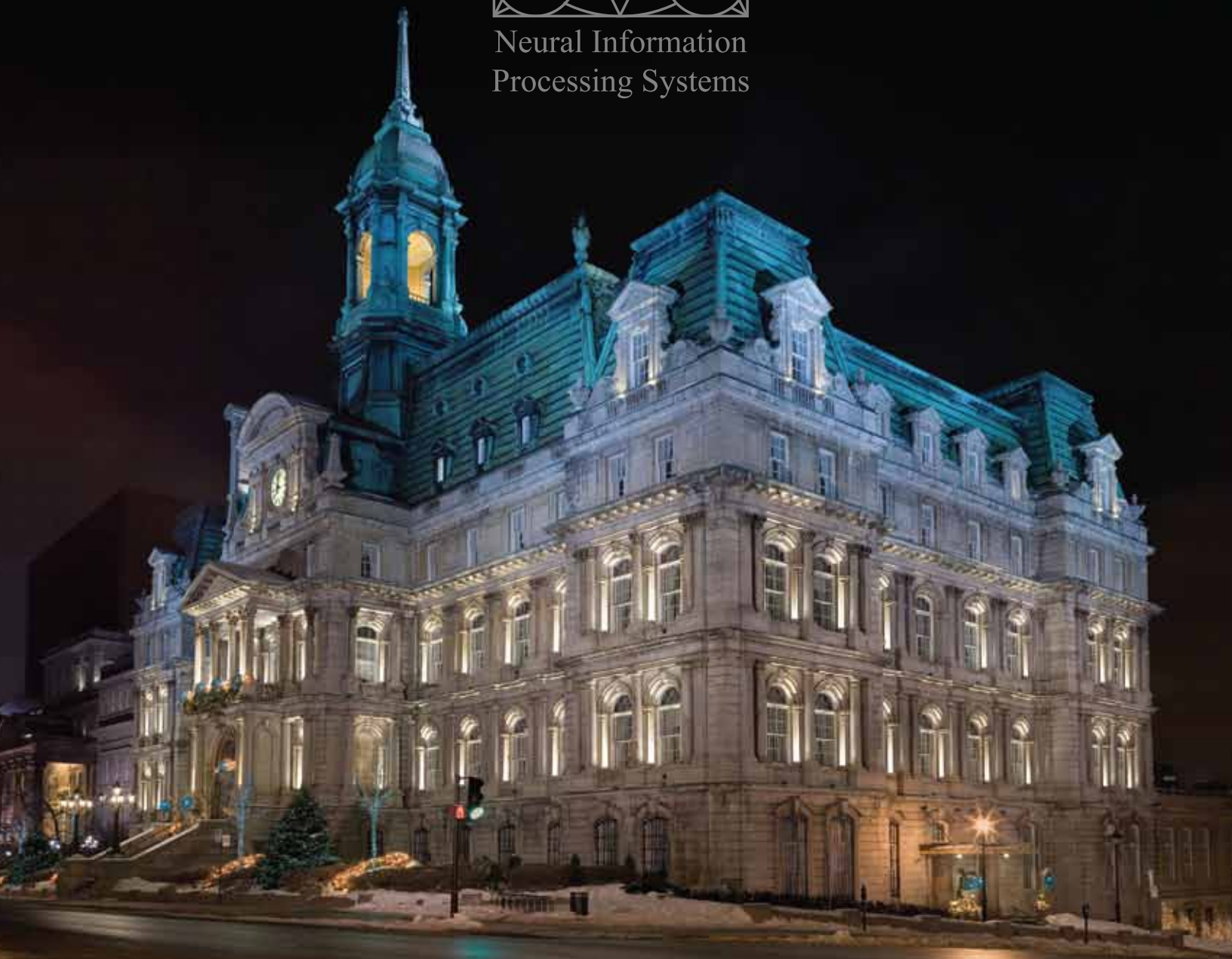
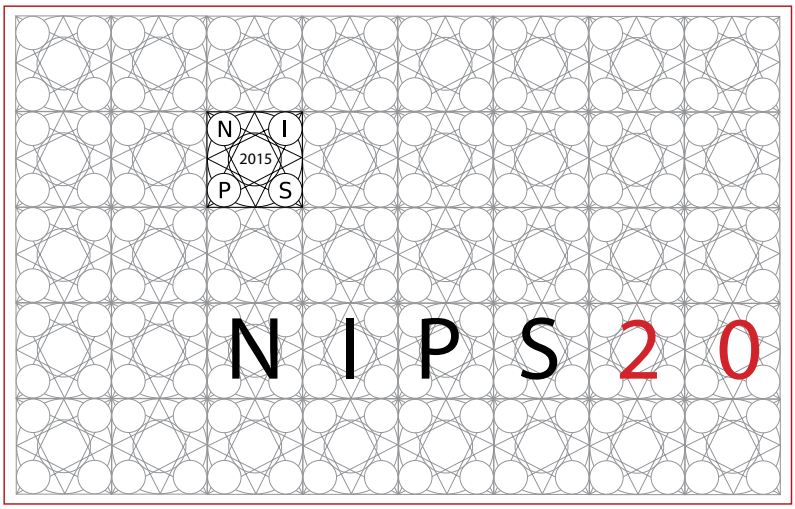
NIPS 2015

CONFERENCE BOOK

MONTREAL

Neural Information
Processing Systems





NIPS 2015

Abstracts of Papers

LOCATION
Palais des Congrès de Montréal
Convention and Exhibition Center,
Montreal, Quebec, Canada

TUTORIALS
December 7, 2015

CONFERENCE SESSIONS
December 8 - 10, 2015

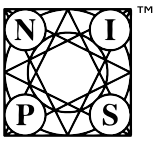
SYMPOSIA
December 10, 2015

WORKSHOPS
December 11-12, 2015

Sponsored by the Neural Information Processing System Foundation, Inc

The technical program includes 6 invited talks and 403 accepted papers, selected from a total of 1838 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 28," edited by edited by Daniel D. Lee, Masashi Sugiyama, Corinna Cortes, Neil Lawrence and Roman Garnett.



Neural Information
Processing Systems
Foundation

TABLE OF CONTENTS

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

NIPS appreciates the taping of tutorials, conference, symposia & select workshops.

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
Mary Ellen Perry - Executive Director
Mike Perry - Administrator
Susan Perry - Volunteer Manager
Jen Perry - Administrator
Lee Maree Fazzio - Administrator

ORGANIZING COMMITTEE

General Chairs: Corinna Cortes (Google Research)
Neil D Lawrence (University of Sheffield)
Program Chairs: Daniel D Lee (University of Pennsylvania)
Masashi Sugiyama (The University of Tokyo)
Tutorials Chair: Ralf Herbrich (Amazon Development Center Germany GmbH)
Workshop Chairs: Borja Balle (McGill University)
Marco Cuturi (Kyoto University)
Demonstration Chair: Marc'Aurelio Ranzato (Facebook)
Publications Chair and Electronic Proceedings Chair
Roman Garnett (Washington Univ. St. Louis)
Program Managers: Pedro A Ortega (Univ. of Pennsylvania)
Yung-Kyun Noh (Seoul National Univ.)

NIPS FOUNDATION OFFICERS & BOARD MEMBERS

PRESIDENT
Terrence Sejnowski, The Salk Institute

TREASURER
Marian Stewart Bartlett, UC San Diego

SECRETARY
Michael Mozer, University of Colorado, Boulder

EXECUTIVE DIRECTOR
Mary Ellen Perry, The Salk Institute

EXECUTIVE BOARD
Max Welling, University of Amsterdam
Zoubin Ghahramani, University of Cambridge
Peter Bartlett, Queensland University and UC Berkeley
Léon Bottou, Microsoft Research
Chris J.C. Burges, Microsoft Research
Fernando Pereira, Google Research
Rich Zemel, University of Toronto

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, Itnl Computer Science Institute, Berkeley & 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
John Shawe-Taylor, University College London
Sara A. Solla, Northwestern University Medical School
Yair Weiss, Hebrew University of Jerusalem
Chris Williams, University of Edinburgh

PROGRAM COMMITTEE

Maria-Florina Balcan (Carnegie Mellon University.)
David Balduzzi (Wellington)
Samy Bengio (Google Research)
Alina Beygelzimer (Yahoo!)
Daniel Braun (MPI Tuebingen)
Emma Brunskill (CMU)
Gal Chechik (Bar Ilan U. / Google)
Kyunghyun Cho (U. Montreal)
Seungjin Choi (POSTECH)
Aaron Courville (U. Montreal)
Marco Cuturi (Kyoto U.)
Florence d'Alche-Buc (Telecom ParisTech)
Marc Deisenroth (Imperial College London)
Inderjit Dhillon (U. of Texas at Austin)
Francesco Dinuzzo (IBM Research)
Mohammad Emtiyaz Khan (EPFL)
Aldo Faisal (Imperial College London)
Emily Fox (U. Washington)
Kenji Fukumizu (Inst. of Statistical Mathematics)
Thomas Gaertner (Fraunhofer IAIS)
Amir Globerson (Hebrew U. of Jerusalem)
Ian Goodfellow (Google)
Moritz Grosse-Wentrup (MPI for Intelligent Systems)
Bohyung Han (Postech)
Elad Hazan (Princeton U.)
Xiaofei He (Zhejiang U.)
Tomoharu Iwata (Nippon Telegraph and Telephone Corporation)
Amos J. Storkey (U. of Edinburgh)
Kee-Eung Kim (KAIST)
Samory Kpotufe (Princeton U.)
Andreas Krause (ETH)
James Kwok (Hong Kong U. of Science and Tech)
Christoph Lampert (Inst. of Science & Tech Austria)
John Langford (Microsoft)
Hugo Larochelle (U. Sherbrooke)
Pavel Laskov (U. of Tuebingen)
Svetlana Lazebnik (U. of Illinois at Urbana-Champaign)
Honglak Lee (U. Michigan)
Deng Li (MSR)
Hang Li (Huawei Technologies)
Chih-Jen Lin (National Taiwan U.)
Hsuan-Tien Lin (National Taiwan U.)
Yuanqing Lin (NEC Labs)
Zhouchen Lin (Peking U.)
Tie-Yan Liu (MSR Asia)
Aurelie Lozano (IBM Research)
David McAllester (Toyota Technological Inst. at Chicago)
Marina Meila (U. of Washington)
Shakir Mohamed (Google)
Claire Monteleoni (George Washington U.)
Greg Mori (Simon Fraser U.)
Remi Munos (INRIA Lille)
Shinichi Nakajima (Technische U. Berlin)
Sebastian Nowozin (MSR Cambridge)
Peter Orbanz (Columbia U.)
Sinno Pan (NTU Singapore)
Barnabas Poczos (Carnegie Mellon U.)
Massimiliano Pontil (U. Collage London)
Novi Quadrianto (U. of Sussex)
Gunnar Raetsch (Memorial Sloan-Kettering Cancer Center)
Alain Rakotomamonjy (U. of Rouen)
Cynthia Rudin (MIT)
Ruslan Salakhutdinov (U. of Toronto)
Issei Sato (U. of Tokyo)
Clayton Scott (U. of Michigan)
Matthias Seeger (Amazon Development Center Germany, Berlin)
Dino Sejdinovic (U. of Oxford)
Yevgeny Seldin (U. of Copenhagen)
Jianbo Shi (U. of Pennsylvania)
Aarti Singh (CMU)
Le Song (Georgia Inst. of Technology)
Cheng Soon Ong (NICTA)
Nati Srebro (Toyota Technological Inst. at Chicago)
Bharath Sriperumbudur (Pennsylvania State U.)
Alan Stocker (U. of Pennsylvania)
Wee Sun Lee (National University of Singapore)
Ilya Sutskever (Google)
Taiji Suzuki (Tokyo Inst. of Technology)
Csaba Szepesvari (U. of Alberta)
Ichiro Takeuchi (Nagoya Inst. of Technology)
Toshiyuki Tanaka (Kyoto U.)
Ryota Tomioka (Toyota Technological Inst. at Chicago)
Ivor Tsang (U. of Technology, Sydney)
Koji Tsuda (U. of Tokyo)
Naonori Ueda (Nippon Telegraph and Telephone Corporation)
Laurens van der Maaten (U.of Delft)
Rene Vidal (Johns Hopkins U.)
S.V.N. Vishwanathan (UC Santa Cruz)
Liwei Wang (Peking U.)
Sinead Williamson (U. of Texas at Austin)
Eric Xing (Carnegie Mellon U.)
Huan Xu (National U. of Singapore)
Eunho Yang (IBM Thomas J. Watson Research Center)
Jieping Ye (U. of Michigan)
Byron Yu (Carnegie Mellon U.)
Jingyi Yu (U. of Delaware)
Xinhua Zhang (National ICT Australia)
Dengyong Zhou (MSR)
Zhi-Hua Zhou (Nanjing U.)
Jun Zhu (Tsinghua U.)
Xiaojin Zhu (U. of Wisconsin-Madison)

SPONSORS

NIPS gratefully acknowledges the generosity of those individuals and organizations who have provided financial support for the NIPS 2015 conference. Their financial support enables us to sponsor student travel and participation, the outstanding paper awards, and the volunteers who assist during NIPS.

PLATINUM SPONSORS

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.

Microsoft

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.

Alibaba Group 阿里巴巴集团

ALIBABA GROUP'S MISSION IS TO MAKE IT EASY TO DO BUSINESS ANYWHERE. We operate leading online and mobile marketplaces in retail and wholesale trade, as well as cloud computing and other services. We provide technology and services to enable consumers, merchants, and other participants to conduct commerce in our ecosystem.

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.

amazon.com

Amazon.com strives to be Earth's most customer-centric company where people can find and discover virtually anything they want to buy online. Amazon's evolution from Web site to e-commerce partner to development platform is driven by the spirit of innovation that is part of the company's DNA. The world's brightest technology minds come to Amazon.com to research and develop technology that improves the lives of shoppers, sellers and developers.

Apple

Apple revolutionized personal technology with the introduction of the Macintosh in 1984. Today, Apple leads the world in innovation with iPhone, iPad, the Mac and Apple Watch. Apple's three software platforms — iOS, OS X and watchOS — provide seamless experiences across all Apple devices and empower people with breakthrough services including the App Store, Apple Music, Apple Pay and iCloud. Apple's 100,000 employees are dedicated to making the best products on earth, and to leaving the world better than we found it. To learn about job opportunities at Apple, visit Jobs at Apple.

Baidu Research

Baidu Research 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. Baidu Research represents three labs: the Silicon Valley AI Lab, the Institute of Deep Learning and the Big Data Lab.

CITADEL

We're Citadel, a worldwide leader in finance that uses next-generation technology and alpha-driven strategies to transform the global economy. We tackle some of the toughest problems in the industry by pushing ourselves to be the best again and again. It's demanding work for the brightest minds, but we wouldn't have it any other way. Here, great ideas can come from anyone. Everyone. You.

twitter

Twitter is the heartbeat of the world: it's the only platform that offers insight into everything that is happening in the last 15 minutes from the macro to very local. Every event happens on Twitter, and our team Cortex is responsible for surfacing the most relevant tweet, photos, videos and live periscopes from the massive and ever changing firehose of content.

SPONSORS

Founded in 2004, Facebook's mission is to give people the power to share and make the world more open and connected. People use Facebook to stay connected with friends and family, to discover what's going on in the world, and to share and express what matters to them.

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. As part of IBM's Cognitive Business initiative, IBM Research is continually augmenting Watson's cognitive capabilities, thereby enabling real-world transformations of many domains of commerce, ranging from healthcare to law, finance, retail and media. 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.

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.

Co-founded in 2007 by two leading scientists, The Voleon Group designs, develops, and applies cutting-edge technologies to investment management. We are a family of companies dedicated to solving large-scale scientific problems with statistical machine learning techniques.

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.

Bloomberg technology helps drive the world's financial markets. We provide communications platforms, data, analytics, trading platforms, news and information for the world's leading financial market participants. We deliver through our unrivaled software, digital platforms, mobile applications and state of the art hardware developed by Bloomberg technologists for Bloomberg customers.

Adobe is the global leader in digital marketing and digital media solutions. Our tools and services allow our customers to create groundbreaking digital content, deploy it across media and devices, measure and optimize it over time and achieve greater business success. We help our customers make, manage, measure and monetize their content across every channel and screen.

AdRoll is a high-performance advertising platform that helps businesses of all sizes execute on marketing campaigns around the world. We are the world's most widely used "real-time bidding" retargeting platform, with over 20,000 customers in more than 100 countries. What is real-time bidding? It means we help companies intelligently bid on advertising spaces billions of times a day in fractions of seconds. Our proprietary large-scale machine learning infrastructure ingests billions of events with billions of sparse features 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.

SPONSORS

Analog Devices | Lyric Labs

Analog Devices Lyric Labs is focused on interfacing with the real world, developing algorithms to process real data, and delivering unique processing solutions in hardware built to run machine learning, statistical inference, and deep learning algorithms, as well as cloud-hosted implementations.

CenturyLink™ Business

CenturyLink Cognilytics is a global leader in big data, decision sciences, and data visualization solutions including predictive asset maintenance, cybersecurity, risk analytics, demand forecasting, sales & marketing analytics, customer churn analytics, pricing and promotion analytics, and supply chain optimization. Our team focuses on leveraging advanced machine learning techniques to solve real world problems.

criteo

About Criteo At Criteo, personalized digital performance advertising is what we do. And it's what we do best. Known for the quality of our digital advertising technologies, developed by the 250+ world-class engineers working for our R&D in Paris, Grenoble and Palo Alto. Our technology takes an algorithmic approach to determining what user we show an ad to, when, and for what products.

Cubist Systematic Strategies

Cubist Systematic Strategies is one of the world's premier investment firms. The firm deploys systematic, computer-driven trading strategies across multiple liquid asset classes, including equities, futures and foreign exchange. The core of our effort is rigorous research into a wide range of market anomalies, fueled by our unparalleled access to a wide range of publicly available data sources.

deep genomics

Deep Genomics is inventing machine learning systems to discover how cells read and interpret the genome, taking computational genome diagnostics to the next level. We combine world-leading expertise in deep learning, genome biology, and precision medicine.

D E Shaw & Co

The D. E. Shaw group is a global investment and technology development firm with more than $37 billion in aggregate investment capital as of July 1, 2015 and offices in North America, Europe, and Asia. Since its organization in 1988, the firm has earned an international reputation for financial innovation, technological leadership, and an extraordinarily distinguished staff.

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.

ebay

eBay is The World's Online Marketplace, enabling trade on a local, national and international basis. With a diverse and passionate community of individuals and small businesses, eBay offers an online platform where millions of items are traded each day.

imagia

Imagia's mission is to layer actionable information atop medical imaging data. Together with our research and industry partners, we will integrate deep learning to oncology practices and products for better productivity and patient outcomes. We are poised to make clinically significant differences across the cancer care continuum in detection, diagnosis, image-guidance and treatment monitoring.

MaluubA

Maluuba is a well-funded startup connecting 50+ million people to technology through A.I. conversations. We are driven by the single purpose of building great products powered by natural language processing. Our goal? To help machines understand the world via human language. Our Montreal lab is a fast, nimble group of scientists tackling cutting-edge problems in areas of deep learning for NLP.

TWO SIGMA

We find connections in all the world's data. These insights guide investment strategies that support retirements, fund research, advance education and benefit philanthropic initiatives. So, if harnessing the power of math, data and technology to help people is something that excites you, let's connect

SPONSORS

At AHL we mix mathematics, computer science, statistics and engineering with terabytes of data to understand and predict financial markets. Led by research and technology, we build models and write programs that invest billions of dollars every day. We are a small flat-structured company that seeks the best.

Tackling Today's Biggest Challenges. 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.

Panasonic is focusing on bringing new solutions to an ever-changing environment, full of cutting edge technologies. We apply Deep Learning as a tool to improve the real-life situations of today and the evolving situations of tomorrow. Deep Learning is just one of the key technologies we employ to understand more about each other and how to navigate through our lives: safely, honestly and happily.

We are a quantitative research and trading group with a strong track record of hiring, challenging, and retaining scientists interested in conducting research where the lab is the financial markets. Using state-of-the-art technology, we develop and deploy model-driven trading strategies. We value depth and expertise, encourage intellectual curiosity and seek constant innovation. We are a passionate community of tech geeks, science fiction fans, LOTR aficionados, musicians, athletes, and foodies. Come talk to us!

Sony's US Research Center (USRC) is engaged in R&D of emerging technologies, and explorations in basic science for next-gen products and new businesses. By utilizing complex engineering and mathematical principles in computer vision, graphics and deep learning, our researchers create core technologies, algorithms and architectures for advanced multimedia & medical imaging platforms and products.

We are the UK's national institute for data science. Our mission is to undertake data science research at the intersection of computer science, maths, statistics and systems engineering; provide technically informed advice to policy makers on the wider implications of algorithms; enable researchers from industry and academia to work together to undertake research with practical applications.

Toyota Research Institute of North America (TRI-NA) (est. 2008), located in Ann Arbor, MI, is one of Toyota's global advanced R&D sites. Research includes the development of intelligent driving systems for advanced safety and automated driving. Our team develops machine learning, computer vision, and robotics technologies applied to advanced sensor processing and decision making systems.

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.

Vatic Labs was built by scientists, engineers, and traders, working in collaboration to solve complex problems that will define the future of algorithmic trading. Through the application of statistical methods and cutting edge technology, our team is transforming novel trading ideas into strategies that boost market efficiency and provide sustainable liquidity for all market participants.

Winton is a new kind of investment management business, driven by a scientific approach to trading financial markets. From humble beginnings, we have grown into one of the largest investment management businesses in Europe with an AUM exceeding $30 billion. We pride ourselves on providing a collaborative environment where original thought and stimulating research can flourish. Our continued success is founded upon the 350 outstanding researchers, engineers and finance professionals working across our offices in Europe, US, Australia and Asia.

Yahoo Labs powers Yahoo's most critical products with innovative science. As Yahoo's research incubator for bold new ideas and laboratory for rigorous experimentation, Yahoo Labs applies its scientific findings in powering products for users and enhancing value for partners and advertisers. The Labs' forward-looking innovation also helps position Yahoo as an industry and scientific leader.

Netflix is the world's leading Internet television network with over 70 million members in over 60 countries enjoying more than 100 million hours of TV shows and movies per day, including original series, documentaries and films. Machine learning drives many aspects of the Netflix product and is core to Netflix's mission to provide the best content tailored to each individual's tastes.

PROGRAM HIGHLIGHTS

SUNDAY DEC 6TH

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

MONDAY DEC 7TH

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

8:00 am – 9:30 am
Breakfast, Level 2, 220A

9:30 am – 11:30 am
Deep Learning
Geoffrey E Hinton, Yoshua Bengio, Yann LeCun

Large-Scale Distributed Systems for Training Neural Networks
Chandra Chekuri

10:45 – 11:15 am - Coffee Break

12:05 – 1:00 pm - Lunch Break

1:00 – 3:00 pm
Monte Carlo Inference Methods
Iain Murray

Probabilistic Programming
Frank Wood

3:00 – 3:30 am - Coffee Break

3:30 – 5:30 pm
Introduction to Reinforcement Learning with Function Approximation
Rich S Sutton

High-Performance Hardware for Machine Learning
Bill Dally

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

7:00 – 11:59 pm
Poster Session

TUESDAY DEC 8TH

7:30 am – 9:00 am
Breakfast, Level 2, 220A

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

9:00 – 9:50 am
INVITED TALK: Probabilistic Machine Learning: Foundations and Frontiers
Zoubin Ghahramani

9:50 – 10:10 am
Oral Session 1
Randomized Block Krylov Methods for Stronger and Faster Approximate Singular Value Decomposition
Cameron Musco · Christopher Musco

10:10 –10:40 am
Spotlight Session 1: Room 210A

10:40 – 11:10 am - Coffee Break

11:10 –11:50 am
Oral Session 2
Sampling from Probabilistic Submodular Models
Alkis Gotovos · Hamed Hassani · Andreas Krause

Solving Random Quadratic Systems of Equations Is Nearly as Easy as Solving Linear Systems
Yuxin Chen · Emmanuel Candes

11:50 – 12:00 pm
Spotlight Session 2:

12:00 – 2:00 pm - Lunch Break

2:00 – 2:50 pm
INVITED TALK: Incremental Methods for Additive Cost Convex Optimization
Asuman Ozdaglar

2:50 – 3:30 pm
Oral Session 3
Probabilistic Line Searches for Stochastic Optimization
Maren Mahsereci · Philipp Hennig

COEVOLVE: A Joint Point Process Model for Information Diffusion and Network Co-evolution
Mehrdad Farajtabar · Yichen Wang · Manuel Rodriguez · Shuang Li · Hongyuan Zha · Le Song

3:30 – 4:00 pm
Spotlight Session 3: Neuroscience and Neural Coding

4:00 – 4:30 pm - Coffee Break

4:30 – 5:30 pm
NIPS award Session
Oral Session 4
Competitive Distribution Estimation: Why is Good-Turing Good
Alon Orlitsky · Ananda Suresh

Fast Convergence of Regularized Learning in Games
Vasilis Syrgkanis · Alekh Agarwal · Haipeng Luo · Robert Schapire

Interactive Control of Diverse Complex Characters with Neural Networks
Igor Mordatch · Kendall Lowrey · Galen Andrew · Zoran Popovic · Emanuel Todorov

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

7:00 – 11:59 pm
Demonstrations
Poster Session

PROGRAM HIGHLIGHTS

WEDNESDAY DEC 9TH

7:30 am – 9:00 am
Breakfast, Level 2, 220A

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

9:00 – 9:50 am
INVITED TALK: Post-selection Inference for Forward Stepwise Regression, Lasso and other Adaptive Statistical procedures
Robert Tibshirani

9:50 – 10:10 am
Oral Session 5
Learning Theory and Algorithms for Forecasting Non-stationary Time Series
Vitaly Kuznetsov · Mehryar Mohri

10:10 –10:40 am
Spotlight Session 5: Regression and time series spotlights

10:40 – 11:10 am - Coffee Break

11:10 –11:50 am
Oral Session 6
Deep Visual Analogy-Making
Scott E Reed · Yi Zhang · Yuting Zhang · Honglak Lee

End-To-End Memory Networks
Sainbayar Sukhbaatar · arthur szlam · Jason Weston · Rob Fergus

11:50 – 12:00 pm
Spotlight Session 6: Learning theory spotlights

12:00 – 2:00 pm - Lunch Break

2:00 – 2:50 pm
INVITED TALK: Diagnosis and Therapy of Psychiatric Disorders Based on Brain Dynamics
Mitsuo Kawato

2:50 – 3:30 pm
Oral Session 7
A Reduced-Dimension fMRI Shared Response Model
Po-Hsuan (Cameron) Chen · Janice Chen · Yaara Yeshurun · Uri Hasson · James Haxby · Peter J Ramadge

Attractor Network Dynamics Enable Preplay and Rapid Path Planning in Maze–like Environments
Dane S Corneil · Wulfram Gerstner

3:30 – 4:00 pm
Spotlight Session 7: Reinforcement learning spotlights

4:00 – 4:30 pm - Coffee Break

4:30 – 5:20 pm
INVITED TALK: Computational Principles for Deep Neuronal Architectures
Haim Sompolinsky

5:20 – 5:40 pm
Oral Session 8
Efficient Exact Gradient Update for training Deep Networks with Very Large Sparse Targets
Pascal Vincent · Alexandre de Brébisson · Xavier Bouthillier

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 10TH

7:30 am – 9:00 am
Breakfast, Level 2, 220A

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

9:00 – 9:50 am
INVITED TALK: Learning with Intelligent Teacher: Similarity Control and Knowledge Transfer
Vladimir Vapnik

9:50 – 10:10 am
Oral Session 9
Less is More: Nyström Computational Regularization
Alessandro Rudi · Raffaello Camoriano · Lorenzo Rosasco

10:10 –10:40 am
Spotlight Session 9: Graphical models spotlights, model selection

10:40 – 10:50 am
Closing Remarks

10:50 am –11:00 am - Coffee Break

11:00 – 3:00 pm
Poster Session

3:00 pm - Symposia

2015 EXHIBITORS

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

From The President

The 2015 NIPS Conference and Workshops is on track to host a record number of participants, topping the record number last year in Montréal by nearly 1000. The success of deep learning has put "N" firmly back into NIPS. However, NIPS covers a wide range of computational approaches to analyzing large data sets and continues to attract researchers from a broad range of disciplines. Over the decades NIPS has overseen the development of many algorithms, including neural networks, support vector machines, and graphical models, and we expect the field to continue to evolve as new algorithms are invented to solve complex problems in all fields of science and engineering.

New additions to NIPS this year are a new named lecture, a new NIPS symposium track, and light snacks at the poster sessions. These are funded by the generous support from a record number of sponsors, which are also providing funding for student travel.

Robert Tibshirani from Stanford will deliver the first in an annual series of Breiman Lectures, which celebrates the close ties between the NIPS and the statistics communities. The NIPS Breiman Lecture is named after Leo Breiman, who served on the NIPS Board from 1994 to 2005. A distinguished statistician at UC Berkeley, his work on classification and regression trees was seminal. "Bagging" was his term for bootstrap aggregation. He bridged the statistics and machine learning communities, and was an important voice on the NIPS Board.

Two Posner Lectures will be given this year by distinguished NIPS researchers, Vladimer Vapnik and Zoubin Ghahramani. The NIPS Posner Lecture is named after Ed Posner, who founded NIPS in 1987. Ed worked on communications and information theory at Caltech and was an early pioneer in neural networks. He organized the first NIPS conference and workshop in Denver in 1989 and incorporated the NIPS Foundation in 1992. He was an inspiring teacher and an effective leader whose influence continues today.

Three NIPS symposia will be held in parallel on Thursday, Dec 10 between the Conference and the Workshops. The goal of these symposia is to present timely lectures on a single broad theme, to complement the single track in the main conference. This revives the Symposium track held at the NIPS Conferences in Vancouver to fill a gap between the end of the conference and the beginning of the workshops in Whistler. The symposia are:

- Algorithms Among Us: the Societal Impacts of Machine Learning
- Brains, Minds, and Machines
- Deep Learning.

In 2016 the NIPS Conference will be held in Barcelona, Spain, returning to Europe after the popular 2011 NIPS Conference in Granada, Spain. In 2017, NIPS will be in Long Beach, California – warmer weather! NIPS will return to Montréal, Canada, in 2018.

Terry Sejnowski

NIPS Foundation President

UPCOMING CONFERENCES

Barcelona 2016

Long Beach 2017

MONDAY SESSIONS

OUR SPONSORS

Google
Microsoft
Alibaba Group 阿里巴巴集团
Ketchum Trading, LLC
amazon.com
Apple
Baidu Research
CITADEL
facebook
IBM Research
NVIDIA
The Voleon Group
Artificial Intelligence
www.elsevier.com/locate/artint
Bloomberg
twitter
AdRoll
Analog Devices | Lyric Labs
CenturyLink Business
criteo
Cubist Systematic Strategies
deep genomics
D E Shaw & Co
NETFLIX
Disney Research
ebay
imagia
MaluubA
Man AHL
ORACLE
Panasonic
PDT PARTNERS
SONY
The Alan Turing Institute
TOYOTA
United Technologies Research Center
Vatic Labs
TWO SIGMA
YAHOO! LABS
WINTON
Adobe

MONDAY TUTORIALS

INTRODUCTORY TOPICS

Tutorial Session A, 9:30 – 11:30 AM
Session Chair: Ralf Herbrich

Large-Scale Distributed Systems for Training Neural Networks
Jeff Dean · Oriol Vinyals
Level 2 room 210 EF

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

Deep Learning
Geoffrey E Hinton · Yoshua Bengio · Yann LeCun
Level 2 room 210 AB

Tutorial Session A, 1:00 – 3:00 PM
Session Chair: Joaquin Qunionero Candela

Monte Carlo Inference Methods Adventures in Simulated Evolution
Iain Murray
Location: Level 2, Room 210 AB

Tutorial Session B, 1:00 – 3:00 PM
Session Chair: Ralf Herbrich

Probabilistic Programming
Frank Wood
Location: Level 2, Room 210 EF

Tutorial Session A, 3:30 – 5:30 PM
Session Chair: Joaquin Qunionero Candela

Introduction to Reinforcement Learning with Function Approximation
Rich S Sutton
Location: Level 2, Room 210 AB

Tutorial Session B, 3:30 – 5:30 PM
Session Chair: Ralf Herbrich

High-Performance Hardware for Machine Learning
Bill Dally
Location: Level 2, Room 210 EF

ABSTRACTS OF TUTORIALS

Tutorial Session A, 9:30 – 11:30 AM

Session Chair: Ralf Herbrich

Large-Scale Distributed Systems for Training Neural Networks

Jeff Dean jeff@google.com
Google Brain Team
Oriol Vinyals vinyals@google.com
Google

Over the past few years, we have built large-scale computer systems for training neural networks, and then applied these systems to a wide variety of problems that have traditionally been very difficult for computers. We have made significant improvements in the state-of-the-art in many of these areas, and our software systems and algorithms have been used by dozens of different groups at Google to train state-of-the-art models for speech recognition, image recognition, various visual detection tasks, language modeling, language translation, and many other tasks. In this talk,we'll highlight some of the distributed systems and algorithms that we use in order to train large models quickly, and demonstrate some of the software systems we have put together that make it easy to conduct research in large-scale machine learning.

