologically plausible. In these networks, strong -- but transient -- selective amplification of external noise generates the spatially correlated activity fluctuations prescribed by the posterior. Intriguingly, although a detailed balance of excitation and inhibition is dynamically maintained, detailed balance of Markov chain steps in the resulting sampler is violated, consistent with recent findings on how statistical irreversibility can overcome the speed limitation of random walks in other domains.

M46 Information-based learning by agents in unbounded state spaces

Shariq Mobin shariq.mobin@gmail.com
James Arnemann arnemann@berkeley.edu
Fritz Sommer fsommer@berkeley.edu
UC Berkeley

The idea that animals might use information-driven planning to explore an unknown environment and build an internal model of it has been proposed for quite some time. Recent work has demonstrated that agents using this principle can efficiently learn models of probabilistic environments with discrete, bounded state spaces. However, animals and robots are commonly confronted with unbounded environments. To address this more challenging situation, we study information-based learning strategies of agents in unbounded state spaces using non-parametric Bayesian models. Specifically, we demonstrate that the Chinese Restaurant Process (CRP) model is able to solve this problem and that an Empirical Bayes version is able to efficiently explore bounded and unbounded worlds by relying on little prior information.

M47 On the Computational Efficiency of Training Neural Networks

Roi Livni roi.livni@mail.huji.ac.il
Shai Shalev-Shwartz shai.shwartz@gmail.com
The Hebrew University
Ohad Shamir ohad.shamir@weizmann.ac.il
Weizmann Institute of Science

It is well-known that neural networks are computationally hard to train. On the other hand, in practice, modern day neural networks are trained efficiently using SGD and a variety of tricks that include different activation functions (e.g. ReLU), over-specification (i.e., train networks which are larger than needed), and regularization. In this paper we revisit the computational complexity of training neural networks from a modern perspective. We provide both positive and negative results, some of them yield new provably efficient and practical algorithms for training neural networks.

M48 Attentional Neural Network: Feature Selection Using Cognitive Feedback

Qian Wang qianwang.thu@gmail.com
Sen Song sen.song@gmail.com
Tsinghua University
Jiaxing Zhang jiaxz@microsoft.com
Zheng Zhang zzhang@microsoft.com
Microsoft Reasearch

Attentional Neural Network is a new framework that integrates top-down cognitive bias and bottom-up feature extraction in one coherent architecture. The top-down influence is especially effective when dealing with high noise or difficult segmentation problems. Our system is modular and extensible. It is also easy to train and cheap to run, and yet can accommodate complex behaviors. We obtain classification accuracy better than or competitive with state of art results on the MNIST variation dataset, and successfully disentangle overlaid digits with high success rates. We view such a general purpose framework as an essential foundation for a larger system emulating the cognitive abilities of the whole brain.

M49 Expectation Backpropagation: Parameter-Free Training of Multilayer Neural Networks with Continuous or Discrete Weights

Daniel Soudry daniel.soudry@gmail.com
Columbia University
Itay Hubara itayhubara@gmail.com
Ron Meir rmeir@ee.technion.ac.il
Technion

Multilayer Neural Networks (MNNs) are commonly trained using gradient descent-based methods, such as BackPropagation (BP). Inference in probabilistic graphical models is often done using variational Bayes methods, such as Expectation Propagation (EP). We show how an EP based approach can also be used to train deterministic MNNs. Specifically, we approximate the posterior of the weights given the data using a "mean-field" factorized distribution, in an online setting. Using online EP and the central limit theorem we find an analytical approximation to the Bayes update of this posterior, as well as the resulting Bayes estimates of the weights and outputs. Despite a different origin, the resulting algorithm, Expectation BackPropagation (EBP), is very similar to BP in form and efficiency. However, it has several additional advantages: (1) Training is parameter-free, given initial conditions (prior) and the MNN architecture. This is useful for large-scale problems, where parameter tuning is a major challenge. (2) The weights can be restricted to have discrete values. This is especially useful for implementing trained MNNs in precision limited hardware chips, thus improving their speed and energy efficiency by several orders of magnitude. We test the EBP algorithm numerically in eight binary text classification tasks. In all tasks, EBP outperforms: (1) standard BP with the optimal constant learning rate (2) previously reported state of the art. Interestingly, EBP-trained MNNs with binary weights usually perform better than MNNs with continuous (real) weights - if we average the MNN output using the inferred posterior.

M50 An Autoencoder Approach to Learning Bilingual Word Representations

Sarath Chandar A P apsarathchandar@gmail.com
IBM Research India
Stanislas Lauly Stanislas.Lauly@USherbrooke.ca
Hugo Larochelle hugo.larochelle@usherbrooke.ca
Université de Sherbrooke
Mitesh Khapra mikhapra@in.ibm.com
Amrita Saha amrsaha4@in.ibm.com
IBM India Research Lab
Balaraman Ravindran ravi@cse.iitm.ac.in
Indian Institute of Technology Madras
Vikas Raykar vikasraykar@gmail.com
IBM Research

Cross-language learning allows us to use training data from one language to build models for a different language. Many approaches to bilingual learning require that we have word-level alignment of sentences from parallel corpora. In this work we explore the use of autoencoder-based methods for cross-language learning of vectorial word representations that are aligned between two languages, while not relying on word-level alignments. We show that by simply learning to reconstruct the bag-of-words representations of aligned sentences, within

and between languages, we can in fact learn high-quality representations and do without word alignments. We empirically investigate the success of our approach on the problem of cross-language text classification, where a classifier trained on a given language (e.g., English) must learn to generalize to a different language (e.g., German). In experiments on 3 language pairs, we show that our approach achieves state-of-the-art performance, outperforming a method exploiting word alignments and a strong machine translation baseline.

M51 Pre-training of Recurrent Neural Networks via Linear Autoencoders

Luca Pasa lpasa@math.unipd.it
Alessandro Sperduti sperduti@math.unipd.it
Università degli Studi di Padova

We propose a pre-training technique for recurrent neural networks based on linear autoencoder networks for sequences, i.e. linear dynamical systems modelling the target sequences. We start by giving a closed form solution for the definition of the optimal weights of a linear autoencoder given a training set of sequences. This solution, however, is computationally very demanding, so we suggest a procedure to get an approximate solution for a given number of hidden units. The weights obtained for the linear autoencoder are then used as initial weights for the input-to-hidden connections of a recurrent neural network, which is then trained on the desired task. Using four well known datasets of sequences of polyphonic music, we show that the proposed pre-training approach is highly effective, since it allows to largely improve the state of the art results on all the considered datasets.

M52 Using Convolutional Neural Networks to Recognize Rhythm Stimuli from Electroencephalography Recordings

Sebastian Stober sstober@uwo.ca
Western University
Daniel Cameron dcamer25@uwo.ca
Jessica Grahn jgrahn@uwo.ca

Electroencephalography (EEG) recordings of rhythm perception might contain enough information to distinguish different rhythm types/genres or even identify the rhythms themselves. We apply convolutional neural networks (CNNs) to analyze and classify EEG data recorded within a rhythm perception study in Kigali, Rwanda which comprises 12 East African and 12 Western rhythmic stimuli – each presented in a loop for 32 seconds to 13 participants. We investigate the impact of the data representation and the pre-processing steps for this classification tasks and compare different network structures. Using CNNs, we are able to recognize individual rhythms from the EEG with a mean classification accuracy of 24.4% (chance level 4.17%) over all subjects by looking at less than three seconds from a single channel. Aggregating predictions for multiple channels, a mean accuracy of up to 50% can be achieved for individual subjects.

M53 Neurons as Monte Carlo Samplers: Bayesian Inference and Learning in Spiking Networks

Yanping Huang huangyp@cs.washington.edu
Rajesh Rao rao@cs.washington.edu
University of Washington

We propose a two-layer spiking network capable of performing approximate inference and learning for a hidden Markov model. The lower layer sensory neurons detect noisy measurements of hidden world states. The higher layer neurons with recurrent connections infer a posterior distribution over world states from spike trains generated by sensory neurons. We show how such a neuronal network with synaptic plasticity can implement a form of Bayesian inference similar to Monte Carlo methods such as particle filtering. Each spike in the population of inference neurons represents a sample of a particular hidden world state. The spiking activity across the neural population approximates the posterior distribution of hidden state. The model provides a functional explanation for the Poisson-like noise commonly observed in cortical responses. Uncertainties in spike times provide the necessary variability for sampling during inference. Unlike previous models, the hidden world state is not observed by the sensory neurons, and the temporal dynamics of the hidden state is unknown. We demonstrate how this network can sequentially learn the hidden Markov model using a spike-timing dependent Hebbian learning rule and achieve power-law convergence rates.

M54 A framework for studying synaptic plasticity with neural spike train data

Scott Linderman slinderman@seas.harvard.edu
Christopher Stock cstock@college.harvard.edu
Ryan Adams rpa@seas.harvard.edu
Harvard University

Learning and memory in the brain are implemented by complex, time-varying changes in neural circuitry. The computational rules according to which synaptic weights change over time are the subject of much research, and are not precisely understood. Until recently, limitations in experimental methods have made it challenging to test hypotheses about synaptic plasticity on a large scale. However, as such data become available and these barriers are lifted, it becomes necessary to develop analysis techniques to validate plasticity models. Here, we present a highly extensible framework for modeling arbitrary synaptic plasticity rules on spike train data in populations of interconnected neurons. We treat synaptic weights as a (potentially nonlinear) dynamical system embedded in a fully-Bayesian generalized linear model (GLM). In addition, we provide an algorithm for inferring synaptic weight trajectories alongside the parameters of the GLM and of the learning rules. Using this method, we perform model comparison of two proposed variants of the well-known spike-timing-dependent plasticity (STDP) rule, where nonlinear effects play a substantial role. On synthetic data generated from the biophysical simulator NEURON, we show that we can recover the weight trajectories, the pattern of connectivity, and the underlying learning rules.

M55 Global Belief Recursive Neural Networks

Romain Paulus rpaulus@isep.fr
ISEP
Richard Socher richard@socher.org
Chris Manning manning@cs.stanford.edu
Stanford University

Recursive Neural Networks have recently obtained state of the art performance on several natural language processing tasks. However, because of their feedforward architecture they cannot correctly predict phrase or word labels that are determined by context. This is a problem in tasks such as aspect-specific sentiment classification which tries to, for instance, predict that the word Android is positive in the sentence Android beats iOS. We introduce global belief recursive neural networks (GB-RNNs) which are based on the idea of extending purely feedforward neural networks to include one feedbackward step during inference. This allows phrase level predictions and representations to give feedback to words. We show the effectiveness of this model on the task of contextual sentiment analysis. We also show that dropout can improve RNN training and that a combination of unsupervised and supervised word vector representations performs better than either alone. The feedbackward step improves F1 performance by 3% over the standard RNN on this task, obtains state-of-the-art performance on the SemEval 2013 challenge and can accurately predict the sentiment of specific entities.

M56 Deep Networks with Internal Selective Attention through Feedback Connections

Marijn Stollenga marijn@idsia.ch
IDSIA/SUPSI, USI
Jonathan Masci jonathan@idsia.ch
Faustino Gomez tino@idsia.ch
Jürgen Schmidhuber juergen@idsia.ch
IDSIA

Traditional convolutional neural networks (CNN) are stationary and feedforward. They neither change their parameters during evaluation nor use feedback from higher to lower layers. Real brains, however, do. So does our Deep Attention Selective Network (dasNet) architecture. DasNet's feedback structure can dynamically alter its convolutional filter sensitivities during classification. It harnesses the power of sequential processing to improve classification performance, by allowing the network to iteratively focus its internal attention on some of its convolutional filters. Feedback is trained through direct policy search in a huge million-dimensional parameter space, through scalable natural evolution strategies (SNES). On the CIFAR-10 and CIFAR-100 datasets, dasNet outperforms the previous state-of-the-art model on unaugmented datasets.

M57 Iterative Neural Autoregressive Distribution Estimator NADE-k

Tapani Raiko tapani.raiko@aalto.fi
Aalto University
Yao Li li.yao@umontreal.ca
Yoshua Bengio bengioy@iro.umontreal.ca
Kyunghyun Cho kyunghyun.cho@umontreal.ca
University of Montreal

Training of the neural autoregressive density estimator (NADE) can be viewed as doing one step of probabilistic inference on missing values in data. We propose a new model that extends this inference scheme to multiple steps, arguing that it is easier to learn to improve a reconstruction in k steps rather than to learn to reconstruct in a single inference step. The proposed model is an unsupervised building block for deep learning that combines the desirable properties of NADE and multi-predictive training: (1) Its test likelihood can be computed analytically, (2) it is easy to generate independent samples from it, and (3) it uses an inference engine that is a superset of variational inference for Boltzmann machines. The proposed NADE-k is competitive with the state-of-the-art in density estimation on the two datasets tested.

M58 General Stochastic Networks for Classification

Matthias Zöhrer matthias.zoehrer@tugraz.at
Franz Pernkopf pernkopf@tugraz.at
Graz University of Technology

We extend generative stochastic networks to supervised learning of representations. In particular, we introduce a hybrid training objective considering a generative and discriminative cost function governed by a trade-off parameter lambda. We use a new variant of network training involving noise injection, i.e. walkback training, to jointly optimize multiple network layers. Neither additional regularization constraints, such as l1, l2 norms or dropout variants, nor pooling- or convolutional layers were added. Nevertheless, we are able to obtain state-of-the-art performance on the MNIST dataset, without using permutation invariant digits and outperform baseline models on sub-variants of the MNIST and rectangles dataset significantly.

M59 Improved Multimodal Deep Learning with Variation of Information

Kihyuk Sohn kihyuk.sohn@gmail.com
Wenling Shang shangw@umich.edu
Honglak Lee honglak@eecs.umich.edu
University of Michigan

Deep learning has been successfully applied to multimodal representation learning problems, with a common strategy to learning joint representations that are shared across multiple modalities on top of layers of modality-specific networks. Nonetheless, there still remains a question how to learn a good association between data modalities; in particular, a good generative model of multimodal data should be able to reason about missing data modality given the rest of data modalities. In this paper, we propose a novel multimodal representation learning framework that explicitly aims this goal. Rather than learning with maximum likelihood, we train the model to minimize the variation of information. We provide a theoretical insight why the proposed learning objective is sufficient to estimate the data-generating joint distribution of multimodal data. We apply our method to restricted Boltzmann machines and introduce learning methods based on contrastive divergence and multi-prediction training. In addition, we extend to deep networks with recurrent encoding structure to finetune the whole network. In experiments, we demonstrate the state-of-the-art visual recognition performance on MIR-Flickr database and PASCAL VOC 2007 database with and without text features.

M60 Restricted Boltzmann machines modeling human choice

Takayuki Osogami osogami@jp.ibm.com
Makoto Otsuka motsuka@jp.ibm.com
IBM Research - Tokyo

We extend the multinomial logit model to represent some of the empirical phenomena that are frequently observed in the choices made by humans. These phenomena include the similarity effect, the attraction effect, and the compromise effect. We formally quantify the strength of these phenomena that can be represented by our choice model, which illuminates the flexibility of our choice model. We then show that our choice model can be represented as a restricted Boltzmann machine and that its parameters can be learned effectively from data. Our numerical experiments with real data of human choices suggest that we can train our choice model in such a way that it represents the typical phenomena of choice.

M61 Deep Fragment Embeddings for Bidirectional Image Sentence Mapping

Andrej Karpathy andrej.karpathy@gmail.com
Armand Joulin armand.joulin@gmail.com
Fei Fei Li feifeili@cs.stanford.edu
Stanford University

We introduce a model for bidirectional retrieval of images and sentences through a deep, multi-modal embedding of visual and natural language data. Unlike previous models that directly map images or sentences into a common embedding space, our model works on a finer level and embeds fragments of images (objects) and fragments of sentences (typed dependency tree relations) into a common space. We then introduce a structured max-margin objective that allows our model to explicitly associate these fragments across modalities. Extensive experimental evaluation shows that reasoning on both the global level of images and sentences and the finer level of their respective fragments improves performance on image-sentence retrieval tasks. Additionally, our model provides interpretable predictions for the image-sentence retrieval task since the inferred inter-modal alignment of fragments is explicit.

M62 Recursive Context Propagation Network for Semantic Scene Labeling

Abhishek Sharma bhokaal@cs.umd.edu
University of Maryland
Oncel Tuzel onceltuzel@gmail.com
Ming-Yu Liu mliu@merl.com
MERL

We propose a deep feed-forward neural network architecture for pixel-wise semantic scene labeling. It uses a novel recursive neural network architecture for context propagation, referred to as rCPN. It first maps the local visual features into a semantic space followed by a bottom-up aggregation of local information into a global representation of the entire image. Then a top-down propagation of the aggregated information takes place that enhances the contextual information of each local feature. Therefore, the information from every location in the image is propagated to every other location. Experimental results on Stanford background and SIFT Flow datasets show that the proposed method outperforms previous approaches. It is also orders of magnitude faster than previous methods and takes only 0.07 seconds on a GPU for pixel-wise labeling of a 256x256 image starting from raw RGB pixel values, given the super-pixel mask that takes an additional 0.3 seconds using an off-the-shelf implementation.

M63 Multiscale Fields of Patterns

Pedro Felzenszwalb pff@brown.edu
John Oberlin oberlin@cs.brown.edu
Brown University

We describe a framework for defining high-order image models that can be used in a variety of applications. The approach involves modeling local patterns in a multiscale representation of an image. Local properties of a coarsened image reflect non-local properties of the original image. In the case of binary images local properties are defined by the binary patterns observed over small neighborhoods around each pixel. With the multiscale representation we capture the frequency of patterns observed at different scales of resolution. This framework leads to expressive priors that depend on a relatively small number of parameters. For inference and learning we use an MCMC method for block sampling with very large blocks. We evaluate the approach with two example applications. One involves contour detection. The other involves binary segmentation.

M64 Weakly-supervised Discovery of Visual Pattern Configurations

Hyun Oh Song song@eecs.berkeley.edu
Stefanie Jegelka stefje@eecs.berkeley.edu
Trevor Darrell trevor@eecs.berkeley.edu
UC Berkeley
Yong Jae Lee yjlee@cs.ucdavis.edu
UC Davis

The prominence of weakly labeled data gives rise to a growing demand for object detection methods that can cope with minimal supervision. We propose an approach that automatically identifies discriminative configurations of visual patterns that are characteristic of a given object class. We formulate the problem as a constrained submodular optimization problem and demonstrate the benefits of the discovered configurations in remedying mislocalizations and finding informative positive and negative training examples. Together, these lead to state-of-the-art weakly-supervised detection results on the challenging PASCAL VOC dataset.

M65 Encoding High Dimensional Local Features by Sparse Coding Based Fisher Vectors

Lingqiao Liu liulq83@gmail.com
Univeristy of Adelaide
Chunhua Shen chhshen@gmail.com
NICTA (National ICT Australia)
Lei Wang lei.hs.wang@gmail.com
Chao Wang chaow@uow.edu.au
University of Wollongong
Anton van den Hengel
Anton.vandenHengel@adelaide.edu.au
University of Adelaide

Deriving from the gradient vector of a generative model of local features, Fisher vector coding (FVC) has been identified as an effective coding method for image classification. Most, if not all, FVC implementations employ the Gaussian mixture model (GMM) to characterize the generation process of local features. This choice has shown to be sufficient for traditional low dimensional local features, e.g., SIFT; and typically, good performance can be achieved with only a few hundred Gaussian distributions. However, the same number of Gaussians is insufficient to model the feature space spanned by higher dimensional local features, which have become popular recently. In order to improve the modeling capacity for high dimensional features, it turns out to be inefficient and computationally impractical to simply increase the number of Gaussians. In this paper, we propose a model in which each local feature is drawn from a Gaussian distribution whose mean vector is sampled from a subspace. With certain approximation, this model can be converted to a sparse coding procedure and the learning/inference problems can be readily solved by standard sparse coding methods. By calculating the gradient vector of the proposed model, we derive a new fisher vector encoding strategy, termed Sparse Coding based Fisher Vector Coding (SCFVC). Moreover, we adopt the recently developed Deep Convolutional Neural Network (CNN) descriptor as a high dimensional local feature and implement image classification with the proposed SCFVC. Our experimental evaluations demonstrate that our method not only significantly outperforms the traditional GMM based Fisher vector encoding but also achieves the state-of-the-art performance in generic object recognition, indoor scene, and fine-grained image classification problems.

M66 Self-Adaptable Templates for Feature Coding

Xavier Boix boxavier@vision.ee.ethz.ch
Gemma Roig gemmar@vision.ee.ethz.ch
Salomon Diether sdiether@ethz.ch
Luc Gool vangool@vision.ee.ethz.ch
ETH Zurich

Hierarchical feed-forward networks have been successfully applied in object recognition. At each level of the hierarchy, features are extracted and encoded, followed by a pooling step. Within this processing pipeline, the common trend is to learn the feature coding templates, often referred as codebook entries, filters, or over-complete basis. Recently, an approach that apparently does not use templates has been shown to obtain very promising results. This is the second-order pooling (O2P). In this paper, we analyze O2P as a coding-pooling scheme. We find that at testing phase, O2P automatically adapts the feature coding templates to the input features, rather than using templates learned during the training phase. From this finding, we are able to bring common concepts of coding-pooling schemes to O2P, such as feature quantization. This allows for significant accuracy improvements of O2P in standard benchmarks of image classification, namely Caltech101 and VOC07.

M67 Learning From Weakly Supervised Data by The Expectation Loss SVM (e-SVM) algorithm

Jun Zhu jzh@ucla.edu
Junhua Mao mjhustc@ucla.edu
Alan Yuille yuille@stat.ucla.edu
UCLA

In many situations we have some measurement of confidence on ``positiveness" for a binary label. The ``positiveness" is a continuous value whose range is a bounded interval. It quantifies the affiliation of each training data to the positive class. We propose a novel learning algorithm called**expectation loss SVM** (e-SVM) that is devoted to the problems where only the ``positiveness" instead of a binary label of each training sample is available. Our e-SVM algorithm can also be readily extended to learn segment classifiers under weak supervision where the exact positiveness value of each training example is unobserved. In experiments, we show that the e-SVM algorithm can effectively address the segment proposal classification task under both strong supervision (e.g. the pixel-level annotations are available) and the weak supervision (e.g. only bounding-box annotations are available), and outperforms the alternative approaches. Besides, we further validate this method on two major tasks of computer vision: semantic segmentation and object detection. Our method achieves the state-of-the-art object detection performance on PASCAL VOC 2007 dataset.

M68 Predicting Useful Neighborhoods for Lazy Local Learning

Aron Yu aron.yu@utexas.edu
Kristen Grauman grauman@cs.utexas.edu
UT Austin

Lazy local learning methods train a classifier “on the fly” at test time, using only a subset of the training instances that are most relevant to the novel test example. The goal is to tailor the classifier to the properties of the data surrounding the test example. Existing methods assume that the instances most useful for building the local model are strictly those closest to the test example. However, this fails to account for the fact that the success of the resulting classifier depends on the full distribution of selected training instances. Rather than simply gather the test example’s nearest neighbors, we propose to predict the subset of training data that is jointly relevant to training its local model. We develop an approach to discover patterns between queries and their “good” neighborhoods using large-scale multi-label classification with compressed sensing. Given a novel test point, we estimate both the composition and size of the training subset likely to yield an accurate local model. We demonstrate the approach on image classification tasks on SUN and aPascal and show it outperforms traditional global and local approaches.

M69 A Unified Semantic Embedding: Relating Taxonomies and Attributes

Sung Ju Hwang sungju.hwang@disneyresearch.com
Leonid Sigal lsigal@disneyresearch.com
Disney Research

We propose a method that learns a discriminative yet semantic space for object categorization, where we also embed auxiliary semantic entities such as supercategories and attributes. Contrary to prior work which only utilized them as side information, we explicitly embed the semantic entities into the same space where we embed categories, which enables us to represent a category as their linear combination. By exploiting such a unified model for semantics, we enforce each category to be generated as a sparse combination of a supercategory + attributes, with an additional exclusive regularization to learn discriminative composition. The proposed reconstructive regularization guides the discriminative learning process to learn a better generalizing model, as well as generates compact semantic description of each category, which enables humans to analyze what has been learned.

M70 Zero-shot recognition with unreliable attributes

Dinesh Jayaraman dineshj@cs.utexas.edu
Kristen Grauman grauman@cs.utexas.edu
UT Austin

In principle, zero-shot learning makes it possible to train an object recognition model simply by specifying the category's attributes. For example, with classifiers for generic attributes like striped and four-legged, one can construct a classifier for the zebra category by enumerating which properties it possesses --- even without providing zebra training images. In practice, however, the standard zero-shot paradigm suffers because attribute predictions in novel images are hard to get right. We propose a novel random forest approach to train zero-shot models that explicitly accounts for the unreliability of attribute predictions. By leveraging statistics about each attribute's error tendencies, our method obtains more robust discriminative models for the unseen classes. We further devise extensions to handle the few-shot scenario and unreliable attribute descriptions. On three datasets, we demonstrate the benefit for visual category learning with zero or few training examples, a critical domain for rare categories or categories defined on the fly.

M71 Articulated Pose Estimation by a Graphical Model with Image Dependent Pairwise Relations

Xianjie Chen cxj@ucla.edu
Alan Yuille yuille@stat.ucla.edu
UCLA

We present a method for estimating articulated human pose from a single static image based on a graphical model with novel pairwise relations that make adaptive use of local image measurements. More precisely, we specify a graphical model for human pose which exploits the fact the local image measurements can be used both to detect parts (or joints) and also to predict the spatial relationships between them (Image Dependent Pairwise Relations). These spatial relationships are represented by a mixture model. We use Deep Convolutional Neural Networks (DCNNs) to learn conditional probabilities for the presence of parts and their spatial relationships within image patches. Hence our model combines the representational flexibility of graphical models with the efficiency and statistical power of DCNNs. Our method significantly outperforms the state of the art methods on the LSP and FLIC datasets and also performs very well on the Buffy dataset without any training.

M72 Exact Post Model Selection Inference for Marginal Screening

Jason Lee jdl17@stanford.edu
Jonathan Taylor jonathan.taylor@stanford.edu
Stanford University

We develop a framework for post model selection inference, via marginal screening, in linear regression. At the core of this framework is a result that characterizes the exact distribution of linear functions of the response y, conditional on the model being selected (``condition on selection" framework). This allows us to construct valid confidence intervals and hypothesis tests for regression coefficients that account for the selection procedure. In contrast to recent work in high-dimensional statistics, our results are exact (non-asymptotic) and require no eigenvalue-like assumptions on the design matrix X. Furthermore, the computational cost of marginal regression, constructing confidence intervals and hypothesis testing is negligible compared to the cost of linear regression, thus making our methods particularly suitable for extremely large datasets. Although we focus on marginal screening to illustrate the applicability of the condition on selection framework, this framework is much more broadly applicable. We show how to apply the proposed framework to several other selection procedures including orthogonal matching pursuit and marginal screening+Lasso.

M73 On Iterative Hard Thresholding Methods for High-dimensional M-Estimation

Prateek Jain prajain@microsoft.com
Microsoft Research
Ambuj Tewari tewaria@umich.edu
University of Michigan
Puru Kar purushot@cse.iitk.ac.in
Microsoft Research India

The use of M-estimators in generalized linear regression models in high dimensional settings requires risk minimization with hard L_0 constraints. Of the known methods, the class of projected gradient descent (also known as iterative hard thresholding (IHT)) methods is known to offer the fastest and most scalable solutions. However, the current state-of-the-art is only able to analyze these methods in extremely restrictive settings which do not hold in high dimensional statistical models. In this work we bridge this gap by providing the first analysis for IHT-style methods in the high dimensional statistical setting. Our bounds are tight and match known minimax lower bounds. Our results rely on a general analysis framework that enables us to analyze several popular hard thresholding style algorithms (such as HTP, CoSaMP, SP) in the high dimensional regression setting. Finally, we extend our analysis to the problem of low-rank matrix recovery.

M74 A Safe Screening Rule for Sparse Logistic Regression

Jie Wang jie.wang.ustc@asu.edu
Jiayu Zhou jiayu.zhou@asu.edu
Jieping Ye jieping.ye@asu.edu
Arizona State University
Jun Liu jun.liu@sas.com
SAS Institute
Peter Wonka peter.wonka@asu.edu
KAUST

The l1-regularized logistic regression (or sparse logistic regression) is a widely used method for simultaneous classification and feature selection. Although many recent efforts have been devoted to its efficient implementation, its application to high dimensional data still poses significant challenges. In this paper, we present a fast and effective sparse logistic regression screening rule (Slores) to identify the zero components in the solution vector, which may lead to a substantial reduction in the number of features to be entered to the optimization. An appealing feature of Slores is that the data set needs to be scanned only once to run the screening and its computational cost is negligible compared to that of solving the sparse logistic regression problem. Moreover, Slores is independent of solvers for sparse logistic regression, thus Slores can be integrated with any existing solver to improve the efficiency. We have evaluated Slores using high-dimensional data sets from different applications. Extensive experimental results demonstrate that Slores outperforms the existing state-of-the-art screening rules and the efficiency of solving sparse logistic regression is improved by one magnitude in general.

M75 Generalized Dantzig Selector: Application to the k-support norm

Soumyadeep Chatterjee chat0129@umn.edu
Sheng Chen shengc@cs.umn.edu
Arindam Banerjee banerjee@cs.umn.edu
University of Minnesota, Twin Cites

We propose a Generalized Dantzig Selector (GDS) for linear models, in which any norm encoding the parameter structure can be leveraged for estimation. We investigate both computational and statistical aspects of the GDS. Based on conjugate proximal operator, a flexible inexact ADMM framework is designed for solving GDS. Thereafter, non-asymptotic high-probability bounds are established on the estimation error, which rely on Gaussian widths of the unit norm ball and the error set. Further, we consider a non-trivial example of the GDS using k-support norm. We derive an efficient method to compute the proximal operator for k-support norm since existing methods are inapplicable in this setting. For statistical analysis, we provide upper bounds for the Gaussian widths needed in the GDS analysis, yielding the first statistical recovery guarantee for estimation with the k-support norm. The experimental results confirm our theoretical analysis.

M76 Parallel Feature Selection Inspired by Group Testing

Yingbo Zhou yingbozh@buffalo.edu
State University of New York at Buffalo
Utkarsh Porwal utkarshp@buffalo.edu
Ce Zhang czhang@cs.wisc.edu
UW-Madison
Hung Ngo hungngo@buffalo.edu
University at Buffalo, SUNY
Long Nguyen xuanlong@umich.edu
University of Michigan
Christopher Ré chrismre@cs.stanford.edu
Stanford University
Venu Govindaraju govind@buffalo.edu
SUNY Buffalo

This paper presents a parallel feature selection method for classification that scales up to very high dimensions and large data sizes. Our original method is inspired by group testing theory, under which the feature selection procedure consists of a collection of randomized tests to be performed in parallel. Each test corresponds to a subset of features, for which a scoring function may be applied to measure the relevance of the features in a classification task. We develop a general theory providing sufficient conditions under which true features are guaranteed to be correctly identified. Superior performance of our method is demonstrated on a challenging relation extraction task from a very large data set that have both redundant features and sample size in the order of millions. We present comprehensive comparisons with state-of-the-art feature selection methods on a range of data sets, for which our method exhibits competitive performance in terms of running time and accuracy. Moreover, it also yields substantial speedup when used as a pre-processing step for most other existing methods.

M77 Spectral k-Support Norm Regularization

Andrew McDonald andrewmmcd@gmail.com
University College London
Massimiliano Pontil m.pontil@cs.ucl.ac.uk
Dimitris Stamos dimitrios.i.stamos@gmail.com
UCL

The k-support norm has successfully been applied to sparse vector prediction problems. We observe that it belongs to a wider class of norms, which we call the box-norms. Within this framework we derive an efficient algorithm to compute the proximity operator of the squared norm, improving upon the original method for the k-support norm. We extend the norms from the vector to the matrix setting and we introduce the spectral k-support norm. We study its properties and show that it is closely related to the multitask learning cluster norm. We apply the norms to real and synthetic matrix completion datasets. Our findings indicate that spectral k-support norm regularization gives state of the art performance, consistently improving over trace norm regularization and the matrix elastic net.

M78 Beta-Negative Binomial Process and Exchangeable Random Partitions for Mixed-Membership Modeling

Mingyuan Zhou mingyuan.zhou@mccombs.utexas.edu
University of Texas at Austin

The beta-negative binomial process (BNBP), an integer-valued stochastic process, is employed to partition a count vector into a latent random count matrix. As the marginal probability distribution of the BNBP that governs the exchangeable random partitions of grouped data has not yet been developed, current inference for the BNBP has to truncate the number of atoms of the beta process. This paper introduces an exchangeable partition probability function to explicitly describe how the BNBP clusters the data points of each group into a random number of exchangeable partitions, which are shared across all the groups. A fully collapsed Gibbs sampler is developed for the BNBP, leading to a novel nonparametric Bayesian topic model that is distinct from existing ones, with simple implementation, fast convergence, good mixing, and state-of-the-art predictive performance.

M79 The Infinite Mixture of Infinite Gaussian Mixtures

Halid Yerebakan hzyereba@cs.iupui.edu
Murat Dundar murat.dundar@siemens.com
IUPUI
Bartek Rajwa brajwa@purdue.edu
Purdue University

Dirichlet process mixture of Gaussians (DPMG) has been used in the literature for clustering and density estimation problems. However, many real-world data exhibit cluster distributions that cannot be captured by a single Gaussian. Modeling such data sets by DPMG creates several extraneous clusters even when clusters are relatively well-defined. Herein, we present the infinite mixture of infinite Gaussian mixtures (I2GMM) for more flexible modeling of data sets with skewed and multi-modal cluster distributions. Instead of using a single Gaussian for each cluster as in the standard DPMG model, the generative model of I2GMM uses a single DPMG for each cluster. The individual DPMGs are linked together through centering of their base distributions at the atoms of a higher level DP prior. Inference is performed by a collapsed Gibbs sampler that also enables partial parallelization. Experimental results on several artificial and real-world data sets suggest the proposed I2GMM model can predict clusters more accurately than existing variational Bayes and Gibbs sampler versions of DPMG.

M80 Capturing Semantically Meaningful Word Dependencies with an Admixture of Poisson MRFs

David Inouye dinouye@cs.utexas.edu
Pradeep Ravikumar pradeepr@cs.utexas.edu
Inderjit Dhillon inderjit@cs.utexas.edu
University of Texas

We develop a fast algorithm for the Admixture of Poisson MRFs (APM) topic model and propose a novel metric to directly evaluate this model. The APM topic model recently introduced by Inouye et al. (2014) is the first topic model that allows for word dependencies within each topic unlike in previous topic models like LDA that assume independence between words within a topic. Research in both the semantic coherence of a topic models (Mimno et al. 2011, Newman et al. 2010) and measures of model fitness (Mimno & Blei 2011) provide strong support that explicitly modeling word dependencies---as in APM---could be both semantically meaningful and essential for appropriately modeling real text data. Though APM shows significant promise for providing a better topic model, APM has a high computational complexity because $O(p_2)$ parameters must be estimated where p is the number of words (Inouye et al. could only provide results for datasets with p=200). In light of this, we develop a parallel alternating Newton-like algorithm for training the APM model that can handle p=10_4 as an important step towards scaling to large datasets. In addition, Inouye et al. only provided tentative and inconclusive results on the utility of APM. Thus, motivated by simple intuitions and previous evaluations of topic models, we propose a novel evaluation metric based on human evocation scores between word pairs (i.e. how much one word “brings to mind” another word (Boyd-Graber et al. 2006)). We provide compelling quantitative and qualitative results on the BNC corpus that demonstrate the superiority of APM over previous topic models for identifying semantically meaningful word dependencies. (MATLAB code available at: http://bigdata.ices.utexas.edu/software/apm/)

M81 Dynamic Rank Factor Model for Text Streams

Shaobo Han shaobo.han@duke.edu
Lin Du lin.du@duke.edu
Esther Salazar esther.salazar@duke.edu
Lawrence Carin lcarin@duke.edu
Duke University

We propose a semi-parametric and dynamic rank factor model for topic modeling, capable of (1) discovering topic prevalence over time, and (2) learning contemporary multi-scale dependence structures, providing topic and word correlations as a byproduct. The high-dimensional and time-evolving ordinal/rank observations (such as word counts), after an arbitrary monotone transformation, are well accommodated through an underlying dynamic sparse factor model. The framework naturally admits heavy-tailed innovations, capable of inferring abrupt temporal jumps in the importance of topics. Posterior inference is performed through straightforward Gibbs sampling, based on the forward-filtering backward-sampling algorithm. Moreover, an efficient data subsampling scheme is leveraged to speed up inference on massive datasets. The modeling framework is illustrated on two real datasets: the US State of the Union Address and the JSTOR collection from Science.

M82 A provable SVD-based algorithm for learning topics in dominant admixture corpus

Trapit Bansal trapitbansal@gmail.com
Chiranjib Bhattacharyya chiranjib.bhattacharyya@gmail.com
Ravindran Kannan kannan@csa.iisc.ernet.in
Indian Institute of Science

Topic models, such as Latent Dirichlet Allocation (LDA), posit that documents are drawn from admixtures of distributions over words, known as topics. The inference problem of recovering topics from such a collection of documents drawn from admixtures, is NP-hard. Making a strong assumption called separability, [4] gave the first provable algorithm for inference. For the widely used LDA model, [6] gave a provable algorithm using clever tensor-methods. But [4, 6] do not learn topic vectors with bounded l_1 error (a natural measure for probability vectors). Our aim is to develop a model which makes intuitive and empirically supported assumptions and to design an algorithm with natural, simple components such as SVD, which provably solves the inference problem for the model with bounded l_1 error. A topic in LDA and other models is essentially characterized by a group of co-occurring words. Motivated by this, we introduce topic specific Catchwords, a group of words which occur with strictly greater frequency in a topic than any other topic individually and are required to have high frequency together rather than individually. A major contribution of the paper is to show that under this more realistic assumption, which is empirically verified on real corpora, a singular value decomposition (SVD) based algorithm with a crucial pre-processing step of thresholding, can provably recover the topics from a collection of documents drawn from Dominant admixtures. Dominant admixtures are convex combination of distributions in which one distribution has a significantly higher contribution than the others. Apart from the simplicity of the algorithm, the sample complexity has near optimal dependence on w_0, the lowest probability that a topic is dominant, and is better than [4]. Empirical evidence shows that on several real world corpora, both Catchwords and Dominant admixture assumptions hold and the proposed algorithm substantially outperforms the state of the art [5].

M83 Learning a Concept Hierarchy from Multi-labeled Documents

Viet-An Nguyen vietan@cs.umd.edu
Jordan Boyd-Graber jbg@umiacs.umd.edu
Philip Resnik resnik@umd.edu
University of Maryland
Jonathan Chang jonchang@fb.com
Facebook

While topic models can discover patterns of word usage in large corpora, it is difficult to meld this unsupervised structure with noisy, human-provided labels, especially when the label space is large. In this paper, we present a model-Label to Hierarchy (L2H)-that can induce a hierarchy of user-generated labels and the topics associated with those labels from a set of multi-labeled documents. The model is robust enough to account for missing labels from untrained, disparate annotators and provide an interpretable summary of an otherwise unwieldy label set. We show empirically the effectiveness of L2H in predicting held-out words and labels for unseen documents.

M84 On a Theory of Nonparametric Pairwise Similarity for Clustering: Connecting Clustering to Classification

Yingzhen Yang yyang58@illinois.edu
Zhangyang Wang zwang119@illinois.edu
Thomas Huang huang@ifp.illinois.edu
Feng Liang liangf@illinois.edu
Univ. of Illinois Urbana-Champaign Statistics
Shuicheng Yan eleyans@nus.edu.sg
National University of Singapore

Pairwise clustering methods partition the data space into clusters by the pairwise similarity between data points. The success of pairwise clustering largely depends on the pairwise similarity function defined over the data points, where kernel similarity is broadly used. In this paper, we present a novel pairwise clustering framework by bridging the gap between clustering and multi-class classification. This pairwise clustering framework learns an unsupervised nonparametric classifier from each data partition, and search for the optimal partition of the data by minimizing the generalization error of the learned classifiers associated with the data partitions. We consider two nonparametric classifiers in this framework, i.e. the nearest neighbor classifier and the plug-in classifier. Modeling the underlying data distribution by nonparametric kernel density estimation, the generalization error bounds for both unsupervised nonparametric classifiers are the sum of nonparametric pairwise similarity terms between the data points for the purpose of clustering. Under uniform distribution, the nonparametric similarity terms induced by both unsupervised classifiers exhibit a well known form of kernel similarity. We also prove that the generalization error bound for the unsupervised plug-in classifier is asymptotically equal to the weighted volume of cluster boundary for Low Density Separation, a widely used criteria for semi-supervised learning and clustering. Based on the derived nonparametric pairwise similarity using the plug-in classifier, we propose a new nonparametric exemplar-based clustering method with enhanced discriminative capability, whose superiority is evidenced by the experimental results.

M85 Robust Bayesian Max-Margin Clustering

Changyou Chen cchangyou@gmail.com
Duke University
Jun Zhu dcszj@mail.tsinghua.edu.cn
Tsinghua University
Xinhua Zhang xinhua.zhang.cs@gmail.com
NICTA and Australian National University

We present max-margin Bayesian clustering (BMC), a general and robust framework that incorporates the max-margin criterion into Bayesian clustering models, as well as two concrete models of BMC to demonstrate its flexibility and effectiveness in dealing with different clustering tasks. The Dirichlet process max-margin Gaussian mixture is a nonparametric Bayesian clustering model that relaxes the underlying Gaussian assumption of Dirichlet process Gaussian mixtures by incorporating max-margin posterior constraints, and is able to infer the number of clusters from data. We further extend the ideas to present max-margin clustering topic model, which can learn the latent topic representation of each document while at the same time cluster documents in the max-margin fashion. Extensive experiments are performed on a number of real datasets, and the results indicate superior clustering performance of our methods compared to related baselines.

M86 On Integrated Clustering and Outlier Detection

Lionel Ott lott4241@uni.sydney.edu.au
Linsey Pang qlinsey@it.usyd.edu.au
Fabio Ramos fabio.ramos@sydney.edu.au
University of Sydney
Sanjay Chawla chawsanj@gmail.com
Sydney University

We model the joint clustering and outlier detection problem using an extension of the facility location formulation. The advantages of combining clustering and outlier selection include: (i) the resulting clusters tend to be compact and semantically coherent (ii) the clusters are more robust against data perturbations and (iii) the outliers are contextualised by the clusters and more interpretable. We provide a practical subgradient-based algorithm for the problem and also study the theoretical properties of algorithm in terms of approximation and convergence. Extensive evaluation on synthetic and real data sets attest to both the quality and scalability of our proposed method.

M87 Convex Optimization Procedure for Clustering: Theoretical Revisit

Changbo Zhu zhuchangbo@gmail.com
Shuicheng Yan eleyans@nus.edu.sg
Huan Xu mpexuh@nus.edu.sg
National University of Singapore
Chenlei Leng c.leng@warwick.ac.uk
University of Warwick

In this paper, we present theoretical analysis of SON~--~a convex optimization procedure for clustering using a sum-of-norms (SON) regularization recently proposed in \cite{ICML2011Hocking_419,SON, Lindsten650707, pelckmans2005convex}. In particular, we show if the samples are drawn from two cubes, each being one cluster, then SON can provably identify the cluster membership provided that the distance between the two cubes is larger than a threshold which (linearly) depends on the size of the cube and the ratio of numbers of samples in each cluster. To the best of our knowledge, this paper is the first to provide a rigorous analysis to understand why and when SON works. We believe this may provide important insights to develop novel convex optimization based algorithms for clustering.

M88 Greedy Subspace Clustering

Dohyung Park dhpark@utexas.edu
Constantine Caramanis constantine@utexas.edu
Sujay Sanghavi sanghavi@mail.utexas.edu
UT Austin

We consider the problem of subspace clustering: given points that lie on or near the union of many low-dimensional linear subspaces, recover the subspaces. To this end, one first identifies sets of points close to the same subspace and uses the sets to estimate the subspaces. As the geometric structure of the clusters (linear subspaces) forbids proper performance of general distance based approaches such as K-means, many model-specific methods have been proposed. In this paper, we provide new simple and efficient algorithms for this problem. Our statistical analysis shows that the algorithms are guaranteed exact (perfect) clustering performance under certain conditions on the number of points and the affinity be- tween subspaces. These conditions are weaker than those considered in the standard statistical literature. Experimental results on synthetic data generated from the standard unions of subspaces model demonstrate our theory. We also show that our algorithm performs competitively against state-of-the-art algorithms on real-world applications such as motion segmentation and face clustering, with much simpler implementation and lower computational cost.

M89 Graph Clustering With Missing Data: Convex Algorithms and Analysis

Ramya Korlakai Vinayak ramya@caltech.edu
California Institute of Technology
Samet Oymak soymak@caltech.edu
Babak Hassibi hassibi@systems.caltech.edu
Caltech

We consider the problem of finding clusters in an unweighted graph, when the graph is partially observed. We analyze two programs, one which works for dense graphs and one which works for both sparse and dense graphs, but requires some a priori knowledge of the total cluster size, that are based on the convex optimization approach for low-rank matrix recovery using nuclear norm minimization. For the commonly used Stochastic Block Model, we obtain **explicit** bounds on the parameters of the problem (size and sparsity of clusters, the amount of observed data) and the regularization parameter characterize the success and failure of the programs. We corroborate our theoretical findings through extensive simulations. We also run our algorithm on a real data set obtained from crowdsourcing an image classification task on the Amazon Mechanical Turk, and observe significant performance improvement over traditional methods such as k-means.

M90 Dimensionality Reduction with Subspace Structure Preservation

Devansh Arpit devansha@buffalo.edu
Ifeoma Nwogu inwogu@buffalo.edu
Venu Govindaraju govind@buffalo.edu
SUNY Buffalo

Modeling data as being sampled from a union of independent subspaces has been widely applied to a number of real world applications. However, dimensionality reduction approaches that theoretically preserve this independence assumption have not been well studied. Our key contribution is to show that $2K$ projection vectors are sufficient for the independence preservation of any K class data sampled from a union of independent subspaces. It is this non-trivial observation that we use for designing our dimensionality reduction technique. In this paper, we propose a novel dimensionality reduction algorithm that theoretically preserves this structure for a given dataset. We support our theoretical analysis with empirical results on both synthetic and real world data achieving \textit{state-of-the-art} results compared to popular dimensionality reduction techniques.

M91 Distributed Balanced Clustering via Mapping Coresets

Mohammadhossein Bateni bateni@google.com
Aditya Bhaskara bhaskaraaditya@gmail.com
Silvio Lattanzi silviol@google.com
Vahab Mirrokni mirrokni@google.com
Google Research NYC

Large-scale clustering of data points in metric spaces is an important problem in mining big data sets. For many applications, we face explicit or implicit size constraints for each cluster which leads to the problem of clustering under capacity constraints or the ``balanced clustering'' problem. Although the balanced clustering problem has been widely studied, developing a theoretically sound distributed algorithm remains an open problem. In the present paper we develop a general framework based on ``mapping coresets'' to tackle this issue. For a wide range of clustering objective functions such as k-center, k-median, and k-means, our techniques give distributed algorithms for balanced clustering that match the best known single machine approximation ratios.

M92 Zeta Hull Pursuits: Learning Nonconvex Data Hulls

Yuanjun Xiong xy012@ie.cuhk.edu.hk
Xiaoou Tang xtang@ie.cuhk.edu.hk
The Chinese University of Hong Kong
Wei Liu wliu@ee.columbia.edu
IBM T. J. Watson Research Center
Deli Zhao zhaodeli@gmail.com

Selecting a small informative subset from a given dataset, also called column sampling, has drawn much attention in machine learning. For incorporating structured data information into column sampling, research efforts were devoted to the cases where data points are fitted with clusters, simplices, or general convex hulls. This paper aims to study nonconvex hull learning which has rarely been investigated in the literature. In order to learn data-adaptive nonconvex hulls, a novel approach is proposed based on a graph-theoretic measure that leverages graph cycles to characterize the structural complexities of input data points. Employing this measure, we present a greedy algorithmic framework, dubbed Zeta Hulls, to perform structured column sampling. The process of pursuing a Zeta hull involves the computation of matrix inverse. To accelerate the matrix inversion computation and reduce its space complexity as well, we exploit a low-rank approximation to the graph adjacency matrix by using an efficient anchor graph technique. Extensive experimental results show that data representation learned by Zeta Hulls can achieve state-of-the-art accuracy in text and image classification tasks.

M93 The Bayesian Case Model: A Generative Approach for Case-Based Reasoning and Prototype Classification

Been Kim beenkim@csail.mit.edu
Cynthia Rudin crudin@princeton.edu
Julie Shah julie_a_shah@csail.mit.edu
Massachusetts Institute of Technology

We present the Bayesian Case Model (BCM), a general framework for Bayesian case-based reasoning (CBR) and prototype classification and clustering. BCM brings the intuitive power of CBR to a Bayesian generative framework. The BCM learns prototypes, the ``quintessential'' observations that best represent clusters in a dataset, by performing joint inference on cluster labels, prototypes and important features. Simultaneously, BCM pursues sparsity by learning subspaces, the sets of features that play important roles in the characterization of the prototypes. The prototype and subspace representation provides quantitative benefits in interpretability while preserving classification accuracy. Human subject experiments verify statistically significant improvements to participants' understanding when using explanations produced by BCM, compared to those given by prior art.

M94 Consistency of Spectral Partitioning of Uniform Hypergraphs under Planted Partition Model

Debarghya Ghoshdastidar debarghya.g@csa.iisc.ernet.in
Ambedkar Dukkipati ad@csa.iisc.ernet.in
Indian Institute of Science

Spectral graph partitioning methods have received significant attention from both practitioners and theorists in computer science. Some notable studies have been carried out regarding the behavior of these methods for infinitely large sample size (von Luxburg et al., 2008; Rohe et al., 2011), which provide sufficient confidence to practitioners about the effectiveness of these methods. On the other hand, recent developments in computer vision have led to a plethora of applications, where the model deals with multi-way affinity relations and can be posed as uniform hypergraphs. In this paper, we view these models as random m-uniform hypergraphs and establish the consistency of spectral algorithm in this general setting. We develop a planted partition model or stochastic blockmodel for such problems using higher order tensors, present a spectral technique suited for the purpose and study its large sample behavior. The analysis reveals that the algorithm is consistent for m-uniform hypergraphs for larger values of m, and also the rate of convergence improves for increasing m. Our result provides the first theoretical evidence that establishes the importance of m-way affinities.

M95 Projecting Markov Random Field Parameters for Fast Mixing

Xianghang Liu xianghang.liu@nicta.com.au
Justin Domke justin.domke@nicta.com.au
NICTA and Australian National University

Markov chain Monte Carlo (MCMC) algorithms are simple and extremely powerful techniques to sample from almost arbitrary distributions. The flaw in practice is that it can take a large and/or unknown amount of time to converge to the stationary distribution. This paper gives sufficient conditions to guarantee that univariate Gibbs sampling on Markov Random Fields (MRFs) will be fast mixing, in a precise sense. Further, an algorithm is given to project onto this set of fast-mixing parameters in the Euclidean norm. Following recent work, we give an example use of this to project in various divergence measures, comparing of univariate marginals obtained by sampling after projection to common variational methods and Gibbs sampling on the original parameters.

M96 Blossom Tree Graphical Models

Zhe Liu zheliu@uchicago.edu
John Lafferty lafferty@gmail.com
University of Chicago

We combine the ideas behind trees and Gaussian graphical models to form a new nonparametric family of graphical models. Our approach is to attach nonparanormal "blossoms", with arbitrary graphs, to a collection of nonparametric trees. The tree edges are chosen to connect variables that most violate joint Gaussianity. The non-tree edges are partitioned into disjoint groups, and assigned to tree nodes using a nonparametric partial correlation statistic. A nonparanormal blossom is then "grown" for each group using established methods based on the graphical lasso. The result is a factorization with respect to the union of the tree branches and blossoms, defining a high-dimensional joint density that can be efficiently estimated and evaluated on test points. Theoretical properties and experiments with simulated and real data demonstrate the effectiveness of blossom trees.

M97 Distributed Parameter Estimation in Probabilistic Graphical Models

Yariv Mizrahi yariv@math.ubc.ca
University of British Columbia
Misha Denil misha.denil@cs.ox.ac.uk
Nando de Freitas nando@cs.ox.ac.uk
University of Oxford

This paper presents foundational theoretical results on distributed parameter estimation for undirected probabilistic graphical models. It introduces a general condition on composite likelihood decompositions of these models which guarantees the global consistency of distributed estimators, provided the local estimators are consistent.

M98 Elementary Estimators for Graphical Models

Eunho Yang eunho@cs.utexas.edu
Pradeep Ravikumar pradeepr@cs.utexas.edu
UT Austin
Aurelie Lozano aclozano@us.ibm.com
IBM T.J. Watson Research Center

We propose a class of closed-form estimators for sparsity-structured graphical models, expressed as exponential family distributions, under high-dimensional settings. Our approach builds on observing the precise manner in which the classical graphical model MLE ``breaks down'' under high-dimensional settings. Our estimator uses a carefully constructed, well-defined and closed-form backward map, and then performs thresholding operations to ensure the desired sparsity structure. We provide a rigorous statistical analysis that shows that surprisingly our simple class of estimators recovers the same asymptotic convergence rates as those of the ℓ_1-regularized MLEs that are much more difficult to compute. We corroborate this statistical performance, as well as significant computational advantages via simulations of both discrete and Gaussian graphical models.

M99 Structure learning of antiferromagnetic Ising models

Guy Bresler gbresler@mit.edu
David Gamarnik gamarnik@mit.edu
Devavrat Shah devavrat@mit.edu
Massachusetts Institute of Technology

In this paper we investigate the computational complexity of learning the graph structure underlying a discrete undirected graphical model from i.i.d. samples. Our first result is an unconditional computational lower bound of $\Omega(p^{d/2})$ for learning general graphical models on p nodes of maximum degree d, for the class of statistical algorithms recently introduced by Feldman et al. The construction is related to the notoriously difficult learning parities with noise problem in computational learning theory. Our lower bound shows that the $\tilde{O}(p^{d+2})$ runtime required by Bresler, Mossel, and Sly's exhaustive-search algorithm cannot be significantly improved without restricting the class of models. Aside from structural assumptions on the graph such as it being a tree, hypertree, tree-like, etc., most recent papers on structure learning assume that the model has the correlation decay property. Indeed, focusing on ferromagnetic Ising models, Bento and Montanari showed that all known low-complexity algorithms fail to learn simple graphs when the interaction strength exceeds a number related to the correlation decay threshold. Our second set of results gives a class of repelling (antiferromagnetic) models that have the **opposite** behavior: very strong repelling allows efficient learning in time $\tilde{O}(p^2)$. We provide an algorithm whose performance interpolates between $\tilde{O}(p^2)$and $\tilde{O}(p^{d+2})$ depending on the strength of the repulsion.

M100 On Sparse Gaussian Chain Graph Models

Calvin McCarter calvin199@gmail.com
Seyoung Kim sssykim@cs.cmu.edu
Carnegie Mellon University

In this paper, we address the problem of learning the structure of Gaussian chain graph models in a high-dimensional space. Chain graph models are generalizations of undirected and directed graphical models that contain a mixed set of directed and undirected edges. While the problem of sparse structure learning has been studied extensively for Gaussian graphical models and more recently for conditional Gaussian graphical models (CGGMs), there has been little previous work on the structure recovery of Gaussian chain graph models. We consider linear regression models and a re-parameterization of the linear regression models using CGGMs as building blocks of chain graph models. We argue that when the goal is to recover model structures, there are many advantages of using CGGMs as chain component models over linear regression models, including convexity of the optimization problem, computational efficiency, recovery of structured sparsity, and ability to leverage the model structure for semi-supervised learning. We demonstrate our approach on simulated and genomic datasets.

TUESDAY SESSIONS

ORAL SESSION

SESSION 1 - 9:00 – 10:10 AM

Session Chair: Jeff Blimes

INVITED TALK: Subgradient Methods for Huge-Scale Optimization Problems

Yurii Nesterov Yurii.Nesterov@uclouvain.be
Catholic University of Louvain (UCL)

We consider a new class of huge-scale problems, the problems with sparse subgradients. The most important functions of this type are piece-wise linear. For optimization problems with uniform sparsity of corresponding linear operators, we suggest a very efficient implementation of subgradient iterations, which total cost depends logarithmically in the dimension. This technique is based on a recursive update of the results of matrix/vector products and the values of symmetric functions. It works well, for example, for matrices with few nonzero diagonals and for max-type functions. We show that the updating technique can be efficiently coupled only with the simplest subgradient methods. Similar results can be obtained for a new nonsmooth random variant of a coordinate descent scheme. We present also the promising results of preliminary computational experiments.

Yurii Nesterov is a professor at Center for Operations Research and Econometrics (CORE) in Catholic University of Louvain (UCL), Belgium. He received Ph.D. degree (Applied Mathematics) in 1984 at Institute of Control Sciences, Moscow. Starting from 1993 he works at CORE. His research interests are related to complexity issues and efficient methods for solving various optimization problems. The main results are obtained in Convex Optimization (optimal methods for smooth problems, polynomial-time interior-point methods, smoothing technique for structural optimization, cubic regularization of Newton method, optimization methods for huge-scale problems). He is an author of 4 monographs and more than 80 refereed papers in the leading optimization journals. He got the Dantzig Prize from SIAM and Mathematical Programming society in 2000, von Neumann Theory Prize from INFORMS in 2009, and SIAM Outstanding paper award in 2014.

Provable Submodular Minimization using Wolfe's Algorithm

Deeparnab Chakrabarty dechakr@microsoft.com
Microsoft
Prateek Jain prajain@microsoft.com
Microsoft Research
Pravesh Kothari kothari@cs.utexas.edu
UT Austin

Owing to several applications in large scale learning and vision problems, fast submodular function minimization (SFM) has become a critical problem. Theoretically, unconstrained SFM can be performed in polynomial time (Iwata and Orlin 2009), however these algorithms are not practical. In 1976, Wolfe proposed an algorithm to find the minimum Euclidean norm point in a polytope, and in 1980, Fujishige showed how Wolfe's algorithm can be used for SFM. For general submodular functions, the Fujishige-Wolfe minimum norm algorithm seems to have the best empirical performance. Despite its good practical performance, theoretically very little is known about Wolfe's minimum norm algorithm -- to our knowledge the only result is an exponential time analysis due to Wolfe himself. In this paper we give a maiden convergence analysis of Wolfe's algorithm. We prove that in t iterations, Wolfe's algorithm returns a O(1/t)-approximate solution to the min-norm point. We also prove a robust version of Fujishige's theorem which shows that an O(1/n^2)-approximate solution to the min-norm point problem implies exact submodular minimization. As a corollary, we get the first pseudo-polynomial time guarantee for the Fujishige-Wolfe minimum norm algorithm for submodular function minimization. In particular, we show that the min-norm point algorithm solves SFM in O(n^7F^2)-time, where *F* is an upper bound on the maximum change a single element can cause in the function value.

SPOTLIGHT SESSION

SESSION 1: 10:10 – 10:30 AM

- **A Differential Equation for Modeling Nesterov's Accelerated Gradient Method: Theory and Insights**
  W. Su, S. Boyd, E. Candes, Stanford University
  See abstract T4

- **Greedy Algorithms for Finding Diverse Subsets in Exponentially-Large Structured Item Sets**
  A. Prasad, University of Texas at Austin; S. Jegelka, UC Berkeley; D. Batra, Virginia Tech
  See abstract T

- **Learning Distributed Representations for Structured Output Prediction**
  V. Srikumar, C. Manning, Stanford University
  See abstract T40

- **Predictive Entropy Search for Efficient Global Optimization of Black-box Functions**
  J. Hernández-Lobato, Harvard University; M. Hoffman, Z. Ghahramani, University of Cambridge
  See abstract T7

- **The Noisy Power Method: A Meta Algorithm with Applications**
  M. Hardt, IBM Research Almaden; E. Price, IBM
  See abstract T10

- **Two-Layer Feature Reduction for Sparse-Group Lasso via Decomposition of Convex Sets**
  J. Wang, J. Ye, Arizona State University
  See abstract T9

ORAL SESSION

SESSION 2: 10:55 – 11:40 AM

Session Chair: D. Sculley

NIPS Award Session

Median selection subset aggregation for parallel inference

Xiangyu Wang xw56@stat.duke.edu
David Dunson dunson@stat.duke.edu
Duke University
Peichao Peng ppeichao@wharton.upenn.edu
University of Pennsylvania

For massive data sets, efficient computation commonly relies on distributed algorithms that store and process subsets of the data on different machines, minimizing communication costs. Our focus is on regression and classification problems involving many features. A variety of distributed algorithms have been proposed in this context, but challenges arise in defining an algorithm with low communication, theoretical guarantees and excellent practical performance in general settings. We propose a MEdian Selection Subset Aggregation Estimator (message) algorithm, which attempts to solve these problems. The algorithm applies feature selection in parallel for each subset using Lasso or another method, calculates the `median' feature inclusion index, estimates coefficients for the selected features in parallel for each subset, and then averages these estimates. The algorithm is simple, involves very minimal communication, scales efficiently in both sample and feature size, and has theoretical guarantees. In particular, we show model selection consistency and coefficient estimation efficiency. Extensive experiments show excellent performance in variable selection, estimation, prediction, and computation time relative to usual competitors.

Asymmetric LSH (ALSH) for Sublinear Time Maximum Inner Product Search (MIPS)

Anshumali Shrivastava anshu@cs.cornell.edu
Cornell University
Ping Li pingli@stat.rutgers.edu
Rutgers University

We present the first provably sublinear time algorithm for approximate **Maximum Inner Product Search** (MIPS). Searching with (un-normalized) inner product as the underlying similarity measure is a known difficult problem and finding hashing schemes for MIPS was considered hard. While the existing Locality Sensitive Hashing (LSH) framework is insufficient for solving MIPS, in this paper we extend the LSH framework to allow asymmetric hashing schemes. Our proposal is based on an interesting mathematical phenomenon that, the problem of searching for inner products, after independent asymmetric transformations, can be converted into the problem of approximate near neighbor search. This key observation makes efficient sublinear hashing scheme for MIPS possible. In the extended asymmetric LSH (ALSH) framework, we provide an explicit construction of provably fast hashing scheme for MIPS. Our proposed algorithm is simple and easy to implement. The proposed hashing scheme outperforms the two popular existing LSH heuristics (i) Signed Random Projection (SRP) and (ii) hashing based on p-stable distributions for L2 norm (L2LSH), in the collaborative filtering task of item recommendations on Netflix and Movielens (10M) datasets.

SPOTLIGHT SESSION

SESSION 2: 11:40 AM – 12:00 PM

- **Large-scale L-BFGS using MapReduce**
  W. Chen, Z. Wang, J. Zhou, Microsoft
  See abstract T5

- **A Latent Source Model for Online Collaborative Filtering**
  G. Bresler, G. Chen, D. Shah, Massachusetts Institute of Technology
  See abstract T11

- **Discrete Graph Hashing**
  W. Liu, IBM T. J. Watson Research Center; C. Mu, S. Chang, Columbia University; S. Kumar, Google
  See abstract T60

- **Clustering from Labels and Time-Varying Graphs**
  S. Lim, National University of Singapore; Y. Chen, University of California, Berkeley; H. Xu, NUS
  See abstract T68

- **Scalable methods for nonnegative matrix factorizations of near-separable tall-and-skinny matrices**
  A. Benson, J. Lee, Stanford University; B. Rajwa, D. Gleich, Purdue University
  See abstract T75

- **Recovery of Coherent Data via Low-Rank Dictionary Pursuit**
  G. Liu, Rutgers; P. Li, Rutgers University
  See abstract T78

ORAL SESSION

SESSION 3: 2:00 – 3:30 PM

Session Chair: Hugo Larochelle

INVITED TALK: Role of Coupled Networks in 21st Century Energy Infrastructure

Arunava Majumdar arunava@google.com
Google Inc. & Stanford University

Our modern economy overwhelmingly depends on the electricity infrastructure or the grid. The architecture of the grid owes its origins to Tesla, Edison and their industrial partners, and has remained largely unchanged since then. But this paradigm is being challenged by a confluence of factors, namely, age of the physical assets, cyber-physical security, weather-related stresses, and rapidly reducing cost of renewable electricity and storage. These are producing some unmistakable trends towards increased integration of distributed generation and storage. Furthermore, the integration of communication, computing and control to automate system operation is at its early stages. At scale, the resulting architecture will produce coupling of two large networks - electricity and information. Finally, the abundance and low cost of natural gas has already led to increased dependence of natural gas for electricity generation. The coupling between natural gas and electricity networks is producing dramatic fluctuations in prices with severe economic impact. This talk will

discuss the changing landscape of the electricity infrastructure, the challenges that may arise from network coupling, and opportunities to use machine learning to address these challenges.

Arun Majumdar is the Jay Precourt Provostial Chair Professor at Stanford University, where he serves on the faculty of the Department of Mechanical Engineering and is a Senior Fellow of the Precourt Institute for Energy. Prior to joining Stanford, he was the Vice President for Energy at Google, where he created several energy technology initiatives and advised the company on its broader energy strategy. He continues to be a consultant to Google on energy. In October 2009, Dr. Arun Majumdar was nominated by President Obama and confirmed by the Senate to become the Founding Director of the Advanced Research Projects Agency - Energy (ARPA-E), where he served till June 2012. Between March 2011 and June 2012, Dr. Majumdar also served as the Acting Under Secretary of Energy, and a Senior Advisor to the Secretary of Energy. Prior to joining the Department of Energy, Dr. Majumdar was the Almy and Agnes Maynard Chair Professor of Mechanical Engineering and Materials Science and Engineering at the University of California, Berkeley and the Associate Laboratory Director for Energy and Environment at Lawrence Berkeley National Laboratory. His research career includes the science and engineering of nanoscale materials and devices as well as large engineered systems. Dr. Majumdar is a member of the National Academy of Engineering and the American Academy of Arts and Sciences. He received his bachelor's degree in Mechanical Engineering at the Indian Institute of Technology, Bombay in 1985 and his Ph.D. from the University of California, Berkeley in 1989. He currently serves on the US Secretary of Energy's Advisory Board, the Council for the National Academy of Engineering and the Electric Power Research Institute, as well as the Science Board of the Stanford Linear Accelerator Center (SLAC) and the Oak Ridge National Laboratory. He is a member of the International Advisory Panel for Energy of the Singapore Ministry of Trade and Industry and the US delegation for the US-India Track II dialogue on climate change and energy.

Analog Memories in a Balanced Rate-Based Network of E-I Neurons

Dylan Festa df325@cam.ac.uk
Mate Lengyel m.lengyel@eng.cam.ac.uk
University of Cambridge
Guillaume Hennequin guillaume.hennequin@epfl.ch
EPFL

The dynamics of brain circuits often settles in persistent and graded activity states that are sometimes seen as signatures of the autoassociative retrieval of memorized items. Despite decades of theoretical work on the subject, the mechanisms that support the storage and retrieval of memories remain unclear. Previous proposals concerning the dynamics of memory networks have fallen short of incorporating some key physiological constraints in a unified way. Specifically, some models violate Dale's law (i.e. allow neurons to be both excitatory (E) and inhibitory (I)), while some others restrict the representation of memories to a binary format, or induce recall states in which most neurons fire at saturation. We propose a novel control-theoretic framework to build functioning attractor networks that meet all of the above physiological constraints. We directly optimize networks of E and I neurons to force collections of arbitrary, analog activity patterns to become stable fixed points of the dynamics. The resulting networks operate in the balanced regime, are robust to corruptions of the memory cue as well as ongoing noise, and incidentally explain the reduction of trial-to-trial variability following stimulus onset that is ubiquitously observed in the sensory and motor cortices. Our results constitute a step forward in our growing understanding of the neural substrate of memory.

Feedforward Learning of Mixture Models

Matthew Lawlor matthew.lawlor@yale.edu
Steven Zucker zucker@cs.yale.edu
Yale University

We develop a biologically-plausible learning rule that provably converges to the class means of general mixture models. It generalizes the classical BCM neural rule within a tensor framework, incorporates a multi-view assumption, and shows how learning of higher-order structure is possible. Together our model provides a novel information processing interpretation to spike-timing-dependent plasticity.

SPOTLIGHT SESSION

SESSION 3: 3:30 – 3:50 PM

- **Spatio-temporal Representations of Uncertainty in Spiking Neural Networks**
  C. Savin, University of Cambridge; S. Denève, École Normale Supérieure
  See abstract T47

- **A Bayesian model for identifying hierarchically organised states in neural population activity**
  P. Putzky, G. Bassetto, Max Planck Institute for Biological Cybernetics; F. Franzen, ; J. Macke, MPI for Biological Cybernetics
  See abstract T49

- **Clustered factor analysis of multineuronal spike data**
  L. Buesing, Gatsby Unit, UCL; J. Cunningham, T. Machado, L. Paninski, Columbia University
  See abstract T50

- **Sparse Space-Time Deconvolution for Calcium Image Analysis**
  F. Diego Andilla, F. Hamprecht, University of Heidelberg
  See abstract T63

- **Scalable Inference for Neuronal Connectivity from Calcium Imaging**
  A. Fletcher, University of California, Santa Cruz; S. Rangan, Polytechnic Institute of New York University
  See abstract T70

- **Design Principles of the Hippocampal Cognitive Map**
  K. Stachenfeld, M. Botvinick, Princeton University; S. Gershman, Massachusetts Institute of Technology
  See abstract T73

ORAL SESSION

SESSION 4: 4:15 – 5:40 PM

Session Chair: Ranzato Marc'Aurelio

NIPS Award Session

Conditional Random Field Autoencoders for Unsupervised Structured Prediction

Waleed Ammar waleed.ammar@gmail.com
CMU
Chris Dyer cdyer@cs.cmu.edu
Noah Smith nasmith@cs.cmu.edu
Carnegie Mellon University

We introduce a framework for unsupervised learning of structured predictors with overlapping, global features. Each input's latent representation is predicted conditional on the observable data using a feature-rich conditional random field. Then a reconstruction of the input is (re)generated, conditional on the latent structure, using models for which maximum likelihood estimation has a closed-form. Our autoencoder formulation enables efficient learning without making unrealistic independence assumptions or restricting the kinds of features that can be used. We illustrate insightful connections to traditional autoencoders, posterior regularization and multi-view learning. We show competitive results with instantiations of the model for two canonical NLP tasks: part-of-speech induction and bitext word alignment, and show that training our model can be substantially more efficient than comparable feature-rich baselines.

Learning Generative Models with Visual Attention

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

Attention has long been proposed by psychologists to be important for efficiently dealing with the massive amounts of sensory stimulus in the neocortex. Inspired by the attention models in visual neuroscience and the need for object-centered data for generative models, we propose a deep-learning based generative framework using attention. The attentional mechanism propagates signals from the region of interest in a scene to an aligned canonical representation for generative modeling. By ignoring scene background clutter, the generative model can concentrate its resources on the object of interest. A convolutional neural net is employed to provide good initializations during posterior inference which uses Hamiltonian Monte Carlo. Upon learning images of faces, our model can robustly attend to face regions of novel test subjects. More importantly, our model can learn generative models of new faces from a novel dataset of large images where the face locations are not known.