Session Chair: Joaquin Qunionero Candela

Deep Learning

Geoffrey E Hinton hinton@cs.toronto.edu
Google & University of Toronto
Yoshua Bengio bengioy@iro.umontreal.ca
University of Montreal
Yann LeCun yann@cs.nyu.edu
New York University

Deep Learning allows computational models composed of multiple processing layers to learn representations of data with multiple levels of abstraction. These methods have dramatically improved the state-of-the-art in speech recognition, visual object recognition, object detection, and many other domains such as drug discovery and genomics. Deep learning discovers intricate structure in large datasets by using the back-propagation algorithm to indicate how a machine should change its internal parameters that are used to compute the representation in each layer from the representation in the previous layer. Deep convolutional nets have brought about dramatic improvements in processing images, video, speech and audio, while recurrent nets have shone on sequential data such as text and speech. Representation learning is a set of methods that allows a machine to be fed with raw data and to automatically discover the representations needed for detection or classification. Deep learning methods are representation learning methods with multiple levels of representation, obtained by composing simple but non-linear modules that each transform the representation at one level (starting with the raw input) into a representation at a higher, slightly more abstract level. This tutorial will introduce the fundamentals of deep learning, discuss applications, and close with challenges ahead.

ABSTRACTS OF TUTORIALS

Tutorial Session A, 1:00 – 3:30 PM

Session Chair: Joaquin Qunionero Candela

Monte Carlo Inference Methods
Level 2 room 210 AB

Iain Murray i.murray@ed.ac.uk
University of Edinburgh

"Monte Carlo" methods use random sampling to understand a system, estimate averages, or compute integrals. Monte Carlo methods were amongst the earliest applications run on electronic computers in the 1940s, and continue to see widespread use and research as our models and computational power grow. In the NIPS community, random sampling is widely used within optimization methods, and as a way to perform inference in probabilistic models. Here "inference" simply means obtaining multiple plausible settings of model parameters that could have led to the observed data. Obtaining a range of explanations tells us both what we can and cannot know from our data, and prevents us from making overconfident (wrong) predictions.

This introductory-level tutorial will describe some of the fundamental Monte Carlo algorithms, and examples of how they can be combined with models in different ways. We'll see that Monte Carlo methods are sometimes a quick and easy way to perform inference in a new model, but also what can go wrong, and some treatment of how to debug these randomized algorithms.

Session Chair: Ralf Herbrich

Probabilistic Programming
Level 2 room 210 E,F

Frank Wood fwood@robots.ox.ac.uk
University of Oxford

Probabilistic programming is a general-purpose means of expressing and automatically performing model-based inference. A key characteristic of many probabilistic programming systems is that models can be compactly expressed in terms of executable generative procedures, rather than in declarative mathematical notation. For this reason, along with automated or programmable inference, probabilistic programming has the potential to increase the number of people who can build and understand their own models. It also could make the development and testing of new general-purpose inference algorithms more efficient, and could accelerate the exploration and development of new models for application-specific use.

The primary goals of this tutorial will be to introduce probabilistic programming both as a general concept and in terms of how current systems work, to examine the historical academic context in which probabilistic programming arose, and to expose some challenges unique to probabilistic programming.

Tutorial Session A, 3:30 – 5:30 PM

Session Chair: Joaquin Qunionero Candela

Introduction to Reinforcement Learning with Function Approximation
Level 2 room 210 AB

Rich S Sutton rich@richsutton.com
University of Alberta

Reinforcement learning is a body of theory and techniques for optimal sequential decision making developed in the last thirty years primarily within the machine learning and operations research communities, and which has separately become important in psychology and neuroscience. This tutorial will develop an intuitive understanding of the underlying formal problem (Markov decision processes) and its core solution methods, including dynamic programming, Monte Carlo methods, and temporal-difference learning. It will focus on how these methods have been combined with parametric function approximation, including deep learning, to find good approximate solutions to problems that are otherwise too large to be addressed at all. Finally, it will briefly survey some recent developments in function approximation, eligibility traces, and off-policy learning.

Session Chair: Ralf Herbrich

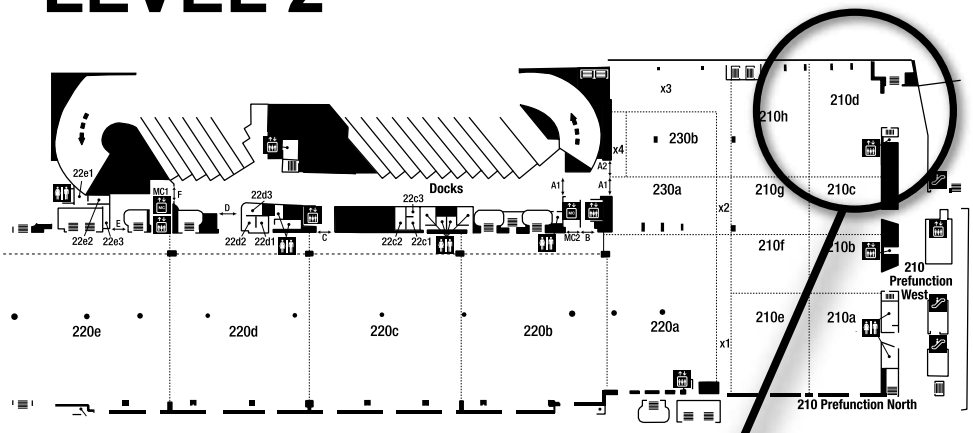
High-Performance Hardware for Machine Learning
Level 2 room 210 E,F

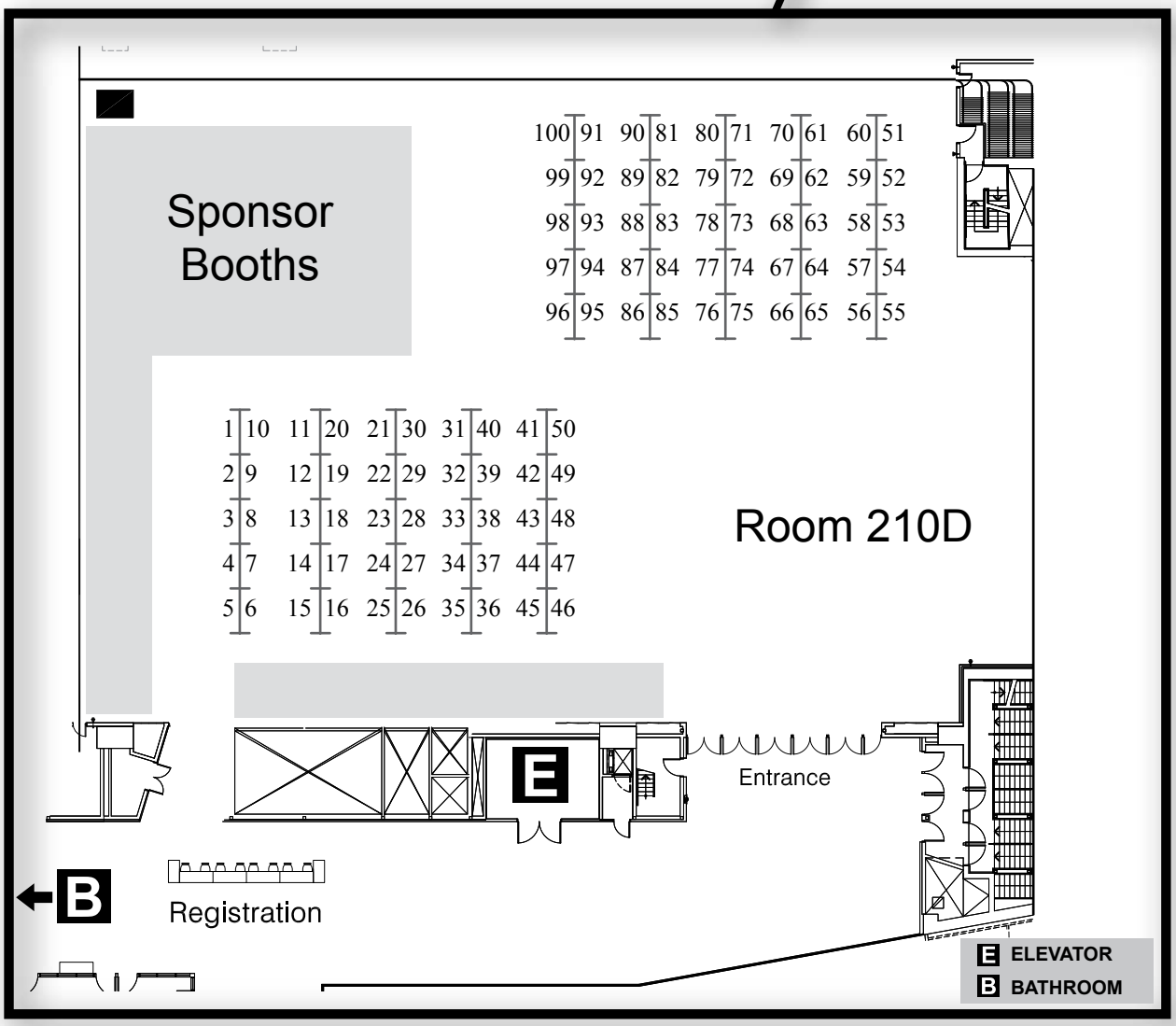
Bill Dally dally@stanford.edu
Stanford University

This tutorial will survey the state of the art in high-performance hardware for machine learning with an emphasis on hardware for training and deployment of deep neural networks (DNNs). We establish a baseline by characterizing the performance and efficiency (perf/W) of DNNs implemented on conventional CPUs. GPU implementations of DNNs make substantial improvements over this baseline. GPU implementations perform best with moderate batch sizes. We examine the sensitivity of performance to batch size. Training of DNNs can be accelerated further using both model and data parallelism, at the cost of inter-processor communication. We examine common parallel formulations and the communication traffic they induce. Training and deployment can also be accelerated by using reduced precision for weights and activations. We will examine the tradeoff between accuracy and precision in these networks. We close with a discussion of dedicated hardware for machine learning. We survey recent publications on this topic and make some general observations about the relative importance of arithmetic and memory bandwidth in such dedicated hardware.

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

1 Texture Synthesis Using Convolutional Neural Networks
Leon Gatys · Alexander S Ecker · Matthias Bethge

2 Convolutional Neural Networks with Intra-Layer Recurrent Connections for Scene Labeling
Ming Liang · Xiaolin Hu · Bo Zhang

3 Grammar as a Foreign Language
Oriol Vinyals · Łukasz Kaiser · Terry Koo · Slav Petrov · Ilya Sutskever · Geoffrey Hinton

4 Recursive Training of 2D-3D Convolutional Networks for Neuronal Boundary Prediction
Kisuk Lee · Aleks Zlateski · Vishwanathan Ashwin · H. Sebastian Seung

5 Generative Image Modeling Using Spatial LSTMs
Lucas Theis · Matthias Bethge

6 Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks
Shaoqing Ren · Kaiming He · Ross Girshick · Jian Sun

7 Weakly-supervised Disentangling with Recurrent Transformations for 3D View Synthesis
Jimei Yang · Scott E Reed · Ming-Hsuan Yang · Honglak Lee

8 Exploring Models and Data for Image Question Answering
Mengye Ren · Ryan Kiros · Richard Zemel

9 Are You Talking to a Machine? Dataset and Methods for Multilingual Image Question
Haoyuan Gao · Junhua Mao · Jie Zhou · Zhiheng Huang · Lei Wang · Wei Xu

10 Parallel Multi-Dimensional LSTM, With Application to Fast Biomedical Volumetric Image Segmentation
Marijn F Stollenga · Wonmin Byeon · Marcus Liwicki · Juergen Schmidhuber

11 Learning-Curve Analysis of Simple Decision Heuristics
Özgür Şimşek · Marcus Buckmann

12 3D Object Proposals for Accurate Object Class Detection
Xiaozhi Chen · Kaustav Kundu · Yukun Zhu · Andrew G Berneshawi · Huimin Ma · Sanja Fidler · Raquel Urtasun

13 The Poisson Gamma Belief Network
Mingyuan Zhou · Yulai Cong · Bo Chen

14 Semi-Supervised Factored Logistic Regression for High-Dimensional Neuroimaging Data
Danilo Bzdok · Michael Eickenberg · Olivier Grisel · Bertrand Thirion · Gael Varoquaux

15 BinaryConnect: Training Deep Neural Networks with binary weights during propagations
Matthieu Courbariaux · Yoshua Bengio · Jean-Pierre David

16 Learning to Transduce with Unbounded Memory
Edward Grefenstette · Karl Moritz Hermann · Mustafa Suleyman · Phil Blunsom

17 Spectral Representations for Convolutional Neural Networks
Oren Rippel · Jasper Snoek · Ryan P Adams

18 A Theory of Decision Making Under Dynamic Context
Michael Shvartsman · Vaibhav Srivastava · Jonathan D Cohen

19 Bidirectional Recurrent Neural Networks as Generative Models
Mathias Berglund · Tapani Raiko · Mikko Honkala · Leo Kärkkäinen · Akos Vetek · Juha T Karhunen

20 Recognizing retinal ganglion cells in the dark
Emile Richard · Georges A Goetz · E. J. Chichilnisky

21 A Recurrent Latent Variable Model for Sequential Data
Junyoung Chung · Kyle Kastner · Laurent Dinh · Kratarth Goel · Aaron C Courville · Yoshua Bengio

22 Deep Knowledge Tracing
Chris Piech · Jonathan Bassen · Jonathan Huang · Surya Ganguli · Mehran Sahami · Leonidas J Guibas · Jascha Sohl-Dickstein

23 Deep Temporal Sigmoid Belief Networks for Sequence Modeling
Zhe Gan · Chunyuan Li · Ricardo Henao · David E Carlson · Lawrence Carin

24 Hidden Technical Debt in Machine Learning Systems
D. Sculley · Gary Holt · Daniel Golovin · Eugene Davydov · Todd Phillips · Dietmar Ebner · Vinay Chaudhary · Michael Young · JF Crespo · Dan Dennison

25 Statistical Model Criticism using Kernel Two Sample Tests
James R Lloyd · Zoubin Ghahramani

26 Calibrated Structured Prediction
Volodymyr Kuleshov · Percy S Liang

27 A Bayesian Framework for Modeling Confidence in Perceptual Decision Making
Koosha Khalvati · Rajesh P Rao

28 Dependent Multinomial Models Made Easy: Stick-Breaking with the Polya-gamma Augmentation
Scott Linderman · Matthew Johnson · Ryan P Adams

29 Scalable Adaptation of State Complexity for Nonparametric Hidden Markov Models
Mike C Hughes · Will T Stephenson · Erik Sudderth

30 Robust Feature-Sample Linear Discriminant Analysis for Brain Disorders Diagnosis
Ehsan Adeli-Mosabbeb · Kim-Han Thung · Le An · Feng Shi · Dinggang Shen

31 Learning spatiotemporal trajectories from manifold-valued longitudinal data
Jean-Baptiste SCHIRATTI · Stéphanie ALLASSONNIERE · Olivier Colliot · Stanley DURRLEMAN

32 Hessian-free Optimization for Learning Deep Multidimensional Recurrent Neural Networks
Minhyung Cho · Chandra Dhir · Jaehyung Lee

33 Scalable Inference for Gaussian Process Models with Black-Box Likelihoods
Amir Dezfouli · Edwin Bonilla · Edwin V Bonilla

34 Variational Dropout and the Local Reparameterization Trick
Diederik P Kingma · Tim Salimans · Max Welling

35 Infinite Factorial Dynamical Model
Isabel Valera · Francisco J. R. Ruiz · Lennart Svensson · Fernando Perez-Cruz

36 Variational Information Maximisation for Intrinsically Motivated Reinforcement Learning
Shakir Mohamed · Danilo Jimenez Rezende

37 Variational inference with copula augmentation
Dustin Tran · David Blei · Edo M Airoldi

38 Fast Second Order Stochastic Backpropagation for Variational Inference
Kai Fan · Ziteng Wang · Jeff Beck · James Kwok · Katherine A Heller

39 Rethinking LDA: Moment Matching for Discrete ICA
Anastasia Podosinnikova · Francis Bach · Simon Lacoste-Julien

40 Model-Based Relative Entropy Stochastic Search
Abbas Abdolmaleki · Rudolf Lioutikov · Jan R Peters · Nuno Lau · Luis Pualo Reis · Gerhard Neumann

41 On Predictive Belief Methods for Dynamical System Learning
Ahmed Hefny · Carlton Downey · Geoffrey J Gordon

42 Expectation Particle Belief Propagation
Thibaut Lienart · Yee Whye Teh · Arnaud Doucet

43 Embedding Inference for Structured Multilabel Prediction
Farzaneh Mirzazadeh · Siamak Ravanbakhsh · Nan Ding · Dale Schuurmans

44 Tractable Learning for Complex Probability Queries
Jessa Bekker · Guy Van den Broeck · Arthur Choi · Adnan Darwiche · Jesse Davis

45 Double or Nothing: Multiplicative Incentive Mechanisms for Crowdsourcing
Nihar Bhadresh Shah · Denny Zhou

46 Local Expectation Gradients for Black Box Variational Inference
Michalis Titsias · Miguel Lázaro-Gredilla

47 Learning with a Wasserstein Loss
Charlie Frogner · Chiyuan Zhang · Hossein Mobahi · Mauricio Araya · Tomaso A Poggio

48 Principal Geodesic Analysis for Probability Measures under the Optimal Transport Metric
Vivien Seguy · Marco Cuturi

49 Fast and Accurate Inference of Plackett–Luce Models
Lucas Maystre · Matthias Grossglauser

50 BACKSHIFT: Learning causal cyclic graphs from unknown shift interventions
Dominik Rothenhäusler · Christina Heinze · Jonas Peters · Nicolai Meinshausen

51 Learning with Relaxed Supervision
Jacob Steinhardt · Percy S Liang

52 M-Statistic for Kernel Change-Point Detection
Shuang Li · Yao Xie · Hanjun Dai · Le Song

53 Fast Two-Sample Testing with Analytic Representations of Probability Measures
Kacper P Chwialkowski · Aaditya Ramdas · Dino Sejdinovic · Arthur Gretton

54 Adversarial Prediction Games for Multivariate Losses
Hong Wang · Wei Xing · Kaiser Asif · Brian Ziebart

55 Regressive Virtual Metric Learning
Michaël Perrot · Amaury Habrard

56 Halting in Random Walk Kernels
Mahito Sugiyama · Karsten Borgwardt

57 Rate-Agnostic (Causal) Structure Learning
Sergey Plis · David Danks · Cynthia Freeman · Vince Calhoun

58 Online Prediction at the Limit of Zero Temperature
Mark Herbster · Stephen Pasteris · Shaona Ghosh

59 Lifted Symmetry Detection and Breaking for MAP Inference
Tim Kopp · Parag Singla · Henry Kautz

60 Bandits with Unobserved Confounders: A Causal Approach
Elias Bareinboim · Andrew Forney · Judea Pearl

61 Sample Complexity Bounds for Iterative Stochastic Policy Optimization
Marin Kobilarov

62 Basis refinement strategies for linear value function approximation in MDPs
Gheorghe Comanici · Doina Precup · Prakash Panangaden

63 Probabilistic Variational Bounds for Graphical Models
Qiang Liu · John W Fisher III · Alex T Ihler

64 On the Convergence of Stochastic Gradient MCMC Algorithms with High-Order Integrators
Changyou Chen · Nan Ding · Lawrence Carin

65 An Active Learning Framework using Sparse-Graph Codes for Sparse Polynomials and Graph Sketching
Xiao Li · Kannan Ramchandran

66 Discrete Rényi Classifiers
Meisam Razaviyayn · Farzan Farnia · David Tse

67 GAP Safe screening rules for sparse multi-task and multi-class models
Eugene Ndiaye · Olivier Fercoq · Alexandre Gramfort · Joseph Salmon

68 Decomposition Bounds for Marginal MAP
Wei Ping · Qiang Liu · Alex T Ihler

69 Anytime Influence Bounds and the Explosive Behavior of Continuous-Time Diffusion Networks
Kevin Scaman · Rémi Lemonnier · Nicolas Vayatis

70 Estimating Mixture Models via Mixtures of Polynomials
Sida Wang · Arun Tejasvi Chaganty · Percy S Liang

71 Robust Gaussian Graphical Modeling with the Trimmed Graphical Lasso
Eunho Yang · Aurelie C Lozano

72 Matrix Completion from Fewer Entries: Spectral Detectability and Rank Estimation
Alaa Saade · Florent Krzakala · Lenka Zdeborová

73 Robust PCA with compressed data
Wooseok Ha · Rina Foygel Barber

74 Mixed Robust/Average Submodular Partitioning: Fast Algorithms, Guarantees, and Applications
Kai Wei · Rishabh K Iyer · Shengjie Wang · Wenruo Bai · Jeff A Bilmes

75 Subspace Clustering with Irrelevant Features via Robust Dantzig Selector
Chao Qu · Huan Xu

76 A class of network models recoverable by spectral clustering
Yali Wan · Marina Meila

77 Monotone k-Submodular Function Maximization with Size Constraints
Naoto Ohsaka · Yuichi Yoshida

78 Smooth and Strong: MAP Inference with Linear Convergence
Ofer P Meshi · Mehrdad Mahdavi · Alex Schwing

79 StopWasting My Gradients: Practical SVRG
Reza Harikandeh · Mohamed Osama Ahmed · Alim Virani · Mark Schmidt · Jakub Konečný · Scott Sallinen

80 Spectral Norm Regularization of Orthonormal Representations for Graph Transduction
Rakesh Shivanna · Bibaswan K Chatterjee · Raman Sankaran · Chiranjib Bhattacharyya · Francis Bach

81 Differentially Private Learning of Structured Discrete Distributions
Ilias Diakonikolas · Moritz Hardt · Ludwig Schmidt

82 Robust Portfolio Optimization
Huitong Qiu · Fang Han · Han Liu · Brian Caffo

83 Bayesian Optimization with Exponential Convergence
Kenji Kawaguchi · Leslie Kaelbling · Tomás Lozano-Pérez

84 Fast Randomized Kernel Ridge Regression with Statistical Guarantees
Ahmed Alaoui · Michael W Mahoney

85 Taming the Wild: A Unified Analysis of Hogwild-Style Algorithms
Christopher M De Sa · Ce Zhang · Kunle Olukotun · Christopher Ré · Chris Ré

86 Beyond Convexity: Stochastic Quasi-Convex Optimization
Elad Hazan · Kfir Levy · Shai Shalev-Shwartz

87 On the Limitation of Spectral Methods: From the Gaussian Hidden Clique Problem to Rank-One Perturbations of Gaussian Tensors
Andrea Montanari · Daniel Reichman · Ofer Zeitouni

88 Regularized EM Algorithms: A Unified Framework and Statistical Guarantees
Xinyang Yi · Constantine Caramanis

89 Black-box optimization of noisy functions with unknown smoothness
jean-bastien grill · Michal Valko · Remi Munos

90 Combinatorial Cascading Bandits
Branislav Kveton · Zheng Wen · Azin Ashkan · Csaba Szepesvari

91 Adaptive Primal-Dual Splitting Methods for Statistical Learning and Image Processing
Tom Goldstein · Min Li · Xiaoming Yuan

92 Sum-of-Squares Lower Bounds for Sparse PCA
Tengyu Ma · Avi Wigderson

93 Online Gradient Boosting
Alina Beygelzimer · Elad Hazan · Satyen Kale · Haipeng Luo

94 Regularization-Free Estimation in Trace Regression with Symmetric Positive Semidefinite Matrices
Martin Slawski · Ping Li · Matthias Hein

95 Convergence Analysis of Prediction Markets via Randomized Subspace Descent
Rafael M Frongillo · Mark D Reid

96 Accelerated Proximal Gradient Methods for Nonconvex Programming
Li Huan · Zhouchen Lin

97 Nearly Optimal Private LASSO
Kunal Talwar · Abhradeep Thakurta · Li Zhang

98 Minimax Time Series Prediction
Wouter M Koolen · Alan Malek · Peter L Bartlett · Yasin Abbasi

99 Communication Complexity of Distributed Convex Learning and Optimization
Yossi Arjevani · Ohad Shamir

100 Explore no more: Improved high-probability regret bounds for non-stochastic bandits
Gergely Neu

101 A Nonconvex Optimization Framework for Low Rank Matrix Estimation
Tuo Zhao · Zhaoran Wang · Han Liu

102 Individual Planning in Infinite-Horizon Multiagent Settings: Inference, Structure and Scalability
Xia Qu · Prashant Doshi

1 Texture Synthesis Using Convolutional Neural Networks

Leon Gatys leon.gatys@bethgelab.org
Alexander S Ecker alexander.ecker@uni-tuebingen.de
Matthias Bethge matthias@bethgelab.org
University of Tuebingen

Here we introduce a new model of natural textures based on the feature spaces of convolutional neural networks optimised for object recognition. Samples from the model are of high perceptual quality demonstrating the generative power of neural networks learned in a purely discriminative fashion. Within the model, textures are represented by the correlations between feature maps in several layers of the network. We show that across layers the texture representations increasingly capture the statistical properties of natural images while making object information more and more explicit. The model provides a new tool to generate stimuli for neuroscience and might offer insights into the deep representations learned by convolutional neural networks.

2 Convolutional Neural Networks with Intra-Layer Recurrent Connections for Scene Labeling

Ming Liang liangm07@mails.tsinghua.edu.cn
Xiaolin Hu xlhu@tsinghua.edu.cn
Bo Zhang dcszb@tsinghua.edu.cn
Tsinghua University

Scene labeling is a challenging computer vision task. It requires the use of both local discriminative features and global context information. We adopt a deep recurrent convolutional neural network (RCNN) for this task, which is originally proposed for object recognition. Different from traditional CNN, this model has intra-layer recurrent connections in the convolutional layers. Therefore each convolutional layer becomes a two-dimensional recurrent neural network. The units receive constant feed-forward inputs from the previous layer and recurrent inputs from their neighborhoods. While recurrent iterations proceed, the region of context captured by each unit expands. In this way, feature extraction and context modulation are seamlessly integrated, which is different from typical methods that entail separate modules for the two steps. To further utilize the context, a multi-scale RCNN is proposed. Over two benchmark datasets, Standford Background and Sift Flow, the model outperforms the state-of-the-art models and takes less time to fully label an image.

3 Grammar as a Foreign Language

Oriol Vinyals vinyals@google.com
Łukasz Kaiser lukaszkaiser@google.com
Terry Koo terrykoo@google.com
Slav Petrov slav@google.com
Ilya Sutskever ilyasu@google.com
Geoffrey Hinton geoffhinton@google.com
Google

Syntactic constituency parsing is a fundamental problem in naturallanguage processing which has been the subject of intensive researchand engineering for decades. As a result, the most accurate parsersare domain specific, complex, and inefficient. In this paper we showthat the domain agnostic attention-enhanced sequence-to-sequence modelachieves state-of-the-art results on the most widely used syntacticconstituency parsing dataset, when trained on a large synthetic corpusthat was annotated using existing parsers. It also matches theperformance of standard parsers when trained on a smallhuman-annotated dataset, which shows that this model is highlydata-efficient, in contrast to sequence-to-sequence models without theattention mechanism. Our parser is also fast, processing over ahundred sentences per second with an unoptimized CPU implementation.

4 Recursive Training of 2D-3D Convolutional Networks for Neuronal Boundary Prediction

Kisuk Lee kisuklee@mit.edu
Aleks Zlateski zlateski@mit.edu
MIT
Vishwanathan Ashwin ashwinv@princeton.edu
H. Sebastian Seung sseung@princeton.edu
Princeton University

Efforts to automate the reconstruction of neural circuits from 3D electron microscopic (EM) brain images are critical for the field of connectomics. An important computation for reconstruction is the detection of neuronal boundaries. Images acquired by serial section EM, a leading 3D EM technique, are highly anisotropic, with inferior quality along the third dimension. For such images, the 2D max-pooling convolutional network has set the standard for performance at boundary detection. Here we achieve a substantial gain in accuracy through three innovations. Following the trend towards deeper networks for object recognition, we use a much deeper network than previously employed for boundary detection. Second, we incorporate 3D as well as 2D filters, to enable computations that use 3D context. Finally, we adopt a recursively trained architecture in which a first network generates a preliminary boundary map that is provided as input along with the original image to a second network that generates a final boundary map. Backpropagation training is accelerated by ZNN, a new implementation of 3D convolutional networks that uses multicore CPU parallelism for speed. Our hybrid 2D-3D architecture could be more generally applicable to other types of anisotropic 3D images, including video, and our recursive framework for any image labeling problem.

5 Generative Image Modeling Using Spatial LSTMs

Lucas Theis lucas@bethgelab.org
U.Tuebingen
Matthias Bethge matthias@bethgelab.org
CIN, University Tübingen

Modeling the distribution of natural images is challenging, partly because of strong statistical dependencies which can extend over hundreds of pixels. Recurrent neural networks have been successful in capturing long-range dependencies in a number of problems but only recently have found their way into generative image models. We here introduce a recurrent image model based on multi-dimensional long short-term memory units which are particularly suited for image modeling due to their spatial structure. Our model scales to images of arbitrary size and its likelihood is computationally tractable. We find that it outperforms the state of the art in quantitative comparisons on several image datasets and produces promising results when used for texture synthesis and inpainting.

6 Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks

Shaoqing Ren sqren@mail.ustc.edu.cn
USTC
Kaiming He kahe@microsoft.com
Ross Girshick rbg@microsoft.com
Jian Sun jiansun@microsoft.com
Microsoft Research Next

State-of-the-art object detection networks depend on region proposal algorithms to hypothesize object locations. Advances like SPPnet and Fast R-CNN have reduced the running time of these detection networks, exposing region proposal computation as a bottleneck. In this work, we introduce a Region Proposal Network (RPN) that shares full-image convolutional features with the detection network, thus enabling nearly cost-free region proposals. An RPN is a fully-convolutional network that simultaneously predicts object bounds and objectness scores at each position. RPNs are trained end-to-end to generate high-quality region proposals, which are used by Fast R-CNN for detection. With a simple alternating optimization, RPN and Fast R-CNN can be trained to share convolutional features. For the very deep VGG-16 model, our detection system has a frame rate of 5fps (including all steps) on a GPU, while achieving state-of-the-art object detection accuracy on PASCAL VOC 2007 (73.2% mAP) and 2012 (70.4% mAP) using 300 proposals per image. Code is available at https://github.com/ShaoqingRen/faster_rcnn.

7 Weakly-supervised Disentangling with Recurrent Transformations for 3D View Synthesis

Jimei Yang jyang44@ucmerced.edu
Ming-Hsuan Yang mhyang@ucmerced.edu
UC Merced
Scott E Reed reedscot@umich.edu
Honglak Lee honglak@eecs.umich.edu
University of Michigan

An important problem for both graphics and vision is to synthesize novel views of a 3D object from a single image. This is in particular challenging due to the partial observability inherent in projecting a 3D object onto the image space, and the ill-posedness of inferring object shape and pose. However, we can train a neural network to address the problem if we restrict our attention to specific object classes (in our case faces and chairs) for which we can gather ample training data. In this paper, we propose a novel recurrent convolutional encoder-decoder network that is trained end-to-end on the task of rendering rotated objects starting from a single image. The recurrent structure allows our model to capture long-term dependencies along a sequence of transformations, and we demonstrate the quality of its predictions for human faces on the Multi-PIE dataset and for a dataset of 3D chair models, and also show its ability of disentangling latent data factors without using object class labels.