Sequence to Sequence Learning with Neural Networks

Ilya Sutskever ilyasu@google.com
Oriol Vinyals oriol18@gmail.com
Google Research
Quoc V. Le qvl@google.com
Google

Deep Neural Networks (DNNs) are powerful models that have shown significant success on difficult supervised learning tasks. Although generally their performance improves as dataset and model size increases, they are applicable only to problems whose inputs and outputs are encoded as vectors of fixed dimensionality, so they cannot straightforwardly learn to map sequences to sequences. Existing approaches for sequence problems (such as speech recognition and machine translation) use problem specific multistage pipelines that are highly complex. In this paper, we present a general end-to-end approach to sequence learning. Our method uses a Long Short-Term Memory (LSTM) to convert the input sequence into a vector of a fixed dimensionality, and then another LSTM to extract the target sequence from input's fixed-size representation. We evaluate the feasibility of our approach on a machine translation problem. Our main result is that on the WMT dataset, our system achieves near state-of-the-art BLUE score on the German to English translation when all the words in the input sentences are frequent. We also obtained an improvement of 1.5 BLUE points on an English to French translation over a strong baseline by rescoring the 1000 most-likely hypotheses. Finally, we show that our LSTM learns sensible phrase and sentence representations that are sensitive to word order and are fairly invariant to the active and the passive voice.

Quantifying the transferability of features in deep neural networks

Jason Yosinski nps.jyo@0sg.net
Hod Lipson hod.lipson@cornell.edu
Cornell University
Jeff Clune jeffclune@uwyo.edu
Yoshua Bengio bengioy@iro.umontreal.ca
University of Montreal

A high percentage of recently reported deep networks trained on natural images exhibit a curious aspect in common: they all learn features similar to Gabor filters on the first layer. Such features appear to be general, as opposed to specific for a particular task. Because higher layers are difficult to visualize, not as much is known about whether the are general or specific. In this paper we experimentally quantify the generality of neurons on each layer of a deep convolutional neural network and through this expose a few surprising aspects. First, we find that transferability is negatively affected by two distinct issues — not only the expected specialization of higher layer neurons to their original task at the expense of the target task, but also optimization difficulties related to splitting networks in the middle of co-adapted neurons. In an example network trained on ImageNet, we demonstrate that either of these two issues may dominate, depending on whether features are transferred from the bottom, middle, or top of the network. We also show how the transferability gap grows as the distance between base task and target task increases, but how even transfer from distant tasks can be better than using random features. Finally, we show a surprising result that initializing a network with transferred features from almost any number of layers can produce a boost to generalization that lingers even after fine-tuning to the target dataset.

SPOTLIGHT SESSION

SESSION 4: 5:40 – 6:00 PM

- **Learning Deep Features for Scene Recognition using PLACES Database**
  B. Zhou, A. Garcia, A. Oliva, A. Torralba, Massachusetts Institute of Technology; J. Xiao, Princeton University
  See abstract T91

- **Semi-supervised Learning with Deep Generative Models**
  D. Kingma, M. Welling, University of Amsterdam; S. Mohamed, Google DeepMind; D. Jimenez Rezende, Google Deepmind
  See abstract T93

- **Two-Stream Convolutional Networks for Action Recognition in Videos**
  K. Simonyan, A. Zisserman, University of Oxford
  See abstract T99

- **Learning to Discover Efficient Mathematical Identities**
  W. Zaremba, New York University; K. Kurach, Google Zurich / University of Warsaw; R. Fergus, NYU
  See abstract T100

- **Searching for Higgs Boson Decay Modes with Deep Learning**
  P. Sadowski, D. Whiteson, University of California Irvine; P. Baldi, UC Irvine
  See abstract 103

- **Convolutional Kernel Networks**
  J. Mairal, INRIA; P. Koniusz, Z. Harchaoui, Inria; C. Schmid, Inria Alpes
  See abstract T104

POSTER SESSION

POSTERS 7:00 – 11:59 PM

T1 **Rounding-based Moves for Metric Labeling**
M. Kumar

T2 **Provable Submodular Minimization using Wolfe's Algorithm**
D. Chakrabarty, P. Jain, P. Kothari

T3 **Stochastic Gradient Descent, Weighted Sampling, and the Randomized Kaczmarz algorithm**
D. Needell, R. Ward, N. Srebro

T4 **A Differential Equation for Modeling Nesterov's Accelerated Gradient Method: Theory and Insights**
W. Su, S. Boyd, E. Candes

T5 **Large-scale L-BFGS using MapReduce**
W. Chen, Z. Wang, J. Zhou

T6 **An Accelerated Proximal Coordinate Gradient Method**
Q. Lin, Z. Lu, L. Xiao

T7 **Predictive Entropy Search for Efficient Global Optimization of Black-box Functions**
J. Hernández-Lobato, M. Hoffman, Z. Ghahramani

T8 **Inference by Learning: Speeding-up Graphical Model Optimization via a Coarse-to-Fine Cascade of Pruning Classifiers**
B. Conejo, N. Komodakis, S. Leprince, J. Avouac

T9 **Two-Layer Feature Reduction for Sparse-Group Lasso via Decomposition of Convex Sets**
J. Wang, J. Ye

T10 **The Noisy Power Method: A Meta Algorithm with Applications**
M. Hardt, E. Price

T11 **A Latent Source Model for Online Collaborative Filtering**
G. Bresler, G. Chen, D. Shah

T12 **Probabilistic low-rank matrix completion on finite alphabets**
J. Lafond, O. Klopp, E. Moulines, J. Salmon

T13 **Controlling privacy in recommender systems**
Y. Xin, T. Jaakkola

T14 **Content-based recommendations with Poisson factorization**
P. Gopalan, L. Charlin, D. Blei

T15 **Minimax-optimal Inference from Partial Rankings**
B. Hajek, S. Oh, J. Xu

T16 **Efficient Optimization for Average Precision SVM**
P. Mohapatra, C. Jawahar, M. Kumar

T17 **Ranking via Robust Binary Classification**
H. Yun, P. Raman, S. Vishwanathan

T18 **Tight Bounds for Influence in Diffusion Networks and Application to Bond Percolation and Epidemiology**
R. Lemonnier, K. Scaman, N. Vayatis

T19 **Shaping Social Activity by Incentivizing Users**
M. Farajtabar, N. Du, M. Gomez-Rodriguez, I. Valera, H. Zha, L. Song

T20 **Learning Time-Varying Coverage Functions**
N. Du, Y. Liang, M. Balcan, L. Song

T21 **Online and Stochastic Gradient Methods for Non-decomposable Loss Functions**
P. Kar, H. Narasimhan, P. Jain

T22 **Optimistic Planning in Markov Decision Processes Using a Generative Model**
B. Szörényi, G. Kedenburg, R. Munos

T23 **Conditional Swap Regret and Conditional Correlated Equilibrium**
M. Mohri, S. Yang

T24 Efficient Partial Monitoring with Prior Information
H. Vanchinathan, G. Bartók, A. Krause

T25 Nonparametric Bayesian inference on multivariate exponential families
W. Vega-Brown, M. Doniec, N. Roy

T26 Fast Kernel Learning for Multidimensional Pattern Extrapolation
A. Wilson, E. Gilboa, J. Cunningham, A. Nehorai

T27 Mind the Nuisance: Gaussian Process Classification using Privileged Noise
D. Hernández-lobato, V. Sharmanska, K. Kersting, C. Lampert, N. Quadrianto

T28 Automated Variational Inference for Gaussian Process Models
T. Nguyen, E. Bonilla

T29 Variational Gaussian Process State-Space Models
R. Frigola, Y. Chen, C. Rasmussen

T30 Gaussian Process Volatility Model
Y. Wu, J. Hernández-Lobato, Z. Ghahramani

T31 Bandit Convex Optimization: Towards Tight Bounds
E. Hazan, K. Levy

T32 Stochastic Multi-Armed-Bandit Problem with Non-stationary Rewards
Y. Gur, A. Zeevi, O. Besbes

T33 Extreme bandits
A. Carpentier, M. Valko

T34 Discovering, Learning and Exploiting Relevance
C. Tekin, M. Van Der Schaar

T35 Online combinatorial optimization with stochastic decision sets and adversarial losses
G. Neu, M. Valko

T36 Multilabel Structured Output Learning with Random Spanning Trees of Max-Margin Markov Networks
M. Marchand, H. Su, E. Morvant, J. Rousu, J. Shawe-Taylor

T37 Metric Learning for Temporal Sequence Alignment
R. Lajugie, D. Garreau, F. Bach, S. Arlot

T38 Proximal Quasi-Newton for Computationally Intensive L1-regularized M-estimators
K. Zhong, E. Yen, I. Dhillon, P. Ravikumar

T39 Submodular meets Structured: Finding Diverse Subsets in Exponentially-Large Structured Item Sets
A. Prasad, S. Jegelka, D. Batra

T40 Learning Distributed Representations for Structured Output Prediction
V. Srikumar, C. Manning

T41 Conditional Random Field Autoencoders for Unsupervised Structured Prediction
W. Ammar, C. Dyer, N. Smith

T42 Discriminative Metric Learning by Neighborhood Gerrymandering
S. Trivedi, D. Mcallester, G. Shakhnarovich

T43 Fundamental Limits of Online and Distributed Algorithms for Statistical Learning and Estimation
O. Shamir

T44 Optimal rates for k-NN density and mode estimation
S. Dasgupta, S. Kpotufe

T45 Learning on graphs using Orthonormal Representation is Statistically Consistent
R. Shivanna, C. Bhattacharyya

T46 Optimal prior-dependent neural population coding under shared input noise
A. Grabska-Barwinska, J. Pillow

T47 Spatio-temporal Representations of Uncertainty in Spiking Neural Networks
C. Savin, S. Denève

T48 Optimal Neural Codes for Control and Estimation
A. Susemihl, R. Meir, M. Opper

T49 A Bayesian model for identifying hierarchically organised states in neural population activity
P. Putzky, F. Franzen, G. Bassetto, J. Macke

T50 Clustered factor analysis of multineuronal spike data
L. Buesing, T. Machado, J. Cunningham, L. Paninski

T51 The Large Margin Mechanism for Differentially Private Maximization
K. Chaudhuri, D. Hsu, S. Song

T52 Extremal Mechanisms for Local Differential Privacy
P. Kairouz, S. Oh, P. Viswanath

T53 Reputation-based Worker Filtering in Crowdsourcing
S. Jagabathula, L. Subramanian, A. Venkataraman

T54 Feedback Detection for Live Predictors
S. Wager, N. Chamandy, O. Muralidharan, A. Najmi

T55 DFacTo: Distributed Factorization of Tensors
J. Choi, S. Vishwanathan

T56 Distributed Power-law Graph Computing: Theoretical and Empirical Analysis
C. Xie, L. Yan, W. Li, Z. Zhang

T57 Scalable Non-linear Learning with Adaptive Polynomial Expansions
A. Agarwal, A. Beygelzimer, D. Hsu, J. Langford, M. Telgarsky

T58 Median Selection Subset Aggregation for Parallel Inference
X. Wang, P. Peng, D. Dunson

T59 Asymmetric LSH (ALSH) for Sublinear Time Maximum Inner Product Search (MIPS)
A. Shrivastava, P. Li

T60 Discrete Graph Hashing
W. Liu, C. Mu, S. Kumar, S. Chang

T61 Orbit Regularization
R. Negrinho, A. Martins

T62 Covariance shrinkage for autocorrelated data
D. Bartz, K. Müller

T63 Sparse Space-Time Deconvolution for Calcium Image Analysis
F. Diego Andilla, F. Hamprecht

T64 A Dual Algorithm for Olfactory Computation in the Locust Brain
S. Tootoonian, M. Lengyel

T65 Online Optimization for Max-Norm Regularization
J. Shen, H. Xu, P. Li

T66 Finding a sparse vector in a subspace: Linear sparsity using alternating directions
Q. Qu, J. Sun, J. Wright

T67 Compressive Sensing of Signals from a GMM with Sparse Precision Matrices
J. Yang, X. Liao, M. Chen, L. Carin

T68 Clustering from Labels and Time-Varying Graphs
S. Lim, Y. Chen, H. Xu

T69 On the relations of LFPs & Neural Spike Trains
D. Carlson, J. Schaich Borg, K. Dzirasa, L. Carin

T70 Scalable Inference for Neuronal Connectivity from Calcium Imaging
A. Fletcher, S. Rangan

T71 A Synaptical Story of Persistent Activity with Graded Lifetime in a Neural System
Y. Mi, L. Li, D. Wang, S. Wu

T72 Analog Memories in a Balanced Rate-Based Network of E-I Neurons
D. Festa, G. Hennequin, M. Lengyel

T73 Design Principles of the Hippocampal Cognitive Map
K. Stachenfeld, M. Botvinick, S. Gershman

T74 Feedforward Learning of Mixture Models
M. Lawlor, S. Zucker

T75 Scalable Methods for Nonnegative Matrix Factorizations of Near-separable Tall-and-skinny Matrices
A. Benson, J. Lee, B. Rajwa, D. Gleich

T76 Sparse PCA via Covariance Thresholding
Y. Deshpande, A. Montanari

T77 Low Rank Approximation Lower Bounds in Row-Update Streams
D. Woodruff

T78 Recovery of Coherent Data via Low-Rank Dictionary Pursuit
G. Liu, P. Li

T79 Tight convex relaxations for sparse matrix factorization
E. Richard, G. Obozinski, J. Vert

T80 Robust Tensor Decomposition with Gross Corruption
Q. Gu, H. Gui, J. Han

T81 PEWA: Patch-based Exponentially Weighted Aggregation for image denoising
C. Kervrann

T82 A Multi-World Approach to Question Answering about Real-World Scenes based on Uncertain Input
M. Malinowski, M. Fritz

T83 Quantized Kernel Learning for Feature Matching
D. Qin, X. Chen, M. Guillaumin, L. Gool

T84 Diverse Sequential Subset Selection for Supervised Video Summarization
B. Gong, W. Chao, K. Grauman, F. Sha

T85 Grouping-Based Low-Rank Trajectory Completion and 3D Reconstruction
K. Fragkiadaki, M. Salas, P. Arbelaez, J. Malik

T86 Learning Mixtures of Submodular Functions for Image Collection Summarization
S. Tschiatschek, R. Iyer, H. Wei, J. Bilmes

T87 Deep Learning Face Representation by Joint Identification-Verification
Y. Sun, Y. Chen, X. Wang, X. Tang

T88 Fast Training of Pose Detectors in the Fourier Domain
J. Henriques, P. Martins, R. Caseiro, J. Batista

T89 LSDA: Large Scale Detection through Adaptation
J. Hoffman, S. Guadarrama, E. Tzeng, R. Hu, J. Donahue, R. Girshick, T. Darrell, K. Saenko

T90 Local Decorrelation For Improved Pedestrian Detection
W. Nam, P. Dollar, J. Han

T91 Learning Deep Features for Scene Recognition using Places Database
B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, A. Oliva

T92 Do Convnets Learn Correspondence?
J. Long, N. Zhang, T. Darrell

T93 Semi-supervised Learning with Deep Generative Models
D. Kingma, S. Mohamed, D. Jimenez Rezende, M. Welling

T94 Deep Learning for Real-Time Atari Game Play Using Offline Monte-Carlo Tree Search Planning
X. Guo, S. Singh, H. Lee, R. Lewis, X. Wang

T95 On the Number of Linear Regions of Deep Neural Networks
G. Montufar, R. Pascanu, K. Cho, Y. Bengio

T96 Generative Adversarial Nets
I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio

T97 Deep Symmetry Networks
R. Gens, P. Domingos

T98 Sequence to Sequence Learning with Neural Networks
I. Sutskever, O. Vinyals, Q. Le

T99 Two-Stream Convolutional Networks for Action Recognition in Videos
K. Simonyan, A. Zisserman

T100 Learning to Discover Efficient Mathematical Identities
W. Zaremba, K. Kurach, R. Fergus

T101 Learning Generative Models with Visual Attention
Y. Tang, N. Srivastava, R. Salakhutdinov

T102 A Multiplicative Model for Learning Distributed Text-Based Attribute Representations
R. Kiros, R. Zemel, R. Salakhutdinov

T103 Searching for Higgs Boson Decay Modes with Deep Learning
P. Sadowski, D. Whiteson, P. Baldi

T104 Convolutional Kernel Networks
J. Mairal, P. Koniusz, Z. Harchaoui, C. Schmid

DEMONSTRATIONS

Demonstrations, 7:00 – 11:59 PM

BIDMach: High Performance Machine Learning through Codesign and Rooflining, H. Zhao, B. Jiang, J. Canny

Massively Distributed Machine Learning in the Browser, T. Meeds, R. Hendriks, S. Al Faraby, T. Schoegje, M. Bruntink, M. Welling

Neural Machine Translation, B. van Merriënboer, K. Cho, D. Bahdanau, Y. Bengio

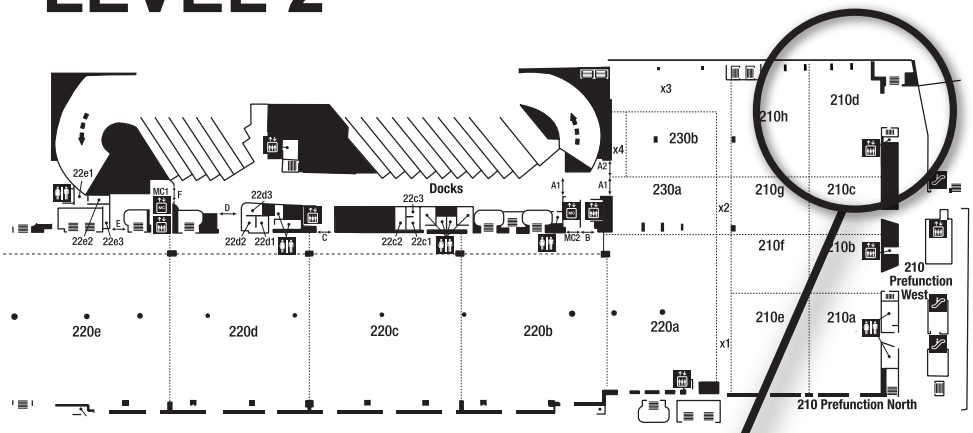
Playing with Convnets, J. Yosinski, H. Lipson

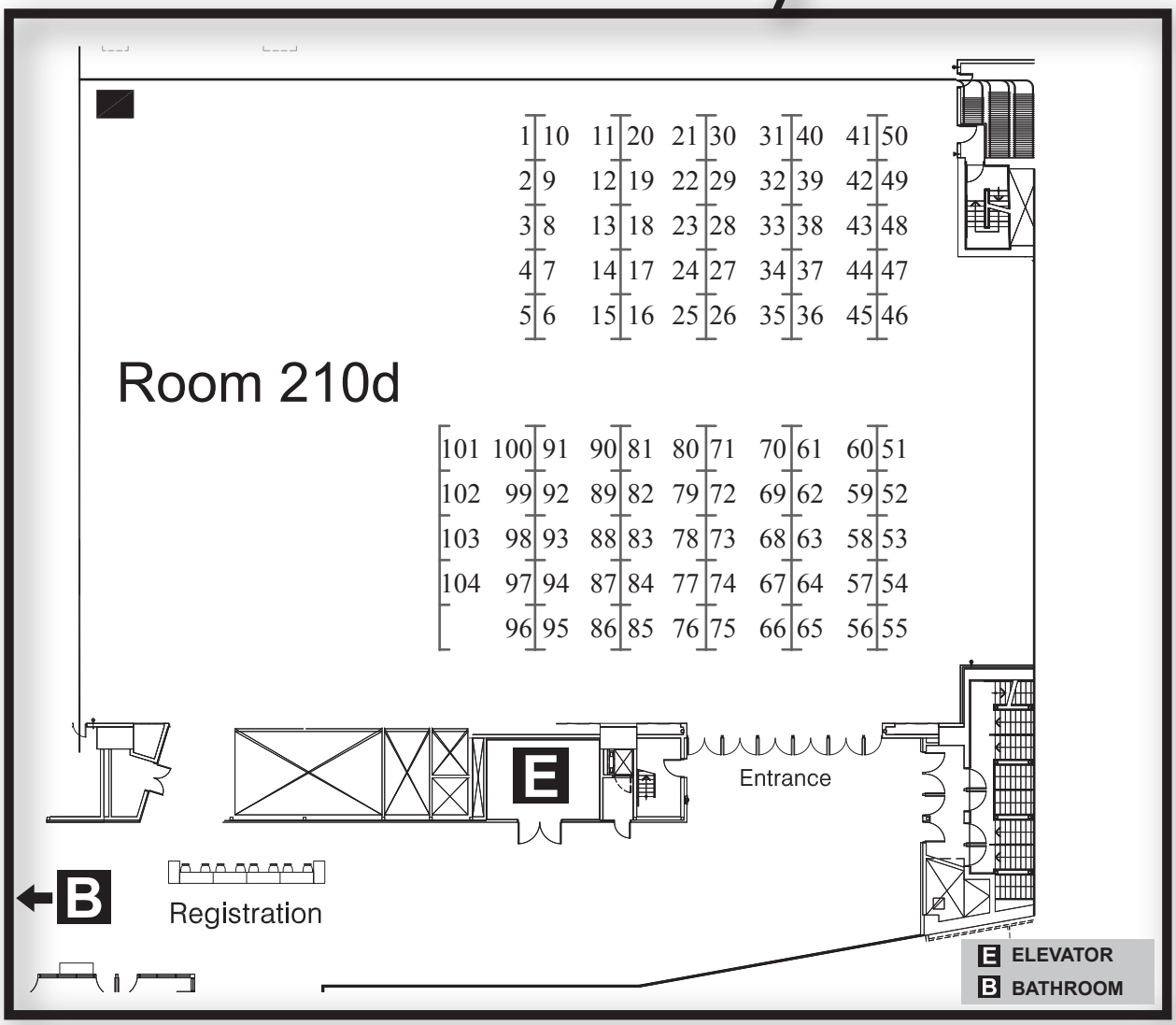
Real-time machine-vision applications on a million-spiking-neuron chip, P. Merolla, R. Alvarez-Icaza, A. Amir, L. Longinotti, S. Bamford, B. Taba, D. Modha

Toronto Deep Learning, R. Kiros, R. Salakhutdinov, N. Srivastava, Y. Tang

TUESDAY POSTER FLOORPLAN

LEVEL 2





T1 Rounding-based Moves for Metric Labeling

M. Pawan Kumar pawan.kumar@ecp.fr
Ecole Centrale Paris

Metric labeling is a special case of energy minimization for pairwise Markov random fields. The energy function consists of arbitrary unary potentials, and pairwise potentials that are proportional to a given metric distance function over the label set. Popular methods for solving metric labeling include (i) move-making algorithms, which iteratively solve a minimum st-cut problem; and (ii) the linear programming (LP) relaxation based approach. In order to convert the fractional solution of the LP relaxation to an integer solution, several randomized rounding procedures have been developed in the literature. We consider a large class of parallel rounding procedures, and design move-making algorithms that closely mimic them. We prove that the multiplicative bound of a move-making algorithm exactly matches the approximation factor of the corresponding rounding procedure for any arbitrary distance function. Our analysis includes all known results for move-making algorithms as special cases.

T2 Provable Submodular Minimization using Wolfe's Algorithm

Deeparnab Chakrabarty dechakr@microsoft.com
Microsoft
Prateek Jain prajain@microsoft.com
Microsoft Research
Pravesh Kothari kothari@cs.utexas.edu
UT Austin

Owing to several applications in large scale learning and vision problems, fast submodular function minimization (SFM) has become a critical problem. Theoretically, unconstrained SFM can be performed in polynomial time (Iwata and Orlin 2009), however these algorithms are not practical. In 1976, Wolfe proposed an algorithm to find the minimum Euclidean norm point in a polytope, and in 1980, Fujishige showed how Wolfe's algorithm can be used for SFM. For general submodular functions, the Fujishige-Wolfe minimum norm algorithm seems to have the best empirical performance. Despite its good practical performance, theoretically very little is known about Wolfe's minimum norm algorithm -- to our knowledge the only result is an exponential time analysis due to Wolfe himself. In this paper we give a maiden convergence analysis of Wolfe's algorithm. We prove that in t iterations, Wolfe's algorithm returns a O(1/t)-approximate solution to the min-norm point. We also prove a robust version of Fujishige's theorem which shows that an O(1/n^2)-approximate solution to the min-norm point problem implies exact submodular minimization. As a corollary, we get the first pseudo-polynomial time guarantee for the Fujishige-Wolfe minimum norm algorithm for submodular function minimization. In particular, we show that the min-norm point algorithm solves SFM in O(n^7F^2)-time, where F is an upper bound on the maximum change a single element can cause in the function value.

T3 Stochastic Gradient Descent, Weighted Sampling, and the Randomized Kaczmarz algorithm

Deanna Needell dneedell@cmc.edu
Claremont McKenna College
Rachel Ward rachel@math.utexas.edu
Univ. of Texas Austin
Nati Srebro nati@ttic.edu
TTI Chicago

We improve a recent gurantee of Bach and Moulines on the linear convergence of SGD for smooth and strongly convex objectives, reducing a quadratic dependence on the strong convexity to a linear dependence. Furthermore, we show how reweighting the sampling distribution (i.e. importance sampling) is necessary in order to further improve convergence, and obtain a linear dependence on average smoothness, dominating previous results, and more broadly discus how importance sampling for SGD can improve convergence also in other scenarios. Our results are based on a connection we make between SGD and the randomized Kaczmarz algorithm, which allows us to transfer ideas between the separate bodies of literature studying each of the two methods.

T4 A Differential Equation for Modeling Nesterov's Accelerated Gradient Method: Theory and Insights

Weijie Su wjsu@stanford.edu
Stephen Boyd boyd@stanford.edu
Emmanuel Candes candes@stanford.edu
Stanford University

We derive a second-order ordinary differential equation (ODE), which is the limit of Nesterov's accelerated gradient method. This ODE exhibits approximate equivalence to Nesterov's scheme and thus can serve as a tool for analysis. We show that the continuous time ODE allows for a better understanding of Nesterov's scheme. As a byproduct, we obtain a family of schemes with similar convergence rates. The ODE interpretation also suggests restarting Nesterov's scheme leading to an algorithm, which can be rigorously proven to converge at a linear rate whenever the objective is strongly convex.

T5 Large-scale L-BFGS using MapReduce

Weizhu Chen wzchen@microsoft.com
Zhenghao Wang zhwang@microsoft.com
Jingren Zhou jrzhou@microsoft.com
Microsoft

L-BFGS has been applied as an effective parameter estimation method for various machine learning algorithms since 1980s. With an increasing demand to deal with massive instances and variables, it is important to scale up and parallelize L-BFGS effectively in a distributed system. In this paper, we study the problem of parallelizing the L-BFGS algorithm in large clusters of tens of thousands of shared-nothing commodity machines. First, we show that a naive implementation of L-BFGS using Map-Reduce requires either a significant amount of memory or a large number of map-reduce steps with negative performance impact. Second, we propose a new L-BFGS algorithm, called Vector-free L-BFGS, which avoids the expensive dot product operations in the two loop recursion and greatly improves computation efficiency with a great degree of parallelism. The algorithm scales very well and enables a variety of machine learning algorithms to handle a massive number of variables over large datasets. We prove the mathematical equivalence of the new Vector-free L-BFGS and demonstrate its excellent performance and scalability using real-world machine learning problems with billions of variables in production clusters.

T6 An Accelerated Proximal Coordinate Gradient Method

Qihang Lin qihangl@andrew.cmu.edu
University of Iowa
Zhaosong Lu zhaosong@sfu.ca
Simon Frasor University
Lin Xiao lin.xiao@microsoft.com
MSR Redmond

We develop an accelerated randomized proximal coordinate gradient (APCG) method, for solving a broad class of composite convex optimization problems. In particular, our method achieves faster linear convergence rates for minimizing strongly convex functions than existing randomized proximal coordinate gradient methods. We show how to apply the APCG method to solve the dual of the regularized empirical risk minimization (ERM) problem, and devise efficient implementations that can avoid full-dimensional vector operations. For ill-conditioned ERM problems, our method obtains improved convergence rates than the state-of-the-art stochastic dual coordinate ascent (SDCA) method.

T7 Predictive Entropy Search for Efficient Global Optimization of Black-box Functions

José Miguel Hernández-Lobato jmh@seas.harvard.edu
Harvard University
Matthew Hoffman mwh30@cam.ac.uk
Zoubin Ghahramani zoubin@eng.cam.ac.uk
University of Cambridge

We propose a novel information-theoretic approach for Bayesian optimization called Predictive Entropy Search (PES). At each iteration, PES selects the next evaluation point that maximizes the expected information gained with respect to the global maximum. PES codifies this intractable acquisition function in terms of the expected reduction in the differential entropy of the predictive distribution. This reformulation allows PES to obtain approximations that are both more accurate and efficient than other alternatives such as Entropy Search (ES). Furthermore, PES can easily perform a fully Bayesian treatment of the model hyperparameters while ES cannot. We evaluate PES in both synthetic and real-world applications, including optimization problems in machine learning, finance, biotechnology, and robotics. We show that the increased accuracy of PES leads to significant gains in optimization performance.

T8 Inference by Learning: Speeding-up Graphical Model Optimization via a Coarse-to-Fine Cascade of Pruning Classifiers

Bruno Conejo bconejo@caltech.edu
Caltech / Ecole des Ponts ParisTech
Nikos Komodakis komod@csd.uoc.gr
Ecole des Ponts ParisTech
Sebastien Leprince leprincs@caltech.edu
Jean Philippe Avouac avouac@gps.caltech.edu
Caltech

We propose a general and versatile framework that significantly speeds-up graphical model optimization while maintaining an excellent solution accuracy. The proposed approach, refereed as Inference by Learning or IbyL, relies on a multi-scale pruning scheme that progressively reduces the solution space by use of a coarse-to-fine cascade of learnt classifiers. We thoroughly experiment with classic computer vision related MRF problems, where our novel framework constantly yields a significant time speed-up (with respect to the most efficient inference methods) and obtains a more accurate solution than directly optimizing the MRF. We make our code available on-line.

T9 Two-Layer Feature Reduction for Sparse-Group Lasso via Decomposition of Convex Sets

Jie Wang jie.wang.ustc@asu.edu
Jieping Ye jieping.ye@asu.edu
Arizona State University

Sparse-Group Lasso (SGL) has been shown to be a powerful regression technique for simultaneously discovering group and within-group sparse patterns by using a combination of the l1 and l2 norms. However, in large-scale applications, the complexity of the regularizers entails great computational challenges. In this paper, we propose a novel two-layer feature reduction method (TLFre) for SGL via a decomposition of its dual feasible set. The two-layer reduction is able to quickly identify the inactive groups and the inactive features, respectively, which are guaranteed to be absent from the sparse representation and can be removed from the optimization. Existing feature reduction methods are only applicable for sparse models with one sparsity-inducing regularizer. To our best knowledge, TLFre is the first one that is capable of dealing with multiple sparsity-inducing regularizers. Moreover, TLFre has a very low computational cost and can be integrated with any existing solvers. Experiments on both synthetic and real data sets show that TLFre improves the efficiency of SGL by orders of magnitude.

T10 The Noisy Power Method: A Meta Algorithm with Applications

Moritz Hardt m@mrtz.org
IBM Research Almaden
Eric Price ecprice@mit.edu
IBM

We provide a new robust convergence analysis of the well-known power method for computing the dominant singular vectors of a matrix that we call noisy power method. Our result characterizes the convergence behavior of the algorithm when a large amount noise is introduced after each matrix-vector multiplication. The noisy power method can be seen as a meta-algorithm that has recently found a number of important applications in a broad range of machine learning problems including alternating minimization for matrix completion, streaming principal component analysis (PCA), and privacy-preserving spectral analysis. Our general analysis subsumes several existing ad-hoc convergence bounds and resolves a number of open problems in multiple applications. A recent work of Mitliagkas et al.~(NIPS 2013) gives a space-efficient algorithm for PCA in a streaming model where samples are drawn from a spiked covariance model. We give a simpler and more general analysis that applies to arbitrary distributions. Moreover, even in the spiked covariance model our result gives quantitative improvements in a natural parameter regime. As a second application, we provide an algorithm for differentially private principal component analysis that runs in nearly linear time in the input sparsity and achieves nearly tight worst-case error bounds. Complementing our worst-case bounds, we show that the error dependence of our algorithm on the matrix dimension can be replaced by an essentially tight dependence on the coherence of the matrix. This result resolves the main problem left open by Hardt and Roth (STOC 2013) and leads to strong average-case improvements over the optimal worst-case bound.

T11 A Latent Source Model for Online Collaborative Filtering

Guy Bresler gbresler@mit.edu
George Chen georgehc@mit.edu
Devavrat Shah devavrat@mit.edu
Massachusetts Institute of Technology

Despite the prevalence of collaborative filtering in recommendation systems, there has been little theoretical development on why and how well it works, especially in the ``online'' setting, where items are recommended to users over time. We address this theoretical gap by introducing a model for online recommendation systems, cast item recommendation under the model as a learning problem, and analyze the performance of a cosine-similarity collaborative filtering method. In our model, each of n users either likes or dislikes each of m items. We assume there to be k types of users, and all the users of a given type share a common string of probabilities determining the chance of liking each item. At each time step, we recommend an item to each user, where a key distinction from related bandit literature is that once a user consumes an item (e.g., watches a movie), then that item cannot be recommended to the same user again. The goal is to maximize the number of likable items recommended to users over time. Our main result establishes that after nearly $\log(km)$ initial learning time steps, a simple collaborative filtering algorithm achieves essentially optimal performance without knowing k. The algorithm has an exploitation step that uses cosine similarity and two types of exploration steps, one to explore the space of items (standard in the literature) and the other to explore similarity between users (novel to this work).

T12 Probabilistic low-rank matrix completion on finite alphabets

Jean Lafond jean.lafond@telecom-paristech.fr
Joseph Salmon joseph.salmon@telecom-paristech.fr
Télécom ParisTech
Olga Klopp olga.klopp@math.cnrs.fr
Université Paris Ouest
Eric Moulines eric.moulines@telecom-paristech.fr
Telecom ParisTech

The task of reconstructing a matrix given a sample of observed entries is known as the **matrix completion problem**. Such a consideration arises in a wide variety of problems, including recommender systems, collaborative filtering, dimensionality reduction, image processing, quantum physics or multi-class classification to name a few. Most works have focused on recovering an unknown real-valued low-rank matrix from randomly sub-sampling its entries. Here, we investigate the case where the observations take a finite numbers of values, corresponding for examples to ratings in recommender systems or labels in multi-class classification. We also consider a general sampling scheme (non-necessarily uniform) over the matrix entries. The performance of a nuclear-norm penalized estimator is analyzed theoretically. More precisely, we derive bounds for the Kullback-Leibler divergence between the true and estimated distributions. In practice, we have also proposed an efficient algorithm based on lifted coordinate gradient descent in order to tackle potentially high dimensional settings.

T13 Controlling privacy in recommender systems

Yu Xin yuxin@mit.edu
Tommi Jaakkola tommi@csail.mit.edu
Massachusetts Institute of Technology

Recommender systems involve an inherent trade-off between accuracy of recommendations and the extent to which users are willing to release information about their preferences. In this paper, we explore a two-tiered notion of privacy where there is a small set of ``public'' users who are willing to share their preferences openly, and a large set of ``private'' users who require privacy guarantees. We show theoretically and demonstrate empirically that a moderate number of public users with no access to private user information already suffices for reasonable accuracy. Moreover, we introduce a new privacy concept for gleaning relational information from private users while maintaining a first order deniability. We demonstrate gains from controlled access to private user preferences.

T14 Content-based recommendations with Poisson factorization

Prem Gopalan pgopalan@cs.princeton.edu
Princeton University
Laurent Charlin lcharlin@cs.columbia.edu
David Blei blei@cs.princeton.edu
Columbia University

We develop collaborative topic Poisson factorization (CTPF), a generative model of articles and reader preferences. CTPF can be used to build recommender systems by learning from reader histories and content to recommend personalized articles of interest. In detail, CTPF models both reader behavior and article texts with Poisson distributions, connecting the latent topics that represent the texts with the latent preferences that represent the readers. This provides better recommendations than competing methods and gives an interpretable latent space for understanding patterns of readership. Further, we exploit stochastic variational inference to model massive real-world datasets. For example, we can fit CPTF to the full arXiv usage dataset, which contains over 43 million ratings and 42 million word counts, within a day. We demonstrate empirically that our model outperforms several baselines, including the previous state-of-the-art approach.

T15 Minimax-optimal Inference from Partial Rankings

Bruce Hajek b-hajek@illinois.edu
Sewoong Oh swoh@illinois.edu
Jiaming Xu jxu18@illinois.edu
University of Illinois at Urbana Champaign

This paper studies the problem of rank aggregation under the Plackett-Luce model. The goal is to infer a global ranking and related scores of the items, based on partial rankings provided by multiple users over multiple subsets of items. A question of particular interest is how to optimally assign items to users for ranking and how many item assignments are needed to achieve a target estimation error. Without any assumptions on how the items are assigned to users, we derive an oracle lower bound and the Cram\'er-Rao lower bound of the estimation error. We prove an upper bound on the estimation error achieved by the maximum likelihood estimator, and show that both the upper bound and the Cram\'er-Rao lower bound inversely depend on the spectral gap of the Laplacian of an appropriately defined comparison graph. Since random comparison graphs are known to have large spectral gaps, this suggests the use of random assignments when we have the control. Precisely, the matching oracle lower bound and the upper bound on the estimation error imply that the maximum likelihood estimator together with a random assignment is minimax-optimal up to a logarithmic factor. We further analyze a popular rank-breaking scheme that decompose partial rankings into pairwise comparisons. We show that even if one applies the mismatched maximum likelihood estimator that assumes independence (on pairwise comparisons that are now dependent due to rank-breaking), minimax optimal performance is still achieved up to a logarithmic factor.

T16 Efficient Optimization for Average Precision SVM

Pritish Mohapatra pritish.mohapatra@research.iiit.ac.in
C.V. Jawahar jawahar@iiit.ac.in
International Institute of Information Technology
M. Pawan Kumar pawan.kumar@ecp.fr
Ecole Centrale Paris

The accuracy of information retrieval systems is often measured using average precision (AP). Given a set of positive (relevant) and negative (non-relevant) samples, the parameters of a retrieval system can be estimated using the AP-SVM framework, which minimizes a regularized convex upper bound on the empirical AP loss. However, the high computational complexity of loss-augmented inference, which is required for learning an AP-SVM, prohibits its use with large training datasets. To alleviate this deficiency, we propose three complementary approaches. The first approach guarantees an asymptotic decrease in the computational complexity of loss-augmented inference by exploiting the problem structure. The second approach takes advantage of the fact that we do not require a full ranking during loss-augmented inference. This helps us to avoid the expensive step of sorting the negative samples according to their individual scores. The third approach approximates the AP loss over all samples by the AP loss over difficult samples (for example, those that are incorrectly classified by a binary SVM), while ensuring the correct classification of the remaining samples. Using the PASCAL VOC action classification and object detection datasets, we show that our approaches provide significant speed-ups during training without degrading the test accuracy of AP-SVM.

T17 Ranking via Robust Binary Classification

Hyokun Yun yungilbert@gmail.com
Amazon
Parameswaran Raman praman1@ucsc.edu
University of California - Santa Cruz
S. Vishwanathan vishy@stat.purdue.edu
Purdue University

We propose RoBiRank, a ranking algorithm that is motivated by observing a close connection between evaluation metrics for learning to rank and loss functions for robust classification. The algorithm shows a very competitive performance on standard benchmark datasets against other representative algorithms in the literature. Further, in large scale problems where explicit feature vectors and scores are not given, our algorithm can be efficiently parallelized across a large number of machines; for a task that requires 386,133 x 49,824,519 pairwise interactions between items to be ranked, our algorithm finds solutions that are of dramatically higher quality than that can be found by a state-of-the-art competitor algorithm, given the same amount of wall-clock time for computation.

T18 Tight Bounds for Influence in Diffusion Networks and Application to Bond Percolation and Epidemiology

Remi Lemonnier remi.lemonnier@ens-cachan.fr
Kevin Scaman scaman@cmla.ens-cachan.fr
Nicolas Vayatis nicolas.vayatis@cmla.ens-cachan.fr
Ecole Normale Supérieure de Cachan

In this paper, we derive theoretical bounds for the long-term influence of a node in an Independent Cascade Model (ICM). We relate these bounds to the spectral radius of a particular matrix and show that the behavior is sub-critical when this spectral radius is lower than 1. More specifically, we point out that, in general networks, the sub-critical regime behaves in O(sqrt(n)) where n is the size of the network, and that this upper bound is met for star-shaped networks. We apply our results to epidemiology and percolation on arbitrary networks, and derive a bound for the critical value beyond which a giant connected component arises. Finally, we show empirically the tightness of our bounds for a large family of networks.

T19 Shaping Social Activity by Incentivizing Users

Mehrdad Farajtabar mehrdad@gatech.edu
Nan Du dunan@gatech.edu
Hongyuan Zha zha@cc.gatech.edu
Le Song lsong@cc.gatech.edu
Georgia Institute of Technolog
Manuel Gomez-Rodriguez manuelgr@**T**bingen.mpg.de
MPI for Software Systems
Isabel Valera ivalera@tsc.uc3m.es
UC3M

Events in an online social network can be categorized roughly into endogenous events, where users just respond to the actions of their neighbors within the network, or exogenous events, where users take actions due to drives external to the network. How much external drive should be provided to each user, such that the network activity can be steered towards a target state? In this paper, we model social events using multivariate Hawkes processes, which can capture both endogenous and exogenous event intensities, and derive a time dependent linear relation between the intensity of exogenous events and the overall network activity. Exploiting this connection, we develop a convex optimization framework for determining the required level of external drive in order for the network to reach a desired activity level. We experimented with event data gathered from Twitter, and show that our method can steer the activity of the network more accurately than alternatives.

T20 Learning Time-Varying Coverage Functions

Nan Du dunan@gatech.edu
Maria-Florina Balcan ninamf@cc.gatech.edu
Le Song lsong@cc.gatech.edu
Georgia Tech
Yingyu Liang yingyul@cs.princeton.edu
Princeton University

Coverage functions are an important class of discrete functions that capture laws of diminishing returns. In this paper, we propose a new problem of learning time-varying coverage functions which arise naturally from applications in social network analysis, machine learning, and algorithmic game theory. We develop a novel parametrization of the time-varying coverage function by illustrating the connections with counting processes. We present an efficient algorithm to learn the parameters by maximum likelihood estimation, and provide a rigorous theoretic analysis of its sample complexity. Empirical experiments from information diffusion in social network analysis demonstrate that with few assumptions about the underlying diffusion process, our method performs significantly better than existing approaches on both synthetic and real world data.

T21 Online and Stochastic Gradient Methods for Non-decomposable Loss Functions

Puru Kar purushot@cse.iitk.ac.in
Microsoft Research India
Harikrishna Narasimhan nhari88@gmail.com
Indian Institute of Science
Prateek Jain prajain@microsoft.com
Microsoft Research

Modern applications in sensitive domains such as biometrics and medicine frequently require the use of non-decomposable loss functions such as precision@k, F-measure etc. Compared to point loss functions such as hinge-loss, these offer much more fine grained control over prediction, but at the same time present novel challenges in terms of algorithm design and analysis. In this work we initiate a study of online learning techniques for such non-decomposable loss functions with an aim to enable incremental learning as well as design scalable solvers for batch problems. To this end, we propose an online learning framework for such loss functions. Our model enjoys several nice properties, chief amongst them being the existence of efficient online learning algorithms with sublinear regret and online to batch conversion bounds. Our model is a provable extension of existing online learning models for point loss functions. We instantiate two popular losses, Prec @k and pAUC, in our model and prove sublinear regret bounds for both of them. Our proofs require a novel structural lemma over ranked lists which may be of independent interest. We then develop scalable stochastic gradient descent solvers for non-decomposable loss functions. We show that for a large family of loss functions satisfying a certain uniform convergence property (that includes Prec @k, pAUC, and F-measure), our methods provably converge to the empirical risk minimizer. Such uniform convergence results were not known for these losses and we establish these using novel proof techniques. We then use extensive experimentation on real life and benchmark datasets to establish that our method can be orders of magnitude faster than a recently proposed cutting plane method.

T22 Optimistic Planning in Markov Decision Processes Using a Generative Model

Balázs Szörényi szorenyi@inf.u-szeged.hu
Gunnar Kedenburg gunnar@kede.de
Remi Munos remi.munos@inria.fr
INRIA / MSR

We consider the problem of online planning in a Markov decision process with discounted rewards for any given initial state. We consider the PAC sample complexity problem of computing, with probability $1-\delta$, an ϵ-optimal action using the smallest possible number of calls to the generative model (which provides reward and next-state samples). We design an algorithm, called StOP (for Stochastic-Optimistic Planning), based on the "optimism in the face of uncertainty" principle. StOP can be used in the general setting, requires only a generative model, and enjoys a complexity bound that only depends on the local structure of the MDP.

T23 Conditional Swap Regret and Conditional Correlated Equilibrium

Mehryar Mohri mohri@google.com
Courant Institute, NYU & Google
Scott Yang yangs@cims.nyu.edu
New York University

We introduce a natural extension of the notion of swap regret, conditional swap regret, that allows for action modifications conditioned on the player's action history. We prove a series of new results for conditional swap regret minimization. We present algorithms for minimizing conditional swap regret with bounded conditioning history. We further extend these results to the case where conditional swaps are considered only for a subset of actions. We also define a new notion of equilibrium, conditional correlated equilibrium, that is tightly connected to the notion of conditional swap regret: when all players follow conditional swap regret minimization strategies, then the empirical distribution approaches this equilibrium. Finally, we extend our results to the multi-armed bandit scenario.

T24 Efficient Partial Monitoring with Prior Information

Hasta Vanchinathan hastagiri@inf.ethz.ch
Andreas Krause krausea@ethz.ch
ETH Zurich
Gábor Bartók bartok@google.com
Google Zürich

Partial monitoring is a general model for online learning with limited feedback: a learner chooses actions in a sequential manner while an opponent chooses outcomes. In every round, the learner suffers some loss and receives some feedback based on the action and the outcome. The goal of the learner is to minimize her cumulative loss. Applications range from dynamic pricing to label-efficient prediction to dueling bandits. In this paper, we assume that we are given some prior information about the distribution based on which the opponent generates the outcomes. We propose BPM, a family of new efficient algorithms whose core is to track the outcome distribution with an ellipsoid centered around the estimated distribution. We show that our algorithm provably enjoys near-optimal regret rate for locally observable partial-monitoring problems against stochastic opponents. As demonstrated with experiments on synthetic as well as real-world data, the algorithm outperforms previous approaches, even for very uninformed priors, with an order of magnitude smaller regret and lower running time.

T25 Nonparametric Bayesian inference on multivariate exponential families

William Vega-Brown wrvb@mit.edu
Marek Doniec doniec@mit.edu
Nicholas Roy nickroy@mit.edu
Massachusetts Institute of Technology

We develop a model by choosing the maximum entropy distribution from the set of models satisfying certain smoothness and independence criteria; we show that inference on this model generalizes local kernel estimation to the context of Bayesian inference on stochastic processes. Our model enables Bayesian inference in contexts when standard techniques like Gaussian process inference are too expensive to apply. Exact inference on our model is possible for any likelihood function from the exponential family. Inference is then highly efficient, requiring only O(log N) time and O(N) space at run time. We demonstrate our algorithm on several problems and show quantifiable improvement in both speed and performance relative to models based on the Gaussian process.

T26 Fast Kernel Learning for Multidimensional Pattern Extrapolation

Andrew Wilson agw38@cam.ac.uk
University of Cambridge
Elad Gilboa egilboa@gmail.com
Technion
John Cunningham jpc74@cam.ac.uk
Columbia University
Arye Nehorai nehorai@ese.wustl.edu
Washington University in St. Louis

The ability to automatically discover patterns and perform extrapolation is an essential quality of intelligent systems. Kernel methods, such as Gaussian processes, have great potential for pattern extrapolation, since the kernel flexibly and interpretably controls the generalisation properties of these methods. However, automatically extrapolating large scale multidimensional patterns is in general difficult, and developing Gaussian process models for this purpose involves several challenges. A vast majority of kernels, and kernel learning methods, currently only succeed in smoothing and interpolation. This difficulty is compounded by the fact that Gaussian processes are typically only tractable for small datasets, and scaling an expressive kernel learning approach poses different challenges than scaling a standard Gaussian process model. One faces additional computational constraints, and the need to retain significant model structure for expressing the rich information available in a large dataset. In this paper, we propose a Gaussian process approach for large scale multidimensional pattern extrapolation. We recover sophisticated out of class kernels, perform texture extrapolation, inpainting, and video extrapolation, and long range forecasting of land surface temperatures, all on large multidimensional datasets, including a problem with 383,400 training points. The proposed method significantly outperforms alternative scalable and flexible Gaussian process methods, in speed and accuracy. Moreover, we show that a distinct combination of expressive kernels, a fully

non-parametric representation, and scalable inference which exploits existing model structure, are critical for large scale multidimensional pattern extrapolation.

T27 Mind the Nuisance: Gaussian Process Classification using Privileged Noise

Daniel Hernández-lobato danielhernandezlobato@gmail.com
Universidad Autonoma de Madrid
Viktoriia Sharmanska sharmanska.v@gmail.com
Christoph Lampert chl@ist.ac.at
IST Austria
Kristian Kersting kristian.kersting@iais.fraunhofer.de
University of Bonn and Fraunhofer IAIS
Novi Quadrianto novi.quadrianto@gmail.com
University of Sussex

The learning with privileged information setting has recently attracted a lot of attention within the machine learning community, as it allows the integration of additional knowledge into the training process of a classifier, even when this comes in the form of a data modality that is not available at test time. Here, we show that privileged information can naturally be treated as noise in the latent function of a Gaussian process classifier (GPC). That is, in contrast to the standard GPC setting, the latent function is not just a nuisance but a feature: it becomes a natural measure of confidence about the training data by modulating the slope of the GPC probit likelihood function. Extensive experiments on public datasets show that the proposed GPC method using privileged noise, called GPC+, improves over a standard GPC without privileged knowledge, and also over the current state-of-the-art SVM-based method, SVM+. Moreover, we show that advanced neural networks and deep learning methods can be compressed as privileged information.

T28 Automated Variational Inference for Gaussian Process Models

Trung Nguyen vantrung.nguyen@nicta.com.au
ANU; NICTA
Edwin Bonilla edwinbonilla@gmail.com
The University of New South Wales

We develop an automated variational method for approximate inference in Gaussian process (GP) models whose posteriors are often intractable. Using a mixture of Gaussians as the variational distribution, we show that (i) the variational objective and its gradients can be approximated efficiently via sampling from univariate Gaussian distributions and (ii) the gradients of the GP hyperparameters can be obtained analytically regardless of the model likelihood. We further propose two instances of the variational distribution whose covariance matrices can be parametrized linearly in the number of observations. These results allow gradient-based optimization to be done efficiently in a black-box manner. Our approach is thoroughly verified on 5 models using 6 benchmark datasets, performing as well as the exact or hard-coded implementations while running orders of magnitude faster than the alternative MCMC sampling approaches. Our method can be a valuable tool for practitioners and researchers to investigate new models with minimal effort in deriving model-specific inference algorithms.

T29 Variational Gaussian Process State-Space Models

Roger Frigola rf342@cam.ac.uk
Yutian Chen yutian.chen@eng.cam.ac.uk
Carl Rasmussen cer54@cam.ac.uk
University of Cambridge

State-space models have been successfully used for more than fifty years in different areas of science and engineering. We present a procedure for efficient variational Bayesian learning of nonlinear state-space models based on sparse Gaussian processes. The result of learning is a tractable posterior over nonlinear dynamical systems. In comparison to conventional parametric models, we offer the possibility to straightforwardly trade off model capacity and computational cost whilst avoiding overfitting. Our main algorithm uses a hybrid inference approach combining variational Bayes and sequential Monte Carlo. We also present stochastic variational inference and online learning approaches for fast learning with long time series.

T30 Gaussian Process Volatility Model

Yue Wu wu5@post.harvard.edu
Cambridge
José Miguel Hernández-Lobato jmh@seas.harvard.edu
Harvard University
Zoubin Ghahramani zoubin@eng.cam.ac.uk
University of Cambridge

The prediction of time-changing variances is an important task in the modeling of financial data. Standard econometric models are often limited as they assume rigid functional relationships for the evolution of the variance. Moreover, functional parameters are usually learned by maximum likelihood, which can lead to overfitting. To address these problems we introduce GP-Vol, a novel non-parametric model for time-changing variances based on Gaussian Processes. This new model can capture highly flexible functional relationships for the variances. Furthermore, we introduce a new online algorithm for fast inference in GP-Vol. This method is much faster than current offline inference procedures and it avoids overfitting problems by following a fully Bayesian approach. Experiments with financial data show that GP-Vol performs significantly better than current standard alternatives.

T31 Bandit Convex Optimization: Towards Tight Bounds

Elad Hazan ehazan@ie.technion.ac.il
Kfir Levy kfiryehud@gmail.com
Technion

Bandit Convex Optimization (BCO) is a fundamental framework for decision making under uncertainty, which generalizes many problems from the realm of online and statistical learning. While the special case of linear cost functions is well understood, a gap on the attainable regret for BCO with nonlinear losses remains an important open question. In this paper we take a step towards understanding the best attainable regret bounds for BCO: we give an efficient and near-optimal regret algorithm for BCO with strongly-convex and smooth loss functions. In contrast to previous works on BCO that use time invariant exploration schemes, our method employs an exploration scheme that shrinks with time.

T32 Stochastic Multi-Armed-Bandit Problem with Non-stationary Rewards

Yonatan Gur ygur@stanford.edu
Stanford Graduate School of Business
Assaf Zeevi assaf@gsb.columbia.edu
Omar Besbes ob2105@columbia.edu
Columbia University

In a multi-armed bandit (MAB) problem a gambler needs to choose at each round of play one of K arms, each characterized by an unknown reward distribution. Reward realizations are only observed when an arm is selected, and the gambler's objective is to maximize his cumulative expected earnings over some given horizon of play T. To do this, the gambler needs to acquire information about arms (exploration) while simultaneously optimizing immediate rewards (exploitation); the price paid due to this trade off is often referred to as the regret, and the main question is how small can this price be as a function of the horizon length T. This problem has been studied extensively when the reward distributions do not change over time; an assumption that supports a sharp characterization of the regret, yet is often violated in practical settings. In this paper, we focus on a MAB formulation which allows for a broad range of temporal uncertainties in the rewards, while still maintaining mathematical tractability. We fully characterize the (regret) complexity of this class of MAB problems by establishing a direct link between the extent of allowable reward "variation" and the minimal achievable regret, and by establishing a connection between the adversarial and the stochastic MAB frameworks.

T33 Extreme bandits

Alexandra Carpentier a.carpentier@statslab.cam.ac.uk
StatsLab Cambridge
Misho Valko michal.valko@inria.fr
INRIA Lille - Nord Europe

In many areas of medicine, security, and life sciences, we want to allocate limited resources to different sources in order to detect extreme values. In this paper, we study an efficient way to allocate these resources sequentially under limited feedback. While sequential design of experiments is well studied in bandit theory, the most commonly optimized property is the regret with respect to the maximum mean reward. However, in other problems such as network intrusion detection, we are interested in detecting the most extreme value output by the sources. Therefore, in our work we study extreme regret which measures the efficiency of an algorithm compared to the oracle policy selecting the source with the heaviest tail. We propose the ExtremeHunter algorithm, provide its analysis, and evaluate it empirically on synthetic and real-world experiments.

T34 Discovering, Learning and Exploiting Relevance

Cem Tekin cem.tekin.dr@gmail.com
Mihaela Van Der Schaar mihaela@ee.ucla.edu
University of California, Los Angeles

In this paper we consider the problem of learning online what is the information to consider when making sequential decisions. We formalize this as a contextual multi-armed bandit problem where a high dimensional (D-dimensional) context vector arrives to a learner which needs to select an action to maximize its expected reward at each time step. Each dimension of the context vector is called a type. We assume that there exists an unknown relation between actions and types, called the relevance relation, such that the reward of an action only depends on the contexts of the relevant types. When the relation is a function, i.e., the reward of an action only depends on the context of a single type, and the expected reward of an action is Lipschitz continuous in the context of its relevant type, we propose an algorithm that achieves $O\tilde{}(T^{\gamma})$ regret with a high probability, where $\gamma=2/(1+2\sqrt{})$. Our algorithm achieves this by learning the unknown relevance relation, whereas prior contextual bandit algorithms that do not exploit the existence of a relevance relation will have $O\tilde{}(T^{(D+1)/(D+2)})$ regret. Our algorithm alternates between exploring and exploiting, it does not require reward observations in exploitations, and it guarantees with a high probability that actions with suboptimality greater than ϵ are never selected in exploitations. Our proposed method can be applied to a variety of learning applications including medical diagnosis, recommender systems, popularity prediction from social networks, network security etc., where at each instance of time vast amounts of different types of information are available to the decision maker, but the effect of an action depends only on a single type.

T35 Online combinatorial optimization with stochastic decision sets and adversarial losses

Gergo Neu neu.gergely@gmail.com
Misho Valko michal.valko@inria.fr
INRIA Lille - Nord Europe

Most work on sequential learning assumes a fixed set of actions that are available all the time. However, in practice, actions can consist of picking subsets of readings from sensors that may break from time to time, road segments that can be blocked or goods that are out of stock. In this paper we study learning algorithms that are able to deal with stochastic availability of such unreliable composite actions. We propose and analyze algorithms based on the Follow-The-Perturbed-Leader prediction method for several learning settings differing in the feedback provided to the learner. Our algorithms rely on a novel loss estimation technique that we call Counting Asleep Times. We deliver regret bounds for our algorithms for the previously studied full information and (semi-)bandit settings, as well as a natural middle point between the two that we call the restricted information setting. A special consequence of our results is a significant improvement of the best known performance guarantees achieved by an efficient algorithm for the sleeping bandit problem with stochastic availability. Finally, we evaluate our algorithms empirically and show their improvement over the known approaches.

T36 Multilabel Structured Output Learning with Random Spanning Trees of Max-Margin Markov Networks

Mario Marchand mario.marchand@ift.ulaval.ca
Université Laval
Hongyu Su hongyu.su@aalto.fi
Juho Rousu juho.rousu@aalto.fi
Aalto University
Emilie Morvant emilie.morvant@univ-st-etienne.fr
LaHC, University of Saint-Etienne
John Shawe-Taylor jst@cs.ucl.ac.uk
UCL

We show that the usual score function for conditional Markov networks can be written as the expectation over the scores of their spanning trees. We also show that a small random sample of these output trees can attain a significant fraction of the margin obtained by the complete graph and we provide conditions under which we can perform tractable inference. The experimental results confirm that practical learning is scalable to realistic datasets using this approach.

T37 Metric Learning for Temporal Sequence Alignment

Rémi Lajugie remi.lajugie@ens.fr
Damien Garreau damien.garreau@ens.fr
Francis Bach francis.bach@mines.org
INRIA & ENS
Sylvain Arlot sylvain.arlot@ens.fr
CNRS

In this paper, we propose to learn a Mahalanobis distance to perform alignment of multivariate time series. The learning examples for this task are time series for which the true alignment is known. We cast the alignment problem as a structured prediction task, and propose realistic losses between alignments for which the optimization is tractable. We provide experiments on real data in the audio-to-audio context, where we show that the learning of a similarity measure leads to improvements in the performance of the alignment task. We also propose to use this metric learning framework to perform feature selection and, from basic audio features, build a combination of these with better alignment performance.

T38 Proximal Quasi-Newton for Computationally Intensive L1-regularized M-estimators

Kai Zhong kaizhong89@gmail.com
En-Hsu Yen a061105@gmail.com
Inderjit Dhillon inderjit@cs.utexas.edu
Pradeep Ravikumar pradeepr@cs.utexas.edu
University of Texas - Austin

We consider the class of optimization problems arising from computationally intensive L1-regularized M-estimators, where the function or gradient values are very expensive to compute. A particular instance of interest is the L1-regularized MLE for learning Conditional Random Fields (CRFs), which are a popular class of statistical models for varied structured prediction problems such as sequence labeling, alignment, and classification with label taxonomy. L1-regularized MLEs for CRFs are particularly expensive to optimize since computing the gradient values requires an expensive inference step. In this work, we propose the use of a carefully constructed proximal quasi-Newton algorithm for such computationally intensive M-estimation problems, where we employ an aggressive active set selection technique. In a key contribution of the paper, we show that our proximal quasi-Newton algorithm is provably super-linearly convergent, even in the absence of strong convexity, by leveraging a restricted variant of strong convexity. In our experiments, the proposed algorithm converges considerably faster than current state-of-the-art on the problems of sequence labeling and hierarchical classification.

T39 Submodular meets Structured: Finding Diverse Subsets in Exponentially-Large Structured Item Sets

Adarsh Prasad adarsh@cs.utexas.edu
University of Texas at Austin
Stefanie Jegelka stefje@eecs.berkeley.edu
UC Berkeley
Dhruv Batra dbatra@vt.edu
Virginia Tech

To cope with the high level of ambiguity faced in domains such as Computer Vision or Natural Language processing, robust prediction methods often search for a diverse set of high-quality candidate solutions or proposals. In structured prediction problems, this becomes a daunting task, as the solution space (image labelings, sentence parses, etc.) is exponentially large. We study greedy algorithms for finding a diverse subset of solutions in structured-output spaces by drawing new connections between submodular functions over combinatorial item sets and High-Order Potentials (HOPs) studied for graphical models. Specifically, we show via examples that when marginal gains of submodular diversity functions allow structured representations, this enables efficient (sub-linear time) approximate maximization by reducing the greedy augmentation step to inference in a factor graph with appropriately constructed HOPs. We discuss benefits, tradeoffs, and show that our constructions lead to significantly better proposals.

T40 Learning Distributed Representations for Structured Output Prediction

Vivek Srikumar svivek@cs.stanford.edu
Christopher Manning manning@stanford.edu
Stanford University

In recent years, distributed representations of inputs have led to performance gains in many applications by allowing statistical information to be shared across inputs. However, the predicted outputs (labels, and more generally structures) are still treated as discrete objects even though outputs are often not discrete units of meaning. In this paper, we present a new formulation for structured prediction where we represent individual labels in a structure as dense vectors and allow semantically similar labels to share parameters. We extend this representation to larger structures by defining compositionality using tensor products to give a natural generalization of standard structured prediction approaches. We define a learning objective for jointly learning the model parameters and the label vectors and propose an alternating minimization algorithm for learning. We show that our formulation outperforms structural SVM baselines in two tasks: multiclass document classification and part-of-speech tagging.

T41 Conditional Random Field Autoencoders for Unsupervised Structured Prediction

Waleed Ammar waleed.ammar@gmail.com
Chris Dyer cdyer@cs.cmu.edu
Noah Smith nasmith@cs.cmu.edu
Carnegie Mellon University

We introduce a framework for unsupervised learning of structured predictors with overlapping, global features. Each input's latent representation is predicted conditional on the observed data using a feature-rich conditional random field (CRF). Then a reconstruction of the input is (re)generated, conditional on the latent structure, using a generative model which factorizes similarly to the CRF. The autoencoder formulation enables efficient exact inference without resorting to unrealistic independence assumptions or restricting the kinds of features that can be used. We illustrate insightful connections to traditional autoencoders, posterior regularization and multi-view learning. Finally, we show competitive results with instantiations of the framework for two canonical tasks in natural language processing: part-of-speech induction and bitext word alignment, and show that training our model can be substantially more efficient than comparable feature-rich baselines.

T42 Discriminative Metric Learning by Neighborhood Gerrymandering

Shubhendu Trivedi shubhendu@ttic.edu
David Mcallester mcallester@ttic.edu
Greg Shakhnarovich greg@ttic.edu
Toyota Technological Institute at Chicago

We formulate the problem of metric learning for k nearest neighbor classification as a large margin structured prediction problem, with a latent variable representing the choice of neighbors and the task loss directly corresponding to classification error. We describe an efficient algorithm for exact loss augmented inference,and a fast gradient descent algorithm for learning in this model. The objective drives the metric to establish neighborhood boundaries that benefit the true class labels for the training points. Our approach, reminiscent of gerrymandering (redrawing of political boundaries to provide advantage to certain parties), is more direct in its handling of optimizing classification accuracy than those previously proposed. In experiments on a variety of data sets our method is shown to achieve excellent results compared to current state of the art in metric learning.

T43 Fundamental Limits of Online and Distributed Algorithms for Statistical Learning and Estimation

Ohad Shamir ohad.shamir@weizmann.ac.il
Weizmann Institute of Science

Many machine learning approaches are characterized by information constraints on how they interact with the training data. These include memory and sequential access constraints (e.g. fast first-order methods to solve stochastic optimization problems); communication constraints (e.g. distributed learning); partial access to the underlying data (e.g. missing features and multi-armed bandits) and more. However, currently we have little understanding how such information constraints fundamentally affect our performance, independent of the learning problem semantics. For example, are there learning problems where any algorithm which has small memory footprint (or can use any bounded number of bits from each example, or has certain communication constraints) will perform worse than what is possible without such constraints? In this paper, we describe how a single set of results implies positive answers to the above, for several different settings.

T44 Optimal Rates for k-NN Density and mode Estimation

Sanjoy Dasgupta dasgupta@eng.ucsd.edu
UC San Diego
Samory Kpotufe samory@princeton.edu
Princeton University

We present two related contributions of independent interest: (1) high-probability finite sample rates for k-NN density estimation, and (2) practical mode estimators -- based on k-NN -- which attain minimax-optimal rates under surprisingly general distributional conditions.

T45 Learning on graphs using Orthonormal Representation is Statistically Consistent

Rakesh Shivanna rakeshsmysore@gmail.com
Chiranjib Bhattacharyya chiranjib.bhattacharyya@gmail.com
Indian Institute of Science

Existing research \cite{reg} suggests that embedding graphs on a unit sphere can be beneficial in learning labels on the vertices of a graph. However the choice of optimal embedding remains an open issue. **Orthonormal representation** of graphs, a class of embeddings over the unit sphere, was introduced by Lov\'asz \cite{lovasz_shannon}. In this paper, we show that there exists orthonormal representations which are statistically consistent over a large class of graphs, including power law and random graphs. This result is achieved by extending the notion of consistency designed in the inductive setting to graph transduction. As part of the analysis, we explicitly derive relationships between the Rademacher complexity measure and structural properties of graphs, such as the chromatic number. We further show the fraction of vertices of a graph G, on n nodes, that need to be labelled for the learning algorithm to be consistent, also known as labelled sample complexity, is $\Omega\left(\frac{\vartheta(G)}{n}\right)$ where $\vartheta(G)$ is the famous Lov\'asz~ϑ function of the graph. This, for the first time, relates labelled sample complexity to graph connectivity properties, such as the density of graphs. In the multiview setting, whenever individual views are expressed by a graph, it is a well known heuristic that a convex combination of Laplacians \cite{lap_mv1} tend to improve accuracy. The analysis presented here easily extends to Multiple graph transduction, and helps develop a sound statistical understanding of the heuristic, previously unavailable.

T46 Optimal prior-dependent neural population coding under shared input noise

Agnieszka Grabska-Barwinska agnieszka@gatsby.ucl.ac.uk
Gatsby Unit, UCL
Jonathan Pillow pillow@mail.utexas.edu
UT Austin

The brain uses population codes to form distributed, noise-tolerant representations of sensory and motor variables. Recent work has explored the optimality of such codes in order to understand the principles governing population

codes found in the brain. However, the majority of this literature considers either conditionally independent neurons or neurons with noise governed by a stimulus-independent covariance matrix. Here we analyze population coding under a simple alternative model in which latent ``input noise'' corrupts the stimulus prior to encoding by the population. This provides a convenient and tractable description for irreducible uncertainty that cannot be overcome by adding neurons, and induces stimulus-dependent correlations that mimic certain aspects of the correlations observed in real populations. We examine prior-dependent, Bayesian optimal coding in such populations using exact analyses of cases in which the posterior is approximately Gaussian. These analyses extend previous results on independent Poisson population codes, many of which relied on approximate formulas involving Fisher information, and yield an analytic expression for squared loss and a tight upper bound for mutual information. We show that, for homogeneous populations, optimal tuning curve width depends on the prior, the loss function, and grows with amount of input noise. Finally, we examine the implications for the correlations in multi-neuron responses induced by input noise.

T47 Spatio-temporal Representations of Uncertainty in Spiking Neural Networks

Cristina Savin cs664@cam.ac.uk
University of Cambridge
Sophie Denève sophie.deneve@ens.fr
École Normale Supérieure

It has been long argued that, because of inherent ambiguity and noise, the brain needs to represent uncertainty in the form of probability distributions. The neural encoding of such distributions remains however highly controversial. Here we present a novel circuit model for representing multidimensional real-valued distributions using a spike based spatio-temporal code. Our model combines the computational advantages of the currently competing models for probabilistic codes and exhibits realistic neural responses along a variety of classic measures. Furthermore, the model highlights the challenges associated with interpreting neural activity in relation to behavioral uncertainty and points to alternative population-level approaches for the experimental validation of distributed representations.

T48 Optimal Neural Codes for Control and Estimation

Alex Susemihl alex.susemihl@bccn-berlin.de
Berlin Institute of Technology
Ron Meir rmeir@ee.technion.ac.il
Technion
Manfred Opper manfred.opper@tu-berlin.de
TU Berlin

Agents acting in the natural world aim at selecting appropriate actions based on noisy and partial sensory observations. Many behaviors leading to decision making and action selection in a closed loop setting are naturally phrased within a control theoretic framework. Within the framework of optimal Control Theory, one is usually given a cost function which is minimized by selecting a control law based on the observations. While in standard control settings the sensors are assumed fixed, biological systems often gain from the extra flexibility of optimizing the sensors themselves. However, this sensory adaptation is geared towards control rather than perception, as is often assumed. In this work we show that sensory adaptation for control differs from sensory adaptation for perception, even for simple control setups. This implies, consistently with recent experimental results, that when studying sensory adaptation, it is essential to account for the task being performed.

T49 A Bayesian model for identifying hierarchically organised states in neural population activity

Patrick Putzky patrick.putzky@**T**bingen.mpg.de
Giacomo Bassetto giacomo.bassetto@gmail.com
Florian Franzen florian.franzen@**T**bingen.mpg.de
Jakob Macke Jakob.Macke@gmail.com
MPI for Biological Cybernetics

Neural population activity in cortical circuits is not solely driven by external inputs, but is also modulated by endogenous states which vary on multiple time-scales. To understand information processing in cortical circuits, we need to understand the statistical structure of internal states and their interaction with sensory inputs. Here, we present a statistical model for extracting hierarchically organised neural population states from multi-channel recordings of neural spiking activity. Population states are modelled using a hidden Markov decision tree with state-dependent tuning parameters and a generalised linear observation model. We present a variational Bayesian inference algorithm for estimating the posterior distribution over parameters from neural population recordings. On simulated data, we show that we can identify the underlying sequence of population states and reconstruct the ground truth parameters. Using population recordings from visual cortex, we find that a model with two levels of population states outperforms both a one-state and a two-state generalised linear model. Finally, we find that modelling of state-dependence also improves the accuracy with which sensory stimuli can be decoded from the population response.