8 Exploring Models and Data for Image Question Answering

Mengye Ren mren@cs.toronto.edu
Ryan Kiros rkiros@cs.toronto.edu
Richard Zemel zemel@cs.toronto.edu
University of Toronto

This work aims to address the problem of image-based question-answering (QA) with new models and datasets. In our work, we propose to use neural networks and visual semantic embeddings, without intermediate stages such as object detection and image segmentation, to predict answers to simple questions about images. Our model performs 1.8 times better than the only published results on an existing image QA dataset. We also present a question generation algorithm that converts image descriptions, which are widely available, into QA form. We used this algorithm to produce an order-of-magnitude larger dataset, with more evenly distributed answers. A suite of baseline results on this new dataset are also presented.

9 Are You Talking to a Machine? Dataset and Methods for Multilingual Image Question

Haoyuan Gao gaohaoyuan@baidu.com
Jie Zhou zhoujie01@baidu.com
Zhiheng Huang huangzhiheng@baidu.com
Lei Wang wanglei22@baidu.com
Wei Xu wei.xu@baidu.com
Baidu
Junhua Mao mjhustc@ucla.edu
UCLA

In this paper, we present the mQA model, which is able to answer questions about the content of an image. The answer can be a sentence, a phrase or a single word. Our model contains four components: a Long-Short Term Memory (LSTM) to extract the question representation, a Convolutional Neural Network (CNN) to extract the visual representation, a LSTM for storing the linguistic context in an answer, and a fusing component to combine the information from the first three components and generate the answer. We construct a Freestyle Multilingual Image Question Answering (FM-IQA) dataset to train and evaluate our mQA model. It contains over 120,000 images and 250,000 freestyle Chinese question-answer pairs and their English translations. The quality of the generated answers of our mQA model on this dataset are evaluated by human judges through a Turing Test. Specifically, we mix the answers provided by humans and our model. The human judges need to distinguish our model from the human. They will also provide a score (i.e. 0, 1, 2, the larger the better) indicating the quality of the answer. We propose strategies to monitor the quality of this evaluation process. The experiments show that in 64.7% of cases, the human judges cannot distinguish our model from humans. The average score is 1.454 (1.918 for human). The details of this work, including the release of the FM-IQA dataset, can be found in \url{http://idl.baidu.com/FM-IQA.html}

10 Parallel Multi-Dimensional LSTM, With Application to Fast Biomedical Volumetric Image Segmentation

Marijn F Stollenga marijn@idsia.ch
Wonmin Byeon wonmin.byeon@dfki.de
IDSIA
Marcus Liwicki liwicki@cs.uni-kl.de
TU Kaiserslautern
Juergen Schmidhuber juergen@idsia.ch

Convolutional Neural Networks (CNNs) can be shifted across 2D images or 3D videos to segment them. They have a fixed input size and typically perceive only small local contexts of the pixels to be classified as foreground or background. In contrast, Multi-Dimensional Recurrent NNs (MD-RNNs) can perceive the entire spatio-temporal context of each pixel in a few sweeps through all pixels, especially when the RNN is a Long Short-Term Memory (LSTM). Despite these theoretical advantages, however, unlike CNNs, previous MD-LSTM variants were hard to parallelise on GPUs. Here we re-arrange the traditional cuboid order of computations in MD-LSTM in pyramidal fashion. The resulting PyraMiD-LSTM is easy to parallelise, especially for 3D data such as stacks of brain slice images. PyraMiD-LSTM achieved best known pixel-wise brain image segmentation results on MRBrainS13 (and competitive results on EM-ISBI12).

11 Learning-Curve Analysis of Simple Decision Heuristics

Özgür Şimşek ozgur@mpib-berlin.mpg.de
Marcus Buckmann buckmann@mpib-berlin.mpg.de
Max Planck Institute

Simple decision heuristics are models of human and animal behavior that use few pieces of information---perhaps only a single piece of information---and integrate the pieces in simple ways, for example by considering them sequentially, one at a time, or by giving them equal weight. We present analytical and empirical results on the effectiveness of simple decision heuristics in paired comparison problems. We examine three families of heuristics: single-cue decision making, lexicographic heuristics, and tallying. Our empirical analysis is the most extensive to date, employing 63 natural data sets on diverse subjects.

12 3D Object Proposals for Accurate Object Class Detection

Xiaozhi Chen chenxz12@mails.tsinghua.edu.cn
Huimin Ma mhmpub@tsinghua.edu.cn
Tsinghua University
Kaustav Kundu kkundu@cs.toronto.edu
Yukun Zhu yukun@cs.toronto.edu
Andrew G Berneshawi andrew.berneshawi@mail.utoronto.ca
Sanja Fidler fidler@cs.toronto.edu
Raquel Urtasun urtasun@cs.toronto.edu
University of Toronto

The goal of this paper is to generate high-quality 3D object proposals in the context of autonomous driving. Our method exploits stereo imagery to place proposals in the form of 3D bounding boxes. We formulate the problem as minimizing an energy function encoding object size priors, ground plane as well as several depth informed features that reason about free space, point cloud densities and distance to the ground. Our experiments show significant performance gains over existing RGB and RGB-D object proposal methods on the challenging KITTI benchmark. Combined with convolutional neural net (CNN) scoring, our approach outperforms all existing results on *Car and Cyclist*, and is competitive for the *Pedestrian* class.

13 The Poisson Gamma Belief Network

Mingyuan Zhou mzhou@utexas.edu
University of Texas at Austin
Yulai Cong yulai_cong@163.com
Bo Chen bchen@mail.xidian.edu.cn
Xidian University

To infer a multilayer representation of high-dimensional count vectors, we propose the Poisson gamma belief network (PGBN) that factorizes each of its layers into the product of a connection weight matrix and the nonnegative real hidden units of the next layer. The PGBN's hidden layers are jointly trained with an upward-downward Gibbs sampler, each iteration of which upward samples Dirichlet distributed connection weight vectors starting from the first layer (bottom data layer), and then downward samples gamma distributed hidden units starting from the top hidden layer. The gamma-negative binomial process combined with a layer-wise training strategy allows the PGBN to infer the width of each layer given a fixed budget on the width of the first layer. The PGBN with a single hidden layer reduces to Poisson factor analysis. Example results on text analysis illustrate interesting relationships between the width of the first layer and the inferred network structure, and demonstrate that the PGBN, whose hidden units are imposed with correlated gamma priors, can add more layers to increase its performance gains over Poisson factor analysis, given the same limit on the width of the first layer.

14 Semi-Supervised Factored Logistic Regression for High-Dimensional Neuroimaging Data

Danilo Bzdok danilo.bzdok@inria.fr
Michael Eickenberg michael.eickenberg@gmail.com
Olivier Grisel olivier.grisel@ensta.org
Bertrand Thirion bertrand.thirion@inria.fr
Gael Varoquaux gael.varoquaux@inria.fr
INRIA

Imaging neuroscience links human behavior to aspects of brain biology in ever-increasing datasets. Existing neuroimaging methods typically perform either discovery of unknown neural structure or testing of neural structure associated with mental tasks. However, testing hypotheses on the neural correlates underlying larger sets of mental tasks necessitates adequate representations for the observations. We therefore propose to blend representation modelling and task classification into a unified statistical learning problem. A multinomial logistic regression is introduced that is constrained by factored coefficients and coupled with an autoencoder. We show that this approach yields more accurate and interpretable neural models of psychological tasks in a reference dataset, as well as better generalization to other datasets.

15 BinaryConnect: Training Deep Neural Networks with binary weights during propagations

Matthieu Courbariaux matthieu.courbariaux@gmail.com
Jean-Pierre David jean-pierre.david@polymtl.ca
Polytechnique Montréal
Yoshua Bengio yoshua.bengio@gmail.com
U. Montreal

Deep Neural Networks (DNN) have achieved state-of-the-art results in a wide range of tasks, with the best results obtained with large training sets and large models. In the past, GPUs enabled these breakthroughs because of their greater computational speed. In the future, faster computation at both training and test time is likely to be crucial for further progress and for consumer applications on low-power devices. As a result, there is much interest in research and development of dedicated hardware for Deep Learning (DL). Binary weights, i.e., weights which are constrained to only two possible values (e.g. 0 or 1), would bring great benefits to specialized DL hardware by replacing many multiply-accumulate operations by simple accumulations, as multipliers are the most space and power-hungry components of the digital implementation of neural networks. We introduce BinaryConnect, a method which consists in training a DNN with binary weights during the forward and backward propagations, while retaining precision of the stored weights in which gradients are accumulated. Like other dropout schemes, we show that BinaryConnect acts as regularizer and we obtain near state-of-the-art results with BinaryConnect on the permutation-invariant MNIST. Lastly, we discuss in detail how BinaryConnect would be beneficial for DL hardware.

16 Learning to Transduce with Unbounded Memory

Edward Grefenstette etg@google.com
Karl Moritz Hermann kmh@google.com
Mustafa Suleyman mustafasul@google.com
Phil Blunsom pblunsom@google.com
Google DeepMind

Recently, strong results have been demonstrated by Deep Recurrent Neural Networks on natural language transduction problems. In this paper we explore the representational power of these models using synthetic grammars designed to exhibit phenomena similar to those found in real transduction problems such as machine translation. These experiments lead us to propose new memory-based recurrent networks that implement continuously differentiable analogues of traditional data structures such as Stacks, Queues, and DeQues. We show that these architectures exhibit superior generalisation performance to Deep RNNs and are often able to learn the underlying generating algorithms in our transduction experiments.

17 Spectral Representations for Convolutional Neural Networks

Oren Rippel rippel@math.mit.edu
MIT, Harvard
Jasper Snoek jsnoek@seas.harvard.edu
Ryan P Adams rpa@seas.harvard.edu
Harvard

Discrete Fourier transforms provide a significant speedup in the computation of convolutions in deep learning. In this work, we demonstrate that, beyond its advantages for efficient computation, the spectral domain also provides a powerful representation in which to model and train convolutional neural networks (CNNs). We employ spectral representations to introduce a number of innovations to CNN design. First, we propose spectral pooling, which performs dimensionality reduction by truncating the representation in the frequency domain. This approach preserves considerably more information per parameter than other pooling strategies and enables flexibility in the choice of pooling output dimensionality. This representation also enables a new form of stochastic regularization by randomized modification of resolution. We show that these methods achieve competitive results on classification and approximation tasks, without using any dropout or max-pooling. Finally, we demonstrate the effectiveness of complex-coefficient spectral parameterization of convolutional filters. While this leaves the underlying model unchanged, it results in a representation that greatly facilitates optimization. We observe on a variety of popular CNN configurations that this leads to significantly faster convergence during training.

18 A Theory of Decision Making Under Dynamic Context

Michael Shvartsman ms44@princeton.edu
Vaibhav Srivastava vaibhavs@princeton.edu
Jonathan D Cohen jdc@princeton.edu
Princeton University

The dynamics of simple decisions are well understood and modeled as a class of random walk models (e.g. Laming, 1968; Ratcliff, 1978; Busemeyer and Townsend, 1993; Usher and McClelland, 2001; Bogacz et al., 2006). However, most real-life decisions include a rich and dynamically-changing influence of additional information we call context. In this work, we describe a computational theory of decision making under dynamically shifting context. We show how the model generalizes the dominant existing model of fixed-context decision making (Ratcliff, 1978) and can be built up from a weighted combination of fixed-context decisions evolving simultaneously. We also show how the model generalizes re- cent work on the control of attention in the Flanker task (Yu et al., 2009). Finally, we show how the model recovers qualitative data patterns in another task of longstanding psychological interest, the AX Continuous Performance Test (Servan-Schreiber et al., 1996), using the same model parameters.

19 Bidirectional Recurrent Neural Networks as Generative Models

Mathias Berglund mathias.berglund@aalto.fi
Juha T Karhunen juha.karhunen@aalto.fi
Aalto University
Tapani Raiko tapani.raiko@aalto.fi
Aalto University, The Curious AI Company
Mikko Honkala mikko.honkala@nokia.com
Leo Kärkkäinen leo.m.karkkainen@nokia.com
Akos Vetek akos.vetek@nokia.com
Nokia Labs

Bidirectional recurrent neural networks (RNN) are trained to predict both in the positive and negative time directions simultaneously. They have not been used commonly in unsupervised tasks, because a probabilistic interpretation of the model has been difficult. Recently, two different frameworks, GSN and NADE, provide a connection between reconstruction and probabilistic modeling, which makes the interpretation possible. As far as we know, neither GSN or NADE have been studied in the context of

time series before. As an example of an unsupervised task, we study the problem of filling in gaps in high-dimensional time series with complex dynamics. Although unidirectional RNNs have recently been trained successfully to model such time series, inference in the negative time direction is non-trivial. We propose two probabilistic interpretations of bidirectional RNNs that can be used to reconstruct missing gaps efficiently. Our experiments on text data show that both proposed methods are much more accurate than unidirectional reconstructions, although a bit less accurate than a computationally complex bidirectional Bayesian inference on the unidirectional RNN. We also provide results on music data for which the Bayesian inference is computationally infeasible, demonstrating the scalability of the proposed methods.

20 Recognizing retinal ganglion cells in the dark

Emile Richard r.emile.richard@gmail.com
Georges A Goetz ggoetz@stanford.edu
E. J. Chichilnisky ej@stanford.edu
Stanford University

Real neural networks are composed of numerous distinct cell types that perform different operations on their inputs, and send their outputs to distinct targets. Therefore, a key step in understanding neural systems is to reliably distinguish cell types. An important example is the retina. Present-day techniques for identifying cell types in the retina are accurate, but very labor-intensive. Here, we develop automated classifiers for functional identification of retinal ganglion cells, the output neurons of the retina, based solely on recorded voltage patterns on a large scale array. We use per-cell classifiers based on features extracted from electrophysiological images (spatiotemporal voltage waveforms) and interspike intervals (autocorrelations). These classifiers achieve high performance in distinguishing between the major ganglion cell classes of the primate retina, but fail in achieving the same accuracy on predicting cell polarities (ON vs. OFF). We then show how to use indicators of functional coupling within populations of ganglion cells (cross-correlation) to infer these polarities with a matrix completion algorithm. Together these approaches result in nearly ideal, fully automated performance for cell type classification.

21 A Recurrent Latent Variable Model for Sequential Data

Junyoung Chung elecegg@gmail.com
Kyle Kastner kyle.kastner@umontreal.ca
Laurent Dinh laurent.dinh@umontreal.ca
Kratarth Goel kratarthgoel@gmail.com
Aaron C Courville aaron.courville@gmail.com
Yoshua Bengio yoshua.bengio@gmail.com
University of Montreal

In this paper, we explore the inclusion of random variables into the dynamics latent state of a recurrent neural networks (RNN) by combining elements of the variational autoencoder. We argue that through the use of high-level latent random variables, our variational RNN (VRNN) is able to learn to model the kind of variability observed in highly-structured sequential data (such as speech). We empirically evaluate the proposed model against related sequential models on five sequence datasets, four of speech and one of handwriting. Our results show the importance of the role random variables can play in the RNN dynamic latent state.

22 Deep Knowledge Tracing

Chris Piech piech@cs.stanford.edu
Jonathan Bassen jspencer@cs.stanford.edu
Surya Ganguli sganguli@stanford.edu
Mehran Sahami sahami@cs.stanford.edu
Leonidas J Guibas guibas@cs.stanford.edu
Jascha Sohl-Dickstein jascha@stanford.edu
Stanford University
Jonathan Huang jonathanhuang@google.com
google.com

Knowledge tracing, where a machine models the knowledge of a student as they interact with coursework, is an established and significantly unsolved problem in computer supported education. In this paper we explore the benefit of using recurrent neural networks to model student learning.This family of models have important advantages over current state of the art methods in that they do not require the explicit encoding of human domain knowledge,and have a far more flexible functional form which can capture substantially more complex student interactions.We show that these neural networks outperform the current state of the art in prediction on real student data,while allowing straightforward interpretation and discovery of structure in the curriculum.These results suggest a promising new line of research for knowledge tracing.

23 Deep Temporal Sigmoid Belief Networks for Sequence Modeling

Zhe Gan zg27@duke.edu
Chunyuan Li cl319@duke.edu
Ricardo Henao ricardo.henao@duke.edu
David E Carlson david.carlson@duke.edu
Lawrence Carin lcarin@duke.edu
Duke University

Deep dynamic generative models are developed to learn sequential dependencies in time-series data. The multi-layered model is designed by constructing a hierarchy of temporal sigmoid belief networks (TSBNs), defined as a sequential stack of sigmoid belief networks (SBNs). Each SBN has a contextual hidden state, inherited from the previous SBNs in the sequence, and is used to regulate its hidden bias. Scalable learning and inference algorithms are derived by introducing a recognition model that yields fast sampling from the variational posterior. This recognition model is trained jointly with the generative model, by maximizing its variational lower bound on the log-likelihood. Experimental results on bouncing balls, polyphonic music, motion capture, and text streams show that the proposed approach achieves state-of-the-art predictive performance, and has the capacity to synthesize various sequences.

24 Hidden Technical Debt in Machine Learning Systems

D. Sculley dsculley@google.com
Gary Holt gholt@google.com
Daniel Golovin dgg@google.com
Eugene Davydov edavydov@google.com
Todd Phillips toddphillips@google.com
Dietmar Ebner ebner@google.com
Vinay Chaudhary vchaudhary@google.com
Michael Young mwyoung@google.com
JF Crespo jfcrespo@google.com
Dan Dennison dennison@google.com
Google, Inc.

Machine learning offers a fantastically powerful toolkit for building useful complex prediction systems quickly. This paper argues it is dangerous to think of these quick wins as coming for free. Using the software engineering framework of technical debt, we find it is common to incur massive ongoing maintenance costs in real-world ML systems. We explore several ML-specific risk factors to account for in ML system design. These include boundary erosion, entanglement, hidden feedback loops, undeclared consumers, data dependencies, configuration issues, changes in the external world, and a variety of system-level anti-patterns.

25 Statistical Model Criticism using Kernel Two Sample Tests

James R Lloyd jrl44@cam.ac.uk
Zoubin Ghahramani zoubin@eng.cam.ac.uk
University of Cambridge

We propose an exploratory approach to statistical model criticism using maximum mean discrepancy (MMD) two sample tests. Typical approaches to model criticism require a practitioner to select a statistic by which to measure discrepancies between data and a statistical model. MMD two sample tests are instead constructed as an analytic maximisation over a large space of possible statistics and therefore automatically select the statistic which most shows any discrepancy. We demonstrate on synthetic data that the selected statistic, called the witness function, can be used to identify where a statistical model most misrepresents the data it was trained on. We then apply the procedure to real data where the models being assessed are restricted Boltzmann machines, deep belief networks and Gaussian process regression and demonstrate the ways in which these models fail to capture the properties of the data they are trained on.

26 Calibrated Structured Prediction

Volodymyr Kuleshov kuleshov@stanford.edu
Percy S Liang pliang@cs.stanford.edu
Stanford University

In user-facing applications,displaying predictions alongside calibrated confidence measures---probabilities that correspondto true frequency---can be as important as obtaining high accuracy. We are interested in calibration for structured prediction problemssuch as speech recognition, optical character recognition, or medical diagnosis.Calibrating probabilities in structured prediction presents new challenges:the output space is large,and users may want to issue many types of probability queries (e.g., marginals) on the structured output.We first discuss some of the subtleties with calibration in this setting,and then provide a simple recalibration method that trains a binary classifier to predict event probabilities.We then explore a range of recalibration features appropriate for the structured prediction setting,and demonstrate their efficacy on three real-world datasets.

27 A Bayesian Framework for Modeling Confidence in Perceptual Decision Making

Koosha Khalvati koosha@cs.washington.edu
Rajesh P Rao rao@cs.washington.edu
University of Washington

The degree of confidence in one's choice or decision is a critical aspect of perceptual decision making. Attempts to quantify a decision maker's confidence by measuring accuracy in a task have yielded limited success because confidence and accuracy are typically not equal. In this paper, we introduce a Bayesian framework to model confidence in perceptual decision making. We show that this model, based on partially observable Markov decision processes (POMDPs), is able to predict confidence of a decision maker based only on the data available to the experimenter. We test our model on two experiments on confidence-based decision making involving the well-known random dots motion discrimination task. In both experiments, we show that our model's predictions closely match experimental data. Additionally, our model is also consistent with other phenomena such as the hard-easy effect in perceptual decision making.

28 Dependent Multinomial Models Made Easy: Stick-Breaking with the Polya-gamma Augmentation

Scott Linderman slinderman@seas.harvard.edu
Ryan P Adams rpa@seas.harvard.edu
Harvard University
Matthew Johnson mattjj@csail.mit.edu
MIT

Many practical modeling problems involve discrete data that are best represented as draws from multinomial or categorical distributions. For example, nucleotides in a DNA sequence, children's names in a given state and year, and text documents are all commonly modeled with multinomial distributions. In all of these cases, we expect some form of dependency between the draws: the nucleotide at one position in the DNA strand may depend on the preceding nucleotides, children's names are highly correlated from year to year, and topics in text may be correlated and dynamic. These dependencies are not naturally captured by the typical Dirichlet-multinomial formulation. Here, we leverage a logistic stick-breaking representation and recent innovations in P\'{o}lya-gamma augmentation to reformulate the multinomial distribution in terms of latent variables with jointly Gaussian likelihoods, enabling us to take advantage of a host of Bayesian inference techniques for Gaussian models with minimal overhead.

29 Scalable Adaptation of State Complexity for Nonparametric Hidden Markov Models

Mike C Hughes mhughes@cs.brown.edu
Will T Stephenson wtstephe@gmail.com
Erik Sudderth sudderth@cs.brown.edu
Brown University

Bayesian nonparametric hidden Markov models are typically learned via fixed truncations of the infinite state space or local Monte Carlo proposals that make small changes to the state space. We develop an inference algorithm for the sticky hierarchical Dirichlet process hidden Markov model that scales to

big datasets by processing a few sequences at a time yet allows rapid adaptation of the state space cardinality. Unlike previous point-estimate methods, our novel variational bound penalizes redundant or irrelevant states and thus enables optimization of the state space. Our birth proposals use observed data statistics to create useful new states that escape local optima. Merge and delete proposals remove ineffective states to yield simpler models with more affordable future computations. Experiments on speaker diarization, motion capture, and epigenetic chromatin datasets discover models that are more compact, more interpretable, and better aligned to ground truth segmentations than competitors. We have released an open-source Python implementation which can parallelize local inference steps across sequences.

30 Robust Feature-Sample Linear Discriminant Analysis for Brain Disorders Diagnosis

Ehsan Adeli-Mosabbeb eadeli@unc.edu
Kim-Han Thung khthung@email.unc.edu
Le An lan004@unc.edu
Feng Shi fengshi@med.unc.edu
Dinggang Shen dinggang_shen@med.unc.edu
UNC-Chapel Hill

A wide spectrum of discriminative methods is increasingly used in diverse applications for classification or regression tasks. However, many existing discriminative methods assume that the input data is nearly noise-free, which limits their applications to solve real-world problems. Particularly for disease diagnosis, the data acquired by the neuroimaging devices are always prone to different noise factors. Robust discriminative models are somewhat scarce and only a few attempts have been made to make them robust against noise or outliers. These methods focus on detecting either the sample-outliers or feature-noises. Moreover, they usually use unsupervised de-noising procedures, or separately de-noise the training and the testing data. All these factors may induce biases in the learning process. In this paper, we propose a method based on the least-squares formulation of linear discriminant analysis, which simultaneously detects the sample-outliers and feature-noises. The proposed method operates under a semi-supervised setting, in which both labeled training and unlabeled testing data are incorporated to form the intrinsic geometry of the sample space. Therefore, the violating samples or feature values are identified as sample-outliers or feature-noises, respectively. We test our algorithm on synthetic and two brain neurodegenerative databases (Parkinson's and Alzheimer's disease databases). The results demonstrate that our method outperforms all baseline and state-of-the-art methods, in terms of both accuracy and the area under the ROC curve.

31 Learning spatiotemporal trajectories from manifold-valued longitudinal data

Jean-Baptiste Schiratti
jean-baptiste.schiratti@cmap.polytechnique.fr
Stéphanie Allassonniere
stephanie.allassonniere@polytechnique.edu
Ecole Polytechnique
Olivier Colliot olivier.colliot@upmc.fr
Université Pierre et Marie Curie (UPMC)
Stanley Durrleman stanley.durrleman@inria.fr
INRIA

We propose a Bayesian mixed-effects model to learn typical scenarios of changes from longitudinal manifold-valued data, namely repeated measurements of the same objects or individuals at several points in time. The model allows to estimate a group-average trajectory in the space of measurements. Random variations of this trajectory result from spatiotemporal transformations, which allow changes in the direction of the trajectory and in the pace at which trajectories are followed. The use of the tools of Riemannian geometry allows to derive a generic algorithm for any kind of data with smooth constraints, which lie therefore on a Riemannian manifold. Stochastic approximations of the Expectation-Maximization algorithm is used to estimate the model parameters in this highly non-linear setting.The method is used to estimate a data-driven model of the progressive impairments of cognitive functions during the onset of Alzheimer's disease. Experimental results show that the model correctly put into correspondence the age at which each individual was diagnosed with the disease, thus validating the fact that it effectively estimated a normative scenario of disease progression. Random effects provide unique insights into the variations in the ordering and timing of the succession of cognitive impairments across different individuals.

32 Hessian-free Optimization for Learning Deep Multidimensional Recurrent Neural Networks

Minhyung Cho mhyung.cho@gmail.com
Chandra Dhirshekhardhir@gmail.com
Jaehyung Lee jaehyung.lee@kaist.ac.kr
Gracenote

Multidimensional recurrent neural networks (MDRNNs) have shown a remarkable performance in the area of speech and handwriting recognition. The performance of an MDRNN is improved by further increasing its depth, and the difficulty of learning the deeper network is overcome by using Hessian-free (HF) optimization. Given that connectionist temporal classification (CTC) is utilized as an objective of learning an MDRNN for sequence labeling, the non-convexity of CTC poses a problem when applying HF to the network. As a solution, a convex approximation of CTC is formulated and its relationship with the EM algorithm and the Fisher information matrix is discussed. An MDRNN up to a depth of 15 layers is successfully trained using HF, resulting in an improved performance for sequence labeling.

33 Scalable Inference for Gaussian Process Models with Black-Box Likelihoods

Amir Dezfouli akdezfuli@gmail.com
Edwin Bonilla edwinb@cse.unsw.edu.au
Edwin V Bonilla edwinbonilla@gmail.com
The University of New South Wales

We propose a sparse method for scalable automated variational inference (AVI) in a large class of models with Gaussian process (GP) priors, multiple latent functions, multiple outputs and non-linear likelihoods. Our approach maintains the statistical efficiency property of the original AVI method, requiring only expectations over univariate Gaussian distributions to approximate the posterior with a mixture of Gaussians. Experiments on small datasets for various problems including regression, classification, Log Gaussian Cox processes, and warped GPs show that our method can perform as well as the full method under high levels of sparsity. On larger experiments using the MNIST and the SARCOS datasets we show that our method can provide superior performance to previously published scalable approaches that have been handcrafted to specific likelihood models.

34 Variational Dropout and the Local Reparameterization Trick

Diederik P Kingma dpkingma@gmail.com
Max Welling welling.max@gmail.com
University of Amsterdam
Tim Salimans salimanstim@gmail.com
Algoritmica

We explore an as yet unexploited opportunity for drastically improving the efficiency of stochastic gradient variational Bayes (SGVB) with global model parameters. Regular SGVB estimators rely on sampling of parameters once per minibatch of data, and have variance that is constant w.r.t. the minibatch size. The efficiency of such estimators can be drastically improved upon by translating uncertainty about global parameters into local noise that is independent across datapoints in the minibatch. Such reparameterizations with local noise can be trivially parallelized and have variance that is inversely proportional to the minibatch size, generally leading to much faster convergence.We find an important connection with regularization by dropout: the original Gaussian dropout objective corresponds to SGVB with local noise, a scale-invariant prior and proportionally fixed posterior variance. Our method allows inference of more flexibly parameterized posteriors; specifically, we propose \emph{variational dropout}, a generalization of Gaussian dropout, but with a more flexibly parameterized posterior, often leading to better generalization. The method is demonstrated through several experiments.

35 Infinite Factorial Dynamical Model

Isabel Valera ivalera@mpi-sws.org
MPI-SWS
Francisco J. R. Ruiz franrruiz@tsc.uc3m.es
Fernando Perez-Cruz fernando@tsc.uc3m.es
University Carlos III, Madrid
Lennart Svensson lennart.svensson@chalmers.se
Chalmers University of Technology, Göteborg

We propose the infinite factorial dynamic model (iFDM), a general Bayesian nonparametric model for source separation. Our model builds on the Markov Indian buffet process to consider a potentially unbounded number of hidden Markov chains (sources) that evolve independently according to some dynamics, in which the state space can be either discrete or continuous. For posterior inference, we develop an algorithm based on particle Gibbs with ancestor sampling that can be efficiently applied to a wide range of source separation problems. We evaluate the performance of our iFDM on four well-known applications: multitarget tracking, cocktail party, power disaggregation, and multiuser detection. Our experimental results show that our approach for source separation does not only outperform previous approaches, but it can also handle problems that were computationally intractable for existing approaches.

36 Variational Information Maximisation for Intrinsically Motivated Reinforcement Learning

Shakir Mohamed shakir@google.com
Danilo Jimenez Rezende danilor@google.com
Google DeepMind

The mutual information is a core statistical quantity that has applications in all areas of machine learning, whether this is in training of density models over multiple data modalities, in maximising the efficiency of noisy transmission channels, or when learning behaviour policies for exploration by artificial agents. Most learning algorithms that involve optimisation of the mutual information rely on the Blahut-Arimoto algorithm --- an enumerative algorithm with exponential complexity that is not suitable for modern machine learning applications. This paper provides a new approach for scalable optimisation of the mutual information by merging techniques from variational inference and deep learning. We develop our approach by focusing on the problem of intrinsically-motivated learning, where the mutual information forms the definition of a well-known internal drive known as empowerment. Using a variational lower bound on the mutual information, combined with convolutional networks for handling visual input streams, we develop a stochastic optimisation algorithm that allows for scalable information maximisation and empowerment-based reasoning directly from pixels to actions.

37 Variational inference with copula augmentation

Dustin Tran dtran@g.harvard.edu
Edo M Airoldi airoldi@fas.harvard.edu
Harvard University
David Blei david.blei@columbia.edu
Columbia University

We develop a general methodology for variational inference which preserves dependency among the latent variables. This is done by augmenting the families of distributions used in mean-field and structured approximation with copulas. Copulas allow one to separately model the dependency given a factorization of the variational distribution, and can guarantee us better approximations to the posterior as measured by KL divergence. We show that inference on the augmented distribution is highly scalable using stochastic optimization. Furthermore, the addition of a copula is generic and can be applied straightforwardly to any inference procedure using the original mean-field or structured approach. This reduces bias, sensitivity to local optima, sensitivity to hyperparameters, and significantly helps characterize and interpret the dependency among the latent variables.

38 Fast Second Order Stochastic Backpropagation for Variational Inference

Kai Fan kai.fan@duke.edu
Jeff Beck jeff.beck@duke.edu
Katherine A Heller kheller@gmail.com
Duke University
Ziteng Wang wangzt2012@gmail.com
James Kwok jamesk@cse.ust.hk
Hong Kong University of Science and Technology

We propose a second-order (Hessian or Hessian-free) based optimization method for variational inference inspired by Gaussian backpropagation, and argue that quasi-Newton optimization can be developed as well. This is accomplished by generalizing the gradient computation in stochastic backpropagation via a reparametrization trick with lower complexity. As an illustrative example, we apply this approach to the problems of Bayesian logistic regression and variational auto-encoder (VAE). Additionally, we compute bounds on the estimator variance of intractable expectations for the family of Lipschitz continuous function. Our method is practical, scalable and model free. We demonstrate our method on several real-world datasets and provide comparisons with other stochastic gradient methods to show substantial enhancement in convergence rates.

39 Rethinking LDA: Moment Matching for Discrete ICA

Anastasia Podosinnikova anastasia.podosinnikova@ens.fr
Francis Bach francis.bach@ens.fr
Simon Lacoste-Julien simon.lacoste-julien@ens.fr
INRIA

We consider moment matching techniques for estimation in Latent Dirichlet Allocation (LDA). By drawing explicit links between LDA and discrete versions of independent component analysis (ICA), we first derive a new set of cumulant-based tensors, with an improved sample complexity. Moreover, we reuse standard ICA techniques such as joint diagonalization of tensors to improve over existing methods based on the tensor power method. In an extensive set of experiments on both synthetic and real datasets, we show that our new combination of tensors and orthogonal joint diagonalization techniques outperforms existing moment matching methods.