T50 Clustered factor analysis of multineuronal spike data

Lars Buesing lars@gatsby.ucl.ac.uk
Gatsby Unit, UCL
Timothy Machado tam2138@columbia.edu
John Cunningham jpc74@cam.ac.uk
Liam Paninski liam@stat.columbia.edu
Columbia University

High-dimensional, simultaneous recordings of neural spiking activity are often explored, analyzed and visualized with the help of latent variable or factor models. Such models are however ill-equipped to extract structure beyond shared, distributed aspects of firing activity across multiple cells. Here, we extend unstructured factor models by proposing a model that discovers subpopulations or groups of cells from the pool of recorded neurons. The model combines aspects of mixture of factor analyzer models for capturing clustering structure, and aspects of latent dynamical system models for capturing temporal dependencies. In the resulting model, we infer the subpopulations and the latent factors from data using variational inference and model parameters are estimated by Expectation Maximization (EM). We also address the crucial problem of initializing parameters for EM by extending a sparse subspace clustering algorithm to integer-valued spike count observations. We illustrate the merits of the proposed model by applying it to calcium-imaging data from spinal cord neurons, and we show that it uncovers meaningful clustering structure in the data.

T51 The Large Margin Mechanism for Differentially Private Maximization

Kamalika Chaudhuri kamalika@cs.ucsd.edu
Shuang Song shs037@eng.ucsd.edu
UC San Diego
Daniel Hsu danielhsu@gmail.com
Columbia University

A basic problem in the design of privacy-preserving algorithms is the **private maximization problem**: the goal is to pick an item from a universe that (approximately) maximizes a data-dependent function, all under the constraint of differential privacy. This problem has been used as a sub-routine in many privacy-preserving algorithms for statistics and machine learning. Previous algorithms for this problem are either range-dependent---i.e., their utility diminishes with the size of the universe---or only apply to very restricted function classes. This work provides the first general purpose, range-independent algorithm for private maximization that guarantees approximate differential privacy. Its applicability is demonstrated on two fundamental tasks in data mining and machine learning.

T52 Extremal Mechanisms for Local Differential Privacy

Peter Kairouz kayrouzp@gmail.com
Sewoong Oh swoh@illinois.edu
Pramod Viswanath pramodv@illinois.edu
University of Illinois at Urbana Champaign

Local differential privacy has recently surfaced as a strong measure of privacy in contexts where personal information remains private even from data analysts. Working in a setting where the data providers and data analysts want to maximize the utility of statistical inferences performed on the released data, we study the fundamental tradeoff between local differential privacy and information theoretic utility functions. We introduce a family of extremal privatization mechanisms, which we call staircase mechanisms, and prove that it contains the optimal privatization mechanism that maximizes utility. We further show that for all information theoretic utility functions studied in this paper, maximizing utility is equivalent to solving a linear program, the outcome of which is the optimal staircase mechanism. However, solving this linear program can be computationally expensive since it has a number of variables that is exponential in the data size. To account for this, we show that two simple staircase mechanisms, the binary and randomized response mechanisms, are universally optimal in the high and low privacy regimes, respectively, and well approximate the intermediate regime.

T53 Reputation-based Worker Filtering in Crowdsourcing

Srikanth Jagabathula jskanth@mit.edu
Lakshminarayanan Subramanian lakshmi@cs.nyu.edu
Ashwin Venkataraman ashwin@cs.nyu.edu
New York University

In this paper, we study the problem of aggregating noisy labels from crowd workers to infer the underlying true labels of binary tasks. Unlike most prior work which has examined this problem under the random worker paradigm, we consider a much broader class of {\em adversarial} workers with no specific assumptions on their labeling strategy. Our key contribution is the design of a computationally efficient reputation algorithm to identify and filter out these adversarial workers in crowdsourcing systems. Our algorithm uses the concept of optimal semi-matchings in conjunction with worker penalties based on label disagreements, to assign a reputation score for every worker. We provide strong theoretical guarantees for deterministic adversarial strategies as well as the extreme case of {\em sophisticated} adversaries where we analyze the worst-case behavior of our algorithm. Finally, we show that our reputation algorithm can significantly improve the accuracy of existing label aggregation algorithms in real-world crowdsourcing datasets.

T54 Feedback Detection for Live Predictors

Stefan Wager swager@stanford.edu
Stanford University
Nick Chamandy chamandy@google.com
Omkar Muralidharan omuralidharan@google.com
Amir Najmi amir@google.com
Google

A predictor that is deployed in a live production system may perturb the features it uses to make predictions. Such a feedback loop can occur, for example, when a model that predicts a certain type of behavior ends up causing the behavior it predicts, thus creating a self-fulfilling prophecy. In this paper we analyze predictor feedback detection as a causal inference problem, and introduce a local randomization scheme that can be used to detect non-linear feedback in real-world problems. We conduct a pilot study for our proposed methodology using a predictive system currently deployed as a part of a search engine.

T55 DFacTo: Distributed Factorization of Tensors

Joon Hee Choi choi240@purdue.edu
S. Vishwanathan vishy@stat.purdue.edu
Purdue University

We present a technique for significantly speeding up Alternating Least Squares (ALS) and Gradient Descent (GD), two widely used algorithms for tensor factorization. By exploiting properties of the Khatri-Rao product, we show how to efficiently address a computationally challenging sub-step of both algorithms. Our algorithm, DFacTo, only requires two matrix-vector products and is easy to parallelize. DFacTo is not only scalable but also on average 4 to 10 times faster than competing algorithms on a variety of datasets. For instance, DFacTo only takes 480 seconds on 4 machines to perform one iteration of the ALS algorithm and 1,143 seconds to perform one iteration of the GD algorithm on a 6.5 million x 2.5 million x 1.5 million dimensional tensor with 1.2 billion non-zero entries.

T56 Distributed Power-law Graph Computing: Theoretical and Empirical Analysis

Cong Xie xcgoner1108@gmail.com
Ling Yan yling0718@gmail.com
Zhihua Zhang zhzhang@gmail.com
Shanghai Jiao Tong University
Wu-Jun Li liwujun@nju.edu.cn
Nanjing University

With the emergence of big graphs in a variety of real applications like social networks, machine learning based on distributed graph-computing~(DGC) frameworks has attracted much attention from big data machine learning community. In DGC

frameworks, the graph partitioning~(GP) strategy plays a key role to affect the performance, including the workload balance and communication cost. Typically, the degree distributions of natural graphs from real applications follow skewed power laws, which makes GP a challenging task. Recently, many methods have been proposed to solve the GP problem. However, the existing GP methods cannot achieve satisfactory performance for applications with power-law graphs. In this paper, we propose a novel vertex-cut method, called **degree-based hashing**~(DBH), for GP. DBH makes effective use of the skewed degree distributions for GP. We theoretically prove that DBH can achieve lower communication cost than existing methods and can simultaneously guarantee good workload balance. Furthermore, empirical results on several large power-law graphs also show that DBH can outperform the state of the art.

T57 Scalable Non-linear Learning with Adaptive Polynomial Expansions

Alekh Agarwal alekha@microsoft.com
John Langford jl@hunch.net
Microsoft Research
Alina Beygelzimer beygel@yahoo-inc.com
Yahoo
Daniel Hsu danielhsu@gmail.com
Columbia University
Matus Telgarsky mtelgars@cs.ucsd.edu
University of Michigan

Can we effectively learn a nonlinear representation in time comparable to linear learning? We describe a new algorithm that explicitly and adaptively expands higher-order interaction features over base linear representations. The algorithm is designed for extreme computational efficiency, and an extensive experimental study shows that its computation/prediction tradeoff ability compares very favorably against strong baselines.

T58 Median Selection Subset Aggregation for Parallel Inference

Xiangyu Wang xw56@stat.duke.edu
David Dunson dunson@stat.duke.edu
Duke University
Peichao Peng ppeichao@wharton.upenn.edu
University of Pennsylvania

For massive data sets, efficient computation commonly relies on distributed algorithms that store and process subsets of the data on different machines, minimizing communication costs. Our focus is on regression and classification problems involving many features. A variety of distributed algorithms have been proposed in this context, but challenges arise in defining an algorithm with low communication, theoretical guarantees and excellent practical performance in general settings. We propose a MEdian Selection Subset AGgregation Estimator (message) algorithm, which attempts to solve these problems. The algorithm applies feature selection in parallel for each subset using Lasso or another method, calculates the `median' feature inclusion index, estimates coefficients for the selected features in parallel for each subset, and then averages these estimates. The algorithm is simple, involves very minimal communication, scales efficiently in both sample and feature size, and has theoretical guarantees. In particular, we show model selection consistency and coefficient estimation efficiency. Extensive experiments show excellent performance in variable selection, estimation, prediction, and computation time relative to usual competitors.

T59 Asymmetric LSH (ALSH) for Sublinear Time Maximum Inner Product Search (MIPS)

Anshumali Shrivastava anshu@cs.cornell.edu
Cornell University
Ping Li pingli@stat.rutgers.edu
Rutgers University

We present the first provably sublinear time hashing algorithm for approximate **Maximum Inner Product Search** (MIPS). Searching with (un-normalized) inner product as the underlying similarity measure is a known difficult problem and finding hashing schemes for MIPS was considered hard. While the existing Locality Sensitive Hashing (LSH) framework is insufficient for solving MIPS, in this paper we extend the LSH framework to allow asymmetric hashing schemes. Our proposal is based on a key observation that the problem of finding maximum inner products, after independent asymmetric transformations, can be converted into the problem of approximate near neighbor search in classical settings. This key observation makes efficient sublinear hashing scheme for MIPS possible. Under the extended asymmetric LSH (ALSH) framework, this paper provides an example of explicit construction of provably fast hashing scheme for MIPS. Our proposed algorithm is simple and easy to implement. The proposed hashing scheme leads to significant computational savings over the two popular conventional LSH schemes: (i) Sign Random Projection (SRP) and (ii) hashing based on p-stable distributions for L_2 norm (L2LSH), in the collaborative filtering task of item recommendations on Netflix and Movielens (10M) datasets.

T60 Discrete Graph Hashing

Wei Liu wliu@ee.columbia.edu
IBM T. J. Watson Research Center
Cun Mu cm3052@columbia.edu
Shih-Fu Chang sfchang@ee.columbia.edu
Columbia University
Sanjiv Kumar sanjivk@google.com
Google

Hashing has emerged as a popular technique for fast nearest neighbor search in gigantic databases. In particular, learning based hashing has received considerable attention due to its appealing storage and search efficiency. However, the performance of most unsupervised learning based hashing methods deteriorates rapidly as the hash code length increases. We argue that the degraded performance is due to inferior optimization procedures used to achieve discrete binary codes. This paper presents a graph-based unsupervised hashing model to preserve the neighborhood structure of massive data in a discrete code space. We cast the graph hashing problem into a discrete optimization framework which directly learns the binary codes. A tractable alternating maximization algorithm is then proposed to explicitly deal with the discrete constraints, yielding high-quality codes to well capture the local neighborhoods. Extensive experiments performed on four large datasets with up to one million samples show that our discrete optimization based graph hashing method obtains superior search accuracy over state-of-the-art unsupervised hashing methods, especially for longer codes.

T61 Orbit Regularization

Renato Negrinho renato.negrinho@gmail.com
Instituto de Telecomunicações
Andre Martins afm@cs.cmu.edu
CMU

We propose a general framework for regularization based on group majorization. In this framework, a group is defined to act on the parameter space and an orbit is fixed; to control complexity, the model parameters are confined to lie in the convex hull of this orbit (the orbitope). Common regularizers are recovered as particular cases, and a connection is revealed between the recent sorted 1 -norm and the hyperoctahedral group. We derive the properties a group must satisfy for being amenable to optimization with conditional and projected gradient algorithms. Finally, we suggest a continuation strategy for orbit exploration, presenting simulation results for the symmetric and hyperoctahedral groups.

T62 Covariance shrinkage for autocorrelated data

Daniel Bartz daniel.bartz@tu-berlin.de
Klaus-Robert Müller Klaus-Robert.Mueller@tu-berlin.de
TU Berlin

The accurate estimation of covariance matrices is essential for many signal processing and machine learning algorithms. In high dimensional settings the sample covariance is known to perform poorly, hence regularization strategies such as analytic shrinkage of Ledoit/Wolf are applied. In the standard setting, i.i.d. data is assumed, however, in practice, time series typically exhibit strong autocorrelation structure, which introduces a pronounced estimation bias. Recent work by Sancetta has extended the shrinkage framework beyond i.i.d. data. We contribute in this work by showing that the Sancetta estimator, while being consistent in the high-dimensional limit, suffers from a high bias in finite sample sizes. We propose an alternative estimator, which is (1) unbiased, (2) less sensitive to hyperparameter choice and (3) yields superior performance in simulations on toy data and on a real world data set from an EEG-based Brain-Computer-Interfacing experiment.

T63 Sparse Space-Time Deconvolution for Calcium Image Analysis

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

We describe a unified formulation and algorithm to find an extremely sparse representation for Calcium image sequences in terms of cell locations, cell shapes, spike timings and impulse responses. Solution of a single optimization problem yields cell segmentations and activity estimates that are on par with the state of the art, without the need for heuristic pre- or postprocessing. Experiments on real and synthetic data demonstrate the viability of the proposed method.

T64 A Dual Algorithm for Olfactory Computation in the Locust Brain

Sina Tootoonian st582@eng.cam.ac.uk
Mate Lengyel m.lengyel@eng.cam.ac.uk
University of Cambridge

We study the early locust olfactory system in an attempt to explain its well-characterized structure and dynamics. We first propose its computational function as recovery of high-dimensional sparse olfactory signals from a small number of measurements. Detailed experimental knowledge about this system rules out standard algorithmic solutions to this problem. Instead, we show that solving a dual formulation of the corresponding optimisation problem yields structure and dynamics in good agreement with biological data. Further biological constraints lead us to a reduced form of this dual formulation in which the system uses independent component analysis to continuously adapt to its olfactory environment to allow accurate sparse recovery. Our work demonstrates the challenges and rewards of attempting detailed understanding of experimentally well-characterized systems.

T65 Online Optimization for Max-Norm Regularization

Jie Shen jieshen.sjtu@gmail.com
Ping Li pingli@stat.rutgers.edu
Rutgers University
Huan Xu mpexuh@nus.edu.sg
NUS

Max-norm regularizer has been extensively studied in the last decade as it promotes an effective low rank estimation of the underlying data. However, max-norm regularized problems are typically formulated and solved in a batch manner, which prevents it from processing big data due to possible memory bottleneck. In this paper, we propose an online algorithm for solving max-norm regularized problems that is scalable to large problems. Particularly, we consider the matrix decomposition problem as an example, although our analysis can also be applied in other problems such as matrix completion. The key technique in our algorithm is to reformulate the max-norm into a matrix factorization form, consisting of a basis component and a coefficients one. In this way, we can solve the optimal basis and coefficients alternatively. We prove that the basis produced by our algorithm converges to a stationary point asymptotically. Experiments demonstrate encouraging results for the effectiveness and robustness of our algorithm. See the full paper at arXiv:1406.3190.

T66 Finding a sparse vector in a subspace: Linear sparsity using alternating directions

Qing Qu qingqu87@gmail.com
Ju Sun js4038@columbia.edu
John Wright johnwright@ee.columbia.edu
Columbia University

We consider the problem of recovering the sparsest vector in a subspace $\mathcal{S}\in\mathbb{R}_p$ with $\dim(\mathcal{S})=n$. This problem can be considered a homogeneous variant of the sparse recovery problem, and finds applications in sparse dictionary learning, sparse PCA, and other problems in signal processing and machine learning. Simple convex heuristics for this problem provably break down when the fraction of nonzero entries in the target sparse vector substantially exceeds $1/n--\sqrt{}$. In

contrast, we exhibit a relatively simple nonconvex approach based on alternating directions, which provably succeeds even when the fraction of nonzero entries is $\Omega(1)$. To our knowledge, this is the first practical algorithm to achieve this linear scaling. This result assumes a planted sparse model, in which the target sparse vector is embedded in an otherwise random subspace. Empirically, our proposed algorithm also succeeds in more challenging data models arising, e.g., from sparse dictionary learning.

T67 Compressive Sensing of Signals from a GMM with Sparse Precision Matrices

Jianbo Yang jbyang.email@gmail.com
Xuejun Liao xjliao@ee.duke.edu
Lawrence Carin lcarin@duke.edu
Duke University
Minhua Chen minhua@uchicago.edu
University of Chicago

This paper is concerned with compressive sensing of signals drawn from a Gaussian mixture model (GMM) with sparse precision matrices. Previous work has shown: (i) a signal drawn from a given GMM can be perfectly reconstructed from r noise-free measurements if the (dominant) rank of each covariance matrix is less than r; (ii) a sparse Gaussian graphical model can be efficiently estimated from fully-observed training signals using graphical lasso. This paper addresses a problem more challenging than both (i) and (ii), by assuming that the GMM is unknown and each signal is only partially observed through incomplete linear measurements. Under these challenging assumptions, we develop a hierarchical Bayesian method to simultaneously estimate the GMM and recover the signals using solely the incomplete measurements and a Bayesian shrinkage prior that promotes sparsity of the Gaussian precision matrices. In addition, we provide theoretical performance bounds to relate the reconstruction error to the number of signals for which measurements are available, the sparsity level of precision matrices, and the "incompleteness" of measurements. The proposed method is demonstrated extensively on compressive sensing of imagery and video, and the results with simulated and hardware-acquired real measurements show significant performance improvement over state-of-the-art methods.

T68 Clustering from Labels and Time-Varying Graphs

Shiau-Hong Lim shonglim@gmail.com
Huan Xu mpexuh@nus.edu.sg
National University of Singapore
Yudong Chen ydchen@utexas.edu
University of California, Berkeley

We present a general framework for graph clustering where a label is observed to each pair of nodes. This allows a very rich encoding of various types of pairwise interactions between nodes. We propose a new tractable approach to this problem based on maximum likelihood estimator and convex optimization. We analyze our algorithm under a general generative model, and provide both necessary and sufficient conditions for successful recovery of the underlying clusters. Our theoretical results cover and subsume a wide range of existing graph clustering results including planted partition, weighted clustering and partially observed graphs. Furthermore, the result is applicable to novel settings including time-varying graphs such that new insights can be gained on solving these problems. Our theoretical findings are further supported by empirical results on both synthetic and real data.

T69 On the relations of LFPs & Neural Spike Trains

David Carlson david.carlson@duke.edu
Jana Schaich Borg janaschaichborg@gmail.com
Kafui Dzirasa kafui.dzirasa@duke.edu
Lawrence Carin lcarin@duke.edu
Duke University

One of the goals of neuroscience is to identify neural networks that correlate with important behaviors, environments, or genotypes. This work proposes a strategy for identifying neural networks characterized by time- and frequency-dependent connectivity patterns, using convolutional dictionary learning that links spike-train data to local field potentials (LFPs) across multiple areas of the brain. Analytical contributions are: (i) modeling dynamic relationships between LFPs and spikes; (ii) describing the relationships between spikes and LFPs, by analyzing the ability to predict LFP data from one region based on spiking information from across the brain; and (iii) development of a clustering methodology that allows inference of similarities in neurons from multiple regions. Results are based on data sets in which spike and LFP data are recorded simultaneously from up to 16 brain regions in a mouse.

T70 Scalable Inference for Neuronal Connectivity from Calcium Imaging

Alyson Fletcher afletcher@soe.ucsc.edu
University of California, Santa Cruz
Sundeep Rangan srangan@poly.edu
Polytechnic Institute of New York University

Fluorescent calcium imaging provides a potentially powerful tool for inferring connectivity in neural circuits with up to thousands of neurons. However, a key challenge in using calcium imaging for connectivity detection is that current systems often have a temporal response and frame rate that can be orders of magnitude slower than the underlying neural spiking process. Bayesian inference based on expectation-maximization (EM) have been proposed to overcome these limitations, but they are often computationally demanding since the E-step in the EM procedure typically involves state estimation in a high-dimensional nonlinear dynamical system. In this work, we propose a computationally fast method for the state estimation based on a hybrid of loopy belief propagation and approximate message passing (AMP). The key insight is that a neural system as viewed through calcium imaging can be factorized into simple scalar dynamical systems for each neuron with linear interconnections between the neurons. Using the structure, the updates in the proposed hybrid AMP methodology can be computed by a set of one-dimensional state estimation procedures and linear transforms with the connectivity matrix. This yields a computationally scalable method for inferring connectivity of large neural circuits. Simulations of the method on realistic neural networks demonstrate good accuracy with computation times that are potentially significantly faster than current approaches based on Markov Chain Monte Carlo methods.

T71 A Synaptical Story of Persistent Activity with Graded Lifetime in a Neural System

Yuanyuan Mi miyuanyuan0102@163.com
Weizmann Institute of Science
Luozheng Li luozhengli@gmail.com
Si Wu wusi@bnu.edu.cn
Dahui Wang wangdh@bnu.edu.cn
Beijnng Normal University

Persistent activity refers to the phenomenon that cortical neurons keep firing even after the stimulus triggering the initial neuronal responses is moved. Persistent activity is widely believed to be the substrate for a neural system retaining a memory trace of the stimulus information. In a conventional view, persistent activity is regarded as an attractor of the network dynamics, but it faces a challenge of how to be closed properly. Here, in contrast to the view of attractor, we consider that the stimulus information is encoded in a marginally unstable state of the network which decays very slowly and exhibits persistent firing for a prolonged duration. We propose a simple yet effective mechanism to achieve this goal, which utilizes the property of short-term plasticity (STP) of neuronal synapses. STP has two forms, short-term depression (STD) and short-term facilitation (STF), which have opposite effects on retaining neuronal responses. We find that by properly combining STF and STD, a neural system can hold persistent activity of graded lifetime, and that persistent activity fades away naturally without relying on an external drive. The implications of these results on neural information representation are discussed.

T72 Analog Memories in a Balanced Rate-Based Network of E-I Neurons

Dylan Festa df325@cam.ac.uk
Guillaume Hennequin gjeh2@cam.ac.uk
Mate Lengyel m.lengyel@eng.cam.ac.uk
University of Cambridge

The persistent and graded activity often observed in cortical circuits is sometimes seen as a signature of autoassociative retrieval of memories stored earlier in synaptic efficacies. However, despite decades of theoretical work on the subject, the mechanisms that support the storage and retrieval of memories remain unclear. Previous proposals concerning the dynamics of memory networks have fallen short of incorporating some key physiological constraints in a unified way. Specifically, some models violate Dale's law (i.e. allow neurons to be both excitatory and inhibitory), while some others restrict the representation of memories to a binary format, or induce recall states in which some neurons fire at rates close to saturation. We propose a novel control-theoretic framework to build functioning attractor networks that satisfy a set of relevant physiological constraints. We directly optimize networks of excitatory and inhibitory neurons to force sets of arbitrary analog patterns to become stable fixed points of the dynamics. The resulting networks operate in the balanced regime, are robust to corruptions of the memory cue as well as to ongoing noise, and incidentally explain the reduction of trial-to-trial variability following stimulus onset that is ubiquitously observed in sensory and motor cortices. Our results constitute a step forward in our understanding of the neural substrate of memory.

T73 Design Principles of the Hippocampal Cognitive Map

Kimberly Stachenfeld kls4@princeton.edu
Matthew Botvinick matthewb@princeton.edu
Princeton University
Samuel Gershman sjgershm@mit.edu
Massachusetts Institute of Technology

Hippocampal place fields have been shown to reflect behaviorally relevant aspects of space. For instance, place fields tend to be skewed along commonly traveled directions, they cluster around rewarded locations, and they are constrained by the geometric structure of the environment. We hypothesize a set of design principles for the hippocampal cognitive map that explain how place fields represent space in a way that facilitates navigation and reinforcement learning. In particular, we suggest that place fields encode not just information about the current location, but also predictions about future locations under the current transition distribution. Under this model, a variety of place field phenomena arise naturally from the structure of rewards, barriers, and directional biases as reflected in the transition policy. Furthermore, we demonstrate that this representation of space can support efficient reinforcement learning. We also propose that grid cells compute the eigendecomposition of place fields in part because is useful for segmenting an enclosure along natural boundaries. When applied recursively, this segmentation can be used to discover a hierarchical decomposition of space. Thus, grid cells might be involved in computing subgoals for hierarchical reinforcement learning.

T74 Feedforward Learning of Mixture Models

Matthew Lawlor matthew.lawlor@yale.edu
Steven Zucker zucker@cs.yale.edu
Yale University

We develop a biologically-plausible learning rule that provably converges to the class means of general mixture models. This rule generalizes the classical BCM neural rule within a tensor framework, substantially increasing the generality of the learning problem it solves. It achieves this by incorporating triplets of samples from the mixtures, which provides a novel information processing interpretation to spike-timing-dependent plasticity. We provide both proofs of convergence, and a close fit to experimental data on STDP.

T75 Scalable Methods for Nonnegative Matrix Factorizations of Near-separable Tall-and-skinny Matrices

Austin Benson arbenson@stanford.edu
Jason Lee jdl17@stanford.edu
Stanford University
Bartek Rajwa brajwa@purdue.edu
David Gleich dfgleic@sandia.gov
Purdue University

Numerous algorithms are used for nonnegative matrix factorization under the assumption that the matrix is nearly separable. In this paper, we show how to make these algorithms scalable for data matrices that have many more rows than columns, so-called "tall-and-skinny matrices." One key component to these improved methods is an orthogonal matrix transformation that preserves the

separability of the NMF problem. Our final methods need to read the data matrix only once and are suitable for streaming, multi-core, and MapReduce architectures. We demonstrate the efficacy of these algorithms on terabyte-sized matrices from scientific computing and bioinformatics.

T76 Sparse PCA via Covariance Thresholding

Yash Deshpande yashd@stanford.edu
Andrea Montanari montanari@stanford.edu
Stanford University

In sparse principal component analysis we are given noisy observations of a low-rank matrix of dimension $n\times p$ and seek to reconstruct it under additional sparsity assumptions. In particular, we assume here that the principal components $\bv_1,\dots,\bv_r$ have at most $k_1,\cdots,k_q$ non-zero entries respectively, and study the high-dimensional regime in which p is of the same order as n. In an influential paper, Johnstone and Lu \cite{johnstone2004sparse} introduced a simple algorithm that estimates the support of the principal vectors $\bv_1,\dots,\bv_r$ by the largest entries in the diagonal of the empirical covariance. This method can be shown to succeed with high probability if $k_q\le C_1 n/\log p$, and to fail with high probability if $k_q\ge C_2 n/\log p$ for two constants $0<C_1,C_2<\infty$. Despite a considerable amount of work over the last ten years, no practical algorithm exists with provably better support recovery guarantees. Here we analyze a covariance thresholding algorithm that was recently proposed by Krauthgamer, Nadler and Vilenchik \cite{KrauthgamerSPCA}. We confirm empirical evidence presented by these authors and rigorously prove that the algorithm succeeds with high probability for k of order n $\sqrt{}$. Recent conditional lower bounds \cite{berthet2013computational} suggest that it might be impossible to do significantly better. The key technical component of our analysis develops new bounds on the norm of kernel random matrices, in regimes that were not considered before.

T77 Low Rank Approximation Lower Bounds in Row-Update Streams

David Woodruff dpwoodru@us.ibm.com
IBM Research

We study low-rank approximation in the streaming model in which the rows of an $n\times d$ matrix $A$are presented one at a time in an arbitrary order. At the end of the stream, the streaming algorithm should output a $k\times d$ matrix R so that $\|A-AR^\dagger R\|_{2F}\le(1+\eps)\|A-A_k\|_{2F}$, where A_k is the best rank-k approximation to A. A deterministic streaming algorithm of Liberty (KDD, 2013), with an improved analysis of Ghashami and Phillips (SODA, 2014), provides such a streaming algorithm using $O(dk/\epsilon)$ words of space. A natural question is if smaller space is possible. We give an almost matching lower bound of $\Omega(dk/\epsilon)$ bits of space, even for randomized algorithms which succeed only with constant probability. Our lower bound matches the upper bound of Ghashami and Phillips up to the word size, improving on a simple $\Omega(dk)$ space lower bound.

T78 Recovery of Coherent Data via Low-Rank Dictionary Pursuit

Guangcan Liu gcliu1982@gmail.com
Ping Li pingli@stat.rutgers.edu
Rutgers University

The recently established RPCA method provides a convenient way to restore low-rank matrices from grossly corrupted observations. While elegant in theory and powerful in reality, RPCA is not an ultimate solution to the low-rank matrix recovery problem. Indeed, its performance may not be perfect even when data are strictly low-rank. This is because RPCA ignores clustering structures of the data which are ubiquitous in applications. As the number of cluster grows, the coherence of data keeps increasing, and accordingly, the recovery performance of RPCA degrades. We show that the challenges raised by coherent data (i.e., data with high coherence) could be alleviated by Low-Rank Representation (LRR)~\cite{tpami_2013_lrr}, provided that the dictionary in LRR is configured appropriately. More precisely, we mathematically prove that if the dictionary itself is low-rank then LRR is immune to the coherence parameter which increases with the underlying cluster number. This provides an elementary principle for dealing with coherent data and naturally leads to a practical algorithm for obtaining proper dictionaries in unsupervised environments. Experiments on randomly generated matrices and real motion sequences verify our claims. See the full paper at arXiv:1404.4032.

T79 Tight convex relaxations for sparse matrix factorization

Emile Richard r.emile.richard@gmail.com
Stanford
Guillaume Obozinski Guillaume.Obozinski@imagine.enpc.fr
Ecole des Ponts - ParisTech
Jean-Philippe Vert Jean-Philippe.Vert@mines.org
Mines ParisTech

Based on a new atomic norm, we propose a new convex formulation for sparse matrix factorization problems in which the number of nonzero elements of the factors is assumed fixed and known. The formulation counts sparse PCA with multiple factors, subspace clustering and low-rank sparse bilinear regression as potential applications. We compute slow rates and an upper bound on the statistical dimension of the suggested norm for rank 1 matrices, showing that its statistical dimension is an order of magnitude smaller than the usual l_1-norm, trace norm and their combinations. Even though our convex formulation is in theory hard and does not lead to provably polynomial time algorithmic schemes, we propose an active set algorithm leveraging the structure of the convex problem to solve it and show promising numerical results.

T80 Robust Tensor Decomposition with Gross Corruption

Quanquan Gu qgu@princeton.edu
Princeton University
Huan Gui huangui2@illinois.edu
Jiawei Han hanj@cs.uiuc.edu
University of Illinois at Urbana-Champaign

In this paper, we study the statistical performance of robust tensor decomposition with gross corruption. The observations are noisy realization of the superposition of a low-rank tensor $\mathcal{W}_*$ and an entrywise sparse corruption tensor $\mathcal{V}_*$. Unlike conventional noise with bounded variance in previous convex tensor decomposition analysis, the magnitude of the gross corruption can be arbitrary large. We show that under certain conditions, the true low-rank tensor as well as the sparse corruption tensor can be recovered simultaneously. Our theory yields nonasymptotic Frobenius-norm estimation error bounds for each tensor separately. We show through numerical experiments that our theory can precisely predict the scaling behavior in practice.

T81 PEWA: Patch-based Exponentially Weighted Aggregation for image denoising

Charles Kervrann charles.kervrann@inria.fr
Inria

Patch-based methods have been widely used for noise reduction in recent years. In this paper, we propose a general statistical aggregation method which combines image patches denoised with several commonly-used algorithms. We show that weakly denoised versions of the input image obtained with standard methods, can serve to compute an efficient patch-based aggregated estimd aggregation (EWA) estimator. The resulting approach (PEWA) is based on a MCMC sampling and has a nice statistical foundation while producing denoising results that are comparable to the current state-of-the-art. We demonstrate the performance of the denoising algorithm on real images and we compare the results to several competitive methods.

T82 A Multi-World Approach to Question Answering about Real-World Scenes based on Uncertain Input

Mateusz Malinowski mmalinow@mpi-inf.mpg.de
Mario Fritz mfritz@mpi-inf.mpg.de
Max Planck Institute for Informatics

We propose a method for automatically answering questions about images by bringing together recent advances from natural language processing and computer vision. We combine discrete reasoning with uncertain predictions by a multi-world approach that represents uncertainty about the perceived world in a bayesian framework. Our approach can handle human questions of high complexity about realistic scenes and replies with range of answer like counts, object classes, instances and lists of them. The system is directly trained from question-answer pairs. We establish a first benchmark for this task that can be seen as a modern attempt at a visual turing test.

T83 Quantized Kernel Learning for Feature Matching

Danfeng Qin qind@vision.ee.ethz.ch
Luc Gool vangool@vision.ee.ethz.ch
Matthieu Guillaumin guillaumin@vision.ee.ethz.ch
Computer Vision Lab, ETH Zurich
Xuanli Chen xuanli.chen@tum.de
TU Munich

Matching local visual features is a crucial problem in computer vision and its accuracy greatly depends on the choice of similarity measure. As it is generally very difficult to design by hand a similarity or a kernel perfectly adapted to the data of interest, learning it automatically with as few assumptions as possible is preferable. However, available techniques for kernel learning suffer from several limitations, such as restrictive parametrization or scalability. In this paper, we introduce a simple and flexible family of non-linear kernels which we refer to as Quantized Kernels (QK). QKs are arbitrary kernels in the index space of a data quantizer, i.e., piecewise constant similarities in the original feature space. Quantization allows to compress features and keep the learning tractable. As a result, we obtain state-of-the-art matching performance on a standard benchmark dataset with just a few bits to represent each feature dimension. QKs also have explicit non-linear, low-dimensional feature mappings that grant access to Euclidean geometry for uncompressed features.

T84 Diverse Sequential Subset Selection for Supervised Video Summarization

Boqing Gong boqinggo@usc.edu
Fei Sha feisha@usc.edu
Wei-Lun Chao weilunchao760414@gmail.com
University of Southern California
Kristen Grauman grauman@cs.utexas.edu
UT Austin

Video summarization is a challenging problem with great application potential. Whereas prior approaches, largely unsupervised in nature, focus on sampling useful frames and assembling them as summaries, we consider video summarization as a supervised subset selection problem. Our idea is to teach the system to learn from human-created summaries how to select informative and diverse subsets, so as to best meet evaluation metrics derived from human-perceived quality. To this end, we propose the sequential determinantal point process (seqDPP), a probabilistic model for diverse sequential subset selection. Our novel seqDPP heeds the inherent sequential structures in video data, thus overcoming the deficiency of the standard DPP, which treats video frames as randomly permutable items. Meanwhile, seqDPP retains the power of modeling diverse subsets, essential for summarization. Our extensive results of summarizing videos from 3 datasets demonstrate the superior performance of our method, compared to not only existing unsupervised methods but also naive applications of the standard DPP model.

T85 Grouping-Based Low-Rank Trajectory Completion and 3D Reconstruction

Katerina Fragkiadaki katef@berkeley.edu
Pablo Arbelaez arbelaez@berkeley.edu
Jitendra Malik malik@cs.berkeley.edu
University of California, Berkeley
Marta Salas msalasg@berkeley.edu
Universidad de Zaragoza

Extracting 3D shape of deforming objects in monocular videos, a task known as non-rigid structure-from-motion (NRSfM), has so far been studied only on synthetic datasets and controlled environments. Typically, the objects to reconstruct are pre-segmented, they exhibit limited rotations and occlusions, or full-length trajectories are assumed. In order to integrate NRSfM into current video analysis pipelines, one needs to consider as input realistic -thus incomplete- tracking, and perform spatio-temporal grouping to segment the objects from their surroundings. Furthermore, NRSfM needs to be robust to noise in both segmentation and tracking, e.g., drifting, segmentation ``leaking'', optical flow ``bleeding'' etc. In this paper, we make a first attempt towards this goal, and propose a method that combines dense optical flow tracking, motion trajectory clustering and NRSfM for 3D reconstruction of objects in videos. For each trajectory cluster, we compute multiple reconstructions by minimizing the reprojection error and the rank of the 3D shape under different rank bounds of the trajectory matrix. We show that dense 3D shape is extracted and trajectories are completed across occlusions and low textured regions, even under mild relative motion between the object and the camera. We achieve competitive results on a public NRSfM benchmark while using fixed parameters across all sequences and handling incomplete trajectories, in contrast to existing approaches. We further test our approach on popular video segmentation datasets. To the best of our knowledge, our method is the first to extract dense object models from realistic videos, such as those found in Youtube or Hollywood movies, without object-specific priors.

T86 Learning Mixtures of Submodular Functions for Image Collection Summarization

Sebastian Tschiatschektschiatschek@tugraz.at
Graz University of Technology
Rishabh Iyer rkiyer@u.washington.edu
Haochen Wei weihch90@gmail.com
Jeff Bilmes bilmes@ee.washington.edu
University of Washington, Seattle

We address the problem of image collection summarization by learning mixtures of submodular functions. We argue that submodularity is very natural to this problem, and we show that a number of previously used scoring functions are submodular — a property not explicitly mentioned in these publications. We provide classes of submodular functions capturing the necessary properties of summaries, namely coverage, likelihood, and diversity. To learn mixtures of these submodular functions as scoring functions, we formulate summarization as a supervised learning problem using large-margin structured prediction. Furthermore, we introduce a novel evaluation metric, which we call V-ROUGE, for automatic summary scoring. While a similar metric called ROUGE has been successfully applied to document summarization [14], no such metric was known for quantifying the quality of image collection summaries. We provide a new dataset consisting of 14 real-world image collections along with many human-generated ground truth summaries collected using mechanical turk. We also extensively compare our method with previously explored methods for this problem and show that our learning approach outperforms all competitors on this new dataset. This paper provides, to our knowledge, the first systematic approach for quantifying the problem of image collection summarization, along with a new dataset of image collections and human summaries.

T87 Deep Learning Face Representation by Joint Identification-Verification

Yi Sun sy011@ie.cuhk.edu.hk
Xiaogang Wang xgwang@ee.cuhk.edu.hk
Xiaoou Tang xtang@ie.cuhk.edu.hk
The Chinese University of Hong Kong
Yuheng Chen chyh1990@gmail.com
Tsinghua University

The key challenge of face recognition is to develop effective feature representations for reducing intra-personal variations while enlarging inter-personal differences. In this paper, we show that it can be well solved with deep learning and using both face identification and verification signals as supervision. The Deep IDentification-verification features (DeepID2) are learned with carefully designed deep convolutional networks. The face identification task increases the inter-personal variations by drawing DeepID2 features extracted from different identities apart, while the face verification task reduces the intra-personal variations by pulling DeepID2 features extracted from the same identity together, both of which are essential to face recognition. The learned DeepID2 features can be well generalized to new identities unseen in the training data. On the challenging LFW dataset, 99.15% face verification accuracy is achieved. Compared with the best previous deep learning result on LFW, the error rate has been significantly reduced by 67%.

T88 Fast Training of Pose Detectors in the Fourier Domain

João F. Henriques henriques@isr.uc.pt
Pedro Martins pedromartins@isr.uc.pt
Rui Caseiro ruicaseiro@isr.uc.pt
Jorge Batista batista@isr.uc.pt
University of Coimbra

In many datasets, the samples are related by a known image transformation, such as rotation, or a repeatable non-rigid deformation. This applies to both datasets with the same objects under different viewpoints, and datasets augmented with virtual samples. Such datasets possess a high degree of redundancy, because geometrically-induced transformations should preserve intrinsic properties of the objects. Likewise, ensembles of classifiers used for pose estimation should also share many characteristics, since they are related by a geometric transformation. By assuming that this transformation is norm-preserving and cyclic, we propose a closed-form solution in the Fourier domain that can eliminate most redundancies. It can leverage off-the-shelf solvers with no modification (e.g. libsvm), and train several pose classifiers simultaneously at no extra cost. Our experiments show that training a sliding-window object detector and pose estimator can be sped up by orders of magnitude, for transformations as diverse as planar rotation, the walking motion of pedestrians, and out-of-plane rotations of cars.

T89 LSDA: Large Scale Detection through Adaptation

Judy Hoffman jhoffman@eecs.berkeley.edu
Sergio Guadarrama sguada@eecs.berkeley.edu
Eric Tzeng etzeng@eecs.berkeley.edu
Jeff Donahue jdonahue@eecs.berkeley.edu
Ross Girshick rbg@eecs.berkeley.edu
Trevor Darrell trevor@eecs.berkeley.edu
UC Berkeley
Ronghang Hu hrh11@mails.tsinghua.edu.cn
Tsinghua University
Kate Saenko saenko@eecs.berkeley.edu
UMass Lowell

A major challenge in scaling object detection is the difficulty of obtaining labeled images for large numbers of categories. Recently, deep convolutional neural networks (CNNs) have emerged as clear winners on object classification benchmarks, in part due to training with 1.2M+ labeled classification images. Unfortunately, only a small fraction of those labels are available for the detection task. It is much cheaper and easier to collect large quantities of image-level labels from search engines than it is to collect detection data and label it with precise bounding boxes. In this paper, we propose Large Scale Detection through Adaptation (LSDA), an algorithm which learns the difference between the two tasks and transfers this knowledge to classifiers for categories without bounding box annotated data, turning them into detectors. Our method has the potential to enable detection for the tens of thousands of categories that lack bounding box annotations, yet have plenty of classification data. Evaluation on the ImageNet LSVRC-2013 detection challenge demonstrates the efficacy of our approach. This algorithm enables us to produce a >7.6K detector by using available classification data from leaf nodes in the ImageNet tree. We additionally demonstrate how to modify our architecture to produce a fast detector (running at 2fps for the 7.6K detector).

T90 Local Decorrelation For Improved Pedestrian Detection

Woonhyun Nam whnam82@gmail.com
StradVision
Piotr Dollar pdollar@gmail.com
Facebook AI Research
Joon Hee Han joonhan@postech.ac.kr
POSTECH

Even with the advent of more sophisticated, data-hungry methods, boosted decision trees remain extraordinarily successful for fast rigid object detection, achieving top accuracy on numerous datasets. While effective, most boosted detectors use decision trees with orthogonal (single feature) splits, and the topology of the resulting decision boundary may not be well matched to the natural topology of the data. Given highly correlated data, decision trees with oblique (multiple feature) splits can be effective. Use of oblique splits, however, comes at considerable computational expense. Inspired by recent work on discriminative decorrelation of HOG features, we instead propose an efficient feature transform that removes correlations in local neighborhoods. The result is an overcomplete but locally decorrelated representation ideally suited for use with orthogonal decision trees. In fact, orthogonal trees with our locally decorrelated features outperform oblique trees trained over the original features at a fraction of the computational cost. The overall improvement in accuracy is dramatic: on the Caltech Pedestrian Dataset, we reduce false positives nearly tenfold over the previous state-of-the-art.

T91 Learning Deep Features for Scene Recognition using Places Database

Bolei Zhou bolei@mit.edu
Antonio Torralba fernd@mit.edu
Aude Oliva oliva@mit.edu
Massachusetts Institute of Technology
Agata Lapedriza alapedriza@uoc.edu
MIT / Universitat Oberta de Catalunya
Jianxiong Xiao xj@princeton.edu
Princeton University

Scene recognition is one of the hallmark tasks of computer vision, allowing definition of a context for object recognition. Whereas the tremendous recent progress in object recognition tasks is due to the availability of large datasets like ImageNet and the rise of Convolutional Neural Networks (CNNs) for learning high-level features, performance at scene recognition has not attained the same level of success. This may be because current deep features trained from ImageNet are not competitive enough for such tasks. Here, we introduce a new scene-centric database called Places with over 7 million labeled pictures of scenes. We propose new methods to compare the density and diversity of image datasets and show that Places is as dense as other scene datasets and has more diversity. Using CNN, we learn deep features for scene recognition tasks, and establish new state-of-the-art results on several scene-centric datasets. A visualization of the CNN layers' responses allows us to show differences in the internal representations of object-centric and scene-centric networks.

T92 Do Convnets Learn Correspondence?

Jon Long jonlong@cs.berkeley.edu
Ning Zhang nzhang@eecs.berkeley.edu
Trevor Darrell trevor@eecs.berkeley.edu
UC Berkeley

Convolutional neural nets (convnets) trained from massive labeled datasets have substantially improved the state-of-the-art in image classification and object detection. However, visual understanding requires establishing correspondence on a finer level than object category. Given their large pooling regions and training from whole-image labels, it is not clear that convnets derive their success from an accurate correspondence model which could be used for precise localization. In this paper, we study the effectiveness of convnet activation features for tasks requiring correspondence. We present evidence that convnet features localize at a much finer scale than their receptive field sizes, that they can be used to perform intraclass aligment as well as conventional hand-engineered features, and that they outperform conventional features in keypoint prediction on objects from PASCAL VOC 2011.

T93 Semi-supervised Learning with Deep Generative Models

Diederik Kingma dpkingma@gmail.com
Max Welling welling.max@gmail.com
University of Amsterdam
Shakir Mohamed shakir@google.com
Danilo Jimenez Rezende danilor@google.com
Google Deepmind

The ever-increasing size of modern data sets combined with the difficulty of obtaining label information has made semi-supervised learning one of the problems of significant practical importance in modern data analysis. We revisit the approach to semi-supervised learning with generative models and develop new models that allow for effective generalisation from small labelled data sets to large unlabelled ones. Generative approaches have thus far been either inflexible, inefficient or non-scalable. We show that deep generative models and approximate Bayesian inference exploiting recent advances in variational methods can be used to provide significant improvements, making generative approaches highly competitive for semi-supervised learning.

T94 Deep Learning for Real-Time Atari Game Play Using Offline Monte-Carlo Tree Search Planning

Xiaoxiao Guo guoxiao@umich.edu
Satinder Singh baveja@umich.edu
Honglak Lee honglak@eecs.umich.edu
Richard Lewis rickl@umich.edu
Xiaoshi Wang xiaoshiw@umich.edu
University of Michigan

The combination of modern Reinforcement Learning and Deep Learning approaches holds the promise of making significant progress on challenging applications requiring both rich perception and policy-selection. The Arcade Learning Environment (ALE) provides a set of Atari games that represent a useful benchmark set of such applications. A recent breakthrough in combining model-free reinforcement learning with deep learning, called DQN, achieves the best real-time agents thus far. Planning-based approaches achieve far higher scores than the best model-free approaches, but they exploit information that is not available to human players, and they are orders of magnitude slower than needed for real-time play. Our main goal in this work is to build a better real-time Atari game playing agent than DQN. The central idea is to use the slow planning-based agents to provide training data for a deep-learning architecture capable of real-time play. We proposed new agents based on this idea and show that they outperform DQN.

T95 On the Number of Linear Regions of Deep Neural Networks

Guido Montufar montufar@mis.mpg.de
Max Planck Institute for Mathematics in the Sciences
Razvan Pascanu r.pascanu@gmail.com
Kyunghyun Cho kyunghyun.cho@umontreal.ca
Yoshua Bengio bengioy@iro.umontreal.ca
University of Montreal

We study the complexity of functions computable by deep feedforward neural networks with piecewise linear activations in terms of the symmetries and the number of linear regions that they have. Deep networks are able to sequentially map portions of each layer's input-space to the same output. In this way, deep models compute functions that react equally to complicated patterns of different inputs. The compositional structure of these functions enables them to re-use pieces of computation exponentially often in terms of the network's depth. This paper investigates the complexity of such compositional maps and contributes new theoretical results regarding the advantage of depth for neural networks with piecewise linear activation functions. In particular, our analysis is not specific to a single family of models, and as an example, we employ it for rectifier and maxout networks. We improve complexity bounds from pre-existing work and investigate the behavior of units in higher layers.

T96 Generative Adversarial Nets

Ian Goodfellow goodfellow@google.com
Google
Jean Pouget-Abadie jeanpougetabadie@g.harvard.edu
Harvard University
Mehdi Mirza mirzamom@iro.umontreal.ca
Aaron Courville Aaron.Courville@ umontreal.ca
Yoshua Bengio bengioy@iro.umontreal.ca
David Warde-Farley wardefar@iro.umontreal.ca
University of Montreal
Bing Xu antinucleon@gmail.com
University of Alberta
Sherjil Ozair sherjilozair@gmail.com
Indian Institute of Technology Delhi

We propose a new framework for estimating generative models via adversarial nets, in which we simultaneously train two models: a generative model G that captures the data distribution, and a discriminative model D that estimates the probability that a sample came from the training data rather than G. The training procedure for G is to maximize the probability of D making a mistake. This framework corresponds to a minimax two-player game. In the space of arbitrary functions G and D, a unique solution exists, with G recovering the training data distribution and D equal to 1/2 everywhere. In the case where G and D are defined by multilayer perceptrons, the entire system can be trained with backpropagation. There is no need for any Markov chains or unrolled approximate inference networks during either training or generation of samples. Experiments demonstrate the potential of the framework through qualitative and quantitatively evaluation of the generated samples.

T97 Deep Symmetry Networks

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

The chief difficulty in object recognition is that objects' classes are obscured by a large number of extraneous sources of variability, such as pose and part deformation. These sources of variation can be represented by symmetry groups, sets of composable transformations that preserve object identity. Convolutional neural networks (convnets) achieve a degree of translational invariance by computing feature maps over the translation group, but cannot handle other groups. As a result, these groups' effects have to be approximated by small translations, which often requires augmenting datasets and leads to high sample complexity. In this paper, we introduce deep symmetry networks (symnets), a generalization of convnets that forms feature maps over arbitrary symmetry groups. Symnets use kernel-based interpolation to tractably tie parameters and pool over symmetry spaces of any dimension. Like convnets, they are trained with backpropagation. The composition of feature transformations through the layers of a symnet provides a new approach to deep learning. Experiments on NORB and MNIST-rot show that symnets over the affine group greatly reduce sample complexity relative to convnets by better capturing the symmetries in the data.

T98 Sequence to Sequence Learning with Neural Networks

Ilya Sutskever ilyasu@google.com
Quoc V. Le qvl@google.com
Oriol Vinyals oriol18@gmail.com
Google Research

Deep Neural Networks (DNNs) are powerful models that have achieved excellent performance on difficult learning tasks. Although DNNs work well whenever large labeled training sets are available, they cannot be used to map sequences to sequences. In this paper, we present a general end-to-end approach to sequence learning that makes minimal assumptions on the sequence structure. Our method uses a multilayered Long Short-Term Memory (LSTM) to map the input sequence to a vector of a fixed dimensionality, and then another deep LSTM to decode the target sequence from the vector. Our main result is that on an English to French translation task from the WMT-14 dataset, the translations produced by the LSTM achieve a BLEU score of 34.8 on the entire test set, where the LSTM's BLEU score was penalized on out-of-vocabulary words. Additionally, the LSTM did not have difficulty on long sentences. For comparison, a phrase-based SMT system achieves a BLEU score of 33.3 on the same dataset. When we used the LSTM to rerank the 1000 hypotheses produced by the aforementioned SMT system, its BLEU score increases to 36.5, which is close to the previous state of the art. The LSTM also learned sensible phrase and sentence representations that are sensitive to word order and are relatively invariant to the active and the passive voice. Finally, we found that reversing the order of the words in all source sentences (but not target sentences) improved the LSTM's performance markedly, because doing so introduced many short term dependencies between the source and the target sentence which made the optimization problem easier.

T99 Two-Stream Convolutional Networks for Action Recognition in Videos

Karen Simonyan karen@robots.ox.ac.uk
University of Oxford / Google DeepMind
Andrew Zisserman az@robots.ox.ac.uk
University of Oxford

We investigate architectures of discriminatively trained deep Convolutional Networks (ConvNets) for action recognition in video. The challenge is to capture the complementary information on appearance from still frames and motion between frames. We also aim to generalise the best performing hand-crafted features within a data-driven learning framework. Our contribution is three-fold. First, we propose a two-stream ConvNet architecture which incorporates spatial and temporal networks. Second, we demonstrate that a ConvNet trained on multi-frame dense optical flow is able to achieve very good performance in spite of limited training data. Finally, we show that multi-task learning, applied to two different action classification datasets, can be used to increase the amount of training data and improve the performance on both. Our architecture is trained and evaluated on the standard video actions benchmarks of UCF-101 and HMDB-51, where it is competitive with the state of the art. It also exceeds by a large margin previous attempts to use deep nets for video classification.

T100 Learning to Discover Efficient Mathematical Identities

Wojciech Zaremba woj.zaremba@gmail.com
New York University
Karol Kurach kkurach@gmail.com
Google Zurich / University of Warsaw
Rob Fergus fergus@cs.nyu.edu
NYU

In this paper we explore how machine learning techniques can be applied to the discovery of efficient mathematical identities. We introduce an attribute grammar framework for representing symbolic expressions. Given a grammar of math operators, we build trees that combine them in different ways, looking for compositions that are analytically equivalent to a target expression but of lower computational complexity. However, as the space of trees grows exponentially with the complexity of the target expression, brute force search is impractical for all but the simplest of expressions. Consequently, we introduce two novel learning approaches that are able to learn from simpler expressions to guide the tree search. The first of these is a simple n-gram model, the other being a recursive neural-network. We show how these approaches enable us to derive complex identities, beyond reach of brute-force search, or human derivation.

T101 Learning Generative Models with Visual Attention

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

Attention has long been proposed by psychologists to be important for efficiently dealing with the massive amounts of sensory stimulus in the neocortex. Inspired by the attention models in visual neuroscience and the need for object-centered data for generative models, we propose a deep-learning based generative framework using attention. The attentional mechanism propagates signals from the region of interest in a scene to an aligned canonical representation for generative modeling. By ignoring scene background clutter, the generative model can concentrate its resources on the object of interest. A convolutional neural net is employed to provide good initializations during posterior inference which uses Hamiltonian Monte Carlo. Upon learning images of faces, our model can robustly attend to the face region of novel test subjects. More importantly, our model can learn generative models of new faces from a novel dataset of large images where the face locations are not known.

T102 A Multiplicative Model for Learning Distributed Text-Based Attribute Representations

Ryan Kiros rkiros@cs.toronto.edu
Richard Zemel zemel@cs.toronto.edu
Russ Salakhutdinov rsalakhu@cs.toronto.edu
University of Toronto

In this paper we propose a general framework for learning distributed representations of attributes: characteristics of text whose representations can be jointly learned with word embeddings. Attributes can correspond to a wide variety of concepts, such as document indicators (to learn sentence vectors), language indicators (to learn distributed language representations), meta-data and side information (such as the age, gender and industry of a blogger) or representations of authors. We describe a third-order model where word context and attribute vectors interact multiplicatively to predict the next word in a sequence. This leads to the notion of conditional word similarity: how meanings of words change when conditioned on different attributes. We perform several experimental tasks including sentiment classification, cross-lingual document classification, and blog authorship attribution. We also qualitatively evaluate conditional word neighbours and attribute-conditioned text generation.

T103 Searching for Higgs Boson Decay Modes with Deep Learning

Peter Sadowski peter.j.sadowski@uci.edu
Daniel Whiteson daniel@uci.edu
Pierre Baldi pfbaldi@ics.uci.edu
University of California Irvine

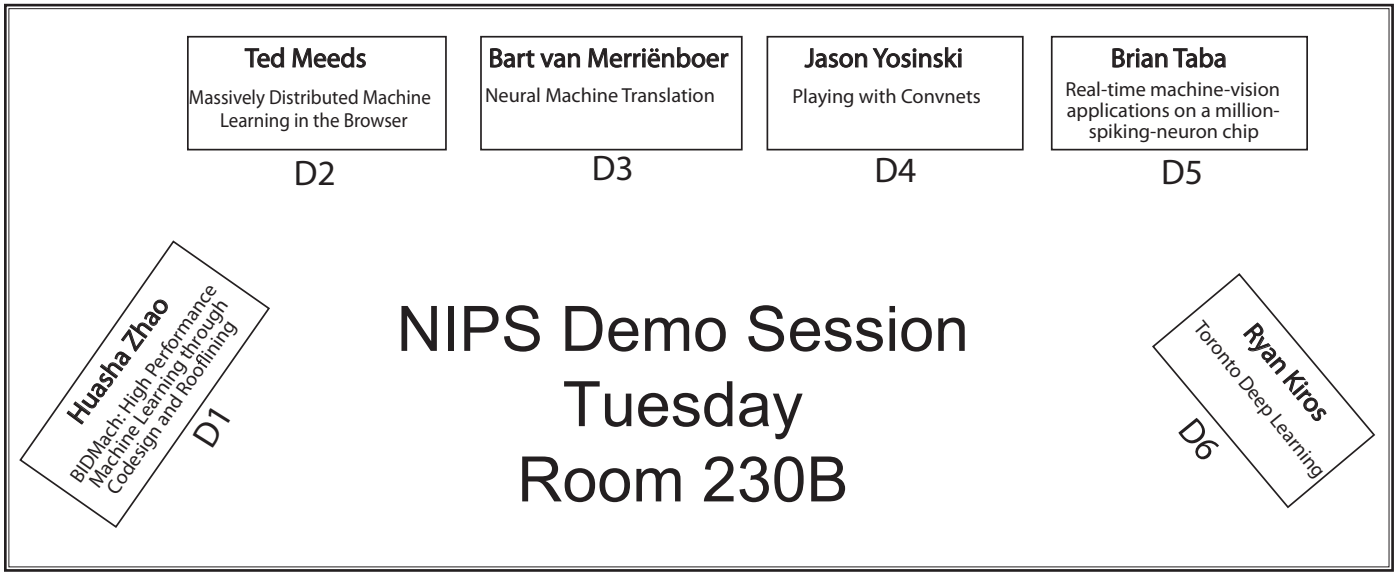
Particle colliders enable us to probe the fundamental nature of matter by observing exotic particles produced by high-energy collisions. Because the experimental measurements from these collisions are necessarily incomplete and imprecise, machine learning algorithms play a major role in the analysis of experimental data. The high-energy physics community typically relies on standardized machine learning software packages for this analysis, and devotes substantial effort towards improving statistical power by hand crafting high-level features derived from the raw collider measurements. In this paper, we train artificial neural networks to detect the decay of the Higgs boson to tau leptons on a dataset of 82 million simulated collision events. We demonstrate that deep neural network architectures are particularly well-suited for this task with the ability to automatically discover high-level features from the data and increase discovery significance.

T104 Convolutional Kernel Networks

Julien Mairal julien.mairal@m4x.org
Piotr Koniusz piotr.koniusz@inria.fr
Zaid Harchaoui zaid.harchaoui@inria.fr
Cordelia Schmid cordelia.schmid@inria.fr
INRIA

An important goal in visual recognition is to devise image representations that are invariant to particular transformations. In this paper, we address this goal with a new type of convolutional neural network (CNN) whose invariance is encoded by a reproducing kernel. Unlike traditional approaches where neural networks are learned either to represent data or for solving a classification task, our network learns to approximate the kernel feature map on training data. Such an approach enjoys several benefits over classical ones. First, by teaching CNNs to be invariant, we obtain simple network architectures that achieve a similar accuracy to more complex ones, while being easy to train and robust to overfitting. Second, we bridge a gap between the neural network literature and kernels, which are natural tools to model invariance. We evaluate our methodology on visual recognition tasks where CNNs have proven to perform well, e.g., digit recognition with the MNIST dataset, and the more challenging CIFAR-10 and STL-10 datasets, where our accuracy is competitive with the state of the art.

DEMONSTRATIONS ABSTRACTS



D1 BIDMach: High Performance Machine Learning through Codesign and Rooflining

Huasha Zhao, Biye Jiang, John Canny
UC Berkeley

Machine learning (ML) is a cornerstone of the new data revolution. Most attempts to scale machine learning to massive datasets focus on parallelization on computer clusters. The BIDMach project instead explores the untapped potential (especially from GPU and SIMD hardware) inside individual machines. Through careful codesign of algorithms and "rooflining", we have demonstrated multiple orders of magnitude speedup over other systems. In fact, BIDMach running on a single machine exceeds the performance of cluster systems on most common ML tasks, and has run computer-intensive tasks on 10-terabyte datasets. We can further show that BIDMach runs at close to the theoretical limits imposed by CPU, memory or network bandwidth. BIDMach includes several innovations to make the data modeling process more agile and effective: likelihood "mixins" and interactive modeling using Gibbs sampling. These results are very encouraging but the greatest potential for future hardware-leveraged machine learning appears to be on MCMC algorithms: Using some old techniques (multiple imputation), we can bring the performance of sample-based Bayesian inference up close to symbolic methods. This opens the possibility for a general-purpose "engine" for machine learning whose performance matches specialized methods. We demonstrate this approach on a specific problem (Latent Dirichlet Allocation), and discuss the general case. Finally we explore scaling ML to clusters. In order to benefit from parallelization, rooflined nodes require very high network bandwidth. We show that the aggregators (reducers) on other systems do not scale, and are not adequate for this task. We describe two new approaches, butterfly mixing and "Kylix" which cover the requirements of machine learning and graph algorithms respectively. We give roofline bounds for both approaches.

D2 Machine Learning in the Browser

Ted Meeds, Remco Hendriks, Said Al Faraby,
Magiel Bruntink, Max Welling
University of Amsterdam

The ubiquity of the browser as a computational engine makes it an ideal platform for the development of massively distributed and collaborative machine learning. This will be demonstrated with our software framework written in Javascript, called MLitB. Our first use case is capable of performing large-scale distributed stochastic gradient descent learning of a DNN using heterogeneous compute devices, including, but not limited to, smart phones and tablets. Demo participants will be able to collaboratively train a DNN for image classification during the poster session and then use the model on their phones. Implications of this software paradigm include cheap, distributed computing on a massive scale, truly collaborative and reproducible machine learning research, publicly accessible machine learning models and algorithms, real privacy preserving computation, field research, green computing, and more. This software is open source and freely available.

D3 Neural Machine Translation

Bart van Merriënboer, Kyunghyun Cho, Yoshua Bengio
Université de Montréal
Dzmitry Bahdanau
Jacobs University Bremen

Neural machine translation is a recently proposed approach to machine translation. Unlike the traditional statistical machine translation, the neural machine translation aims at building a single neural network that can be jointly tuned to maximize the translation performance. In this demo, we present the working system of the state-of-the-art neural machine translation system which implements a novel attention mechanism. This system has achieved a translation performance comparable to the existing state-of-the-art phrase-based system on the task of English-to-French translation.

D4 Playing with Convnets

Jason Yosinski, Hod Lipson
Cornell University

The seminal work by Krizhevsky, Sutskever, and Hinton at NIPS two years ago kicked off a craze of training convnets on ImageNet. These networks have since been trained, improved, tweaked, and adapted to new tasks by those in the field dozens of times over. However, although a stupendous amount of recent progress has been made, the elegant inner workings of these networks are generally hidden behind scalar error metrics. This is a shame. In this demo, we attempt to expose some of these inner workings by allowing people to interact with a video camera hooked to a trained convnet. Audience members can see in real time the activations they cause throughout the network and the classification decisions that are made, which we have found surprisingly fun and informative.

D5 Real-time machine-vision applications on a million-spiking-neuron chip

Paul Merolla, Rodrigo Alvarez-Icaza, Brian Taba,
Dharmendra Modha, Arnon Amir
IBM
Luca Longinotti, Sim Bamford
iniLabs

Interactive examples of spiking-neural-network machine-vision applications running in real-time on IBM's TrueNorth hardware platform and frameless Dynamic Vision Sensors (iniLabs). Each application runs on a single TrueNorth chip that can be configured with a network of up to a million spiking neurons, using the scalable architecture described in Merolla et al. (2014). This is the first demonstration of spiking-neural-network applications running in hardware at this scale and power in real-time.

D6 Toronto Deep Learning

Ryan Kiros, Russ Salakhutdinov, Nitish Srivastava,
Yichuan Tang
University of Toronto

We demonstrate an interactive system for tagging, retrieving and generating sentence descriptions for images. Our models are based on learning a multimodal vector space using deep convolutional networks and long short-term memory (LSTM) recurrent networks for encoding images and sentences. A highly structured multimodal neural language model is used for decoding and generating image descriptions from scratch. Alongside this, we will also showcase a mobile app where a user can take pictures with their phone (such as objects in the demonstration room) and have these images be classified in real time

WEDNESDAY SESSIONS

ORAL SESSION

SESSION 5: - 9:00 – 10:10 AM

Session Chair: Byron M Yu

INVITED TALK: Networks in Climate Science

John Carlos Baez baez@math.ucr.edu
University of California, Riverside

The El Niño is a powerful but irregular climate cycle that has huge consequences for agriculture and perhaps global warming. Predicting its arrival more than 6 months ahead of time has been difficult. A recent paper by Ludescher et al caused a stir by using ideas from network theory to predict the start of an El Niño toward the end of 2014 with a 3-in-4 likelihood. We critically analyze their technique, related applications of network theory, and also attempts to use neural networks to help model the Earth's climate.

John Baez is a professor of mathematics at U.C. Riverside. Until recently he worked on higher category theory and quantum gravity. His internet column "This Week's Finds" dates back to to 1993 and is sometimes called the world's first blog. In 2010, concerned about climate change and the future of the planet, he switched to working on more practical topics and started the Azimuth Project, a collaboration to create a focal point for scientists and engineers interested in environmental issues. One part of this project is developing the mathematics of networks of all kinds: chemical reaction networks, belief networks, signal-flow diagrams, and so on.

Sparse Polynomial Learning and Graph Sketching

Murat Kocaoglu mkocaoglu@utexas.edu
Adam Klivans klivans@cs.utexas.edu
UT Austin
Karthikeyan Shanmugam karthiksh@utexas.edu
Alexandros Dimakis dimakis@austin.utexas.edu
The University of Texas at Austin

Let $f:\{-1,1\}n\rightarrow\mathbb{R}$ be a polynomial with at most s non-zero real coefficients. We give an algorithm for exactly reconstructing f given random examples only from the uniform distribution on $\{-1,1\}$ n that runs in time polynomial in n and $2s$ and succeeds if the function satisfies {unique sign property}: there is one output value which corresponds to a unique set of values of the participating parities. This sufficient condition is satisfied when every coefficient of f is perturbed by a small random noise, or satisfied with high probability when s parity functions are chosen randomly or when all the coefficients are positive. Learning sparse polynomials over the Boolean domain in time polynomial in n and $2s$ is considered a notoriously hard problem in the worst-case. Our result shows that the problem is tractable in some special but generic settings. Then, we show an application of this result to hypergraph sketching which is the problem of learning a sparse (both in the number of hyperedges and the size of the hyperedges) hypergraph from uniformly drawn random cuts. We also provide experimental results on a real world dataset.

SPOTLIGHT SESSION

SESSION 5: 0:10 – 10:30 AM

- **Fast and Robust Least Squares Estimation in Corrupted Linear Models**
  B. McWilliams, G. Krummenacher, M. Lucic, J. Buhmann, ETH Zurich
  See abstract W48

- **A Residual Bootstrap for High-Dimensional Regression with Low-Rank Designs**
  M. Lopes, UC Berkeley
  See abstract W49

- **Spectral Methods meet EM: A Provably Optimal Algorithm for Crowdsourcing**
  Y. Zhang, M. Jordan, UC Berkeley; X. Chen, NYU; D. Zhou, Microsoft Research
  See abstract W95

- **Unsupervised Transcription of Piano Music**
  T. Berg-Kirkpatrick, J. Andreas, D. Klein, UC Berkeley
  See abstract W98

  Non-convex Robust PCA
  P. Netrapalli, P. Jain, Microsoft Research; N. U N, A. Anandkumar, UC Irvine; S. Sanghavi, UT Austin
  See abstract W101

- **Fast Multivariate Spatio-temporal Analysis via Low Rank Tensor Learning**
  Q. Yu, M. Bahadori, U of Southern California; Y. Liu, University of Southern California
  See abstract W103

ORAL SESSION

SESSION 6: 11:00 – 11:40 AM

Session Chair: Mehryar Mohri

Combinatorial Pure Exploration of Multi-Armed Bandits

Shouyuan Chen chenshouyuan@gmail.com
Michael Lyu lyu@cse.cuhk.edu.hk
Irwin King king@cse.cuhk.edu.hk
Chinese University of Hong Kong
Tian Lin lint10@mails.tsinghua.edu.cn
Tsinghua University
Wei Chen weic@microsoft.com
Microsoft Research

We study the *combinatorial pure exploration* (CPE)} problem in the stochastic multi-armed bandit setting, where a learner explores a set of arms with the objective of identifying the optimal member of a **decision class**, which is a collection of subsets of arms with certain combinatorial structures such as size-K subsets, matchings, spanning trees or paths, etc. The CPE problem represents a rich class of pure exploration tasks which covers not only many existing models but also novel cases where the object of interest has a non-trivial combinatorial structure. In this paper, we provide a

series of results for the general CPE problem. We present general learning algorithms which work for all decision classes that admit offline maximization oracles in both fixed confidence and fixed budget settings. We prove problem-dependent upper bounds of our algorithms. Our analysis exploits the combinatorial structures of the decision classes and introduces a new analytic tool. We also establish a general problem-dependent lower bound for the CPE problem. Our results show that the proposed algorithms achieve the optimal sample complexity (within logarithmic factors) for many decision classes. In addition, applying our results back to the problems of top-K arms identification and multiple bandit best arms identification, we recover the best available upper bounds up to constant factors and partially resolve a conjecture on the lower bounds.

From Stochastic Mixability to Fast Rates

Nishant Mehta nishant.mehta@anu.edu.au
Australian National University
Robert Williamson Bob.Williamson@anu.edu.au
NICTA

Empirical risk minimization (ERM) is a fundamental algorithm for statistical learning problems where the data is generated according to some unknown distribution P and returns a hypothesis f chosen from a fixed class $\mathcal{F}$ with small loss ℓ. In the parametric setting, depending upon $(\ell,\mathcal{F},\mathrm{P})$ ERM can have slow $(1/n\sqrt{})$ or fast $(1/n)$ rates of convergence of the excess risk as a function of the sample size n. There exist several results that give sufficient conditions for fast rates in terms of joint properties of ℓ, $\mathcal{F}$, and P, such as the margin condition and the Bernstein condition. In the non-statistical prediction with experts setting, there is an analogous slow and fast rate phenomenon, and it is entirely characterized in terms of the **mixability** of the loss ℓ (there being no role there for $\mathcal{F}$ or P). The notion of **stochastic mixability** builds a bridge between these two models of learning, reducing to classical mixability in a special case. The present paper presents a direct proof of fast rates for ERM in terms of stochastic mixability of $(\ell,\mathcal{F},\mathrm{P})$, and in so doing provides new insight into the fast-rates phenomenon. The proof exploits an old result of Kemperman on the solution to the generalized moment problem. We also show a partial converse that suggests a characterization of fast rates for ERM in terms of stochastic mixability is possible.

SPOTLIGHT SESSION

SESSION 6: 11:40 AM – 12:05 PM

- **Beyond Disagreement-Based Agnostic Active Learning**
  C. Zhang, University of California San Diego; K. Chaudhuri, UC San Diego
  See abstract W11

- **Exploiting easy data in online optimization**
  A. Sani, G. Neu, INRIA; A. Lazaric, INRIA Lille-Nord Europe
  See abstract W13

- **Optimal Regret Minimization in Posted-Price Auctions with Strategic Buyers**
  M. Mohri, Courant Institute, NYU & Google; A. Munoz, Courant Institute of Mathematical Sciences
  See abstract W15

- **Learning Mixtures of Ranking Models**
  P. Awasthi, Princeton University; A. Blum, CMU; O. Sheffet, A. Vijayaraghavan, Carnegie Mellon University
  See abstract W45

- **The limits of squared Euclidean distance regularization**
  M. Derezinski, University of California, Santa Cruz; M. Warmuth, Univ. of Calif. at Santa Cruz
  See abstract W46

- **Rates of Convergence for Nearest Neighbor Classification**
  K. Chaudhuri, S. Dasgupta, UC San Diego
  See abstract W75

ORAL SESSION

SESSION 7: 2:00 – 3:30 PM

Session Chair: Geoffrey J Gordon

POSNER LECTURE: Using the Emergent Dynamics of Attractor Networks for Computation

John J. Hopfield hopfield@Princeton.edu
Princeton University

In higher animals such as mammals, complex collective behaviors emerge from the microscopic properties of large structured ensembles of neurons. I will describe a model example of emergent computational dynamics, based on old-brain cortical properties. This collective (or emergent) description is derivable from the dynamical activity of neurons but has a completely different mathematical structure from the underlying neural network dynamics. The utility of understanding collective dynamics will first be illustrated by showing how it generates a natural solution to the 'time-warp' problem that occurs in recognizing time-varying stimulus patterns having a substantial variation in cadence (e.g. spoken words). The model of emergent dynamics will be shown to be capable of producing goal-directed motor behavior, object-based attention, and rudimentary thinking.

BA Swarthmore 1954; PhD Cornell (theoretical physics) 1958. Member of technical staff Bell Laboratories 1958-1960 & 1973-1996; Faculty positions at UCBerkeley (physics) 1961-1964, Princeton Univ. (physics) 1964-1980, Caltech (chemistry and biology) 1980-1996, Princeton Univ. (molecular biology) 1997-2008, Institute for Advanced Study (2010-2013), now emeritus at Princeton Neuroscience Institute. Served as Chairman of the Faculty, Caltech; President of the American Physical Society; Executive Officer for Computation and Neural Systems, Caltech. Honors include Buckley Prize in Solid State Physics; APS prize in biophysics; Dirac Medal; Einstein Award; MacArthur Fellow; IEEE Rosenblatt Award; Swartz Prize in Computational Neuroscience. Member, National Academy of Science; American Philosophical Society. Research on the interaction of light with solids 1956-1970; biomolecular physics and kinetic proofreading 1970-1980; neural network dynamics and neurobiology 1980-.

"How hard is my MDP?" The distribution-norm to the rescue

Timothy Mann mann@ee.technion.ac.il
Odalric-Ambrym Maillard odalricambrym.maillard@gmail.com
The Technion
Shie Mannor shie@ee.technion.ac.il
Technion

In Reinforcement Learning (RL), state-of-the-art algorithms require a large number of samples per state-action pair to estimate the transition kernel p. In many problems, a good approximation of p is not needed. For instance, if from one state-action pair (s,a), one can only transit to states with the same value, learning $p(\cdot|s,a)$ accurately is irrelevant (only its support matters). This paper aims at capturing such behavior by defining a novel hardness measure for Markov Decision Processes (MDPs) we call the {\em distribution-norm}. The distribution-norm w.r.t.~a measure ν is defined on zero ν-mean functions f by the standard variation of f with respect to ν. We first provide a concentration inequality for the dual of the distribution-norm. This allows us to replace the generic but loose $||\cdot||1$ concentration inequalities used in most previous analysis of RL algorithms, to benefit from this new hardness measure. We then show that several common RL benchmarks have low hardness when measured using the new norm. The distribution-norm captures finer properties than the number of states or the diameter and can be used to assess the difficulty of MDPs.

On Communication Cost of Distributed Statistical Estimation and Dimensionality

Ankit Garg garg@cs.princeton.edu
Princeton
Tengyu Ma tengyu@cs.princeton.edu
Princeton University
Huy Nguyen huylenguyen@gmail.com
University of California, Berkeley

We explore the connection between dimensionality and communication cost in distributed learning problems. Specifically we study the problem of estimating the mean $theta\rightarrow$ of an unknown d dimensional gaussian distribution in the distributed setting. In this problem, the samples from the unknown distribution are distributed among m different machines. The goal is to estimate the mean $theta\rightarrow$ at the optimal minimax rate while communicating as few bits as possible. We show that in this setting, the communication cost scales linearly in the number of dimensions i.e. one needs to deal with different dimensions individually. Applying this result to previous lower bounds for one dimension in the interactive setting by Zhang et al. NIPS'13 and to our improved bounds for the simultaneous setting, we prove new lower bounds of $\Omega(md/\log(m))$ and $\Omega(md)$ for the bits of communication needed to achieve the minimax squared loss, in the interactive and simultaneous settings respectively. To complement, we also demonstrate an interactive protocol achieving the minimax squared loss with $O(md)$ bits of communication. Given the strong lower bounds in the general setting, we initiate the study of the distributed parameter estimation problems with structured parameters. Specifically, when the parameter is known to be s-sparse, we show a protocol achieving the minimax squared loss with high probability and with communication cost proportional to s rather than the dimension d of the ambient space.

SPOTLIGHT SESSION

SESSION 7: 3:30 – 3:50 PM

- **Difference of Convex Functions Programming for Reinforcement Learning**
  B. Piot, M. Geist, SUPELEC; O. Pietquin, Université de Lille
  See abstract W69

- **Near-optimal Reinforcement Learning in Factored MDPs**
  I. Osband, Stanford; B. Van Roy, Stanford University
  See abstract W65

- **Optimizing Energy Production Using Policy Search and Predictive State Representations**
  Y. Grinberg, D. Precup, McGill University; M. Gendreau, CIRRELT
  See abstract W56

- **RAAM: The Benefits of Robustness in Approximating Aggregated MDPs in Reinforcement Learning**
  M. Petrik, IBM Research; D. Subramanian,
  See abstract W68

- **Reducing the Rank in Relational Factorization Models by Including Observable Patterns**
  M. Nickel, Massachusetts Institute of Technology; X. Jiang, Ludwig-Maximilians-Universität München; V. Tresp, Siemens AG
  See abstract W25

- **Trajectory Optimization under Unknown Dynamics for Policy Search**
  S. Levine, UC Berkeley; P. Abbeel, Berkeley
  See abstract W

ORAL SESSION

SESSION 8: 4:20 – 5:40 PM

Session Chair: Sanjiv Kumar

A* Sampling

Chris J Maddison cmaddis@cs.toronto.edu
University of Toronto
Daniel Tarlow dtarlow@cs.toronto.edu
Tom Minka minka@microsoft.com
Microsoft Research

The problem of drawing exact samples from a discrete distribution can be converted into a discrete optimization problem. In this work, we show how sampling from general continuous distributions can be converted into an optimization problem over continuous space. Central to the method is a special case of a stochastic process recently described in mathematical statistics that we call a Gumbel process. We introduce a novel construction of Gumbel processes and A* Sampling, a general sampling algorithm that searches for the maximum of a Gumbel process using A* search. When the global optimum is provably reached, an exact sample is returned. We analyze the convergence time of the algorithms and demonstrate empirically that they make more efficient use of bound and likelihood evaluations than adaptive rejection sampling-based algorithms.

Asynchronous Anytime Sequential Monte Carlo

Brooks Paige brooks@robots.ox.ac.uk
Frank Wood fwood@robots.ox.ac.uk
Arnaud Doucet doucet@stats.ox.ac.uk
Yee Whye Teh y.w.teh@stats.ox.ac.uk
University of Oxford

We introduce a new sequential Monte Carlo algorithm we call the particle cascade. The particle cascade is an asynchronous, anytime alternative to traditional particle filtering algorithms. It uses no barrier-type synchronizations which leads to improved particle throughput and memory efficiency. It is an anytime algorithm in the sense that it can be run forever to emit an unbounded number of particles while keeping within a fixed memory budget. We prove that the particle cascade is an unbiased marginal likelihood estimator which means that it can be straightforwardly plugged into existing pseudomarginal methods.

Probabilistic ODE Solvers with Runge-Kutta Means

Michael Schober michael.schober@tuebingen.mpg.de
MPI for Intelligent Systems
David Duvenaud d duvenaud@seas.harvard.edu
Harvard University
Philipp Hennig phennig@tuebingen.mpg.de
MPI Tübingen

Runge-Kutta methods are the classic family of solvers for ordinary differential equations (ODEs), and the basis for the state-of-the art. Like most numerical methods, they return point estimates. We construct a family of probabilistic numerical methods that instead return a Gauss-Markov process defining a probability distribution over the ODE solution. In contrast to prior work, we construct this family such that posterior means match the outputs of the Runge-Kutta family exactly, thus inheriting their proven good properties. Remaining degrees of freedom not identified by the match to Runge-Kutta are chosen such that the posterior probability measure fits the observed structure of the ODE. Our results shed light on the structure of Runge-Kutta solvers from a new direction, provide a richer, probabilistic output, have low computational cost, and raise new research questions.

A Wild Bootstrap for Degenerate Kernel Tests

Kacper Chwialkowski kacper.chwialkowski@gmail.com
University College London
Dino Sejdinovic dino.sejdinovic@gmail.com
University of Oxford
Arthur Gretton arthur.gretton@gmail.com
UCL

A wild bootstrap method for nonparametric hypothesis tests based on kernel distribution embeddings is proposed. This bootstrap method is used to construct provably consistent tests that apply to random processes, for which the naive permutation-based bootstrap fails. It applies to a large group of kernel tests based on V-statistics, which are degenerate under the null hypothesis, and non-degenerate elsewhere. To illustrate this approach, we construct a two-sample test, an instantaneous independence test and a multiple lag independence test for time series. In experiments, the wild bootstrap gives strong performance on synthetic examples, on audio data, and in performance benchmarking for the Gibbs sampler.

SPOTLIGHT SESSION

SESSION 8: 5:40 – 6:05 PM

- **Log-Hilbert-Schmidt metric between positive definite operators on Hilbert spaces**
  M. Ha Quang, Istituto Italiano di Tecnologia; M. San Biagio, Istituto Italiano di Tecnologia; V. Murino, Italian Institute of Technology
  See abstract W27

- **Extended and Unscented Gaussian Processes**
  D. Steinberg, NICTA; E. Bonilla, The University of New South Wales
  See abstract W33

- **(Almost) No Label No Cry**
  Tree-structured Gaussian Process Approximations
  T. Bui, University of Cambridge; R. Turner, neuroscience
  See abstract W34

- G. Patrini, Australian National University / NICTA; R. Nock, P. Rivera, ; T. Caetano, NICTA Canberra
  See abstract ,W47

- **Hamming Ball Auxiliary Sampling for Factorial Hidden Markov Models**
  M. Titsias, Athens University of Economics and Business; C. Yau, University of Oxford
  See abstract W62

- **Robust Classification Under Sample Selection Bias**
  A. Liu, University of Illinois at Chicago; B. Ziebart, Univ. of Illinois at Chicago
  See abstract W77

- **Consistent Binary Classification with Generalized Performance Metrics**
  N. Natarajan, P. Ravikumar, UT Austin; O. Koyejo, Stanford University; I. Dhillon, University of Texas
  See abstract W80

DEMONSTRATIONS

Demonstrations, 7:00 – 11:59 PM

A 3D Simulator for Evaluating Reinforcement and Imitation Learning Algorithms on Complex Tasks, L. Lefakis, F. Fleuret

A Visual and Interactive IDE for Probabilistic Programming, S. Singh, L. Hewitt, T. Rocktäschel, S. Riedel

ICE: Interactive Classification and Entity Extraction, P. Simard, M. Chickering, A. Lakshmiratan, C. Garcia Jurado Suarez, S. Amershi, J. Verwey, J. Suh

Learning for Tactile Manipulation, T. Hermans, F. Veiga, J. Hölscher, H. van Hoof, J. Peters

SmartWheeler – A smart robotic wheelchair platform, M. Gerdzhev, J. Pineau, A. Leigh, A. Sutcliffe

Unsupervised Transcription of Piano Music, T. Berg-Kirkpatrick, J. Andreas, D. Klein

POSTER SESSION

POSTERS 7:00 – 11:59 PM

W1 **Best-Arm Identification in Linear Bandits**
M. Soare, A. Lazaric, R. Munos

W2 **Combinatorial Pure Exploration of Multi-Armed Bandits**
S. Chen, T. Lin, I. King, M. Lyu, W. Chen

W3 **Bounded Regret for Finite-Armed Structured Bandits**
T. Lattimore, R. Munos

W4 **Efficient learning by implicit exploration in bandit problems with side observations**
T. Kocák, G. Neu, M. Valko, R. Munos

W5 **Learning to Optimize via Information-Directed Sampling**
D. Russo, B. Van Roy

W6 **Bayesian Inference for Structured Spike and Slab Priors**
M. Andersen, O. Winther, L. Hansen

W7 **Estimation with Norm Regularization**
A. Banerjee, S. Chen, F. Fazayeli, V. Sivakumar

W8 **Efficient Sampling for Learning Sparse Additive Models in High Dimensions**
H. Tyagi, B. Gärtner, A. Krause

W9 **Deterministic Symmetric Positive Semidefinite Matrix Completion**
W. Bishop, B. Yu

W10 **Sparse Polynomial Learning and Graph Sketching**
M. Kocaoglu, K. Shanmugam, A. Dimakis, A. Klivans

W11 **Beyond Disagreement-Based Agnostic Active Learning**
C. Zhang, K. Chaudhuri

W12 **Active Regression by Stratification**
S. Sabato, R. Munos

W13 **Exploiting easy data in online optimization**
A. Sani, G. Neu, A. Lazaric

W14 **A Drifting-Games Analysis for Online Learning and Applications to Boosting**
H. Luo, R. Schapire

W15 **Optimal Regret Minimization in Posted-Price Auctions with Strategic Buyers**
M. Mohri, A. Munoz

W16 **Distance-Based Network Recovery under Feature Correlation**
D. Adametz, V. Roth

W17 **Decomposing Parameter Estimation Problems**
K. Refaat, A. Choi, A. Darwiche

W18 **Global Sensitivity Analysis for MAP Inference in Graphical Models**
J. De Bock, C. de Campos, A. Antonucci

W19 **Multi-scale Graphical Models for Spatio-Temporal Processes**
f. janoos, H. Denli, N. Subrahmanya

W20 **Active Learning and Best-Response Dynamics**
M. Balcan, C. Berlind, A. Blum, E. Cohen, K. Patnaik, L. Song

W21 **Provable Tensor Factorization with Missing Data**
P. Jain, S. Oh

W22 **Generalized Higher-Order Orthogonal Iteration for Tensor Decomposition and Completion**
Y. Liu, F. Shang, W. Fan, J. Cheng, H. Cheng

W23 **Neural Word Embedding as Implicit Matrix Factorization**
O. Levy, Y. Goldberg

W24 **Scaling-up Importance Sampling for Markov Logic Networks**
D. Venugopal, V. Gogate

W25 **Reducing the Rank in Relational Factorization Models by Including Observable Patterns**
M. Nickel, X. Jiang, V. Tresp

W26 **Sparse Random Feature Algorithm as Coordinate Descent in Hilbert Space**
E. Yen, T. Lin, S. Lin, P. Ravikumar, I. Dhillon

W27 **Log-Hilbert-Schmidt metric between positive definite operators on Hilbert spaces**
M. Ha Quang, M. San Biagio, V. Murino

W28 **Latent Support Measure Machines for Bag-of-Words Data Classification**
Y. Yoshikawa, T. Iwata, H. Sawada

W29 **A Wild Bootstrap for Degenerate Kernel Tests**
K. Chwialkowski, D. Sejdinovic, A. Gretton

W30 **Fast Prediction for Large-Scale Kernel Machines**
C. Hsieh, S. Si, I. Dhillon

W31 **Testing Unfaithful Gaussian Graphical Models**
D. Soh, S. Tatikonda

W32 **Probabilistic ODE Solvers with Runge-Kutta Means**
M. Schober, D. Duvenaud, P. Hennig

W33 **Extended and Unscented Gaussian Processes**
D. Steinberg, E. Bonilla

W34 **Tree-structured Gaussian Process Approximations**
T. Bui, R. Turner

W35 **Sampling for Inference in Probabilistic Models with Fast Bayesian Quadrature**
T. Gunter, M. Osborne, R. Garnett, P. Hennig, S. Roberts

W36 Do Deep Nets Really Need to be Deep?
J. Ba, R. Caruana

W37 How transferable are features in deep neural networks?
J. Yosinski, J. Clune, Y. Bengio, H. Lipson

W38 Deep Convolutional Neural Network for Image Deconvolution
L. Xu, J. Ren, C. Liu, J. Jia

W39 Identifying and attacking the saddle point problem in high-dimensional non-convex optimization
Y. Dauphin, R. Pascanu, C. Gulcehre, K. Cho, S. Ganguli, Y. Bengio

W40 Learning with Pseudo-Ensembles
P. Bachman, O. Alsharif, D. Precup

W41 On the Information Theoretic Limits of Learning Ising Models
R. Tandon, K. Shanmugam, P. Ravikumar, A. Dimakis

W42 A Probabilistic Framework for Multimodal Retrieval using Integrative Indian Buffet Process
B. Ozdemir, L. Davis

W43 Multivariate f-divergence Estimation With Confidence
K. Moon, A. Hero

W44 On Communication Cost of Distributed Statistical Estimation and Dimensionality
A. Garg, T. Ma, H. Nguyen

W45 Learning Mixtures of Ranking Models
P. Awasthi, A. Blum, O. Sheffet, A. Vijayaraghavan

W46 The limits of squared Euclidean distance regularization
M. Derezinski, M. Warmuth

W47 (Almost) No Label No Cry
G. Patrini, R. Nock, T. Caetano, P. Rivera

W48 Fast and Robust Least Squares Estimation in Corrupted Linear Models
B. McWilliams, G. Krummenacher, M. Lucic, J. Buhmann

W49 A Residual Bootstrap for High-Dimensional Regression with Near Low-Rank Designs
M. Lopes

W50 Parallel Double Greedy Submodular Maximization
X. Pan, S. Jegelka, J. Gonzalez, J. Bradley, M. Jordan

W51 From MAP to Marginals: Variational Inference in Bayesian Submodular Models
J. Djolonga, A. Krause

W52 Stochastic Network Design in Bidirected Trees
x. wu, D. Sheldon, S. Zilberstein

W53 Constrained convex minimization via model-based excessive gap
Q. Tran-Dinh, V. Cevher

W54 Learning to Search in Branch and Bound Algorithms
H. He, H. Daume III, J. Eisner

W55 Convex Deep Learning via Normalized Kernels
Ö. Aslan, X. Zhang, D. Schuurmans

W56 Optimizing Energy Production Using Policy Search and Predictive State Representations
Y. Grinberg, D. Precup, M. Gendreau

W57 A Block-Coordinate Descent Approach for Large-scale Sparse Inverse Covariance Estimation
E. Treister, J. Turek

W58 New Rules for Domain Independent Lifted MAP Inference
H. Mittal, P. Goyal, V. Gogate, P. Singla

W59 A* Sampling
C. Maddison, D. Tarlow, T. Minka

W60 An Integer Polynomial Programming Based Framework for Lifted MAP Inference
S. Sarkhel, D. Venugopal, P. Singla, V. Gogate

W61 Positive Curvature and Hamiltonian Monte Carlo
C. Seiler, S. Rubinstein-Salzedo, S. Holmes

W62 Hamming Ball Auxiliary Sampling for Factorial Hidden Markov Models
M. Titsias, C. Yau

W63 Asynchronous Anytime Sequential Monte Carlo
B. Paige, F. Wood, A. Doucet, Y. Teh

W64 Bayes-Adaptive Simulation-based Search with Value Function Approximation
A. Guez, N. Heess, D. Silver, P. Dayan

W65 Near-optimal Reinforcement Learning in Factored MDPs
I. Osband, B. Van Roy

W66 Learning Neural Network Policies with Guided Policy Search under Unknown Dynamics
S. Levine, P. Abbeel

W67 "How hard is my MDP?" The distribution-norm to the rescue
O. Maillard, T. Mann, S. Mannor

W68 RAAM: The Benefits of Robustness in Approximating Aggregated MDPs in Reinforcement Learning
M. Petrik, D. Subramanian

W69 Difference of Convex Functions Programming for Reinforcement Learning
B. Piot, M. Geist, O. Pietquin

W70 Altitude Training: Strong Bounds for Single-Layer Dropout
S. Wager, W. Fithian, S. Wang, P. Liang

W71 Simultaneous Model Selection and Optimization through Parameter-free Stochastic Learning
F. Orabona

W72 From Stochastic Mixability to Fast Rates
N. Mehta, R. Williamson

W73 On the Statistical Consistency of Plug-in Classifiers for Non-decomposable Performance Measures
H. Narasimhan, R. Vaish, S. Agarwal

W74 Exponential Concentration of a Density Functional Estimator
S. Singh, B. Poczos

W75 Rates of Convergence for Nearest Neighbor Classification
K. Chaudhuri, S. Dasgupta

W76 Deconvolution of High Dimensional Mixtures via Boosting, with Application to Diffusion-Weighted MRI of Human Brain
C. Zheng, F. Pestilli, A. Rokem

W77 Robust Classification Under Sample Selection Bias
A. Liu, B. Ziebart

W78 Bayesian Nonlinear Support Vector Machines and Discriminative Factor Modeling
R. Henao, X. Yuan, L. Carin

W79 Optimizing F-Measures by Cost-Sensitive Classification
S. Puthiya Parambath, N. Usunier, Y. Grandvalet

W80 Consistent Binary Classification with Generalized Performance Metrics
O. Koyejo, N. Natarajan, P. Ravikumar, I. Dhillon

W81 Analysis of Learning from Positive and Unlabeled Data
M. du Plessis, G. Niu, M. Sugiyama

W82 Feature Cross-Substitution in Adversarial Classification
B. Li, Y. Vorobeychik

W83 Large-Margin Convex Polytope Machine
A. Kantchelian, M. Tschantz, L. Huang, P. Bartlett, A. Joseph, J. Tygar

W84 A Boosting Framework on Grounds of Online Learning
T. Naghibi Mohamadpoor, B. Pfister

W85 Multi-Resolution Cascades for Multiclass Object Detection
M. Saberian, N. Vasconcelos

W86 Multi-Class Deep Boosting
V. Kuznetsov, M. Mohri, U. Syed

W87 Robust Logistic Regression and Classification
J. Feng, H. Xu, S. Mannor, S. Yan

W88 Spectral Methods for Indian Buffet Process Inference
H. Tung, A. Smola

W89 Spectral Methods for Supervised Topic Models
Y. Wang, J. Zhu

W90 Spectral Learning of Mixture of Hidden Markov Models
C. Subakan, J. Traa, P. Smaragdis

W91 Multi-Scale Spectral Decomposition of Massive Graphs
S. Si, D. Shin, I. Dhillon, B. Parlett

W93 Spectral Clustering of graphs with the Bethe Hessian
A. Saade, F. Krzakala, L. Zdeborova

W94 Permutation Diffusion Maps (PDM) with Application to the Image Association Problem in Computer Vision
D. Pachauri, R. Kondor, G. Sargur, V. Singh

W95 Spectral Methods meet EM: A Provably Optimal Algorithm for Crowdsourcing
Y. Zhang, X. Chen, D. Zhou, M. Jordan

W96 Low-Rank Time-Frequency Synthesis
C. Févotte, M. Kowalski

W97 A State-Space Model for Decoding Auditory Attentional Modulation from MEG in a Competing-Speaker Environment
S. Akram, J. Simon, S. Shamma, B. Babadi

W98 Unsupervised Transcription of Piano Music
T. Berg-Kirkpatrick, J. Andreas, D. Klein

W99 Efficient Structured Matrix Rank Minimization
A. Yu, W. Ma, Y. Yu, J. Carbonell, S. Sra

W100 Efficient Minimax Signal Detection on Graphs
J. Qian, V. Saligrama

W101 Non-convex Robust PCA
P. Netrapalli, N. U N, S. Sanghavi, A. Anandkumar, P. Jain

W102 Signal Aggregate Constraints in Additive Factorial HMMs, with Application to Energy Disaggregation
M. Zhong, N. Goddard, C. Sutton

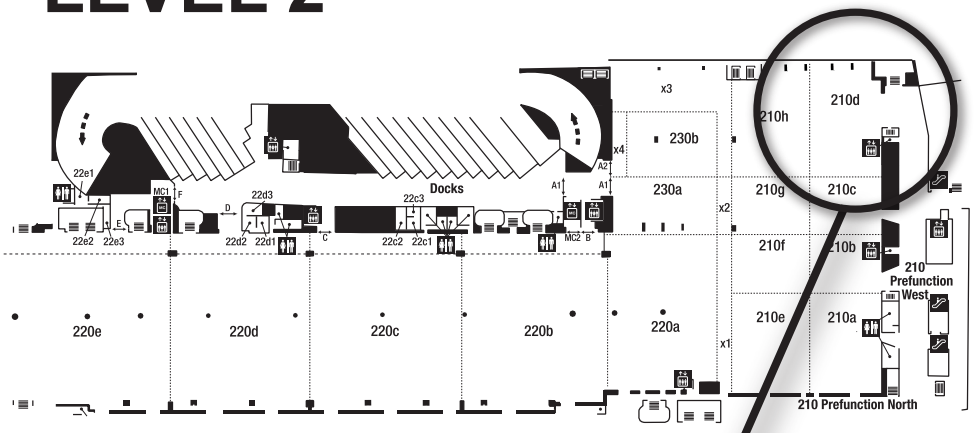
W103 Fast Multivariate Spatio-temporal Analysis via Low Rank Tensor Learning
M. Bahadori, Q. Yu, Y. Liu

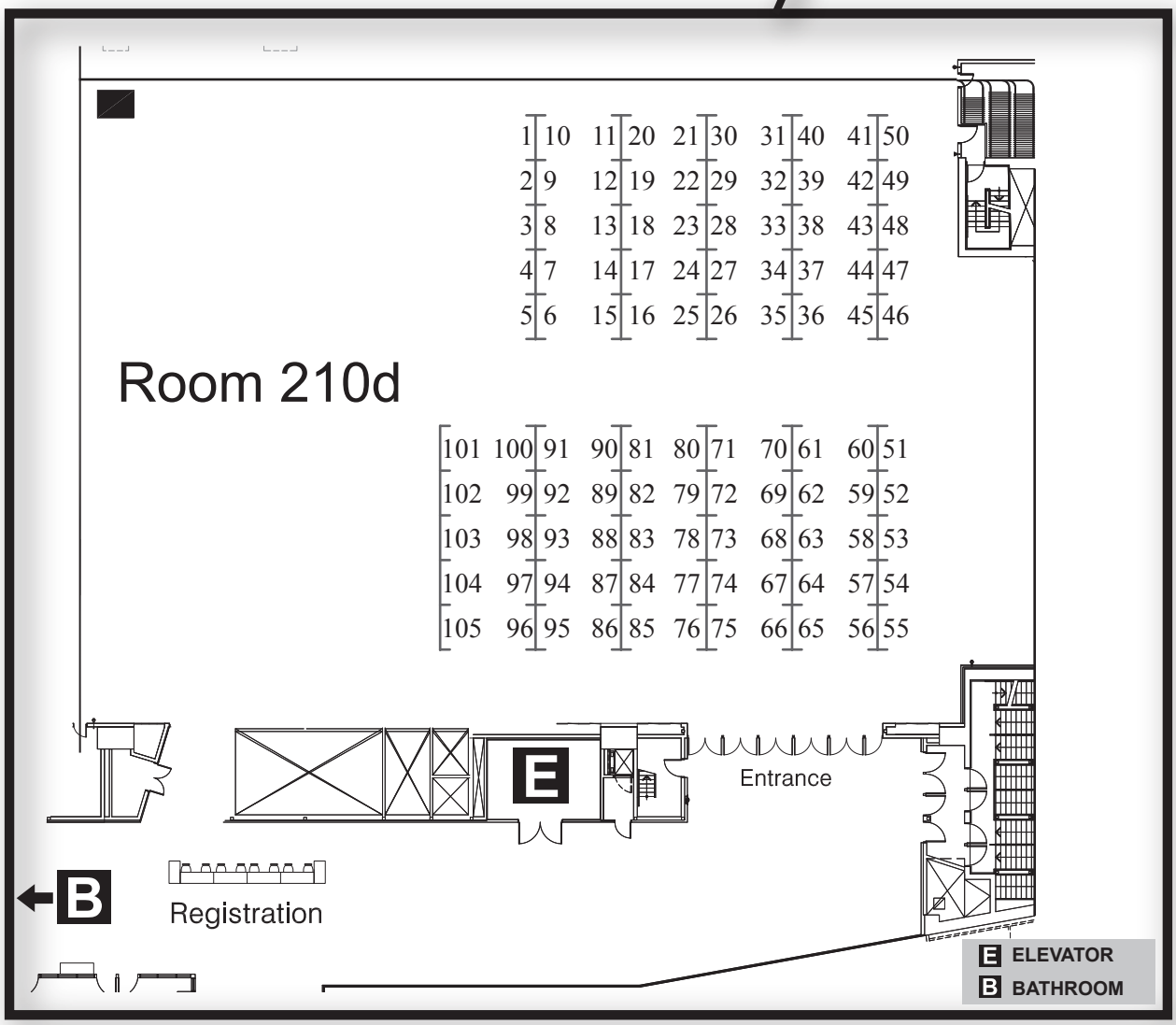
W104 The Blinded Bandit: Learning with Adaptive Feedback
O. Dekel, E. Hazan, T. Koren

W105 Near-optimal sample compression for nearest neighbors
L. Gottlieb, A. Kontorovitch, P. Nisnevitch

WEDNESDAY POSTER FLOORPLAN

LEVEL 2





W1 Best-Arm Identification in Linear Bandits

Marta Soare marta.soare@inria.fr
Alessandro Lazaric alessandro.lazaric@inria.fr
Remi Munos remi.munos@inria.fr
INRIA Lille-Nord Europe

We study the best-arm identification problem in linear bandit, where the rewards of the arms depend linearly on an unknown parameter θ_* and the objective is to return the arm with the largest reward. We characterize the complexity of the problem and introduce sample allocation strategies that pull arms to identify the best arm with a fixed confidence, while minimizing the sample budget. In particular, we show the importance of exploiting the global linear structure to improve the estimate of the reward of near-optimal arms. We analyze the proposed strategies and compare their empirical performance. Finally, as a by-product of our analysis, we point out the connection to the G-optimality criterion used in optimal experimental design.

W2 Combinatorial Pure Exploration of Multi-Armed Bandits

Shouyuan Chen chenshouyuan@gmail.com
Michael Lyu lyu@cse.cuhk.edu.hk
Irwin King king@cse.cuhk.edu.hk
Chinese University of Hong Kong
Tian Lin lint10@mails.tsinghua.edu.cn
Tsinghua University
Wei Chen weic@microsoft.com
Microsoft Research

We study the {\em combinatorial pure exploration (CPE)} problem in the stochastic multi-armed bandit setting, where a learner explores a set of arms with the objective of identifying the optimal member of a **decision class**, which is a collection of subsets of arms with certain combinatorial structures such as size-K subsets, matchings, spanning trees or paths, etc. The CPE problem represents a rich class of pure exploration tasks which covers not only many existing models but also novel cases where the object of interest has a non-trivial combinatorial structure. In this paper, we provide a series of results for the general CPE problem. We present general learning algorithms which work for all decision classes that admit offline maximization oracles in both fixed confidence and fixed budget settings. We prove problem-dependent upper bounds of our algorithms. Our analysis exploits the combinatorial structures of the decision classes and introduces a new analytic tool. We also establish a general problem-dependent lower bound for the CPE problem. Our results show that the proposed algorithms achieve the optimal sample complexity (within logarithmic factors) for many decision classes. In addition, applying our results back to the problems of top-K arms identification and multiple bandit best arms identification, we recover the best available upper bounds up to constant factors and partially resolve a conjecture on the lower bounds.

W3 Bounded Regret for Finite-Armed Structured Bandits

Tor Lattimore tor.lattimore@gmail.com
University of Alberta
Remi Munos remi.munos@inria.fr
INRIA / MSR

We study a new type of K-armed bandit problem where the expected return of one arm may depend on the returns of other arms. We present a new algorithm for this general class of problems and show that under certain circumstances it is possible to achieve finite expected cumulative regret. We also give problem-dependent lower bounds on the cumulative regret showing that at least in special cases the new algorithm is nearly optimal.

W4 Efficient learning by implicit exploration in bandit problems with side observations

Tomáš Kocák tomas.kocak@inria.fr
Gergo Neu neu.gergely@gmail.com
Misho Valko michal.valko@inria.fr
Remi Munos remi.munos@inria.fr
INRIA / MSR

We consider online learning problems under a a partial observability model capturing situations where the information conveyed to the learner is between full information and bandit feedback. In the simplest variant, we assume that in addition to its own loss, the learner also gets to observe losses of some other actions. The revealed losses depend on the learner's action and a directed observation system chosen by the environment. For this setting, we propose the first algorithm that enjoys near-optimal regret guarantees without having to know the observation system before selecting its actions. Along similar lines, we also define a new partial information setting that models online combinatorial optimization problems where the feedback received by the learner is between semi-bandit and full feedback. As the predictions of our first algorithm cannot be always computed efficiently in this setting, we propose another algorithm with similar properties and with the benefit of always being computationally efficient, at the price of a slightly more complicated tuning mechanism. Both algorithms rely on a novel exploration strategy called implicit exploration, which is shown to be more efficient both computationally and information-theoretically than previously studied exploration strategies for the problem.

W5 Learning to Optimize via Information-Directed Sampling

Dan Russo dan.joseph.russo@gmail.com
Benjamin Van Roy bvr@stanford.edu
Stanford University

We propose information-directed sampling -- a new algorithm for online optimization problems in which a decision-maker must balance between exploration and exploitation while learning from partial feedback. Each action is sampled in a manner that minimizes the ratio between the square of expected single-period regret and a measure of information gain: the mutual information between the optimal action and the next observation. We establish an expected regret bound for information-directed sampling that applies across a very general class of models and scales with the entropy of the optimal action distribution. For the widely studied Bernoulli and linear bandit models, we demonstrate simulation performance surpassing popular approaches, including upper confidence bound algorithms, Thompson sampling, and knowledge gradient. Further, we present simple analytic examples illustrating that information-directed sampling can dramatically outperform upper confidence bound algorithms and Thompson sampling due to the way it measures information gain.

W6 Bayesian Inference for Structured Spike and Slab Priors

Michael Andersen michael.riis@gmail.com
Ole Winther owi@imm.dtu.dk
Lars Hansen lkh@imm.dtu.dk
Technical University of Denmark

Sparse signal recovery addresses the problem of solving underdetermined linear inverse problems subject to a sparsity constraint. We propose a novel prior formulation, the structured spike and slab prior, which allows to incorporate a priori knowledge of the sparsity pattern by imposing a spatial Gaussian process on the spike and slab probabilities. Thus, prior information on the structure of the sparsity pattern can be encoded using generic covariance functions. Furthermore, we provide a Bayesian inference scheme for the proposed model based on the expectation propagation framework. Using numerical experiments on synthetic data, we demonstrate the benefits of the model.

W7 Estimation with Norm Regularization

Arindam Banerjee banerjee@cs.umn.edu
Sheng Chen shengc@cs.umn.edu
Vidyashankar Sivakumar sivak017@umn.edu
Farideh Fazayeli farideh@cs.umn.edu
University of Minnesota-Twin Cities

Analysis of estimation error and associated structured statistical recovery based on norm regularized regression, e.g., Lasso, needs to consider four aspects: the norm, the loss function, the design matrix, and the noise vector. This paper presents generalizations of such estimation error analysis on all four aspects, compared to the existing literature. We characterize the restricted error set, establish relations between error sets for the constrained and regularized problems, and present an estimation error bound applicable to {\em any} norm. Precise characterizations of the bound are presented for a variety of noise vectors, design matrices, including sub-Gaussian, anisotropic, and dependent samples, and loss functions, including least squares and generalized linear models. Gaussian widths, as a measure of size of suitable sets, and associated tools play a key role in our generalized analysis.

W8 Efficient Sampling for Learning Sparse Additive Models in High Dimensions

Hemant Tyagi htyagi@inf.ethz.ch
Bernd Gärtner gaertner@inf.ethz.ch
Andreas Krause krausea@ethz.ch
ETH Zurich

We consider the problem of learning sparse additive models, i.e., functions of the form: $f(\vecx)=\sum_{l\in S}\phi_l(x_l)$, $\vecx\in\matR_d$ from point queries of f. Here S is an unknown subset of coordinate variables with $\abs S=k\ll d$. Assuming ϕ_l's to be smooth, we propose a set of points at which to sample f and an efficient randomized algorithm that recovers a \textit{uniform approximation} to each unknown ϕ_l. We provide a rigorous theoretical analysis of our scheme along with sample complexity bounds. Our algorithm utilizes recent results from compressive sensing theory along with a novel convex quadratic program for recovering robust uniform approximations to univariate functions, from point queries corrupted with arbitrary bounded noise. Lastly we theoretically analyze the impact of noise -- either arbitrary but bounded, or stochastic -- on the performance of our algorithm.

W9 Deterministic Symmetric Positive Semidefinite Matrix Completion

William Bishop wbishop@cs.cmu.edu
Byron Yu byronyu@cmu.edu
Carnegie Mellon University

We consider the problem of recovering a symmetric, positive semidefinite (SPSD) matrix from a subset of its entries, possibly corrupted by noise. In contrast to previous matrix recovery work, we drop the assumption of a random sampling of entries in favor of a deterministic sampling of principal submatrices of the matrix. We develop a set of sufficient conditions for the recovery of a SPSD matrix from a set of its principal submatrices, present necessity results based on this set of conditions and develop an algorithm that can exactly recover a matrix when these conditions are met. The proposed algorithm is naturally generalized to the problem of noisy matrix recovery, and we provide a worst-case bound on reconstruction error for this scenario. Finally, we demonstrate the algorithm's utility on noiseless and noisy simulated datasets.

W10 Sparse Polynomial Learning and Graph Sketching

Murat Kocaoglu mkocaoglu@utexas.edu
Adam Klivans klivans@cs.utexas.edu
Karthikeyan Shanmugam karthiksh@utexas.edu
Alexandros Dimakis dimakis@austin.utexas.edu
The University of Texas at Austin

Let f:$\{-1,1\}_n\to\mathbb{R}$ be a polynomial with at most s non-zero real coefficients. We give an algorithm for exactly reconstructing f given random examples from the uniform distribution on $\{-1,1\}_n$ that runs in time polynomial in n and 2_s and succeeds if the function satisfies the \ textit{unique sign property}: there is one output value which corresponds to a unique set of values of the participating parities. This sufficient condition is satisfied when every coefficient of f is perturbed by a small random noise, or satisfied with high probability when s parity functions are chosen randomly or when all the coefficients are positive. Learning sparse polynomials over the Boolean domain in time polynomial in n and 2_s is considered notoriously hard in the worst-case. Our result shows that the problem is tractable for almost all sparse polynomials. Then, we show an application of this result to hypergraph sketching which is the problem of learning a sparse (both in the number of hyperedges and the size of the hyperedges) hypergraph from uniformly drawn random cuts. We also provide experimental results on a real world dataset.

W11 Beyond Disagreement-Based Agnostic Active Learning

Chicheng Zhang chichengzhang@ucsd.edu
Kamalika Chaudhuri kamalika@cs.ucsd.edu
UC San Diego

We study agnostic active learning, where the goal is to learn a classifier in a pre-specified hypothesis class interactively with as few label queries as possible, while making no assumptions on the true function generating the labels. The main algorithms for this problem are {\ em{disagreement-based active learning}}, which has a high label requirement, and {\em{margin-based active learning}}, which only applies to fairly restricted settings. A major challenge is to find an algorithm which achieves better label complexity, is consistent in an agnostic setting, and applies to general classification problems. In this paper, we provide such an algorithm. Our solution is based on two novel contributions -- a reduction from consistent active learning to confidence-rated prediction with guaranteed error, and a novel confidence-rated predictor.

W12 Active Regression by Stratification

Sivan Sabato sivan.sabato@gmail.com
Ben Gurion University
Remi Munos remi.munos@inria.fr
INRIA / MSR

We propose a new active learning algorithm for parametric linear regression with random design. We provide finite sample convergence guarantees for general distributions in the misspecified model. This is the first active learner for this setting that provably can improve over passive learning. Unlike other learning settings (such as classification), in regression the passive learning rate of O(1/epsilon) cannot in general be improved upon. Nonetheless, the so-called `constant' in the rate of convergence, which is characterized by a distribution-dependent risk, can be improved in many cases. For a given distribution, achieving the optimal risk requires prior knowledge of the distribution. Following the stratification technique advocated in Monte-Carlo function integration, our active learner approaches a the optimal risk using piecewise constant approximations.

W13 Exploiting easy data in online optimization

Amir Sani reachme@amirsani.com
Gergo Neu neu.gergely@gmail.com
Alessandro Lazaric alessandro.lazaric@inria.fr
INRIA Lille-Nord Europe

We consider the problem of online optimization, where a learner chooses a decision from a given decision set and suffers some loss associated with the decision and the state of the environment. The learner's objective is to minimize its cumulative regret against the best fixed decision in hindsight. Over the past few decades numerous variants have been considered, with many algorithms designed to achieve sub-linear regret in the worst case. However, this level of robustness comes at a cost. Proposed algorithms are often over-conservative, failing to adapt to the actual complexity of the loss sequence which is often far from the worst case. In this paper we introduce a general algorithm that, provided with a safe learning algorithm and an opportunistic benchmark, can effectively combine good worst-case guarantees with much improved performance on easy data. We derive general theoretical bounds on the regret of the proposed algorithm and discuss its implementation in a wide range of applications, notably in the problem of learning with shifting experts (a recent COLT open problem). Finally, we provide numerical simulations in the setting of prediction with expert advice with comparisons to the state of the art.

W14 A Drifting-Games Analysis for Online Learning and Applications to Boosting

Haipeng Luo haipengl@cs.princeton.edu
Princeton University
Robert Schapire schapire@cs.princeton.edu
MIcrosoft Research

We provide a general mechanism to design online learning algorithms based on a minimax analysis within a drifting-games framework. Different online learning settings (Hedge, multi-armed bandit problems and online convex optimization) are studied by converting into various kinds of drifting games. The original minimax analysis for drifting games is then used and generalized by applying a series of relaxations, starting from choosing a convex surrogate of the 0-1 loss function. With different choices of surrogates, we not only recover existing algorithms, but also propose new algorithms that are totally parameter-free and enjoy other useful properties. Moreover, our drifting-games framework naturally allows us to study high probability bounds without resorting to any concentration results, and also a generalized notion of regret that measures how good the algorithm is compared to all but the top small fraction of candidates. Finally, we translate our new Hedge algorithm into a new adaptive boosting algorithm that is computationally faster as shown in experiments, since it ignores a large number of examples on each round.

W15 Optimal Regret Minimization in Posted-Price Auctions with Strategic Buyers

Mehryar Mohri mohri@google.com
Courant Institute, NYU & Google
Andres Munoz amunoz88@gmail.com
Courant Institute of Mathematical Sciences

We study revenue optimization learning algorithms for posted-price auctions with strategic buyers. We analyze a very broad family of monotone regret minimization algorithms for this problem, which includes the previous best known algorithm, and show that no algorithm in that family admits a strategic regret more favorable than $\Omega(T--\sqrt{})$. We then introduce a new algorithm that achieves a strategic regret differing from the lower bound only by a factor in $O(\log T)$, an exponential improvement upon the previous best algorithm. Our new algorithm admits a natural analysis and simpler proofs, and the ideas behind its design are general. We also report the results of empirical evaluations comparing our algorithm with the previous best algorithm and show a consistent exponential improvement in several different scenarios.

W16 Distance-Based Network Recovery under Feature Correlation

David Adametz david.adametz@unibas.ch
Volker Roth volker.roth@unibas.ch
University of Basel

We present an inference method for Gaussian graphical models when only pairwise distances of n objects are observed. Formally, this is a problem of estimating an n x n covariance matrix from the Mahalanobis distances dMH(xi, xj), where object xi lives in a latent feature space. We solve the problem in fully Bayesian fashion by integrating over the Matrix-Normal likelihood and a Matrix-Gamma prior; the resulting Matrix-T posterior enables network recovery even under strongly correlated features. Hereby, we generalize TiWnet, which assumes Euclidean distances with strict feature independence. In spite of the greatly increased flexibility, our model neither loses statistical power nor entails more computational cost. We argue that the extension is highly relevant as it yields significantly better results in both synthetic and real-world experiments, which is successfully demonstrated for a network of biological pathways in cancer patients.

W17 Decomposing Parameter Estimation Problems

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

We propose a technique for decomposing the parameter learning problem in Bayesian networks into independent learning problems. Our technique applies to incomplete datasets and exploits variables that are either hidden or observed in the given dataset. We show empirically that the proposed technique can lead to orders-of-magnitude savings in learning time. We explain, analytically and empirically, the reasons behind our reported savings, and compare the proposed technique to related ones that are sometimes used by inference algorithms.

W18 Global Sensitivity Analysis for MAP Inference in Graphical Models

Jasper De Bock jasper.debock@ugent.be
Ghent University, SYSTeMS
Cassio de Campos cassiopc@acm.org
Queen's University Belfast
Alessandro Antonucci alessandro@idsia.ch
IDSIA

We study the sensitivity of a MAP configuration of a discrete probabilistic graphical model with respect to perturbations of its parameters. These perturbations are global, in the sense that simultaneous perturbations of all the parameters (or any chosen subset of them) are allowed. Our main contribution is an exact algorithm that can check whether the MAP configuration is robust with respect to given perturbations. Its complexity is essentially the same as that of obtaining the MAP configuration itself, so it can be promptly used with minimal effort. We use our algorithm to identify the largest global perturbation that does not induce a change in the MAP configuration, and we successfully apply this robustness measure in two practical scenarios: the prediction of facial action units with posed images and the classification of multiple real public data sets. A strong correlation between the proposed robustness measure and accuracy is verified in both scenarios.

W19 Multi-scale Graphical Models for Spatio-Temporal Processes

firdaus janoos fjanoos@bwh.harvard.edu
Huseyin Denli huseyin.denli@exxonmobil.com
Niranjan Subrahmanya
niranjan.a.subrahmanya@exxonmobil.com
ExxonMobil Research

Learning the dependency structure between spatially distributed observations of a spatio-temporal process is an important problem in many fields such as geology, geophysics, atmospheric sciences, oceanography, etc. . However, estimation of such systems is complicated by the fact that they exhibit dynamics at multiple scales of space and time arising due to a combination of diffusion and convection/advection. As we show, time-series graphical models based on vector auto-regressive processes are inefficient in capturing such multi-scale structure. In this paper, we present a hierarchical graphical model with physically derived priors that better represents the multi-scale character of these dynamical systems. We also propose algorithms to efficiently estimate the interaction structure from data. We demonstrate results on a general class of problems arising in exploration geophysics by discovering graphical structure that is physically meaningful and provide evidence of its advantages over alternative approaches.

W20 Active Learning and Best-Response Dynamics

Maria-Florina Balcan ninamf@cc.gatech.edu
Le Song lsong@cc.gatech.edu
Christopher Berlind cberlind@gatech.edu
Emma Cohen ecohen@gatech.edu
Kaushik Patnaik kpatnaik3@gatech.edu
Georgia Institute of Technology
Avrim Blum avrim@cs.cmu.edu
CMU

We consider a setting in which low-power distributed sensors are each making highly noisy measurements of some unknown target function. A center wants to accurately learn this function by querying a small number of sensors, which ordinarily would be impossible due to the high noise rate. The question we address is whether local communication among sensors, together with natural best-response dynamics in an appropriately-defined game, can denoise the system without destroying the true signal and allow the center to succeed from only a small number of active queries. We prove positive (and negative) results on the denoising power of several natural dynamics, and also show experimentally that when combined with recent agnostic active learning algorithms, this process can achieve low error from very few queries, performing substantially better than active or passive learning without these denoising dynamics as well as passive learning with denoising.

W21 Provable Tensor Factorization with Missing Data

Prateek Jain prajain@microsoft.com
Microsoft Research
Sewoong Oh swoh@illinois.edu
UIUC

We study the problem of low-rank tensor factorization in the presence of missing data. We ask the following question: how many sampled entries do we need, to efficiently and exactly reconstruct a tensor with a low-rank orthogonal decomposition? We propose a novel alternating minimization based method which iteratively refines estimates of the singular vectors. We show that under certain standard assumptions, our method can recover a three-mode $n \times n \times n$ dimensional rank-r tensor exactly from $O(n^{3/2}r^5 \log^4 n)$ randomly sampled entries. In the process of proving this result, we solve two challenging sub-problems for tensors with missing data. First, in analyzing the initialization step, we prove a generalization of a celebrated result by Szemer\'edie et al. on the spectrum of random graphs. Next, we prove global convergence of alternating minimization with a good initialization. Simulations suggest that the dependence of the sample size on dimensionality n is indeed tight.

W22 Generalized Higher-Order Orthogonal Iteration for Tensor Decomposition and Completion

Yuanyuan Liu liuyuanyuan0917@hotmail.com
Fanhua Shang shangfanhua@hotmail.com
James Cheng jcheng@cse.cuhk.edu.hk
Hong Cheng hongc@cs.cmu.edu
The Chinese University of Hong Kong
Wei Fan weifan@us.ibm.com
Huawei Noah's Ark Lab, Hong Kong

Low-rank tensor estimation has been frequently applied in many real-world problems. Despite successful applications, existing Schatten 1-norm minimization (SNM) methods may become very slow or even not applicable for large-scale problems. To address this difficulty, we therefore propose an efficient and scalable core tensor Schatten 1-norm minimization method for simultaneous tensor decomposition and completion, with a much lower computational complexity. We first induce the equivalence relation of Schatten 1-norm of a low-rank tensor and its core tensor. Then the Schatten 1-norm of the core tensor is used to replace that of the whole tensor, which leads to a much smaller-scale matrix SNM problem. Finally, an efficient algorithm with a rank-increasing scheme is developed to solve the proposed problem with a convergence guarantee. Extensive experimental results show that our method is usually more accurate than the state-of-the-art methods, and is orders of magnitude faster.

W23 Neural Word Embedding as Implicit Matrix Factorization

Omer Levy omerlevy@gmail.com
Yoav Goldberg yoav.goldberg@gmail.com
Bar-Ilan University

We analyze skip-gram with negative-sampling (SGNS), a word embedding method introduced by Mikolov et al., and show that it is implicitly factorizing a word-context matrix, whose cells are the pointwise mutual information (PMI) of the respective word and context pairs, shifted by a global constant. We find that another embedding method, NCE, is implicitly factorizing a similar matrix, where each cell is the (shifted) log conditional probability of a word given its context. We show that using a sparse Shifted Positive PMI word-context matrix to represent words improves results on two word similarity tasks and one of two analogy tasks. When dense low-dimensional vectors are preferred, exact factorization with SVD can achieve solutions that are at least as good as SGNS's solutions for word similarity tasks. On analogy questions SGNS remains superior to SVD. We conjecture that this stems from the weighted nature of SGNS's factorization.

W24 Scaling-up Importance Sampling for Markov Logic Networks

Deepak Venugopal dxv021000@utdallas.edu
Vibhav Gogate vgogate@hlt.utdallas.edu
UT Dallas

Markov Logic Networks (MLNs) are weighted first-order logic templates for generating large (ground) Markov networks. Lifted inference algorithms for them bring the power of logical inference to probabilistic inference. These algorithms operate as much as possible at the compact first-order level, grounding or propositionalizing the MLN only as necessary. As a result, lifted inference algorithms can be much more scalable than propositional algorithms that operate directly on the much larger ground network. Unfortunately, existing lifted inference algorithms suffer from two interrelated problems, which severely affects their scalability in practice. First, for most real-world MLNs having complex structure, they are unable to exploit symmetries and end up grounding most atoms (the grounding problem). Second, they suffer from the evidence problem, which arises because evidence breaks symmetries, severely diminishing the power of lifted inference. In this paper, we address both problems by presenting a scalable, lifted importance sampling-based approach that never grounds the full MLN. Specifically, we show how to scale up the two main steps in importance sampling: sampling from the proposal distribution and weight computation. Scalable sampling is achieved by using an informed, easy-to-sample proposal distribution derived from a compressed MLN-representation. Fast weight computation is achieved by only visiting a small subset of the sampled groundings of each formula instead of all of its possible groundings. We show that our new algorithm yields an asymptotically unbiased estimate. Our experiments on several MLNs clearly demonstrate the promise of our approach.

W25 Reducing the Rank in Relational Factorization Models by Including Observable Patterns

Maximilian Nickel mnick@mit.edu
Massachusetts Institute of Technology
Xueyan Jiang jiangxueyan@hotmail.com
Ludwig-Maximilians-Universität München
Volker Tresp volker.tresp@siemens.com
Siemens AG

Tensor factorizations have become popular methods for learning from multi-relational data. In this context, the rank of a factorization is an important parameter that determines runtime as well as generalization ability. To determine conditions under which factorization is an efficient approach for learning from relational data, we derive upper and lower bounds on the rank required to recover adjacency tensors. Based on our findings, we propose a novel additive tensor factorization model for learning from latent and observable patterns in multi-relational data and present a scalable algorithm for computing the factorization. Experimentally, we show that the proposed approach does not only improve the predictive performance over pure latent variable methods but that it also reduces the required rank --- and therefore runtime and memory complexity --- significantly.

W26 Sparse Random Feature Algorithm as Coordinate Descent in Hilbert Space

En-Hsu Yen a061105@gmail.com
UT-Austin
Ting-Wei Lin b97083@csie.ntu.edu.tw
Shou-De Lin sdlin@csie.ntu.edu.tw
National Taiwan University
Pradeep Ravikumar pradeepr@cs.utexas.edu
Inderjit Dhillon inderjit@cs.utexas.edu
University of Texas

In this paper, we propose a Sparse Random Feature algorithm, which learns a sparse non-linear predictor by minimizing an ℓ_1-regularized objective function over the Hilbert Space induced from kernel function. By interpreting the algorithm as Randomized Coordinate Descent in the infinite-dimensional space, we show the proposed approach converges to a solution comparable within $\backslash \mathrm{eps}$-precision to exact kernel method by drawing $O(1/\backslash \mathrm{eps})$ number of random features, contrasted to the $O(1/\backslash \mathrm{eps}_2)$-type convergence achieved by Monte-Carlo analysis in current Random Feature literature. In our experiments, the Sparse Random Feature algorithm obtains sparse solution that requires less memory and prediction time while maintains comparable performance on tasks of regression and classification. In the meantime, as an approximate solver for infinite-dimensional ℓ_1-regularized problem, the randomized approach converges to better solution than Boosting approach when the greedy step of Boosting cannot be performed exactly.

W27 Log-Hilbert-Schmidt metric between positive definite operators on Hilbert spaces

Minh Ha Quang minh.haquang@iit.it
Vittorio Murino vittorio.murino@univr.it
Marco San Biagio marco.sanbiagio@iit.it
Istituto Italiano di Tecnologia

This paper introduces a novel mathematical and computational framework, namely {\it Log-Hilbert-Schmidt metric} between positive definite operators on a Hilbert space. This is a generalization of the Log-Euclidean metric on the Riemannian manifold of positive definite matrices to the infinite-dimensional setting. The general framework is applied in particular to compute distances between covariance operators on a Reproducing Kernel Hilbert Space (RKHS), for which we obtain explicit formulas via the corresponding Gram matrices. Empirically, we apply our formulation to the task of multi-category image classification, where each image is represented by an infinite-dimensional RKHS covariance operator. On several challenging datasets, our method significantly outperforms approaches based on covariance matrices computed directly on the original input features, including those using the Log-Euclidean metric, Stein and Jeffreys divergences, achieving new state of the art results.

W28 Latent Support Measure Machines for Bag-of-Words Data Classification

Yuya Yoshikawa yuya-y@is.naist.jp
Nara Institute of Science and Technology
Tomoharu Iwata iwata.tomoharu@lab.ntt.co.jp
Hiroshi Sawada sawada.hiroshi@lab.ntt.co.jp
NTT Service Evolution Laboratories

In many classification problems, the input is represented as a set of features, e.g., the bag-of-words (BoW) representation of documents. Support vector machines (SVMs) are widely used tools for such classification problems. The performance of the SVMs is generally determined by whether kernel values between data points can be defined properly. However, SVMs for BoW representations have a major weakness in that the co-occurrence of different but semantically similar words cannot be reflected in the kernel calculation. To overcome the weakness, we propose a kernel-based discriminative classifier for BoW data, which we call the latent support measure machine (latent SMM). With the latent SMM, a latent vector is associated with each vocabulary term, and each document is represented as a distribution of the latent vectors for words appearing in the document. To represent the distributions efficiently, we use the kernel embeddings of distributions that hold high order moment information about distributions. Then the latent SMM finds a separating hyperplane that maximizes the margins between distributions of different classes while estimating latent vectors for words to improve the classification performance. In the experiments, we show that the latent SMM achieves state-of-the-art accuracy for BoW text classification, is robust with respect to its own hyper-parameters, and is useful to visualize words.

W29 A Wild Bootstrap for Degenerate Kernel Tests

Chwialkowski Kacper kacper.chwialkowski@gmail.com
University College London
Dino Sejdinovic dino.sejdinovic@gmail.com
University of Oxford
Arthur Gretton arthur.gretton@gmail.com
University College London

A wild bootstrap method for nonparametric hypothesis tests based on kernel distribution embeddings is proposed. This bootstrap method is used to construct provably consistent tests that apply to random processes, for which the naive permutation-based bootstrap fails. It applies to a large group of kernel tests based on V-statistics, which are degenerate under the null hypothesis, and non-degenerate elsewhere. To illustrate this approach, we construct a two-sample test, an instantaneous independence test and a multiple lag independence test for time series. In experiments, the wild bootstrap gives strong performance on synthetic examples, on audio data, and in performance benchmarking for the Gibbs sampler. The code is available at https://github.com/kacperChwialkowski/wildBootstrap.

W30 Fast Prediction for Large-Scale Kernel Machines

Cho-Jui Hsieh cjhsieh@cs.utexas.edu
Si Si ssi@cs.utexas.edu
Inderjit Dhillon inderjit@cs.utexas.edu
University of Texas

Kernel machines such as kernel SVM and kernel ridge regression usually construct high quality models; however, their use in real-world applications remains limited due to the high prediction cost. In this paper, we present two novel insights for improving the prediction efficiency of kernel machines. First, we show that by adding "pseudo landmark points" to the classical Nystr¨om kernel approximation in an elegant way, we can significantly reduce the prediction error without much additional prediction cost. Second, we provide a new theoretical analysis on bounding the error of the solution computed by using Nystr¨om kernel approximation method, and show that the error is related to the weighted kmeans objective function where the weights are given by the model computed from the original kernel. This theoretical insight suggests a new landmark point selection technique for the situation where we have knowledge of the original model. Based on these two insights, we provide a divide-and-conquer framework for improving the prediction speed. First, we divide the whole problem into smaller local subproblems to reduce the problem size. In the second phase, we develop a kernel approximation based fast prediction approach within each subproblem. We apply our algorithm to real world large-scale classification and regression datasets, and show that the proposed algorithm is consistently and significantly better than other competitors. For example, on the Covertype classification problem, in terms of prediction time, our algorithm achieves more than 10000 times speedup over the full kernel SVM, and a two-fold speedup over the state-of-the-art LDKL approach, while obtaining much higher prediction accuracy than LDKL (95.2% vs. 89.53%).

W31 Testing Unfaithful Gaussian Graphical Models

De Wen Soh dewen.soh@yale.edu
Sekhar Tatikonda sekhar.tatikonda@yale.edu
Yale University

The global Markov property for Gaussian graphical models ensures graph separation implies conditional independence. Specifically if a node set S graph separates nodes u and v then X_u is conditionally independent of X_v given X_S. The opposite direction need not be true, that is, $X_u \perp X_v | X_S$ need not imply S is a node separator of u and v. When it does, the relation $X_u \perp X_v | X_S$ is called faithful. In this paper we provide a characterization of faithful relations and then provide an algorithm to test faithfulness based only on knowledge of other conditional relations of the form $X_i \perp X_j | X_S$.

W32 Probabilistic ODE Solvers with Runge-Kutta Means

Michael Schober michael.schober@tuebingen.mpg.de
Philipp Hennig phennig@tuebingen.mpg.de
MPI for Intelligent Systems
David Duvenaud dduvenaud@seas.harvard.edu
Harvard University

Runge-Kutta methods are the classic family of solvers for ordinary differential equations (ODEs), and the basis for the state of the art. Like most numerical methods, they return point estimates. We construct a family of probabilistic numerical methods that instead return a Gauss-Markov process defining a probability distribution over the ODE solution. In contrast to prior work, we construct this family such that posterior means match the outputs of the Runge-Kutta family exactly, thus inheriting their proven good properties. Remaining degrees of freedom not identified by the match to Runge-Kutta are chosen such that the posterior probability measure fits the observed structure of the ODE. Our results shed light on the structure of Runge-Kutta solvers from a new direction, provide a richer, probabilistic output, have low computational cost, and raise new research questions.

W33 Extended and Unscented Gaussian Processes

Daniel Steinberg daniel.steinberg@nicta.com.au
NICTA
Edwin Bonilla edwinbonilla@gmail.com
The University of New South Wales

We present two new methods for inference in Gaussian process (GP) models with general nonlinear likelihoods. Inference is based on a variational framework where a Gaussian posterior is assumed and the likelihood is linearized about the variational posterior mean using either a Taylor series expansion or statistical linearization. We show that the parameter updates obtained by these algorithms are equivalent to the state update equations in the iterative extended and unscented Kalman filters respectively, hence we refer to our algorithms as extended and unscented GPs. The unscented GP treats the likelihood as a 'black-box' by not requiring its derivative for inference, so it also applies to non-differentiable likelihood models. We evaluate the performance of our algorithms on a number of synthetic inversion problems and a binary classification dataset.

W34 Tree-structured Gaussian Process Approximations

Thang Bui tdb40@cam.ac.uk
Richard Turner ret26@cam.ac.uk
University of Cambridge

Gaussian process regression can be accelerated by constructing a small pseudo-dataset to summarise the observed data. This idea sits at the heart of many approximation schemes, but such an approach requires the number of pseudo-datapoints to be scaled with the range of the input space if the accuracy of the approximation is to be maintained. This presents problems in time-series settings or in spatial datasets where large numbers of pseudo-datapoints are required since computation typically scales quadratically with the pseudo-dataset size. In this paper we devise an approximation whose complexity grows linearly with the number of pseudo-datapoints. This is achieved by imposing a tree or chain structure on the pseudo-datapoints and calibrating the approximation using a Kullback-Leibler (KL) minimisation. Inference and learning can then be performed efficiently using the Gaussian belief propagation algorithm. We demonstrate the validity of our approach on a set of challenging regression tasks including missing data imputation for audio and spatial datasets. We trace out the speed-accuracy trade-off for the new method and show that the frontier dominates those obtained from a large number of existing approximation techniques.

W35 Sampling for Inference in Probabilistic Models with Fast Bayesian Quadrature

Tom Gunter tgunter@robots.ox.ac.uk
Michael Osborne mosb@robots.ox.ac.uk
Stephen Roberts sjrob@robots.ox.ac.uk
University of Oxford
Roman Garnett rgarnett@uni-bonn.de
University of Bonn
Philipp Hennig phennig@tuebingen.mpg.de
MPI Tübingen

We propose a novel sampling framework for inference in probabilistic models: an active learning approach that converges more quickly (in wall-clock time) than Markov chain Monte Carlo (MCMC) benchmarks. The central challenge in probabilistic inference is numerical integration, to average over ensembles of models or unknown (hyper-) parameters (for example to compute marginal likelihood or a partition function). MCMC has provided approaches to numerical integration that deliver state-of-the-art inference, but can suffer from sample inefficiency and poor convergence diagnostics. Bayesian quadrature techniques offer a model-based solution to such problems, but their uptake has been hindered by prohibitive computation costs. We introduce a warped model for probabilistic integrands (likelihoods) that are known to be non-negative, permitting a cheap active learning scheme to optimally select sample locations. Our algorithm is demonstrated to offer faster convergence (in seconds) relative to simple Monte Carlo and annealed importance sampling on both synthetic and real-world examples.

W36 Do Deep Nets Really Need to be Deep?

Jimmy Ba jimmy@psi.utoronto.ca
University of Toronto
Rich Caruana rcaruana@microsoft.com
Microsoft

Currently, deep neural networks are the state of the art on problems such as speech recognition and computer vision. In this paper we empirically demonstrate that shallow feed-forward nets can learn the complex functions previously learned by deep nets and achieve accuracies previously only achievable with deep models. Moreover, in some cases the shallow nets can learn these deep functions using the same number of parameters as the original deep models. On the TIMIT phoneme recognition and CIFAR-10 image recognition tasks, shallow nets can be trained that perform similarly to complex, well-engineered, deeper convolutional models.

W37 How transferable are features in deep neural networks?

Jason Yosinski nps.jyo@0sg.net
Hod Lipson hod.lipson@cornell.edu
Cornell University
Jeff Clune jeffclune@uwyo.edu
Yoshua Bengio bengioy@iro.umontreal.ca
University of Montreal

Many deep neural networks trained on natural images exhibit a curious phenomenon in common: on the first layer they learn features similar to Gabor filters and color blobs. Such first-layer features appear not to be specific to a particular dataset or task, but general in that they are applicable to many datasets and tasks. Features must eventually transition from general to specific by the last layer of the network, but this transition has not been studied extensively. In this paper we experimentally quantify the generality versus specificity of neurons in each layer of a deep convolutional neural network and report a few surprising results. Transferability is negatively affected by two distinct issues: (1) the specialization of higher layer neurons to their original task at the expense of performance on the target task, which was expected, and (2) optimization difficulties related to splitting networks between co-adapted neurons, which was not expected. In an example network trained on ImageNet, we demonstrate that either of these two issues may dominate, depending on whether features are transferred from the bottom, middle, or top of the network. We also document that the transferability of features decreases as the distance between the base task and target task increases, but that transferring features even from distant tasks can be better than using random features. A final surprising result is that initializing a network with transferred features from almost any number of layers can produce a boost to generalization that lingers even after fine-tuning to the target dataset.

W38 Deep Convolutional Neural Network for Image Deconvolution

Li Xu xuli@cse.cuhk.edu.hk
Jimmy Ren jimmy.sj.ren@gmail.com
Lenovo Research
Ce Liu celiu@mit.edu
Microsoft Research
Jiaya Jia leojia@cse.cuhk.edu.hk
CUHK

Many fundamental image-related problems involve deconvolution operators. Real blur degradation seldom complies with an deal linear convolution model due to camera noise, saturation, image compression, to name a few. Instead of perfectly modeling outliers, which is rather challenging from a generative model perspective, we develop a deep convolutional neural network to capture the characteristics of degradation. We note directly applying existing deep neural networks does not produce reasonable results. Our solution is to establish the connection between traditional optimization-based schemes and a neural network architecture where a novel, separable structure is introduced as a reliable support for robust deconvolution against artifacts. Our network contains two submodules, both trained in a supervised manner with proper initialization. They yield decent performance on non-blind image deconvolution compared to previous generative-model based methods.

W39 Identifying and attacking the saddle point problem in high-dimensional non-convex optimization

Yann Dauphin dhaemon@gmail.com
Caglar Gulcehre gulcehrc@iro.umontreal.ca
Yoshua Bengio bengioy@iro.umontreal.ca
Razvan Pascanu r.pascanu@gmail.com
Kyunghyun Cho kyunghyun.cho@umontreal.ca
University of Montreal
Surya Ganguli sganguli@stanford.edu
Stanford University

A central challenge to many fields of science and engineering involves minimizing non-convex error functions over continuous, high dimensional spaces. Gradient descent or quasi-Newton methods are almost ubiquitously used to perform such minimizations, and it is often thought that a main source of difficulty for these local methods to find the global minimum is the proliferation of local minima with much higher error than the global minimum. Here we argue, based on results from statistical physics, random matrix theory, neural network theory, and empirical evidence, that a deeper and more profound difficulty originates from the proliferation of saddle points, not local minima, especially in high dimensional problems of practical interest. Such saddle points are surrounded by high error plateaus that can dramatically slow down learning, and give the illusory impression of the existence of a local minimum. Motivated by these arguments, we propose a new approach to second-order optimization, the saddle-free Newton method, that can rapidly escape high dimensional saddle points, unlike gradient descent and quasi-Newton methods. We apply this algorithm to deep or recurrent neural network training, and provide numerical evidence for its superior optimization performance.

W40 Learning with Pseudo-Ensembles

Phil Bachman phil.bachman@gmail.com
Ouais Alsharif ouais.alsharif@gmail.com
Doina Precup dprecup@cs.mcgill.ca
McGill University

We formalize the notion of a pseudo-ensemble, a (possibly infinite) collection of child models spawned from a parent model by perturbing it according to some noise process. E.g., dropout (Hinton et al, 2012) in a deep neural network trains a pseudo-ensemble of child subnetworks generated by randomly masking nodes in the parent network. We examine the relationship of pseudo-ensembles, which involve perturbation in model-space, to standard ensemble methods and existing notions of robustness, which focus on perturbation in observation-space. We present a novel regularizer based on making the behavior of a pseudo-ensemble robust with respect to the noise process generating it. In the fully-supervised setting, our regularizer matches the performance of dropout. But, unlike dropout, our regularizer naturally extends to the semi-supervised setting, where it produces state-of-the-art results. We provide a case study in which we transform the Recursive Neural Tensor Network of (Socher et al, 2013) into a pseudo-ensemble, which significantly improves its performance on a real-world sentiment analysis benchmark.

W41 On the Information Theoretic Limits of Learning Ising Models

Rashish Tandon rashish@cs.utexas.edu
Karthikeyan Shanmugam karthiksh@utexas.edu
Alexandros Dimakis dimakis@austin.utexas.edu
Pradeep Ravikumar pradeepr@cs.utexas.edu
The University of Texas at Austin

We provide a general framework for computing lower-bounds on the sample complexity of recovering the underlying graphs of Ising models, given i.i.d. samples. While there have been recent results for specific graph classes, these involve fairly extensive technical arguments that are specialized to each specific graph class. In contrast, we isolate two key graph-structural ingredients that can then be used to specify sample complexity lower-bounds. Presence of these structural properties makes the graph class hard to learn. We derive corollaries of our main result that not only recover existing recent results, but also provide lower bounds for novel graph classes not considered previously. We also extend our framework to the random graph setting and derive corollaries for Erdos-Renyi graphs in a certain dense setting.

W42 A Probabilistic Framework for Multimodal Retrieval using Integrative Indian Buffet Process

Bahadir Ozdemir ozdemir@cs.umd.edu
Larry Davis lsd@cs.umd.edu
University of Maryland

We propose a multimodal retrieval procedure based on latent feature models. The procedure consists of a nonparametric Bayesian framework for learning underlying semantically meaningful abstract features in a multimodal dataset, a probabilistic retrieval model that allows cross-modal queries and an extension model for relevance feedback. Experiments on two multimodal datasets, PASCAL-Sentence and SUN-Attribute, demonstrate the effectiveness of the proposed retrieval procedure in comparison to the state-of-the-art algorithms for learning binary codes.

W43 Multivariate f-divergence Estimation With Confidence

Kevin Moon krmoon@umich.edu
Alfred Hero hero@eecs.umich.edu
University of Michigan

The problem of f-divergence estimation is important in the fields of machine learning, information theory, and statistics. While several divergence estimators exist, relatively few have known convergence properties. In particular, even for those estimators whose MSE convergence rates are known, the asymptotic distributions are unknown. We establish the asymptotic normality of a recently proposed ensemble estimator of f-divergence between two distributions from a finite number of samples. This estimator has MSE convergence rate of O(1/T), is simple to implement, and performs well in high dimensions. This theory enables us to perform divergence-based inference tasks such as testing equality of pairs of distributions based on empirical samples. We experimentally validate our theoretical results and, as an illustration, use them to empirically bound the best achievable classification error.