40 Model-Based Relative Entropy Stochastic Search

Abbas Abdolmaleki abbas.a@ua.pt
Nuno Lau nunolau@ua.pt
University of Aveiro
Rudolf Lioutikov lioutikov@ias.tu-darmstadt.de
Jan R Peters mail@jan-peters.net
Gerhard Neumann geri@robot-learning.de
TU Darmstadt
Luis Pualo Reis lpreis@dsi.uminho.pt
University of Minho

Stochastic search algorithms are general black-box optimizers. Due to their ease of use and their generality, they have recently also gained a lot of attention in operations research, machine learning and policy search. Yet, these algorithms require a lot of fitness evaluations and have poor scalability with problem dimension, they may perform poorly in case of highly noisy fitness functions and they may converge prematurely. To alleviate these problems, we introduce a new surrogate-based stochastic search approach. We learn simple, quadratic surrogate models of the fitness function. As the quality of such a quadratic approximation is limited, we do not greedily exploit the learned models because the algorithm can be misled by a inaccurate optimum introduced by the surrogate. Instead, we use information theoretic constraints to bound the `distance' between the new and old data distribution while maximizing the reward. Additionally the new method is able to sustain the exploration of the search distribution to avoid premature convergence. We compare our method with the state of art black-box optimization methods, on standard uni-modal and multi-modal optimization functions, on simulated robotic planar tasks and a complex robot ball throwing task.The proposed method considerably outperforms the existing approaches.

41 On Predictive Belief Methods for Dynamical System Learning

Ahmed Hefny ahefny@cs.cmu.edu
Carlton Downey cmdowney@cs.cmu.edu
Geoffrey J Gordon ggordon@cs.cmu.edu
Carnegie Mellon UNiversity

Recently there has been substantial interest in predictive state methods for learning dynamical systems. These algorithms are popular since they often offer a good tradeoff between computational speed and statistical efficiency. Despite their desirable properties, predictive state methods can be difficult to use in practice.There is a rich literature on supervised learning methods, where we can choose from an extensive menu of models and algorithms to suit the prior beliefs we have about thefunction to be learned. In contrast predictive state dynamical system learningmethods are comparatively inflexible: It is as if we were restricted to use only linear regression instead of being allowed to choose decision trees, nonparametric regression, or the lasso. To address this problem, we propose a new view of predictive state methods in terms of instrumental-variable regression. This view allows us to construct a wide variety of dynamical system learners simply by swapping in different supervised learning methods.We demonstrate the effectiveness of our proposed methods byexperimenting with non-linear regression to learn a hidden Markov model. We show that the resulting algorithm outperforms its linear counterpart; the correctness of this algorithm follows directly from our general analysis.

42 Expectation Particle Belief Propagation

Thibaut Lienart lienart@stats.ox.ac.uk
Yee Whye Teh y.w.teh@stats.ox.ac.uk
Arnaud Doucet doucet@stats.ox.ac.uk
University of Oxford

We propose an original particle-based implementation of the Loopy Belief Propagation (LPB) algorithm for pairwise Markov Random Fields (MRF) on a continuous state space. The algorithm constructs adaptively efficient proposal distributions approximating the local beliefs at each note of the MRF. This is achieved by considering proposal distributions in the exponential family whose parameters are updated iterately in an Expectation Propagation (EP) framework. The proposed particle scheme provides consistent estimation of the LBP marginals as the number of particles increases. We demonstrate that it provides more accurate results than the Particle Belief Propagation (PBP) algorithm of Ihler and McAllester (2009) at a fraction of the computational cost and is additionally more robust empirically. The computational complexity of our algorithm at each iteration is quadratic in the number of particles. We also propose an accelerated implementation with sub-quadratic computational complexity which still provides consistent estimates of the loopy BP marginal distributions and performs almost as well as the original procedure.

43 Embedding Inference for Structured Multilabel Prediction

Farzaneh Mirzazadeh mirzazad@ualberta.ca
Siamak Ravanbakhsh ravanbakhsh@gmail.com
Dale Schuurmans dale@cs.ualberta.ca
University of Alberta
Nan Ding dingnan@google.com
Google

A key bottleneck in structured output prediction is the need for inference during training and testing, usually requiring some form of dynamic program. Rather than approximate inference or tailor a specialized inference method for a particular structure---standard responses to the scaling challenge---we instead propose to embed prediction constraints directly in the learned representation. By eliminating the need for explicit inference a more scalable approach to structured output prediction can be achieved, particularly at test time. We demonstrate this idea for multi-label prediction under subsumption and mutual exclusion constraints, where equivalence to maximum margin structured output prediction is established. Experiments demonstrate that the benefits of structured output training can still be realized even after inference has been eliminated.

44 Tractable Learning for Complex Probability Queries

Jessa Bekker jessa.bekker@cs.kuleuven.be
Jesse Davis jesse.davis@cs.kuleuven.be
KU Leuven
Guy Van den Broeck guy.vandenbroeck@cs.kuleuven.be
Arthur Choi aychoi@cs.ucla.edu
Adnan Darwiche darwiche@cs.ucla.edu
UCLA

Tractable learning's goal is to learn probabilistic graphical models where inference is guaranteed to be efficient. However, the particular class of queries that is tractable depends on the model and underlying representation. Usually this class is MPE or conditional probabilities Pr(x|y) for joint assignments x,y. We propose LearnSDD: a tractable learner that guarantees efficient inference for a broader class of queries. It simultaneously learns a Markov network and its tractable circuit representation, in order to guarantee and measure tractability. A key difference with earlier work is that LearnSDD uses Sentential Decision Diagrams (SDDs) as the tractable language instead of Arithmetic Circuits (AC). SDDs have desirable properties that are absent in more general representations such as ACs. Their additional properties enable basic primitives for Boolean circuit compilation, which allows us to support a broader class of complex probability queries, including counting, threshold, and parity, all in polytime.

45 Double or Nothing: Multiplicative Incentive Mechanisms for Crowdsourcing

Nihar Bhadresh Shah nihar@eecs.berkeley.edu
UC Berkeley
Denny Zhou dengyong.zhou@microsoft.com
MSR

Crowdsourcing has gained immense popularity in machine learning applications for obtaining large amounts of labeled data. It is however typically plagued by the problem of low quality. To address this fundamental challenge in crowdsourcing, we propose a simple payment rule ("double or nothing") to incentivize workers to answer only the questions that they are sure of and skip the rest. We show that surprisingly, under a mild and natural "no-free-lunch" requirement, this is the one and only incentive-compatible payment rule possible. We also show that among all possible incentive-compatible mechanisms (that may or may not satisfy no-free-lunch), our payment rule makes the smallest possible payment to spammers. We further extend our results to a more general setting in which workers are required to provide a quantized confidence for each question. In preliminary experiments involving over 900 worker-task pairs, we observe a considerable drop in the error rates under this unique scoring rule for the same or lower monetary expenditure.

46 Local Expectation Gradients for Black Box Variational Inference

Michalis Titsias mtitsias@aueb.gr
Athens University of Economics and Business
Miguel Lázaro-Gredilla miguel@vicarious.com
Vicarious

We introduce local expectation gradients which is a general purpose stochastic variational inference algorithm for constructing stochastic gradients by sampling from the variational distribution. This algorithm divides the problem of estimating the stochastic gradients over multiple variational parameters into smaller sub-tasks so that each sub-task explores intelligently the most relevant part of the variational distribution. This is achieved by performing an exact expectation over the single random variable that most correlates with the variational parameter of interest resulting in a Rao-Blackwellized estimate that has low variance. Our method works efficiently for both continuous and discrete random variables. Furthermore, the proposed algorithm has interesting similarities with Gibbs sampling but at the same time, unlike Gibbs sampling, can be trivially parallelized.

47 Learning with a Wasserstein Loss

Charlie Frogner frogner@mit.edu
Chiyuan Zhang pluskid@gmail.com
Hossein Mobahi hmobahi@csail.mit.edu
Tomaso A Poggio tp@ai.mit.edu
MIT
Mauricio Araya mauricio.araya@shell.com
Shell Intl. E&P Inc.

Learning to predict multi-label outputs is challenging, but in many problems there is a natural metric on the outputs that can be used to improve predictions. In this paper we develop a loss function for multi-label learning, based on the Wasserstein distance. The Wasserstein distance provides a natural notion of dissimilarity for probability measures. Although optimizing with respect to the exact Wasserstein distance is costly, recent work has described a regularized approximation that is efficiently computed. We describe efficient learning algorithms based on this regularization, extending the Wasserstein loss from probability measures to unnormalized measures. We also describe a statistical learning bound for the loss and show connections with the total variation norm and the Jaccard index. The Wasserstein loss can encourage smoothness of the predictions with respect to a chosen metric on the output space. We demonstrate this property on a real-data tag prediction problem, using the Yahoo Flickr Creative Commons dataset, achieving superior performance over a baseline that doesn't use the metric.

48 Principal Geodesic Analysis for Probability Measures under the Optimal Transport Metric

Vivien Seguy vivien.seguy@iip.ist.i.kyoto-u.ac.jp
Marco Cuturi mcuturi@i.kyoto-u.ac.jp
Kyoto University

We consider in this work the space of probability measures P(X) on a Hilbert space X endowed with the 2-Wasserstein metric. Given a finite family of probability measures in P(X), we propose an iterative approach to compute geodesic principal components that summarize efficiently that dataset. The 2-Wasserstein metric provides P(X) with a Riemannian structure and associated concepts (Fr\'echet mean, geodesics, tangent vectors) which prove crucial to follow the intuitive approach laid out by standard principal component analysis. To make our approach feasible, we propose to use an alternative parameterization of geodesics proposed by \citet[\S 9.2]{ambrosio2006gradient}. These \textit{generalized} geodesics are parameterized with two velocity fields defined on the support of the Wasserstein mean of the data, each pointing towards an ending point of the generalized geodesic. The resulting optimization problem of finding principal components is solved by adapting a projected gradient descend method. Experiment results show the ability of the computed principal components to capture axes of variability on histograms and probability measures data.

49 Fast and Accurate Inference of Plackett–Luce Models

Lucas Maystre lucas.maystre@epfl.ch
Matthias Grossglauser matthias.grossglauser@epfl.ch
EPFL

We show that the maximum-likelihood (ML) estimate of models derived from Luce's choice axiom (e.g., the Plackett-Luce model) can be expressed as the stationary distribution of a Markov chain. This conveys insight into several recently proposed spectral inference algorithms. We take advantage of this perspective and formulate a new spectral algorithm that is significantly more accurate than previous ones for the Plackett--Luce model. With a simple adaptation, this algorithm can be used iteratively, producing a sequence of estimates that converges to the ML estimate. The ML version runs faster than competing approaches on a benchmark of five datasets. Our algorithms are easy to implement, making them relevant for practitioners at large.

50 BACKSHIFT: Learning causal cyclic graphs from unknown shift interventions

Dominik Rothenhäusler rothenhaeusler@stat.math.ethz.ch
Christina Heinze heinze@stat.math.ethz.ch
Nicolai Meinshausen meinshausen@stat.math.ethz.ch
ETH Zurich
Jonas Peters jonas.peters@tuebingen.mpg.de
MPI Tübingen

We propose a simple method to learn linear causal cyclic models in the presence of latent variables. The method relies on equilibrium data of the model recorded under a specific kind of interventions (``shift interventions''). The location and strength of these interventions do not have to be known and can be estimated from the data. Our method, called BACKSHIFT, only uses second moments of the data and performs simple joint matrix diagonalization, applied to differences between covariance matrices. We give a sufficient and necessary condition for identifiability of the system, which is fulfilled almost surely under some quite general assumptions if and only if there are at least three distinct experimental settings, one of which can be pure observational data. We demonstrate the performance on some simulated data and applications in flow cytometry and financial time series.

51 Learning with Relaxed Supervision

Jacob Steinhardt jacob.steinhardt@gmail.com
Percy S Liang pliang@cs.stanford.edu
Stanford University

For partially-supervised problems with deterministic constraints between the latent variables and output, inference can be intractable no matter how simple the model family is. Even finding a single latent variable setting that satisfies the constraints may be difficult; for instance, the observed output may be the result of a latent database query or graphics program which must be inferred. For such problems, the deterministic nature of the constraints seems to clash with the statistical nature of the model. We resolve this tension by parameterizing a family of relaxations of the constraints, which can be interpreted as a marginal likelihood in an extended model family. First, we show that by simultaneously optimizing both the likelihood and the parametrized relaxation, we obtain an asymptotically consistent estimator of the model parameters. Next, we formulate joint constraints on the model and relaxation parameters that ensure efficient inference. Together, these allow us to optimize a relaxed objective that is always tractable, such that the relaxation becomes increasingly tight as the model parameters improve.

52 M-Statistic for Kernel Change-Point Detection

Shuang Li sli370@gatech.edu
Yao Xie yxie77@isye.gatech.edu
Hanjun Dai hanjundai@gatech.edu
Le Song lsong@cc.gatech.edu
Georgia Institute of Technology

Detecting the emergence of an abrupt change-point is a classic problem in statistics and machine learning. Kernel-based nonparametric statistics have been proposed for this task which make fewer assumptions on the distributions than traditional parametric approach. However, none of the existing kernel statistics has provided a computationally efficient way to characterize the extremal behavior of the statistic. Such characterization is crucial for setting the detection threshold, to control the significance level in the offline case as well as the average run length in the online case. In this paper we propose two related computationally efficient M-statistics for kernel-based change-point detection when the amount of background data is large. A novel theoretical result of the paper is the characterization of the tail probability of these statistics using a new technique based on change-of-measure. Such characterization provides us accurate detection thresholds for both offline and online cases in computationally efficient manner, without the need to resort to the more expensive simulations such as bootstrapping. We show that our methods perform well in both synthetic and real world data.

53 Fast Two-Sample Testing with Analytic Representations of Probability Measures

Kacper P Chwialkowski kacper.chwialkowski@gmail.com
Arthur Gretton arthur.gretton@gmail.com
University Collage London
Aaditya Ramdas aramdas@cs.cmu.edu
Carnegie Mellon University
Dino Sejdinovic dino.sejdinovic@gmail.com
University of Oxford

We propose a class of nonparametric two-sample tests with a cost linear in the sample size. Two tests are given, both based on an ensemble of distances between analytic functions representing each of the distributions. The first test uses smoothed empirical characteristic functions to represent the distributions, the second uses distribution embeddings in a reproducing kernel Hilbert space. Analyticity implies that differences in the distributions may be detected almost surely at a finite number of randomly chosen locations/frequencies. The new tests are consistent against a larger class of alternatives than the previous linear-time tests based on the (non-smoothed) empirical characteristic functions, while being much faster than the current state-of-the-art quadratic-time kernel-based or energy distance-based tests. Experiments on artificial benchmarks and on challenging real-world testing problems demonstrate that our tests give a better power/time tradeoff than competing approaches, and in some cases, better outright power than even the most expensive quadratic-time tests. This performance advantage is retained even in high dimensions, and in cases where the difference in distributions is not observable with low order statistics.

54 Adversarial Prediction Games for Multivariate Losses

Hong Wang hwang207@uic.edu
Wei Xing wxing3@uic.edu
Kaiser Asif kasif2@uic.edu
Brian Ziebart bziebart@uic.edu
University of Illinois at Chicago

Multivariate loss functions are used to assess performance in many modern prediction tasks, including information retrieval and ranking applications. Convex approximations are typically optimized in their place to avoid NP-hard empirical risk minimization problems. We propose to approximate the training data instead of the loss function by posing multivariate prediction as an adversarial game between a loss-minimizing prediction player and a loss-maximizing evaluation player constrained to match specified properties of training data. This avoids the non-convexity of empirical risk minimization, but game sizes are exponential in the number of predicted variables. We overcome this intractability using the double oracle constraint generation method. We demonstrate the efficiency and predictive performance of our approach on tasks evaluated using the precision at k, F-score and discounted cumulative gain.

55 Regressive Virtual Metric Learning

Michaël Perrot michael.perrot@univ-st-etienne.fr
Amaury Habrard amaury.habrard@univ-st-etienne.fr
University of Saint-Etienne

We are interested in supervised metric learning of Mahalanobis like distances. Existing approaches mainly focus on learning a new distance using similarity and dissimilarity constraints between examples. In this paper, instead of bringing closer examples of the same class and pushing far away examples of different classes we propose to move the examples with respect to virtual points. Hence, each example is brought closer to a a priori defined virtual point reducing the number of constraints to satisfy. We show that our approach admits a closed form solution which can be kernelized. We provide a theoretical analysis showing the consistency of the approach and establishing some links with other classical metric learning methods. Furthermore we propose an efficient solution to the difficult problem of selecting virtual points based in part on recent works in optimal transport. Lastly, we evaluate our approach on several state of the art datasets.

56 Halting in Random Walk Kernels

Mahito Sugiyama mahito@ar.sanken.osaka-u.ac.jp
Osaka University
Karsten Borgwardt karsten.borgwardt@bsse.ethz.ch
ETH Zurich

Random walk kernels measure graph similarity by counting matching walks in two graphs. In their most popular instance, geometric random walk kernels, longer walks of length k are downweighted by a factor λk (λ<1) to ensure convergence of the corresponding geometric series. It is known from the field of link prediction that this downweighting often leads to a phenomenon, which we refer to as halting: Longer walks are downweighted so much that the similarity score is completely dominated by the comparison of walks of length 1, which is a naive kernel between edges and vertices. We here show theoretically that halting may occur in geometric random walk kernels and quantify its impact empirically in simulated datasets and popular graph classification benchmark datasets. Our findings promise to be instrumental in future graph kernel development and applications of random walk kernels.

57 Rate-Agnostic (Causal) Structure Learning

Sergey Plis s.m.plis@gmail.com
Cynthia Freeman cfreeman@mrn.org
Vince Calhoun vcalhoun@mrn.org
The Mind Research Network
David Danks ddanks@cmu.edu
Carnegie Mellon University

Causal structure learning from time series data is a major scientific challenge. Existing algorithms assume that measurements occur sufficiently quickly; more precisely, they assume that the system and measurement timescales are approximately equal. In many scientific domains, however, measurements occur at a significantly slower rate than the underlying system changes. Moreover, the size of the mismatch between timescales is often unknown. This paper provides three distinct causal structure learning algorithms, all of which discover all dynamic graphs that could explain the observed measurement data as arising from undersampling at some rate. That is, these algorithms all learn causal structure without assuming any particular relation between the measurement and system timescales; they are thus ``rate-agnostic.'' We apply these algorithms to data from simulations. The results provide insight into the challenge of undersampling.

58 Online Prediction at the Limit of Zero Temperature

Mark Herbster m.herbster@cs.ucl.ac.uk
Stephen Pasteris s.pasteris@cs.ucl.ac.uk
University College London
Shaona Ghosh shaona.ghosh@ecs.soton.ac.uk
University of Southhampton

We design an online algorithm to classify the vertices of a graph. Underpinning the algorithm is the probability distribution of an Ising model isomorphic to the graph. Each classification is based on predicting the label with maximum marginal probability in the limit of zero-temperature with respect to the labels and vertices seen so far. Computing these classifications is unfortunately based on a #P-complete problem. This motivates us to develop an algorithm for which we give a sequential guarantee in the online mistake bound framework. Our algorithm is optimal when the graph is a tree matching the prior results in [1].For a general graph, the algorithm exploits the additional connectivity over a tree to provide a per-cluster bound. The algorithm is efficient as the cumulative time to sequentially predict all of the vertices of the graph is quadratic in the size of the graph.

59 Lifted Symmetry Detection and Breaking for MAP Inference

Tim Kopp tkopp@cs.rochester.edu
Henry Kautz kautz@cs.rochester.edu
University of Rochester
Parag Singla parags@cse.iitd.ac.in
Indian Institute of Technology

Symmetry breaking is a technique for speeding up propositional satisfiability testing by adding constraints to the formula that restrict the search space while preserving satisfiability. In this work, we extend symmetry breaking to the problem of model finding in weighted and unweighted relational theories, a class of problems that includes MAP inference in Markov Logic and similar statistical-relational languages. We then present methods for finding and breaking symmetries directly in quantified theories. We introduce term symmetries, which are induced by an evidence set and extend to symmetries over a relational theory. We give the important special case of term equivalent symmetries, showing that such symmetries can be found in low-degree polynomial time and can be completely broken by constraints that are linear in the size of the theory. We show the effectiveness of these techniques through experiments in two domains with high symmetry. Finally, we discuss connections between relational symmetry breaking and work on lifted inference in statistical-relational reasoning.

60 Bandits with Unobserved Confounders: A Causal Approach

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

The Multi-Armed Bandit problem constitutes an archetypal setting for sequential decision-making, cutting across multiple domains including engineering, business, and medicine. One of the key features in a bandit setting is the capability of exploring the environment through active interventions, which contrasts with the ability to collect passive data through the use of random sampling. The existence of unobserved confounders, namely unmeasured variables affecting both the action and the outcome variables, implies that these two data-collection modes will in general not coincide. In this paper, we show that understanding and formalizing this distinction has conceptual and algorithmic implications to the bandit setting. To demonstrate this fact, we first encode the standard bandit setting into a language capable of formally distinguishing observational and interventional distributions. We then show that the current generation of bandit algorithms are implicitly trying to maximize rewards based on the estimation of the experimental distribution, which is not always the best strategy to pursue (i.e., may yield sub-optimal performance and never converge). After this realization, we elicit what would be an adequate optimization metric that bandits should pursue when unobserved confounders are present. We finally construct an algorithm exploiting this metric and demonstrate the results through extensive simulations.

61 Sample Complexity Bounds for Iterative Stochastic Policy Optimization

Marin Kobilarov marin@jhu.edu
Johns Hopkins University

This paper is concerned with robustness analysis of decision making under uncertainty. We consider a class of iterative stochastic policy optimization problems and analyze the resulting expected performance for updating the policy at each iteration. In particular, we employ concentration-of-measure inequalities to compute future expected performance and probability of constraint violation using empirical runs. A novel inequality is derived that accounts for the possibly unbounded change-of-measure likelihood ratio resulting from the iterative policy adaptation. The approach is illustrated with a simple robot control scenario and initial steps towards applications to challenging aerial vehicle navigation problems are presented

62 Basis refinement strategies for linear value function approximation in MDPs

Gheorghe Comanici gcoman@cs.mcgill.ca
Doina Precup dprecup@cs.mcgill.ca
Prakash Panangaden prakash@cs.mcgill.ca
McGill University, Montreal

We provide a theoretical framework for analyzing basis function construction for linear value function approximation in Markov Decision Processes (MDPs). We show that important existing methods, such as Krylov bases and Bellman-error-based methods are a special case of the general framework we develop. We provide a general algorithmic framework for computing basis function refinements which “respect” the dynamics of the environment, and we derive approximation error bounds that apply for any algorithm respecting this general framework. We also show how, using ideas related to bisimulation metrics, one can translate basis refinement into a process of finding “prototypes” that are diverse enough to represent the given MDP.

63 Probabilistic Variational Bounds for Graphical Models

Qiang Liu qliu1@csail.mit.edu
John W Fisher III fisher@csail.mit.edu
MIT
Alex T Ihler ihler@ics.uci.edu
UC Irvine

Variational algorithms such as tree-reweighted belief propagation can provide deterministic bounds on the partition function, but are often loose and difficult to use in an ``any-time'' fashion, expending more computation for tighter bounds. On the other hand, Monte Carlo estimators such as importance sampling have excellent any-time behavior, but depend critically on the proposal distribution. We propose a simple Monte Carlo based inference method that augments convex variational bounds by adding importance sampling (IS). We argue that convex variational methods naturally provide good IS proposals that ``cover'' the probability of the target distribution, and reinterpret the variational optimization as designing a proposal to minimizes an upper bound on the variance of our IS estimator. This both provides an accurate estimator and enables the construction of any-time probabilistic bounds that improve quickly and directly on state of-the-art variational bounds, which provide certificates of accuracy given enough samples relative to the error in the initial bound.

64 On the Convergence of Stochastic Gradient MCMC Algorithms with High-Order Integrators

Changyou Chen cchangyou@gmail.com
Lawrence Carin lcarin@duke.edu
Duke University
Nan Ding ssnding@gmail.com
Google

Recent advances in Bayesian learning with large-scale data have witnessed emergence of stochastic gradient MCMC algorithms (SG-MCMC), such as stochastic gradient Langevin dynamics (SGLD), stochastic gradient Hamiltonian MCMC (SGHMC), and the stochastic gradient thermostat. While finite-time convergence properties of the SGLD with a 1st-order Euler integrator have recently been studied, corresponding theory for general SG-MCMCs has not been explored. In this paper we consider general SG-MCMCs with high-order integrators, and develop theory to analyze finite-time convergence properties and their asymptotic invariant measures. Our theoretical results show faster convergence rates and more accurate invariant measures for SG-MCMCs with higher-order integrators. For example, with the proposed efficient 2nd-order symmetric splitting integrator, the mean square error (MSE) of the posterior average for the SGHMC achieves an optimal convergence rate of L−4/5 at L iterations, compared to L−2/3 for the SGHMC and SGLD with 1st-order Euler integrators. Furthermore, convergence results of decreasing-step-size SG-MCMCs are also developed, with the same convergence rates as their fixed-step-size counterparts for a specific decreasing sequence. Experiments on both synthetic and real datasets verify our theory, and show advantages of the proposed method in two large-scale real applications.

65 An Active Learning Framework using Sparse-Graph Codes for Sparse Polynomials and Graph Sketching

Xiao Li xiaoli@berkeley.edu
Kannan Ramchandran kannanr@berkeley.edu
UC Berkeley

Let f:{−1,1}n→R be an n-variate polynomial consisting of 2n monomials over the boolean field, in which s monomial coefficients are non-zero. We introduce a new active learning framework by designing queries to f based on modern coding theory. The key is to relax the worst-case assumption on sparsity for an ensemble-average setting, where the polynomial is assumed to be drawn uniformly at random from the ensemble of all polynomials (of a given degree n and sparsity s). We show how this relaxation allows us to leverage powerful tools from modern coding theory, specifically, the design and analysis of {\it sparse-graph codes}, such as Low-Density-Parity-Check (LDPC) codes, which represent the state-of-the-art in modern communication systems. This leads to significant savings in both query complexity and computational speed. More significantly, we show how this leads to exciting, and to the best of our knowledge, largely unexplored intellectual connections between learning and coding. Our framework succeeds with high probability with respect to the polynomial ensemble with sparsity up to s=O(2δn) for any δ∈(0,1), where f is exactly learned using O(ns) queries in time O(nslogs), even if the queries are perturbed by {\it unbounded Gaussian noise}. We also show that if each active monomial involves at most d variables, the polynomial can be learned using O(dslogslogn) noiseless queries in time O(dslog2slogn). This solution is particularly applicable to graph sketching, which is the problem of inferring sparse graphs by querying graph cuts. Our result approaches the optimal query complexity for learning hidden graphs, where experiments on real datasets show significant reductions in the run-time and query complexity compared with competitive schemes.

66 Discrete Rényi Classifiers

Meisam Razaviyayn meisamr@stanford.edu
Farzan Farnia farnia@stanford.edu
David Tse dntse@stanford.edu
Stanford University

Consider the binary classification problem of predicting a target variable Y from a discrete feature vector X = (X1,...,Xd). When the probability distribution P(X,Y) is known, the optimal classifier, leading to the minimum misclassification rate, is given by the Maximum A-posteriori Probability (MAP) decision rule. However, in practice, estimating the complete joint distribution P(X,Y) is computationally and statistically impossible for large values of d. Therefore, an alternative approach is to first estimate some low order marginals of the joint probability distribution P(X,Y) and then design the classifier based on the estimated low order marginals. This approach is also helpful when the complete training data instances are not available due to privacy concerns. In this work, we consider the problem of designing the optimum classifier based on some estimated low order marginals of (X,Y). We prove that for a given set of marginals, the minimum Hirschfeld-Gebelein-R´enyi (HGR) correlation principle introduced in [1] leads to a randomized classification rule which is shown to have a misclassification rate no larger than twice the misclassification rate of the optimal classifier. Then, we show that under a separability condition, the proposed algorithm is equivalent to a randomized linear regression approach which naturally results in a robust feature selection method selecting a subset of features having the maximum worst case HGR correlation with the target variable. Our theoretical upper-bound is similar to the recent Discrete Chebyshev Classifier (DCC) approach [2], while the proposed algorithm has significant computational advantages since it only requires solving a least square optimization problem. Finally, we numerically compare our proposed algorithm with the DCC classifier and show that the proposed algorithm results in better misclassification rate over various UCI data repository datasets.

67 GAP Safe screening rules for sparse multi-task and multi-class models

Eugene Ndiaye eugene.ndiaye@telecom-paristech.fr
Olivier Fercoq olivier.fercoq@telecom-paristech.fr
Alexandre Gramfort alexandre.gramfort@telecom-paristech.fr
Joseph Salmon joseph.salmon@telecom-paristech.fr
Télécom ParisTech

High dimensional regression benefits from sparsity promoting regularizations. Screening rules leverage the known sparsity of the solution by ignoring some variables in the optimization, hence speeding up solvers. When the procedure is proven not to discard features wrongly the rules are said to be safe. In this paper we derive new safe rules for generalized linear models regularized with L1 and L1/L2 norms. The rules are based on duality gap computations and spherical safe regions whose diameters converge to zero. This allows to discard safely more variables, in particular for low regularization parameters. The GAP Safe rule can cope with any iterative solver and we illustrate its performance on coordinate descent for multi-task Lasso, binary and multinomial logistic regression, demonstrating significant speed ups on all tested datasets with respect to previous safe rules.

68 Decomposition Bounds for Marginal MAP

Wei Ping weiping.thu@gmail.com
Alex T Ihler ihler@ics.uci.edu
UC Irvine
Qiang Liu qliu1@csail.mit.edu
MIT

Marginal MAP inference involves making MAP predictions in systems defined with latent variables or missing information. It is significantly more difficult than pure marginalization and MAP tasks, for which a large class of efficient and convergent variational algorithms, such as dual decomposition, exist. In this work, we generalize dual decomposition to a generic powered-sum inference task, which includes marginal MAP, along with pure marginalization and MAP, as special cases. Our method is based on a block coordinate descent algorithm on a new convex decomposition bound, that is guaranteed to converge monotonically, and can be parallelized efficiently. We demonstrate our approach on various inference queries over real-world problems from the UAI approximate inference challenge, showing that our framework is faster and more reliable than previous methods.

69 Anytime Influence Bounds and the Explosive Behavior of Continuous-Time Diffusion Networks

Kevin Scaman scaman@cmla.ens-cachan.fr
Rémi Lemonnier lemonnier@cmla.ens-cachan.fr
Nicolas Vayatis vayatis@cmla.ens-cachan.fr
ENS Cachan - CMLA

The paper studies transition phenomena in information cascades observed along a diffusion process over some graph. We introduce the Laplace Hazard matrix and show that its spectral radius fully characterizes the dynamics of the contagion both in terms of influence and of explosion time. Using this concept, we prove tight non-asymptotic bounds for the influence of a set of nodes, and we also provide an in-depth analysis of the critical time after which the contagion becomes super-critical. Our contributions include formal definitions and tight lower bounds of critical explosion time. We illustrate the relevance of our theoretical results through several examples of information cascades used in epidemiology and viral marketing models. Finally, we provide a series of numerical experiments for various types of networks which confirm the tightness of the theoretical bounds.

70 Estimating Mixture Models via Mixtures of Polynomials

Sida Wang sidaw@cs.stanford.edu
Arun Tejasvi Chaganty chaganty@cs.stanford.edu
Percy S Liang pliang@cs.stanford.edu
Stanford University

Mixture modeling is a general technique for making any simple model more expressive through weighted combination. This generality and simplicity in part explains the success of the Expectation Maximization (EM) algorithm, in which updates are easy to derive for a wide class of mixture models. However, the likelihood of a mixture model is non-convex, so EM has no global convergence guarantees. Recently, method of moments approaches offer global guarantees for some mixture models, but they do not extend easily to the range of mixture models that exist. In this work, we present Polymom, an unifying framework based on method of moments in which estimation procedures are easily derivable, just as in EM. Polymom is applicable when the moments of a single mixture component are polynomials of the parameters. The key observation is that the moments of the mixture model are a mixture of these polynomials, which allows us to cast estimation as a General Moment Problem. We solve its relaxations using semidefinite optimization, and then extract parameters using ideas from computer algebra. This framework allows us to draw insights and apply tools from convex optimization, computer algebra and the theory of moments to study important issues in statistical estimation such as parameter constraints and noise. Simulations show good empirical performance on several models.