W44 On Communication Cost of Distributed Statistical Estimation and Dimensionality

Ankit Garg garg@cs.princeton.edu
Tengyu Ma tengyu@cs.princeton.edu
Princeton University
Huy Nguyen huylenguyen@gmail.com
University of California, Berkeley

We explore the connection between dimensionality and communication cost in distributed learning problems. Specifically we study the problem of estimating the mean $\vectheta$ of an unknown d dimensional gaussian distribution in the distributed setting. In this problem, the samples from the unknown distribution are distributed among m different machines. The goal is to estimate the mean $\vectheta$ at the optimal minimax rate while communicating as few bits as possible. We show that in this setting, the communication cost scales linearly in the number of dimensions i.e. one needs to deal with different dimensions individually. Applying this result to previous lower bounds for one dimension in the interactive setting \cite{ZDJW13} and to our improved bounds for the simultaneous setting, we prove new lower bounds of $\Omega(md/\log(m))$ and $\Omega(md)$ for the bits of communication needed to achieve the minimax squared loss, in the interactive and simultaneous settings respectively. To complement, we also demonstrate an interactive protocol achieving the minimax squared loss with $O(md)$ bits of communication, which improves upon the simple simultaneous protocol by a logarithmic factor. Given the strong lower bounds in the general setting, we initiate the study of the distributed parameter estimation problems with structured parameters. Specifically, when the parameter is promised to be s-sparse, we show a simple thresholding based protocol that achieves the same squared loss while saving a d/s factor of communication. We conjecture that the tradeoff between communication and squared loss demonstrated by this protocol is essentially optimal up to logarithmic factor.

W45 Learning Mixtures of Ranking Models

Pranjal Awasthi pranjal.iitm@gmail.com
Princeton University
Avrim Blum avrim@cs.cmu.edu
Or Sheffet osheffet@cs.cmu.edu
Aravindan Vijayaraghavan aravindv@cs.cmu.edu
Carnegie Mellon University

This work concerns learning probabilistic models for ranking data in a heterogeneous population. The specific problem we study is learning the parameters of a {\em Mallows Mixture Model}. Despite being widely studied, current heuristics for this problem do not have theoretical guarantees and can get stuck in bad local optima. We present the first polynomial time algorithm which provably learns the parameters of a mixture of two Mallows models. A key component of our algorithm is a novel use of tensor decomposition techniques to learn the top-k prefix in both the rankings. Before this work, even the question of {\ em identifiability} in the case of a mixture of two Mallows models was unresolved.

W46 The limits of squared Euclidean distance regularization

Michal Derezinski michal.derezinski@gmail.com
Manfred Warmuth manfred@ucsc.edu
Univ. of Calif. at Santa Cruz

Some of the simplest loss functions considered in Machine Learning are the square loss, the logistic loss and the hinge loss. The most common family of algorithms, including Gradient Descent (GD) with and without Weight Decay, always predict with a linear combination of the past instances. We give a random construction for sets of examples where the target linear weight vector is trivial to learn but any algorithm from the above family is drastically sub-optimal. Our lower bound on the latter algorithms holds even if the algorithms are enhanced with an arbitrary kernel function. This type of result was known for the square loss. However, we develop new techniques that let us prove such hardness results for any loss function satisfying some minimal requirements on the loss function (including the three listed above). We also show that algorithms that regularize with the squared Euclidean distance are easily confused by random features. Finally, we conclude by discussing related open problems regarding feed forward neural networks. We conjecture that our hardness results

hold for any training algorithm that is based on the squared Euclidean distance regularization (i.e. Back-propagation with the Weight Decay heuristic).

W47 (Almost) No Label No Cry

Giorgio Patrini giorgio.patrini@nicta.com.au
Richard Nock richard.nock@nicta.com.au
Paul Rivera paul.rivera@nicta.com.au
Tiberio Caetano tiberio.caetano@nicta.com.au
Australian National University / NICTA

In Learning with Label Proportions (LLP), the objective is to learn a supervised classifier when, instead of labels, only label proportions for bags of observations are known. This setting has broad practical relevance, in particular for privacy preserving data processing. We first show that the mean operator, a statistic which aggregates all labels, is minimally sufficient for the minimization of many proper scoring losses with linear (or kernelized) classifiers without using labels. We provide a fast learning algorithm that estimates the mean operator via a manifold regularizer with guaranteed approximation bounds. Then, we present an iterative learning algorithm that uses this as initialization. We ground this algorithm in Rademacher-style generalization bounds that fit the LLP setting, introducing a generalization of Rademacher complexity and a Label Proportion Complexity measure. This latter algorithm optimizes tractable bounds for the corresponding bag-empirical risk. Experiments are provided on fourteen domains, whose size ranges up to 300K observations. They display that our algorithms are scalable and tend to consistently outperform the state of the art in LLP. Moreover, in many cases, our algorithms compete with or are just percents of AUC away from the Oracle that learns knowing all labels. On the largest domains, half a dozen proportions can suffice, i.e. roughly 40K times less than the total number of labels.

W48 Fast and Robust Least Squares Estimation in Corrupted Linear Models

Brian McWilliams brian.mcwilliams@inf.ethz.ch
Gabriel Krummenacher kgabriel@ethz.ch
Mario Lucic lucic@inf.ethz.ch
Joachim Buhmann jbuhmann@inf.ethz.ch
ETH Zurich

Subsampling methods have been recently proposed to speed up least squares estimation in large scale settings. However, these algorithms are typically not robust to outliers or corruptions in the observed covariates. The concept of influence that was developed for regression diagnostics can be used to detect such corrupted observations as shown in this paper. This property of influence -- for which we also develop a randomized approximation -- motivates our proposed subsampling algorithm for large scale corrupted linear regression which limits the influence of data points since highly influential points contribute most to the residual error. Under a general model of corrupted observations, we show theoretically and empirically on a variety of simulated and real datasets that our algorithm improves over the current state-of-the-art approximation schemes for ordinary least squares.

W49 A Residual Bootstrap for High-Dimensional Regression with Near Low-Rank Designs

Miles Lopes mlopes@stat.berkeley.edu
UC Berkeley

We study the residual bootstrap (RB) method in the context of high-dimensional linear regression. Specifically, we analyze the distributional approximation of linear contrasts $c^\top(\hat{\beta}_\rho - \beta)$, where $\hat{\beta}_\rho$ is a ridge-regression estimator. When regression coefficients are estimated via least squares, classical results show that RB consistently approximates the laws of contrasts, provided that $p \ll n$, where the design matrix is of size $n \times p$. Up to now, relatively little work has considered how additional structure in the linear model may extend the validity of RB to the setting where $p/n \asymp 1$. In this setting, we propose a version of RB that resamples residuals obtained from ridge regression. Our main structural assumption on the design matrix is that it is nearly low rank --- in the sense that its singular values decay according to a power-law profile. Under a few extra technical assumptions, we derive a simple criterion for ensuring that RB consistently approximates the law of a given contrast. We then specialize this result to study confidence intervals for mean response values $X_i^\top \beta$, where $X_i^\top$ is the ith row of the design. More precisely, we show that conditionally on a Gaussian design with near low-rank structure, RB **simultaneously** approximates all of the laws $X_i^\top(\hat{\beta}_\rho - \beta)$, $i = 1, \ldots, n$. This result is also notable as it imposes no sparsity assumptions on β. Furthermore, since our consistency results are formulated in terms of the Mallows (Kantorovich) metric, the existence of a limiting distribution is not required.

W50 Parallel Double Greedy Submodular Maximization

Xinghao Pan xinghao@eecs.berkeley.edu
Stefanie Jegelka stefje@eecs.berkeley.edu
Joseph Gonzalez jegonzal@eecs.berkeley.edu
Michael Jordan jordan@cs.berkeley.edu
Joseph Bradley josephkb@berkeley.edu
University of California, Berkeley

Many machine learning problems can be reduced to the maximization of submodular functions. Although well understood in the serial setting, the parallel maximization of submodular functions remains an open area of research with recent results only addressing monotone functions. The optimal algorithm for maximizing the more general class of non-monotone submodular functions was introduced by Buchbinder et al. and follows a strongly serial double-greedy logic and program analysis. In this work, we propose two methods to parallelize the double-greedy algorithm. The first, coordination-free approach emphasizes speed at the cost of a weaker approximation guarantee. The second, concurrency control approach guarantees a tight 1/2-approximation, at the quantifiable cost of additional coordination and reduced parallelism. As a consequence we explore the trade off space between guaranteed performance and objective optimality. We implement and evaluate both algorithms on multi-core hardware and billion edge graphs, demonstrating both the scalability and tradeoffs of each approach.

W51 From MAP to Marginals: Variational Inference in Bayesian Submodular Models

Josip Djolonga josip@djolonga.com
Andreas Krause krausea@ethz.ch
ETH Zurich

Submodular optimization has found many applications in machine learning and beyond. We carry out the first systematic investigation of inference in probabilistic models defined through submodular functions, generalizing regular pairwise MRFs and Determinantal Point Processes. In particular, we present L-Field, a variational approach to general log-submodular and log-supermodular distributions based on sub- and supergradients. We obtain both lower and upper bounds on the log-partition function, which enables us to compute probability intervals for marginals, conditionals and marginal likelihoods. We also obtain fully factorized approximate posteriors, at the same computational cost as ordinary submodular optimization. Our framework results in convex problems for optimizing over differentials of submodular functions, which we show how to optimally solve. We provide theoretical guarantees of the approximation quality with respect to the curvature of the function. We further establish natural relations between our variational approach and the classical mean-field method. Lastly, we empirically demonstrate the accuracy of our inference scheme on several submodular models.

W52 Stochastic Network Design in Bidirected Trees

xiaojian wu xiaojian@cs.umass.edu
Daniel Sheldon sheldon@cs.umass.edu
Shlomo Zilberstein shlomo@cs.umass.edu
University of Massachusetts Amherst

We investigate the problem of stochastic network design in bidirected trees. In this problem, an underlying phenomenon (e.g., a behavior, rumor, or disease) starts at multiple sources in a tree and spreads in both directions along its edges. Actions can be taken to increase the probability of propagation on edges, and the goal is to maximize the total amount of spread away from all sources. Our main result is a rounded dynamic programming approach that leads to a fully polynomial-time approximation scheme (FPTAS), that is, an algorithm that can find $(1-\varepsilon)$-optimal solutions for any problem instance in time polynomial in the input size and $1/\varepsilon$. Our algorithm outperforms competing approaches on a motivating problem from computational sustainability to remove barriers in river networks to restore the health of aquatic ecosystems.

W53 Constrained convex minimization via model-based excessive gap

Quoc Tran-Dinh quoc.trandinh@epfl.ch
Volkan Cevher volkan.cevher@epfl.ch
EPFL

We introduce a model-based excessive gap technique to analyze first-order primal- dual methods for constrained convex minimization. As a result, we construct first- order primal-dual methods with optimal convergence rates on the primal objec- tive residual and the primal feasibility gap of their iterates separately. Through a dual smoothing and prox-center selection strategy, our framework subsumes the augmented Lagrangian, alternating direction, and dual fast-gradient methods as special cases, where our rates apply.

W54 Learning to Search in Branch and Bound Algorithms

He He hhe@cs.umd.edu
Hal Daume III hal@umiacs.umd.edu
University of Maryland
Jason Eisner jason@cs.jhu.edu
Johns Hopkins University

Branch-and-bound is a widely used method in combinatorial optimization, including mixed integer programming, structured prediction and MAP inference. While most work has been focused on developing problem-specific techniques, little is known about how to systematically design the node searching strategy on a branch-and-bound tree. We address the key challenge of learning an adaptive node searching order for any class of problem solvable by branch-and-bound. Our strategies are learned by imitation learning. We apply our algorithm to linear programming based branch-and-bound for solving mixed integer programs (MIP). We compare our method with one of the fastest open-source solvers, SCIP; and a very efficient commercial solver, Gurobi. We demonstrate that our approach achieves better solutions faster on four MIP libraries.

W55 Convex Deep Learning via Normalized Kernels

Özlem Aslan ozlem@ualberta.ca
Dale Schuurmans dale@cs.ualberta.ca
University of Alberta
Xinhua Zhang xinhua.zhang.cs@gmail.com
NICTA and Australian National University

Deep learning has been a long standing pursuit in machine learning, which until recently was hampered by unreliable training methods before the discovery of improved heuristics for embedded layer training. A complementary research strategy is to develop alternative modeling architectures that admit efficient training methods while expanding the range of representable structures toward deep models. In this paper, we develop a new architecture for nested nonlinearities that allows arbitrarily deep compositions to be trained to global optimality. The approach admits both parametric and nonparametric forms through the use of normalized kernels to represent each latent layer. The outcome is a fully convex formulation that is able to capture compositions of trainable nonlinear layers to arbitrary depth.

W56 Optimizing Energy Production Using Policy Search and Predictive State Representations

Yuri Grinberg yuri.grinberg@mail.mcgill.ca
Doina Precup dprecup@cs.mcgill.ca
McGill University
Michel Gendreau michel.gendreau@cirrelt.ca
CIRRELT

We consider the challenging practical problem of optimizing the power production of a complex of hydroelectric power plants, which involves control over three continuous action variables, uncertainty in the amount of water inflows and a variety of constraints that need to be satisfied. We propose a policy-search-based approach coupled with predictive modelling to address this problem. This approach has some key advantages compared to other alternatives, such as dynamic programming: the policy representation and search algorithm can conveniently incorporate domain knowledge; the resulting policies are easy to interpret,

and the algorithm is naturally parallelizable. Our algorithm obtains a policy which outperforms the solution found by dynamic programming both quantitatively and qualitatively.

W57 A Block-Coordinate Descent Approach for Large-scale Sparse Inverse Covariance Estimation

Eran Treister eran@cs.technion.ac.il
University Of British Columbia
Javier Turek javiert@cs.technion.ac.il
Technion

The sparse inverse covariance estimation problem arises in many statistical applications in machine learning and signal processing. In this problem, the inverse of a covariance matrix of a multivariate normal distribution is estimated, assuming that it is sparse. An ℓ_1 regularized log-determinant optimization problem is typically solved to approximate such matrices. Because of memory limitations, most existing algorithms are unable to handle large scale instances of this problem. In this paper we present a new block-coordinate descent approach for solving the problem for large-scale data sets. Our method treats the sought matrix block-by-block using quadratic approximations, and we show that this approach has advantages over existing methods in several aspects. Numerical experiments on both synthetic and real gene expression data demonstrate that our approach outperforms the existing state of the art methods, especially for large-scale problems.

W58 New Rules for Domain Independent Lifted MAP Inference

Happy Mittal Happy.Mittal@cse.iitd.ac.in
Parag Singla parags@cse.iitd.ac.in
IIT Delhi
Prasoon Goyal prasoongoyal13@gmail.com
New York University
Vibhav Gogate vgogate@hlt.utdallas.edu
UT Dallas

Lifted inference algorithms for probabilistic first-order logic frameworks such as Markov logic networks (MLNs) have received significant attention in recent years. These algorithms use so called lifting rules to identify symmetries in the first-order representation and reduce the inference problem over a large probabilistic model to an inference problem over a much smaller model. In this paper, we present two new lifting rules, which enable fast MAP inference in a large class of MLNs. Our first rule uses the concept of single occurrence equivalence class of logical variables, which we define in the paper. The rule states that the MAP assignment over an MLN can be recovered from a much smaller MLN, in which each logical variable in each single occurrence equivalence class is replaced by a constant (i.e., an object in the domain of the variable). Our second rule states that we can safely remove a subset of formulas from the MLN if all equivalence classes of variables in the remaining MLN are single occurrence and all formulas in the subset are tautology (i.e., evaluate to true) at extremes (i.e., assignments with identical truth value for groundings of a predicate). We prove that our two new rules are sound and demonstrate via a detailed experimental evaluation that our approach is superior in terms of scalability and MAP solution quality to the state of the art approaches.

W59 A* Sampling

Chris J Maddison cmaddis@cs.toronto.edu
University of Toronto
Daniel Tarlow dtarlow@cs.toronto.edu
Tom Minka minka@microsoft.com
Microsoft Research

The problem of drawing samples from a discrete distribution can be converted into a discrete optimization problem. In this work, we show how sampling from a continuous distribution can be converted into an optimization problem over continuous space. Central to the method is a stochastic process recently described in mathematical statistics that we call the Gumbel process. We present a new construction of the Gumbel process and A* sampling, a practical generic sampling algorithm that searches for the maximum of a Gumbel process using A* search. We analyze the correctness and convergence time of A* sampling and demonstrate empirically that it makes more efficient use of bound and likelihood evaluations than the most closely related adaptive rejection sampling-based algorithms.

W60 An Integer Polynomial Programming Based Framework for Lifted MAP Inference

Somdeb Sarkhel sxs104721@utdallas.edu
Deepak Venugopal dxv021000@utdallas.edu
Vibhav Gogate vgogate@hlt.utdallas.edu
University of Texas at Dallas
Parag Singla parags@cse.iitd.ac.in
IIT Delhi

In this paper, we present a new approach for lifted MAP inference in Markov logic networks (MLNs). The key idea in our approach is to compactly encode the MAP inference problem as an Integer Polynomial Program (IPP) by schematically applying three lifted inference steps to the MLN: lifted decomposition, lifted conditioning, and partial grounding. Our IPP encoding is lifted in the sense that an integer assignment to a variable in the IPP may represent a truth-assignment to multiple indistinguishable ground atoms in the MLN. We show how to solve the IPP by first converting it to an Integer Linear Program (ILP) and then solving the latter using state-of-the-art ILP techniques. Experiments on several benchmark MLNs show that our new algorithm is substantially superior to ground inference and existing methods in terms of computational efficiency and solution quality.

W61 Positive Curvature and Hamiltonian Monte Carlo

Christof Seiler christof.seiler@stanford.edu
Simon Rubinstein-Salzedo simonr@stanford.edu
Susan Holmes sp.holmes@gmail.com
Stanford University

The Jacobi metric introduced in mathematical physics can be used to analyze Hamiltonian Monte Carlo (HMC). In a geometrical setting, each step of HMC corresponds to a geodesic on a Riemannian manifold with a Jacobi metric. Our calculation of the sectional curvature of this HMC manifold allows us to see that it is positive in cases such as sampling from a high dimensional multivariate Gaussian. We show that positive curvature can be used to prove theoretical concentration results for HMC Markov chains.

W62 Hamming Ball Auxiliary Sampling for Factorial Hidden Markov Models

Michalis Titsias mtitsias@aueb.gr
Athens University of Economics and Business
Christopher Yau cyau@well.ox.ac.uk
University of Oxford

We introduce a novel sampling algorithm for Markov chain Monte Carlo-based Bayesian inference for factorial hidden Markov models. This algorithm is based on an auxiliary variable construction that restricts the model space allowing iterative exploration in polynomial time. The sampling approach overcomes limitations with common conditional Gibbs samplers that use asymmetric updates and become easily trapped in local modes. Instead, our method uses symmetric moves that allows joint updating of the latent sequences and improves mixing. We illustrate the application of the approach with simulated and a real data example.

W63 Asynchronous Anytime Sequential Monte Carlo

Brooks Paige brooks@robots.ox.ac.uk
Frank Wood fwood@robots.ox.ac.uk
Arnaud Doucet doucet@stats.ox.ac.uk
Yee Whye Teh y.w.teh@stats.ox.ac.uk
University of Oxford

We introduce a new sequential Monte Carlo algorithm we call the particle cascade. The particle cascade is an asynchronous, anytime alternative to traditional sequential Monte Carlo algorithms that is amenable to parallel and distributed implementations. It uses no barrier synchronizations which leads to improved particle throughput and memory efficiency. It is an anytime algorithm in the sense that it can be run forever to emit an unbounded number of particles while keeping within a fixed memory budget. We prove that the particle cascade provides an unbiased marginal likelihood estimator which can be straightforwardly plugged into existing pseudo-marginal methods.

W64 Bayes-Adaptive Simulation-based Search with Value Function Approximation

Arthur Guez aguez@gatsby.ucl.ac.uk
Google DeepMind
Nicolas Heess nheess@gatsby.ucl.ac.uk
David Silver davidstarsilver@gmail.com
Peter Dayan dayan@gatsby.ucl.ac.uk
Gatsby Unit, UCL

Bayes-adaptive planning offers a principled solution to the exploration-exploitation trade-off under model uncertainty. It finds the optimal policy in belief space, which explicitly accounts for the expected effect on future rewards of reductions in uncertainty. However, the Bayes-adaptive solution is typically intractable in domains with large or continuous state spaces. We present a tractable method for approximating the Bayes-adaptive solution by combining simulation-based search with a novel value function approximation technique that generalises over belief space. Our method outperforms prior approaches in both discrete bandit tasks and simple continuous navigation and control tasks.

W65 Near-optimal Reinforcement Learning in Factored MDPs

Ian Osband iosband@stanford.edu
Benjamin Van Roy bvr@stanford.edu
Stanford University

Any reinforcement learning algorithm that applies to all Markov decision processes (MDPs) will suffer $\Omega(SAT\sqrt{})$ regret on some MDP, where $T$is the elapsed time and S and A are the cardinalities of the state and action spaces. This implies $T{=}\Omega(SA)$ time to guarantee a near-optimal policy. In many settings of practical interest, due to the curse of dimensionality, S and A can be so enormous that this learning time is unacceptable. We establish that, if the system is known to be a **factored** MDP, it is possible to achieve regret that scales polynomially in the number of **parameters** encoding the factored MDP, which may be exponentially smaller than S or A. We provide two algorithms that satisfy near-optimal regret bounds in this context: posterior sampling reinforcement learning (PSRL) and an upper confidence bound algorithm (UCRL-Factored).

W66 Learning Neural Network Policies with Guided Policy Search under Unknown Dynamics

Sergey Levine svlevine@eecs.berkeley.edu
UC Berkeley
Pieter Abbeel pabbeel@cs.berkeley.edu
Berkeley

We present a policy search method that uses iteratively refitted local linear models to optimize trajectory distributions for large, continuous problems. These trajectory distributions can be used within the framework of guided policy search to learn policies with an arbitrary parameterization. Our method fits time-varying linear dynamics models to speed up learning, but does not rely on learning a global model, which can be difficult when the dynamics are complex and discontinuous. We show that this hybrid approach requires many fewer samples than model-free methods, and can handle complex, nonsmooth dynamics that can pose a challenge for model-based techniques. We present experiments showing that our method can be used to learn complex neural network policies that successfully execute simulated robotic manipulation tasks in partially observed environments with numerous contact discontinuities and underactuation.

W67 "How hard is my MDP?" The distribution-norm to the rescue

Odalric-Ambrym Maillard
odalricambrym.maillard@gmail.com
Timothy Mann mann@ee.technion.ac.il
Shie Mannor shie@ee.technion.ac.il
Technion

In Reinforcement Learning (RL), state-of-the-art algorithms require a large number of samples per state-action pair to estimate the transition kernel p. In many problems, a good approximation of $p$is not needed. For instance, if from one state-action pair (s,a), one can only transit to states with the same value, learning $p(\cdot|s,a)$ accurately is irrelevant (only its support matters). This paper aims at capturing such behavior by defining a novel hardness measure for Markov Decision Processes (MDPs) we call

the {\em distribution-norm}. The distribution-norm w.r.t.~a measure ν is defined on zeroν-mean functions f by the standard variation of f with respect to ν. We first provide a concentration inequality for the dual of the distribution-norm. This allows us to replace the generic but loose $||\cdot||_1$concentration inequalities used in most previous analysis of RL algorithms, to benefit from this new hardness measure. We then show that several common RL benchmarks have low hardness when measured using the new norm. The distribution-norm captures finer properties than the number of states or the diameter and can be used to assess the difficulty of MDPs.

W68 RAAM: The Benefits of Robustness in Approximating Aggregated MDPs in Reinforcement Learning

Marek Petrik marekpetrik@gmail.com
IBM Research
Dharmashankar Subramanian dharmash@us.ibm.com

We describe how to use robust Markov decision processes for value function approximation with state aggregation. The robustness serves to reduce the sensitivity to the approximation error of sub-optimal policies in comparison to classical methods such as fitted value iteration. This results in reducing the bounds on the gamma-discounted infinite horizon performance loss by a factor of 1/(1-gamma) while preserving polynomial-time computational complexity. Our experimental results show that using the robust representation can significantly improve the solution quality with minimal additional computational cost.

W69 Difference of Convex Functions Programming for Reinforcement Learning

Bilal Piot bilal.piot@inria.fr
Université Lille 3
Matthieu Geist matthieu.geist@supelec.fr
SUPELEC
Olivier Pietquin olivier.pietquin@supelec.fr
Université de Lille 1

Large Markov Decision Processes (MDPs) are usually solved using Approximate Dynamic Programming (ADP) methods such as Approximate Value Iteration (AVI) or Approximate Policy Iteration (API). The main contribution of this paper is to show that, alternatively, the optimal state-action value function can be estimated using Difference of Convex functions (DC) Programming. To do so, we study the minimization of a norm of the Optimal Bellman Residual (OBR) T_*Q-Q, where T_* is the so-called optimal Bellman operator. Controlling this residual allows controlling the distance to the optimal action-value function, and we show that minimizing an empirical norm of the OBR is consistant in the Vapnik sense. Finally, we frame this optimization problem as a DC program. That allows envisioning using the large related literature on DC Programming to address the Reinforcement Leaning (RL) problem.

W70 Altitude Training: Strong Bounds for Single-Layer Dropout

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

Dropout training, originally designed for deep neural networks, has been successful on high-dimensional single-layer natural language tasks. This paper proposes a theoretical explanation for this phenomenon: we show that, under a generative Poisson topic model with long documents, dropout training improves the exponent in the generalization bound for empirical risk minimization. Dropout achieves this gain much like a marathon runner who practices at altitude: once a classifier learns to perform reasonably well on training examples that have been artificially corrupted by dropout, it will do very well on the uncorrupted test set. We also show that, under similar conditions, dropout preserves the Bayes decision boundary and should therefore induce minimal bias in high dimensions.

W71 Simultaneous Model Selection and Optimization through Parameter-free Stochastic Learning

Francesco Orabona francesco@orabona.com
Yahoo! Labs

Stochastic gradient descent algorithms for training linear and kernel predictors are gaining more and more importance, thanks to their scalability. While various methods have been proposed to speed up their convergence, the model selection phase is often ignored. In fact, in theoretical works most of the time assumptions are made, for example, on the prior knowledge of the norm of the optimal solution, while in the practical world validation methods remain the only viable approach. In this paper, we propose a new kernel-based stochastic gradient descent algorithm that performs model selection while training, with no parameters to tune, nor any form of cross-validation. The algorithm builds on recent advancement in online learning theory for unconstrained settings, to estimate over time the right regularization in a data-dependent way. Optimal rates of convergence are proved under standard smoothness assumptions on the target function as well as preliminary empirical results.

W72 From Stochastic Mixability to Fast Rates

Nishant Mehta nishant.mehta@anu.edu.au
Robert Williamson Bob.Williamson@anu.edu.au
Australian National University

Empirical risk minimization (ERM) is a fundamental learning rule for statistical learning problems where the data is generated according to some unknown distribution P and returns a hypothesis $f$chosen from a fixed class $\mathcal{F}$ with small loss ℓ. In the parametric setting, depending upon (ℓ,$\mathcal{F}$,P) ERM can have slow $(1/n--\sqrt{})$ or fast $(1/n)$ rates of convergence of the excess risk as a function of the sample size n. There exist several results that give sufficient conditions for fast rates in terms of joint properties of ℓ, $\mathcal{F}$, and P, such as the margin condition and the Bernstein condition. In the non-statistical prediction with expert advice setting, there is an analogous slow and fast rate phenomenon, and it is entirely characterized in terms of the mixability of the loss ℓ (there being no role there for $\mathcal{F}$ or P). The notion of stochastic mixability builds a bridge between these two models of learning, reducing to classical mixability in a special case. The present paper presents a direct proof of fast rates for ERM in terms of stochastic mixability of (ℓ,$\mathcal{F}$,P), and in so doing provides new insight into the fast-rates phenomenon. The proof exploits an old result of Kemperman on the solution to the general moment problem. We also show a partial converse that suggests a characterization of fast rates for ERM in terms of stochastic mixability is possible.

W73 On the Statistical Consistency of Plug-in Classifiers for Non-decomposable Performance Measures

Harikrishna Narasimhan nhari88@gmail.com
Shivani Agarwal shivani@csa.iisc.ernet.in
Rohit Vaish rohit.vaish@csa.iisc.ernet.in
Indian Insititute of Science

We study consistency properties of algorithms for non-decomposable performance measures that cannot be expressed as a sum of losses on individual data points, such as the F-measure used in text retrieval and several other performance measures used in class imbalanced settings. While there has been much work on designing algorithms for such performance measures, there is limited understanding of the theoretical properties of these algorithms. Recently, Ye et al. (2012) showed consistency results for two algorithms that optimize the F-measure, but their results apply only to an idealized setting, where precise knowledge of the underlying probability distribution (in the form of the `true' posterior class probability) is available to a learning algorithm. In this work, we consider plug-in algorithms that learn a classifier by applying an empirically determined threshold to a suitable `estimate' of the class probability, and provide a general methodology to show consistency of these methods for any non-decomposable measure that can be expressed as a continuous function of true positive rate (TPR) and true negative rate (TNR), and for which the Bayes optimal classifier is the class probability function thresholded suitably. We use this template to derive consistency results for plug-in algorithms for the F-measure and for the geometric mean of TPR and precision; to our knowledge, these are the first such results for these measures. In addition, for continuous distributions, we show consistency of plug-in algorithms for any performance measure that is a continuous and monotonically increasing function of TPR and TNR. Experimental results confirm our theoretical findings.

W74 Exponential Concentration of a Density Functional Estimator

Shashank Singh sss1@andrew.cmu.edu
Barnabas Poczos bapoczos@cs.cmu.edu
Carnegie Mellon University

We analyse a plug-in estimator for a large class of integral functionals of one or more continuous probability densities. This class includes important families of entropy, divergence, mutual information, and their conditional versions. For densities on the d-dimensional unit cube [0,1]^d that lie in a beta-Holder smoothness class, we prove our estimator converges at the rate O(n^(1/(beta+d))). Furthermore, we prove that the estimator obeys an exponential concentration inequality about its mean, whereas most previous related results have bounded only expected error of estimators. Finally, we demonstrate our bounds to the case of conditional Renyi mutual information.

W75 Rates of Convergence for Nearest Neighbor Classification

Kamalika Chaudhuri kamalika@cs.ucsd.edu
Sanjoy Dasgupta dasgupta@cs.ucsd.edu
UC San Diego

We analyze the behavior of nearest neighbor classification in metric spaces and provide finite-sample, distribution-dependent rates of convergence under minimal assumptions. These are more general than existing bounds, and enable us, as a by-product, to establish the universal consistency of nearest neighbor in a broader range of data spaces than was previously known. We illustrate our upper and lower bounds by introducing a new smoothness class customized for nearest neighbor classification. We find, for instance, that under the Tsybakov margin condition the convergence rate of nearest neighbor matches recently established lower bounds for nonparametric classification.

W76 Deconvolution of High Dimensional Mixtures via Boosting, with Application to Diffusion-Weighted MRI of Human Brain

Charles Zheng snarles@stanford.edu
Franco Pestilli frk@stanford.edu
Ariel Rokem arokem@stanford.edu
Stanford University

Diffusion-weighted magnetic resonance imaging (DWI) and fiber tractography are the only methods to measure the structure of the white matter in the living human brain. The diffusion signal has been modelled as the combined contribution from many individual fascicles of nerve fibers passing through each location in the white matter. Typically, this is done via basis pursuit, but estimation of the exact directions is limited due to discretization. The difficulties inherent in modeling DWI data are shared by many other problems involving fitting non-parametric mixture models. Ekanadaham et al. proposed an approach, continuous basis pursuit, to overcome discretization error in the 1-dimensional case (e.g., spike-sorting). Here, we propose a more general algorithm that fits mixture models

of any dimensionality without discretization. Our algorithm uses the principles of L2-boost, together with refitting of the weights and pruning of the parameters. The addition of these steps to L2-boost both accelerates the algorithm and assures its accuracy. We refer to the resulting algorithm as elastic basis pursuit, or EBP, since it expands and contracts the active set of kernels as needed. We show that in contrast to existing approaches to fitting mixtures, our boosting framework (1) enables the selection of the optimal bias-variance tradeoff along the solution path, and (2) scales with high-dimensional problems. In simulations of DWI, we find that EBP yields better parameter estimates than a non-negative least squares (NNLS) approach, or the standard model used in DWI, the tensor model, which serves as the basis for diffusion tensor imaging (DTI). We demonstrate the utility of the method in DWI data acquired in parts of the brain containing crossings of multiple fascicles of nerve fibers.

W77 Robust Classification Under Sample Selection Bias

Anqi Liu aliu33@uic.edu
Brian Ziebart bziebart@uic.edu
Univ. of Illinois at Chicago

In many important machine learning applications, the source distribution used to estimate a probabilistic classifier differs from the target distribution on which the classifier will be used to make predictions. Due to its asymptotic properties, sample-reweighted loss minimization is a commonly employed technique to deal with this difference. However, given finite amounts of labeled source data, this technique suffers from significant estimation errors in settings with large sample selection bias. We develop a framework for robustly learning a probabilistic classifier that adapts to different sample selection biases using a minimax estimation formulation. Our approach requires only accurate estimates of statistics under the source distribution and is otherwise as robust as possible to unknown properties of the conditional label distribution, except when explicit generalization assumptions are incorporated. We demonstrate the behavior and effectiveness of our approach on synthetic and UCI binary classification tasks.

W78 Bayesian Nonlinear Support Vector Machines and Discriminative Factor Modeling

Ricardo Henao r.henao@duke.edu
Xin Yuan xin.yuan@duke.edu
Lawrence Carin lcarin@duke.edu
Duke University

A new Bayesian formulation is developed for nonlinear support vector machines (SVMs), based on a Gaussian process and with the SVM hinge loss expressed as a scaled mixture of normals. We then integrate the Bayesian SVM into a factor model, in which feature learning and nonlinear classifier design are performed jointly; almost all previous work on such discriminative feature learning has assumed a linear classifier. Inference is performed with expectation conditional maximization (ECM) and Markov Chain Monte Carlo (MCMC). An extensive set of experiments demonstrate the utility of using a nonlinear Bayesian SVM within discriminative feature learning and factor modeling, from the standpoints of accuracy and interpretability.

W79 Optimizing F-Measures by Cost-Sensitive Classification

Shameem Puthiya Parambath sputhiya@utc.fr
Nicolas Usunier nicolas.usunier@utc.fr
Yves Grandvalet yves.grandvalet@utc.fr
Université de Technologie de Compiégne

We present a theoretical analysis of F-measures for binary, multiclass and multilabel classification. These performance measures are non-linear, but in many scenarios they are pseudo-linear functions of the per-class false negative/ false positive rate. Based on this observation, we present a general reduction of F-measure maximization to cost-sensitive classification with unknown costs. We then propose an algorithm with provable guarantees to obtain an approximately optimal classifier for the F-measure by solving a series of cost-sensitive classification problems. The strength of our analysis is to be valid on any dataset and any class of classifiers, extending the existing theoretical results on F-measures, which are asymptotic in nature. We present numerical experiments to illustrate the relative importance of cost asymmetry and thresholding when learning linear classifiers on various F-measure optimization tasks.

W80 Consistent Binary Classification with Generalized Performance Metrics

Sanmi Koyejo sanmi.k@gmail.com
Stanford University
Nagarajan Natarajan naga86@cs.utexas.edu
Pradeep Ravikumar pradeepr@cs.utexas.edu
Inderjit Dhillon inderjit@cs.utexas.edu
University of Texas

Performance metrics for binary classification are designed to capture tradeoffs between four fundamental population quantities: true positives, false positives, true negatives and false negatives. Despite significant interest from theoretical and applied communities, little is known about either optimal classifiers or consistent algorithms for optimizing binary classification performance metrics beyond a few special cases. We consider a fairly large family of performance metrics given by ratios of linear combinations of the four fundamental population quantities. This family includes many well known binary classification metrics such as classification accuracy, AM measure, F-measure and the Jaccard similarity coefficient as special cases. Our analysis identifies the optimal classifiers as the sign of the thresholded conditional probability of the positive class, with a performance metric-dependent threshold. The optimal threshold can be constructed using simple plug-in estimators when the performance metric is a linear combination of the population quantities, but alternative techniques are required for the general case. We propose two algorithms for estimating the optimal classifiers, and prove their statistical consistency. Both algorithms are straightforward modifications of standard approaches to address the key challenge of optimal threshold selection, thus are simple to implement in practice. The first algorithm combines a plug-in estimate of the conditional probability of the positive class with optimal threshold selection. The second algorithm leverages recent work on calibrated asymmetric surrogate losses to construct candidate classifiers. We present empirical comparisons between these algorithms on benchmark datasets.

W81 Analysis of Learning from Positive and Unlabeled Data

Marthinus du Plessis christo@sg.cs.titech.ac.jp
Tokyo Institute of Technology
Gang Niu niugang@baidu.com
Baidu, Inc.
Masashi Sugiyama sugi@k.u-tokyo.ac.jp
The University of Tokyo

Learning a classifier from positive and unlabeled data is an important class of classification problems that are conceivable in many practical applications. In this paper, we first show that this problem can be solved by cost-sensitive learning between positive and unlabeled data. We then show that convex surrogate loss functions such as the hinge loss may lead to a wrong classification boundary due to an intrinsic bias, but the problem can be avoided by using non-convex loss functions such as the ramp loss. We next analyze the excess risk when the class prior is estimated from data, and show that the classification accuracy is not sensitive to class prior estimation if the unlabeled data is dominated by the positive data (this is naturally satisfied in inlier-based outlier detection because inliers are dominant in the unlabeled dataset). Finally, we provide generalization error bounds and show that, for an equal number of labeled and unlabeled samples, the generalization error of learning only from positive and unlabeled samples is no worse than $22\sqrt{}$ times the fully supervised case. These theoretical findings are also validated through experiments.

W82 Feature Cross-Substitution in Adversarial Classification

Bo Li bo.li.2@vanderbilt.edu
Yevgeniy Vorobeychik yevgeniy.vorobeychik@vanderbilt.edu
Vanderbilt University

The success of machine learning, particularly in supervised settings, has led to numerous attempts to apply it in adversarial settings such as spam and malware detection. The core challenge in this class of applications is that adversaries are not static data generators, but make a deliberate effort to evade the classifiers deployed to detect them. We investigate both the problem of modeling the objectives of such adversaries, as well as the algorithmic problem of accounting for rational, objective-driven adversaries. In particular, we demonstrate severe shortcomings of feature reduction in adversarial settings using several natural adversarial objective functions, an observation that is particularly pronounced when the adversary is able to substitute across similar features (for example, replace words with synonyms or replace letters in words). We offer a simple heuristic method for making learning more robust to feature cross-substitution attacks. We then present a more general approach based on mixed-integer linear programming with constraint generation, which implicitly trades off overfitting and feature selection in an adversarial setting using a sparse regularizer along with an evasion model. Our approach is the first method for combining an adversarial classification algorithm with a very general class of models of adversarial classifier evasion. We show that our algorithmic approach significantly outperforms state-of-the-art alternatives.

W83 Large-Margin Convex Polytope Machine

Alex Kantchelian akant@eecs.berkeley.edu
Michael Tschantz mct@berkeley.edu
Peter Bartlett bartlett@cs.berkeley.edu
Anthony Joseph adj@cs.berkeley.edu
J. D. Tygar tygar@cs.berkeley.edu
UC Berkeley
Ling Huang ling.huang@intel.com
Intel

We present the Convex Polytope Machine (CPM), a novel non-linear learning algorithm for large-scale binary classification tasks. The CPM finds a large margin convex polytope separator which encloses one class. We develop a stochastic gradient descent based algorithm that is amenable to massive datasets, and augment it with a heuristic procedure to avoid sub-optimal local minima. Our experimental evaluations of the CPM on large-scale datasets from distinct domains (MNIST handwritten digit recognition, text topic, and web security) demonstrate that the CPM trains models faster, sometimes several orders of magnitude, than state-of-the-art similar approaches and kernel-SVM methods while achieving comparable or better classification performance. Our empirical results suggest that, unlike prior similar approaches, we do not need to control the number of sub-classifiers (sides of the polytope) to avoid overfitting.

W84 A Boosting Framework on Grounds of Online Learning

Tofigh Naghibi Mohamadpoor tofigh@gmail.com
Beat Pfister pfister@tik.ee.ethz.ch
ETH Zurich

By exploiting the duality between boosting and online learning, we present a boosting framework which proves to be extremely powerful thanks to employing the vast knowledge available in the online learning area. Using this framework, we develop various algorithms to address multiple practically and theoretically interesting questions including sparse boosting, smooth-distribution boosting, agnostic learning and, as a by-product, some generalization to double-projection online learning algorithms.

W85 Multi-Resolution Cascades for Multiclass Object Detection

Mohammad Saberian mj.saberian@gmail.com
Nuno Vasconcelos nuno@ece.ucsd.edu
UC San Diego

An algorithm for learning fast multiclass object detection cascades is introduced. It produces multi-resolution (MRes) cascades, whose early stages are binary target vs. non-target detectors that eliminate false positives, late stages multiclass classifiers that finely discriminate target classes, and middle stages have intermediate numbers of classes, determined in a data-driven manner. This MRes structure is achieved with a new structurally biased boosting algorithm (SBBoost). SBBost extends previous multiclass boosting approaches, whose boosting mechanisms are shown to implement two complementary data-driven biases: 1) the standard bias towards examples difficult to classify, and 2) a bias towards difficult classes. It is shown that structural biases can be implemented by generalizing this class-

based bias, so as to encourage the desired MRes structure. This is accomplished through a generalized definition of multiclass margin, which includes a set of bias parameters. SBBoost is a boosting algorithm for maximization of this margin. It can also be interpreted as standard multiclass boosting algorithm augmented with margin thresholds or a cost-sensitive boosting algorithm with costs defined by the bias parameters. A stage adaptive bias policy is then introduced to determine bias parameters in a data driven manner. This is shown to produce MRes cascades that have high detection rate and are computationally efficient. Experiments on multiclass object detection show improved performance over previous solutions.

W86 Multi-Class Deep Boosting

Vitaly Kuznetsov vitaly@cims.nyu.edu
Courant Institute
Mehryar Mohri mohri@google.com
Courant Institute, NYU & Google
Umar Syed usyed@google.com
Google Research

We present new ensemble learning algorithms for multi-class classification. Our algorithms can use as a base classifier set a family of deep decision trees or other rich or complex families and yet benefit from strong generalization guarantees. We give new data-dependent learning bounds for convex ensembles in the multi-class classification setting expressed in terms of the Rademacher complexities of the sub-families composing the base classifier set, and the mixture weight assigned to each sub-family. These bounds are finer than existing ones both thanks to an improved dependency on the number of classes and, more crucially, by virtue of a more favorable complexity term expressed as an average of the Rademacher complexities based on the ensemble's mixture weights. We introduce and discuss several new multi-class ensemble algorithms benefiting from these guarantees, prove positive results for the H-consistency of several of them, and report the results of experiments showing that their performance compares favorably with that of multi-class versions of AdaBoost and Logistic Regression and their L1-regularized counterparts.

W87 Robust Logistic Regression and Classification

Jiashi Feng a0066331@nus.edu.sg
UC Berkeley
Shie Mannor shie@ee.technion.ac.il
Technion
Huan Xu mpexuh@nus.edu.sg
Shuicheng Yan eleyans@nus.edu.sg
National University of Singapore

We consider logistic regression with arbitrary outliers in the covariate matrix. We propose a new robust logistic regression algorithm, called RoLR, that estimates the parameter through a simple linear programming procedure. We prove that RoLR is robust to a constant fraction of adversarial outliers. To the best of our knowledge, this is the first result on estimating logistic regression model when the covariate matrix is corrupted with any performance guarantees. Besides regression, we apply RoLR to solving binary classification problems where a fraction of training samples are corrupted.

W88 Spectral Methods for Indian Buffet Process Inference

Hsiao-Yu Tung sfish0101@gmail.com
Alex Smola alex@smola.org
Carnegie Mellon

The Indian Buffet Process is a versatile statistical tool for modeling distributions over binary matrices. We provide an efficient spectral algorithm as an alternative to costly Variational Bayes and sampling-based algorithms. We derive a novel tensorial characterization of the moments of the Indian Buffet Process proper and for two of its applications. We give a computationally efficient iterative inference algorithm, concentration of measure bounds, and reconstruction guarantees. Our algorithm provides superior accuracy and cheaper computation than comparable Variational Bayesian approach on a number of reference problems.

W89 Spectral Methods for Supervised Topic Models

Yining Wang yiningwa@cs.cmu.edu
Carnegie Mellon University
Jun Zhu dcszj@mail.tsinghua.edu.cn
Tsinghua University

Supervised topic models simultaneously model the latent topic structure of large collections of documents and a response variable associated with each document. Existing inference methods are based on either variational approximation or Monte Carlo sampling. This paper presents a novel spectral decomposition algorithm to recover the parameters of supervised latent Dirichlet allocation (sLDA) models. The Spectral-sLDA algorithm is provably correct and computationally efficient. We prove a sample complexity bound and subsequently derive a sufficient condition for the identifiability of sLDA. Thorough experiments on a diverse range of synthetic and real-world datasets verify the theory and demonstrate the practical effectiveness of the algorithm.

W90 Spectral Learning of Mixture of Hidden Markov Models

Cem Subakan csubakan@gmail.com
Johannes Traa traa2@illinois.edu
Paris Smaragdis paris@illinois.edu
University of Illinois Urbana-Champaign

In this paper, we propose a learning approach for the Mixture of Hidden Markov Models (MHMM) based on the Method of Moments (MoM). Computational advantages of MoM make MHMM learning amenable for large data sets. It is not possible to directly learn an MHMM with existing learning approaches, mainly due to a permutation ambiguity in the estimation process. We show that it is possible to resolve this ambiguity using the spectral properties of a global transition matrix even in the presence of estimation noise. We demonstrate the validity of our approach on synthetic and real data.

W91 Multi-Scale Spectral Decomposition of Massive Graphs

Si Si ssi@cs.utexas.edu
Donghyuk Shin dshin@cs.utexas.edu
Inderjit Dhillon inderjit@cs.utexas.edu
University of Texas
Beresford Parlett parlett@math.berkeley.edu
University of California, Berkeley

Computing the k dominant eigenvalues and eigenvectors of massive graphs is a key operation in numerous machine learning applications; however, popular solvers suffer from slow convergence, especially when k is reasonably large. In this paper, we propose and analyze a novel multi-scale spectral decomposition method (MSEIGS), which first clusters the graph into smaller clusters whose spectral decomposition can be computed efficiently and independently. We show theoretically as well as empirically that the union of all cluster's subspaces has significant overlap with the dominant subspace of the original graph, provided that the graph is clustered appropriately. Thus, eigenvectors of the clusters serve as good initializations to a block Lanczos algorithm that is used to compute spectral decomposition of the original graph. We further use hierarchical clustering to speed up the computation and adopt a fast early termination strategy to compute quality approximations. Our method outperforms widely used solvers in terms of convergence speed and approximation quality. Furthermore, our method is naturally parallelizable and exhibits significant speedups in shared-memory parallel settings. For example, on a graph with more than 82 million nodes and 3.6 billion edges, MSEIGS takes less than 3 hours on a single-core machine while Randomized SVD takes more than 6 hours, to obtain a similar approximation of the top-50 eigenvectors. Using 16 cores, we can reduce this time to less than 40 minutes.

W93 Spectral Clustering of graphs with the Bethe Hessian

Alaa Saade alaa.saade89@gmail.com
Florent Krzakala florent.krzakala@ens.fr
Ecole Normale Superieure
Lenka Zdeborova lenka.zdeborova@cea.fr
CEA Saclay and CNRS URA 2306

Spectral clustering is a standard approach to label nodes on a graph by studying the (largest or lowest) eigenvalues of a symmetric real matrix such as e.g. the adjacency or the Laplacian. Recently, it has been argued that using instead a more complicated, non-symmetric and higher dimensional operator, related to the non-backtracking walk on the graph, leads to improved performance in detecting clusters, and even to optimal performance for the stochastic block model. Here, we propose to use instead a simpler object, a symmetric real matrix known as the Bethe Hessian operator, or deformed Laplacian. We show that this approach combines the performances of the non-backtracking operator, thus detecting clusters all the way down to the theoretical limit in the stochastic block model, with the computational, theoretical and memory advantages of real symmetric matrices.

W94 Permutation Diffusion Maps (PDM) with Application to the Image Association Problem in Computer Vision

Deepti Pachauri pachauri@cs.wisc.edu
Vikas Singh vsingh@biostat.wisc.edu
Gautam Sargur gautam@cs.wisc.edu
University of Wisconsin Madison
Risi Kondor risi@uchicago.edu
University of Chicago

Consistently matching keypoints across images, and the related problem of finding clusters of nearby images, are critical components of various tasks in Computer Vision, including Structure from Motion (SfM). Unfortunately, occlusion and large repetitive structures tend to mislead most currently used matching algorithms, leading to characteristic pathologies in the final output. In this paper we introduce a new method, Permutations Diffusion Maps (PDM), to solve the matching problem, as well as a related new affinity measure, derived using ideas from harmonic analysis on the symmetric group. We show that just by using it as a preprocessing step to existing SfM pipelines, PDM can greatly improve reconstruction quality on difficult datasets.

W95 Spectral Methods meet EM: A Provably Optimal Algorithm for Crowdsourcing

Yuchen Zhang yuczhang@eecs.berkeley.edu
Michael Jordan jordan@cs.berkeley.edu
UC Berkeley
Xi Chen xichen@nyu.edu
NYU
Dengyong Zhou dengyong.zhou@microsoft.com
Microsoft Research

The Dawid-Skene estimator has been widely used for inferring the true labels from the noisy labels provided by non-expert crowdsourcing workers. However, since the estimator maximizes a non-convex log-likelihood function, it is hard to theoretically justify its performance. In this paper, we propose a two-stage efficient algorithm for multi-class crowd labeling problems. The first stage uses the spectral method to obtain an initial estimate of parameters. Then the second stage refines the estimation by optimizing the objective function of the Dawid-Skene estimator via the EM algorithm. We show that our algorithm achieves the optimal convergence rate up to a logarithmic factor. We conduct extensive experiments on synthetic and real datasets. Experimental results demonstrate that the proposed algorithm is comparable to the most accurate empirical approach, while outperforming several other recently proposed methods.

W96 Low-Rank Time-Frequency Synthesis

Cédric Févotte cfevotte@unice.fr
CNRS
Matthieu Kowalski kowalski@lss.supelec.fr
Univ Paris-Sud

Many single-channel signal decomposition techniques rely on a low-rank factorization of a time-frequency transform. In particular, nonnegative matrix factorization (NMF) of the spectrogram -- the (power) magnitude of the short-time Fourier transform (STFT) -- has been considered in many audio applications. In this setting, NMF with the Itakura-Saito

divergence was shown to underly a generative Gaussian composite model (GCM) of the STFT, a step forward from more empirical approaches based on ad-hoc transform and divergence specifications. Still, the GCM is not yet a generative model of the raw signal itself, but only of its STFT. The work presented in this paper fills in this ultimate gap by proposing a novel signal synthesis model with low-rank time-frequency structure. In particular, our new approach opens doors to multi-resolution representations, that were not possible in the traditional NMF setting. We describe two expectation-maximization algorithms for estimation in the new model and report audio signal processing results with music decomposition and speech enhancement.

W97 A State-Space Model for Decoding Auditory Attentional Modulation from MEG in a Competing-Speaker Environment

Sahar Akram sakram@umd.edu
Jonathan Simon jzsimon@umd.edu
Shihab Shamma sas@isr.umd.edu
Behtash Babadi behtash@umd.edu
University of Maryland

Humans are able to segregate auditory objects in a complex acoustic scene, through an interplay of bottom-up feature extraction and top-down selective attention in the brain. The detailed mechanism underlying this process is largely unknown and the ability to mimic this procedure is an important problem in artificial intelligence and computational neuroscience. We consider the problem of decoding the attentional state of a listener in a competing-speaker environment from magnetoencephalographic (MEG) recordings from the human brain. We develop a behaviorally inspired state-space model to account for the modulation of the MEG with respect to attentional state of the listener. We construct a decoder based on the maximum a posteriori (MAP) estimate of the state parameters via the Expectation-Maximization (EM) algorithm. The resulting decoder is able to track the attentional modulation of the listener with multi-second resolution using only the envelopes of the two speech streams as covariates. We present simulation studies as well as application to real MEG data from two human subjects. Our results reveal that the proposed decoder provides substantial gains in terms of temporal resolution, complexity, and decoding accuracy.

W98 Unsupervised Transcription of Piano Music

Taylor Berg-Kirkpatrick tberg@cs.berkeley.edu
Jacob Andreas jda@cs.berkeley.edu
Dan Klein klein@cs.berkeley.edu
UC Berkeley

We present a new probabilistic model for transcribing piano music from audio to a symbolic form. Our model reflects the process by which discrete musical events give rise to acoustic signals that are then superimposed to produce the observed data. As a result, the inference procedure for our model naturally resolves the source separation problem introduced by the the piano's polyphony. In order to adapt to the properties of a new instrument or acoustic environment being transcribed, we learn recording specific spectral profiles and temporal envelopes in an unsupervised fashion. Our system outperforms the best published approaches on a standard piano transcription task, achieving a 10.6% relative gain in note onset F1 on real piano audio.

W99 Efficient Structured Matrix Rank Minimization

Adams Wei Yu adamsyuwei@gmail.com
Wanli Ma mawanli@cs.cmu.edu
Yaoliang Yu yaoliang@cs.cmu.edu
Jaime Carbonell jgc@cs.cmu.edu
Carnegie Mellon University
Suvrit Sra suvrit@gmail.com
MPI for Intelligent Systems

We study the problem of finding structured low-rank matrices using nuclear norm regularization where the structure is encoded by a linear map. In contrast to most known approaches for linearly structured rank minimization, we do not (a) use the full SVD; nor (b) resort to augmented Lagrangian techniques; nor (c) solve linear systems per iteration. Instead, we formulate the problem differently so that it is amenable to a generalized conditional gradient method, which results in a practical improvement with low per iteration computational cost. Numerical results show that our approach significantly outperforms state-of-the-art competitors in terms of running time, while effectively recovering low rank solutions in stochastic system realization and spectral compressed sensing problems.

W100 Efficient Minimax Signal Detection on Graphs

Jing Qian jingq@bu.edu
Venkatesh Saligrama srv@bu.edu
Boston University

Several problems such as network intrusion, community detection, and disease outbreak can be described by observations attributed to nodes or edges of a graph. In these applications presence of intrusion, community or disease outbreak is characterized by novel observations on some unknown connected subgraph. These problems can be formulated in terms of optimization of suitable objectives on connected subgraphs, a problem which is generally computationally difficult. We overcome the combinatorics of connectivity by embedding connected subgraphs into linear matrix inequalities (LMI). Computationally efficient tests are then realized by optimizing convex objective functions subject to these LMI constraints. We prove, by means of a novel Euclidean embedding argument, that our tests are minimax optimal for exponential family of distributions on 1-D and 2-D lattices. We show that internal conductance of the connected subgraph family plays a fundamental role in characterizing detectability.

W101 Non-convex Robust PCA

Praneeth Netrapalli praneethn@gmail.com
Prateek Jain prajain@microsoft.com
Microsoft Research
Niranjan U N un.niranjan@gmail.com
Animashree Anandkumar animakumar@gmail.com
UC Irvine
Sujay Sanghavi sanghavi@mail.utexas.edu
UT Austin

We propose a new provable method for robust PCA, where the task is to recover a low-rank matrix, which is corrupted with sparse perturbations. Our method consists of simple alternating projections onto the set of low rank and sparse matrices with intermediate de-noising steps. We prove correct recovery of the low rank and sparse components under tight recovery conditions, which match those for the state-of-art convex relaxation techniques. Our method is extremely simple to implement and has low computational complexity. For a $m \times n$ input matrix (say m \geq n), our method has O(r^2 mn\log(1/\epsilon)) running time, where r is the rank of the low-rank component and ϵ is the accuracy. In contrast, the convex relaxation methods have a running time O(mn^2/\epsilon), which is not scalable to large problem instances. Our running time nearly matches that of the usual PCA (i.e. non robust), which is O(rmn\log (1/\epsilon)). Thus, we achieve ``best of both the worlds", viz low computational complexity and provable recovery for robust PCA. Our analysis represents one of the few instances of global convergence guarantees for non-convex methods.

W102 Signal Aggregate Constraints in Additive Factorial HMMs, with Application to Energy Disaggregation

Mingjun Zhong mzhong@inf.ed.ac.uk
Nigel Goddard nigel.goddard@ed.ac.uk
Charles Sutton csutton@inf.ed.ac.uk
University of Edinburgh

Blind source separation problems are difficult because they are inherently unidentifiable, yet the entire goal is to identify meaningful sources. We introduce a way of incorporating domain knowledge into this problem, called signal aggregate constraints (SACs). SACs encourage the total signal for each of the unknown sources to be close to a specified value. This is based on the observation that the total signal often varies widely across the unknown sources, and we often have a good idea of what total values to expect. We incorporate SACs into an additive factorial hidden Markov model (AFHMM) to formulate the energy disaggregation problems where only one mixture signal is assumed to be observed. A convex quadratic program for approximate inference is employed for recovering those source signals. On a real-world energy disaggregation data set, we show that the use of SACs dramatically improves the original AFHMM, and significantly improves over a recent state-of-the art approach.

W103 Fast Multivariate Spatio-temporal Analysis via Low Rank Tensor Learning

Mohammad Taha Bahadori mohammab@usc.edu
Qi (Rose) Yu qiyu@usc.edu
Yan Liu yanliu.cs@usc.edu
University of Southern California

Accurate and efficient analysis of multivariate spatio-temporal data is critical in climatology, geology, and sociology applications. Existing models usually assume simple inter-dependence among variables, space, and time, and are computationally expensive. We propose a unified low rank tensor learning framework for multivariate spatio-temporal analysis, which can conveniently incorporate different properties in spatio-temporal data, such as spatial clustering and shared structure among variables. We demonstrate how the general framework can be applied to cokriging and forecasting tasks, and develop an efficient greedy algorithm to solve the resulting optimization problem with convergence guarantee. We conduct experiments on both synthetic datasets and real application datasets to demonstrate that our method is not only significantly faster than existing methods but also achieves lower estimation error.

W104 The Blinded Bandit: Learning with Adaptive Feedback

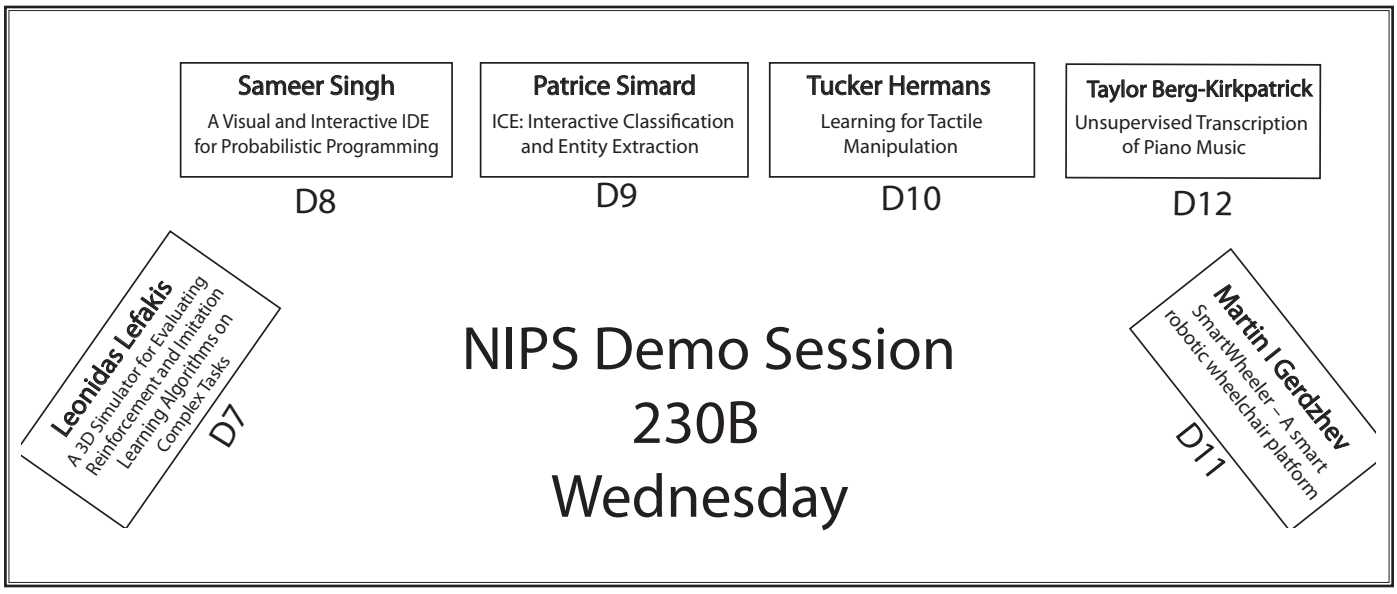
Ofer Dekel oferd@microsoft.com
Microsoft Research
Elad Hazan ehazan@ie.technion.ac.il
Tomer Koren tomerk@technion.ac.il
Technion

We study an online learning setting where the player is temporarily deprived of feedback each time it switches to a different action. Such model of **adaptive feedback** naturally occurs in scenarios where the environment reacts to the player's actions and requires some time to recover and stabilize after the algorithm switches actions. This motivates a variant of the multi-armed bandit problem, which we call the **blinded multi-armed bandit**, in which no feedback is given to the algorithm whenever it switches arms. We develop efficient online learning algorithms for this problem and prove that they guarantee the same asymptotic regret as the optimal algorithms for the standard multi-armed bandit problem. This result stands in stark contrast to another recent result, which states that adding a switching cost to the standard multi-armed bandit makes it substantially harder to learn, and provides a direct comparison of how feedback and loss contribute to the difficulty of an online learning problem. We also extend our results to the general prediction framework of bandit linear optimization, again attaining near-optimal regret bounds.

W105 Near-optimal sample compression for nearest neighbors

Lee-Ad Gottlieb leead@ariel.ac.il
Aryeh Kontorovitch karyeh@cs.bgu.ac.il
Ben Gurion University
Pinhas Nisnevitch pinhasn@gmail.com
Ariel University

We present the first sample compression algorithm for nearest neighbors with non-trivial performance guarantees. We complement these guarantees by demonstrating almost matching hardness lower bounds, which show that our bound is nearly optimal. Our result yields new insight into margin-based nearest neighbor classification in metric spaces and allows us to significantly sharpen and simplify existing bounds. Some encouraging empirical results are also presented.



D7 A 3D Simulator for Evaluating Reinforcement and Imitation Learning Algorithms on Complex Tasks

Leonidas Lefakis
EPFL
François Fleuret, Cijo Jose
Idiap Research Institute

We present the MASH simulator, a benchmarking tool for both reinforcement and imitation learning algorithms. It comprises a 3D simulator which allows for evaluation on tasks based on motions in a physical environment whose complexities go well beyond that of typical 2D mazes. The simulator is composed of a graphics engine, Ogre 3D, to generate images corresponding to an avatar's views, and of a physical engine, Bullet, to simulate the avatar's physical interactions with the environment when moving around. We have developed a series of tasks and environments within this framework. To each is associated a reward function, and for some of them a teacher can be queried for the optimal action. The framework has been designed so that new tasks and environments can be easily created. This simulator has been developed in C++ and the code is publicly available under the GPLv3 open-source license.

D8 A Visual and Interactive IDE for Probabilistic Programming

Sameer Singh
University of Washington
Luke Hewitt, Tim Rocktäschel, Sebastian Riedel
University College London

Probabilistic programming languages (PPLs) such as Church have recently gained interest due to their expressiveness and ease-of-use. These languages support models that can be arbitrary composed and abstracted, and close the gap between traditional programming and probabilistic modeling. However debugging machine learning is still quite hard; it is difficult to identify whether a wrong prediction is due to inaccurate modeling, bug in the data representation, or the wrong choice of the approximate inference algorithm. Corresponding to the IDEs for traditional programming languages, there is a significant need for tools for probabilistic programming that enable the user to quickly create, debug, and iterate on complex models and inference. In this demonstration, we present a first step towards integrated developing, debugging and visualization for PPLs. The proposed IDE is based on the read-eval-print loop (REPL) to allow quick iterations on writing and debugging, and consists of the following elements: (1) Editor (read): Users define the model and inference in a declarative, math-like language (with operators such as argmax) using a syntax highlighted code editor. (2) Build Automation (eval): Our underlying compilation engine transforms this declarative definition into factor graphs, and replaces the operators with efficient and/or approximate versions (such as Gibbs sampling, variational inference, gradient-based approaches, etc.) to generate low-level inference or learning code. (3) Debugging/Visualization (print): Our tool presents the underlying factor graph as an interactive UI element that supports clicking, drag and drop, hover, etc. to explore the structure and the potentials of the model. We visualize the results of inference in a graphical manner that adapts to the type of the result (bar charts for simple distributions, shaded maps for matrix-like objects, circles/arrows for NLP data types, etc.). For further fine-grained debugging, we can also surface intermediate results from inference, for example, visualizing the messages in belief propagation for each edge in the factor graph. The combination of concise definition of the model and the inference, underlying compilation to factor graphs, and efficient implementation of operators using approximate inference enables users to write a variety of models, explore a number of different inference algorithms, and tackle tasks from a range of domains using an easy-to-use, platform independent (browser-based) interface.

D9 ICE: Interactive Classification and Entity Extraction

Patrice Simard, Max Chickering, Saleema Amershi,
Aparna Lakshmiratan, Carlos Garcia Jurado Suarez,
Johan Verwey, Jina Suh
Microsoft Research

Quick interaction between a human teacher and a learning machine presents numerous benefits and challenges when working with web-scale data. The human teacher guides the machine towards accomplishing the task of interest. The learning machine leverages big data to find examples that maximize the training value of its interaction with the teacher. When the teacher is restricted to labeling examples selected by the machine, this problem is an instance of active learning. When the teacher can provide additional information to the machine (e.g., suggestions on what examples or predictive features should be used) as the learning task progresses, then the problem becomes one of interactive learning. To accommodate the two-way communication channel needed for efficient interactive learning, the teacher and the machine need an environment that supports an interaction language. The machine can access, process, and summarize more examples than the teacher can see in a lifetime. Based on the machine's output, the teacher can revise the definition of the task or make it more precise. Both the teacher and the machine continuously learn and benefit from the interaction. We have built a platform (ICE) to (1) produce valuable and deployable models and (2) support research on both the machine learning and user interface challenges of the interactive learning problem. The platform relies on a dedicated, low-latency, distributed, in-memory architecture that allows us to construct web-scale learning machines with quick interaction speed.

D10 Learning for Tactile Manipulation

Tucker Hermans, Filipe Veiga, Janine Hölscher,
Herke van Hoof, Jan Peters
TU Darmstadt

Tactile sensing affords robots the opportunity to dexterously manipulate objects in-hand without the need of strong object models and planning. Our demonstration focuses on learning for tactile, in-hand manipulation by robots. We address learning problems related to the control of objects in-hand, as well as perception problems encountered by a robot exploring its environment with a tactile sensor. We demonstrate applications for three specific learning problems: learning to detect slip for grasp stability, learning to reposition objects in-hand, and learning to identify objects and object properties through tactile exploration. We address the problem of learning to detect slip of grasped objects. We show that the robot can learn a detector for slip events which generalizes to novel objects. We leverage this slip detector to produce a feedback controller that can stabilize objects during grasping and manipulation. Our work compares a number of supervised learning approaches and feature representations in order to achieve reliable slip detection. Tactile sensors provide observations of high enough dimension to cause prob- lems for traditional reinforcement learning methods. As such, we introduce a novel reinforcement learning (RL) algorithm which learns transition functions embedded in a reproducing kernel Hilbert space (RKHS). The resulting policy search algorithm provides robust policy updates which can efficiently deal with high-dimensional sensory input. We demonstrate the method on the problem of repositioning a grasped object in the hand. Finally, we present a method for learning to classify objects through tactile exploration. The robot collects data from a number of objects through various exploratory motions. The robot learns a classifier for each object to be used dur- ing exploration of its environment to detect objects in cluttered environments. Here again we compare a number of learning methods and features present in the literature and synthesize a method to best work in human environments the robot is likely to encounter. Users will be able to interact with a robot hand by giving it objects to grasp and attempting to remove these objects from the robot. The hand will also perform some basic in-hand manipulation tasks such as rolling the object between the fingers and rotating the object about a fixed grasp point. Users will also be able to interact with a touch sensor capable of classifying objects as well as semantic events such as slipping from a stable contact location.

D11 SmartWheeler – A Smart Robotic Wheelchair Platform

Martin Gerdzhev, Joelle Pineau, Angus Leigh,
Andrew Sutcliffe
McGill University

This demo features an interactive autonomous wheelchair robot, called the SmartWheeler, designed with the aim of increasing the autonomy and safety of individuals with severe mobility impairments. The robot platform consists of a commercial power wheelchair, modified to be controlled by an on-board computer. It is further equipped with laser range finders and a Kinect sensor used for sensing the environment and building a map. The wheelchair can be given high-level commands (e.g. setting a location goal on a map), or lower-level commands (e.g. move forward, turn left/right, etc.) via a touchscreen interface. The commands are then interpreted and executed autonomously and safely by the robot, by taking into account static and dynamic obstacles in the environment. Several of the robot's software components rely on machine learning methods. One of the most recent developments is a new algorithm for detecting, tracking and following one or multiple people in real time using planar laser scans acquired onboard the robot.

D12 Unsupervised Transcription of Piano Music

Taylor Berg-Kirkpatrick, Jacob Andreas, Dan Klein
UC Berkeley

We demonstrate a polyphonic transcription system capable of recognizing live music from an acoustic piano and representing it in symbolic form. In addition, we show how the system can be used to resynthesize performances by famous pianists using new instruments. Our transcription system uses a new probabilistic model that reflects the process by which discrete musical events give rise to acoustic signals that are then superimposed to produce the observed data. As a result, the inference procedure for our model naturally resolves the source separation problem introduced by the the piano's polyphony. In order to adapt to the properties of a new instrument or acoustic environment being transcribed, we learn recording-specific spectral profiles and temporal envelopes in an unsupervised fashion.

THURSDAY SESSIONS

ORAL SESSION

SESSION 9: 9:00 - 10:10 AM

Session Chair: Jennifer Wortman Vaughan

INVITED TALK: Privacy in the Land of Plenty

Cynthia Dwork dwork@microsoft.com
Microsoft Research

Privacy-preserving data analysis has a large literature that spans several disciplines. Many early attempts have proved problematic either in practice or on paper. A new approach, “differential privacy” -- a notion tailored to situations in which data are plentiful -- has provided a theoretically sound and powerful framework, and given rise to an explosion of research (see the NIPS 2014 tutorial on December 8). We will review the definition of differential privacy, describe some recent algorithmic advances, and conclude with a surprising application.

Cynthia Dwork, Distinguished Scientist at Microsoft Research, is renowned for placing privacy-preserving data analysis on a mathematically rigorous foundation. A cornerstone of this work is differential privacy, a strong privacy guarantee frequently permitting highly accurate data analysis. Dr. Dwork has also made seminal contributions in cryptography and distributed computing, and is a recipient of the Edsger W. Dijkstra Prize, recognizing some of her earliest work establishing the pillars on which every fault-tolerant system has been built for decades. She is a member of the National Academy of Sciences and the National Academy of Engineering, and a Fellow of the American Academy of Arts and Sciences.