71 Robust Gaussian Graphical Modeling with the Trimmed Graphical Lasso

Eunho Yang yangeh@gmail.com
Aurelie C Lozano aclozano@us.ibm.com
IBM Research

Gaussian Graphical Models (GGMs) are popular tools for studying network structures. However, many modern applications such as gene network discovery and social interactions analysis often involve high-dimensional noisy data with outliers or heavier tails than the Gaussian distribution. In this paper, we propose the Trimmed Graphical Lasso for robust estimation of sparse GGMs. Our method guards against outliers by an implicit trimming mechanism akin to the popular Least Trimmed Squares method used for linear regression. We provide a rigorous statistical analysis of our estimator in the high-dimensional setting. In contrast, existing approaches for robust sparse GGMs estimation lack statistical guarantees. Our theoretical results are complemented by experiments on simulated and real gene expression data which further demonstrate the value of our approach.

72 Matrix Completion from Fewer Entries: Spectral Detectability and Rank Estimation

Alaa Saade alaa.saade@m4x.org
Florent Krzakala florent.krzakala@ens.fr
Ecole Normale Superieure CNRS
Lenka Zdeborová lenka.zdeborova@gmail.com
CEA

The completion of low rank matrices from few entries is a task with many practical applications. We consider here two aspects of this problem: detectability, i.e. the ability to estimate the rank r reliably from the fewest possible random entries, and performance in achieving small reconstruction error. We propose a spectral algorithm for these two tasks called MaCBetH (for Matrix Completion with the Bethe Hessian). The rank is estimated as the number of negative eigenvalues of the Bethe Hessian matrix, and the corresponding eigenvectors are used as initial condition for the minimization of the discrepancy between the estimated matrix and the revealed entries. We analyze the performance in a random matrix setting using results from the statistical mechanics of the Hopfield neural network, and show in particular that MaCBetH efficiently detects the rank r of a large n×m matrix from C(r)rnm−−−√ entries, where C(r) is a constant close to 1. We also evaluate the corresponding root-mean-square error empirically and show that MaCBetH compares favorably to other existing approaches.

73 Robust PCA with compressed data

Wooseok Ha haywse@gmail.com
Rina Foygel Barber rina@uchicago.edu
University of Chicago

The robust principal component analysis (RPCA) problem seeks to separate low-rank trends from sparse outlierswithin a data matrix, that is, to approximate a n×d matrix D as the sum of a low-rank matrix L and a sparse matrix S.We examine the robust principal component analysis (RPCA) problem under data compression, wherethe data Y is approximately given by (L+S)·C, that is, a low-rank + sparse data matrix that has been compressed to size n×m (with m substantially smaller than the original dimension d) via multiplication witha compression matrix C. We give a convex program for recovering the sparse component S along with the compressed low-rank component L·C, along with upper bounds on the error of this reconstructionthat scales naturally with the compression dimension m and coincides with existing results for the uncompressedsetting m=d. Our results can also handle error introduced through additive noise or through missing data. The scaling of dimension, compression, and signal complexity in our theoretical results is verified empirically through simulations, and we also apply our method to a data set measuring chlorine concentration acrossa network of sensors, to test its performance in practice.

74 Mixed Robust/Average Submodular Partitioning: Fast Algorithms, Guarantees, and Applications

Kai Wei kaiwei@uw.edu
Rishabh K Iyer rkiyer@u.washington.edu
Shengjie Wang wangsj@cs.washington.edu
Wenruo Bai wrbai@uw.edu
Jeff A Bilmes bilmes@ee.washington.edu
University of Washington, Seattle

We investigate two novel mixed robust/average-case submodular data partitioning problems that we collectively call Submodular Partitioning. These problems generalize purely robust instances of the problem, namely max-min submodular fair allocation (SFA) and \emph{min-max submodular load balancing} (SLB), and also average-case instances, that is the submodular welfare problem (SWP) and submodular multiway partition (SMP). While the robust versions have been studied in the theory community, existing work has focused on tight approximation guarantees, and the resultant algorithms are not generally scalable to large real-world applications. This contrasts the average case instances, where most of the algorithms are scalable. In the present paper, we bridge this gap, by proposing several new algorithms (including greedy, majorization-minimization, minorization-maximization, and relaxation algorithms) that not only scale to large datasets but that also achieve theoretical approximation guarantees comparable to the state-of-the-art. We moreover provide new scalable algorithms that apply to additive combinations of the robust and average-case objectives, and also allow for more general partitioning constraints (k-cover partitionings). We show that these problems have many applications in machine learning (ML), including data partitioning and load balancing for distributed ML, data clustering, and image segmentation. We empirically demonstrate the efficacy of our algorithms on real-world problems involving data partitioning for distributed optimization (of convex, nearest neighbor, and deep neural network objectives), and also purely unsupervised image segmentation.

75 Subspace Clustering with Irrelevant Features via Robust Dantzig Selector

Chao Qu a0117143@u.nus.edu
Huan Xu mpexuh@nus.edu.sg
National University of Singapore

This paper considers the subspace clustering problem where the data contains irrelevant or corrupted features. We propose a method termed ``robust Dantzig selector'' which can successfully identify the clustering structure even with the presence of irrelevant features. The idea is simple yet powerful: we replace the inner product by its robust counterpart, which is insensitive to the irrelevant features given an upper bound of the number of irrelevant features. We establish theoretical guarantees for the algorithm to identify the correct subspace, and demonstrate the effectiveness of the algorithm via numerical simulations. To the best of our knowledge, this is the first method developed to tackle subspace clustering with irrelevant features.

76 A class of network models recoverable by spectral clustering

Yali Wan yaliwan@uw.edu
Marina Meila mmp@stat.washington.edu
University of Washington

There has been a lot of recent advances in the recovery of communities in networks, under ``block-model'' assumptions. In particular, advances in recovering communities by spectral clustering algorithms. These have been extended to models including node-specific propensities. In this paper, we argue that one can further expand the model class for which recovery by spectral clustering is possible, and describe a model that subsumes a number of existing models , which we call the KPFM model. We show that under the KPFM model, the communities can be recovered with small error, w.h.p. Our results correspond to what termed the ``weak recovery'' regime, in which w.h.p. the fraction of nodes that are mislabeled is o(1)

77 Monotone k-Submodular Function Maximization with Size Constraints

Naoto Ohsaka ohsaka@is.s.u-tokyo.ac.jp
The University of Tokyo
Yuichi Yoshida yyoshida@nii.ac.jp
National Institute of Informatics

A k-submodular function is a generalization of a submodular function, where the input consists of k disjoint subsets, instead of a single subset, of the domain.Many machine learning problems, including influence maximization with k kinds of topics and sensor placement with k kinds of sensors, can be naturally modeled as the problem of maximizing monotone k-submodular functions.In this paper, we give constant-factor approximation algorithms for maximizing monotone k-submodular functions subject to several size constraints.The running time of our algorithms are almost linear in the domain size.We experimentally demonstrate that our algorithms outperform baseline algorithms in terms of the solution quality.

78 Smooth and Strong: MAP Inference with Linear Convergence

Ofer Meshi meshi@ttic.edu
Mehrdad Mahdavi mahdavi@ttic.edu
TTI Chicago
Alex Schwing aschwing@cs.toronto.edu
University of Toronto

Maximum a-posteriori (MAP) inference is an important task for many applications. Although the standard formulation gives rise to a hard combinatorial optimization problem, several effective approximations have been proposed and studied in recent years. We focus on linear programming (LP) relaxations, which have achieved state-of-the-art performance in many applications. However, optimization of the resulting program is in general challenging due to non-smoothness and complex non-separable constraints.Therefore, in this work we study the benefits of augmenting the objective function of the relaxation with strong convexity. Specifically, we introduce strong convexity by adding a quadratic term to the LP relaxation objective. We provide theoretical guarantees for the resulting programs, bounding the difference between their optimal value and the original optimum. Further, we propose suitable optimization algorithms and analyze their convergence.

79 StopWasting My Gradients: Practical SVRG

Reza Harikandeh rezababa@cs.ubc.ca
Mohamed Osama Ahmed moahmed@cs.ubc.ca
Scott Sallinen scotts@ece.ubc.ca
Mark Schmidt schmidtm@cs.ubc.ca
University of British Columbia
Alim Virani alim.virani@gmail.com
Jakub Konečný j.konecny@sms.ed.ac.uk

We present and analyze several strategies for improving the performance ofstochastic variance-reduced gradient (SVRG) methods. We first show that theconvergence rate of these methods can be preserved under a decreasing sequenceof errors in the control variate, and use this to derive a variants of SVRG that usesa growing-batch strategy to reduce the number of gradient calculations requiredin the early iterations. We further (i) show how to exploit support vectors to reducethe number of gradient computations in the later iterations, (ii) prove that thecommonly–used regularized SVRG iteration is justified and improves the convergencerate, (iii) propose a better mini-batch selection strategy, and (iv) considerthe generalization error the method.

80 Spectral Norm Regularization of Orthonormal Representations for Graph Transduction

Rakesh Shivanna rakeshsmysore@gmail.com
Google Inc.
Bibaswan K Chatterjee
bibaswan.chatterjee@csa.iisc.ernet.in
Raman Sankaran raman.cse@gmail.com
Chiranjib Bhattacharyya chiru@csa.iisc.ernet.in
Indian Institute of Science
Francis Bach francis.bach@ens.fr
INRIA - ENS

For learning labels on vertices of a graph, Ando and Zhang~\cite{ando} argued that embedding a graph on a unit sphere leads to better generalization. However, the choice of optimal embedding and an efficient algorithm to compute the same remains an open issue. In this paper, we show that orthonormal representations, a class of unit-sphere graph embeddings are PAC learnable. The existing analysis does not directly apply, and we propose an alternative PAC-based bounds which do not depend on the VC dimension of the underlying function class but are related to the famous \lo~ϑ function. The main contribution of the paper is \form, a spectral regularized orthonormal embedding for graph transduction, derived from the PAC bound. \form~is posed as a non-smooth convex function over an \emph{elliptope}. These problems are usually solved as semi-definite programs (SDPs) with time complexity O(n6). Projecting on elliptopes is a demanding problem and remains the main computational bottleneck. We tackle this challenge by presenting Infeasible Inexact proximal~(IIP) method: an Inexact proximal method which performs subgradient procedure on an approximate projection, not necessarily feasible. We show that IIP converges to the optimal with the rate O(1T√). IIP is generally applicable and we use it to compute \form~where the approximate projection step is computed by an accelerated gradient descent procedure, such as FISTA. We show that if number of iterations in the FISTA step are at least T√, the IIP method yields the same O(1T√)~convergence. The proposed algorithm easily scales to 1000's vertices, while the standard SDP computation does not scale beyond few hundred vertices. As an interesting aside, we can easily compute \lo~ϑ, and show similar speedups. Furthermore, the analysis presented here easily extends to the multiple graphs setting.

81 Differentially Private Learning of Structured Discrete Distributions

Ilias Diakonikolas ilias.d@ed.ac.uk
University of Edinburgh
Moritz Hardt m@mrtz.org
Google
Ludwig Schmidt ludwigs@mit.edu
MIT

We investigate the problem of learning an unknown probability distribution over a discrete population from random samples. Our goal is to design efficient algorithms that simultaneously achieve low error in total variation norm while guaranteeing Differential Privacy to the individuals of the population.We describe a general approach that yields near sample-optimal and computationally efficient differentially private estimators for a wide range of well-studied and natural distribution families. Our theoretical results show that for a wide variety of structured distributions there exist private estimation algorithms that are nearly as efficient---both in terms of sample size and running time---as their non-private counterparts. We complement our theoretical guarantees with an experimental evaluation. Our experiments illustrate the speed and accuracy of our private estimators on both synthetic mixture models, as well as a large public data set.

82 Robust Portfolio Optimization

Huitong Qiu hqiu7@jhu.edu
Fang Han fhan@jhu.edu
Johns Hopkins University
Han Liu hanliu@princeton.edu
Princeton University
Brian Caffo bcaffo@jhu.edu

We propose a robust portfolio optimization approach based on quantile statistics. The proposed method is robust to extreme

events in asset returns, and accommodates large portfolios under limited historical data. Specifically, we show that the risk of the estimated portfolio converges to the oracle optimal risk with parametric rate under weakly dependent asset returns. The theory does not rely on higher order moment assumptions, thus allowing for heavy-tailed asset returns. Moreover, the rate of convergence quantifies that the size of the portfolio under management is allowed to scale exponentially with the sample size of the historical data. The empirical effectiveness of the proposed method is demonstrated under both synthetic and real stock data. Our work extends existing ones by achieving robustness in high dimensions, and by allowing serial dependence.

83 Bayesian Optimization with Exponential Convergence

Kenji Kawaguchi kawaguch@mit.edu
Leslie Kaelbling lpk@csail.mit.edu
Tomás Lozano-Pérez tlp@csail.mit.edu
MIT

This paper presents a Bayesian optimization method with exponential convergence without the need of auxiliary optimization and without the delta-cover sampling. Most Bayesian optimization methods require auxiliary optimization: an additional non-convex global optimization problem, which can be time-consuming and hard to implement in practice. Also, the existing Bayesian optimization method with exponential convergence requires access to the delta-cover sampling, which was considered to be impractical. Our approach eliminates both requirements and achieves an exponential convergence rate.

84 Fast Randomized Kernel Ridge Regression with Statistical Guarantees

Ahmed Alaoui elalaoui@berkeley.edu
Michael W Mahoney mmahoney@stat.berkeley.edu
UC Berkeley

One approach to improving the running time of kernel-based methods is to build a small sketch of the kernel matrix and use it in lieu of the full matrix in the machine learning task of interest. Here, we describe a version of this approach that comes with running time guarantees as well as improved guarantees on its statistical performance.By extending the notion of \emph{statistical leverage scores} to the setting of kernel ridge regression, we are able to identify a sampling distribution that reduces the size of the sketch (i.e., the required number of columns to be sampled) to the \ emph{effective dimensionality} of the problem. This latter quantity is often much smaller than previous bounds that depend on the \emph{maximal degrees of freedom}. We give an empirical evidence supporting this fact. Our second contribution is to present a fast algorithm to quickly compute coarse approximations to thesescores in time linear in the number of samples. More precisely, the running time of the algorithm is O(np2) with p only depending on the trace of the kernel matrix and the regularization parameter. This is obtained via a variant of squared length sampling that we adapt to the kernel setting. Lastly, we discuss how this new notion of the leverage of a data point captures a fine notion of the difficulty of the learning problem.

85 Taming the Wild: A Unified Analysis of Hogwild-Style Algorithms

Christopher M De Sa cdesa@stanford.edu
Kunle Olukotun kunle@stanford.edu
Christopher Ré chrismre@stanford.edu
Chris Ré chrismre@cs.stanford.edu
Stanford
Ce Zhang czhang@cs.wisc.edu
Wisconsin

Stochastic gradient descent (SGD) is a ubiquitous algorithm for a variety of machine learning problems. Researchers and industry have developed several techniques to optimize SGD's runtime performance, including asynchronous execution and reduced precision. Our main result is a martingale-based analysis that enables us to capture the rich noise models that may arise from such techniques. Specifically, we useour new analysis in three ways: (1) we derive convergence rates for the convex case (Hogwild) with relaxed assumptions on the sparsity of the problem; (2) we analyze asynchronous SGD algorithms for non-convex matrix problems including matrix completion; and (3) we design and analyze an asynchronous SGD algorithm, called Buckwild, that uses lower-precision arithmetic. We show experimentally that our algorithms run efficiently for a variety of problems on modern hardware.

86 Beyond Convexity: Stochastic Quasi-Convex Optimization

Elad Hazan ehazan@cs.princeton.edu
Princeton University
Kfir Levy kfiryehud@gmail.com
Technion
Shai Shalev-Shwartz shais@cs.huji.ac.il
Hebrew University

Stochastic convex optimization is a basic and well studied primitive in machine learning. It is well known that convex and Lipschitz functions can be minimized efficiently using Stochastic Gradient Descent (SGD).The Normalized Gradient Descent (NGD) algorithm, is an adaptation of Gradient Descent, which updates according to the direction of the gradients, rather than the gradients themselves. In this paper we analyze a stochastic version of NGD and prove its convergence to a global minimum for a wider class of functions: we require the functions to be quasi-convex and locally-Lipschitz. Quasi-convexity broadens the concept of unimodality to multidimensions and allows for certain types of saddle points, which are a known hurdle for first-order optimization methods such as gradient descent. Locally-Lipschitz functions are only required to be Lipschitz in a small region around the optimum. This assumption circumvents gradient explosion, which is another known hurdle for gradient descent variants. Interestingly, unlike the vanilla SGD algorithm, the stochastic normalized gradient descent algorithm provably requires a minimal minibatch size.

87 On the Limitation of Spectral Methods: From the Gaussian Hidden Clique Problem to Rank-One Perturbations of Gaussian Tensors

Andrea Montanari montanari@stanford.edu
Stanford
Daniel Reichman daniel.reichman@gmail.com
Cornell University
Ofer Zeitouni ofer.zeitouni@gmail.com
Weizmann Institute and Courant Institute

We consider the following detection problem: given a realization of asymmetric matrix X of dimension n, distinguish between the hypothesisthat all upper triangular variables are i.i.d. Gaussians variableswith mean 0 and variance 1 and the hypothesis that there is aplanted principal submatrix B of dimension L for which all upper triangularvariables are i.i.d. Gaussians with mean 1 and variance 1, whereasall other upper triangular elements of X not in B are i.i.d.Gaussians variables with mean 0 and variance 1. We refer to this asthe `Gaussian hidden clique problem'. When $L=(1+\epsilon)n\sqrt{}$ ($\epsilon>0$), it is possible to solve thisdetection problem with probability $1-o_n(1)$ by computing thespectrum of X and considering the largest eigenvalue of X.We prove that when$L<(1-\epsilon)n\sqrt{}$ no algorithm that examines only theeigenvalues of Xcan detect the existence of a hiddenGaussian clique, with error probability vanishing as $n\to\infty$.The result above is an immediate consequence of a more general result on rank-oneperturbations of k-dimensional Gaussian tensors.In this context we establish a lower bound on the criticalsignal-to-noise ratio below which a rank-one signal cannot be detected.

88 Regularized EM Algorithms: A Unified Framework and Statistical Guarantees

Xinyang Yi yixy@utexas.edu
Constantine Caramanis constantine@utexas.edu
UT Austin

Latent models are a fundamental modeling tool in machine learning applications, but they present significant computational and analytical challenges. The popular EM algorithm and its variants, is a much used algorithmic tool; yet our rigorous understanding of its performance is highly incomplete. Recently, work in [1] has demonstrated that for an important class of problems, EM exhibits linear local convergence. In the high-dimensional setting, however, the M-step may not be well defined. We address precisely this setting through a unified treatment using regularization. While regularization for high-dimensional problems is by now well understood, the iterative EM algorithm requires a careful balancing of making progress towards the solution while identifying the right structure (e.g., sparsity or low-rank). In particular, regularizing the M-step using the state-of-the-art high-dimensional prescriptions (e.g., \`a la [19]) is not guaranteed to provide this balance. Our algorithm and analysis are linked in a way that reveals the balance between optimization and statistical errors. We specialize our general framework to sparse gaussian mixture models, high-dimensional mixed regression, and regression with missing variables, obtaining statistical guarantees for each of these examples.

89 Black-box optimization of noisy functions with unknown smoothness

jean-bastien grill jean-bastien.grill@inria.fr
Michal Valko michal.valko@inria.fr
INRIA Lille - Nord Europe
Remi Munos remi.munos@inria.fr
Google DeepMind

We study the problem of black box optimization of a function f given evaluations perturbed by noise. The function is assumed to be locally smooth around one of its global optima, but this smoothness is not assumed to be known. Our contribution is to introduce an adaptive optimization algorithm, called POO, parallel optimistic optimization, which performs almost as well as the best known algorithms requiring the knowledge of the function smoothness, for a larger class of functions than what was previously considered (especially for functions that are 'difficult' to optimize, in a precise sense). We provide a finite-time analysis of the performance of POO which shows that its optimization error after n function evaluations is at most a factor of sqrt(ln n) away from the error of the best known optimization algorithms using the knowledge of the function smoothness.

90 Combinatorial Cascading Bandits

Branislav Kveton kveton@adobe.com
Adobe Research
Zheng Wen zhengwen@yahoo-inc.com
Yahoo Labs
Azin Ashkan azin.ashkan@technicolor.com
Technicolor Research
Csaba Szepesvari szepesva@cs.ualberta.ca
University of Alberta

We propose combinatorial cascading bandits, a class of partial monitoring problems where at each step a learning agent chooses a tuple of ground items subject to constraints and receives a reward if and only if the weights of all chosen items are one. The weights of the items are binary, stochastic, and drawn independently of each other. The agent observes the index of the first chosen item whose weight is zero. This observation model arises in network routing, for instance, where the learning agent may only observe the first link in the routing path which is down, and blocks the path. We propose a UCB-like algorithm for solving our problems, CombCascade; and prove gap-dependent and gap-free upper bounds on its n-step regret. Our proofs build on recent work in stochastic combinatorial semi-bandits but also address two novel challenges of our setting, a non-linear reward function and partial observability. We evaluate CombCascade on two real-world problems and show that it performs well even when our modeling assumptions are violated. We also demonstrate that our setting requires a new learning algorithm.

91 Adaptive Primal-Dual Splitting Methods for Statistical Learning and Image Processing

Tom Goldstein tomg@cs.umd.edu
University of Maryland
Min Li limin@seu.edu.cn
Southeast University
Xiaoming Yuan xmyuan@hkbu.edu.hk
Hong Kong Baptist University

The alternating direction method of multipliers (ADMM) is an important tool for solving complex optimization problems, but it involves minimization sub-steps that are often difficult to solve efficiently. The Primal-Dual Hybrid Gradient (PDHG) method is a powerful alternative that often has simpler substeps than ADMM, thus producing lower complexity solvers. Despite the flexibility of this method, PDHG is often impractical because it requires the careful choice of multiple stepsize parameters. There is often no intuitive way to choose these parameters to maximize efficiency, or even achieve convergence. We propose self-adaptive stepsize rules that automatically tune PDHG parameters for optimal convergence. We rigorously analyze our methods, and identify convergence rates. Numerical experiments show that adaptive PDHG has strong advantages over non-adaptive methods in terms of both efficiency and simplicity for the user.

92 Sum-of-Squares Lower Bounds for Sparse PCA

Tengyu Ma tengyu@cs.princeton.edu
Princeton University
Avi Wigderson avi@ias.edu
Institute for Advanced Study

This paper establishes a statistical versus computational trade-offfor solving a basic high-dimensional machine learning problem via a basic convex relaxation method. Specifically, we consider the {\em Sparse Principal Component Analysis} (Sparse PCA) problem, and the family of {\em Sum-of-Squares} (SoS, aka Lasserre/Parillo) convex relaxations. It was well known that in large dimension p, a planted k-sparse unit vector can be {\em in principle} detected using only n≈klogp (Gaussian or Bernoulli) samples, but all {\em efficient} (polynomial time) algorithms known require n≈k2 samples. It was also known that this quadratic gap cannot be improved by the the most basic {\em semi-definite} (SDP, aka spectral) relaxation, equivalent to a degree-2 SoS algorithms. Here we prove that also degree-4 SoS algorithms cannot improve this quadratic gap. This average-case lower bound adds to the small collection of hardness results in machine learning for this powerful family of convex relaxation algorithms. Moreover, our design of moments (or ``pseudo-expectations'') for this lower bound is quite different than previous lower bounds. Establishing lower bounds for higher degree SoS algorithms for remains a challenging problem.

93 Online Gradient Boosting

Alina Beygelzimer beygel@yahoo-inc.com
Satyen Kale satyen.kale@gmail.com
Yahoo Labs
Elad Hazan ehazan@cs.princeton.edu
Haipeng Luo haipengl@cs.princeton.edu
Princeton University

We extend the theory of boosting for regression problems to the online learning setting. Generalizing from the batch setting for boosting, the notion of a weak learning algorithm is modeled as an online learning algorithm with linear loss functions that competes with a base class of regression functions, while a strong learning algorithm is an online learning algorithm with smooth convex loss functions that competes with a larger class of regression functions. Our main result is an online gradient boosting algorithm which converts a weak online learning algorithm into a strong one where the larger class of functions is the linear span of the base class. We also give a simpler boosting algorithm that converts a weak online learning algorithm into a strong one where the larger class of functions is the convex hull of the base class, and prove its optimality.

94 Regularization-Free Estimation in Trace Regression with Symmetric Positive Semidefinite Matrices

Martin Slawski martin.slawski@rutgers.edu
Ping Li pingli@stat.rutgers.edu
Rugters University
Matthias Hein hein@cs.uni-sb.de
Saarland University

Trace regression models have received considerable attention in the context of matrix completion, quantum state tomography, and compressed sensing. Estimation of the underlying matrix from regularization-based approaches promoting low-rankedness, notably nuclear norm regularization, have enjoyed great popularity. In this paper, we argue that such regularization may no longer be necessary if the underlying matrix is symmetric positive semidefinite (spd) and the design satisfies certain conditions. In this situation, simple least squares estimation subject to an spd constraint may perform as well as regularization-based approaches with a proper choice of regularization parameter, which entails knowledge of the noise level and/or tuning. By contrast, constrained least squaresestimation comes without any tuning parameter and may hence be preferred due to its simplicity.

95 Convergence Analysis of Prediction Markets via Randomized Subspace Descent

Rafael Frongillo raf@cs.berkeley.edu
CU Boulder
Mark D Reid mark.reid@anu.edu.au
Australia National University

Prediction markets are economic mechanisms for aggregating information about future events through sequential interactions with traders. The pricing mechanisms in these markets are known to be related to optimization algorithms in machine learning and through these connections we have some understanding of how equilibrium market prices relate to the beliefs of the traders in a market. However, little is known about rates and guarantees for the convergence of these sequential mechanisms, and two recent papers cite this as an important open question.In this paper we show how some previously studied prediction market trading models can be understood as a natural generalization of randomized coordinate descent which we call randomized subspace descent (RSD). We establish convergence rates for RSD and leverage them to prove rates for the two prediction market models above, answering the open questions. Our results extend beyond standard centralized markets to arbitrary trade networks.

96 Accelerated Proximal Gradient Methods for Nonconvex Programming

Li Huan lihuan_ss@126.com
Zhouchen Lin zlin@pku.edu.cn
Peking University

Nonconvex and nonsmooth problems have recently received considerable attention in signal/image processing, statistics and machine learning. However, solving the nonconvex and nonsmooth optimization problems remains a big challenge. Accelerated proximal gradient (APG) is an excellent method for convex programming. However, it is still unknown whether the usual APG can ensure the convergence to a critical point in nonconvex programming. To address this issue, we introduce a monitor-corrector step and extend APG for general nonconvex and nonsmooth programs. Accordingly, we propose a monotone APG and a non-monotone APG. The latter waives the requirement on monotonic reduction of the objective function and needs less computation in each iteration. To the best of our knowledge, we are the first to provide APG-type algorithms for general nonconvex and nonsmooth problems ensuring that every accumulation point is a critical point, and the convergence rates remain O(1/k2) when the problems are convex, in which k is the number of iterations. Numerical results testify to the advantage of our algorithms in speed.

97 Nearly Optimal Private LASSO

Kunal Talwar kunal@google.com
Li Zhang liqzhang@google.com
Google
Abhradeep Thakurta guhathakurta.abhradeep@gmail.com

We present a nearly optimal differentially private version of the well known LASSO estimator. Our algorithm provides privacy protection with respect to each training data item. The excess risk of our algorithm, compared to the non-private version, is O˜(1/ n2/3), assuming all the input data has bounded ℓ∞ norm. This is the first differentially private algorithm that achieves such a bound without the polynomial dependence on p under no addition assumption on the design matrix. In addition, we show that this error bound is nearly optimal amongst all differentially private algorithms.

98 Minimax Time Series Prediction

Wouter M Koolen wmkoolen@cwi.nl
Centrum Wiskunde & Informatica
Alan Malek alan.malek@gmail.com
Peter L Bartlett bartlett@cs.berkeley.edu
UC Berkeley
Yasin Abbasi yasin.abbasi@gmail.com
Queensland University of Technology

We consider an adversarial formulation of the problem of predicting atime series with square loss. The aim is to predict an arbitrarysequence of vectors almost as well as the best smooth comparatorsequence in retrospect. Our approach allows natural measures ofsmoothness, such as the squared norm of increments. More generally, wecan consider a linear time series model and penalize the comparatorsequence through the energy of the implied driving noise terms. Wederive the minimax strategy for all problems of this type, and we showthat it can be implemented efficiently. The optimal predictions arelinear in the previous observations. We obtain an explicit expressionfor the regret in terms of the parameters defining the problem. Fortypical, simple definitions of smoothness, the computation of theoptimal predictions involves only sparse matrices. In the case ofnorm-constrained data, where the smoothness is defined in terms of thesquared norm of the comparator's increments, we show that the regretgrows as T/λ√, where T is the length of the game andλ specifies the smoothness of the comparator.

99 Communication Complexity of Distributed Convex Learning and Optimization

Yossi Arjevani yossi.arjevani@weizmann.ac.il
Ohad Shamir ohad.shamir@weizmann.ac.il
Weizmann Institute of Science

We study the fundamental limits to communication-efficient distributed methods for convex learning and optimization, under different assumptions on the information available to individual machines, and the types of functions considered. We identify cases where existing algorithms are already worst-case optimal, as well as cases where room for further improvement is still possible. Among other things, our results indicate that without similarity between the local objective functions (due to statistical data similarity or otherwise) many communication rounds may be required, even if the machines have unbounded computational power.

100 Explore no more: Improved high-probability regret bounds for non-stochastic bandits

Gergely Neu neu.gergely@gmail.com
INRIA

This work addresses the problem of regret minimization in non-stochastic multi-armed bandit problems, focusing on performance guarantees that hold with high probability. Such results are rather scarce in the literature since proving them requires a large deal of technical effort and significant modifications to the standard, more intuitive algorithms that come only with guarantees that hold on expectation. One of these modifications is forcing the learner to sample the losses of every arm at least Ω(T√) times over T rounds, which can adversely affect performance if many of the arms are obviously suboptimal. While it is widely conjectured that this property is essential for proving high-probability regret bounds, we show in this paper that it is possible to achieve such strong results without this undesirable exploration component. Our result relies on a simple and intuitive loss-estimation strategy called Implicit eXploration (IX) that allows a remarkably clean analysis. To demonstrate the flexibility of our technique, we derive several improved high-probability bounds for various extensions of the standard multi-armed bandit framework. Finally, we conduct a simple experiment that illustrates the robustness of our implicit exploration technique.

101 A Nonconvex Optimization Framework for Low Rank Matrix Estimation

Tuo Zhao tzhao5@jhu.edu
Zhaoran Wang zhaoran@princeton.edu
Han Liu hanliu@princeton.edu
Princeton University

We study the estimation of low rank matrices via nonconvex optimization. Compared with convex relaxation, nonconvex optimization exhibits superior empirical performance for large scale instances of low rank matrix estimation. However, the understanding of its theoretical guarantees are limited. In this paper, we define the notion of projected oracle divergence based on which we establish sufficient conditions for the success of nonconvex optimization. We illustrate the consequences of this general framework for matrix sensing and completion. In particular, we prove that a broad class of nonconvex optimization algorithms, including alternating minimization and gradient-type methods, geometrically converge to the global optimum and exactly recover the true low rank matrices under standard conditions.

102 Individual Planning in Infinite-Horizon Multiagent Settings: Inference, Structure and Scalability

Xia Qu
Prashant Doshi

This paper provides the first formalization of self-interested planning in multiagent settings using expectation-maximization (EM). Our formalization in the context of infinite-horizon and finitely-nested interactive POMDPs (I-POMDP) is distinct from EM formulations for POMDPs and cooperative multiagent planning frameworks. We exploit the graphical model structure specific to I-POMDPs, and present a new approach based on block-coordinate descent for further speed up. Forward filtering-backward sampling -- a combination of exact filtering with sampling -- is explored to exploit problem structure.