Clamping Variables and Approximate Inference

Adrian Weller adrian@cs.columbia.edu
Tony Jebara jebara@cs.columbia.edu
Columbia University

It was recently proved using graph covers (Ruozzi, 2012) that the Bethe partition function is upper bounded by the true partition function for a binary pairwise model that is attractive. Here we provide a new, arguably simpler proof from first principles. We make use of the idea of clamping a variable to a particular value. For an attractive model, we show that summing over the Bethe partition functions for each sub-model obtained after clamping any variable can only raise (and hence improve) the approximation. In fact, we derive a stronger result that may have other useful implications. Repeatedly clamping until we obtain a model with no cycles, where the Bethe approximation is exact, yields the result. We also provide a related lower bound on a broad class of approximate partition functions of general pairwise multi-label models that depends only on the topology. We demonstrate that clamping a few wisely chosen variables can be of practical value by dramatically reducing approximation error.

SPOTLIGHT SESSION

SESSION 9: 10:10 – 10:35 AM

- **Mode Estimation for High Dimensional Discrete Tree Graphical Models**
  C. Chen, D. Metaxas, Rutgers University; H. Liu, T. Zhao, Princeton University
  See abstract Th38

- **Randomized Experimental Design for Causal Graph Discovery**
  H. Hu, A. Vetta, McGill University; Z. Li, École normale supérieure
  See abstract Th40

- **Advances in Learning Bayesian Networks of Bounded Treewidth**
  S. Nie, Q. Ji, RPI; D. Maua, USP; C. de Campos, Queen’s University Belfast
  See abstract Th44

- **Poisson Process Jumping between an Unknown Number of Rates: Application to Neural Spike Data**
  F. Stimberg, A. Ruttor, M. Opper, TU Berlin
  See abstract Th61

- **Augur: Data-Parallel Probabilistic Modelling**
  J. Tristan, A. Pocock, S. Green, G. Steele, Oracle Labs; D. Huang, Harvard University; J. Tassarotti, Carnegie Mellon University
  See abstract Th65

- **Making Pairwise Binary Graphical Models Attractive**
  N. Ruozzi, T. Jebara, Columbia University
  See abstract Th73

- **Transportability from Multiple Environments with Limited Experiments: Completeness Results**
  E. Bareinboim, J. Pearl, UCLA
  See abstract Th74

ORAL SESSION

SESSION 10: 10:50 AM – 12:00 PM

Session Chair: Aurelie C Lozano

POSNER LECTURE: Games, Networks, and People

Michael Kearns mkearns@cis.upenn.edu
University of Pennsylvania

Beginning with the introduction of graphical games and related models, there is now a rich body of algorithmic connections between probabilistic inference, game theory and microeconomics. Strategic analogues of belief propagation and other inference techniques have been developed for the computation of Nash, correlated and market equilibria, and have played a significant role in the evolution of algorithmic game theory over the past decade. There are also important points of departure between probabilistic and strategic graphical models — perhaps most notably that in the latter, vertices are not random variables, but self-interested humans or organizations. It is thus natural to wonder how social network structures might influence equilibrium outcomes such as social welfare or the relative wealth and power of individuals. One logical path that such questions lead to is human-subject experiments on strategic interaction in social networks.

Michael Kearns is Professor and National Center Chair in the Computer and Information Science department at the University of Pennsylvania. His research interests include topics in machine learning, algorithmic game theory, social networks, and computational finance. Prior to joining the Penn faculty, he spent a decade at AT&T/Bell Labs, where he was head of AI Research. He is co-director of Penn's Warren Center for Network and Data Sciences (warrencenter.upenn.edu), and founder of Penn's Networked and Social Systems Engineering (NETS) undergraduate program (www.nets.upenn.edu). Kearns consults extensively in technology and finance, and is a Fellow of the Association for the Advancement of Artificial Intelligence and the American Academy of Arts and Sciences.

A Statistical Decision-Theoretic Framework for Social Choice

Hossein Azari Soufiani azari@fas.harvard.edu
David Parkes parkes@eecs.harvard.edu
Lirong Xiaxia lirong@gmail.com
Harvard University

In this paper, we take a statistical decision-theoretic viewpoint on social choice, putting a focus on the decision to be made on behalf of a system of agents. In our framework, we are given a statistical ranking model, a decision space, and a loss function defined on (parameter, decision) pairs, and formulate social choice mechanisms as statistical estimators that minimize expected loss. This suggests a general framework for the design and analysis of new social choice mechanisms. We compare Bayesian estimators, which minimize Bayesian expected loss, for two variants of the Mallows model and the Kemeny rule. We consider various normative properties, in addition to computational complexity and asymptotic behavior.

SPOTLIGHT SESSION

SESSION 10: 12:00 – 12:25 PM

- **Recurrent Models of Visual Attention**
  V. Mnih, N. Heess, A. Graves, k. kavukcuoglu, Google DeepMind
  See abstract Th9

- **Unsupervised learning of an efficient short-term memory network**
  P. Vertechi, Champalimaud Center for the Unknown; W. Brendel, Champalimaud Neuroscience Programme; C. Machens, Champalimaud Centre for the Unknown
  See abstract Th10

- **Optimal decision-making with time-varying evidence reliability**
  J. Drugowitsch, A. Pouget, University of Geneva; R. Moreno-Bote, Foundation Sant Joan de Deu
  See abstract Th14

- **Optimal Teaching for Limited-Capacity Human Learners**
  K. Patil, Affective Brain Lab; X. Zhu, UW-Madison; Ł. Kopeć, B. Love, University College London
  See abstract Th15

- **Depth Map Prediction from a Single Image using a Multi-Scale Deep Network**
  D. Eigen, New York University; C. Puhrsch, ; R. Fergus, NYU
  See abstract Th29

- **Causal Strategic Inference in Networked Microfinance Economies**
  M. Irfan, Bowdoin College; L. Ortiz, Stony Brook University
  See abstract Th52

POSTER SESSION

POSTERS 2:00 – 6:00 PM

Th1 **Exploiting Linear Structure Within Convolutional Networks for Efficient Evaluation**
E. Denton, W. Zaremba, J. Bruna, Y. LeCun, R. Fergus

Th2 **Unsupervised Deep Haar Scattering on Graphs**
X. Chen, X. Cheng, S. Mallat

Th3 **Multi-View Perceptron: a Deep Model for Learning Face Identity and View Representations**
Z. Zhu, P. Luo, X. Wang, X. Tang

Th4 **Deep Joint Task Learning for Generic Object Extraction**
X. Wang, L. Zhang, L. Lin, Z. Liang, W. Zuo

Th5 **Discriminative Unsupervised Feature Learning with Convolutional Neural Networks**
A. Dosovitskiy, J. Springenberg, M. Riedmiller, T. Brox

Th6 **Modeling Deep Temporal Dependencies with Recurrent "Grammar Cells"**
V. Michalski, R. Memisevic, K. Konda

Th7 **Convolutional Neural Network Architectures for Matching Natural Language Sentences**
B. Hu, Z. Lu, H. Li, Q. Chen

Th8 **Deep Recursive Neural Networks for Compositionality in Language**
O. Irsoy, C. Cardie

Th9 **Recurrent Models of Visual Attention**
V. Mnih, N. Heess, A. Graves, k. kavukcuoglu

Th10 **Unsupervised learning of an efficient short-term memory network**
P. Vertechi, W. Brendel, C. Machens

Th11 **Algorithm selection by rational metareasoning as a model of human strategy selection**
F. Lieder, D. Plunkett, J. Hamrick, S. Russell, N. Hay, T. Griffiths

Th12 **A Framework for Testing Identifiability of Bayesian Models of Perception**
L. Acerbi, W. Ma, S. Vijayakumar

Th13 **Automatic Discovery of Cognitive Skills to Improve the Prediction of Student Learning**
R. Lindsey, M. Khajah, M. Mozer

Th14 **Optimal decision-making with time-varying evidence reliability**
J. Drugowitsch, R. Moreno-Bote, A. Pouget

Th15 **Optimal Teaching for Limited-Capacity Human Learners**
K. Patil, X. Zhu, Ł. Kopeć, B. Love

Th16 **Message Passing Inference for Large Scale Graphical Models with High Order Potentials**
J. Zhang, A. Schwing, R. Urtasun

Th17 **A Filtering Approach to Stochastic Variational Inference**
N. Houlsby, D. Blei

Th18 **Smoothed Gradients for Stochastic Variational Inference**
S. Mandt, D. Blei

Th19 **Analysis of Variational Bayesian Latent Dirichlet Allocation: Weaker Sparsity Than MAP**
S. Nakajima, I. Sato, M. Sugiyama, K. Watanabe, H. Kobayashi

Th20 **Decoupled Variational Gaussian Inference**
M. Khan

Th21 **Stochastic variational inference for hidden Markov models**
N. Foti, J. Xu, D. Laird, E. Fox

Th22 **Object Localization based on Structural SVM using Privileged Information**
J. Feyereisl, S. Kwak, J. Son, B. Han

Th23 **Efficient Inference of Continuous Markov Random Fields with Polynomial Potentials**
S. Wang, A. Schwing, R. Urtasun

Th24 **Structure Regularization for Structured Prediction**
X. Sun

Th25 **Expectation-Maximization for Learning Determinantal Point Processes**
J. Gillenwater, A. Kulesza, E. Fox, B. Taskar

Th26 **Submodular Attribute Selection for Action Recognition in Video**
J. Zheng, Z. Jiang, R. Chellappa, J. Phillips

Th27 **Scale Adaptive Blind Deblurring**
H. Zhang, J. Yang

Th28 **Shape and Illumination from Shading using the Generic Viewpoint Assumption**
D. Zoran, D. Krishnan, J. Bento, B. Freeman

Th29 **Depth Map Prediction from a Single Image using a Multi-Scale Deep Network**
D. Eigen, C. Puhrsch, R. Fergus

Th30 **Self-Paced Learning with Diversity**
L. Jiang, D. Meng, S. Yu, Z. Lan, S. Shan, A. Hauptmann

Th31 **Hardness of parameter estimation in graphical models**
G. Bresler, D. Gamarnik, D. Shah

Th32 **Sequential Monte Carlo for Graphical Models**
C. Andersson Naesseth, F. Lindsten, T. Schön

Th33 **Concavity of reweighted Kikuchi approximation**
P. Loh, A. Wibisono

Th34 **A Complete Variational Tracker**
R. Turner, S. Bottone, B. Avasarala

Th35 **Joint Training of a Convolutional Network and a Graphical Model for Human Pose Estimation**
J. Tompson, A. Jain, Y. LeCun, C. Bregler

Th36 **Clamping Variables and Approximate Inference**
A. Weller, T. Jebara

Th37 **Divide-and-Conquer Learning by Anchoring a Conical Hull**
T. Zhou, J. Bilmes, C. Guestrin

Th38 **Mode Estimation for High Dimensional Discrete Tree Graphical Models**
C. Chen, H. Liu, D. Metaxas, T. Zhao

Th39 **Graphical Models for Recovering Probabilistic and Causal Queries from Missing Data**
K. Mohan, J. Pearl

Th40 **Randomized Experimental Design for Causal Graph Discovery**
H. Hu, Z. Li, A. Vetta

Th41 **Optimization Methods for Sparse Pseudo-Likelihood Graphical Model Selection**
S. Oh, O. Dalal, K. Khare, B. Rajaratnam

Th42 **Learning Chordal Markov Networks by Dynamic Programming**
K. Kangas, M. Koivisto, T. Niinimäki

Th43 **QUIC & DIRTY: A Quadratic Approximation Approach for Dirty Statistical Models**
C. Hsieh, I. Dhillon, P. Ravikumar, S. Becker, P. Olsen

Th44 **Advances in Learning Bayesian Networks of Bounded Treewidth**
S. Nie, D. Maua, C. de Campos, Q. Ji

Th45 **Recursive Inversion Models for Permutations**
C. Meek, M. Meila

Th46 **Discovering Structure in High-Dimensional Data Through Correlation Explanation**
G. Ver Steeg, A. Galstyan

Th47 **Coresets for k-Segmentation of Streaming Data**
G. Rosman, M. Volkov, D. Feldman, J. Fisher III, D. Rus

Th48 **Approximating Hierarchical MV-sets for Hierarchical Clustering**
A. Glazer, O. Weissbrod, M. Lindenbaum, S. Markovitch

Th49 **Tight Continuous Relaxation of the Balanced k-Cut Problem**
S. Rangapuram, P. Mudrakarta, M. Hein

Th50 **Streaming, Memory Limited Algorithms for Community Detection**
S. Yun, m. lelarge, A. Proutiere

Th51 **Computing Nash Equilibria in Generalized Interdependent Security Games**
H. Chan, L. Ortiz

Th52 **Causal Strategic Inference in Networked Microfinance Economies**
M. Irfan, L. Ortiz

Th53 **A Statistical Decision-Theoretic Framework for Social Choice**
H. Azari Soufiani, D. Parkes, L. Xia

Th54 **Learning Optimal Commitment to Overcome Insecurity**
A. Blum, N. Haghtalab, A. Procaccia

Th55 **Diverse Randomized Agents Vote to Win**
A. Jiang, L. Soriano Marcolino, A. Procaccia, T. Sandholm, N. Shah, M. Tambe

Th56 **Fairness in Multi-Agent Sequential Decision-Making**
C. Zhang, J. Shah

Th57 **Repeated Contextual Auctions with Strategic Buyers**
K. Amin, A. Rostamizadeh, U. Syed

Th58 **Universal Option Models**
h. yao, C. Szepesvari, R. Sutton, J. Modayil, S. Bhatnagar

Th59 **Semi-Separable Hamiltonian Monte Carlo for Inference in Bayesian Hierarchical Models**
Y. Zhang, C. Sutton

Th60 **Bayesian Sampling Using Stochastic Gradient Thermostats**
N. Ding, Y. Fang, R. Babbush, C. Chen, R. Skeel, H. Neven

Th61 **Poisson Process Jumping between an Unknown Number of Rates: Application to Neural Spike Data**
F. Stimberg, A. Ruttor, M. Opper

Th63 **Distributed Variational Inference in Sparse Gaussian Process Regression and Latent Variable Models**
Y. Gal, M. van der Wilk, C. Rasmussen

Th64 **Incremental Local Gaussian Regression**
F. Meier, P. Hennig, S. Schaal

Th65 **Augur: Data-Parallel Probabilistic Modeling**
J. Tristan, D. Huang, J. Tassarotti, A. Pocock, S. Green, G. Steele

Th66 **Just-In-Time Learning for Fast and Flexible Inference**
S. Eslami, D. Tarlow, P. Kohli, J. Winn

Th67 **Distributed Bayesian Posterior Sampling via Moment Sharing**
M. Xu, B. Lakshminarayanan, Y. Teh, J. Zhu, B. Zhang

Th68 **Communication Efficient Distributed Machine Learning with the Parameter Server**
M. Li, D. Andersen, A. Smola, K. Yu

Th69 **On Model Parallelization and Scheduling Strategies for Distributed Machine Learning**
S. Lee, J. Kim, X. Zheng, Q. Ho, G. Gibson, E. Xing

Th70 **Communication-Efficient Distributed Dual Coordinate Ascent**
M. Jaggi, V. Smith, M. Takac, J. Terhorst, S. Krishnan, T. Hofmann, M. Jordan

Th71 **Projective dictionary pair learning for pattern classification**
S. Gu, L. Zhang, W. Zuo, X. Feng

Th72 **Augmentative Message Passing for Traveling Salesman Problem and Graph Partitioning**
S. Ravanbakhsh, R. Rabbany, R. Greiner

Th73 **Making Pairwise Binary Graphical Models Attractive**
N. Ruozzi, T. Jebara

Th74 **Transportability from Multiple Environments with Limited Experiments: Completeness Results**
E. Bareinboim, J. Pearl

Th75 **Causal Inference through a Witness Protection Program**
R. Silva, R. Evans

Th76 **Biclustering Usinig Message Passing**
L. O'Connor, S. Feizi

Th77 **PAC-Bayesian AUC classification and scoring**
J. Ridgway, P. Alquier, N. Chopin, F. Liang

Th78 **Partition-wise Linear Models**
H. Oiwa, R. Fujimaki

Th79 **Near-optimal sample compression for nearest neighbors**
L. Gottlieb, A. Kontorovitch, P. Nisnevitch
Moved to Wednesday, Abstract W105

Th80 **Learning Shuffle Ideals Under Restricted Distributions**
D. Chen

Th81 **Extracting Certainty from Uncertainty: Transductive Pairwise Classification from Pairwise Similarities**
T. Yang, R. Jin

Th82 **Incremental Clustering: The Case for Extra Clusters**
M. Ackerman, S. Dasgupta

Th83 **Parallel Successive Convex Approximation for Nonsmooth Nonconvex Optimization**
M. Razaviyayn, M. Hong, Z. Luo, J. Pang

Th84 **Stochastic Proximal Gradient Descent with Acceleration Techniques**
A. Nitanda

Th85 **Beyond the Birkhoff Polytope: Convex Relaxations for Vector Permutation Problems**
C. Lim, S. Wright

Th86 **Bregman Alternating Direction Method of Multipliers**
H. Wang, A. Banerjee

Th87 **Multi-Step Stochastic ADMM in High Dimensions: Applications to Sparse Optimization and Matrix Decomposition**
H. Sedghi, A. Anandkumar, E. Jonckheere

Th88 **Accelerated Mini-batch Randomized Block Coordinate Descent Method**
T. Zhao, M. Yu, Y. Wang, R. Arora, H. Liu

Th89 **Sparse PCA with Oracle Property**
Q. Gu, Z. Wang, H. Liu

Th90 **A statistical model for tensor PCA**
E. Richard, A. Montanari

Th91 **large scale canonical correlation analysis with iterative least squares**
Y. Lu, D. Foster

Th92 **Cone-Constrained Principal Component Analysis**
Y. Deshpande, A. Montanari, E. Richard

Th93 **Improved Distributed Principal Component Analysis**
Y. Liang, M. Balcan, V. Kanchanapally, D. Woodruff

Th94 **Generalized Unsupervised Manifold Alignment**
Z. Cui, H. Chang, S. Shan, X. Chen

Th95 **On Prior Distributions and Approximate Inference for Structured Variables**
O. Koyejo, R. Khanna, J. Ghosh, R. Poldrack

Th96 **Real-Time Decoding of an Integrate and Fire Encoder**
S. Saxena, M. Dahleh

Th97 **Spike Frequency Adaptation Implements Anticipative Tracking in Continuous Attractor Neural Networks**
Y. Mi, C. Fung, K. Wong, S. Wu

Th98 **Analysis of Brain States from Multi-Region LFP Time-Series**
K. Ulrich, D. Carlson, W. Lian, J. Borg, K. Dzirasa, L. Carin

Th99 **Extracting Latent Structure From Multiple Interacting Neural Populations**
J. Semedo, A. Zandvakili, A. Kohn, C. Machens, B. Yu

Th100 **Learning convolution filters for inverse covariance estimation of neural network connectivity**
G. Mohler

Th101 **On Multiplicative Multitask Feature Learning**
X. Wang, J. Bi, S. Yu, J. Sun

Th102 **Multitask learning meets tensor factorization: task imputation via convex optimization**
K. Wimalawarne, M. Sugiyama, R. Tomioka

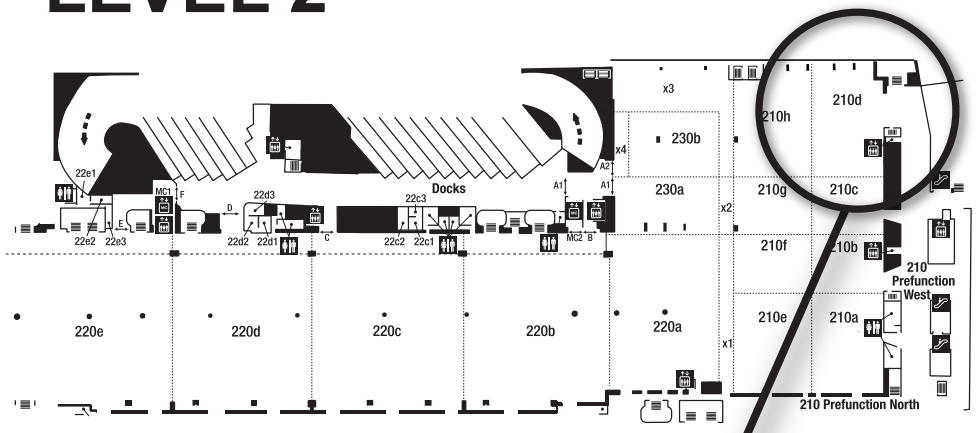
Th103 **Learning Multiple Tasks in Parallel with a Shared Annotator**
H. Cohen, K. Crammer

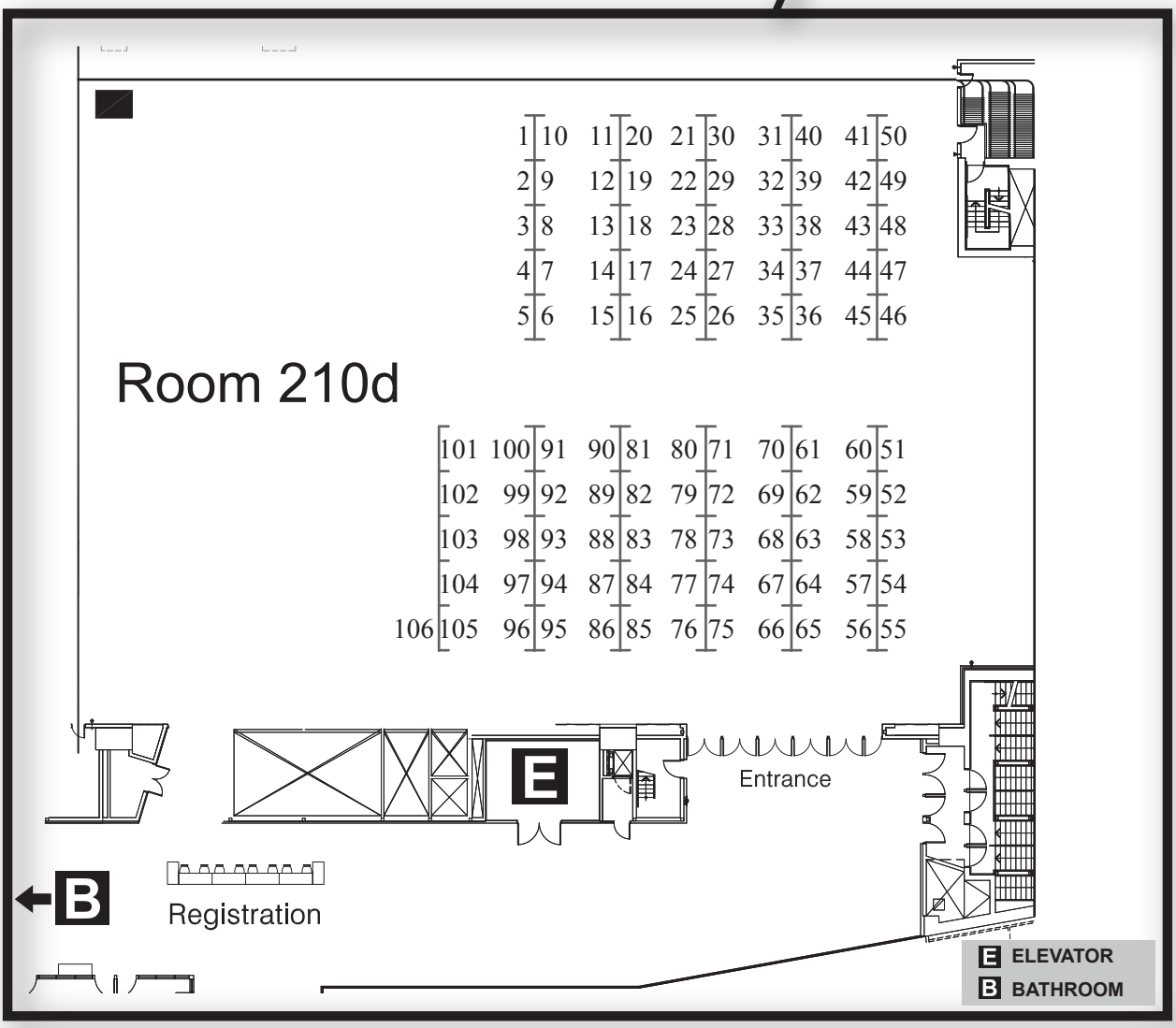
Th104 **Multivariate Regression with Calibration**
H. Liu, L. Wang, T. Zhao

Th105 **Exclusive Feature Learning on Arbitrary Structures via $\ell 1,2$-norm**
D. Kong, R. Fujimaki, J. Liu, F. Nie, C. Ding

Th106 **Flexible Transfer Learning under Support and Model Shift**
X. Wang, J. Schneider

LEVEL 2


x3
210h
210d
230b
x4
A2
22e1
MC1
F
Docks
A1
A1
230a
210g
210c
E
22e2
22e3
D
22d3
22d2
22d1
C
22c3
22c2
22c1
MC2
B
x2
210f
210b
210
Prefunction
West
220e
220d
220c
220b
220a
210e
210a
x1
210 Prefunction North


1 10 11 20 21 30 31 40 41 50
2 9 12 19 22 29 32 39 42 49
3 8 13 18 23 28 33 38 43 48
4 7 14 17 24 27 34 37 44 47
5 6 15 16 25 26 35 36 45 46
Room 210d
101 100 91 90 81 80 71 70 61 60 51
102 99 92 89 82 79 72 69 62 59 52
103 98 93 88 83 78 73 68 63 58 53
104 97 94 87 84 77 74 67 64 57 54
106 105 96 95 86 85 76 75 66 65 56 55
E
Entrance
B
Registration
E ELEVATOR
B BATHROOM

Th1 Exploiting Linear Structure Within Convolutional Networks for Efficient Evaluation

Emily Denton denton@cs.nyu.edu
Wojciech Zaremba woj.zaremba@gmail.com
Joan Bruna bruna@cims.nyu.edu
Rob Fergus fergus@cs.nyu.edu
Yann LeCun yann@cs.nyu.edu
New York University

We present techniques for speeding up the test-time evaluation of large convolutional networks, designed for object recognition tasks. These models deliver impressive accuracy, but each image evaluation requires millions of floating point operations, making their deployment on smartphones and Internet-scale clusters problematic. The computation is dominated by the convolution operations in the lower layers of the model. We exploit the redundancy present within the convolutional filters to derive approximations that significantly reduce the required computation. Using large state-of-the-art models, we demonstrate speedups of convolutional layers on both CPU and GPU by a factor of 2×, while keeping the accuracy within 1% of the original model.

Th2 Unsupervised Deep Haar Scattering on Graphs

Xu Chen xhen@umich.edu
Princeton University
Xiuyuan Cheng xiuyuan.cheng@ens.fr
Stephane Mallat mallat@cmap.polytechnique.fr
Ecole Polytechnique Paris

The classification of high-dimensional data defined on graphs is particularly difficult when the graph geometry is unknown. We introduce a Haar scattering transform on graphs, which computes invariant signal descriptors. It is implemented with a deep cascade of additions, subtractions and absolute values, which iteratively compute orthogonal Haar wavelet transforms. Multiscale neighborhoods of unknown graphs are estimated by minimizing an average total variation, with a pair matching algorithm of polynomial complexity. Supervised classification with dimension reduction is tested on data bases of scrambled images, and for signals sampled on unknown irregular grids on a sphere.

Th3 Multi-View Perceptron: a Deep Model for Learning Face Identity and View Representations

Zhenyao Zhu zz012@ie.cuhk.edu.hk
Ping Luo pluo.lhi@gmail.com
Xiaogang Wang xgwang@ee.cuhk.edu.hk
Xiaoou Tang xtang@ie.cuhk.edu.hk
Chinese University of Hong Kong

Various factors, such as identities, views (poses), and illuminations, are coupled in face images. Disentangling the identity and view representations is a major challenge in face recognition. Existing face recognition systems either use handcrafted features or learn features discriminatively to improve recognition accuracy. This is different from the behavior of human brain. Intriguingly, even without accessing 3D data, human not only can recognize face identity, but can also imagine face images of a person under different viewpoints given a single 2D image, making face perception in the brain robust to view changes. In this sense, human brain has learned and encoded 3D face models from 2D images. To take into account this instinct, this paper proposes a novel deep neural net, named multi-view perceptron (MVP), which can untangle the identity and view features, and infer a full spectrum of multi-view images in the meanwhile, given a single 2D face image. The identity features of MVP achieve superior performance on the MultiPIE dataset. MVP is also capable to interpolate and predict images under viewpoints that are unobserved in the training data.

Th4 Deep Joint Task Learning for Generic Object Extraction

Xiaolong Wang dragonwxl123@gmail.com
Carnegie Mellon University
Liliang Zhang zhangll.level0@gmail.com
Liang Lin linliang@ieee.org
Zhujin Liang alfredtofu@gmail.com
Sun Yat-Sen University
Wangmeng Zuo cswmzuo@gmail.com

This paper investigates how to extract objects-of-interest without relying on hand-craft features and sliding windows approaches, that aims to jointly solve two sub-tasks: (i) rapidly localizing salient objects from images, and (ii) accurately segmenting the objects based on the localizations. We present a general joint task learning framework, in which each task (either object localization or object segmentation) is tackled via a multi-layer convolutional neural network, and the two networks work collaboratively to boost performance. In particular, we propose to incorporate latent variables bridging the two networks in a joint optimization manner. The first network directly predicts the positions and scales of salient objects from raw images, and the latent variables adjust the object localizations to feed the second network that produces pixelwise object masks. An EM-type method is then studied for the joint optimization, iterating with two steps: (i) by using the two networks, it estimates the latent variables by employing an MCMC-based sampling method; (ii) it optimizes the parameters of the two networks unitedly via back propagation, with the fixed latent variables. Extensive experiments demonstrate that our joint learning framework significantly outperforms other state-of-the-art approaches in both accuracy and efficiency (e.g., 1000 times faster than competing approaches).

Th5 Discriminative Unsupervised Feature Learning with Convolutional Neural Networks

Alexey Dosovitskiy dosovits@cs.uni-freiburg.de
Jost Tobias Springenberg springj@tf.uni-freiburg.de
Martin Riedmiller martin.riedmiller@uos.de
Thomas Brox brox@cs.uni-freiburg.de
University of Freiburg

Current methods for training convolutional neural networks depend on large amounts of labeled samples for supervised training. In this paper we present an approach for training a convolutional neural network using only unlabeled data. We train the network to discriminate between a set of surrogate classes. Each surrogate class is formed by applying a variety of transformations to a randomly sampled 'seed' image patch. We find that

this simple feature learning algorithm is surprisingly successful when applied to visual object recognition. The feature representation learned by our algorithm achieves classification results matching or outperforming the current state-of-the-art for unsupervised learning on several popular datasets (STL-10, CIFAR-10, Caltech-101).

Th6 Modeling Deep Temporal Dependencies with Recurrent "Grammar Cells"

Vincent Michalski vmichals@rz.uni-frankfurt.de
Kishore Konda konda@informatik.uni-frankfurt.de
Goethe University Frankfurt
Roland Memisevic roland.memisevic@gmail.com
University of Montreal

We propose modeling time series by representing the transformations that take a frame at time t to a frame at time t+1. To this end we show how a bi-linear model of transformations, such as a gated autoencoder, can be turned into a recurrent network, by training it to predict future frames from the current one and the inferred transformation using backprop-through-time. We also show how stacking multiple layers of gating units in a recurrent pyramid makes it possible to represent the "syntax" of complicated time series, and that it can outperform standard recurrent neural networks in terms of prediction accuracy on a variety of tasks.

Th7 Convolutional Neural Network Architectures for Matching Natural Language Sentences

Baotian Hu baotianchina@126.com
Harbin Institute of Technology
Zhengdong Lu lu.zhengdong@huawei.com
Hang Li hangli.hl@huawei.com
Noah's Ark Lab, Huawei Technologies
Qingcai Chen qingcai.chen@gmail.com
Harbin Institute of Technology

Semantic matching is of central importance to many natural language tasks \cite{bordes2014semantic,RetrievalQA}. A successful matching algorithm needs to adequately model the internal structures of language objects and the interaction between them. As a step toward this goal, we propose convolutional neural network models for matching two sentences, by adapting the convolutional strategy in vision and speech. The proposed models not only nicely represent the hierarchical structures of sentences with their layer-by-layer composition and pooling, but also capture the rich matching patterns at different levels. Our models are rather generic, requiring no prior knowledge on language, and can hence be applied to matching tasks of different nature and in different languages. The empirical study on a variety of matching tasks demonstrates the efficacy of the proposed model on a variety of matching tasks and its superiority to competitor models.

Th8 Deep Recursive Neural Networks for Compositionality in Language

Ozan Irsoy oirsoy@cs.cornell.edu
Claire Cardie cardie@cs.cornell.edu
Cornell University

Recursive neural networks comprise a class of architecture that can operate on structured input. They have been previously successfully applied to model compositionality in natural language using parse-tree-based structural representations. Even though these architectures are deep in structure, they lack the capacity for hierarchical representation that exists in conventional deep feed-forward networks as well as in recently investigated deep recurrent neural networks. In this work we introduce a new architecture --- a deep recursive neural network (deep RNN) --- constructed by stacking multiple recursive layers. We evaluate the proposed model on the task of fine-grained sentiment classification. Our results show that deep RNNs outperform associated shallow counterparts that employ the same number of parameters. Furthermore, our approach outperforms previous baselines on the sentiment analysis task, including a multiplicative RNN variant as well as the recently introduced paragraph vectors, achieving new state-of-the-art results. We provide exploratory analyses of the effect of multiple layers and show that they capture different aspects of compositionality in language.

Th9 Recurrent Models of Visual Attention

Volodymyr Mnih vmnih@google.com
Nicolas Heess heess@google.com
Alex Graves alex.graves@gmail.com
koray kavukcuoglu korayk@google.com
Google DeepMind

Applying convolutional neural networks to large images is computationally expensive because the amount of computation scales linearly with the number of image pixels. We present a novel recurrent neural network model that is capable of extracting information from an image or video by adaptively selecting a sequence of regions or locations and only processing the selected regions at high resolution. Like convolutional neural networks, the proposed model has a degree of translation invariance built-in, but the amount of computation it performs can be controlled independently of the input image size. While the model is non-differentiable, it can be trained using reinforcement learning methods to learn task-specific policies. We evaluate our model on several image classification tasks, where it significantly outperforms a convolutional neural network baseline on cluttered images, and on a dynamic visual control problem, where it learns to track a simple object without an explicit training signal for doing so.

Th10 Unsupervised learning of an efficient short-term memory network

Pietro Vertechi pietro.vertechi@neuro.fchampalimaud.org
Wieland Brendel wieland.brendel@ens.fr
Christian Machens
christian.machens@neuro.fchampalimaud.org
Champalimaud Centre for the Unknown

Learning in recurrent neural networks has been a topic fraught with difficulties and problems. We here report substantial progress in the unsupervised learning of recurrent networks that can keep track of an input signal. Specifically, we show how these networks can learn to efficiently represent their present and past inputs, based on local learning rules only. Our results are based on several key insights. First, we develop a local learning rule for the recurrent weights whose main aim is to drive the network into a regime where, on average, feedforward signal inputs are canceled by recurrent inputs. We show that this learning rule minimizes a cost function. Second, we develop a local learning rule for the feedforward weights that, based on networks in which recurrent inputs already predict feedforward inputs, further minimizes the cost. Third, we show how the learning rules can be modified such that the network can directly encode non-whitened inputs. Fourth, we show that these learning rules can also be applied to a network that feeds a time-delayed version of the network output back into itself. As a consequence, the network starts to efficiently represent both its signal inputs and their history. We develop our main theory for linear networks, but then sketch how the learning rules could be transferred to balanced, spiking networks.

Th11 Algorithm selection by rational metareasoning as a model of human strategy selection

Falk Lieder falk.lieder@gmail.com
Dillon Plunkett dillonplunkett@berkeley.edu
Stuart Russell russell@cs.berkeley.edu
Nicholas Hay nickjhay@cs.berkeley.edu
Thomas Griffiths tom_griffiths@berkeley.edu
Jessica Hamrick jhamrick@berkeley.edu
University of California, Berkeley

Selecting the right algorithm is an important problem in computer science, because the algorithm often has to exploit the structure of the input to be efficient. The human mind faces the same challenge. Therefore, solutions to the algorithm selection problem can inspire models of human strategy selection and vice versa. Here, we view the algorithm selection problem as a special case of metareasoning and derive a solution that outperforms existing methods in sorting algorithm selection. We apply our theory to model how people choose between cognitive strategies and test its prediction in a behavioral experiment. We find that people quickly learn to adaptively choose between cognitive strategies. People's choices in our experiment are consistent with our model but inconsistent with previous theories of human strategy selection. Rational metareasoning appears to be a promising framework for reverse-engineering how people choose among cognitive strategies and translating the results into better solutions to the algorithm selection problem.

Th12 A Framework for Testing Identifiability of Bayesian Models of Perception

Luigi Acerbi luigi.acerbi@nyu.edu
Wei Ji Ma weijima@gmail.com
New York University
Sethu Vijayakumar sethu.vijayakumar@ed.ac.uk
University of Edinburgh

Bayesian observer models are very effective in describing human performance in perceptual tasks, so much so that they are trusted to faithfully recover hidden mental representations of priors, likelihoods, or loss functions from the data. However, the intrinsic degeneracy of the Bayesian framework, as multiple combinations of elements can yield empirically indistinguishable results, prompts the question of model identifiability. We propose a novel framework for a systematic testing of the identifiability of a significant class of Bayesian observer models, with practical applications for improving experimental design. We examine the theoretical identifiability of the inferred internal representations in two case studies. First, we show which experimental designs work better to remove the underlying degeneracy in a time interval estimation task. Second, we find that the reconstructed representations in a speed perception task under a slow-speed prior are fairly robust.

Th13 Automatic Discovery of Cognitive Skills to Improve the Prediction of Student Learning

Robert Lindsey robert.lindsey@colorado.edu
Michael Mozer mozer@colorado.edu
Mohammad Khajah mohammad.khajah@colorado.edu
University of Colorado Boulder

To master a discipline such as algebra or physics, students must acquire a set of cognitive skills. Traditionally, educators and domain experts manually determine what these skills are and then select practice exercises to hone a particular skill. We propose a technique that uses student performance data to automatically discover the skills needed in a discipline. The technique assigns a latent skill to each exercise such that a student's expected accuracy on a sequence of same-skill exercises improves monotonically with practice. Rather than discarding the skills identified by experts, our technique incorporates a nonparametric prior over the exercise-skill assignments that is based on the expert-provided skills and a weighted Chinese restaurant process. We test our technique on datasets from five different intelligent tutoring systems designed for students ranging in age from middle school through college. We obtain two surprising results. First, in three of the five datasets, the skills inferred by our technique support significantly improved predictions of student performance over the expert-provided skills. Second, the expert-provided skills have little value: our technique predicts student performance nearly as well when it ignores the domain expertise as when it attempts to leverage it. We discuss explanations for these surprising results and also the relationship of our skill-discovery technique to alternative approaches.

Th14 Optimal decision-making with time-varying evidence reliability

Jan Drugowitsch jdrugo@gmail.com
Alexandre Pouget alexandre.pouget@unige.ch
University of Geneva
Ruben Moreno-Bote rmoreno@fsjd.org
Foundation Sant Joan de Deu

Previous theoretical and experimental work on optimal decision-making was restricted to the artificial setting of a reliability of the momentary sensory evidence that remained constant within single trials. The work presented here describes the computation and characterization of optimal decision-making in the more realistic case of an evidence reliability that varies across time even within a trial. It shows that, in this case, the optimal behavior is determined by a bound in the decision maker's belief that depends only on the current, but not the past, reliability. We furthermore demonstrate that simpler heuristics fail to match the optimal performance for certain characteristics of the process that determines the time-course of this reliability, causing a drop in reward rate by more than 50%.

Th15 Optimal Teaching for Limited-Capacity Human Learners

Kaustubh Patil kaustubh.patil@gmail.com
Affective Brain Lab
Xiaojin Zhu jerryzhu@cs.wisc.edu
UW-Madison
Łukasz Kopeć l.kopec.12@ucl.ac.uk
Bradley Love b.love@ucl.ac.uk
University College London

Basic decisions, such as judging a person as a friend or foe, involve categorizing novel stimuli. Recent work finds that people's category judgments are guided by a small set of examples that are retrieved from memory at decision time. This limited and stochastic retrieval places limits on human performance for probabilistic classification decisions. In light of this capacity limitation, recent work finds that idealizing training items, such that the saliency of ambiguous cases is reduced, improves human performance on novel test items. One shortcoming of previous work in idealization is that category distributions were idealized in an ad hoc or heuristic fashion. In this contribution, we take a first principles approach to constructing idealized training sets. We apply a machine teaching procedure to a cognitive model that is either limited capacity (as humans are) or unlimited capacity (as most machine learning systems are). As predicted, we find that the machine teacher recommends idealized training sets. We also find that human learners perform best when training recommendations from the machine teacher are based on a limited-capacity model. As predicted, to the extent that the learning model used by the machine teacher conforms to the true nature of human learners, the recommendations of the machine teacher prove effective. Our results provide a normative basis (given capacity constraints) for idealization procedures and offer a novel selection procedure for models of human learning.

Th16 Message Passing Inference for Large Scale Graphical Models with High Order Potentials

Jian Zhang jizhang@student.ethz.ch
ETH Zurich
Alex Schwing aschwing@cs.toronto.edu
Raquel Urtasun urtasun@cs.toronto.edu
University of Toronto

To keep up with the Big Data challenge, parallelized algorithms based on dual decomposition have been proposed to perform inference in Markov random fields. Despite this parallelization, current algorithms struggle when the energy has high order terms and the graph is densely connected. In this paper we propose a partitioning strategy followed by a message passing algorithm which is able to exploit pre-computations. It only updates the high-order factors when passing messages across machines. We demonstrate the effectiveness of our approach on the task of joint layout and semantic segmentation estimation from single images, and show that our approach is orders of magnitude faster than current methods.

Th17 A Filtering Approach to Stochastic Variational Inference

Neil Houlsby nmth2@cam.ac.uk
Cambridge
David Blei blei@cs.princeton.edu
Columbia University

Stochastic variational inference (SVI) uses stochastic optimization to scale up Bayesian computation to massive data. We present an alternative perspective on SVI as approximate parallel coordinate ascent. SVI trades-off bias and variance to step close to the unknown true coordinate optimum given by batch variational Bayes (VB). We define a model to automate this process. The model infers the location of the next VB optimum from a sequence of noisy realizations. As a consequence of this construction, we update the variational parameters using Bayes rule, rather than a hand-crafted optimization schedule. When our model is a Kalman filter this procedure can recover the original SVI algorithm and SVI with adaptive steps. We may also encode additional assumptions in the model, such as heavy-tailed noise. By doing so, our algorithm outperforms the original SVI schedule and a state-of-the-art adaptive SVI algorithm in two diverse domains.

Th18 Smoothed Gradients for Stochastic Variational Inference

Stephan Mandt stephan.mandt@gmail.com
Princeton University
David Blei blei@cs.princeton.edu
Columbia University

Stochastic variational inference (SVI) lets us scale up Bayesian computation to massive data. It uses stochastic optimization to fit a variational distribution, following easy-to-compute noisy natural gradients. As with most traditional stochastic optimization methods, SVI takes precautions to use unbiased stochastic gradients whose expectations are equal to the true gradients. In this paper, we explore the idea of following biased stochastic gradients in SVI. Our method replaces the natural gradient with a similarly constructed vector that uses a fixed-window moving average of some of its previous terms. We will demonstrate the many advantages of this technique. First, its computational cost is the same as for SVI and storage requirements only multiply by a constant factor. Second, it enjoys significant variance reduction over the unbiased estimates, smaller bias than averaged gradients, and leads to smaller mean-squared error against the full gradient. We test our method on latent Dirichlet allocation with three large corpora.

Th19 Analysis of Variational Bayesian Latent Dirichlet Allocation: Weaker Sparsity Than MAP

Shinichi Nakajima shinnkj23@gmail.com
Technische Universität Berlin
Issei Sato isseis@gmail.com
Masashi Sugiyama sugi@k.u-tokyo.ac.jp
The University of Tokyo
Kazuho Watanabe wkazuho@is.naist.jp
Toyohashi Tech
Hiroko Kobayashi hiroko.kobayashi@nikon.com
Nikon Corporation

Latent Dirichlet allocation (LDA) is a popular generative model of various objects such as texts and images, where an object is expressed as a mixture of latent topics. In this paper, we theoretically investigate variational Bayesian (VB) learning in LDA. More specifically, we analytically derive the leading term of the VB free energy under an asymptotic setup, and show that there exist transition thresholds in Dirichlet hyperparameters around which the sparsity-inducing behavior drastically changes. Then we further theoretically reveal the notable phenomenon that VB tends to induce weaker sparsity than MAP in the LDA model, which is opposed to other models. We experimentally demonstrate the practical validity of our asymptotic theory on real-world Last.FM music data.

Th20 Decoupled Variational Gaussian Inference

Emtiyaz Khan emtiyaz@gmail.com
EPFL

Variational Gaussian (VG) inference methods that optimize a lower bound to the marginal likelihood are a popular approach for Bayesian inference. These methods are fast and easy to use, while being reasonably accurate. A difficulty remains in computation of the lower bound when the latent dimensionality L is large. Even though the lower bound is concave for many models, its computation requires optimization over $O(L^2)$ variational parameters. Efficient reparameterization schemes can reduce the number of parameters, but give inaccurate solutions or destroy concavity leading to slow convergence. We propose decoupled variational inference that brings the best of both worlds together. First, it maximizes a Lagrangian of the lower bound reducing the number of parameters to $O(N)$, where N is the number of data examples. The reparameterization obtained is unique and recovers maxima of the lower-bound even when the bound is not concave. Second, our method maximizes the lower bound using a sequence of convex problems, each of which is parallellizable over data examples and computes gradient efficiently. Overall, our approach avoids all direct computations of the covariance, only requiring its linear projections. Theoretically, our method converges at the same rate as existing methods in the case of concave lower bounds, while remaining convergent at a reasonable rate for the non-concave case.

Th21 Stochastic variational inference for hidden Markov models

Nicholas Foti nfoti@uw.edu
Jason Xu jasonxu@uw.edu
Dillon Laird dillonl2@cs.washington.edu
Emily Fox ebfox@uw.edu
University of Washington

Variational inference algorithms have proven successful for Bayesian analysis in large data settings, with recent advances using stochastic variational inference (SVI). However, such methods have largely been studied in independent or exchangeable data settings. We develop an SVI algorithm to learn the parameters of hidden Markov models (HMMs) in a time-dependent data setting. The challenge in applying stochastic optimization in this setting arises from dependencies in the chain, which must be broken to consider minibatches of observations. We propose an algorithm that harnesses the memory decay of the chain to adaptively bound errors arising from edge effects. We demonstrate the effectiveness of our algorithm on synthetic experiments and a large genomics dataset where a batch algorithm is computationally infeasible.

Th22 Object Localization based on Structural SVM using Privileged Information

Jan Feyereisl thefillm@gmail.com
Postech, Samsung Electronics
Suha Kwak mercury3@postech.ac.kr
INRIA
Jeany Son jeany@postech.ac.kr
Bohyung Han bhhan@postech.ac.kr
POSTECH

We propose a structured prediction algorithm for object localization based on Support Vector Machines (SVMs) using privileged information. Privileged information provides useful high-level knowledge for image understanding and facilitates learning a reliable model even with a small number of training examples. In our setting, we assume that such information is available only at training time since it may be difficult to obtain

from visual data accurately without human supervision. Our goal is to improve performance by incorporating privileged information into ordinary learning framework and adjusting model parameters for better generalization. We tackle object localization problem based on a novel structural SVM using privileged information, where an alternating loss-augmented inference procedure is employed to handle the term in the objective function corresponding to privileged information. We apply the proposed algorithm to the Caltech-UCSD Birds 200-2011 dataset, and obtain encouraging results suggesting further investigation into the benefit of privileged information in structured prediction.

Th23 Efficient Inference of Continuous Markov Random Fields with Polynomial Potentials

Shenlong Wang shenlong.wang@gmail.com
Alex Schwing aschwing@cs.toronto.edu
Raquel Urtasun urtasun@cs.toronto.edu
University of Toronto

In this paper, we prove that every multivariate polynomial with even degree can be decomposed into a sum of convex and concave polynomials. Motivated by this property, we exploit the concave-convex procedure to perform inference on continuous Markov random fields with polynomial potentials. In particular, we show that the concave-convex decomposition of polynomials can be expressed as a sum-of-squares optimization, which can be efficiently solved via semidefinite programming. We demonstrate the effectiveness of our approach in the context of 3D reconstruction, shape from shading and image denoising, and show that our approach significantly outperforms existing approaches in terms of efficiency as well as the quality of the retrieved solution.

Th24 Structure Regularization for Structured Prediction

Xu Sun xusun@pku.edu.cn
Peking University

While there are many studies on weight regularization, the study on structure regularization is rare. Many existing systems on structured prediction focus on increasing the level of structural dependencies within the model. However, this trend could have been misdirected, because our study suggests that complex structures are actually harmful to generalization ability in structured prediction. To control structure-based overfitting, we propose a structure regularization framework via **structure decomposition**, which decomposes training samples into mini-samples with simpler structures, deriving a model with better generalization power. We show both theoretically and empirically that structure regularization can effectively control overfitting risk and lead to better accuracy. As a by-product, the proposed method can also substantially accelerate the training speed. The method and the theoretical results can apply to general graphical models with arbitrary structures. Experiments on well-known tasks demonstrate that our method can easily beat the benchmark systems on those highly-competitive tasks, achieving record-breaking accuracies yet with substantially faster training speed.

Th25 Expectation-Maximization for Learning Determinantal Point Processes

Jennifer Gillenwater jengi@cis.upenn.edu
University of Pennsylvania
Alex Kulesza kulesza@umich.edu
University of Michigan
Emily Fox ebfox@uw.edu
Ben Taskar taskar@cis.upenn.edu
University of Washington

A determinantal point process (DPP) is a probabilistic model of set diversity compactly parameterized by a positive semi-definite kernel matrix. To fit a DPP to a given task, we would like to learn the entries of its kernel matrix by maximizing the log-likelihood of the available data. However, log-likelihood is non-convex in the entries of the kernel matrix, and this learning problem is conjectured to be NP-hard. Thus, previous work has instead focused on more restricted convex learning settings: learning only a single weight for each row of the kernel matrix, or learning weights for a linear combination of DPPs with fixed kernel matrices. In this work we propose a novel algorithm for learning the full kernel matrix. By changing the kernel parameterization from matrix entries to eigenvalues and eigenvectors, and then lower-bounding the likelihood in the manner of expectation-maximization algorithms, we obtain an effective optimization procedure. We test our method on a real-world product recommendation task, and achieve relative gains of up to 16.5% in test log-likelihood compared to the naive approach of maximizing likelihood by projected gradient ascent on the entries of the kernel matrix.

Th26 Submodular Attribute Selection for Action Recognition in Video

Jingjing Zheng zjngjng@gmail.com
Rama Chellappa Rama@umiacs.umd.edu
University of Maryland
Zhuolin Jiang zhuolin.jiang@huawei.com
Noah's Ark Lab, Huawei Technologies
Jonathon Phillips jonathon@nist.gov
National Institute of Standards and Technology

In real-world action recognition problems, low-level features cannot adequately characterize the rich spatial-temporal structures in action videos. In this work, we encode actions based on attributes that describes actions as high-level concepts: \textit{e.g.}, jump forward and motion in the air. We base our analysis on two types of action attributes. One type of action attributes is generated by humans. The second type is data-driven attributes, which is learned from data using dictionary learning methods. Attribute-based representation may exhibit high variance due to noisy and redundant attributes. We propose a discriminative and compact attribute-based representation by selecting a subset of discriminative attributes from a large attribute set. Three attribute selection criteria are proposed and formulated as a submodular optimization problem. A greedy optimization algorithm is presented and guaranteed to be at least (1-1/e)-approximation to the optimum. Experimental results on the Olympic Sports and UCF101 datasets demonstrate that the proposed attribute-based representation can significantly boost the performance of action recognition algorithms and outperform most recently proposed recognition approaches.

Th27 Scale Adaptive Blind Deblurring

Haichao Zhang hczhang1@gmail.com
Duke University
Jianchao Yang jiayang@adobe.com
Adobe Research

The presence of noise and small scale structures usually leads to large kernel estimation errors in blind image deblurring empirically, if not a total failure. We present a scale space perspective on blind deblurring algorithms, and introduce a cascaded scale space formulation for blind deblurring. This new formulation suggests a natural approach robust to noise and small scale structures through tying the estimation across multiple scales and balancing the contributions of different scales automatically by learning from data. The proposed formulation also allows to handle non-uniform blur with a straightforward extension. Experiments are conducted on both benchmark dataset and real-world images to validate the effectiveness of the proposed method. One surprising finding based on our approach is that blur kernel estimation is not necessarily best at the finest scale.

Th28 Shape and Illumination from Shading using the Generic Viewpoint Assumption

Daniel Zoran danielzoran@gmail.com
Dilip Krishnan dilipkay@mit.edu
Bill Freeman billf@mit.edu
Massachusetts Institute of Technology
Jose Bento jose.bento@bc.edu
Boston College

The Generic Viewpoint Assumption (GVA) states that the position of the viewer or the light in a scene is not special. Thus, any estimated parameters from an observation should be stable under small perturbations such as object, viewpoint or light positions. The GVA has been analyzed and quantified in previous works, but has not been put to practical use in actual vision tasks. In this paper, we show how to utilize the GVA to estimate shape and illumination from a single shading image, without the use of other priors. We propose a novel linearized Spherical Harmonics (SH) shading model which enables us to obtain a computationally efficient form of the GVA term. Together with a data term, we build a model whose unknowns are shape and SH illumination. The model parameters are estimated using the Alternating Direction Method of Multipliers embedded in a multi-scale estimation framework. In this prior-free framework, we obtain competitive shape and illumination estimation results under a variety of models and lighting conditions, requiring fewer assumptions than competing methods.

Th29 Depth Map Prediction from a Single Image using a Multi-Scale Deep Network

David Eigen deigen@cs.nyu.edu
Christian Puhrsch cp1995@cs.nyu.edu
Rob Fergus fergus@cs.nyu.edu
New York University

Predicting depth is an essential component in understanding the 3D geometry of a scene. While for stereo images local correspondence suffices for estimation, finding depth relations from a single image is less straightforward, requiring integration of both global and local information from various cues. Moreover, the task is inherently ambiguous, with a large source of uncertainty coming from the overall scale. In this paper, we present a new method that addresses this task by employing two deep network stacks: one that makes a coarse global prediction based on the entire image, and another that refines this prediction locally. We also apply a scale-invariant error to help measure depth relations rather than scale. By leveraging the raw datasets as large sources of training data, our method achieves state-of-the-art results on both NYU Depth and KITTI, and matches detailed depth boundaries without the need for superpixelation.

Th30 Self-Paced Learning with Diversity

Lu Jiang lujiang@cs.cmu.edu
Deyu Meng deyum@cs.cmu.edu
Shoou-I Yu iyu@cs.cmu.edu
Zhenzhong Lan lanzhzh@cs.cmu.edu
Alexander Hauptmann alex@cs.cmu.edu
Carnegie Mellon University
Shiguang Shan sgshan@ict.ac.cn

Self-paced learning (SPL) is a recently proposed learning regime inspired by the learning process of humans and animals that gradually incorporates easy to more complex samples into training. Existing methods are limited in that they ignore an important aspect in learning: diversity. To incorporate this information, we propose an approach called self-paced learning with diversity (SPLD) which formalizes the preference for both easy and diverse samples into a general regularizer. This regularization term is independent of the learning objective, and thus can be easily generalized into various learning tasks. Albeit non-convex, the optimization of the variables included in this SPLD regularization term for sample selection can be globally solved in linearithmic time. We demonstrate that our method significantly outperforms the conventional SPL on three real-world datasets. Specifically, SPLD achieves the best MAP so far reported in literature on the Hollywood2 and Olympic Sports datasets.

Th31 Hardness of parameter estimation in graphical models

Guy Bresler gbresler@mit.edu
David Gamarnik gamarnik@mit.edu
Devavrat Shah devavrat@mit.edu
Massachusetts Institute of Technology

We consider the problem of learning the canonical parameters specifying an undirected graphical model (Markov random field) from the mean parameters. For graphical models representing a minimal exponential family, the canonical parameters are uniquely determined by the mean parameters, so the problem is feasible in principle. The goal of this paper is to investigate the computational feasibility of this statistical task. Our main result shows that parameter estimation is in general intractable: no algorithm can learn the canonical parameters of a generic pair-wise binary graphical model from the mean parameters in time bounded by a polynomial in the number of variables (unless RP = NP). Indeed, such a result has been believed to be true (see the monograph by Wainwright and Jordan)

but no proof was known. Our proof gives a polynomial time reduction from approximating the partition function of the hard-core model, known to be hard, to learning approximate parameters. Our reduction entails showing that the marginal polytope boundary has an inherent repulsive property, which validates an optimization procedure over the polytope that does not use any knowledge of its structure (as required by the ellipsoid method and others).

Th32 Sequential Monte Carlo for Graphical Models

Christian Andersson Naesseth christian.a.naesseth@liu.se
Fredrik Lindsten lindsten@isy.liu.se
Linköping University
Thomas Schön thomas.schon@it.uu.se
Uppsala University

We propose a new framework for how to use sequential Monte Carlo (SMC) algorithms for inference in probabilistic graphical models (PGM). Via a sequential decomposition of the PGM we find a sequence of auxiliary distributions defined on a monotonically increasing sequence of probability spaces. By targeting these auxiliary distributions using SMC we are able to approximate the full joint distribution defined by the PGM. One of the key merits of the SMC sampler is that it provides an unbiased estimate of the partition function of the model. We also show how it can be used within a particle Markov chain Monte Carlo framework in order to construct high-dimensional block-sampling algorithms for general PGMs.

Th33 Concavity of reweighted Kikuchi approximation

Po-Ling Loh loh@wharton.upenn.edu
University of Pennsylvania
Andre Wibisono wibisono@berkeley.edu
UC Berkeley

We analyze a reweighted version of the Kikuchi approximation for estimating the log partition function of a product distribution defined over a region graph. We establish sufficient conditions for the concavity of our reweighted objective function in terms of weight assignments in the Kikuchi expansion, and show that a reweighted version of the sum product algorithm applied to the Kikuchi region graph will produce global optima of the Kikuchi approximation whenever the algorithm converges. When the region graph has two layers, corresponding to a Bethe approximation, we show that our sufficient conditions for concavity are also necessary. Finally, we provide an explicit characterization of the polytope of concavity in terms of the cycle structure of the region graph. We conclude with simulations that demonstrate the advantages of the reweighted Kikuchi approach.

Th34 A Complete Variational Tracker

Ryan Turner Ryan.Turner@ngc.com
Steven Bottone steven.bottone@ngc.com
Northrop Grumman
Bhargav Avasarala bhargav.avasarala@ngc.com
Northrop Grumman Corp

We introduce a novel probabilistic tracking algorithm that incorporates combinatorial data association constraints and model-based track management using variational Bayes. We use a Bethe entropy approximation to incorporate data association constraints that are often ignored in previous probabilistic tracking algorithms. Noteworthy aspects of our method include a model-based mechanism to replace heuristic logic typically used to initiate and destroy tracks, and an assignment posterior with linear computation cost in window length as opposed to the exponential scaling of previous MAP-based approaches. We demonstrate the applicability of our method on radar tracking and computer vision problems.

Th35 Joint Training of a Convolutional Network and a Graphical Model for Human Pose Estimation

Jonathan Tompson tompson@cims.nyu.edu
Arjun Jain ajain@nyu.edu
Christoph Bregler chris.bregler@nyu.edu
Yann LeCun yann@cs.nyu.edu
New York University

This paper proposes a new hybrid architecture that consists of a deep Convolutional Network and a Markov Random Field. We show how this architecture is successfully applied to the challenging problem of articulated human pose estimation in monocular images. The architecture can exploit structural domain constraints such as geometric relationships between body joint locations. We show that joint training of these two model paradigms improves performance and allows us to significantly outperform existing state-of-the-art techniques.

Th36 Clamping Variables and Approximate Inference

Adrian Weller adrian@cs.columbia.edu
Tony Jebara jebara@cs.columbia.edu
Columbia University

It was recently proved using graph covers (Ruozzi, 2012) that the Bethe partition function is upper bounded by the true partition function for a binary pairwise model that is attractive. Here we provide a new, arguably simpler proof from first principles. We make use of the idea of clamping a variable to a particular value. For an attractive model, we show that summing over the Bethe partition functions for each sub-model obtained after clamping any variable can only raise (and hence improve) the approximation. In fact, we derive a stronger result that may have other useful implications. Repeatedly clamping until we obtain a model with no cycles, where the Bethe approximation is exact, yields the result. We also provide a related lower bound on a broad class of approximate partition functions of general pairwise multi-label models that depends only on the topology. We demonstrate that clamping a few wisely chosen variables can be of practical value by dramatically reducing approximation error.

Th37 Divide-and-Conquer Learning by Anchoring a Conical Hull

Tianyi Zhou tianyi.david.zhou@gmail.com
Jeff Bilmes bilmes@ee.washington.edu
Carlos Guestrin guestrin@cs.washington.edu
University of Washington

We reduce a broad class of machine learning problems, usually addressed by EM or sampling, to the problem of finding the k extremal rays spanning the conical hull of a data point set. These k``anchors'' lead to a global solution and a more interpretable model that can even outperform EM and sampling on generalization error. To find the $k$ anchors, we propose a novel divide-and-conquer learning scheme ``DCA'' that distributes the problem to $\mathcal{O}(k\log k)$ same-type sub-problems on different low-D random hyperplanes, each can be solved by any solver. For the 2D sub-problem, we present a non-iterative solver that only needs to compute an array of cosine values and its max/min entries. DCA also provides a faster subroutine for other methods to check whether a point is covered in a conical hull, which improves algorithm design in multiple dimensions and brings significant speedup to learning. We apply our method to GMM, HMM, LDA, NMF and subspace clustering, then show its competitive performance and scalability over other methods on rich datasets.

Th8 Mode Estimation for High Dimensional Discrete Tree Graphical Models

Chao Chen chao.chen.cchen@gmail.com
Dimitris Metaxas dnm@cs.rutgers.edu
Rutgers University
Han Liu hanliu@princeton.edu
Tianqi Zhao tianqi@princeton.edu
Princeton University

This paper studies the following problem: given samples from a high dimensional discrete distribution, we want to estimate the leading (δ,ρ)-modes of the underlying distributions. A point is defined to be a (δ,ρ)-mode if it is a local optimum of the density within a δ-neighborhood under metric ρ. As we increase the ``scale'' parameter δ, the neighborhood size increases and the total number of modes monotonically decreases. The sequence of the (δ,ρ)-modes reveal intrinsic topographical information of the underlying distributions. Though the mode finding problem is generally intractable in high dimensions, this paper unveils that, if the distribution can be approximated well by a tree graphical model, mode characterization is significantly easier. An efficient algorithm with provable theoretical guarantees is proposed and is applied to applications like data analysis and multiple predictions.

Th39 Graphical Models for Recovering Probabilistic and Causal Queries from Missing Data

Karthika Mohan karthika@cs.ucla.edu
Judea Pearl judea@cs.ucla.edu
UCLA

We address the problem of deciding whether a causal or probabilistic query is estimable from data corrupted by missing entries, given a model of missingness process. We extend the results of Mohan et al, 2013 by presenting more general conditions for recovering probabilistic queries of the form P(y|x) and P(y,x) as well as causal queries of the form P(y|do(x)). We show that causal queries may be recoverable even when the factors in their identifying estimands are not recoverable. Specifically, we derive graphical conditions for recovering causal effects of the form P(y|do(x)) when Y and its missingness mechanism are not d-separable. Finally, we apply our results to problems of attrition and characterize the recovery of causal effects from data corrupted by attrition.

Th40 Randomized Experimental Design for Causal Graph Discovery

Huining Hu huining.hu@mail.mcgill.ca
Adrian Vetta vetta@math.mcgill.ca
McGill University
Zhentao Li zhentao.li@ens.fr
École normale supérieure

We examine the number of controlled experiments required to discover a causal graph. Hauser and Buhlmann showed that the number of experiments required is logarithmic in the cardinality of maximum undirected clique in the essential graph. Their lower bounds, however, assume that the experiment designer cannot use randomization in selecting the experiments. We show that significant improvements are possible with the aid of randomization – in an adversarial (worst-case) setting, the designer can then recover the causal graph using at most O(log log n) experiments in expectation. This bound cannot be improved; we show it is tight for some causal graphs. We then show that in a non-adversarial (average-case) setting, even larger improvements are possible: if the causal graph is chosen uniformly at random under a Erdös-Rényi model then the expected number of experiments to discover the causal graph is constant. Finally, we present computer simulations to complement our theoretic results. Our work exploits a structural characterization of essential graphs by Andersson et al. Their characterization is based upon a set of orientation forcing operations. Our results show a distinction between which forcing operations are most important in worst-case and average-case settings.

Th41 Optimization Methods for Sparse Pseudo-Likelihood Graphical Model Selection

Sang Oh syoh@lbl.gov
Lawrence Berkeley Lab
Onkar Dalal onkar@alumni.stanford.edu
Kshitij Khare kdkhare@stat.ufl.edu
University of Florida
Bala Rajaratnam brajarat@stanford.edu
Stanford University

Sparse high dimensional graphical model selection is a popular topic in contemporary machine learning. To this end, various useful approaches have been proposed in the context of ℓ_1 penalized estimation in the Gaussian framework. Though many of these approaches are demonstrably scalable and have leveraged recent advances in convex optimization, they still depend on the Gaussian functional form. To address this gap, a convex pseudo-likelihood based partial correlation graph

estimation method (CONCORD) has been recently proposed. This method uses cyclic coordinate-wise minimization of a regression based pseudo-likelihood, and has been shown to have robust model selection properties in comparison with the Gaussian approach. In direct contrast to the parallel work in the Gaussian setting however, this new convex pseudo-likelihood framework has not leveraged the extensive array of methods that have been proposed in the machine learning literature for convex optimization. In this paper, we address this crucial gap by proposing two proximal gradient methods (CONCORD-ISTA and CONCORD-FISTA) for performing ℓ_1-regularized inverse covariance matrix estimation in the pseudo-likelihood framework. We present timing comparisons with coordinate-wise minimization and demonstrate that our approach yields tremendous pay offs for ℓ_1-penalized partial correlation graph estimation outside the Gaussian setting, thus yielding the fastest and most scalable approach for such problems. We undertake a theoretical analysis of our approach and rigorously demonstrate convergence, and also derive rates thereof.

Th42 Learning Chordal Markov Networks by Dynamic Programming

Kustaa Kangas jwkangas@cs.helsinki.fi
Teppo Niinimäki teppo.niinimaki@cs.helsinki.fi
Mikko Koivisto mikko.koivisto@cs.helsinki.fi
Helsinki Institute for Information Technology

We present an algorithm for finding a chordal Markov network that maximizes any given decomposable scoring function. The algorithm is based on a recursive characterization of clique trees, and it runs in O(4^n) time for n vertices. On an eight-vertex benchmark instance, our implementation turns out to be about ten million times faster than a recently proposed, constraint satisfaction based algorithm (Corander et al., NIPS 2013). Within a few hours, it is able to solve instances up to 18 vertices, and beyond if we restrict the maximum clique size. We also study the performance of a recent integer linear programming algorithm (Bartlett and Cussens, UAI 2013). Our results suggest that, unless we bound the clique sizes, currently only the dynamic programming algorithm is guaranteed to solve instances with around 15 or more vertices.

Th43 QUIC & DIRTY: A Quadratic Approximation Approach for Dirty Statistical Models

Cho-Jui Hsieh cjhsieh@cs.utexas.edu
Pradeep Ravikumar pradeepr@cs.utexas.edu
Inderjit Dhillon inderjit@cs.utexas.edu
University of Texas
Stephen Becker stephen.beckr@gmail.com
University of Colorado
Peder Olsen pederao@us.ibm.com
IBM

In this paper, we develop a family of algorithms for optimizing “superposition-structured” or “dirty” statistical estimators for high-dimensional problems involving the minimization of the sum of a smooth loss function with a hybrid regularization. Most of the current approaches are first-order methods, including proximal gradient or Alternating Direction Method of Multipliers (ADMM). We propose a new family of second-order methods where we approximate the loss function using quadratic approximation. The superposition structured regularizer then leads to a subproblem that can be efficiently solved by alternating minimization. We propose a general active subspace selection approach to speed up the solver by utilizing the low-dimensional structure given by the regularizers, and provide convergence guarantees for our algorithm. Empirically, we show that our approach is more than 10 times faster than state-of-the-art first-order approaches for the latent variable graphical model selection problems and multi-task learning problems when there is more than one regularizer. For these problems, our approach appears to be the first algorithm that can extend active subspace ideas to multiple regularizers.

Th44 Advances in Learning Bayesian Networks of Bounded Treewidth

Siqi Nie nies@rpi.edu
Qiang Ji qji@ecse.rpi.edu
RPI
Denis Maua denis@idsia.ch
USP
Cassio de Campos cassiopc@acm.org
Queen’s University Belfast

This work presents novel algorithms for learning Bayesian networks of bounded treewidth. Both exact and approximate methods are developed. The exact method combines mixed integer linear programming formulations for structure learning and treewidth computation. The approximate method consists in sampling k-trees (maximal graphs of treewidth k), and subsequently selecting, exactly or approximately, the best structure whose moral graph is a subgraph of that k-tree. The approaches are empirically compared to each other and to state-of-the-art methods on a collection of public data sets with up to 100 variables.

Th45 Recursive Inversion Models for Permutations

Christopher Meek meek@microsoft.com
Microsoft Research
Marina Meila mmp@stat.washington.edu
University of Washington

We develop a new exponential family probabilistic model for permutations that can capture hierarchical structure, and that has the well known Mallows and generalized Mallows models as subclasses. We describe how one can do parameter estimation and propose an approach to structure search for this class of models. We provide experimental evidence that this added flexibility both improves predictive performance and enables a deeper understanding of collections of permutations.

Th46 Discovering Structure in High-Dimensional Data Through Correlation Explanation

Greg Ver Steeg gregv@isi.edu
Aram Galstyan galstyan@isi.edu
University of Southern California

We introduce a method to learn a hierarchy of successively more abstract representations of complex data based on optimizing an information-theoretic objective. Intuitively, the optimization searches for a set of latent factors that best explain the correlations in the data as measured by multivariate mutual information. The method is unsupervised, requires no model assumptions, and scales linearly with the number of variables which makes it an attractive approach for very high dimensional systems. We demonstrate that Correlation Explanation (CorEx) automatically discovers meaningful structure for data from diverse sources including personality tests, DNA, and human language.

Th47 Coresets for k-Segmentation of Streaming Data

Guy Rosman rosman@csail.mit.edu
Mikhail Volkov mikhail@csail.mit.edu
Dan Feldman dannyf@caltech.edu
John Fisher III fisher@csail.mit.edu
Daniela Rus rus@csail.mit.edu
Massachusetts Institute of Technology

Life-logging video streams, financial time series, and Twitter tweets are a few examples of high-dimensional signals over practically unbounded time. We consider the problem of computing optimal segmentation of such signals by k-piecewise linear function, using only one pass over the data by maintaining a coreset for the signal. The coreset enables fast further analysis such as automatic summarization and analysis of such signals. A coreset (core-set) is a compact representation of the data seen so far, which approximates the data well for a specific task -- in our case, segmentation of the stream. We show that, perhaps surprisingly, the segmentation problem admits coresets of cardinality only linear in the number of segments k, independently of both the dimension d of the signal, and its number n of points. More precisely, we construct a representation of size O(klog n /eps^2) that provides a (1+eps)-approximation for the sum of squared distances to any given k-piecewise linear function. Moreover, such coresets can be constructed in a parallel streaming approach. Our results rely on a novel eduction of statistical estimations to problems in computational geometry. We empirically evaluate our algorithms on very large synthetic and real data sets from GPS, video and financial domains, using 255 machines in Amazon cloud.

Th48 Approximating Hierarchical MV-sets for Hierarchical Clustering

Assaf Glazer assafgr@cs.technion.ac.il
Omer Weissbrod omer.we@gmail.com
Michael Lindenbaum mic@cs.technion.ac.il
Shaul Markovitch shaulm@cs.technion.ac.il
Technion

The goal of hierarchical clustering is to construct a cluster tree, which can be viewed as the modal structure of a density. For this purpose, we use a convex optimization program that can efficiently estimate a family of hierarchical dense sets in high-dimensional distributions. We further extend existing graph-based methods to approximate the cluster tree of a distribution. By avoiding direct density estimation, our method is able to handle high-dimensional data more efficiently than existing density-based approaches. We present empirical results that demonstrate the superiority of our method over existing ones.

Th49 Tight Continuous Relaxation of the Balanced k-Cut Problem

Syama Sundar Rangapuram srangapu@mpi-inf.mpg.de
Matthias Hein hein@cs.uni-saarland.de
Saarland University
Pramod Mudrakarta pramodkm@uchicago.edu
University of Chicago

Spectral Clustering as a relaxation of the normalized/ratio cut has become one of the standard graph-based clustering methods. Existing methods for the computation of multiple clusters, corresponding to a balanced k-cut of the graph, are either based on greedy techniques or heuristics which have weak connection to the original motivation of minimizing the normalized cut. In this paper we propose a new tight continuous relaxation for any balanced k-cut problem and show that a related recently proposed relaxation is in most cases loose leading to poor performance in practice. For the optimization of our tight continuous relaxation we propose a new algorithm for the hard sum-of-ratios minimization problem which achieves monotonic descent. Extensive comparisons show that our method beats all existing approaches for ratio cut and other balanced k-cut criteria.

Th50 Streaming, Memory Limited Algorithms for Community Detection

Se-Young Yun seyoung.yun@inria.fr
marc lelarge marc.lelarge@ens.fr
INRIA ENS
Alexandre Proutiere alepro@kth.se
KTH

In this paper, we consider sparse networks consisting of a finite number of non-overlapping communities, i.e. disjoint clusters, so that there is higher density within clusters than across clusters. Both the intra- and inter-cluster edge densities vanish when the size of the graph grows large, making the cluster reconstruction problem nosier and hence difficult to solve. We are interested in scenarios where the network size is very large, so that the adjacency matrix of the graph is hard to manipulate and store. The data stream model in which columns of the adjacency matrix are revealed sequentially constitutes a natural framework in this setting. For this model, we develop two novel clustering algorithms that extract the clusters asymptotically accurately. The first algorithm is {\ it offline}, as it needs to store and keep the assignments of nodes to clusters, and requires a memory that scales linearly with the network size. The second algorithm is {\it online}, as it may classify a node when the corresponding column is revealed and then discard this information. This algorithm requires a memory growing sub-linearly with the network size. To construct these efficient streaming

memory-limited clustering algorithms, we first address the problem of clustering with partial information, where only a small proportion of the columns of the adjacency matrix is observed and develop, for this setting, a new spectral algorithm which is of independent interest.

Th51 Computing Nash Equilibria in Generalized Interdependent Security Games

Hau Chan hauchan@cs.stonybrook.edu
Luis Ortiz leortiz@cs.stonybrook.edu
Stony Brook University

We study the computational complexity of computing Nash equilibria in generalized interdependent-security (IDS) games. Like traditional IDS games, originally introduced by economists and risk-assessment experts Heal and Kunreuther about a decade ago, generalized IDS games model agents' voluntary investment decisions when facing potential direct risk and transfer risk exposure from other agents. A distinct feature of generalized IDS games, however, is that full investment can reduce transfer risk. As a result, depending on the transfer-risk reduction level, generalized IDS games may exhibit strategic complementarity (SC) or strategic substitutability (SS). We consider three variants of generalized IDS games in which players exhibit only SC, only SS, and both SC+SS. We show that determining whether there is a pure-strategy Nash equilibrium (PSNE) in SC+SS-type games is NP-complete, while computing a single PSNE in SC-type games takes worst-case polynomial time. As for the problem of computing all mixed-strategy Nash equilibria (MSNE) efficiently, we produce a partial characterization. Whenever each agent in the game is indiscriminate in terms of the transfer-risk exposure to the other agents, a case that Kearns and Ortiz originally studied in the context of traditional IDS games in their NIPS 2003 paper, we can compute all MSNE that satisfy some ordering constraints in polynomial time in all three game variants. Yet, there is a computational barrier in the general (transfer) case: we show that the computational problem is as hard as the Pure-Nash-Extension problem, also originally introduced by Kearns and Ortiz, and that it is NP complete for all three variants. Finally, we experimentally examine and discuss the practical impact that the additional protection from transfer risk allowed in generalized IDS games has on MSNE by solving several randomly-generated instances of SC+SS-type games with graph structures taken from several real-world datasets.

Th52 Causal Strategic Inference in Networked Microfinance Economies

Mohammad Irfan mirfan@bowdoin.edu
Bowdoin College
Luis Ortiz leortiz@cs.stonybrook.edu
Stony Brook University

Performing interventions is a major challenge in economic policy-making. We propose **causal strategic inference** as a framework for conducting interventions and apply it to large, networked microfinance economies. The basic solution platform consists of modeling a microfinance market as a networked economy, learning the parameters of the model from the real-world microfinance data, and designing algorithms for various computational problems in question. We adopt Nash equilibrium as the solution concept for our model. For a special case of our model, we show that an equilibrium point always exists and that the equilibrium interest rates are unique. For the general case, we give a constructive proof of the existence of an equilibrium point. Our empirical study is based on the microfinance data from Bangladesh and Bolivia, which we use to first learn our models. We show that causal strategic inference can assist policy-makers by evaluating the outcomes of various types of interventions, such as removing a loss-making bank from the market, imposing an interest rate cap, and subsidizing banks.

Th53 A Statistical Decision-Theoretic Framework for Social Choice

Hossein Azari Soufiani azari@fas.harvard.edu
David Parkes parkes@eecs.harvard.edu
Lirong Xia xialirong@gmail.com
Harvard University

In this paper, we take a statistical decision-theoretic viewpoint on social choice, putting a focus on the decision to be made on behalf of a system of agents. In our framework, we are given a statistical ranking model, a decision space, and a loss function defined on (parameter, decision) pairs, and formulate social choice mechanisms as decision rules that minimize expected loss. This suggests a general framework for the design and analysis of new social choice mechanisms. We compare Bayesian estimators, which minimize Bayesian expected loss, for the Mallows model and the Condorcet model respectively, and the Kemeny rule. We consider various normative properties, in addition to computational complexity and asymptotic behavior. In particular, we show that the Bayesian estimator for the Condorcet model satisfies some desired properties such as anonymity, neutrality, and monotonicity, can be computed in polynomial time, and is asymptotically different from the other two rules when the data are generated from the Condorcet model for some ground truth parameter.

Th54 Learning Optimal Commitment to Overcome Insecurity

Avrim Blum avrim@cs.cmu.edu
Nika Haghtalab nika@cmu.edu
Ariel Procaccia arielpro@cs.cmu.edu
Carnegie Mellon University

Game-theoretic algorithms for physical security have made an impressive real-world impact. These algorithms compute an optimal strategy for the defender to commit to in a Stackelberg game, where the attacker observes the defender's strategy and best-responds. In order to build the game model, though, the payoffs of potential attackers for various outcomes must be estimated; inaccurate estimates can lead to significant inefficiencies. We design an algorithm that optimizes the defender's strategy with no prior information, by observing the attacker's responses to randomized deployments of resources and learning his priorities. In contrast to previous work, our algorithm requires a number of queries that is polynomial in the representation of the game.

Th55 Diverse Randomized Agents Vote to Win

Albert Jiang jiangx@usc.edu
Milind Tambe tambe@usc.edu
Leandro Soriano Marcolino sorianom@usc.edu
University of Southern California
Ariel Procaccia arielpro@cs.cmu.edu
Tuomas Sandholm sandholm@cs.cmu.edu
Nisarg Shah nkshah@cs.cmu.edu
Carnegie Mellon University

We investigate the power of voting among diverse, randomized software agents. With teams of computer Go agents in mind, we develop a novel theoretical model of two-stage noisy voting that builds on recent work in machine learning. This model allows us to reason about a collection of agents with different biases (determined by the first-stage noise models), which, furthermore, apply randomized algorithms to evaluate alternatives and produce votes (captured by the second-stage noise models). We analytically demonstrate that a uniform team, consisting of multiple instances of any single agent, must make a significant number of mistakes, whereas a diverse team converges to perfection as the number of agents grows. Our experiments, which pit teams of computer Go agents against strong agents, provide evidence for the effectiveness of voting when agents are diverse.

Th56 Fairness in Multi-Agent Sequential Decision-Making

Chongjie Zhang chongjie@csail.mit.edu
Julie Shah julie_a_shah@csail.mit.edu
Massachusetts Institute of Technology

We define a fairness solution criterion for multi-agent decision-making problems, where agents have local interests. This new criterion aims to maximize the worst performance of agents with consideration on the overall performance. We develop a simple linear programming approach and a more scalable game-theoretic approach for computing an optimal fairness policy. This game-theoretic approach formulates this fairness optimization as a two-player, zero-sum game and employs an iterative algorithm for finding a Nash equilibrium, corresponding to an optimal fairness policy. We scale up this approach by exploiting problem structure and value function approximation. Our experiments on resource allocation problems show that this fairness criterion provides a more favorable solution than the utilitarian criterion, and that our game-theoretic approach is significantly faster than linear programming.

Th57 Repeated Contextual 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

Motivated by real-time advertising exchanges, we analyze the problem of pricing inventory in a repeated posted-price auction. We consider both the cases of a truthful and surplus-maximizing buyer, where the former makes decisions myopically on every round, and the latter may strategically react to our algorithm, forgoing short-term surplus in order to trick the algorithm into setting better prices in the future. We further assume a buyer's valuation of a good is a function of a context vector that describes the good being sold. We give the first algorithm attaining sublinear (O(T^{2/3})) regret in the contextual setting against a surplus-maximizing buyer. We also extend this result to repeated second-price auctions with multiple buyers.

Th58 Universal Option Models

hengshuai yao hengshuai@gmail.com
Csaba Szepesvari szepesva@cs.ualberta.ca
Rich Sutton rich@richsutton.com
Joseph Modayil jmodayil@cs.ualberta.ca
University of Alberta
Shalabh Bhatnagar shalabh@csa.iisc.ernet.in
Indian Institute of Science

We consider the problem of learning models of options for real-time abstract planning, in the setting where reward functions can be specified at any time and their expected returns must be efficiently computed. We introduce a new model for an option that is independent of any reward function, called the {\it universal option model (UOM)}. We prove that the UOM of an option can construct a traditional option model given a reward function, and the option-conditional return is computed directly by a single dot-product of the UOM with the reward function. We extend the UOM to linear function approximation, and we show it gives the TD solution of option returns and value functions of policies over options. We provide a stochastic approximation algorithm for incrementally learning UOMs from data and prove its consistency. We demonstrate our method in two domains. The first domain is document recommendation, where each user query defines a new reward function and a document's relevance is the expected return of a simulated random-walk through the document's references. The second domain is a real-time strategy game, where the controller must select the best game unit to accomplish dynamically-specified tasks. Our experiments show that UOMs are substantially more efficient in evaluating option returns and policies than previously known methods.

Th59 Semi-Separable Hamiltonian Monte Carlo for Inference in Bayesian Hierarchical Models

Yichuan Zhang Y.Zhang-60@sms.ed.ac.uk
Charles Sutton csutton@inf.ed.ac.uk
University of Edinburgh

Sampling from hierarchical Bayesian models is often difficult for MCMC methods, because of the strong correlations between the model parameters and the hyperparameters. Recent Riemannian manifold Hamiltonian Monte Carlo (RMHMC) methods have significant potential advantages in this setting, but are computationally expensive. We introduce a new RMHMC method, which we call semi-separable Hamiltonian Monte Carlo, which uses a specially designed mass matrix that allows the joint Hamiltonian over model parameters and hyperparameters to decompose into two simpler Hamiltonians. This structure is exploited by

a new integrator which we call the alternating blockwise leapfrog algorithm. The resulting method can mix faster than simpler Gibbs sampling while being simpler and more efficient than previous instances of RMHMC.

Th60 Bayesian Sampling Using Stochastic Gradient Thermostats

Nan Ding dingnan@google.com
Google Inc.
Youhan Fang yfang@purdue.edu
Robert Skeel skeel@cs.purdue.edu
Purdue University
Ryan Babbush ryanbabbush@gmail.com
Harvard University
Changyou Chen cchangyou@gmail.com
Duke University
Hartmut Neven neven@google.com
Google

Dynamics-based sampling methods, such as Hybrid Monte Carlo (HMC) and Langevin dynamics (LD), are commonly used to sample target distributions. Recently, such approaches have been combined with stochastic gradient techniques to increase sampling efficiency when dealing with large datasets. An outstanding problem with this approach is that the stochastic gradient introduces an unknown amount of noise which can prevent proper sampling after discretization. To remedy this problem, we show that one can leverage a small number of additional variables in order to stabilize momentum fluctuations induced by the unknown noise. Our method is inspired by the idea of a thermostat in statistical physics and is justified by a general theory.

Th61 Poisson Process Jumping between an Unknown Number of Rates: Application to Neural Spike Data

Florian Stimberg florian.stimberg@tu-berlin.de
Andreas Ruttor andreas.ruttor@tu-berlin.de
Manfred Opper manfred.opper@tu-berlin.de
TU Berlin

We introduce a model where the rate of an inhomogeneous Poisson process is modified by a Chinese restaurant process. Applying a MCMC sampler to this model allows us to do posterior Bayesian inference about the number of states in Poisson-like data. Our sampler is shown to get accurate results for synthetic data and we apply it to V1 neuron spike data to find discrete firing rate states depending on the orientation of a stimulus.

Th63 Distributed Variational Inference in Sparse Gaussian Process Regression and Latent Variable Models

Yarin Gal yg279@cam.ac.uk
Mark van der Wilk mv310@cam.ac.uk
Carl Rasmussen cer54@cam.ac.uk
University of Cambridge

Gaussian processes (GPs) are a powerful tool for probabilistic inference over functions. They have been applied to both regression and non-linear dimensionality reduction, and offer desirable properties such as uncertainty estimates, robustness to over-fitting, and principled ways for tuning hyper-parameters. However the scalability of these models to big datasets remains an active topic of research. We introduce a novel re-parametrisation of variational inference for sparse GP regression and latent variable models that allows for an efficient distributed algorithm. This is done by exploiting the decoupling of the data given the inducing points to re-formulate the evidence lower bound in a Map-Reduce setting. We show that the inference scales well with data and computational resources, while preserving a balanced distribution of the load among the nodes. We further demonstrate the utility in scaling Gaussian processes to big data. We show that GP performance improves with increasing amounts of data in regression (on flight data with 2 million records) and latent variable modelling (on MNIST). The results show that GPs perform better than many common models often used for big data.

Th64 Incremental Local Gaussian Regression

Franziska Meier fmeier@usc.edu
Stefan Schaal sschaal@usc.edu
USC
Philipp Hennig phennig@tuebingen.mpg.de
MPI Tübingen

Locally weighted regression (LWR) was created as a nonparametric method that can approximate a wide range of functions, is computationally efficient, and can learn continually from very large amounts of incrementally collected data. As an interesting feature, LWR can regress on non-stationary functions, a beneficial property, for instance, in control problems. However, it does not provide a proper generative model for function values, and existing algorithms have a variety of manual tuning parameters that strongly influence bias, variance and learning speed of the results. Gaussian (process) regression, on the other hand, does provide a generative model with rather black-box automatic parameter tuning, but it has higher computational cost, especially for big data sets and if a non-stationary model is required. In this paper, we suggest a path from Gaussian (process) regression to locally weighted regression, where we retain the best of both approaches. Using a localizing function basis and approximate inference techniques, we build a Gaussian (process) regression algorithm of increasingly local nature and similar computational complexity to LWR. Empirical evaluations are performed on several synthetic and real robot datasets of increasing complexity and (big) data scale, and demonstrate that we consistently achieve on par or superior performance compared to current state-of-the-art methods while retaining a principled approach to fast incremental regression with minimal manual tuning parameters.

Th65 Augur: Data-Parallel Probabilistic Modeling

Jean-Baptiste Tristan jean.baptiste.tristan@oracle.com
Adam Pocock adam.pocock@oracle.com
Stephen Green stephen.x.green@oracle.com
Guy Steele guy.steele@oracle.com
Oracle Labs
Daniel Huang dehuang@fas.harvard.edu
Harvard University
Joseph Tassarotti jtassaro@cs.cmu.edu
Carnegie Mellon University

Implementing inference procedures for each new probabilistic model is time-consuming and error-prone. Probabilistic programming addresses this problem by allowing a user to specify the model and then automatically generating the inference procedure. To make this practical it is important to generate high performance inference code. In turn, on modern architectures, high performance requires parallel execution. In this paper we present Augur, a probabilistic modeling language and compiler for Bayesian networks designed to make effective use of data-parallel architectures such as GPUs. We show that the compiler can generate data-parallel inference code scalable to thousands of GPU cores by making use of the conditional independence relationships in the Bayesian network.

Th66 Just-In-Time Learning for Fast and Flexible Inference

Ali Eslami s.m.eslami@sms.ed.ac.uk
Daniel Tarlow dtarlow@cs.toronto.edu
Pushmeet Kohli pkohli@microsoft.com
John Winn jwinn@microsoft.com
Microsoft Research

Much of research in machine learning has centered around the search for inference algorithms that are both general-purpose and efficient. The problem is extremely challenging and general inference remains computationally expensive. We seek to address this problem by observing that in most specific applications of a model, we typically only need to perform a small subset of all possible inference computations. Motivated by this, we introduce just-in-time learning, a framework for fast and flexible inference that learns to speed up inference at run-time. Through a series of experiments, we show how this framework can allow us to combine the flexibility of sampling with the efficiency of deterministic message-passing.

TH67 Distributed Bayesian Posterior Sampling via Moment Sharing

Minjie Xu chokkyvista06@gmail.com
Jun Zhu dcszj@mail.tsinghua.edu.cn
Bo Zhang dcszb@mail.tsinghua.edu.cn
Tsinghua University
Balaji Lakshminarayanan balaji@gatsby.ucl.ac.uk
Gatsby Unit, University College London
Yee Whye Teh y.w.teh@stats.ox.ac.uk
University of Oxford

We propose a distributed Markov chain Monte Carlo (MCMC) inference algorithm for large scale Bayesian posterior simulation. We assume that the dataset is partitioned and stored across nodes of a cluster. Our procedure involves an independent MCMC posterior sampler at each node based on its local partition of the data. Moment statistics of the local posteriors are collected from each sampler and propagated across the cluster using expectation propagation message passing with low communication costs. The moment sharing scheme improves posterior estimation quality by enforcing agreement among the samplers. We demonstrate the speed and inference quality of our method with empirical studies on Bayesian logistic regression and sparse linear regression with a spike-and-slab prior.

Th68 Communication Efficient Distributed Machine Learning with the Parameter Server

Mu Li muli@cs.cmu.edu
David Andersen dga@cs.cmu.edu
Alex Smola alex@smola.org
Carnegie Mellon University
Kai Yu kai.yu.cool@gmail.com
Baidu

This paper describes a third-generation parameter server framework for distributed machine learning. This framework offers two relaxations to balance system performance and algorithm efficiency. We propose a new algorithm that takes advantage of this framework to solve non-convex non-smooth problems with convergence guarantees. We present an in-depth analysis of two large scale machine learning problems ranging from ℓ_1-regularized logistic regression on CPUs to reconstruction ICA on GPUs, using 636TB of real data with hundreds of billions of samples and dimensions. We demonstrate using these examples that the parameter server framework is an effective and straightforward way to scale machine learning to larger problems and systems than have been previously achieved.

Th69 On Model Parallelization and Scheduling Strategies for Distributed Machine Learning

Seunghak Lee seunghak@cs.cmu.edu
Xun Zheng xunzheng@cs.cmu.edu
Jin Kyu Kim jinkyuk@cs.cmu.edu
Garth Gibson garth@cs.cmu.edu
Eric Xing epxing@cs.cmu.edu
Carnegie Mellon University
Qirong Ho hoqirong@gmail.com
Institute for Infocomm Research, A*STAR

Distributed machine learning has typically been approached from a data parallel perspective, where big data are partitioned to multiple workers and an algorithm is executed concurrently over different data subsets under various synchronization schemes to ensure speed-up and/or correctness. A sibling problem that has received relatively less attention is how to ensure efficient and correct model parallel execution of ML algorithms, where parameters of an ML program are partitioned to different workers and undergone concurrent iterative updates. We argue that model and data parallelisms impose rather different challenges for system design, algorithmic adjustment, and theoretical analysis. In this paper, we develop a system for model-parallelism, STRADS, that provides a programming abstraction

for scheduling parameter updates by discovering and leveraging changing structural properties of ML programs. STRADS enables a flexible tradeoff between scheduling efficiency and fidelity to intrinsic dependencies within the models, and improves memory efficiency of distributed ML. We demonstrate the efficacy of model-parallel algorithms implemented on STRADS versus popular implementations for topic modeling, matrix factorization, and Lasso.

Th70 Communication-Efficient Distributed Dual Coordinate Ascent

Martin Jaggi mjaggi@gmail.com
Thomas Hofmann thomas.hofmann@inf.ethz.ch
ETH Zurich
Ginger Smith vsmith@berkeley.edu
Jonathan Terhorst terhorst@stat.berkeley.edu
Michael Jordan jordan@cs.berkeley.edu
UC Berkeley
Martin Takac martin.taki@gmail.com
Lehigh University
Sanjay Krishnan sanjay@eecs.berkeley.edu
University of California Berkeley

Communication remains the most significant bottleneck in the performance of distributed optimization algorithms for large-scale machine learning. In this paper, we propose a communication-efficient framework, COCOA, that uses local computation in a primal-dual setting to dramatically reduce the amount of necessary communication. We provide a strong convergence rate analysis for this class of algorithms, as well as experiments on real-world distributed datasets with implementations in Spark. In our experiments, we find that as compared to state-of-the-art mini-batch versions of SGD and SDCA algorithms, COCOA converges to the same .001-accurate solution quality on average 25× as quickly.

Th71 Projective dictionary pair learning for pattern classification

Shuhang Gu shuhanggu@gmail.com
Lei Zhang cslzhang@comp.polyu.edu.hk
The Hong Kong Polytechnic Univ
Wangmeng Zuo cswmzuo@gmail.com
Xiangchu Feng xcfeng@mail.xidian.edu.cn
Xidian University

Discriminative dictionary learning (DL) has been widely studied in various pattern classification problems. Most of the existing DL methods aim to learn a synthesis dictionary to represent the input signal while enforcing the representation coefficients and/or representation residual to be discriminative. However, the ℓ_0 or ℓ_1-norm sparsity constraint on the representation coefficients adopted in many DL methods makes the training and testing phases time consuming. We propose a new discriminative DL framework, namely projective dictionary pair learning (DPL), which learns a synthesis dictionary and an analysis dictionary jointly to achieve the goal of signal representation and discrimination. Compared with conventional DL methods, the proposed DPL method can not only greatly reduce the time complexity in the training and testing phases, but also lead to very competitive accuracies in a variety of visual classification tasks.

Th72 Augmentative Message Passing for Traveling Salesman Problem and Graph Partitioning

Siamak Ravanbakhsh mravanba@ualberta.ca
Reihaneh Rabbany rabbanyk@ualberta.ca
Russell Greiner rgreiner@ualberta.ca
University of Alberta

The cutting plane method is an augmentative constrained optimization procedure that is often used with continuous-domain optimization techniques such as linear and convex programs. We investigate the viability of a similar idea within message passing -- for integral solutions -- in the context of two combinatorial problems: 1) For Traveling Salesman Problem (TSP), we propose a factor-graph based on Held-Karp formulation, with an exponential number of constraint factors, each of which has an exponential but sparse tabular form. 2) For graph-partitioning (a.k.a. community mining) using modularity optimization, we introduce a binary variable model with a large number of constraints that enforce formation of cliques. In both cases we are able to derive surprisingly simple message updates that lead to competitive solutions on benchmark instances. In particular for TSP we are able to find near-optimal solutions in the time that empirically grows with N_3, demonstrating that augmentation is practical and efficient.

Th73 Making Pairwise Binary Graphical Models Attractive

Nicholas Ruozzi nr2493@columbia.edu
Tony Jebara jebara@cs.columbia.edu
Columbia University

Computing the partition function (i.e., the normalizing constant) of a given pairwise binary graphical model is NP-hard in general. As a result, the partition function is typically estimated by approximate inference algorithms such as belief propagation (BP) and tree-reweighted belief propagation (TRBP). The former provides reasonable estimates in practice but has convergence issues. The later has better convergence properties but typically provides poorer estimates. In this work, we propose a novel scheme that has better convergence properties than BP and provably provides better partition function estimates in many instances than TRBP. In particular, given an arbitrary pairwise binary graphical model, we construct a specific ``attractive'' 2-cover. We explore the properties of this special cover and show that it can be used to construct an algorithm with the desired properties.

Th74 Transportability from Multiple Environments with Limited Experiments: Completeness Results

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

This paper addresses the problem of mz-transportability, that is, transferring causal knowledge collected in several heterogeneous domains to a target domain in which only passive observations and limited experimental data can be collected. The paper first establishes a necessary and sufficient condition for deciding the feasibility of mz-transportability, i.e., whether causal effects in the target domain are estimable from the information available. It further proves that a previously established algorithm for computing transport formula is in fact complete, that is, failure of the algorithm implies non-existence of a transport formula. Finally, the paper shows that the do-calculus is complete for the mz-transportability class.

Th75 Causal Inference through a Witness Protection Program

Ricardo Silva ricardo@stats.ucl.ac.uk
UCL
Robin Evans evans@stats.ox.ac.uk
University of Oxford

One of the most fundamental problems in causal inference is the estimation of a causal effect when variables are confounded. This is difficult in an observational study because one has no direct evidence that all confounders have been adjusted for. We introduce a novel approach for estimating causal effects that exploits observational conditional independencies to suggest ``weak'' paths in a unknown causal graph. The widely used faithfulness condition of Spirtes et al. is relaxed to allow for varying degrees of ``path cancellations'' that will imply conditional independencies but do not rule out the existence of confounding causal paths. The outcome is a posterior distribution over bounds on the average causal effect via a linear programming approach and Bayesian inference. We claim this approach should be used in regular practice to complement other default tools in observational studies.

Th76 Biclustering Usinig Message Passing

Luke O'Connor loconnor@college.harvard.edu
Harvard University
Soheil Feizi sfeizi@mit.edu
Massachusetts Institute of Technology

Biclustering is the analog of clustering on a bipartite graph. Existent methods infer biclusters through local search strategies that find one cluster at a time; a common technique is to update the row memberships based on the current column memberships, and vice versa. We propose a biclustering algorithm that maximizes a global objective function using message passing. Our objective function closely approximates a general likelihood function, separating a cluster size penalty term into row- and column-count penalties. Because we use a global optimization framework, our approach excels at resolving the overlaps between biclusters, which are important features of biclusters in practice. Moreover, Expectation-Maximization can be used to learn the model parameters if they are unknown. In simulations, we find that our method outperforms two of the best existing biclustering algorithms, ISA and LAS, when the planted clusters overlap. Applied to three gene expression datasets, our method finds coregulated gene clusters that have high quality in terms of cluster size and density.

Th77 PAC-Bayesian AUC classification and scoring

James Ridgway james.ridgway@ensae.fr
Crest-Ensae and Dauphine
Pierre Alquier alquier@ensae.fr
ENSAE
Nicolas Chopin nicolas.chopin@ensae.fr
CREST
Feng Liang liangf@illinois.edu
Univ. of Illinois Urbana-Champaign Statistics

We develop a scoring and classification procedure based on the PAC-Bayesian approach and the AUC (Area Under Curve) criterion. We focus initially on the class of linear score functions. We derive PAC-Bayesian non-asymptotic bounds for two types of prior for the score parameters: a Gaussian prior, and a spike-and-slab prior; the latter makes it possible to perform feature selection. One important advantage of our approach is that it is amenable to powerful Bayesian computational tools. We derive in particular a Sequential Monte Carlo algorithm, as an efficient method which may be used as a gold standard, and an Expectation-Propagation algorithm, as a much faster but approximate method. We also extend our method to a class of non-linear score functions, essentially leading to a nonparametric procedure, by considering a Gaussian process prior.

Th78 Partition-wise Linear Models

Hidekazu Oiwa oiwa@r.dl.itc.u-tokyo.ac.jp
The University of Tokyo
Ryohei Fujimaki rfujimaki@nec-labs.com
NEC Laboratories America

Region-specific linear models are widely used in practical applications because of their non-linear but highly interpretable model representations. One of the key challenges in their use is non-convexity in simultaneous optimization of regions and region-specific models. This paper proposes novel convex region-specific linear models, which we refer to as partition-wise linear models. Our key ideas are 1) assigning linear models not to regions but to partitions (region-specifiers) and representing region-specific linear models by linear combinations of partition-specific models, and 2) optimizing regions via partition selection from a large number of given partition candidates by means of convex structured regularizations. In addition to providing initialization-free globally-optimal solutions, our convex formulation makes it possible to derive a generalization bound and to use such advanced optimization techniques as proximal methods and decomposition of the proximal maps for sparsity-inducing regularizations. Experimental results demonstrate that our partition-wise linear models perform better than or are at least competitive with state-of-the-art region-specific or locally linear models.

Th79 Near-optimal sample compression for nearest neighbors

This poster has been moved to Wednesday, W105

Th80 Learning Shuffle Ideals Under Restricted Distributions

Dongqu Chen dongqu.chen@yale.edu
Yale University

The class of shuffle ideals is a fundamental sub-family of regular languages. The shuffle ideal generated by a string set U is the collection of all strings containing some string $u \in U$ as a (not necessarily contiguous) subsequence. In spite of its apparent simplicity, the problem of learning a shuffle ideal from given data is known to be computationally intractable. In this paper, we study the PAC learnability of shuffle ideals and present positive results on this learning problem under element-wise independent and identical distributions and Markovian distributions in the statistical query model. A constrained generalization to learning shuffle ideals under product distributions is also provided. In the empirical direction, we propose a heuristic algorithm for learning shuffle ideals from given labeled strings under general unrestricted distributions. Experiments demonstrate the advantage for both efficiency and accuracy of our algorithm.

Th81 Extracting Certainty from Uncertainty: Transductive Pairwise Classification from Pairwise Similarities

Tianbao Yang yangtia1@gmail.com
University of Iowa
Rong Jin rong+@cs.cmu.edu
Michigan State University (MSU)

In this work, we study the problem of transductive pairwise classification from pairwise similarities~\footnote{The pairwise similarities are usually derived from some side information instead of the underlying class labels.}. The goal of transductive pairwise classification from pairwise similarities is to infer the pairwise class relationships, to which we refer as pairwise labels, between all examples given a subset of class relationships for a small set of examples, to which we refer as labeled examples. We propose a very simple yet effective algorithm that consists of two simple steps: the first step is to complete the sub-matrix corresponding to the labeled examples and the second step is to reconstruct the label matrix from the completed sub-matrix and the provided similarity matrix. Our analysis exhibits that under several mild preconditions we can recover the label matrix with a small error, if the top eigen-space that corresponds to the largest eigenvalues of the similarity matrix covers well the column space of label matrix and is subject to a low coherence, and the number of observed pairwise labels is sufficiently enough. We demonstrate the effectiveness of the proposed algorithm by several experiments.

Th82 Incremental Clustering: The Case for Extra Clusters

Margareta Ackerman mackerman@fsu.edu
Florida State University
Sanjoy Dasgupta dasgupta@eng.ucsd.edu
UC San Diego

The explosion in the amount of data available for analysis often necessitates a transition from batch to incremental clustering methods, which process one element at a time and typically store only a small subset of the data. In this paper, we initiate the formal analysis of incremental clustering methods focusing on the types of cluster structure that they are able to detect. We find that the incremental setting is strictly weaker than the batch model, proving that a fundamental class of cluster structures that can readily be detected in the batch setting is impossible to identify using any incremental method. Furthermore, we show how the limitations of incremental clustering can be overcome by allowing additional clusters.

Th83 Parallel Successive Convex Approximation for Nonsmooth Nonconvex Optimization

Meisam Razaviyayn razav002@umn.edu
University of Minnesota
Mingyi Hong mingyi@iastate.edu
Iowa State University
Zhi-Quan Luo luozq@ece.umn.edu
University of Minnesota, Twin Cites
Jong-Shi Pang jongship@usc.edu

Consider the problem of minimizing the sum of a smooth (possibly non-convex) and a convex (possibly nonsmooth) function involving a large number of variables. A popular approach to solve this problem is the block coordinate descent (BCD) method whereby at each iteration only one variable block is updated while the remaining variables are held fixed. With the recent advances in the developments of the multi-core parallel processing technology, it is desirable to parallelize the BCD method by allowing multiple blocks to be updated simultaneously at each iteration of the algorithm. In this work, we propose an inexact parallel BCD approach where at each iteration, a subset of the variables is updated in parallel by minimizing convex approximations of the original objective function. We investigate the convergence of this parallel BCD method for both randomized and cyclic variable selection rules. We analyze the asymptotic and non-asymptotic convergence behavior of the algorithm for both convex and non-convex objective functions. The numerical experiments suggest that for a special case of Lasso minimization problem, the cyclic block selection rule can outperform the randomized rule.

Th84 Stochastic Proximal Gradient Descent with Acceleration Techniques

Atsushi Nitanda nitanda@msi.co.jp
NTT DATA Mathematical Systems Inc.

Proximal gradient descent (PGD) and stochastic proximal gradient descent (SPGD) are popular methods for solving regularized risk minimization problems in machine learning and statistics. In this paper, we propose and analyze an accelerated variant of these methods in the mini-batch setting. This method incorporates two acceleration techniques: one is Nesterov's acceleration method, and the other is a variance reduction for the stochastic gradient. Accelerated proximal gradient descent (APG) and proximal stochastic variance reduction gradient (Prox-SVRG) are in a trade-off relationship. We show that our method, with the appropriate mini-batch size, achieves lower overall complexity than both APG and Prox-SVRG.

Th85 Beyond the Birkhoff Polytope: Convex Relaxations for Vector Permutation Problems

Cong Han Lim conghan@cs.wisc.edu
Stephen Wright swright@cs.wisc.edu
UW-Madison

The Birkhoff polytope (the convex hull of the set of permutation matrices), which is represented using $\Theta(n_2)$ variables and constraints, is frequently invoked in formulating relaxations of optimization problems over permutations. Using a recent construction of Goemans (2010), we show that when optimizing over the convex hull of the permutation vectors (the permutahedron), we can reduce the number of variables and constraints to $\Theta(n\log n)$ in theory and $\Theta(n\log_2 n)$ in practice. We modify the recent convex formulation of the 2-SUM problem introduced by Fogel et al. (2013) to use this polytope, and demonstrate how we can attain results of similar quality in significantly less computational time for large n. To our knowledge, this is the first usage of Goemans' compact formulation of the permutahedron in a convex optimization problem. We also introduce a simpler regularization scheme for this convex formulation of the 2-SUM problem that yields good empirical results.

Th86 Bregman Alternating Direction Method of Multipliers

Huahua Wang huwang@cs.umn.edu
Arindam Banerjee banerjee@cs.umn.edu
University of Minnesota, Twin Cites

The mirror descent algorithm (MDA) generalizes gradient descent by using a Bregman divergence to replace squared Euclidean distance. In this paper, we similarly generalize the alternating direction method of multipliers (ADMM) to Bregman ADMM (BADMM), which allows the choice of different Bregman divergences to exploit the structure of problems. BADMM provides a unified framework for ADMM and its variants, including generalized ADMM, inexact ADMM and Bethe ADMM. We establish the global convergence and the $O(1/T)$ iteration complexity for BADMM. In some cases, BADMM can be faster than ADMM by a factor of $O(n/\ln n)$ where n is the dimensionality. In solving the linear program of mass transportation problem, BADMM leads to massive parallelism and can easily run on GPU. BADMM is several times faster than highly optimized commercial software Gurobi.

Th87 Multi-Step Stochastic ADMM in High Dimensions: Applications to Sparse Optimization and Matrix Decomposition

Hanie Sedghi hsedghi@usc.edu
University of Southern California
Anima Anandkumar a.anandkumar@uci.edu
UC Irvine
Edmond Jonckheere jonckhee@usc.edu

In this paper, we consider a multi-step version of the stochastic ADMM method with efficient guarantees for high-dimensional problems. We first analyze the simple setting, where the optimization problem consists of a loss function and a single regularizer (e.g. sparse optimization), and then extend to the multi-block setting with multiple regularizers and multiple variables (e.g. matrix decomposition into sparse and low rank components). For the sparse optimization problem, our method achieves the minimax rate of $O(s\log d/T)$ for s-sparse problems in d dimensions in T steps, and is thus, unimprovable by any method up to constant factors. For the matrix decomposition problem with a general loss function, we analyze the multi-step ADMM with multiple blocks. We establish $O(1/T)$ rate and efficient scaling as the size of matrix grows. For natural noise models (e.g. independent noise), our convergence rate is minimax-optimal. Thus, we establish tight convergence guarantees for multi-block ADMM in high dimensions. Experiments show that for both sparse optimization and matrix decomposition problems, our algorithm outperforms the state-of-the-art methods.

Th88 Accelerated Mini-batch Randomized Block Coordinate Descent Method

Tuo Zhao tzhao5@jhu.edu
Princeton University and Johns Hopkins University
Mo Yu gflfof@gmail.com
Yiming Wang freewym@gmail.com
Raman Arora arora@cs.jhu.edu
Johns Hopkins University
Han Liu hanliu@princeton.edu
Princeton University

We consider regularized empirical risk minimization problems. In particular, we minimize the sum of a smooth empirical risk function and a nonsmooth regularization function. When the regularization function is block separable, we can solve the minimization problems in a randomized block coordinate descent (RBCD) manner. Existing RBCD methods usually decrease the objective value by exploiting the partial gradient of a randomly selected block of coordinates in each iteration. Thus they need all data to be accessible so that the partial gradient of the block gradient can be exactly obtained. However, such a ``batch" setting may be computationally expensive in practice. In this paper, we propose a mini-batch randomized block coordinate descent (MRBCD) method, which estimates the partial gradient of the selected block

based on a mini-batch of randomly sampled data in each iteration. We further accelerate the MRBCD method by exploiting the semi-stochastic optimization scheme, which effectively reduces the variance of the partial gradient estimators. Theoretically, we show that for strongly convex functions, the MRBCD method attains lower overall iteration complexity than existing RBCD methods. As an application, we further trim the MRBCD method to solve the regularized sparse learning problems. Our numerical experiments shows that the MRBCD method naturally exploits the sparsity structure and achieves better computational performance than existing methods.

Th89 Sparse PCA with Oracle Property

Quanquan Gu qgu@princeton.edu
Zhaoran Wang zhaoran@princeton.edu
Han Liu hanliu@princeton.edu
Princeton University

In this paper, we study the estimation of the k-dimensional sparse principal subspace of covariance matrix Σ in the high-dimensional setting. We aim to recover the oracle principal subspace solution, i.e., the principal subspace estimator obtained assuming the true support is known a priori. To this end, we propose a family of estimators based on the semidefinite relaxation of sparse PCA with novel regularizations. In particular, under a weak assumption on the magnitude of the population projection matrix, one estimator within this family exactly recovers the true support with high probability, has exact rank-k, and attains a $s/n---\sqrt{}$ statistical rate of convergence with s being the subspace sparsity level and n the sample size. Compared to existing support recovery results for sparse PCA, our approach does not hinge on the spiked covariance model or the limited correlation condition. As a complement to the first estimator that enjoys the oracle property, we prove that, another estimator within the family achieves a sharper statistical rate of convergence than the standard semidefinite relaxation of sparse PCA, even when the previous assumption on the magnitude of the projection matrix is violated. We validate the theoretical results by numerical experiments on synthetic datasets.

Th90 A statistical model for tensor PCA

Emile Richard r.emile.richard@gmail.com
Andrea Montanari montanari@stanford.edu
Stanford University

We consider the Principal Component Analysis problem for large tensors of arbitrary order k under a single-spike (or rank-one plus noise) model. On the one hand, we use information theory, and recent results in probability theory to establish necessary and sufficient conditions under which the principal component can be estimated using unbounded computational resources. It turns out that this is possible as soon as the signal-to-noise ratio beta becomes larger than C\sqrt{k log k} (and in particular beta can remain bounded has the problem dimensions increase). On the other hand, we analyze several polynomial-time estimation algorithms, based on tensor unfolding, power iteration and message passing ideas from graphical models. We show that, unless the signal-to-noise ratio diverges in the system dimensions, none of these approaches succeeds. This is possibly related to a fundamental limitation of computationally tractable estimators for this problem. For moderate dimensions, we propose an hybrid approach that uses unfolding together with power iteration, and show that it outperforms significantly baseline methods. Finally, we consider the case in which additional side information is available about the unknown signal. We characterize the amount of side information that allow the iterative algorithms to converge to a good estimate.

Th91 large scale canonical correlation analysis with iterative least squares

Yichao Lu yichaolu@wharton.upenn.edu
Dean Foster foster@wharton.upenn.edu
University of Pennsylvania

Canonical Correlation Analysis (CCA) is a widely used statistical tool with both well established theory and favorable performance for a wide range of machine learning problems. However, computing CCA for huge datasets can be very slow since it involves implementing QR decomposition or singular value decomposition of huge matrices. In this paper we introduce L-CCA, an iterative algorithm which can compute CCA fast on huge sparse datasets. Theory on both the asymptotic convergence and finite time accuracy of L-CCA are established. The experiments also show that L-CCA outperform other fast CCA approximation schemes on two real datasets.

Th92 Cone-Constrained Principal Component Analysis

Yash Deshpande yashd@stanford.edu
Andrea Montanari montanari@stanford.edu
Emile Richard r.emile.richard@gmail.com
Stanford University

Estimating a vector from noisy quadratic observations is a task that arises naturally in many contexts, from dimensionality reduction, to synchronization and phase retrieval problems. It is often the case that additional information is available about the unknown vector (for instance, sparsity, sign or magnitude of its entries). Many authors propose non-convex quadratic optimization problems that aim at exploiting optimally this information. However, solving these problems is typically NP-hard. We consider a simple model for noisy quadratic observation of an unknown vector $\backslash\mathrm{bvz}$. The unknown vector is constrained to belong to a cone $\backslash\mathrm{Cone}\ni\backslash\mathrm{bvz}$. While optimal estimation appears to be intractable for the general problems in this class, we provide evidence that it is tractable when $\backslash\mathrm{Cone}$is a convex cone with an efficient projection. This is surprising, since the corresponding optimization problem is non-convex and --from a worst case perspective-- often NP hard. We characterize the resulting minimax risk in terms of the statistical dimension of the cone $\delta(\backslash\mathrm{Cone})$. This quantity is already known to control the risk of estimation from gaussian observations and random linear measurements. It is rather surprising that the same quantity plays a role in the estimation risk from quadratic measurements.

Th93 Improved Distributed Principal Component Analysis

Yingyu Liang yingyul@cs.princeton.edu
Princeton University
Maria-Florina Balcan ninamf@cc.gatech.edu
Georgia Tech
Vandana Kanchanapally vvandana@gatech.edu
David Woodruff dpwoodru@us.ibm.com
IBM Research

We study the distributed computing setting in which there are multiple servers, each holding a set of points, who wish to compute functions on the union of their point sets. A key task in this setting is Principal Component Analysis (PCA), in which the servers would like to compute a low dimensional subspace capturing as much of the variance of the union of their point sets as possible. Given a procedure for approximate PCA, one can use it to approximately solve problems such as k-means clustering and low rank approximation. The essential properties of an approximate distributed PCA algorithm are its communication cost and computational efficiency for a given desired accuracy in downstream applications. We give new algorithms and analyses for distributed PCA which lead to improved communication and computational costs for k-means clustering and related problems. Our empirical study on real world data shows a speedup of orders of magnitude, preserving communication with only a negligible degradation in solution quality. Some of these techniques we develop, such as input-sparsity subspace embeddings with high correctness probability with a dimension and sparsity independent of the error probability, may be of independent interest.

Th94 Generalized Unsupervised Manifold Alignment

Zhen Cui zhen.cui@vipl.ict.ac.cn
Vision and Machine Learning Lab
Hong Chang hong.chang@vipl.ict.ac.cn
Chinese Academy of Sciences
Shiguang Shan sgshan@ict.ac.cn
Xilin Chen xlchen@ict.ac.cn

In this paper, we propose a generalized Unsupervised Manifold Alignment (GUMA) method to build the connections between different but correlated datasets without any known correspondences. Based on the assumption that datasets of the same theme usually have similar manifold structures, GUMA is formulated into an explicit integer optimization problem considering the structure matching and preserving criteria, as well as the feature comparability of the corresponding points in the mutual embedding space. The main benefits of this model include: (1) simultaneous discovery and alignment of manifold structures; (2) fully unsupervised matching without any pre-specified correspondences; (3) efficient iterative alignment without computations in all permutation cases. Experimental results on dataset matching and real-world applications demonstrate the effectiveness and the practicability of our manifold alignment method.

Th95 On Prior Distributions and Approximate Inference for Structured Variables

Sanmi Koyejo sanmi.k@gmail.com
Stanford University
Rajiv Khanna rajivak@utexas.edu
Joydeep Ghosh ghosh@ece.utexas.edu
Russell Poldrack poldrack@utexas.edu
University of Texas

We present a general framework for constructing prior distributions with structured variables. The prior is defined as the information projection of a base distribution onto distributions supported on the constraint set of interest. In cases where this projection is intractable, we propose a family of parameterized approximations indexed by subsets of the domain. We further analyze the special case of sparse structure. While the optimal prior is intractable in general, we show that approximate inference using convex subsets is tractable, and is equivalent to maximizing a submodular function subject to cardinality constraints. As a result, inference using greedy forward selection provably achieves within a factor of (1-1/e) of the optimal objective value. Our work is motivated by the predictive modeling of high-dimensional functional neuroimaging data. For this task, we employ the Gaussian base distribution induced by local partial correlations and consider the design of priors to capture the domain knowledge of sparse support. Experimental results on simulated data and high dimensional neuroimaging data show the effectiveness of our approach in terms of support recovery and predictive accuracy.

Th96 Real-Time Decoding of an Integrate and Fire Encoder

Shreya Saxena ssaxena@mit.edu
Munther Dahleh dahleh@mit.edu
Massachusetts Institute of Tehcnology

Neuronal encoding models range from the detailed biophysically-based Hodgkin Huxley model, to the statistical linear time invariant model specifying firing rates in terms of the extrinsic signal. Decoding the former becomes intractable, while the latter does not adequately capture the nonlinearities present in the neuronal encoding system. For use in practical applications, we wish to record the output of neurons, namely spikes, and decode this signal fast in order to drive a machine, for example a prosthetic device. Here, we introduce a causal, real-time decoder of the biophysically-based Integrate and Fire encoding neuron model. We show that the upper bound of the real-time reconstruction error decreases polynomially in time, and that the L2 norm of the error is bounded by a constant that depends on the density of the spikes, as well as the bandwidth and the decay of the input signal. We numerically validate the effect of these parameters on the reconstruction error.

Th97 Spike Frequency Adaptation Implements Anticipative Tracking in Continuous Attractor Neural Networks

Yuanyuan Mi miyuanyuan0102@163.com
Weizmann Institute of Science
C. C. Alan Fung phccfung@ust.hk
K. Y. Michael Wong phkywong@ust.hk
The Hong Kong University of Science and Technology
Si Wu wusi@bnu.edu.cn
Beijing Normal University

To extract motion information, the brain needs to compensate for time delays that are ubiquitous in neural signal transmission and processing. Here we propose a simple yet effective mechanism to implement anticipative tracking in neural systems. The proposed mechanism utilizes the property of spike-frequency adaptation (SFA), a feature widely observed in neuronal responses. We employ continuous attractor neural networks (CANNs) as the model to describe the tracking behaviors in neural systems. Incorporating SFA, a CANN exhibits intrinsic mobility, manifested by the ability of the CANN to hold self-sustained travelling waves. In tracking a moving stimulus, the interplay between the external drive and the intrinsic mobility of the network determines the tracking performance. Interestingly, we find that the regime of anticipation effectively coincides with the regime where the intrinsic speed of the travelling wave exceeds that of the external drive. Depending on the SFA amplitudes, the network can achieve either perfect tracking, with zero-lag to the input, or perfect anticipative tracking, with a constant leading time to the input. Our model successfully reproduces experimentally observed anticipative tracking behaviors, and sheds light on our understanding of how the brain processes motion information in a timely manner.

Th98 Analysis of Brain States from Multi-Region LFP Time-Series

Kyle Ulrich kyle.ulrich@duke.edu
David Carlson david.carlson@duke.edu
Wenzhao Lian wl89@duke.edu
Jana Borg borg@neuro.duke.edu
Kafui Dzirasa kafui.dzirasa@duke.edu
Lawrence Carin lcarin@duke.edu
Duke University

The local field potential (LFP) is a source of information about the broad patterns of brain activity, and the frequencies present in these time-series measurements are often highly correlated between regions. It is believed that these regions may jointly constitute a ``brain state,'' relating to cognition and behavior. An infinite hidden Markov model (iHMM) is proposed to model the evolution of brain states, based on electrophysiological LFP data measured at multiple brain regions. A brain state influences the spectral content of each region in the measured LFP. A new state-dependent tensor factorization is employed across brain regions, and the spectral properties of the LFPs are characterized in terms of Gaussian processes (GPs). The LFPs are modeled as a mixture of GPs, with state- and region-dependent mixture weights, and with the spectral content of the data encoded in GP spectral mixture covariance kernels. The model is able to estimate the number of brain states and the number of mixture components in the mixture of GPs. A new variational Bayesian split-merge algorithm is employed for inference. The model infers state changes as a function of external covariates in two novel electrophysiological datasets, using LFP data recorded simultaneously from multiple brain regions in mice; the results are validated and interpreted by subject-matter experts.

Th99 Extracting Latent Structure From Multiple Interacting Neural Populations

Joao Semedo jsemedo@andrew.cmu.edu
Byron Yu byronyu@cmu.edu
Carnegie Mellon University
Amin Zandvakili zandvakili@gmail.com
Adam Kohn akohn@aecom.yu.edu
Albert Einstein College of Medicine
Christian Machens
christian.machens@neuro.fchampalimaud.org
Champalimaud Centre for the Unknown

Developments in neural recording technology are rapidly enabling the recording of populations of neurons in multiple brain areas simultaneously, as well as the identification of the types of neurons being recorded (e.g., excitatory vs. inhibitory). There is a growing need for statistical methods to study the interaction among multiple, labeled populations of neurons. Rather than attempting to identify direct interactions between neurons (where the number of interactions grows with the number of neurons squared), we propose to extract a smaller number of latent variables from each population and study how the latent variables interact. Specifically, we propose extensions to probabilistic canonical correlation analysis (pCCA) to capture the temporal structure of the latent variables, as well as to distinguish within-population dynamics from across-population interactions (termed Group Latent Auto-Regressive Analysis, gLARA). We then applied these methods to populations of neurons recorded simultaneously in visual areas V1 and V2, and found that gLARA provides a better description of the recordings than pCCA. This work provides a foundation for studying how multiple populations of neurons interact and how this interaction supports brain function.

Th100 Learning convolution filters for inverse covariance estimation of neural network connectivity

George Mohler gmohler@scu.edu
Santa Clara University

We consider the problem of inferring direct neural network connections from Calcium imaging time series. Inverse covariance estimation has proven to be a fast and accurate method for learning macro- and micro-scale network connectivity in the brain and in a recent Kaggle Connectomics competition inverse covariance was the main component of several top ten solutions, including our own and the winning team's algorithm. However, the accuracy of inverse covariance estimation is highly sensitive to signal preprocessing of the Calcium fluorescence time series. Furthermore, brute force optimization methods such as grid search and coordinate ascent over signal processing parameters is a time intensive process, where learning may take several days and parameters that optimize one network may not generalize to networks with different size and parameters. In this paper we show how inverse covariance estimation can be dramatically improved using a simple convolution filter prior to applying sample covariance. Furthermore, these signal processing parameters can be learned quickly using a supervised optimization algorithm. In particular, we maximize a binomial log-likelihood loss function with respect to a convolution filter of the time series and the inverse covariance regularization parameter. Our proposed algorithm is relatively fast on networks the size of those in the competition (1000 neurons), producing AUC scores with similar accuracy to the winning solution in training time under 2 hours on a cpu. Prediction on new networks of the same size is carried out in less than 15 minutes, the time it takes to read in the data and write out the solution.

Th101 On Multiplicative Multitask Feature Learning

Xin Wang wangxin@engr.uconn.edu
Jinbo Bi jinbo@engr.uconn.edu
Jiangwen Sun javon@engr.uconn.edu
University of Connecticut
Shipeng Yu shipeng.yu@siemens.com
Siemens

We investigate a general framework of multiplicative multitask feature learning which decomposes each task's model parameters into a multiplication of two components. One of the components is used across all tasks and the other component is task-specific. Several previous methods have been proposed as special cases of our framework. We study the theoretical properties of this framework when different regularization conditions are applied to the two decomposed components. We prove that this framework is mathematically equivalent to the widely used multitask feature learning methods that are based on a joint regularization of all model parameters, but with a more general form of regularizers. Further, an analytical formula is derived for the across-task component as related to the task-specific component for all these regularizers, leading to a better understanding of the shrinkage effect. Study of this framework motivates new multitask learning algorithms. We propose two new learning formulations by varying the parameters in the proposed framework. Empirical studies have revealed the relative advantages of the two new formulations by comparing with the state of the art, which provides instructive insights into the feature learning problem with multiple tasks.

Th102 Multitask learning meets tensor factorization: task imputation via convex optimization

Kishan Wimalawarne kishan@sg.cs.titech.ac.jp
Tokyo Institute of Technology
Masashi Sugiyama sugi@k.u-tokyo.ac.jp
The University of Tokyo
Ryota Tomioka tomioka@ttic.edu
TTI Chicago

We study a multitask learning problem in which each task is parametrized by a weight vector and indexed by a pair of indices, which can be e.g, (consumer, time). The weight vectors can be collected into a tensor and the (multilinear-)rank of the tensor controls the amount of sharing of information among tasks. Two types of convex relaxations have recently been proposed for the tensor multilinear rank. However, we argue that both of them are not optimal in the context of multitask learning in which the dimensions or multilinear rank are typically heterogeneous. We propose a new norm, which we call the scaled latent trace norm and analyze the excess risk of all the three norms. The results apply to various settings including matrix and tensor completion, multitask learning, and multilinear multitask learning. Both the theory and experiments support the advantage of the new norm when the tensor is not equal-sized and we do not a priori know which mode is low rank.

Th103 Learning Multiple Tasks in Parallel with a Shared Annotator

Haim Cohen haimco10@gmail.com
Koby Crammer koby@ee.technion.ac.il
The Technion

We introduce a new multi-task framework, in which K online learners are sharing a single annotator with limited bandwidth. On each round, each of the K learners receives an input, and makes a prediction about the label of that input. Then, a shared (stochastic) mechanism decides which of the K inputs will be annotated. The learner that receives the feedback (label) may update its prediction rule, and we proceed to the next round. We develop an online algorithm for multi-task binary classification that learns in this setting, and bound its performance in the worst-case setting. Additionally, we show that our algorithm can be used to solve two bandits problems: contextual bandits, and dueling bandits with context, both allowed to decouple exploration and exploitation. Empirical study with OCR data, vowel prediction (VJ project) and document classification, shows that our algorithm outperforms other algorithms, one of which uses uniform allocation, and essentially makes more (accuracy) for the same labour of the annotator.

Th104 Multivariate Regression with Calibration

Han Liu hanliu@princeton.edu
Princeton University
Lie Wang liewang@math.mit.edu
Massachusetts Institute of Technology
Tuo Zhao tzhao5@jhu.edu
Princeton University and Johns Hopkins University

We propose a new method named calibrated multivariate regression (CMR) for fitting high dimensional multivariate regression models. Compared to existing methods, CMR calibrates the regularization for each regression task with respect to its noise level so that it is simultaneously tuning insensitive and achieves an improved finite-sample performance. Computationally, we develop an efficient smoothed proximal gradient algorithm which has a worst-case iteration complexity $O(1/\epsilon)$, where ϵ is a pre-specified numerical accuracy. Theoretically, we prove that CMR achieves the optimal rate of convergence in parameter estimation. We illustrate the usefulness of CMR by thorough numerical simulations and show that CMR consistently outperforms other high dimensional multivariate regression methods. We also apply CMR on a brain activity prediction problem and find that CMR is as competitive as the handcrafted model created by human experts.

Th105 Exclusive Feature Learning on Arbitrary Structures via $\ell_{1,2}$-norm

Deguang Kong doogkong@gmail.com
UT Arlington
Ryohei Fujimaki rfujimaki@nec-labs.com
NEC Laboratories America
Ji Liu ji.liu.uwisc@gmail.com
UW-Madison
Feiping Nie feipingnie@gmail.com
Chris Ding chqding@uta.edu
University of Texas at Arlington

Group lasso is widely used to enforce the structural sparsity, which achieves the sparsity at inter-group level. In this paper, we propose a new formulation called ``exclusive group lasso'', which brings out sparsity at intra-group level in the context of feature selection. The proposed exclusive group lasso is applicable on any feature structures, regardless of their overlapping or non-overlapping structures. We give analysis on the properties of exclusive group lasso, and propose an effective iteratively re-weighted algorithm to solve the corresponding optimization problem with rigorous convergence analysis. We show applications of exclusive group lasso for uncorrelated feature selection. Extensive experiments on both synthetic and real-world datasets indicate the good performance of proposed methods.

Th106 Flexible Transfer Learning under Support and Model Shift

Xuezhi Wang xzwang07@gmail.com
Jeff Schneider schneide@cs.cmu.edu
Carnegie Mellon University

Transfer learning algorithms are used when one has sufficient training data for one supervised learning task (the source/training domain) but only very limited training data for a second task (the target/test domain) that is similar but not identical to the first. Previous work on transfer learning has focused on relatively restricted settings, where specific parts of the model are considered to be carried over between tasks. Recent work on covariate shift focuses on matching the marginal distributions on observations X across domains. Similarly, work on target/conditional shift focuses on matching marginal distributions on labels Y and adjusting conditional distributions $P(X|Y)$, such that $P(X)$ can be matched across domains. However, covariate shift assumes that the support of test $P(X)$ is contained in the support of training $P(X)$, i.e., the training set is richer than the test set. Target/conditional shift makes a similar assumption for $P(Y)$. Moreover, not much work on transfer learning has considered the case when a few labels in the test domain are available. Also little work has been done when all marginal and conditional distributions are allowed to change while the changes are smooth. In this paper, we consider a general case where both the support and the model change across domains. We transform both X and Y by a location-scale shift to achieve transfer between tasks. Since we allow more flexible transformations, the proposed method yields better results on both synthetic data and real-world data.

REVIEWERS

Yasin Abbasi
Ahmad Abdulkader
Jayadev Acharya
Margareta Ackerman
Alekh Agarwal
Deepak Agarwal
Shivani Agarwal
Shipra Agrawal
Amr Ahmed
Karteek Alahari
Morteza Alamgir
Qi Alan
Christian Albers
Daniel Alexander
Karim Ali
Alexandre Allauzen
Genevera Allen
Yasemin Altun
Marco Alvarez
Mauricio Alvarez
Carlos Alzate
Kareem Amin
Arash Amini
Massih Amini
Eyal Amir
Martin Andersen
Charles Anderson
Bjoern Andres
Christophe Andrieu
Elaine Angelino
Dragomir Anguelov
Ognjen Arandjelovic
Cedric Archambeau
Evan Archer
Raman Arora
Thierry Artières
Özlem Aslan
Arthur Asuncion
Hagai Attias
Erik Aurell
Joseph Austerweil
Haim Avron
Hossein Azari Soufiani
Chloe-Agathe Azencott
Javad Azimi
Martin Azizyan
S. Derin Babacan
Francis Bach
Stephen Bach
Ashwinkumar Badanidiyuru
Bing Bai
Xiang Bai
Raphael Bailly
Krishnakumar Balasubramanian
Christopher Baldassano
Pierre Baldi
David Balduzzi
Borja Balle
Arindam Banerjee
Ying-Ze Bao
Yoseph Barash
Remi Bardenet
Andre Barreto
Jon Barron
Simon Barthelmé
Gabor Bartok
Sumit Basu
Dhruv Batra
Francoise Beaufays
Stephen Becker
Behrouz Behmardi
Oscar Beijbom
Peter Bell
Marc Bellemare
Hani Ben-Amor
Shai Ben-David
Yoshua Bengio
Jos_ Bento
Philipp Berens
Wicher Bergsma
James Bergstra
Michel Besserve
Chiranjib Bhattacharyya
Jinbo Bi
Manuele Bicego
Felix Biessmann
Misha Bilenko
Johannes Bill
Mathieu Blondel
Avrim Blum
Charles Blundell
Phil Blunsom
Liefeng Bo
Natalia Bochkina
Botond Bocsi
Sander Bohte
Danushka Bollegala
Edwin Bonilla
Byron Boots
Antoine Bordes
Ali Borji
Luke Bornn
Jorg Bornschein

Reza Bosagh Zadeh
Leon Bottou
Alexandre Bouchard-Cote
Abdeslam Boularias
YLan Boureau
Herve Bourlard
Christos Boutsidis
Jake Bouvrie
Jordan Boyd-Graber
Levi Boyles
Steve Branson
Daniel Braun
Ulf Brefeld
Xavier Bresson
John Bridle
Mark Briers
Tamara Broderick
Marcus Brubaker
Joan Bruna
Emma Brunskill
Lars Buesing
Joachim Buhmann
Yaroslav Bulatov
Wray Buntine
Wolfram Burgard
R?bert Busa-Fekete
Florian Büttner
Laiwan CHAN
Ricardo Cabral
Tiberio Caetano
Deng Cai
Roberto Calandra
Clément Calauzènes
Ben Calderhead
Colin Campbell
William Campbell
Guillermo Canas
Stephane Canu
Liangliang Cao
Olivier Cappe
Constantine Caramanis
Jaime Carbonell
David Carlson
Francois Caron
Alexandra Carpentier
Miguel Carreira-Perpinan
Roberto Casarin
Robert Castelo
Gert Cauwenberghs
Gavin Cawley
Asli Celikyilmaz
Taylan Cemgil
Nicole Cesa-Bianchi
Volkan Cevher
Hakan Cevikalp
Arun Chaganty
Brahim Chaib-draa
Edward Challis
Jason Chang
Jonathan Chang
Kai-Wei Chang
Denis Charles
Laurent Charlin
Kamalika Chaudhuri
Sougata Chaudhuri
Sourish Chaudhuri
Xuan Che
Gal Chechik
Rama Chellappa
Bee-Chung Chen
Bo Chen
Jianhui Chen
Kewei Chen
Minmin Chen
Ning Chen
Tsuhan Chen
Wenlin Chen
Xi Chen
Yudong Chen
Yutian Chen
Yuxin Chen
Zhitang Chen
Weiwei Cheng
Michael Chertkov
Max Chickering
Hai Leong Chieu
Gillian Chin
Julien Chiquet
Kyunghyun Cho
Arthur Choi
Seungjin Choi
Nicolas Chopin
Krzysztof Choromanski
Christos Christoudias
Colin Cilfford
Tom Claassen
Stephan Clemencon
Mark Coates
Ruben Coen-cagli
Jonathan Cohen
Olivier Collier
Ronan Collobert
Greg Corrado

Andrew Cotter
Garrison Cottrell
Aaron Courville
Emanuele Coviello
Gabor Csardi
Aron Culotta
James Cussens
Marco Cuturi
William Dabney
George Dahl
Bo Dai
Zhenwen Dai
Andreas Damianou
Christian Daniel
Christoph Dann
Trevor Darrell
Sanjoy Dasgupta
Emmanuel Dauce
Hal Daume III
Yann Dauphin
Ian Davidson
Jesse Davis
Peter Dayan
Tijl De Bie
Giulia De Salvo
Nema Dean
Eyal Dechter
Dennis Decoste
Marc Deisenroth
Alessio Del Bue
Krzysztof Dembczynski
Vasil Denchev
Jia Deng
Li Deng
Xinwei Deng
Brian Depasquale
Guillaume Desjardins
Inderjit Dhillon
Paramveer Dhillon
Tom Diethe
Thomas Dietterich
Laura Dietz
Onur Dikmen
Chris Ding
Hu Ding
Nan Ding
Francesco Dinuzzo
Carlos Diuk
Santosh Divvala
Chuong Do
Eizaburo Doi
Justin Domke
Frank Dondelinger
Nan Du
Christopher DuBois
Lixin Duan
Artur Dubrawski
John Duchi
Miroslav Dudik
Haimonti Dutta
David Duvenaud
Jennifer Dy
Chris Dyer
Frederik Eaton
Elad Eban
Sandra Ebert
Douglas Eck
Alexander Ecker
David Eigen
Jacob Eisenstein
Jason Eisner
Carl Henrik Ek
Chaitanya Ekanadham
Khalid El-Arini
James Elder
Ehsan Elhamifar
Charles Elkan
Lloyd Elliott
Dan Ellis
Micha Elsner
Frank Emmert-Streib
Dominik Endres
Peter Englert
Tom Erez
Dumitru Erhan
Stefano Ermon
Tim van Erven
S. M. Ali Eslami
Robin Evans
Richard Everitt
Ludger Evers
Georgios Exarchakis
Shobeir Fakhraei
Clement Farabet
Amir-massoud Farahmand
Alireza Fathi
Paul Fearnhead
Vitaly Feldman
Jiashi Feng
Aasa Feragen
Rob Fergus
Xiaoli Fern
Carlos Fernandez-Granda

Vittorio Ferrari
Sanja Fidler
Mario Figueiredo
Maurizio Filippone
Marcelo Fiori
Asja Fischer
Jozsef Fiser
John Fisher III
Madalina Fiterau
Boris Flach
Peter Flach
Alyson Fletcher
Raphael Fonteneau
Florence Forbes
David Forsyth
Dean Foster
Nick Foti
James Foulds
Rina Foygel
Vojtech Franc
Eibe Frank
Peter Frazier
William Freeman
Oren Freifeld
Antonino Freno
Abe Friesen
Roger Frigola
Karl Friston
Mario Fritz
Rafael Frongillo
Johannes Fuernkranz
Satoru Fujishige
Kenji Fukumizu
Thomas Furmston
Nicolo Fusi
Alona Fyshe
Raphaël Féraud
Victor Gabillon
Thomas Gaertner
Stéphane Gaiffas
Juergen Gall
Marcus Gallagher
Patrick Gallinari
Joao Gama
Ravi Ganti
Jing Gao
Tianshi Gao
Xin Gao
Vikas Garg
Roman Garnett
Marta Garrido
Pierre Garrigues
Gilles Gasso
Rong Ge
Peter Gehler
Andreas Geiger
Philipp Geiger
Matthieu Geist
Michael Gelbart
Andrew Gelfand
Robert Gens
Sebastien Gerchinovitz
Sean Gerrish
Samuel Gershman
Pierre Geurts
Mohammad Gheshlaghi azar
Soumya Ghosh
Arnab Ghoshal
Bryan Gibson
Fabian Gieseke
Ran Gilad-Bachrach
Elad Gilboa
Jennifer Gillenwater
Mark Girolami
Michele Giugliano
Inmar Givoni
Tobias Glasmachers
Simon Godsill
Nico Goernitz
Vibhav Gogate
Hanlin Goh
Anna Goldenberg
Andrew Golightly
Daniel Golovin
Carla Gomes
Manuel Gomez-Rodriguez
Boqing Gong
Pinghua Gong
Yunchao Gong
Javier Gonzalez
Ian Goodfellow
Siddharth Gopal
Nakul Gopalan
Prem Gopalan
Raghuraman Gopalan
Sergiu Goschin
Stephen Gould
Amit Goyal
Navin Goyal
Agnieszka Grabska-Barwinska
Hans-Peter Graf
Alexandre Gramfort

Edouard Grave
Mihajlo Grbovic
Derek Greene
Karol Gregor
Russ Greiner
Arthur Gretton
Shane Griffith
Yuri Grinberg
Roger Grosse
Moritz Grosse-Wentrup
Peter Grunwald
Quanquan Gu
Arthur Guez
Vincent Guigue
Michele Guindani
Asela Gunawardana
Jiacheng Guo
Shengbo Guo
Yuhong Guo
Abhinav Gupta
Maya Gupta
Michael Gutmann
Isabelle Guyon
Mehmet Gönen
Minh Ha Quang
Michael Habeck
Hirotaka Hachiya
Ralf Haefner
Stefan Haeusler
Patrick Haffner
Richard Hahn
Will Hamilton
Jihun Hamm
Fred Hamprecht
Fang Han
Shaobo Han
Steve Hanneke
Lars-Kai Hansen
Satoshi Hara
Tatsuya Harada
Zaid Harchaoui
Kohei Hatano
Kohei Hayashi
Tamir Hazan
He He
Jingrui He
Creighton Heaukulani
Martial Hebert
Fares Hedayati
Nicolas Heess
Chinmay Hegde
Hoda Heidari
Matthias Hein
Uri Heinemann
Markus Heinonen
Saied Hemati
James Henderson
Guillaume Hennequin
James Hensman
Mark Herbster
Michiel Hermans
Daniel Hernandez-Lobato
José Miguel Hernández-Lobato
Alfred Hero
Aaron Hertzmann
Hideitsu Hino
Geoff Hinton
Michael Hirsch
Vaclav Hlavac
Chien-Ju Ho
Qirong Ho
Matt Hoffman
Matthew Hoffman
Steven Hoi
Sean Holden
Gary Holt
Antti Honkela
Xiaodi Hou
Jonathan How
Cho-Jui Hsieh
Chun-Nan Hsu
Daniel Hsu
Yuening Hu
Heng Huang
Jonathan Huang
Junzhou Huang
Sheng-Jun Huang
Shuai Huang
Tzu-Kuo Huang
Eyke Huellermeier
Jonathan Huggins
Michael Hughes
Dongsung Huh
Bui Hung
Jonny Hunt
Frank Hutter
Sung Ju Hwang
Aapo Hyvarinen
Morteza Ibrahimi
Tsuyoshi Ide
Kazushi Ikeda
Shiro Ikeda
Giacomo Indiveri