TUESDAY SESSIONS

OUR SPONSORS

Google

Microsoft

Alibaba Group 阿里巴巴集团

Ketchum Trading, LLC

amazon.com

Apple

Baidu Research

CITADEL

facebook

IBM Research

NVIDIA

The Voleon Group

Artificial Intelligence
www.elsevier.com/locate/artint

Bloomberg

twitter

AdRoll

Analog Devices | Lyric Labs

CenturyLink Business

criteo

Cubist Systematic Strategies

deep genomics

D E Shaw & Co

NETFLIX

Disney Research

ebay

imagia

MaluubA

Man AHL

ORACLE

Panasonic

PDT PARTNERS

SONY

The Alan Turing Institute

TOYOTA

United Technologies Research Center

Vatic Labs

TWO SIGMA

YAHOO! LABS

WINTON

Adobe

ORAL SESSION

SESSION 1 - 9:00 – 10:10 AM

INVITED TALK: POSNER LECTURE
Probabilistic Machine Learning: Foundations and Frontiers

Zoubin Ghahramani
zoubin@eng.cam.ac.uk
University of Cambridge

Probabilistic modelling provides a mathematical framework for understanding what learning is, and has therefore emerged as one of the principal approaches for designing computer algorithms that learn from data acquired through experience. I will review the foundations of this field, from basics to Bayesian nonparametric models and scalable inference. I will then highlight some current areas of research at the frontiers of machine learning, leading up to topics such as probabilistic programming, Bayesian optimisation, the rational allocation of computational resources, and the Automatic Statistician.

Randomized Block Krylov Methods for Stronger and Faster Approximate Singular Value Decomposition

Cameron Musco cnmusco@mit.edu
Christopher Musco cpmusco@mit.edu
Massachusetts Institute of Technology

Since being analyzed by Rokhlin, Szlam, and Tygert \ cite{RokhlinTygertPCA} and popularized by Halko, Martinsson, and Tropp \cite{Halko:2011}, randomized Simultaneous Power Iteration has become the method of choice for approximate singular value decomposition. It is more accurate than simpler sketching algorithms, yet still converges quickly for \emph{any} matrix, independently of singular value gaps. After $O\tilde{}(1/\epsilon)$ iterations, it gives a low-rank approximation within $(1+\epsilon)$ of optimal for spectral norm error.We give the first provable runtime improvement on Simultaneous Iteration. A simple randomized block Krylov method, closely related to the classic Block Lanczos algorithm, gives the same guarantees in just $O\tilde{}(1/\epsilon\sqrt{})$ iterations and performs substantially better experimentally. Despite their long history, our analysis is the first of a Krylov subspace method that does not depend on singular value gaps, which are unreliable in practice.Furthermore, while it is a simple accuracy benchmark, even $(1+\epsilon)$ error for spectral norm low-rank approximation does not imply that an algorithm returns high quality principal components, a major issue for data applications. We address this problem for the first time by showing that both block Krylov methods and Simultaneous Iteration give nearly optimal PCA for any matrix. This result further justifies their strength over non-iterative sketching methods.

SPOTLIGHT SESSION

SESSION 1: 10:10 – 10:40 AM

- **Minimum Weight Perfect Matching via Blossom Belief Propagation**
  Sungsoo Ahn KAIST
  Sung-Soo Ahn, Sejun Park, KAIST
  Misha Chertkov, Jinwoo Shin, KAIST

- **Super-Resolution Off the Grid**
  Qingqing Huang, MIT
  Sham Kakade, University of Washington

- **b-bit Marginal Regression**
  Martin Slawski, Rutgers University
  Ping Li, Rugters University

- **LASSO with Non-linear Measurements is Equivalent to One With Linear Measurements**
  Christos Thrampoulidis, Caltech
  Ehsan Abbasi, Caltech
  Babak Hassibi, Caltech

- **Optimal Rates for Random Fourier Features**
  Bharath Sriperumbudur, Pennsylvania State University
  Zoltan Szabo, Gatsby Unit, UCL

- **Submodular Hamming Metrics**
  Jennifer Gillenwater, University of Washington
  Rishabh K Iyer, University of Washington, Seattle
  Bethany Lusch, University of Washington
  Rahul Kidambi, University of Washington
  Jeff A Bilmes, University of Washington

- **Top-k Multiclass SVM**
  Maksim Lapin, Max Planck Institute for Informatics
  Matthias Hein, Saarland University
  Bernt Schiele, Max Planck Institute for Informatics

ORAL SESSION

SESSION 2: 11:10 – 11:50 AM

Sampling from Probabilistic Submodular Models

Alkis Gotovos alkisg@inf.ethz.ch
Hamed Hassani hamed@inf.ethz.ch
Andreas Krause krausea@ethz.ch
ETH Zurich

Submodular and supermodular functions have found wide applicability in machine learning, capturing notions such as diversity and regularity, respectively. These notions have deep consequences for optimization, and the problem of (approximately) optimizing submodular functions has received much attention. However, beyond optimization, these notions allow specifying expressive probabilistic models that can be used to quantify predictive uncertainty via marginal inference. Prominent, well-studied special cases include Ising models and determinantal point processes, but the general class of log-submodular and log-supermodular models is much richer and little studied. In this paper, we investigate the use of Markov chain Monte Carlo sampling to perform approximate inference in general log-submodular and log-supermodular models. In particular, we consider a simple Gibbs sampling procedure, and establish two sufficient conditions, the first guaranteeing polynomial-time, and the second fast (O(nlogn)) mixing. We also evaluate the efficiency of the Gibbs sampler on three examples of such models, and compare against a recently proposed variational approach.

Solving Random Quadratic Systems of Equations Is Nearly as Easy as Solving Linear Systems

Yuxin Chen yxchen@stanford.edu
Emmanuel Candes candes@stanford.edu
Stanford University

This paper is concerned with finding a solution x to a quadratic system of equations $y_i = |< a_i, x >|^2$, i = 1, 2, ..., m. We prove that it is possible to solve unstructured quadratic systems in n variables exactly from O(n) equations in linear time, that is, in time proportional to reading and evaluating the data. This is accomplished by a novel procedure, which starting from an initial guess given by a spectral initialization procedure, attempts to minimize a non-convex objective. The proposed algorithm distinguishes from prior approaches by regularizing the initialization and descent procedures in an adaptive fashion, which discard terms bearing too much influence on the initial estimate or search directions. These careful selection rules---which effectively serve as a variance reduction scheme---provide a tighter initial guess, more robust descent directions, and thus enhanced practical performance. Further, this procedure also achieves a near-optimal statistical accuracy in the presence of noise. Finally, we demonstrate empirically that the computational cost of our algorithm is about four times that of solving a least-squares problem of the same size.

SPOTLIGHT SESSION

SESSION 2: 11:40 AM – 12:00 PM

- **Distributionally Robust Logistic Regression**
  Soroosh Shafieezadeh Abadeh, EPFL
  Peyman Esfahani, EPFL
  Daniel Kuhn, EPFL

- **On some provably correct cases of variational inference for topic models**
  Pranjal Awasthi, Princeton
  Andrej Risteski, Princeton

- **Extending Gossip Algorithms to Distributed Estimation of U-statistics**
  Igor Colin, Télécom ParisTech
  Aurélien Bellet, Telecom ParisTech
  Joseph Salmon, Télécom ParisTech
  Stéphan Clémençon, Telecom ParisTech

- **The Self-Normalized Estimator for Counterfactual Learning**
  Adith Swaminathan, Cornell University
  Thorsten Joachims, Cornell

- **Frank-Wolfe Bayesian Quadrature: Probabilistic Integration with Theoretical Guarantees**
  François-Xavier Briol, University of Warwick
  Chris Oates, University of Tech., Sydney
  Mark Girolami, Univeresity of Warwick
  Michael A Osborne, U Oxford

- **Newton-Stein Method: A Second Order Method for GLMs via Stein's Lemma**
  Murat A. Erdogdu, Stanford University

- **Asynchronous Parallel Stochastic Gradient for Nonconvex Optimization**
  Xiangru Lian, University of Rochester
  Yijun Huang, University of Rochester
  Yuncheng Li, University of Rochester
  Ji Liu, University of Rochester

- **Distributed Submodular Cover: Succinctly Summarizing Massive Data**
  Baharan Mirzasoleiman, ETHZ
  Amin Karbasi, Yale
  Ashwinkumar Badanidiyuru, Google Research
  Andreas Krause, ETHZ

ORAL SESSION

SESSION 3: 2:00 – 3:30 PM

INVITED TALK: Incremental Methods for Additive Cost Convex Optimization

Asuman Ozdaglar
asuman@mit.edu
Massachusetts Institute of Technology

Motivated by machine learning problems over large data sets and distributed optimization over networks, we consider the problem of minimizing the sum of a large number of convex component functions. We study incremental gradient methods for solving such problems, which use information about a single component function at each iteration. We provide new convergence rate results under some assumptions. We also consider incremental aggregated gradient methods, which compute a single component function gradient at each iteration while using outdated gradients of all component functions to approximate the entire global cost function, and provide new linear rate results.

This is joint work with Mert Gurbuzbalaban and Pablo Parrilo.

Probabilistic Line Searches for Stochastic Optimization

Maren Mahsereci mmahsereci@tue.mpg.de
Philipp Hennig phennig@tue.mpg.de
MPI for Intelligent Systems, Tübingen

In deterministic optimization, line searches are a standard tool ensuring stability and efficiency. Where only stochastic gradients are available, no direct equivalent has so far been formulated, because uncertain gradients do not allow for a strict sequence of decisions collapsing the search space. We construct a probabilistic line search by combining the structure of existing deterministic methods with notions from Bayesian optimization. Our method retains a Gaussian process surrogate of the univariate optimization objective, and uses a probabilistic belief over the Wolfe conditions to monitor the descent. The algorithm has very low computational cost, and no user-controlled parameters. Experiments show that it effectively removes the need to define a learning rate for stochastic gradient descent.

OCOEVOLVE: A Joint Point Process Model for Information Diffusion and Network Co-evolution

Mehrdad Farajtabar mehrdad@gatech.edu
Yichen Wang ywang737@math.gatech.edu
Shuang Li sli370@gatech.edu
Hongyuan Zha zha@cc.gatech.edu
Le Song lsong@cc.gatech.edu
Georgia Institute of Technology
Manuel Rodriguez manuelgr@mpi-sws.org
MPI SWS

Information diffusion in online social networks is affected by the underlying network topology, but it also has the power to change it. Online users are constantly creating new links when exposed to new information sources, and in turn these links are alternating the way information spreads. However, these two highly intertwined stochastic processes, information diffusion and network evolution, have been predominantly studied separately, ignoring their co-evolutionary dynamics.We propose a temporal point process model, COEVOLVE, for such joint dynamics, allowing the intensity of one process to be modulated by that of the other. This model allows us to efficiently simulate interleaved diffusion and network events, and generate traces obeying common diffusion and network patterns observed in real-world networks. Furthermore, we also develop a convex optimization framework to learn the parameters of the model from historical diffusion and network evolution traces. We experimented with both synthetic data and data gathered from Twitter, and show that our model provides a good fit to the data as well as more accurate predictions than alternatives.

SPOTLIGHT SESSION

SESSION 3: 3:30 – 4:00 PM

- **Linear Response Methods for Accurate Covariance Estimates from Mean Field Variational Bayes**
  Ryan J Giordano, UC Berkeley
  Tamara Broderick, MIT
  Michael I Jordan, UC Berkeley

- **Latent Bayesian melding for integrating individual and population models**
  Mingjun Zhong, University of Edinburgh
  Nigel Goddard, Charles Sutton, University of Edinburgh

- **Rapidly Mixing Gibbs Sampling for a Class of Factor Graphs Using Hierarchy Width**
  Christopher M De Sa, Stanford
  Kunle Olukotun, Stanford
  Christopher Ré, Stanford
  Chris Ré, Stanford
  Ce Zhang, Wisconsin

- **Automatic Variational Inference in Stan**
  Alp Kucukelbir, Princeton University
  Rajesh Ranganath, Princeton University
  Andrew Gelman, Columbia University
  David Blei, Columbia University

- **Data Generation as Sequential Decision Making**
  Philip Bachman, McGill University
  Doina Precup, University of McGill

- **Stochastic Expectation Propagation**
  Yingzhen Li, University of Cambridge
  José Miguel Hernández-Lobato, Harvard
  Richard E Turner, University of Cambridge

- **Deep learning with Elastic Averaging SGD**
  Sixin Zhang, New York University
  Anna E Choromanska, Courant Institute, NYU
  Yann LeCun, New York University

ORAL SESSION

SESSION 4: 4:30 – 5:30 PM

Competitive Distribution Estimation: Why is Good-Turing Good

Alon Orlitsky alon@ucsd.edu
Ananda Suresh s.theertha@gmail.com
University of California, San Diego

Estimating distributions over large alphabets is a fundamental tenet of machine learning. Yet no estimator is known to estimate all distributions well. For example, add-constant estimators are nearly min-max optimal, but perform poorly in practice, while practical estimators such as Jelinek-Mercer, absolute discounting, and Good-Turing, are not known to be near optimal for essentially any distribution.We provide the first uniform optimality proof for any distribution estimator. We show that a variant of Good-Turing estimators is nearly best for all distributions in two competitive ways. First it estimates every distribution nearly as well as the best estimator designed with prior knowledge of the distribution up to a permutation. Second, it estimates every distribution nearly as well as the best estimator designed with prior knowledge of the exact distribution but restricted, as all natural estimators, to assign the same probability to all symbols appearing the same number of times. Specifically, we show that for both comparisons, the KL divergence of the Good-Turing variant is always within $O\sim(\min(k/n,1/n\sqrt{}))$ of the best estimator. Conversely, any estimator must have a KL divergence $\geq\Omega\sim(\min(k/n,1/n2/3))$ over the best estimator for the first comparison, and $\geq\Omega\sim(\min(k/n,1/n\sqrt{}))$ for the second.

Fast Convergence of Regularized Learning in Games

Vasilis Syrgkanis vasy@microsoft.com
Alekh Agarwal alekha@microsoft.com
Robert Schapire schapire@microsoft.com
MIcrosoft Research
Haipeng Luo haipengl@cs.princeton.edu
Princeton University

We show that natural classes of regularized learning algorithms with a form of recency bias achieve faster convergence rates to approximate efficiency and to coarse correlated equilibria in multiplayer normal form games. When each player in a game uses an algorithm from our class, their individual regret decays at $O(T^{-3/4})$, while the sum of utilities converges to an approximate optimum at $O(T^{-1})$--an improvement upon the worst case $O(T^{-1/2})$ rates. We show a black-box reduction for any algorithm in the class to achieve $O(T^{-1/2})$ rates against an adversary, while maintaining the faster rates against algorithms in the class. Our results extend those of Rakhlin and Shridharan~\cite{Rakhlin2013} and Daskalakis et al.~\cite{Daskalakis2014}, who only analyzed two-player zero-sum games for specific algorithms.

Interactive Control of Diverse Complex Characters with Neural Networks

Igor Mordatch igor.mordatch@gmail.com
Kendall Lowrey kendall.lowrey@gmail.com
Galen Andrew gmandrew@uw.edu
Zoran Popovic zoran@cs.washington.edu
Emanuel Todorov todorov@cs.washington.edu
University of Washington

We present a method for training recurrent neural networks to act as near-optimal feedback controllers. It is able to generate stable and realistic behaviors for a range of dynamical systems and tasks -- swimming, flying, biped and quadruped walking with different body morphologies. It does not require motion capture or task-specific features or state machines. The controller is a neural network, having a large number of feed-forward units that learn elaborate state-action mappings, and a small number of recurrent units that implement memory states beyond the physical system state. The action generated by the network is defined as velocity. Thus the network is not learning a control policy, but rather the dynamics under an implicit policy. Essential features of the method include interleaving supervised learning with trajectory optimization, injecting noise during training, training for unexpected changes in the task specification, and using the trajectory optimizer to obtain optimal feedback gains in addition to optimal actions.

SPOTLIGHT SESSION

SESSION 4: 5:40 – 6:00 PM

- **The Human Kernel**
  Andrew G Wilson, Carnegie Mellon University
  Christoph Dann, Carnegie Mellon University
  Chris Lucas, University of Edinburgh
  Eric P Xing, Carnegie Mellon University

- **On the Pseudo-Dimension of Nearly Optimal Auctions**
  Jamie H Morgenstern, University of Pennsylvania
  Tim Roughgarden, Stanford University

- **High-dimensional neural spike train analysis with generalized count linear dynamical systems**
  YUANJUN GAO, Columbia University
  Lars Busing, Columbia University
  Krishna V Shenoy,Stanford University
  John Cunningham, University of Columbia

- **Measuring Sample Quality with Stein's Method**
  Jackson Gorham, Stanford University
  Lester Mackey, Stanford

- **Biologically Inspired Dynamic Textures for Probing Motion Perception**
  Jonathan Vacher, Université Paris Dauphine
  Andrew Isaac Meso, Institut des neurosciences de la Timone
  Laurent U Perrinet, Institut des neurosciences de la Timone
  Gabriel Peyré, CNRS & Université Paris-Dauphine

- **Large-Scale Bayesian Multi-Label Learning via Topic-Based Label Embeddings**
  Piyush Rai, IIT Kanpur
  Changwei Hu,
  Ricardo Henao, Duke University
  Lawrence Carin, Duke University

- **Closed-form Estimators for High-dimensional Generalized Linear Models**
  Eunho Yang, IBM Thomas J. Watson Research Center
  Aurelie C Lozano, IBM Research
  Pradeep K Ravikumar, University of Texas at Austin

- **Learning Stationary Time Series using Gaussian Processes with Nonparametric Kernels**
  Felipe Tobar, Universidad de Chile
  Thang Bui, University of Cambridge
  Richard E Turner, University of Cambridge

POSTER SESSION

POSTERS 7:00 – 11:59 PM

1 **Deep Generative Image Models using a Laplacian Pyramid of Adversarial Networks**
Emily L Denton · Rob Fergus · arthur szlam · Soumith Chintala

2 **Shepard Convolutional Neural Networks**
Jimmy SJ Ren · Li Xu · Qiong Yan · Wenxiu Sun

3 **Learning Structured Output Representation using Deep Conditional Generative Models**
Kihyuk Sohn · Honglak Lee · Xinchen Yan

4 **Expressing an Image Stream with a Sequence of Natural Sentences**
Cesc C Park · Gunhee Kim

5 **Visalogy: Answering Visual Analogy Questions**
Fereshteh Sadeghi · C. Lawrence Zitnick · Ali Farhadi

6 **Bidirectional Recurrent Convolutional Networks for Multi-Frame Super-Resolution**
Yan Huang · Wei Wang · Liang Wang

7 **SubmodBoxes: Near-Optimal Search for a Set of Diverse Object Proposals**
Qing Sun · Dhruv Batra

8 **Galileo: Perceiving Physical Object Properties by Integrating a Physics Engine with Deep Learning**
Jiajun Wu · Ilker Yildirim · Joseph J Lim · Bill Freeman · Josh Tenenbaum

9 **Learning visual biases from human imagination**
Carl Vondrick · Hamed Pirsiavash · Aude Oliva · Antonio Torralba

10 **Character-level Convolutional Networks for Text Classification**
Xiang Zhang · Junbo Zhao · Yann LeCun

11 Winner-Take-All Autoencoders
Alireza Makhzani · Brendan J Frey

12 Learning both Weights and Connections for Efficient Neural Network
Song Han · Jeff Pool · John Tran · Bill Dally

13 Interactive Control of Diverse Complex Characters with Neural Networks
Igor Mordatch · Kendall Lowrey · Galen Andrew · Zoran Popovic · Emanuel Todorov

14 Biologically Inspired Dynamic Textures for Probing Motion Perception
Jonathan Vacher · Andrew Isaac Meso · Laurent U Perrinet · Gabriel Peyré

15 Unsupervised Learning by Program Synthesis
Kevin Ellis · Armando Solar-Lezama · Josh Tenenbaum

16 Deep Poisson Factor Modeling
Ricardo Henao · Zhe Gan · James Lu · Lawrence Carin

17 Large-Scale Bayesian Multi-Label Learning via Topic-Based Label Embeddings
Piyush Rai · Changwei Hu · Ricardo Henao · Lawrence Carin

18 Tensorizing Neural Networks
Alexander Novikov · Dmitrii Podoprikhin · Anton Osokin · Dmitry P Vetrov

19 Training Restricted Boltzmann Machine via the Thouless-Anderson-Palmer free energy
Marylou Gabrie · Eric W Tramel · Florent Krzakala

20 The Brain Uses Reliability of Stimulus Information when Making Perceptual Decisions
Sebastian Bitzer · Stefan Kiebel

21 Unlocking neural population non-stationarities using hierarchical dynamics models
Mijung Park · Gergo Bohner · Jakob H Macke

22 Deeply Learning the Messages in Message Passing Inference
Guosheng Lin · Chunhua Shen · Ian Reid · Anton van den Hengel

23 COEVOLVE: A Joint Point Process Model for Information Diffusion and Network Co-evolution
Mehrdad Farajtabar · Yichen Wang · Manuel Rodriguez · Shuang Li · Hongyuan Zha · Le Song

24 The Human Kernel
Andrew G Wilson · Christoph Dann · Chris Lucas · Eric P Xing

25 Latent Bayesian melding for integrating individual and population models
Mingjun Zhong · Nigel Goddard · Charles Sutton

26 High-dimensional neural spike train analysis with generalized count linear dynamical systems
Yuanjun Gao · Lars Busing · Krishna V Shenoy · John Cunningham

27 Efficient Continuous-Time Hidden Markov Model for Disease Modeling
Yu-Ying Liu · Fuxin Li · Shuang Li · Le Song · James M Rehg

28 The Population ior and Bayesian Modeling on Streams
James McInerney · Rajesh Ranganath · David Blei

29 Probabilistic Curve Learning: Coulomb Repulsion and the Electrostatic Gaussian Process
Ye Wang · David B Dunson

30 Preconditioned Spectral Descent for Deep Learning
David E Carlson · Edo Collins · Ya-Ping Hsieh · Lawrence Carin · Volkan Cevher

31 Learning Continuous Control Policies by Stochastic Value Gradients
Nicolas Heess · Greg Wayne · David Silver · Tim Lillicrap · Tom Erez · Yuval Tassa

32 Learning Stationary Time Series using Gaussian Processes with Nonparametric Kernels
Felipe Tobar · Thang Bui · Richard E Turner

33 Path-SGD: Path-Normalized Optimization in Deep Neural Networks
Behnam Neyshabur · Russ R Salakhutdinov · Nati Srebro

34 Automatic Variational Inference in Stan
Alp Kucukelbir · Rajesh Ranganath · Andrew Gelman · David Blei

35 Data Generation as Sequential Decision Making
Philip Bachman · Doina Precup

36 Stochastic Expectation Propagation
Yingzhen Li · José Miguel Hernández-Lobato · Richard E Turner

37 Deep learning with Elastic Averaging SGD
Sixin Zhang · Anna E Choromanska · Yann LeCun

38 Learning with Group Invariant Features: A Kernel Perspective.
Youssef Mroueh · Stephen Voinea · Tomaso A Poggio

39 Linear Response Methods for Accurate Covariance Estimates from Mean Field Variational Bayes
Ryan J Giordano · Tamara Broderick · Michael I Jordan

40 Probabilistic Line Searches for Stochastic Optimization
Maren Mahsereci · Philipp Hennig

41 A hybrid sampler for Poisson-Kingman mixture models
Maria Lomeli-Garcia · Stefano Favaro · Yee Whye Teh

42 Tree-Guided MCMC Inference for Normalized Random Measure Mixture Models
Juho Lee · Seungjin Choi

43 Reflection, Refraction, and Hamiltonian Monte Carlo
Hadi Mohasel Afshar · Justin Domke

44 Planar Ultrametrics for Image Segmentation
Julian E Yarkony · Charless Fowlkes

45 Learning Bayesian Networks with Thousands of Variables
Mauro Scanagatta · Cassio P de Campos · Giorgio Corani · Marco Zaffalon

46 Parallel Predictive Entropy Search for Batch Global Optimization of Expensive Objective Functions
Amar Shah · Zoubin Ghahramani

47 Rapidly Mixing Gibbs Sampling for a Class of Factor Graphs Using Hierarchy Width
Christopher M De Sa · Ce Zhang · Kunle Olukotun · Chris Ré

48 On some provably correct cases of variational inference for topic models
Pranjal Awasthi · Andrej Risteski

49 Large-scale probabilistic predictors with and without guarantees of validity
Vladimir Vovk · Ivan Petej · Valentina Fedorova

50 On the Accuracy of Self-Normalized Log-Linear Models
Jacob Andreas · Maxim Rabinovich · Michael I Jordan · Dan Klein

51 Policy Evaluation Using the Ω-Return
Philip S Thomas · Scott Niekum · Georgios Theocharous · George Konidaris

52 Community Detection via Measure Space Embedding
Mark Kozdoba · Shie Mannor

53 The Consistency of Common Neighbors for Link Prediction in Stochastic Blockmodels
Purnamrita Sarkar · Deepayan Chakrabarti · peter j bickel

54 Inference for determinantal point processes without spectral knowledge
Rémi Bardenet · Michalis Titsias

55 Sample Complexity of Learning Mahalanobis Distance Metrics
Nakul Verma · Kristin Branson

56 Manifold Optimization for Gaussian Mixture Models
Reshad Hosseini · Suvrit Sra

57 Frank-Wolfe Bayesian Quadrature: Probabilistic Integration with Theoretical Guarantees
François-Xavier Briol · Chris Oates · Mark Girolami · Michael A Osborne

58 Scale Up Nonlinear Component Analysis with Doubly Stochastic Gradients
Bo Xie · Yingyu Liang · Le Song

59 The Self-Normalized Estimator for Counterfactual Learning
Adith Swaminathan · Thorsten Joachims

60 Distributionally Robust Logistic Regression
Soroosh Shafieezadeh Abadeh · Peyman Esfahani · Daniel Kuhn

61 Top-k Multiclass SVM
Maksim Lapin · Matthias Hein · Bernt Schiele

62 Measuring Sample Quality with Stein's Method
Jackson Gorham · Lester Mackey

63 Asynchronous Parallel Stochastic Gradient for Nonconvex Optimization
Xiangru Lian · Yijun Huang · Yuncheng Li · Ji Liu

64 Solving Random Quadratic Systems of Equations Is Nearly as Easy as Solving Linear Systems
Yuxin Chen · Emmanuel Candes

65 Distributed Submodular Cover: Succinctly Summarizing Massive Data
Baharan Mirzasoleiman · Amin Karbasi · Ashwinkumar Badanidiyuru · Andreas Krause

66 Parallel Correlation Clustering on Big Graphs
Xinghao Pan · Dimitris Papailiopoulos · Benjamin Recht · Kannan Ramchandran · Michael Jordan

67 Fast Bidirectional Probability Estimation in Markov Models
Sid Banerjee · Peter Lofgren

68 Evaluating the statistical significance of biclusters
Jason D Lee · Yuekai Sun · Jonathan E Taylor

69 Regularization Path of Cross-Validation Error Lower Bounds
Atsushi Shibagaki · Yoshiki Suzuki · Masayuki Karasuyama · Ichiro Takeuchi

70 Sampling from Probabilistic Submodular Models
Alkis Gotovos · Hamed Hassani · Andreas Krause

71 Submodular Hamming Metrics
Jennifer Gillenwater · Rishabh K Iyer · Bethany Lusch · Rahul Kidambi · Jeff A Bilmes

72 Extending Gossip Algorithms to Distributed Estimation of U-statistics
Igor Colin · Aurélien Bellet · Joseph Salmon · Stéphan Clémençon

73 Newton-Stein Method: A Second Order Method for GLMs via Stein's Lemma
Murat A. Erdogdu

74 Collaboratively Learning Preferences from Ordinal Data
Sewoong Oh · Kiran K Thekumparampil · Jiaming Xu

75 SGD Algorithms based on Incomplete U-statistics: Large-Scale Minimization of Empirical Risk
Guillaume Papa · Stéphan Clémençon · Aurélien Bellet

76 Alternating Minimization for Regression Problems with Vector-valued Outputs
Prateek Jain · Ambuj Tewari

77 On Variance Reduction in Stochastic Gradient Descent and its Asynchronous Variants
Sashank J. Reddi · Ahmed Hefny · Suvrit Sra · Barnabas Poczos · Alex J Smola

78 Subset Selection by Pareto Optimization
Chao Qian · Yang Yu · Zhi-Hua Zhou

79 Interpolating Convex and Non-Convex Tensor Decompositions via the Subspace Norm
Qinqing Zheng · Ryota Tomioka

80 Minimum Weight Perfect Matching via Blossom Belief Propagation
Sung-Soo Ahn · Sejun Park · Misha Chertkov · Jinwoo Shin

81 b-bit Marginal Regression
Martin Slawski · Ping Li

82 LASSO with Non-linear Measurements is Equivalent to One With Linear Measurements
Christos Thrampoulidis · Ehsan Abbasi · Babak Hassibi

83 Randomized Block Krylov Methods for Stronger and Faster Approximate Singular Value Decomposition
Cameron Musco · Christopher Musco

84 On the Pseudo-Dimension of Nearly Optimal Auctions
Jamie H Morgenstern · Tim Roughgarden

85 Closed-form Estimators for High-dimensional Generalized Linear Models
Eunho Yang · Aurelie C Lozano · Pradeep K Ravikumar

86 Fast, Provable Algorithms for Isotonic Regression in all l_p-norms
Rasmus Kyng · Anup Rao · Sushant Sachdeva

87 Semi-proximal Mirror-Prox for Nonsmooth Composite Minimization
Niao He · Zaid Harchaoui

88 Competitive Distribution Estimation: Why is Good-Turing Good
Alon Orlitsky · Ananda Suresh

89 A Universal Primal-Dual Convex Optimization Framework
Alp Yurtsever · Quoc Tran Dinh · Volkan Cevher

90 Sample Complexity of Episodic Fixed-Horizon Reinforcement Learning
Christoph Dann · Emma Brunskill

91 Private Graphon Estimation for Sparse Graphs
Christian Borgs · Jennifer Chayes · Adam Smith

92 HONOR: Hybrid Optimization for NOn-convex Regularized problems
Pinghua Gong · Jieping Ye

93 A Convergent Gradient Descent Algorithm for Rank Minimization and Semidefinite Programming from Random Linear Measurements
Qinqing Zheng · John Lafferty

94 Super-Resolution Off the Grid
Qingqing Huang · Sham Kakade

95 Optimal Rates for Random Fourier Features
Bharath Sriperumbudur · Zoltan Szabo

96 Combinatorial Bandits Revisited
Richard Combes · Mohammad Sadegh Talebi Mazraeh Shahi · Alexandre Proutiere · marc lelarge

97 Fast Convergence of Regularized Learning in Games
Vasilis Syrgkanis · Alekh Agarwal · Haipeng Luo · Robert Schapire

98 On Elicitation Complexity
Rafael M Frongillo · Ian Kash

99 Online Learning with Adversarial Delays
Kent Quanrud · Daniel Khashabi

100 Structured Estimation with Atomic Norms: General Bounds and Applications
Sheng Chen · Arindam Banerjee

101 Subsampled Power Iteration: a Unified Algorithm for Block Models and Planted CSP's
Vitaly Feldman · Will Perkins · Santosh Vempala

DEMONSTRATIONS

Demonstrations, 7:00 – 11:59 PM

D1 Vitruvian Science: a visual editor for quickly building neural networks in the cloud
Markus Beissinger, Vitruvian Science
Sherjil Ozair, Indian Institute of Technology Delhi

D2 DIANNE - Distributed Artificial Neural Networks
Steven Bohez, Ghent University - iMinds
Tim Verbelen, Ghent University - iMinds

D3 Fast sampling with neuromorphic hardware
Mihai A Petrovici, University of Heidelberg
David Stöckel, Heidelberg University
Ilja Bytschok, Kirchhoff-Institute for Physics
Johannes Bill
Thomas Pfeil, Kirchhoff-Institute for Physics
Johannes Schemmel, University of Heidelberg
Karlheinz Meier, Heidelberg University

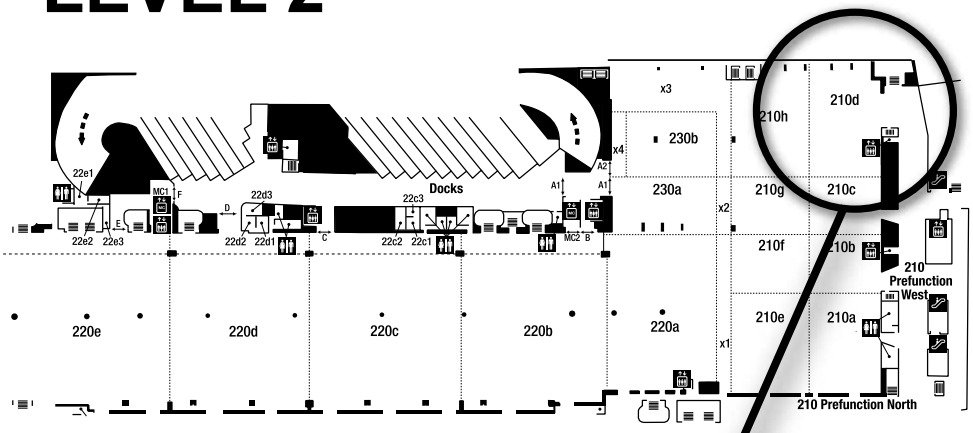
D4 Deep Learning using Approximate Hardware
Joseph Bates, Singular Computing LLC