Marius C?t?lin Iordan
Charles Isbell
Masakazu Ishihata
Tomoharu Iwata
Rishabh Iyer
Shahin Jabbari
Laurent Jacob
Pierre Jacob
Robert Jacobs
Martin Jaggi
Ashesh Jain
Prateek Jain
Kevin Jamieson
Jeremy Jancsary
Majid Janzamin
Bruno Jedynak
Stefanie Jegelka
Rodolphe Jenatton
Alan Jern
Qiang Ji
Shuiwang Ji
Yangqing Jia
Jiarong Jiang
Jing Jiang
Ke Jiang
Tian Jiang
Luo Jie
Rong Jin
Adam Johansen
Matthew Johnson
Nebojsa Jojic
Vladimir Jojic
Galin Jones
Frederic Jurie
Pasi Jylänki
Ata Kaban
Hachem Kadri
Alfredo Kalaitzis
Satyen Kale
Yuri Kalnishkan
Alexandros Kalousis
Hirokazu Kameoka
Ulugbek Kamilov
Varun Kanade
Takafumi Kanamori
Purushottam Kar
Theofanis Karaletsos
Nikos Karampatziakis
Masayuki Karasuyama
Sergey Karayev
Amin Karbasi
Yan Karklin
Dimitris Karlis
Zohar Karnin
Hisashi Kashima
Samuel Kaski
Emilie Kaufmann
Koray Kavukcuoglu
Yoshinobu Kawahara
Jaya Kawale
Sathiya Keerthi
Balazs Kegl
Mikaela Keller
Kristian Kersting
Azadeh Khaleghi
Mohammad Emtiyaz Khan
Aditya Khosla
Bryan Kian Hsiang Low
Martin Kiefel
Dongho Kim
Kee-Eung Kim
Akisato Kimura
Irwin King
Franz Kiraly
Ryan Kiros
Negar Kiyavash
Hedvig Kjellström
Arto Klami
Edgar Klenske
Marius Kloft
David Knowles
Ryota Kobayashi
Jens Kober
Mikko Koivisto
Iasonas Kokkinos
Vladimir Kolmogorov
J. Zico Kolter
Vladlen Koltun
Mamoru Komachi
George Konidaris
Aryeh Kontorovitch
Wouter Koolen
Hema Koppula
Anoop Korattikara
Filip Korc
Nathaniel Korda
Wojciech Kotlowski
Adriana Kovashka
Matthieu Kowalski
Danica Kragic
Oswin Krause
Bartosz Krawczyk
Akshay Krishnamurthy
Dilip Krishnan

Balaji Krishnapuram
Oliver Kroemer
Rui Kuang
Alp Kucukelbir
Daniel Kuhn
Alex Kulesza
Abhishek Kumar
M. Pawan Kumar
Anshul Kundaje
Vitaly Kuznetsov
Branislav Kveton
James Kwok
Theo Kypraios
Arnd König
Urs Köster
Simon Lacoste-Julien
Kevin Lai
Balaji Lakshminarayanan
Diane Lambert
Christoph Lampert
Tian Lan
Yanyan Lan
Marc Lanctot
Niels Landwehr
Tobias Lang
Jan Larsen
Pavel Laskov
Longin Jan Latecki
Aurel Lazar
Alessandro Lazaric
Nevena Lazic
Hai-Son Le
Nicolas Le Roux
Chansoo Lee
Daniel Lee
Dongryeol Lee
Jason Lee
Joonseok Lee
Sangkyun Lee
Seunghak Lee
Tai Sing Lee
Wee Sun Lee
Leonidas Lefakis
Jing Lei
Joel Leibo
Victor Lempitsky
Chenlei Leng
Ian Lenz
Vincent Lepetit
Gilad Lerman
Sergey Levine
Fuxin Li
Hang Li
Jing Li
Li-Jia Li
Limin Li
Ping Li
Weixin Li
Wu-Jun Li
Yu-feng Li
Yujia Li
Zhenguo Li
Dawen Liang
Feng Liang
Li Liao
Xuejun Liao
Binbin Lin
Chih-Jen Lin
Dahua Lin
Hsuan-Tien Lin
Yuanqing Lin
Zhouchen Lin
Scott Linderman
Fredrik Lindsten
Christoph Lippert
Jennifer Listgarten
Michael Littman
Bo Liu
Ce Liu
Che-Yu Liu
Han Liu
Ji Liu
Jie Liu
Jun Liu
Qiang Liu
Shih-Chii Liu
Tie-Yan Liu
Wei Liu
Wenyu Liu
Xianghang Liu
Xiaoming Liu
Yan Liu
Ying Liu
Yufeng Liu
Zhandong Liu
James Lloyd
Po-Ling Loh
Maria Lomeli
Ben London
Bo Long
David Lopez-Paz
Daniel Lowd
Zhengdong Lu
Chris Lucas

REVIEWERS

Elliot Ludvig
Gediminas Luksys
Dijun Luo
Michael Lyu
Siwei Lyu
Miguel Lázaro-Gredilla
Rui Ma
Shiqian Ma
Yifei Ma
Liam MacDermed
Pierre Machart
Jakob Macke
Dougal Maclaurin
Guilherme Maeda
Sridhar Mahadevan
Vijay Mahadevan
Mehrdad Mahdavi
Michael Mahoney
Odalric-Ambrym Maillard
Julien Mairal
Alan Malek
Arian Maleki
Stephane Mallat
Hiroshi Mamitsuka
Chris Manning
Shie Mannor
Ioanna Manolopoulou
Vikash Mansinghka
Yishay Mansour
Oded Margalit
Benjamin Marlin
James Martens
Radoslaw Martin Cichy
Andre Martins
Pekka Marttinen
Jonathan Masci
Tomoko Matsui
Rahul Mazumdar
David McAllester
Julian McAuley
Andrew McCallum
Brian McFee
Andrew McHutchon
Scott McQuade
Brian McWilliams
Ted Meeds
Ron Meir
aditya Menon
João Messias
Elad Mezuman
Tomas Mikolov
Andrew Miller
David Mimno
Tom Minka
Mehdi Mirza
Baharan Mirzasoleiman
Ioannis Mitliagkas
Dunja Mladenic
Andriy Mnih
Volodymyr Mnih
Hossein Mobahi
Daichi Mochihashi
Joseph Modayil
Shakir Mohamed
Mehryar Mohri
Andrea Montanari
Gregoire Montavon
Claire Monteleoni
Guido Montufar
Joris Mooij
Sebastian Moreno
Greg Mori
Tetsuro Morimura
Edward Moroshko
Quaid Morris
Morten Morup
Alessandro Moschitti
Michael Mozer
Youssef Mroueh
Krikamol Muandet
Andreas Mueller
Klaus-Robert Mueller
Sayan Mukherjee
Remi Munos
Andres Munoz Medina
Vittorio Murino
Brendan Murphy
Robert Murphy
Alejandro Murua
Pablo Muse
Gautham Mysore
Boaz Nadler
Sri Nagarajan
Vinod Nair
Shinichi Nakajima
Nils Napp
Mukund Narasimhan
Hari Narayanan
Nagarajan Natarajan
Saketha Nath
Willie Neiswanger
Blaine Nelson
James Nelson
Shamim Nemati
Bernhard Nessler
Gergely Neu
James Neufeld
Gerhard Neumann
Behnam Neyshabur
Huy Nguyen
Minh Hoai Nguyen
Trung Nguyen
Hannes Nickisch
Alexandru Niculescu-Mizil
Feiping Nie
Juan Carlos Niebles
Gang Niu
William Noble
Richard Nock
Yung-Kyun Noh
David Nott
Rob Nowak
Rebecca Nugent
Tim Oates
Guillaume Obozinski
Masato Okada
Aude Oliva
Peder Olsen
Bruno Olshausen
Takshi Onoda
Luis Ortiz
Ronald Ortner
Sarah Osentoski
Takayuki Osogami
FRANCESCA PETRALIA
Deepti Pachauri
Adam Packer
Benjamin Packer
David Page
Jean-Francois Paiement
Brooks Paige
Ari Pakman
Christopher Pal
David Pal
Konstantina Palla
Jason Palmer
Liam Paninski
Tivadar Papai
Dimitris Papailiopoulos
George Papandreou
Ulrich Paquet
Alexandros Paraschos
Ankur Parikh
Il Park
Mijung Park
Charles Parker
Johannes Partzsch
Razvan Pascanu
alexandre Passos
Genevieve Patterson
Vladimir Pavlovic
Klaus Pawelzik
Barak Pearlmutter
Marcello Pelillo
Tomi Peltola
Jaakko Peltonen
Jian Peng
Vianney Perchet
Daniel Percival
Fernando Pereira
Julien Perez
Fernando Perez-Cruz
Alessandro Perina
Gareth Peters
Jan Peters
Marek Petrik
Nico Pfeifer
Joseph Pfeiffer
Natesh Pillai
Jonathan Pillow
Joelle Pineau
Hamed Pirsiavash
John Platt
Patrick Pletscher
Eftychios Pnevmatikakis
Barnabas Poczos
Florian Pokorny
Jan Poland
Jean-Baptiste Poline
Hoifung Poon
Pascal Poupart
Daniel Povey
Denis Prangle
Doina Precup
Philippe Preux
Alexandre Proutiere
Jay Pujara
Guido Pusiol
Yanjun Qi
Tao Qin
Novi Quadrianto
Qichao Que
Neil Rabinowitz
Guillaume Rabusseau
Luis Rademacher
Gunnar Raetsch
Anna Rafferty
Ali Rahimi
Piyush Rai
Alain Rakotomamonjy
Liva Ralaivola
Peter Ramadge
Subramanian Ramamoorthy
Vignesh Ramanathan
Harish Ramaswamy
Aaditya Ramdas
Arti Ramesh
Jan Ramon
Rajesh Ranganath
Vinayak Rao
Carl Rasmussen
Vibhor Rastogi
Magnus Rattray
Sujith Ravi
Pradeep Ravikumar
Balaraman Ravindran
Debajyoti Ray
Soumya Ray
Vikas Raykar
Christopher Re
Khaled Refaat
Roi Reichart
David Reichert
Mark Reid
Joseph Reisinger
Thodoris Rekatsinas
Marcello Restelli
Manuel Reyes-Gomez
Lev Reyzin
Iead Rezek
Emile Richard
Thomas Richardson
Sebastian Riedel
Ludovic Righetti
Michael Riley
Irina Rish
Yaacov Ritov
Christian Robert
St?phane Robin
Abel Rodriguez
Simon Rogers
Marcus Rohrbach
Jaldert Rombouts
Teemu Roos
Romer Rosales
Lorenzo Rosasco
David Rosenberg
Arun Ross
Fabrice Rossi
Andrew Roth
Volker Roth
Constantin Rothkopf
Tim Roughgarden
Juho Rousu
Benjamin Van Roy
Daniel Roy
Nicholas Roy
Ben Rubinstein
Alessandro Rudi
Nicholas Ruozzi
Alexander Rush
Olga Russakovsky
Bryan Russell
Daniel Russo
Andreas Ruttor
Paul Ruvolo
Daniil Ryabko
Sivan Sabato
Regis Sabbadin
Ashish Sabharwal
Robin Sabhnani
Mohammad Amin Sadeghi
Ankan Saha
Hiroto Saigo
Tara Sainath
Jun Sakuma
Ruslan Salakhutdinov
Venkatesh Saligrama
Joseph Salmon
Michael Salter-Townshend
Mathieu Salzmann
Wojciech Samek
Guido Sanguinetti
Aswin Sankaranarayanan
Sriram Sankararaman
Scott Sanner
George Saon
Guillermo Sapiro
Purnamrita Sarkar
Anand Sarwate
Issei Sato
Lawrence Saul
Richard Savage
Cristina Savin
Christoph Sawade
Andrew Saxe
Stefan Schaal
Robert Schapire
Tobias Scheffer
Bruno Scherrer
Bernt Schiele
Alexander Schliep
Tanya Schmah
Jürgen Schmidhuber
Jeff Schneider
Thomas Schon
Hannes Schulz
Haim Schweitzer
Alex Schwing
Clayton Scott
Michele Sebag
Hanie Sedghi
Jeffrey Seely
Dino Sejdinovic
Yevgeny Seldin
Sundararajan Sellamanickam
Bart Selman
Mike Seltzer
Andrew Senior
Walter Senn
Pierre Sermanet
Simon Setzer
Eleni Sgouritsa
Ben Shababo
Patrick Shafto
Mohak Shah
Uri Shalit
Cosma Shalizi
Ohad Shamir
Nataliya Shapovalova
Tatyana Sharpee
James Sharpnack
Madhusudana Shashanka
Abdul Saboor Sheikh
Daniel Sheldon
Christian Shelton
Jacquelyn Shelton
Bin Shen
Chunhua Shen
Li Shen
Xiaotong Shen
Lei Shi
Qinfeng Shi
Motoki Shiga
Nobuyuki Shimizu
Shohei Shimizu
Shigeru Shinomoto
Pannaga Shivaswamy
Jon Shlens
Susan Shortreed
Lavi Shpigelman
Abhinav Shrivastava
Anshumali Shrivastava
Si Si
Leonid Sigal
Marco Signoretto
Tomi Silander
David Silver
Khe Chai Sim
Noah Simon
Karen Simonyan
Ozgur Simsek
Vikas Sindhwani
Ajit Singh
Sameer Singh
Vikas Singh
Adish Singla
Kaushik Sinha
Mathieu Sinn
Fabian Sinz
Scott Sisson
Josef Sivic
Martin Slawski
Kevin Small
Paris Smaragdis
Cristian Sminchisescu
Adam Smith
Padhraic Smyth
Jasper Snoek
Anthony Man-Cho So
Richard Socher
Eric Sodomka
Jascha Sohl-Dickstein
Kihyuk Sohn
Kyung-Ah Sohn
Arno Solin
Justin Solomon
Mahdi Soltanolkotabi
Hagen Soltau
Fritz Sommer
Diego Sona
Le Song
Cheng Soon Ong
Daniel Soudry
Luigi Spezia
Luciano Spinello
Pablo Sprechmann
Karthik Sridharan
Nitish Srivastava
Daniel Stefankovic
Jacob Steinhardt
Florian Steinke
Ingo Steinwart
Philip Sterne
Alan Stocker
Greg Stoddard
Jack Stokes
Dmitry Storcheus
Heiko Strathmann
Karl Stratos
Andreas Stuhlmller
Erik Sudderth
Mahito Sugiyama
Masashi Sugiyama
Dennis Sun
Liang Sun
Tingni Sun
Yuekai Sun
Matyas Sustik
Ilya Sutskever
Charles Sutton
Richard Sutton
Johan Suykens
Joe Suzuki
Taiji Suzuki
Kevin Swersky
Pawel Swietojanski
Zeeshan Syed
Adam Sykulski
Zoltan Szabo
Marie Szafranski
Christian Szegedy
Arthur Szlam
Raphael Sznitman
Balázs Szörényi
Simo Särkkä
Yasuo Tabei
Nima Taghipour
Akiko Takeda
Takashi Takenouchi
Ichiro Takeuchi
Ichigaku Takigawa
Eiji Takimoto
Partha Talukdar
Erik Talvitie
Ameet Talwalkar
Pablo Tamayo
Cheston Tan
Toshiyuki Tanaka
Cheng Tang
Kui Tang
Yichuan Tang
Dacheng Tao
Daniel Tarlow
Nikolaj Tatti
Graham Taylor
Georgios Theocharous
Romain Thibaux
Bertrand Thirion
Philip Thomas
Nicolas Thome
Ryan Tibshirani
Robert Tillman
Diego Tipaldi
Ivan Titov
Michalis Titsias
Sinisa Todorovic
Ryota Tomioka
Hanghang Tong
Luis Torgo
Alexander Toshev
Ivana Tosic
Behrouz Touri
Kristina Toutanova
Volker Tresp
Bill Triggs
Ivor Tsang
Evgeni Tsivtsivadze
Richard Turner
Ryan Turner
Hemant Tyagi
Stephen Tyree
Naonori Ueda
Balazs Ujfalussy
Lyle Ungar
Ruth Urner
Raquel Urtasun
Benigno Uría
Gregory Valiant
Michal Valko
Jurgen Van Gael
Herke Van Hoof
Guy Van den Broeck
Vincent Vanhoucke
Lav Varshney
Eleni Vasilaki
Andrea Vedaldi
Sandro Vega Pons
Shankar Vembu
Archana Venkataraman
Suresh katasubramanian
Dan Ventura
Deepak Venugopal
Nakul Verma
Paul Vernaza
Jean-Philippe Vert
Rene Vidal
Matthieu Vignes
Sudheendra Vijayanarasimhan
Silvia Villa
Brett Vinch
Vibhav Vineet
Oriol Vinyals
Fabio Vitale
Max Vladymyrov
Nikos Vlassis
Joshua Vogelstein
Julia Vogt
Maksims Volkovs
Vincent Vu
Slobodan Vucetic
Ed Vul
Vladislav Vyshemirsky
Willem Waegeman
Stefan Wager
Christian Walder
Thomas Wallis
Thomas Walsh
Tom Walters
Chong Wang
Hua Wang
Huahua Wang
Huayan Wang
Jack Wang
Jingdong Wang
Joseph Wang
Jun Wang
Jun Wang
Junhui Wang
Lei Wang
Liming Wang
Liwei Wang
Mengdi Wang
Michelle Wang
Shaojun Wang
Shusen Wang
Sida Wang
Wei Wang
Weiran Wang
Yang Wang
Yiwen Wang
Yu-Xiang Wang
Zhaoran Wang
Zhuo Wang
Ziyu Wang
Zuoguan Wang
Takashi Washio
Larry Wasserman
Kazuho Watanabe
Shinji Watanabe
Chris Watkins
Kevin Waugh
Fabian Wauthier
Chu Wei
Kai Wei
Markus Weimer
David Weiss
Zaiwen Wen
Zheng Wen
Tomas Werner
Martha White
Nick Whiteley
Shimon Whiteson
Andre Wibisono
Michael Wick
Jenna Wiens
Rich Wilkinson
Chris Williams
Sinead Williamson
Andrew Wilson
Robert Wilson
David Wingate
John Winn
Ole Winther
Ernst Wit
Lei Wu
Qiang Wu
Si Wu
Teresa Wu
Xiao-Ming Wu
Jason Wyse
Lirong Xia
Jing Xiang
Jianxiong Xiao
Lin Xiao
Lexing Xie
Yu Xin
Fei Xu
Huan Xu
Linli Xu
Miao Xu
Min Xu
Minjie Xu
Zenglin Xu
Zhixiang (Eddie) Xu
Oksana Yakhnenko
Makoto Yamada
Feng Yan
shuicheng Yan
Keiji Yanai
Pinar Yanardag
Jimei Yang
Liu Yang
Ming Yang
Ming-Hsuan Yang
Qiang Yang
Scott Yang
Shuanghong Yang
Shulin Yang
Tianbao Yang
Weilong Yang
Yi Yang
Zhi Yang
Zhirong Yang
Angela Yao
Bangpeng Yao
Yuan Yao
Christopher Yau
Jieping Ye
Nan Ye
Dit-Yan Yeung
Florian Yger
Jinfeng Yi
Scott Yih
Junming Yin
Yiming Ying
Chang D. Yoo
Junichiro Yoshimoto
Chun-Nam Yu
Felix Yu
Hsiang-Fu Yu
Jia Yuan Yu
Yao-Liang Yu
Yisong Yue
Hyokun Yun
Bianca Zadrozny
Zahra Zamani
Thorsten Zander
Giovanni Zappella
Wojciech Zaremba
Mikhail Zaslavskiy
Hongyuan Zha
Changshui Zhang
Chiyuan Zhang
Haichao Zhang
Hao Helen Zhang
Jiji Zhang
Kun Zhang
Lei Zhang
Lijun Zhang
Liqing Zhang
Shunan Zhang
Wenhao Zhang
Xianxing Zhang
Xinhua Zhang
Yi Zhang
Yichuan Zhang
Yu Zhang
Yuchen Zhang
Zhihua Zhang
Tuo Zhao
Zheng Zhao
Yi Zhen
Chunxiao Zhou
Dengyong Zhou
Guang-Tong Zhou
Jiayu Zhou
Ke Zhou
Mingyuan Zhou
Zhi-Hua Zhou
Jun Zhu
Shenghuo Zhu
Song-Chun Zhu
Xiaojin Zhu
Hankz Hankui Zhuo
Brian Ziebart
Andrew Ziegler
Martin Zinkevich
Justin Zobel
Onno Zoeter
Daniel Zoran
Will Zou
Or Zuk
Alon Zweig
Geoffrey Zweig
Florence d'Alche-Buc
Cassio de Campos
Teofilo de Campos
Nando de Freitas
Fernando de la Torre
Jan Willem van de Meent
Laurens van der Maaten
James von Brecht=

AUTHOR INDEX

Abbeel, Pieter: Spotlight Session 7: Reinforcement learning spotlights, W66
Acerbi, Luigi: Th12
Acharya, Jayadev: M31
Ackerman, Margareta: Th82
Adametz, David: W16
Adams, Ryan: M54
Agarwal, Shivani: M24, W73
Agarwal, Alekh: T57
Aitchison, Laurence: M45
Akram, Sahar: W97
Al Faraby, Said: DeMstrations
Alquier, Pierre: Th77
Alsharif, Ouais: W40
Alvarez-Icaza, Rodrigo: DeMstrations
Amershi, Saleema: DeMstrations
Amin, Kareem: Th57
Amir, Arnon: DeMstrations
Ammar, Waleed: Oral Session 4, T41
Anandkumar, Animashree: Spotlight Session 5: Regression and time series spotlights, W101
Anandkumar, Anima: Th87
Andersen, David: Th68
Andersen, Michael: W6
Andersson Naesseth, Christian: Th32
Andreas, Jacob: Spotlight Session 5: Regression and time series spotlights, W98
Andreas, Jacob: DeMstrations
Antonucci, Alessandro: W18
Arbelaez, Pablo: T85
Archer, Evan: M43
Arlot, Sylvain: T37
Arnemann, James: M46
Arora, Raman: Th88
Arpit, Devansh: M90
Aslan, Özlem: W55
Avasarala, Bhargav: Th34
Avouac, Jean Philippe: T8
Avron, Haim: M19
Awasthi, Pranjal: Spotlight Session 6: Learning theory spotlights, W45
Azari Soufiani, Hossein: Oral Session 10, Th53
Babadi, Behtash: W97
Babbush, Ryan: Th60
Bachman, Phil: W40
Bach, Francis: M38, T37
Baez, John Carlos: Oral Session 5
Bahadori, Mohammad Taha: Spotlight Session 5: Regression and time series spotlights, W103
Bahdanau, Dzmitry: DeMstrations
Ba, Jimmy: W36
Balcan, Maria-Florina: M17, T20, W20, Th93
Baldi, Pierre: Spotlight Session 4: Deep spotlights, T103
Bamford, Sim: DeMstrations
Banerjee, Arindam: Climate Change: Challenges for Machine Learning, M36, M75, W7, Th86
Bansal, Trapit: M82
Bareinboim, Elias: Spotlight Session 9: Graphical models spotlights, model selection, Th74
Bartlett, Peter: M22, W83
Bartz, Daniel: T62
Bartók, Gábor: T24
Bassetto, Giacomo: Spotlight Session 3: Neuroscience and Neural Coding spotlights, T49
Bateni, Mohammadhossein: M91
Batista, Jorge: T88
Batra, Dhruv: Spotlight Session 1: Optimization spotlights, T39
Becker, Stephen: M39, Th43
Belkin, Mikhail: M16
Bengio, Yoshua: M57, Oral Session 4, DeMstrations, T95, T96, W37, W39
Benson, Austin: Spotlight Session 2: Large scale learning spotlights, T75
Bento, Jose: Th28
Berend, Daniel: M33
Berg-Kirkpatrick, Taylor: Spotlight Session 5: Regression and time series spotlights, DeMstrations, W98
Berlind, Christopher: W20
Besbes, Omar: T32
Beygelzimer, Alina: T57
Bhaskara, Aditya: M91
Bhatnagar, Shalabh: Th58
Bhattacharyya, Chiranjib: M82, T45
Bi, Jinbo: Th101
Bilmes, Jeff: Tutorial Session B, Tutorial Session B, Tutorial Session B, Oral Session 1, T86, Th37
Bishop, William: W9
Blei, David: T14, Th17, Th18
Blum, Avrim: Spotlight Session 6: Learning theory spotlights, W20, W45, Th54
Boix, Xavier: M66
Bonilla, Edwin: T28, Spotlight Session 8: GP spotlights, kernel spotlights, sampling spotlights, classification spotlights, W33
Borg, Jana: Th98
Bottone, Steven: Th34
Botvinick, Matthew: Spotlight Session 3: Neuroscience and Neural Coding spotlights, T73
Boyd-Graber, Jordan: M83
Boyd, Stephen: Spotlight Session 1: Optimization spotlights, T4
Bradley, Joseph: W50
Bregler, Christoph: Th35
Brendel, Wieland: Spotlight Session 10: Cognitive science spotlights, Th10
Bresler, Guy: M99, Spotlight Session 2: Large scale learning spotlights, T11, Th31
Brox, Thomas: Th5
Bruer, John: M39
Bruna, Joan: Th1
Bruntink, Magiel: DeMstrations
Buesing, Lars: Spotlight Session 3: Neuroscience and Neural Coding spotlights, T50
Buhmann, Joachim: Spotlight Session 5: Regression and time series spotlights, W48
Bui, Thang: Spotlight Session 8: GP spotlights, kernel spotlights, sampling spotlights, classification spotlights, W34
Caetano, Tiberio: Spotlight Session 8: GP spotlights, kernel spotlights, sampling spotlights, classification spotlights, W47
Calandriello, Daniele: M27
Cameron, Daniel: M52
Candes, Emmanuel: Spotlight Session 1: Optimization spotlights, T4
Canny, John: DeMstrations
Caramanis, Constantine: M88
Carbonell, Jaime: W99
Cardie, Claire: Th8
Carin, Lawrence: M81, T67, T69, W78, Th98
Carlson, David: T69, Th98
Carpentier, Alexandra: T33
Caruana, Rich: W36
Caseiro, Rui: T88
Cevher, Volkan: M39, W53
Chakrabarty, Deeparnab: Oral Session 1, T2
Chamandy, Nick: T54
Chandar A P, Sarath: M50
Chang, Shih-Fu: Spotlight Session 2: Large scale learning spotlights, T60
Chang, Jonathan: M83
Chang, Hong: Th94
Chang, Jason: M14
Chan, Hau: Th51
Chan, Siu On: M5
Chao, Wei-Lun: T84
Charlin, Laurent: T14
Chatterjee, Soumyadeep: M75
Chaudhuri, Kamalika: T51, Spotlight Session 6: Learning theory spotlights, Spotlight Session 6: Learning theory spotlights, W11, W75
Chawla, Sanjay: M86
Chekuri, Chandra: Recent Progress in the Structure of Large-Treewidth Graphs and Some Applications
Chellappa, Rama: Th26
Chen, Changyou: M85, Th60
Chen, George: Spotlight Session 2: Large scale learning spotlights, T11
Chen, Chao: Spotlight Session 9: Graphical models spotlights, model selection, Th38
Cheng, James: W22
Cheng, Xiuyuan: Th2
Cheng, Hong: W22
Chen, Xianjie: M71
Chen, Yutian: T29
Chen, Sheng: M75, W7
Chen, Minhua: T67
Chen, Xuanli: T83
Chen, Shouyuan: Oral Session 6, W2
Chen, Wei: Oral Session 6, W2
Chen, Xi: Spotlight Session 5: Regression and time series spotlights, W95
Chen, Xilin: Th94
Chen, Weizhu: Spotlight Session 2: Large scale learning spotlights, T5
Chen, Dongqu: Th80
Chen, Yuheng: T87
Chen, Yudong: Spotlight Session 2: Large scale learning spotlights, T68
Chen, Qingcai: Th7
Chen, Xu: Th2
Chichilnisky, E. J.: M42
Chickering, Max: DeMstrations
Choi, ArThr: W17
Choi, Joon Hee: T55
Chow, Yinlam: M26
Cho, Kyunghyun: M57, DeMstrations, T95, W39
Chopin, Nicolas: Th77
Chwialkowski, Kacper: Oral Session 8, W29
Clune, Jeff: Oral Session 4, W37
Cohen, Emma: W20
Cohen, Haim: Th103
Conejo, Bruno: T8
Conitzer, Vincent: Computing Game-Theoretic Solutions
Courville, Aaron: T96
Crammer, Koby: Th103
Cui, Zhen: Th94
Cunningham, John: Spotlight Session 3: Neuroscience and Neural Coding spotlights, T26, T50
d'AspreMt, Alexandre: M2
Dahleh, Munther: Th96
Dai, Bo: M17
Dalal, Onkar: Th41
Darrell, Trevor: M64, T89, T92
Darwiche, Adnan: W17
Dasgupta, Sanjoy: Spotlight Session 6: Learning theory spotlights, W75
Dasgupta, Sanjoy: T44, Th82
Daume III, Hal: W54
Dauphin, Yann: W39
Davis, Larry: W42
Dayan, Peter: W64
De Bock, Jasper: W18
de Campos, Cassio: W18, Spotlight Session 9: Graphical models spotlights, model selection, Th44
Defazio, Aaron: M38
de Freitas, Nando: M97
Dekel, Ofer: W104
Denil, Misha: M97
Denli, Huseyin: W19
Denton, Emily: Th1
Denève, Sophie: Spotlight Session 3: Neuroscience and Neural Coding spotlights, T47
Derezinski, Michal: Spotlight Session 6: Learning theory spotlights, W46
Deshpande, Yash: T76, Th92
Dhillon, Inderjit: M37, M80, T38, Spotlight Session 8: GP spotlights, kernel spotlights, sampling spotlights, classification spotlights, W26, W30, W80, W91, Th43
Diakonikolas, Ilias: M5
Diego Andilla, Ferran: Spotlight Session 3: Neuroscience and Neural Coding spotlights, T63
Diether, SaloM: M66
Dimakis, Alexandros: Oral Session 5, W10, W41
Ding, Chris: Th105
Ding, Nan: Th60
Djolonga, Josip: W51
Dollar, Piotr: T90
Domingos, Pedro: T97
Domke, Justin: M95
Donahue, Jeff: T89
Doniec, Marek: T25
Dosovitskiy, Alexey: Th5
Doucet, Arnaud: Oral Session 8, W63
Drugowitsch, Jan: Spotlight Session 10: Cognitive science spotlights, Th14
Dubey, Kumar: M11
Du, Lin: M81
Du, Nan: T19, T20
Dukkipati, Ambedkar: M94
Dundar, Murat: M79
Dunson, David: Oral Session 2, T58
du Plessis, Marthinus: W81
Duvenaud, David: Oral Session 8, W32
Dwork, Cynthia: Oral Session 9
Dyer, Chris: Oral Session 4, T41
Dzirasa, Kafui: T69, Th98

AUTHOR INDEX

Eigen, David: Spotlight Session 10: Cognitive science spotlights, Th29
Eisner, Jason: W54
Eslami, S. M. Ali: Th66
Evans, Robin: Th75
Fadili, Jalal: M40
Fang, Youhan: Th60
Fan, Wei: W22
Farajtabar, Mehrdad: T19
Fazayeli, Farideh: W7
Feizi, Soheil: Th76
Feldman, Dan: Th47
Felzenszwalb, Pedro: M63
Feng, Xiangchu: Th71
Feng, Jiashi: W87
Fergus, Rob: Spotlight Session 4: Deep spotlights, T100, Spotlight Session 10: Cognitive science spotlights, Th1, Th29
Festa, Dylan: Oral Session 3, T72
Feyereisl, Jan: Th22
Fisher III, John: M14, Th47
Fithian, William: W70
Fletcher, Alyson: Spotlight Session 3: Neuroscience and Neural Coding spotlights, T70
Fleuret, François: DeMstrations
Fogel, Fajwel: M2
Foster, Dean: Th91
Foti, Nicholas: Th21
Fox, Emily: Th21, Th25
Fragkiadaki, Katerina: T85
Franzen, Florian: Spotlight Session 3: Neuroscience and Neural Coding spotlights, T49
Freeman, Bill: Th28
Frigola, Roger: T29
Fritz, Mario: T82
Frostig, Roy: M41
Fujimaki, Ryohei: Th78, Th105
Fung, C. C. Alan: Th97
Févotte, Cédric: W96
Galstyan, Aram: Th46
Gal, Yarin: Th63
Gamarnik, David: M99, Th31
Ganguli, Surya: W39
Garcia Jurado Suarez, Carlos: DeMstrations
Garg, Ankit: Oral Session 7, W44
Garnett, Roman: W35
Garreau, Damien: T37
Geist, Matthieu: Spotlight Session 7: Reinforcement learning spotlights, W69
Gendreau, Michel: Spotlight Session 7: Reinforcement learning spotlights, W56
Gens, Robert: T97
Gerdzhev, Martin: DeMstrations
Gershman, Samuel: Spotlight Session 3: Neuroscience and Neural Coding spotlights, T73
Ghahramani, Zoubin: M10, Spotlight Session 1: Optimization spotlights, T7, T30
Ghavamzadeh, Mohammad: M26
Ghoshdastidar, Debarghya: M94
Ghosh, Joydeep: Th95
Gibson, Garth: Th69
Gilboa, Elad: T26
Gillenwater, Jennifer: Th25
Girshick, Ross: T89
Glazer, Assaf: Th48
Gleich, David: Spotlight Session 2: Large scale learning spotlights, T75
Goddard, Nigel: W102
Gogate, Vibhav: W24, W58, W60
Goldberg, Yoav: W23
Gomez-Rodriguez, Manuel: T19
Gomez, Faustino: M56
Gong, Boqing: T84
Gonzalez, Joseph: W50
Goodfellow, Ian: T96
Gool, Luc: M66, T83
Gopalan, Prem: T14
Gordon, Geoffrey: Oral Session 7
Gottlieb, Lee-Ad: Th79
Govindaraju, Venu: M76, M90
Goyal, Prasoon: W58
Grabska-Barwinska, Agnieszka: T46
Grahn, Jessica: M52
Grandvalet, Yves: W79
Grauman, Kristen: M68, M70, T84
Graves, Alex: Spotlight Session 10: Cognitive science spotlights, Th9
Green, Stephen: Spotlight Session 9: Graphical models spotlights, model selection, Th65
Greiner, Russell: Th72
Gretton, ArThr: Oral Session 8, W29
Griffiths, Thomas: Th11
Grinberg, Yuri: Spotlight Session 7: Reinforcement learning spotlights, W56
Grünwald, Peter: M20
Guadarrama, Sergio: T89
Guez, ArThr: W64
Guillaumin, Matthieu: T83
Gunter, Tom: W35
Guo, Xiaoxiao: T94
Gu, Shuhang: Th71
Gu, Quanquan: T80, Th89
Guestrin, Carlos: Th37
Gui, Huan: T80
Gulcehre, Caglar: W39
Gur, Yonatan: T32
Gärtner, Bernd: W8
Gönen, Mehmet: M15
Haghtalab, Nika: Th54
Hajek, Bruce: T15
Hamprecht, Fred: Spotlight Session 3: Neuroscience and Neural Coding spotlights, T63
Hamrick, Jessica: Th11
Han, Joon Hee: T90
Han, Bohyung: Th22
Han, Shaobo: M81
Han, Jiawei: T80
Hansen, Lars: W6
Ha Quang, Minh: Spotlight Session 8: GP spotlights, kernel spotlights, sampling spotlights, classification spotlights, W27
Harchaoui, Zaid: Spotlight Session 4: Deep spotlights, T104
Hardt, Moritz: Spotlight Session 1: Optimization spotlights, T10
Hassibi, Babak: M89
Hauptmann, Alexander: Th30
Hay, Nicholas: Th11
Hazan, Elad: T31, W104
Heess, Nicolas: W64
Heess, Nicolas: Spotlight Session 10: Cognitive science spotlights, Th9
He, He: W54
Hein, Matthias: Th49
Henao, Ricardo: W78
Hendriks, Remco: DeMstrations
He, Niao: M17
Hennequin, Guillaume: M45, Oral Session 3, T72
Hennig, Philipp: Oral Session 8, W32, W35, Th64
Henriques, João F.: T88
Hermans, Tucker: DeMstrations
Hernández-Lobato, José Miguel: Spotlight Session 1: Optimization spotlights, T7, T30
Hernández-lobato, Daniel: T27
Hero, Alfred: W43
Hewitt, Luke: DeMstrations
Hoffman, Judy: T89
Hoffman, Matthew: Spotlight Session 1: Optimization spotlights, T7
Hofmann, Thomas: Th70
Holmes, Susan: W61
Hong, Mingyi: Th83
Hopfield, John J.: Oral Session 7
Ho, Qirong: M11, Th69
Houlsby, Neil: Th17
Hsieh, Cho-Jui: M37, W30, Th43
Hsu, Daniel: T51, T57
Huang, Daniel: Spotlight Session 9: Graphical models spotlights, model selection, Th65
Huang, Yanping: M53
Huang, Ling: W83
Huang, Thomas: M84
Hubara, Itay: M49
Hu, Huining: Spotlight Session 9: Graphical models spotlights, model selection, Th40
Hu, Baotian: Th7
Hu, Ronghang: T89
Huk, Alexander: M44
Hwang, Sung Ju: M69
Hölscher, Janine: DeMstrations
Ihler, Alex: M8
Inouye, David: M80
Irfan, Mohammad: Spotlight Session 10: Cognitive science spotlights, Th52
Irsoy, Ozan: Th8
Iwata, Tomoharu: W28
Iyer, Rishabh: T86
Jaakkola, Tommi: T13
Jafarpour, Ashkan: M31
JagabaThla, Srikanth: T53
Jaggi, Martin: Th70
Jain, Arjun: Th35
Jain, Prateek: M73, Oral Session 1, T2, T21, Spotlight Session 5: Regression and time series spotlights, W21, W101
janoos, firdaus: W19
Jawahar, C.V.: T16
Jayaraman, Dinesh: M70
Jebara, Tony: Oral Session 9, Spotlight Session 9: Graphical models spotlights, model selection, Th36, Th73
Jegelka, Stefanie: M35, M64, Spotlight Session 1: Optimization spotlights, T39, W50
Jia, Jiaya: W38
Jiang, Biye: DeMstrations
Jiang, Xueyan: Spotlight Session 7: Reinforcement learning spotlights, W25
Jiang, Zhuolin: Th26
Jiang, Albert: Th55
Jiang, Lu: Th30
Jin, Rong: M1, Th81
Ji, Qiang: Spotlight Session 9: Graphical models spotlights, model selection, Th44
Jimenez Rezende, Danilo: Spotlight Session 4: Deep spotlights, T93
Jonckheere, EdMd: Th87
Jordan, Michael: M35, Spotlight Session 5: Regression and time series spotlights, W50, W95, Th70
Jose, Cijo: DeMstrations
Joseph, Anthony: W83
Joulin, Armand: M61
Kairouz, Peter: T52
Kanchanapally, Vandana: Th93
Kangas, Kustaa: Th42
Kannan, Ravindran: M82
Kantchelian, Alex: W83
Kar, Purushottam: M73, T21
Karpathy, Andrej: M61
kavukcuoglu, koray: Spotlight Session 10: Cognitive science spotlights, Th9
Kearns, Michael: Oral Session 10
Kedenburg, Gunnar: T22
Kersting, Kristian: T27
Kervrann, Charles: T81
Khajah, Mohammad: Th13
Khanna, Rajiv: Th95
Khan, Mohammad: Th20
Khapra, Mitesh: M50
Khare, Kshitij: Th41
Kim, Been: M93
Kim, Jin Kyu: Th69
Kim, Seyoung: M100
King, Irwin: Oral Session 6, W2
Kingma, Diederik: Spotlight Session 4: Deep spotlights, T93
Kiros, Ryan: DeMstrations, T102
Klein, Dan: Spotlight Session 5: Regression and time series spotlights, DeMstrations, W98
Klivans, Adam: Oral Session 5, W10
Klopp, Olga: T12
Knudson, Karin: M44
Kobayashi, Hiroko: Th19
Kocaoglu, Murat: Oral Session 5, W10
Koch, Christof: Integrated Information Theory of Consciousness - Conceptual Formulation & Adventures in Simulated Evolution
Kocák, Tomáš: W4
Kohli, Pushmeet: Th66
Kohn, Adam: Th99
Koivisto, Mikko: Th42
Komodakis, Nikos: T8
Konda, Kishore: Th6
Kondor, Risi: W94
Kong, Deguang: Th105
Koniusz, Piotr: Spotlight Session 4: Deep spotlights, T104
Kontorovitch, Aryeh: M33, Th79
Koolen, Wouter: M20, M22
Kopeć, Łukasz: Spotlight Session 10: Cognitive science spotlights, Th15
Koren, Tomer: W104
Korlakai Vinayak, Ramya: M89
Koster, Urs: M43
Kothari, Pravesh: Oral Session 1, T2
Kowalski, Matthieu: W96
Koyejo, Oluwasanmi: M12, Spotlight Session 8: GP spotlights, kernel spotlights, sampling spotlights, classification spotlights, W80, Th95
Kpotufe, Samory: T44
Krause, Andreas: T24, W8, W51
Krishnan, Dilip: Th28
Krishnan, Sanjay: Th70
Krummenacher, Gabriel: Spotlight Session 5: Regression and time series spotlights, W48
Krzakala, Florent: W93
Kulesza, Alex: Th25

AUTHOR INDEX

Kumar, Sanjiv: Spotlight Session 2: Large scale learning spotlights, T60, Oral Session 8
Kumar, M. Pawan: T1, T16
Kurach, Karol: Spotlight Session 4: Deep spotlights, T100
Kuznetsov, Vitaly: W86
Kwak, Suha: Th22
Lacoste-Julien, SiM: M38
Lafferty, John: M34, M96
Lafond, Jean: T12
Laird, Dillon: Th21
Lajugie, Rémi: T37
Lakshminarayanan, Balaji: M13, Th67
Lakshmiratan, Aparna: DeMstrations
Lampert, Christoph: T27
Langford, John: T57
Lan, Zhenzhong: Th30
Lapedriza, Agata: Spotlight Session 4: Deep spotlights, T91
Larochelle, Hugo: M50, Oral Session 3
Latimer, Kenneth: M42
Lattanzi, Silvio: M91
Lattimore, Tor: W3
Lauly, Stanislas: M50
Lawlor, Matthew: Oral Session 3, T74
Lazaric, Alessandro: M27, Spotlight Session 6: Learning theory spotlights, W1, W13
LeCun, Yann: Th1, Th35
Lee, Jason: M72, Spotlight Session 2: Large scale learning spotlights, T75
Lee, Seunghak: Th69
Lee, Honglak: M59, T94
Lee, Yong Jae: M64
Lefakis, Leonidas: DeMstrations
Le, Quoc V.: Oral Session 4, T98
Leigh, Angus: DeMstrations
lelarge, marc: Th50
LeMnier, Remi: T18
Leng, Chenlei: M87
Lengyel, Mate: M45, Oral Session 3, T64, T72
Leprince, Sebastien: T8
Levine, Sergey: Spotlight Session 7: Reinforcement learning spotlights, W66
Levy, Kfir: T31
Levy, Omer: W23
Lewis, Richard: T94
Liang, Jingwei: M40
Liang, Yingyu: M17, T20, Th93
Liang, Zhujin: Th4
Liang, Percy: M41, W70
Lian, Wenzhao: Th98
Liang, Feng: M84, Th77
Liao, Xuejun: T67
Ligett, Katrina: Differential Privacy and Learning: The Tools, The Results, and The Frontier
Li, Luozheng: T71
Li, Mu: Th68
Li, Fei Fei: M61
Li, Bo: W82
Lieder, Falk: Th11
Li, Hang: Th7
Lim, Shiau Hong: Spotlight Session 2: Large scale learning spotlights, T68
Lim, Cong Han: Th85
Lindenbaum, Michael: Th48
Linderman, Scott: M54
Lindsey, Robert: Th13
Lindsten, Fredrik: Th32
Lin, Shou-De: W26
Lin, Qihang: T6
Lin, Ting-Wei: W26
Lin, Tian: Oral Session 6, W2
Lin, Liang: Th4
Lipson, Hod: Oral Session 4, DeMstrations, W37
Liu, Ming-Yu: M62
Liu, Anqi: Spotlight Session 8: GP spotlights, kernel spotlights, sampling spotlights, classification spotlights, W77
Liu, Yan: Spotlight Session 5: Regression and time series spotlights, W103
Li, Ping: Oral Session 2, Spotlight Session 2: Large scale learning spotlights, T59, T65, T78
Liu, Guangcan: Spotlight Session 2: Large scale learning spotlights, T78
Liu, Han: M32, Spotlight Session 9: Graphical models spotlights, model selection, Th38, Th88, Th89, Th104
Liu, Wei: M92, Spotlight Session 2: Large scale learning spotlights, T60
Li, Wu-Jun: T56
Li, Nan: M1
Liu, Lingqiao: M65
Liu, Ji: Th105
Liu, Yuanyuan: W22
Liu, Xianghang: M95
Li, Yao: M57
Liu, Ce: W38
Liu, Zhe: M96
Liu, Qiang: M8
Li, Zhentao: Spotlight Session 9: Graphical models spotlights, model selection, Th40
Liu, Jun: M74
Livni, Roi: M47
Loh, Po-Ling: Th33
Longinotti, Luca: DeMstrations
Long, Jonathan: T92
Lopes, Miles: Spotlight Session 5: Regression and time series spotlights, W49
Love, Bradley: Spotlight Session 10: Cognitive science spotlights, Th15
Lozano, Aurelie: M98, Oral Session 10
Lucic, Mario: Spotlight Session 5: Regression and time series spotlights, W48
Luo, Haipeng: W14
Luo, Ping: Th3
Luo, Zhi-Quan: M36, Th83
Lu, Zhaosong: T6
Lu, Yichao: Th91
Lu, Huanran: M32
Lu, Zhengdong: Th7
Lyu, Michael: Oral Session 6, W2
Machado, Timothy: Spotlight Session 3: Neuroscience and Neural Coding spotlights, T50
Machens, Christian: Spotlight Session 10: Cognitive science spotlights, Th10, Th99
Macke, Jakob: M43, Spotlight Session 3: Neuroscience and Neural Coding spotlights, T49
Maddison, Chris: Oral Session 8, W59
Mahmood, A. Rupam: M29
Maillard, Odalric-Ambrym: Oral Session 7, W67
Mairal, Julien: Spotlight Session 4: Deep spotlights, T104
Majumdar, Arunava: Oral Session 3
Makin, Joseph: M9
Malek, Alan: M22
Malik, Jitendra: T85
Malinowski, Mateusz: T82
Mallat, Stephane: Th2
Mandt, Stephan: Th18
Manning, Christopher: M41, Spotlight Session 1: Optimization spotlights, T40
Mannor, Shie: Oral Session 7, W67, W87
Mann, Timothy: Oral Session 7, W67
Mao, Junhua: M67
Marchand, Mario: T36
Margolin, Adam: M15
Markovitch, Shaul: Th48
Martins, Pedro: T88
Ma, Wei Ji: Th12
Martins, Andre: T61
Ma, Tengyu: Oral Session 7, W44
Manning, Christopher: M55
Marc'Aurelio, Ranzato: Oral Session 4
Masci, Jonathan: M56
Maua, Denis: Spotlight Session 9: Graphical models spotlights, model selection, Th44
Ma, Wanli: W99
Mcallester, David: T42
McCarter, Calvin: M100
McDonald, Andrew: M77
McMahan, Brendan: M21
McWilliams, Brian: Spotlight Session 5: Regression and time series spotlights, W48
Meeds, Ted: DeMstrations
Meek, Christopher: Th45
Mehta, Nishant: Oral Session 6, W72
Meier, Franziska: Th64
Meila, Marina: Th45
Meir, Ron: M49, T48
Memisevic, Roland: Th6
Meng, Deyu: Th30
Merolla, Paul: DeMstrations
Metaxas, Dimitris: Spotlight Session 9: Graphical models spotlights, model selection, Th38
Michalski, Vincent: Th6
Minka, Tom: Oral Session 8, W59
Mirrokni, Vahab: M91
Mirza, Mehdi: T96
Mittal, Happy: W58
Mi, Yuanyuan: T71, Th97
Mizrahi, Yariv: M97
Mnih, Volodymyr: Spotlight Session 10: Cognitive science spotlights, Th9
Mobin, Shariq: M46
Modayil, Joseph: Th58
Modha, Dharmendra: DeMstrations
Mohamed, Shakir: Spotlight Session 4: Deep spotlights, T93
Mohan, Karthika: Th39
Mohapatra, Pritish: T16
Mohler, George: Th100
Mohri, Mehryar: T23, Oral Session 6, Spotlight Session 6: Learning theory spotlights, W15, W86
Montanari, Andrea: T76, Th90, Th92
Monteleoni, Claire: Climate Change: Challenges for Machine Learning
Montufar, Guido: T95
Moon, Kevin: W43
Moreno-Bote, Ruben: Spotlight Session 10: Cognitive science spotlights, Th14
Morvant, Emilie: T36
Moulines, Eric: T12
Mozer, Michael: Th13
Muandet, Krikamol: M18
Mu, Cun: Spotlight Session 2: Large scale learning spotlights, T60
Mudrakarta, Pramod Kaushik: Th49
Munos, Remi: T22, W1, W3, W4, W12
Munoz, Andres: Spotlight Session 6: Learning theory spotlights, W15
Muralidharan, Omkar: T54
Murino, Vittorio: Spotlight Session 8: GP spotlights, kernel spotlights, sampling spotlights, classification spotlights, W27
Murphy, Kevin: Tutorial Session A, Tutorial Session A, Tutorial Session A
Müller, Klaus-Robert: T62
Naghibi Mohamadpoor, Tofigh: W84
Najmi, Amir: T54
Nakajima, Shinichi: Th19
Nam, Woonhyun: T90
Narasimhan, Harikrishna: T21, W73
Natarajan, Nagarajan: Spotlight Session 8: GP spotlights, kernel spotlights, sampling spotlights, classification spotlights, W80
Needell, Deanna: T3
Negrinho, Renato: T61
Nehorai, Arye: T26
Nesterov, Yurii: Oral Session 1
Netrapalli, Praneeth: Spotlight Session 5: Regression and time series spotlights, W101
Neu, Gergely: T35, Spotlight Session 6: Learning theory spotlights, W4, W13
Neven, Hartmut: Th60
Ngo, Hung: M76
Nguyen, Huy: M19, Oral Session 7, W44
Nguyen, Long: M76
Nguyen, Trung: T28
Nguyen, Viet-An: M83
Nickel, Maximilian: Spotlight Session 7: Reinforcement learning spotlights, W25
Nie, Siqi: Spotlight Session 9: Graphical models spotlights, model selection, Th44
Nie, Feiping: Th105
Niinimäki, Teppo: Th42
Nishihara, Robert: M35
Nisnevitch, Pinhas: Th79
Nitanda, Atsushi: Th84
Niu, Gang: W81
Nock, Richard: Spotlight Session 8: GP spotlights, kernel spotlights, sampling spotlights, classification spotlights, W47
Nwogu, Ifeoma: M90
O'Connor, Luke: Th76
Oberlin, John: M63
Obozinski, Guillaume: T79
Oh, Sewoong: M4, T15, T52, W21
Oh, Sang: Th41
Oiwa, Hidekazu: Th78
Oliva, Aude: Spotlight Session 4: Deep spotlights, T91
Olsen, Peder: Th43

AUTHOR INDEX

Opper, Manfred: T48, Spotlight Session 9: Graphical models spotlights, model selection, Th61
Orabona, Francesco: W71
Orlitsky, Alon: M31
Ortiz, Luis: Spotlight Session 10: Cognitive science spotlights, Th51, Th52
Osband, Ian: M25, Spotlight Session 7: Reinforcement learning spotlights, W65
Osborne, Michael: W35
Osogami, Takayuki: M60
Otsuka, Makoto: M60
Ott, Lionel: M86
Oymak, Samet: M89
Ozair, Sherjil: T96
Ozdemir, Bahadir: W42
Pachauri, Deepti: W94
Paige, Brooks: Oral Session 8, W63
Pang, Linsey: M86
Pang, Jong-Shi: Th83
Paninski, Liam: Spotlight Session 3: Neuroscience and Neural Coding spotlights, T50
Pan, Yunpeng: M28
Pan, Xinghao: W50
Pareek, Harsh: M30
Park, Dohyung: M88
Park, Mijung: M12
Parkes, David: Oral Session 10, Th53
Parlett, Beresford: W91
Pasa, Luca: M51
Pascanu, Razvan: T95, W39
Patil, Kaustubh: Spotlight Session 10: Cognitive science spotlights, Th15
Patnaik, Kaushik: W20
Patrini, Giorgio: Spotlight Session 8: GP spotlights, kernel spotlights, sampling spotlights, classification spotlights, W47
Paulus, Romain: M55
Pearl, Judea: Spotlight Session 9: Graphical models spotlights, model selection, Th39, Th74
Peng, Peichao: Oral Session 2, T58
Pernkopf, Franz: M58
Pestilli, Franco: W76
Peters, Jan: DeMstrations
Petrik, Marek: Spotlight Session 7: Reinforcement learning spotlights, W68
Peyré, Gabriel: M40
Pfister, Beat: W84
Phillips, Jonathon: Th26
Pietquin, Olivier: Spotlight Session 7: Reinforcement learning spotlights, W69
Pillow, Jonathan: M12, M42, M43, M44, T46
Pineau, Joelle: DeMstrations
Piot, Bilal: Spotlight Session 7: Reinforcement learning spotlights, W69
Plunkett, Dillon: Th11
Pocock, Adam: Spotlight Session 9: Graphical models spotlights, model selection, Th65
Poczos, Barnabas: W74
Poldrack, Russell: Th95
Pontil, Massimiliano: M77
Porwal, Utkarsh: M76
Pouget-Abadie, Jean: T96
Pouget, Alexandre: Spotlight Session 10: Cognitive science spotlights, Th14
Prasad, Adarsh: Spotlight Session 1: Optimization spotlights, T39
Precup, Doina: Spotlight Session 7: Reinforcement learning spotlights, W40, W56
Price, Eric: Spotlight Session 1: Optimization spotlights, T10
Procaccia, Ariel: Th54, Th55
Proutiere, Alexandre: Th50
Puhrsch, Christian: Spotlight Session 10: Cognitive science spotlights, Th29
Puthiya Parambath, Shameem: W79
Putzky, Patrick: Spotlight Session 3: Neuroscience and Neural Coding spotlights, T49
Qian, Jing: W100
Qin, Danfeng: T83
Quadrianto, Novi: T27
Que, Qichao: M16
Qu, Qing: T66
Rabbany, Reihaneh: Th72
Raiko, Tapani: M57
Rajaratnam, Bala: Th41
Raj, Anant: M17
Rajkumar, Arun: M24
Rajwa, Bartek: M79, Spotlight Session 2: Large scale learning spotlights, T75
Raman, Parameswaran: T17
Ramos, Fabio: M86
Rangan, Sundeep: Spotlight Session 3: Neuroscience and Neural Coding spotlights, T70
Rangapuram, Syama Sundar: Th49
Rao, Rajesh: M53
Rasmussen, Carl: T29, Th63
Ravanbakhsh, Siamak: Th72
Ravikumar, Pradeep: M30, M37, M80, M98, T38, Spotlight Session 8: GP spotlights, kernel spotlights, sampling spotlights, classification spotlights, W26, W41, W80, Th43
Ravindran, Balaraman: M50
Raykar, Vikas: M50
Razaviyayn, Meisam: Th83
Refaat, Khaled: W17
Ren, Jimmy: W38
Resnik, Philip: M83
Restelli, Marcello: M27
Richardson, Thomas: Non-Parametric Causal Models
Richard, Emile: T79, Th90, Th92
Ridgway, James: Th77
Riedel, Sebastian: DeMstrations
Riedmiller, Martin: Th5
Rieke, Fred: M42
Rivera, Paul: Spotlight Session 8: GP spotlights, kernel spotlights, sampling spotlights, classification spotlights, W47
Roberts, Stephen: W35
Rocktäschel, Tim: DeMstrations
Roig, Gemma: M66
Rokem, Ariel: W76
Rosman, Guy: Th47
Rostamizadeh, Afshin: Th57
Roth, Volker: W16
Rousu, Juho: T36
Roy, Daniel: M13
Roy, Nicholas: T25
Rubinstein-Salzedo, SiM: W61
Rudin, Cynthia: M93
Ruozzi, Nicholas: Spotlight Session 9: Graphical models spotlights, model selection, Th73
Rus, Daniela: Th47
Russell, Stuart: Th11
Russo, Dan: W5
Ruttor, Andreas: Spotlight Session 9: Graphical models spotlights, model selection, Th61
Ré, Christopher: M76
Saade, Alaa: W93
Sabato, Sivan: W12
Saberian, Mohammad: W85
Sabes, Philip: M9
Sadowski, Peter: Spotlight Session 4: Deep spotlights, T103
Saenko, Kate: T89
Saha, Amrita: M50
Salakhutdinov, Ruslan: Oral Session 4, DeMstrations, T101, T102
Salas, Marta: T85
Salazar, Esther: M81
Saligrama, Venkatesh: W100
SalM, Joseph: T12
San Biagio, Marco: Spotlight Session 8: GP spotlights, kernel spotlights, sampling spotlights, classification spotlights, W27
Sandholm, Tuomas: Th55
Sanghavi, Sujay: M88, Spotlight Session 5: Regression and time series spotlights, W101
Sani, Amir: Spotlight Session 6: Learning theory spotlights, W13
Sargur, Gautam: W94
Sarkhel, Somdeb: W60
Sato, Issei: Th19
Savin, Cristina: Spotlight Session 3: Neuroscience and Neural Coding spotlights, T47
Sawada, Hiroshi: W28
Saxena, Shreya: Th96
Scaman, Kevin: T18
Schaal, Stefan: Th64
Schaich Borg, Jana: T69
Schapire, Robert: W14
Schmidhuber, Jürgen: M56
Schmid, Cordelia: Spotlight Session 4: Deep spotlights, T104
Schneider, Jeff: Th106
Schober, Michael: Oral Session 8, W32
Schrater, Paul: M3
Schrauwen, Benjamin: M6
Schuurmans, Dale: W55
Schwing, Alex: Th16, Th23
Schölkopf, Bernhard: M18
Schön, Thomas: Th32
Scott, Clayton: M7
Sculley, D.: Oral Session 2
Sedghi, Hanie: Th87
Seiler, Christof: W61
Sejdinovic, Dino: Oral Session 8, W29
Semedo, Joao: Th99
Servedio, Rocco: M5
Shah, Julie: M93, Th56
Sha, Fei: T84
Shah, Nisarg: Th55
Shah, Devavrat: M4, M99, Spotlight Session 2: Large scale learning spotlights, T11, Th31
Shakhnarovich, Greg: T42
Shalev-Shwartz, Shai: M47
Shamir, Ohad: M47, T43
Shamma, Shihab: W97
Shang, Wenling: M59
Shang, Fanhua: W22
Shan, Shiguang: Th30, Th94
Shanmugam, Karthikeyan: Oral Session 5, W10, W41
Sharma, Abhishek: M62
Sharmanska, Viktoriia: T27
Shawe-Taylor, John: T36
Sheffet, Or: Spotlight Session 6: Learning theory spotlights, W45
Sheldon, Daniel: W52
Shen, Chunhua: M65
Shen, Jie: T65
Shin, Donghyuk: W91
Shivanna, Rakesh: T45
Shrivastava, Anshumali: Oral Session 2, T59
Sigal, Leonid: M69
Silva, Ricardo: Th75
Silver, David: W64
Simard, Patrice: DeMstrations
SiM, Jonathan: W97
SiMyan, Karen: Spotlight Session 4: Deep spotlights, T99
Singh, Satinder: T94
Singh, Sameer: DeMstrations
Singh, Vikas: W94
Singh, Shashank: W74
Singla, Parag: W58, W60
Sivakumar, Vidyashankar: W7
Si, Si: W30, W91
Skeel, Robert: Th60
Smaragdis, Paris: W90
Smith, Virginia: Th70
Smith, Noah: Oral Session 4, T41
Smola, Alex: W88, Th68
Soare, Marta: W1
Socher, Richard: M55
Soh, De Wen: W31
Sohn, Kihyuk: M59
Sommer, Fritz: M46
Song, Hyun Oh: M64
Song, Sen: M48
Song, Le: M17, T19, T20, W20
Song, Shuang: T51
Son, Jeany: Th22
Soriano Marcolino, Leandro: Th55
Soudry, Daniel: M49
Sperduti, Alessandro: M51
Springenberg, Jost Tobias: Th5
Sra, Suvrit: W99
Srebro, Nati: T3
Srikumar, Vivek: Spotlight Session 1: Optimization spotlights, T40
Sriperumbudur, Bharath: M18
Srivastava, Nitish: Oral Session 4, DeMstrations, T101
Srivastava, Nisheeth: M3
Stachenfeld, Kimberly: Spotlight Session 3: Neuroscience and Neural Coding spotlights, T73
Stamos, Dimitris: M77
Steele, Guy: Spotlight Session 9: Graphical models spotlights, model selection, Th65
Steinberg, Daniel: Spotlight Session 8: GP spotlights, kernel spotlights, sampling spotlights, classification spotlights, W33
Stimberg, Florian: Spotlight Session 9: Graphical models spotlights, model selection, Th61
Stober, Sebastian: M52
Stock, Christopher: M54
Stollenga, Marijn: M56
Streeter, Matthew: M21
Subakan, Cem: W90
Subrahmanya, Niranjan: W19
Subramanian, Dharmashankar: Spotlight Session 7: Reinforcement learning spotlights, W68

AUTHOR INDEX

Subramanian, Lakshminarayanan: T53
Sugiyama, Masashi: W81, Th19, Th102
Suh, Jina: DeMstrations
Sun, Xu: Th24
Sun, Xiaorui: M5
Su, Hongyu: T36
Sun, Jiangwen: Th101
Sun, Yi: T87
Sun, Ju: T66
Suresh, Ananda Theertha: M31
Susemihl, Alex: T48
Sutskever, Ilya: Oral Session 4, T98
Sutton, Richard: M29, Th58
Sutton, Charles: W102, Th59
Su, Weijie: Spotlight Session 1: Optimization spotlights, T4
Sutcliffe, Andrew: DeMstrations
Syed, Umar: W86, Th57
Szepesvari, Csaba: Th58
Szörényi, Balázs: T22
Taba, Brian: DeMstrations
Takac, Martin: Th70
Tambe, Milind: Th55
Tandon, Rashish: W41
Tang, Yichuan: Oral Session 4, DeMstrations, T101
Tang, Xiaoou: M92, T87, Th3
Tarlow, Daniel: Oral Session 8, W59, Th66
Taskar, Ben: Th25
Tassarotti, Joseph: Spotlight Session 9: Graphical models spotlights, model selection, Th65
Tatikonda, Sekhar: W31
Taylor, Jonathan: M72
Teh, Yee Whye: M13, Oral Session 8, W63, Th67
Tekin, Cem: T34
Telgarsky, Matus: T57
Terhorst, Jonathan: Th70
Tewari, Ambuj: M73
Theodorou, Evangelos: M28
Titsias, Michalis: Spotlight Session 8: GP spotlights, kernel spotlights, sampling spotlights, classification spotlights, W62
Tomioka, Ryota: Th102
Tompson, Jonathan: Th35
Tootoonian, Sina: T64
Torralba, Antonio: Spotlight Session 4: Deep spotlights, T91
Traa, Johannes: W90
Tran-Dinh, Quoc: W53
Treister, Eran: W57
Tresp, Volker: Spotlight Session 7: Reinforcement learning spotlights, W25
Tristan, Jean-Baptiste: Spotlight Session 9: Graphical models spotlights, model selection, Th65
Trivedi, Shubhendu: T42
Tropp, Joel: M39
Tschantz, Michael: W83
Tschiatschek, Sebastian: T86
Tung, Hsiao-Yu: W88
Turek, Javier: W57
Turner, Ryan: Th34
Turner, Richard: Spotlight Session 8: GP spotlights, kernel spotlights, sampling spotlights, classification spotlights, W34
Tuzel, Oncel: M62
Tyagi, Hemant: W8
Tygar, J. D.: W83
Tzeng, Eric: T89
Ulrich, Kyle: Th98
U N, Niranjan: Spotlight Session 5: Regression and time series spotlights, W101
Urtasun, Raquel: Th16, Th23
Usunier, Nicolas: W79
Vaish, Rohit: W73
Valera, Isabel: M10, T19
Valko, Michal: T33, T35, W4
Vanchinathan, Hastagiri: T24
van den Hengel, Anton: M65
van den Oord, Aaron: M6
Vandermeulen, Robert: M7
Van Der Schaar, Mihaela: T34
van der Wilk, Mark: Th63
van Erven, Tim: M20
van Hasselt, Hado: M29
van Hoof, Herke: DeMstrations
van Merriënboer, Bart: DeMstrations
Van Roy, Benjamin: M25, Spotlight Session 7: Reinforcement learning spotlights, W5, W65
Vasconcelos, Nuno: W85
Vaughan, Jennifer Wortman: Oral Session 9
Vayatis, Nicolas: T18
Vega-Brown, William: T25
Veiga, Filipe: DeMstrations
Venkataraman, Ashwin: T53
Venugopal, Deepak: W24, W60
Ver Steeg, Greg: Th46
Vertechi, Pietro: Spotlight Session 10: Cognitive science spotlights, Th10
Vert, Jean-Philippe: T79
Verwey, Johan: DeMstrations
Vetta, Adrian: Spotlight Session 9: Graphical models spotlights, model selection, Th40
Vijayakumar, SeTh: Th12
Vijayaraghavan, Aravindan: Spotlight Session 6: Learning theory spotlights, W45
Vinyals, Oriol: Oral Session 4, T98
Vishwanathan, S.: T17, T55
Viswanath, Pramod: T52
Vojnovic, Milan: M2
Volkov, Mikhail: Th47
Vorobeychik, Yevgeniy: W82
Vul, Ed: M3
Wager, Stefan: T54, W70
Wang, Jie: M74, Spotlight Session 1: Optimization spotlights, T9
Wang, Sida: M41, W70
Wang, Dahui: T71
Wang, Zhangyang: M84
Wang, Chao: M65
Wang, Lie: Th104
Wang, Xiaoshi: T94
Wang, Shenlong: Th23
Wang, Zhenghao: Spotlight Session 2: Large scale learning spotlights, T5
Wang, Zhaoran: M32, Th89
Wang, Huahua: M36, Th86
Wang, Lei: M65
Wang, Xiaolong: Th4
Wang, Xuezhi: Th106
Wang, Qian: M48
Wang, Yining: W89
Wang, Xin: Th101
Wang, Xiaogang: T87, Th3
Wang, Xiangyu: Oral Session 2, T58
Wang, Yiming: Th88
Wang, Yusu: M16
Warde-Farley, David: T96
Ward, Rachel: T3
Warmuth, Manfred: Spotlight Session 6: Learning theory spotlights, W46
Watanabe, Kazuho: Th19
Wei, Haochen: T86
Weissbrod, Omer: Th48
Weller, Adrian: Oral Session 9, Th36
Welling, Max: Spotlight Session 4: Deep spotlights, DeMstrations, T93
Whiteson, Daniel: Spotlight Session 4: Deep spotlights, T103
Wibisono, Andre: Th33
Williamson, Sinead: M11
Williamson, Robert: Oral Session 6, W72
Wilson, Andrew: T26
Wimalawarne, Kishan: Th102
Winn, John: Th66
Winther, Ole: W6
Wong, K. Y. Michael: Th97
Wonka, Peter: M74
Wood, Frank: Oral Session 8, W63
Woodruff, David: M19, T77, Th93
Wright, John: T66
Wright, Stephen: Th85
wu, xiaojian: W52
Wu, Yue: T30
Wu, Anqi: M12
Wu, Si: T71, Th97
Xia, Lirong: Oral Session 10, Th53
Xiao, Lin: T6
Xiao, Jianxiong: Spotlight Session 4: Deep spotlights, T91
Xie, Cong: T56
Xie, Bo: M17
Xing, Eric: M11, Th69
Xin, Yu: T13
Xiong, Yuanjun: M92
Xu, Huan: M87, Spotlight Session 2: Large scale learning spotlights, T65, T68, W87
Xu, Bing: T96
Xu, Li: W38
Xu, Minjie: Th67
Xu, Jason: Th21
Xu, Jiaming: T15
Yang, Yingzhen: M84
Yang, Tianbao: Th81
Yang, Jianchao: Th27
Yang, Scott: T23
Yang, Eunho: M98
Yang, Jianbo: T67
Yan, Shuicheng: M84, M87, W87
Yan, Ling: T56
yao, hengshuai: Th58
Yates, Jacob: M44
Yau, Christopher: Spotlight Session 8: GP spotlights, kernel spotlights, sampling spotlights, classification spotlights, W62
Ye, Jieping: M74, Spotlight Session 1: Optimization spotlights, T9
Yen, En-Hsu: M37, T38, W26
Yerebakan, Halid: M79
Yoshikawa, Yuya: W28
Yosinski, Jason: Oral Session 4, DeMstrations, W37
Yu, Aron: M68
Yu, Qi (Rose): Spotlight Session 5: Regression and time series spotlights, W103
Yu, Byron: Oral Session 5, W9, Th99
Yu, Adams Wei: W99
Yuan, Xin: W78
Yu, Kai: Th68
Yu, Shoou-I: Th30
Yu, Shipeng: Th101
Yuille, Alan: M67, M71
Yun, Se-Young: Th50
Yu, Yaoliang: W99
Yu, Mo: Th88
Yun, Hyokun: T17
Zandvakili, Amin: Th99
Zaremba, Wojciech: Spotlight Session 4: Deep spotlights, T100, Th1
Zdeborova, Lenka: W93
Zeevi, Assaf: T32
Zemel, Richard: T102
Zha, Hongyuan: T19
Zhang, Ce: M76
Zhang, Jiaxing: M48
Zhang, Zheng: M48
Zhang, Liliang: Th4
Zhang, Ning: T92
Zhang, Lei: Th71
Zhang, Jian: Th16
Zhang, Bo: Th67
Zhang, Yichuan: Th59
Zhang, Chongjie: Th56
Zhang, Xinhua: M85, W55
Zhang, Yuchen: Spotlight Session 5: Regression and time series spotlights, W95
Zhang, Haichao: Th27
Zhang, Zhihua: T56
Zhang, Chicheng: Spotlight Session 6: Learning theory spotlights, W11
Zhao, Deli: M92
Zhao, Tianqi: Spotlight Session 9: Graphical models spotlights, model selection, Th38
Zhao, Huasha: DeMstrations
Zhao, Tuo: Th88, Th104
Zheng, Charles: W76
Zheng, Jingjing: Th26
Zheng, Xun: Th69
Zhong, Mingjun: W102
Zhong, Kai: T38
Zhou, Tianyi: Th37
Zhou, Jingren: Spotlight Session 2: Large scale learning spotlights, T5
Zhou, Dengyong: Spotlight Session 5: Regression and time series spotlights, W95
Zhou, Jiayu: M74
Zhou, Mingyuan: M78
Zhou, Bolei: Spotlight Session 4: Deep spotlights, T91
Zhou, Zhi-Hua: M1
Zhou, Yingbo: M76
Zhu, Jun: M85, W89, Th67
Zhu, Yuancheng: M34
Zhu, Changbo: M87
Zhu, Xiaojin: Spotlight Session 10: Cognitive science spotlights, Th15
Zhu, Jun: M67
Zhu, Zhenyao: Th3
Ziebart, Brian: Spotlight Session 8: GP spotlights, kernel spotlights, sampling spotlights, classification spotlights, W77
Zilberstein, Shlomo: W52
Zisserman, Andrew: Spotlight Session 4: Deep spotlights, T99
Zoran, Daniel: Th28
Zucker, Steven: Oral Session 3, T74
Zuo, Wangmeng: Th4, Th71
Zöhrer, Matthias: M58