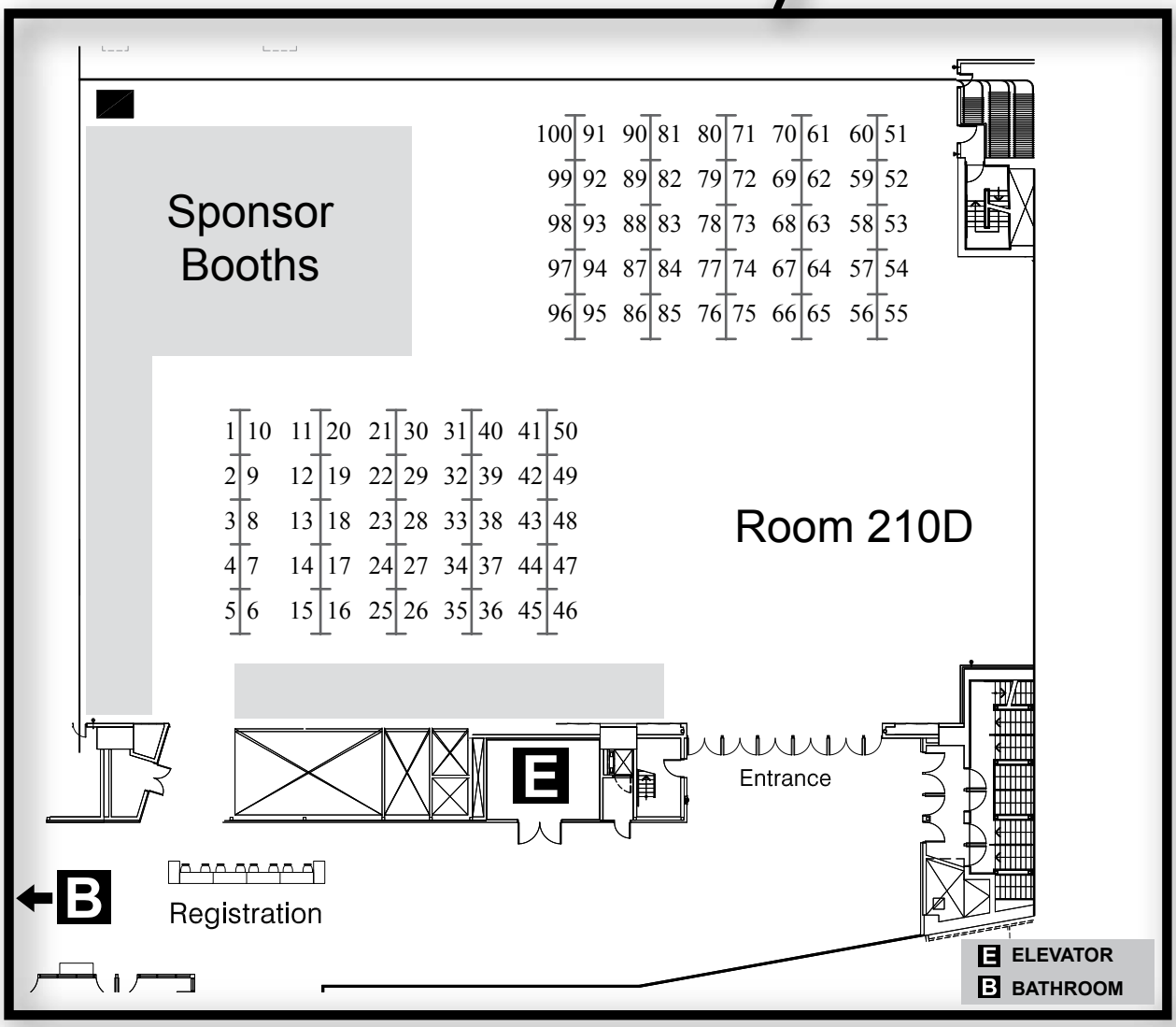
D5 An interactive system for the extraction of meaningful visualizations from high-dimensional data
Madalina Fiterau, Stanford University
Artur Dubrawski, Carnegie Mellon University
Donghan Wang, CMU

D6 Claudico: The World's Strongest No-Limit Texas Hold'em Poker AI
Noam Brown, Carnegie Mellon University
Tuomas Sandholm, Carnegie Mellon University

TUESDAY POSTER FLOORPLAN

LEVEL 2





1 Deep Generative Image Models using a Laplacian Pyramid of Adversarial Networks

Emily L Denton denton@cs.nyu.edu
New York University
Rob Fergus robfergus@fb.com
Facebook AI Research
arthur szlam aszlam@fb.com
Facebook
Soumith Chintala soumith@fb.com
Facebook AI Research

In this paper we introduce a generative model capable of producing high quality samples of natural images. Our approach uses a cascade of convolutional networks (convnets) within a Laplacian pyramid framework to generate images in a coarse-to-fine fashion. At each level of the pyramid a separate generative convnet model is trained using the Generative Adversarial Nets (GAN) approach. Samples drawn from our model are of significantly higher quality than existing models. In a quantitive assessment by human evaluators our CIFAR10 samples were mistaken for real images around 40% of the time, compared to 10% for GAN samples. We also show samples from more diverse datasets such as STL10 and LSUN.

2 Shepard Convolutional Neural Networks

Jimmy SJ Ren jimmy.sj.ren@gmail.com
Li Xu nathan.xuli@gmail.com
Qiong Yan yanqiong@sensetime.com
Wenxiu Sun sunwenxiu@sensetime.com
SenseTime Group Limited

Deep learning has recently been introduced to the field of low-level computer vision and image processing. Promising results have been obtained in a number of tasks including super-resolution, inpainting, denoising, deconvolution, etc. However, previously adopted neural network approaches such as convolutional neural networks and sparse auto-encoders are inherently with translation invariant operators. We found this property prevents the deep learning approaches from outperforming the state-of-the-art if the task itself requires translation variant interpolation (TVI). In this paper, we drew on Shepard interpolation and designed Shepard Convolutional Neural Networks (ShCNN) which efficiently realizes end-to-end trainable TVI operators in the network. With this new ability, ShCNN generated very promising results in image interpolation tasks which demonstrated its capability to achieve high quality TVI. We show that by adding only a few feature maps in the new TVI layer, the network is able to achieve stronger results than a much deeper architecture. Superior performance on both image inpainting and super-resolution was obtained where our super-resolution system outperfromed all the previous systems while keeping the running time competitive.

3 Learning Structured Output Representation using Deep Conditional Generative Models

Kihyuk Sohn kihyuks@umich.edu
University of Michigan
Honglak Lee honglak@eecs.umich.edu
U. Michigan
Xinchen Yan xcyan@umich.edu
UMich

Supervised deep learning has been successfully applied for many recognition problems in machine learning and computer vision. Although it can approximate a complex many-to-one function very well when large number of training data is provided, the lack of probabilistic inference of the current supervised deep learning methods makes it difficult to model a complex structured output representations. In this work, we develop a scalable deep conditional generative model for structured output variables using Gaussian latent variables. The model is trained efficiently in the framework of stochastic gradient variational Bayes, and allows a fast prediction using stochastic feed-forward inference. In addition, we provide novel strategies to build a robust structured prediction algorithms, such as recurrent prediction network architecture, input noise-injection and multi-scale prediction training methods. In experiments, we demonstrate the effectiveness of our proposed algorithm in comparison to the deterministic deep neural network counterparts in generating diverse but realistic output representations using stochastic inference. Furthermore, the proposed schemes in training methods and architecture design were complimentary, which leads to achieve strong pixel-level object segmentation and semantic labeling performance on Caltech-UCSD Birds 200 and the subset of Labeled Faces in the Wild dataset.

4 Expressing an Image Stream with a Sequence of Natural Sentences

Cesc C Park cs.park@vision.snu.ac.kr
Gunhee Kim gunhee@snu.ac.kr
Seoul National University

We propose an approach for generating a sequence of natural sentences for an image stream. Since general users usually take a series of pictures on their special moments, much online visual information exists in the form of image streams, for which it would better take into consideration of the whole set to generate natural language descriptions. While almost all previous studies have dealt with the relation between a single image and a single natural sentence, our work extends both input and output dimension to a sequence of images and a sequence of sentences. To this end, we design a novel architecture called coherent recurrent convolutional network (CRCN), which consists of convolutional networks, bidirectional recurrent networks, and entity-based local coherence model. Our approach directly learns from vast user-generated resource of blog posts as text-image parallel training data. We demonstrate that our approach outperforms other state-of-the-art candidate methods, using both quantitative measures (e.g. BLEU and top-K recall) and user studies via Amazon Mechanical Turk.

5 Visalogy: Answering Visual Analogy Questions

Fereshteh Sadeghi fsadeghi@cs.washington.edu
University of Washington
C. Lawrence Zitnick larryz@microsoft.com
Microsoft Research
Ali Farhadi ali@cs.washington.edu
University of Washington

In this paper, we study the problem of answering visual analogy questions. The visual analogy questions are similar to SAT analogy questions but using images. These questions typically take the form of image A is to image B as image C is to what. Answering these questions entails discovering the mapping from image A to image B and then extending the mapping to image C and searching for the image D such that the relation from A to B holds for C to D. We pose this problem as learning an embedding that encourages pairs of analogous images with similar transformation to be close together using convolutional neural networks with siamese quadruple architecture. We introduce a dataset of visual analogy questions in natural images and show first results of its kind on solving analogy questions on natural images.

6 Bidirectional Recurrent Convolutional Networks for Multi-Frame Super-Resolution

Yan Huang yhuang@nlpr.ia.ac.cn
CRIPAC, CASIA
Wei Wang wangwei@nlpr.ia.ac.cn
NLPR,CASIA
Liang Wang wangliang@nlpr.ia.ac.cn

Super resolving a low-resolution video is usually handled by either single-image super-resolution (SR) or multi-frame SR. Single-Image SR deals with each video frame independently, and ignores intrinsic temporal dependency of video frames which actually plays a very important role in video super-resolution. Multi-Frame SR generally extracts motion information, e.g. optical flow, to model the temporal dependency, which often shows high computational cost. Considering that recurrent neural network (RNN) can model long-term contextual information of temporal sequences well, we propose a bidirectional recurrent convolutional network for efficient multi-frame SR.Different from vanilla RNN, 1) the commonly-used recurrent full connections are replaced with weight-sharing convolutional connections and 2) conditional convolutional connections from previous input layers to current hidden layer are added for enhancing visual-temporal dependency modelling. With the powerful temporal dependency modelling, our model can super resolve videos with complex motions and achieve state-of-the-art performance. Due to the cheap convolution operations, our model has a low computational complexity and runs orders of magnitude faster than other multi-frame methods.

7 SubmodBoxes: Near-Optimal Search for a Set of Diverse Object Proposals

Qing Sun sunqing@vt.edu
Dhruv Batra dbatra@vt.edu
Virginia Tech

This paper formulates the search for a set of bounding boxes (as needed in object proposal generation) as a monotone submodular maximization problem over the space of all possible bounding boxes in an image. Since, the number of possible bounding boxes in an image is very large O(#pixels2), even a single linear scan to perform the greedy augmentation for submodular maximization is intractable. Thus, we formulate the greedy augmentation step as a B\&B scheme. In order to speed up repeated application of B\&B, we propose a novel generalization of Minoux's 'lazy greedy' algorithm to the B\&B tree. Theoretically, our proposed formulation provides a new understanding to the problem, and contains classic heuristic approaches such as Sliding Window+Non-Maximal Suppression (NMS) and and Efficient Subwindow Search (ESS) as special cases. Empirically, we show that our approach leads to a state-of-art performance on object proposal generation via a novel diversity measure.

8 Galileo: Perceiving Physical Object Properties by Integrating a Physics Engine with Deep Learning

Jiajun Wu jiajunwu@mit.edu
Ilker Yildirim ilkery@mit.edu
Joseph J Lim lim@csail.mit.edu
Bill Freeman billf@mit.edu
Josh Tenenbaum jbt@mit.edu
MIT

Humans demonstrate remarkable abilities to predict physical events in dynamic scenes, and to infer the physical properties of objects from static images. We propose a generative model for solving these problems of physical scene understanding from real-world videos and images. At the core of our generative model is a 3D physics engine, operating on an object-based representation of physical properties, including mass, position, 3D shape, and friction. We can infer these latent properties using relatively brief runs of MCMC, which drive simulations in the physics engine to fit key features of visual observations. We further explore directly mapping visual inputs to physical properties, inverting a part of the generative process using deep learning. We name our model Galileo, and evaluate it on a video dataset with simple yet physically rich scenarios. Results show that Galileo is able to infer the physical properties of objects and predict the outcome of a variety of physical events, with an accuracy comparable to human subjects. Our study points towards an account of human vision with generative physical knowledge at its core, and various recognition models as helpers leading to efficient inference.

9 Learning visual biases from human imagination

Carl Vondrick vondrick@mit.edu
Aude Oliva oliva@mit.edu
Antonio Torralba torralba@mit.edu
MIT
Hamed Pirsiavash hpirsiav@mit.edu
UMBC

Although the human visual system can recognize many concepts under challengingconditions, it still has some biases. In this paper, we investigate whether wecan extract these biases and transfer them into a machine recognition system. We introduce a novel method that, inspired by well-known tools in humanpsychophysics, estimates the biases that the human visual system might use forrecognition, but in computer vision feature spaces. Our experiments aresurprising, and suggest that classifiers from the human visual system can betransferred into a machine with some success. Since these classifiers seem tocapture favorable biases in the human visual system, we further present an SVMformulation that constrains the orientation of the SVM hyperplane to agree withthe bias from human visual

system. Our results suggest that transferring thishuman bias into machines may help object recognition systems generalize acrossdatasets and perform better when very little training data is available.

10 Character-level Convolutional Networks for Text Classification

Xiang Zhang xz558@nyu.edu
Junbo Zhao junbo.zhao@cs.nyu.edu
Yann LeCun yann@cs.nyu.edu
New York University

This article offers an empirical exploration on the use of character-level convolutional networks (ConvNets) for text classification. We constructed several large-scale datasets to show that character-level convolutional networks could achieve state-of-the-art or competitive results. Comparisons are offered against traditional models such as bag of words, n-grams and their TFIDF variants, and deep learning models such as word-based ConvNets and recurrent neural networks.

11 Winner-Take-All Autoencoders

Alireza Makhzani a.makhzani@gmail.com
Brendan J Frey frey@psi.toronto.edu
University of Toronto

In this paper, we propose a winner-take-all method for learning hierarchical sparse representations in an unsupervised fashion. We first introduce fully-connected winner-take-all autoencoders which use mini-batch statistics to directly enforce a lifetime sparsity in the activations of the hidden units. We then propose the convolutional winner-take-all autoencoder which combines the benefits of convolutional architectures and autoencoders for learning shift-invariant sparse representations. We describe a way to train convolutional autoencoders layer by layer, where in addition to lifetime sparsity, a spatial sparsity within each feature map is achieved using winner-take-all activation functions. We will show that winner-take-all autoencoders can be used to to learn deep sparse representations from the MNIST, CIFAR-10, ImageNet, Street View House Numbers and Toronto Face datasets, and achieve competitive classification performance.

12 Learning both Weights and Connections for Efficient Neural Network

Song Han songhan@stanford.edu
Bill Dally dally@stanford.edu
Stanford University
Jeff Pool jpool@nvidia.com
John Tran johntran@nvidia.com
NVIDIA

Neural networks are both computationally intensive and memory intensive, making them difficult to deploy on embedded systems. Also, conventional networks fix the architecture before training starts; as a result, training cannot improve the architecture. To address these limitations, we describe a method to reduce the storage and computation required by neural networks by an order of magnitude without affecting their accuracy, by learning only the important connections. Our method prunes redundant connections using a three-step method. First, we train the network to learn which connections are important. Next, we prune the unimportant connections. Finally, we retrain the network to fine tune the weights of the remaining connections. On the ImageNet dataset, our method reduced the number of parameters of AlexNet by a factor of 9×, from 61 million to 6.7 million, without incurring accuracy loss. Similar experiments with VGG-16 found that the number of parameters can be reduced by 13×, from 138 million to 10.3 million, again with no loss of accuracy.

13 Interactive Control of Diverse Complex Characters with Neural Networks

Igor Mordatch igor.mordatch@gmail.com
Kendall Lowrey kendall.lowrey@gmail.com
Galen Andrew gmandrew@uw.edu
Zoran Popovic zoran@cs.washington.edu
Emanuel Todorov todorov@cs.washington.edu
University of Washington

We present a method for training recurrent neural networks to act as near-optimal feedback controllers. It is able to generate stable and realistic behaviors for a range of dynamical systems and tasks -- swimming, flying, biped and quadruped walking with different body morphologies. It does not require motion capture or task-specific features or state machines. The controller is a neural network, having a large number of feed-forward units that learn elaborate state-action mappings, and a small number of recurrent units that implement memory states beyond the physical system state. The action generated by the network is defined as velocity. Thus the network is not learning a control policy, but rather the dynamics under an implicit policy. Essential features of the method include interleaving supervised learning with trajectory optimization, injecting noise during training, training for unexpected changes in the task specification, and using the trajectory optimizer to obtain optimal feedback gains in addition to optimal actions.

14 Biologically Inspired Dynamic Textures for Probing Motion Perception

Jonathan Vacher vacher@ceremade.dauphine.fr
Université Paris Dauphine
Andrew Isaac Meso andrew.meso@univ-amu.fr
Laurent U Perrinet laurent.perrinet@univ-amu.fr
Institut des neurosciences de la Timone
Gabriel Peyré peyre@ceremade.dauphine.fr
CNRS and Ceremade, Université Paris-Dauphine

Perception is often described as a predictive process based on an optimal inference with respect to a generative model. We study here the principled construction of a generative model specifically crafted to probe motion perception. In that context, we first provide an axiomatic, biologically-driven derivation of the model. This model synthesizes random dynamic textures which are defined by stationary Gaussian distributions obtained by the random aggregation of warped patterns. Importantly, we show that this model can equivalently be described as a stochastic partial differential equation. Using this characterization of motion in images, it allows us to recast motion-energy models into a principled Bayesian inference framework. Finally, we apply these textures in order to psychophysically probe speed perception in humans. In this framework, while the likelihood is derived from the generative model, the prior is estimated from the observed results and accounts for the perceptual bias in a principled fashion.

15 Unsupervised Learning by Program Synthesis

Kevin Ellis ellisk@mit.edu
Armando Solar-Lezama asolar@csail.mit.edu
Josh Tenenbaum jbt@mit.edu
MIT

We introduce an unsupervised learning algorithmthat combines probabilistic modeling with solver-based techniques for program synthesis.We apply our techniques to both a visual learning domain and a language learning problem,showing that our algorithm can learn many visual concepts from only a few examplesand that it can recover some English inflectional morphology.Taken together, these results give both a new approach to unsupervised learning of symbolic compositional structures,and a technique for applying program synthesis tools to noisy data.

16 Deep Poisson Factor Modeling

Ricardo Henao ricardo.henao@duke.edu
Zhe Gan zg27@duke.edu
James Lu james.lu@duke.edu
Lawrence Carin lcarin@duke.edu
Duke University

We propose a new deep architecture for topic modeling, based on Poisson Factor Analysis (PFA) modules. The model is composed of a Poisson distribution to model observed vectors of counts, as well as a deep hierarchy of hidden binary units. Rather than using logistic functions to characterize the probability that a latent binary unit is on, we employ a Bernoulli-Poisson link, which allows PFA modules to be used repeatedly in the deep architecture. We also describe an approach to build discriminative topic models, by adapting PFA modules. We derive efficient inference via MCMC and stochastic variational methods, that scale with the number of non-zeros in the data and binary units, yielding significant efficiency, relative to models based on logistic links. Experiments on several corpora demonstrate the advantages of our model when compared to related deep models.

17 Large-Scale Bayesian Multi-Label Learning via Topic-Based Label Embeddings

Piyush Rai piyush.rai@duke.edu
IIT Kanpur
Changwei Hu ch237@duke.edu
Ricardo Henao ricardo.henao@duke.edu
Lawrence Carin lcarin@duke.edu
Duke University

We present a scalable Bayesian multi-label learning model based on learning low-dimensional label embeddings. Our model assumes that the label vector of each example is generated as a weighted combination of a set of topics (each topic being a distribution over the labels). The combination weights (the embeddings) are assumed conditioned on the observed feature vector via a (nonlinear) regression model. This construction, coupled with a Bernoulli-Poisson link function for each binary label, leads to a model with a computational cost that scales in the number of positive labels in the label matrix. This makes the model particularly appealing for problems where the label matrix is massive but highly sparse. Furthermore, using a data-augmentation strategy leads to full local conjugacy, facilitating simple and very efficient Gibbs sampling as well as Expectation Maximization algorithms for the proposed model. In addition to bringing in the various benefits of a Bayesian formulation, another appealing aspect of our model is that predicting the label vector at test time does not require inferring the label embeddings and can be done in closed-form. We report experimental results on several benchmark data sets comparing our model with various state-of-the art methods.

18 Tensorizing Neural Networks

Alexander Novikov novikov@bayesgroup.ru
Dmitrii Podoprikhin podoprikhin.dmitry@gmail.com
Skolkovo Institute of Science and Technology
Anton Osokin anton.osokin@gmail.com
INRIA
Dmitry P Vetrov vetrovd@yandex.ru
Skoltech, Moscow

Deep neural networks currently demonstrate state-of-the-art performance in several domains.At the same time, models of this class are very demanding in terms of computational resources. In particular, a large amount of memory is required by commonly used fully-connected layers, making it hard to use the models on low-end devices and stopping the further increase of the model size. In this paper we convert the dense weight matrices of the fully-connected layers to the Tensor Train format such that the number of parameters is reduced by a huge factor and at the same time the expressive power of the layer is preserved.In particular, for the Very Deep VGG networks we report the compression factor of the dense weight matrix of a fully-connected layer up to 200000 times leading to the compression factor of the whole network up to 7 times.

19 Training Restricted Boltzmann Machine via the Thouless-Anderson-Palmer free energy

Marylou Gabrie gabrie@clipper.ens.fr
Eric W Tramel eric.tramel@lps.ens.fr
Florent Krzakala florent.krzakala@ens.fr
Ecole Normale Superieure CNRS

Restricted Boltzmann machines are undirected neural networks which have been shown tobe effective in many applications, including serving as initializations fortraining deep multi-layer neural networks. One of the main reasons for their success is theexistence of efficient and practical stochastic algorithms, such as contrastive divergence,for unsupervised training. We propose an alternative deterministic iterative procedure based on an improved mean field method from statistical physics known as the Thouless-Anderson-Palmer approach. We demonstrate that our algorithm provides performance equal to, and sometimes superior to, persistent contrastive divergence, while also providing a clear and easy to evaluate objective function. We believe that this strategycan be easily generalized to other models as well as to more accurate higher-order approximations, paving the way for systematic improvements in training Boltzmann machineswith hidden units.

20 The Brain Uses Reliability of Stimulus Information when Making Perceptual Decisions

Sebastian Bitzer sebastian.bitzer@tu-dresden.de
Stefan Kiebel stefan.kiebel@tu-dresden.de
TU Dresden

In simple perceptual decisions the brain has to identify a stimulus based on noisy sensory samples from the stimulus. Basic statistical considerations state that the reliability of the stimulus information, i.e., the amount of noise in the samples, should be taken into account when the decision is made. However, for perceptual decision making experiments it has been questioned whether the brain indeed uses the reliability for making decisions when confronted with unpredictable changes in stimulus reliability. We here show that even the basic drift diffusion model, which has frequently been used to explain experimental findings in perceptual decision making, implicitly relies on estimates of stimulus reliability. We then show that only those variants of the drift diffusion model which allow stimulus-specific reliabilities are consistent with neurophysiological findings. Our analysis suggests that the brain estimates the reliability of the stimulus on a short time scale of at most a few hundred milliseconds.

21 Unlocking neural population non-stationarities using hierarchical dynamics models

Mijung Park mijung@gatsby.ucl.ac.uk
Gergo Bohner gbohner@gatsby.ucl.ac.uk
Gatsby Unit, UCL
Jakob H Macke jakob@caesar.de
Research center caesar & BCCN Tubingen

Neural population activity often exhibits rich variability. This variability is thought to arise from single-neuron stochasticity, neural dynamics on short time-scales, as well as from modulations of neural firing properties on long time-scales, often referred to as non-stationarity. To better understand the nature of co-variability in neural circuits and their impact on cortical information processing, we introduce a hierarchical dynamics model that is able to capture inter-trial modulations in firing rates, as well as neural population dynamics. We derive an algorithm for Bayesian Laplace propagation for fast posterior inference, and demonstrate that our model provides a better account of the structure of neural firing than existing stationary dynamics models, when applied to neural population recordings from primary visual cortex.

22 Deeply Learning the Messages in Message Passing Inference

Guosheng Lin guosheng.lin@adelaide.edu.au
Chunhua Shen chunhua.shen@adelaide.edu.au
Ian Reid ian.reid@adelaide.edu.au
Anton van den Hengel anton.vandenhengel@adelaide.edu.au
University of Adelaide

Deep structured output learning shows great promise in tasks like semantic image segmentation. We proffer a new, efficient deep structured model learning scheme, in which we show how deep Convolutional Neural Networks (CNNs) can be used to directly estimate the messages in message passing inference for structured prediction with Conditional Random Fields CRFs). With such CNN message estimators, we obviate the need to learn or evaluate potential functions for message calculation. This confers significant efficiency for learning, since otherwise when performing structured learning for a CRF with CNN potentials it is necessary to undertake expensive inference for every stochastic gradient iteration. The network output dimension of message estimators is the same as the number of classes, rather than exponentially growing in the order of the potentials. Hence it is more scalable for cases that a large number of classes are involved. We apply our method to semantic image segmentation and achieve impressive performance, which demonstrates the effectiveness and usefulness of our CNN message learning method.

23 COEVOLVE: A Joint Point Process Model for Information Diffusion and Network Co-evolution

Mehrdad Farajtabar mehrdad@gatech.edu
Yichen Wang ywang737@math.gatech.edu
Shuang Li sli370@gatech.edu
Hongyuan Zha zha@cc.gatech.edu
Le Song lsong@cc.gatech.edu
Georgia Institute of Technology
Manuel Rodriguez manuelgr@mpi-sws.org
MPI SWS

Information diffusion in online social networks is affected by the underlying network topology, but it also has the power to change it. Online users are constantly creating new links when exposed to new information sources, and in turn these links are alternating the way information spreads. However, these two highly intertwined stochastic processes, information diffusion and network evolution, have been predominantly studied separately, ignoring their co-evolutionary dynamics.We propose a temporal point process model, COEVOLVE, for such joint dynamics, allowing the intensity of one process to be modulated by that of the other. This model allows us to efficiently simulate interleaved diffusion and network events, and generate traces obeying common diffusion and network patterns observed in real-world networks. Furthermore, we also develop a convex optimization framework to learn the parameters of the model from historical diffusion and network evolution traces. We experimented with both synthetic data and data gathered from Twitter, and show that our model provides a good fit to the data as well as more accurate predictions than alternatives.

24 The Human Kernel

Andrew G Wilson aglwilson@gmail.com
Christoph Dann cdann@cdann.net
Eric P Xing epxing@cs.cmu.edu
Carnegie Mellon University
Chris Lucas clucas2@inf.ed.ac.uk
University of Edinburgh

Bayesian nonparametric models, such as Gaussian processes, provide a compelling framework for automatic statistical modelling: these models have a high degree of flexibility, and automatically calibrated complexity. However, automating human expertise remains elusive; for example, Gaussian processes with standard kernels struggle on function extrapolation problems that are trivial for human learners. In this paper, we create function extrapolation problems and acquire human responses, and then design a kernel learning framework to reverse engineer the inductive biases of human learners across a set of behavioral experiments. We use the learned kernels to gain psychological insights and to extrapolate in human-like ways that go beyond traditional stationary and polynomial kernels. Finally, we investigate Occam's razor in human and Gaussian process based function learning.

25 Latent Bayesian melding for integrating individual and population models

Mingjun Zhong mzhong@inf.ed.ac.uk
Nigel Goddard nigel.goddard@ed.ac.uk
Charles Sutton csutton@inf.ed.ac.uk
University of Edinburgh

In many statistical problems, a more coarse-grained model may be suitable for population-level behaviour, whereas a more detailed model is appropriate for accurate modelling of individual behaviour. This raises the question of how to integrate both types of models. Methods such as posterior regularization follow the idea of generalized moment matching, in that they allow matchingexpectations between two models, but sometimes both models are most conveniently expressed as latent variable models. We propose latent Bayesian melding, which is motivated by averaging the distributions over populations statistics of both the individual-level and the population-level models under a logarithmic opinion pool framework. In a case study on electricity disaggregation, which is a type of single-channel blind source separation problem, we show that latent Bayesian melding leads to significantly more accurate predictions than an approach based solely on generalized moment matching.

26 High-dimensional neural spike train analysis with generalized count linear dynamical systems

YUANJUN GAO yg2312@columbia.edu
Lars Busing lars@stat.columbia.edu
John Cunningham jpc2181@columbia.edu
University of Columbia
Krishna V Shenoy shenoy@stanford.edu
Stanford University

Latent factor models have been widely used to analyze simultaneous recordings of spike trains from large, heterogeneous neural populations. These models assume the signal of interest in the population is a low-dimensional latent intensity that evolves over time, which is observed in high dimension via noisy point-process observations. These techniques have been well used to capture neural correlations across a population and to provide a smooth, denoised, and concise representation of high-dimensional spiking data. One limitation of many current models is that the observation model is assumed to be Poisson, which lacks the flexibility to capture under- and over-dispersion that is common in recorded neural data, thereby introducing bias into estimates of covariance. Here we develop the generalized count linear dynamical system, which relaxes the Poisson assumption by using a more general exponential family for count data. In addition to containing Poisson, Bernoulli, negative binomial, and other common count distributions as special cases, we show that this model can be tractably learned by extending recent advances in variational inference techniques. We apply our model to data from primate motor cortex and demonstrate performance improvements over state-of-the-art methods, both in capturing the variance structure of the data and in held-out prediction.

27 Efficient Continuous-Time Hidden Markov Model for Disease Modeling

Yu-Ying Liu yuyingliu0823@gmail.com
Fuxin Li fli@cc.gatech.edu
Shuang Li saliumass@gmail.com
Le Song lsong@cc.gatech.edu
James M Rehg rehg@gatech.edu
Georgia Institute of Technology

Continuous-Time Hidden Markov Model (CT-HMM) is useful for modeling disease progression with noisy observed data arriving irregularly in time. However, the lack of widely-accepted efficient learning algorithm for CT-HMM restricts its use with very small models or requires unrealistic assumptions on the state transition timing. In this paper, we present the first complete characterization of EM-based learning methods in CT-HMM models, which involves two challenges: the estimation of posterior state probabilities, and the computation of end-state conditioned statistics in a continuous-time markov chain. Without loss of generality, we efficiently discretize the estimation of posterior state probabilities into a discrete time-inhomogeneous hidden Markov model, and present two versions using either the forward-backward algorithm or the Viterbi decoding. Both versions are analyzed and compared in conjunction with three recent approaches for efficiently computing the end-state conditioned statistics. Finally, we demonstrate the use of CT-HMM to visualize and predict future disease measurements using a Glaucoma dataset and an Alzheimer disease dataset. Our results show that CT-HMM outperforms the state-of-the-art method [1], for glaucoma prediction.

28 The Population Posterior and Bayesian Modeling on Streams

James McInerney jm4181@columbia.edu
David Blei david.blei@columbia.edu
Columbia University
Rajesh Ranganath rajeshr@cs.princeton.edu
Princeton University

Many modern data analysis problems involve inferences from streaming data. However, streaming data is not easily amenable to the standard probabilistic modeling approaches, which assume that we condition on finite data. We develop population variational Bayes, a new approach for using Bayesian modeling to analyze streams of data. It approximates a new type of distribution, the population posterior, which combines the notion of a population distribution of the data with Bayesian inference in a probabilistic model. We study our method with latent Dirichlet allocation and Dirichlet process mixtures on several large-scale data sets.

29 Probabilistic Curve Learning: Coulomb Repulsion and the Electrostatic Gaussian Process

Ye Wang eric.ye.wang@duke.edu
David B Dunson dunson@stat.duke.edu
Duke University

Learning of low dimensional structure in multidimensional data is a canonical problem in machine learning. One common approach is to suppose that the observed data are close to a lower-dimensional smooth manifold. There are a rich variety of manifold learning methods available, which allow mapping of

data points to the manifold. However, there is a clear lack of probabilistic methods that allow learning of the manifold along with the generative distribution of the observed data. The best attempt is the Gaussian process latent variable model (GP-LVM), but identifiability issues lead to poor performance. We solve these issues by proposing a novel Coulomb repulsive process (Corp) for locations of points on the manifold, inspired by physical models of electrostatic interactions among particles. Combining this process with a GP prior for the mapping function yields a novel electrostatic GP (electroGP) process. Focusing on the simple case of a one-dimensional manifold, we develop efficient inference algorithms, and illustrate substantially improved performance in a variety of experiments including filling in missing frames in video.

30 Preconditioned Spectral Descent for Deep Learning

David E Carlson david.carlson@duke.edu
Lawrence Carin lcarin@duke.edu
Duke University
Edo Collins edo.collins@epfl.ch
Ya-Ping Hsieh ya-ping.hsieh@epfl.ch
Volkan Cevher volkan.cevher@epfl.ch
EPFL

Optimizing objective functions in deep learning is a notoriously difficult task. Classical algorithms, including variants of gradient descent and quasi-Newton methods, can be interpreted as approximations to the objective function in Euclidean norms. However, it has recently been shown that approximations via non-Euclidean norms can significantly improve optimization performance. In this paper, we provide evidences that neither of the above two methods entirely capture the ``geometry'' of the objective functions in deep learning, while a combination of the two does. We theoretically formalize our arguments and derive a novel second-order non-Euclidean algorithm. We implement our algorithms on Restricted Boltzmann Machines, Feedforward Neural Nets, and Convolutional Neural Nets, and demonstrate improvements in computational efficiency as well as model fit.

31 Learning Continuous Control Policies by Stochastic Value Gradients

Nicolas Heess heess@google.com
Greg Wayne gregwayne@google.com
David Silver davidsilver@google.com
Tim Lillicrap countzero@google.com
Tom Erez etom@google.com
Yuval Tassa tassa@google.com
Google DeepMind

Policy gradient methods based on likelihood ratio estimators optimize stochastic policies in stochastic environments but can converge slowly because their gradient estimates have high variance. In continuous state-action spaces, value gradient methods can provide lower-variance gradient estimators but are limited to optimizing deterministic policies in deterministic environments. By considering all stochasticity in the Bellman equation as a deterministic function of exogenous noise, we create a formalism that enables the optimization of stochastic policies in stochastic environments using simple backpropagation. Our formalism leads to a spectrum of new algorithms for policy learning that range from a model-free to a value function critic-free variant. We apply these algorithms first to a toy stochastic control problem and then to a range of difficult physical control problems in simulation. One of these variants, SVG(1), shows the effectiveness of learning a model, a value function, and a policy simultaneously in a continuous domain.

32 Learning Stationary Time Series using Gaussian Processes with Nonparametric Kernels

Felipe Tobar ftobar@dim.uchile.cl
Universidad de Chile
Thang Bui tdb40@cam.ac.uk
Richard E Turner ret26@cam.ac.uk
University of Cambridge

We present a doubly non-parametric generative model for stationary signals based on the convolution between a continuous-time white-noise process and a continuous-time linear filter defined by a Gaussian process. The model is a continuous-time non-parametric generalisation of a moving average process and, conditionally, is itself a Gaussian process with a nonparametric kernel defined in a probabilistic fashion. One of the main contributions of the paper is to develop a novel variational free-energy approach based on inducing points that efficiently learns the continuous-time linear filter and infers the driving white-noise process. In turn, this scheme provides closed-form probabilistic estimates of the kernel and the noise-free signal. The generative model can be equivalently considered in the frequency domain, where the spectral density of the signal is specified using a Gaussian process. The variational inference procedure provides closed-form expressions for the posterior of the spectral density given the observed data, leading to new non-parametric Bayesian approaches to spectrum estimation. The flexibility of the new model is validated using synthetic and real-world signals.

33 Path-SGD: Path-Normalized Optimization in Deep Neural Networks

Behnam Neyshabur bneyshabur@ttic.edu
Nati Srebro nati@ttic.edu
Toyota Technological Institute at Chicago
Russ R Salakhutdinov rsalakhu@cs.toronto.edu
University of Toronto

We revisit the choice of SGD for training deep neural networks by reconsidering the appropriate geometry in which to optimize the weights. We argue for a geometry invariant to rescaling of weights that does not affect the output of the network, and suggest Path-SGD, which is an approximate steepest descent method with respect to a path-wise regularizer related to max-norm regularization. Path-SGD is easy and efficient to implement and leads to empirical gains over SGD and AdaGrad.

34 Automatic Variational Inference in Stan

Alp Kucukelbir alp@cs.columbia.edu
Andrew Gelman gelman@stat.columbia.edu
David Blei david.blei@columbia.edu
Columbia University
Rajesh Ranganath rajeshr@cs.princeton.edu
Princeton University

Variational inference is a scalable technique for approximate Bayesian inference. Deriving variational inference algorithms requires tedious model-specific calculations; this makes it difficult to automate. We propose an automatic variational inference algorithm, automatic differentiation variational inference (ADVI). The user only provides a Bayesian model and a dataset; nothing else. We make no conjugacy assumptions and support a broad class of models. The algorithm automatically determines an appropriate variational family and optimizes the variational objective. We implement ADVI in Stan (code available now), a probabilistic programming framework. We compare ADVI to MCMC sampling across hierarchical generalized linear models, nonconjugate matrix factorization, and a mixture model. We train the mixture model on a quarter million images. With ADVI we can use variational inference on any model we write in Stan.

35 Data Generation as Sequential Decision Making

Philip Bachman phil.bachman@gmail.com
Doina Precup dprecup@cs.mcgill.ca
University of McGill

We connect a broad class of generative models through their shared reliance on sequential decision making. We show how changes motivated by our point of view can improve an already-strong model, and then explore this idea further in the context of data imputation -- perhaps the simplest setting in which to investigate the relation between unconditional and conditional generative modelling. We formulate data imputation as an MDP and develop models capable of representing effective policies for it. We construct our models using neural networks and train them using a form of guided policy search. Our models generate predictions through an iterative process of feedback and refinement. With empirical tests, we show that our approach can learn effective policies for imputation problems of varying difficulty and across multiple datasets.

36 Stochastic Expectation Propagation

Yingzhen Li yl494@cam.ac.uk
Richard E Turner ret26@cam.ac.uk
University of Cambridge
José Miguel Hernández-Lobato jmh@seas.harvard.edu
Harvard

Expectation propagation (EP) is a deterministic approximation algorithm that is often used to perform approximate Bayesian parameter learning. EP approximates the full intractable posterior distribution through a set of local-approximations that are iteratively refined for each datapoint. EP can offer analytic and computational advantages over other approximations, such as Variational Inference (VI), and is the method of choice for a number of models. The local nature of EP appears to make it an ideal candidate for performing Bayesian learning on large models in large-scale datasets settings. However, EP has a crucial limitation in this context: the number approximating factors needs to increase with the number of data-points, N, which often entails a prohibitively large memory overhead. This paper presents an extension to EP, called stochastic expectation propagation (SEP), that maintains a global posterior approximation (like VI) but updates it in a local way (like EP). Experiments on a number of canonical learning problems using synthetic and real-world datasets indicate that SEP performs almost as well as full EP, but reduces the memory consumption by a factor of N. SEP is therefore ideally suited to performing approximate Bayesian learning in the large model, large dataset setting.

37 Deep learning with Elastic Averaging SGD

Sixin Zhang zsx@cims.nyu.edu
Yann LeCun yann@cs.nyu.edu
New York University
Anna E Choromanska achoroma@cims.nyu.edu
Courant Institute, NYU

We study the problem of stochastic optimization for deep learning in the parallel computing environment under communication constraints. A new algorithm is proposed in this setting where the communication and coordination of work among concurrent processes (local workers), is based on an elastic force which links the parameters they compute with a center variable stored by the parameter server (master). The algorithm enables the local workers to perform more exploration, i.e. the algorithm allows the local variables to fluctuate further from the center variable by reducing the amount of communication between local workers and the master. We empirically demonstrate that in the deep learning setting, due to the existence of many local optima, allowing more exploration can lead to the improved performance. We propose synchronous and asynchronous variants of the new algorithm. We provide the stability analysis of the asynchronous variant in the round-robin scheme and compare it with the more common parallelized method ADMM. We show that the stability of EASGD is guaranteed when a simple stability condition is satisfied, which is not the case for ADMM. We additionally propose the momentum-based version of our algorithm that can be applied in both synchronous and asynchronous settings. Asynchronous variant of the algorithm is applied to train convolutional neural networks for image classification on the CIFAR and ImageNet datasets. Experiments demonstrate that the new algorithm accelerates the training of deep architectures compared to DOWNPOUR and other common baseline approaches and furthermore is very communication efficient.

38 Learning with Group Invariant Features: A Kernel Perspective.

Youssef Mroueh ymroueh@mit.edu
IBM
Stephen Voinea voinea@mit.edu
Tomaso A Poggio tp@ai.mit.edu
MIT

We analyze in this paper a random feature map based on a theory of invariance (\emph{I-theory}) introduced in \ cite{AnselmiLRMTP13}. More specifically, a group invariant signal signature is obtained through cumulative distributions of group-transformed random projections. Our analysis bridges invariant feature learning with kernel methods, as we show that this feature map defines an expected Haar-integration kernel that is invariant to the specified group action. We show how this non-linear random feature map approximates this group invariant kernel uniformly on a set of N points. Moreover, we show that it defines a function space that is dense in the equivalent Invariant Reproducing Kernel Hilbert Space. Finally, we quantify error rates of the convergence of the empirical risk minimization, as well as the reduction in the sample complexity of a learning algorithm using such an invariant representation for signal classification, in a classical supervised learning setting

39 Linear Response Methods for Accurate Covariance Estimates from Mean Field Variational Bayes

Ryan J Giordano rgiordano@berkeley.edu
Michael I Jordan jordan@cs.berkeley.edu
UC Berkeley
Tamara Broderick tbroderick@csail.mit.edu
MIT

Mean field variational Bayes (MFVB) is a popular posterior approximation method due to its fast runtime on large-scale data sets. However, it is well known that a major failing of MFVB is that it underestimates the uncertainty of model variables (sometimes severely) and provides no information about model variable covariance. We generalize linear response methods from statistical physics to deliver accurate uncertainty estimates for model variables---both for individual variables and coherently across variables. We call our method linear response variational Bayes (LRVB). When the MFVB posterior approximation is in the exponential family, LRVB has a simple, analytic form, even for non-conjugate models. Indeed, we make no assumptions about the form of the true posterior. We demonstrate the accuracy and scalability of our method on a range of models for both simulated and real data.

40 Probabilistic Line Searches for Stochastic Optimization

Maren Mahsereci mmahsereci@tue.mpg.de
Philipp Hennig phennig@tue.mpg.de
MPI for Intelligent Systems, Tübingen

In deterministic optimization, line searches are a standard tool ensuring stability and efficiency. Where only stochastic gradients are available, no direct equivalent has so far been formulated, because uncertain gradients do not allow for a strict sequence of decisions collapsing the search space. We construct a probabilistic line search by combining the structure of existing deterministic methods with notions from Bayesian optimization. Our method retains a Gaussian process surrogate of the univariate optimization objective, and uses a probabilistic belief over the Wolfe conditions to monitor the descent. The algorithm has very low computational cost, and no user-controlled parameters. Experiments show that it effectively removes the need to define a learning rate for stochastic gradient descent.

41 A hybrid sampler for Poisson-Kingman mixture models

Maria Lomeli mlomeli@gatsby.ucl.ac.uk
Gatsby Unit, UCL
Stefano Favaro stefano.favaro@unito.it
University of Torino and Collegio Carlo Alberto
Yee Whye Teh y.w.teh@stats.ox.ac.uk
University of Oxford

This paper concerns the introduction of a new Markov Chain Monte Carlo scheme for posterior sampling in Bayesian nonparametric mixture models with priors that belong to the general Poisson-Kingman class. We present a novel and compact way of representing the infinite dimensional component of the model such that while explicitly representing this infinite component it has less memory and storage requirements than previous MCMC schemes. We describe comparative simulation results demonstrating the efficacy of the proposed MCMC algorithm against existing marginal and conditional MCMC samplers.

42 Tree-Guided MCMC Inference for Normalized Random Measure Mixture Models

Juho Lee stonecold@postech.ac.kr
Seungjin Choi seungjin@postech.ac.kr
POSTECH

Normalized random measures (NRMs) provide a broad class of discrete random measures that are often used as priors for Bayesian nonparametric models. Dirichlet process is a well-known example of NRMs. Most of posterior inference methods for NRM mixture models rely on MCMC methods since they are easy to implement and their convergence is well studied. However, MCMC often suffers from slow convergence when the acceptance rate is low. Tree-based inference is an alternative deterministic posterior inference method, where Bayesian hierarchical clustering (BHC) or incremental Bayesian hierarchical clustering (IBHC) have been developed for DP or NRM mixture (NRMM) models, respectively. Although IBHC is a promising method for posterior inference for NRMM models due to its efficiency and applicability to online inference, its convergence is not guaranteed since it uses heuristics that simply selects the best solution after multiple trials are made. In this paper, we present a hybrid inference algorithm for NRMM models, which combines the merits of both MCMC and IBHC. Trees built by IBHC outlinespartitions of data, which guides Metropolis-Hastings procedure to employ appropriate proposals. Inheriting the nature of MCMC, our tree-guided MCMC (tgMCMC) is guaranteed to converge, and enjoys the fast convergence thanks to the effective proposals guided by trees. Experiments on both synthetic and real world datasets demonstrate the benefit of our method.

43 Reflection, Refraction, and Hamiltonian Monte Carlo

Hadi Mohasel Afshar h.m.afshar@gmail.com
Australian National University
Justin Domke justin.domke@nicta.com.au
NICTA

Hamiltonian Monte Carlo (HMC) is a successful approach for sampling from continuous densities. However, it has difficulty simulating Hamiltonian dynamics with non-smooth functions, leading to poor performance. This paper is motivated by the behavior of Hamiltonian dynamics in physical systems like optics. We introduce a modification of the Leapfrog discretization of Hamiltonian dynamics on piecewise continuous energies, where intersections of the trajectory with discontinuities are detected, and the momentum is reflected or refracted to compensate for the change in energy. We prove that this method preserves the correct stationary distribution when boundaries are affine. Experiments show that by reducing the number of rejected samples, this method improves on traditional HMC.

44 Planar Ultrametrics for Image Segmentation

Julian E Yarkony julian.e.yarkony@gmail.com
Charless Fowlkes fowlkes@ics.uci.edu
UC Irvine

We study the problem of hierarchical clustering on planar graphs. We formulate this in terms of finding the closest ultrametric to a specified set of distances and solve it using an LP relaxation that leverages minimum cost perfect matching as a subroutine to efficiently explore the space of planar partitions. We apply our algorithm to the problem of hierarchical image segmentation.

45 Learning Bayesian Networks with Thousands of Variables

Mauro Scanagatta mauro@idsia.ch
Giorgio Corani giorgio@idsia.ch
Marco Zaffalon zaffalon@idsia.ch
IDSIA
Cassio P de Campos cassiopc@gmail.com
Queen's University Belfast

We present a method for learning Bayesian networks from data sets containingthousands of variables without the need for structure constraints. Our approachis made of two parts. The first is a novel algorithm that effectively explores thespace of possible parent sets of a node. It guides the exploration towards themost promising parent sets on the basis of an approximated score function thatis computed in constant time. The second part is an improvement of an existingordering-based algorithm for structure optimization. The new algorithm provablyachieves a higher score compared to its original formulation. On very large datasets containing up to ten thousand nodes, our novel approach consistently outper-forms the state of the art.

46 Parallel Predictive Entropy Search for Batch Global Optimization of Expensive Objective Functions

Amar Shah as793@cam.ac.uk
Zoubin Ghahramani zoubin@eng.cam.ac.uk
University of Cambridge

We develop \textit{parallel predictive entropy search} (PPES), a novel algorithm for Bayesian optimization of expensive black-box objective functions. At each iteration, PPES aims to select a \textit{batch} of points which will maximize the information gain about the global maximizer of the objective. Well known strategies exist for suggesting a single evaluation point based on previous observations, while far fewer are known for selecting batches of points to evaluate in parallel. The few batch selection schemes that have been studied all resort to greedy methods to compute an optimal batch. To the best of our knowledge, PPES is the first non-greedy batch Bayesian optimization strategy. We demonstrate the benefit of this approach in optimization performance on both synthetic and real world applications, including problems in machine learning, rocket science and robotics.

47 Rapidly Mixing Gibbs Sampling for a Class of Factor Graphs Using Hierarchy Width

Christopher M De Sa cdesa@stanford.edu
Kunle Olukotun kunle@stanford.edu
Chris Ré chrismre@cs.stanford.edu
Stanford
Ce Zhang czhang@cs.wisc.edu
Wisconsin

Gibbs sampling on factor graphs is a widely used inference technique, which often produces good empirical results. Theoretical guarantees for its performance are weak: even for tree structured graphs, the mixing time of Gibbs may be exponential in the number of variables. To help understand the behavior of Gibbs sampling, we introduce a new (hyper)graph property, called hierarchy width. We show that under suitable conditions on the weights, bounded hierarchy width ensures polynomial mixing time. Our study of hierarchy width is in part motivated by a class of factor graph templates, hierarchical templates, which have bounded hierarchy width—regardless of the data used to instantiate them. We demonstrate a rich application from natural language processing in which Gibbs sampling provably mixes rapidly and achieves accuracy that exceeds human volunteers.

48 On some provably correct cases of variational inference for topic models

Pranjal Awasthi pranjal.iitm@gmail.com
Andrej Risteski risteski@princeton.edu
Princeton

Variational inference is an efficient, popular heuristic used in the context of latent variable models. We provide the first analysis of instances where variational inference algorithms converge to the global optimum, in the setting of topic models. Our initializations are natural, one of them being used in LDA-c, the mostpopular implementation of variational inference.In addition to providing intuition into why this heuristic might work in practice, the multiplicative, rather than additive nature of the variational inference updates forces us to usenon-standard proof arguments, which we believe might be of general theoretical interest.

49 Large-scale probabilistic predictors with and without guarantees of validity

Vladimir Vovk v.vovk@rhul.ac.uk
Royal Holloway, Univ of London
Ivan Petej ivan.petej@gmail.com
Valentina Fedorova alushaf@gmail.com
Yandex

This paper studies theoretically and empirically a method of turning machine-learning algorithms into probabilistic predictors that automatically enjoys a property of validity (perfect calibration), is computationally efficient, and preserves predictive efficiency. The price to pay for perfect calibration is that these probabilistic predictors produce imprecise (in practice, almost precise for large data sets) probabilities. When these imprecise probabilities are merged into precise probabilities, the resulting predictors, while losing the theoretical property of perfect calibration, consistently outperform the existing methods in empirical studies.

50 On the Accuracy of Self-Normalized Log-Linear Models

Jacob Andreas jda@cs.berkeley.edu
Maxim Rabinovich rabinovich@eecs.berkeley.edu
Michael I Jordan jordan@cs.berkeley.edu
Dan Klein klein@cs.berkeley.edu
UC Berkeley

Calculation of the log-normalizer is a major computational obstacle in applications of log-linear models with large output spaces. The problem of fast normalizer computation has therefore attracted significant attention in the theoretical and applied machine learning literature. In this paper, we analyze a recently proposed technique known as ``self-normalization'', which introduces a regularization term in training to penalize log normalizers for deviating from zero. This makes it possible to use unnormalized model scores as approximate probabilities. Empirical evidence suggests that self-normalization is extremely effective, but a theoretical understanding of why it should work, and how generally it can be applied, is largely lacking.We prove upper bounds on the loss in accuracy due to self-normalization, describe classes of input distributionsthat self-normalize easily, and construct explicit examples of high-variance input distributions. Our theoretical results make predictions about the difficulty of fitting self-normalized models to several classes of distributions, and we conclude with empirical validation of these predictions on both real and synthetic datasets.

51 Policy Evaluation Using the Ω-Return

Philip S Thomas pthomas@cs.umass.edu
UMass, Carnegie Mellon University
Scott Niekum sniekum@cs.utexas.edu
UT Austin
Georgios Theocharous theochar@adobe.com
Adobe
George Konidaris gdk@cs.duke.edu
Duke

We propose the Ω-return as an alternative to the λ-return currently used by the TD(λ) family of algorithms. The Ω-return accounts for the correlation of different length returns, and we provide empirical studies that suggest that it is superior to the λ-return and γ-return for a variety of problems.We propose the Ω-return as an alternative to the λ-return currently used by the TD(λ) family of algorithms. The benefit of the Ω-return is that it accounts for the correlation of different length returns. Because it is difficult to compute exactly, we suggest one way of approximating the Ω-return. We provide empirical studies that suggest that it is superior to the λ-return and γ-return for a variety of problems.

52 Community Detection via Measure Space Embedding

Mark Kozdoba mark.kozdoba@gmail.com
Shie Mannor shie@ee.technion.ac.il
Technion

We present a new algorithm for community detection. The algorithm uses random walks to embed the graph in a space of measures, after which a modification of k-means in that space is applied. The algorithm is therefore fast and easily parallelizable. We evaluate the algorithm on standard random graph benchmarks, including some overlapping community benchmarks, and find its performance to be better or at least as good as previously known algorithms. We also prove a linear time (in number of edges) guarantee for the algorithm on a p,q-stochastic block model with where $p \geq c \cdot N^{-1/2+\epsilon}$ and $p-q \geq c'\sqrt{(p N^{-1/2+\epsilon} \log N)}$.

53 The Consistency of Common Neighbors for Link Prediction in Stochastic Blockmodels

Purnamrita Sarkar purna.sarkar@austin.utexas.edu
Deepayan Chakrabarti deepay@utexas.edu
UT Austin
peter j bickel bickel@stat.berkeley.edu
U C Berkeley

Link prediction and clustering are key problems for network-structureddata. While spectral clustering has strong theoretical guaranteesunder the popular stochastic blockmodel formulation of networks, itcan be expensive for large graphs. On the other hand, the heuristic ofpredicting links to nodes that share the most common neighbors withthe query node is much fast, and works very well in practice. We showtheoretically that the common neighbors heuristic can extract clustersw.h.p. when the graph is dense enough, and can do so even in sparsergraphs with the addition of a ``cleaning'' step. Empirical results onsimulated and real-world data support our conclusions.

54 Inference for determinantal point processes without spectral knowledge

Rémi Bardenet remi.bardenet@gmail.com
University of Lille
Michalis Titsias mtitsias@gmail.com
Athens University of Economics and Business

Determinantal point processes (DPPs) are point process models thatnaturally encode diversity between the points of agiven realization, through a positive definite kernel K. DPPs possess desirable properties, such as exactsampling or analyticity of the moments, but learning the parameters ofkernel K through likelihood-based inference is notstraightforward. First, the kernel that appears in thelikelihood is not K, but another kernel L related to K throughan often intractable spectral decomposition. This issue is typically bypassed in machine learning bydirectly parametrizing the kernel L, at the price of someinterpretability of the model parameters. We follow this approachhere. Second, the likelihood has an intractable normalizingconstant, which takes the form of large determinant in the case of aDPP over a finite set of objects, and the form of a Fredholm determinant in thecase of a DPP over a continuous domain. Our main contribution is to derive bounds on the likelihood ofa DPP, both for finite and continuous domains. Unlike previous work, our bounds arecheap to evaluate since they do not rely on approximating the spectrumof a large matrix or an operator. Through usual arguments, these bounds thus yield cheap variationalinference and moderately expensive exact Markov chain Monte Carlo inference methods for DPPs.

55 Sample Complexity of Learning Mahalanobis Distance Metrics

Nakul Verma nakulverma@gmail.com
Kristin Branson bransonk@janelia.hhmi.org
Janelia Research Campus, HHMI

Metric learning seeks a transformation of the feature space that enhances prediction quality for a given task. In this work we provide PAC-style sample complexity rates for supervised metric learning. We give matching lower- and upper-bounds showing that sample complexity scales with the representation dimension when no assumptions are made about the underlying data distribution. In addition, by leveraging the structure of the data distribution, we provide rates fine-tuned to a specific notion of the intrinsic complexity of a given dataset, allowing us to relax the dependence on representation dimension. We show both theoretically and empirically that augmenting the metric learning optimization criterion with a simple norm-based regularization is important and can help adapt to a dataset's intrinsic complexity yielding better generalization, thus partly explaining the empirical success of similar regularizations reported in previous works.

56 Manifold Optimization for Gaussian Mixture Models

Reshad Hosseini reshadh@gmail.com
University of Tehran
Suvrit Sra suvrit@mit.edu
MIT

We take a new look at parameter estimation for Gaussian Mixture Models (GMMs). In particular, we propose using \emph{Riemannian manifold optimization} as a powerful counterpart to Expectation Maximization (EM). An out-of-the-box invocation of manifold optimization, however, fails spectacularly: it converges to the same solution but vastly slower. Driven by intuition from manifold convexity, we then propose a reparamerization that has remarkable empirical consequences. It makes manifold optimization not only match EM---a highly encouraging result in itself given the poor record nonlinear programming methods have had against EM so far---but also outperform EM in many practical settings, while displaying much less variability in running times. We further highlight the strengths of manifold optimization by developing a somewhat tuned manifold LBFGS method that proves even more competitive and reliable than existing manifold optimization tools. We hope that our results encourage a wider consideration of manifold optimization for parameter estimation problems.

57 Frank-Wolfe Bayesian Quadrature: Probabilistic Integration with Theoretical Guarantees

François-Xavier Briol f-x.briol@warwick.ac.uk
University of Warwick
Chris Oates christopher.oates@uts.edu.au
University of Tech., Sydney
Mark Girolami m.girolami@warwick.ac.uk
Univeresity of Warwick
Michael A Osborne mosb@robots.ox.ac.uk
U Oxford

There is renewed interest in formulating integration as an inference problem, motivated by obtaining a full distribution over numerical error that can be propagated through subsequent computation. Current methods, such as Bayesian Quadrature, demonstrate impressive empirical performance but lack theoretical analysis. An important challenge is to reconcile these probabilistic integrators with rigorous convergence guarantees. In this paper, we present the first probabilistic integrator that admits such theoretical treatment, called Frank-Wolfe Bayesian Quadrature (FWBQ). Under FWBQ, convergence to the true value of the integral is shown to be exponential and posterior contraction rates are proven to be superexponential. In simulations, FWBQ is competitive with state-of-the-art methods and out-performs alternatives based on Frank-Wolfe optimisation. Our approach is applied to successfully quantify numerical error in the solution to a challenging model choice problem in cellular biology.

58 Scale Up Nonlinear Component Analysis with Doubly Stochastic Gradients

Bo Xie bo.xie@gatech.edu
Le Song lsong@cc.gatech.edu
Georgia Institute of Technology
Yingyu Liang yingyul@cs.princeton.edu
Princeton University

Nonlinear component analysis such as kernel Principle Component Analysis (KPCA) and kernel Canonical Correlation Analysis (KCCA) are widely used in machine learning, statistics and data analysis, but they can not scale up to big datasets. Recent attempts have employed random feature approximations to convert the problem to the primal form for linear computational complexity. However, to obtain high quality solutions, the number of random features should be the same order of magnitude as the number of data points, making such approach not directly applicable to the regime with millions of data points.We propose a simple, computationally efficient, and memory friendly algorithm based on the ``doubly stochastic gradients'' to scale up a range of

kernel nonlinear component analysis, such as kernel PCA, CCA and SVD. Despite the {non-convex} nature of these problems, our method enjoys theoretical guarantees that it converges at the rate \Otil(1/t) to the global optimum, even for the top k eigen subspace. Unlike many alternatives, our algorithm does not require explicit orthogonalization, which is infeasible on big datasets. We demonstrate the effectiveness and scalability of our algorithm on large scale synthetic and real world datasets.

59 The Self-Normalized Estimator for Counterfactual Learning

Adith Swaminathan adith@cs.cornell.edu
Thorsten Joachims tj@cs.cornell.edu
Cornell University

This paper introduces a new counterfactual estimator for batch learning from logged bandit feedback (BLBF).In this setting, the learner does not receive full-information feedback like in supervised learning, but observes the feedback only for the actions taken by a historical policy.This makes BLBF algorithms particularly attractive for training online systems (e.g., ad placement, web search, recommendation) using their historical logs.The Counterfactual Risk Minimization (CRM) principleoffers a general recipe for designing BLBF algorithms. It requires a counterfactual risk estimator, and virtually all existing works on BLBF have focused on a particular unbiased estimator. However, we show that this conventional estimator suffers from apropensity overfitting problem when used for learning over complex hypothesis spaces.We propose to replace the unbiased risk estimator with a self-normalized estimator, showing that it neatly avoids this problem.This naturally gives rise to a new learning algorithm -- Normalized Policy Optimizer for Exponential Models (Norm-POEM) --for structured output prediction using linear rules.We evaluate the empirical effectiveness of Norm-POEM on severalmulti-label classification problems, finding that it consistently outperforms the conventional unbiased estimator.

60 Distributionally Robust Logistic Regression

Soroosh Shafieezadeh Abadeh soroosh.shafiee@epfl.ch
Peyman Esfahani peyman.mohajerin@epfl.ch
Daniel Kuhn daniel.kuhn@epfl.ch
EPFL

This paper proposes a distributionally robust approach to logistic regression. We use the Wasserstein distance to construct a ball in the space of probability distributions centered at the uniform distribution on the training samples. If the radius of this Wasserstein ball is chosen judiciously, we can guarantee that it contains the unknown data-generating distribution with high confidence. We then formulate a distributionally robust logistic regression model that minimizes a worst-case expected logloss function, where the worst case is taken over all distributions in the Wasserstein ball. We prove that this optimization problem admits a tractable reformulation and encapsulates the classical as well as the popular regularized logistic regression problems as special cases. We further propose a distributionally robust approach based on Wasserstein balls to compute upper and lower confidence bounds on the misclassification probability of the resulting classifier. These bounds are given by the optimal values of two highly tractable linear programs. We validate our theoretical out-of-sample guarantees through simulated and empirical experiments.

61 Top-k Multiclass SVM

Maksim Lapin mlapin@mpi-inf.mpg.de
Bernt Schiele schiele@mpi-inf.mpg.de
Max Planck Institute for Informatics
Matthias Hein hein@cs.uni-sb.de
Saarland University

The issue of class ambiguity is typical in image and scene classification problems with a large number of classes. Therefore, it makes sense to use top-k error to evaluate the classifier instead of the standard zero-one loss. We propose top-k multiclass SVM as a direct method to optimize for top-k performance. Our generalization of the multiclass SVM is based on a tight convex upper bound of the top-k error. We propose a fast optimization scheme based on an efficient projection onto the top-k simplex, which is of its own interest. Experiments on five datasets show consistent improvements in top-k accuracy compared to various baselines.

62 Measuring Sample Quality with Stein's Method

Jackson Gorham jacksongorham@gmail.com
Lester Mackey lmackey@stanford.edu
Stanford University

To improve the efficiency of Monte Carlo estimation, practitioners are turning to biased Markov chain Monte Carlo procedures that trade off asymptotic exactness for computational speed. The reasoning is sound: a reduction in variance due to more rapid sampling can outweigh the bias introduced. However, the inexactness creates new challenges for sampler and parameter selection, since standard measures of sample quality like effective sample size do not account for asymptotic bias. To address these challenges, we introduce a new computable quality measure based on Stein's method that bounds the discrepancy between sample and target expectations over a large class of test functions. We use our tool to compare exact, biased, and deterministic sample sequences and illustrate applications to hyperparameter selection, convergence rate assessment, and quantifying bias-variance tradeoffs in posterior inference.

63 Asynchronous Parallel Stochastic Gradient for Nonconvex Optimization

Xiangru Lian lianxiangru@gmail.com
Yijun Huang huangyj0@gmail.com
Yuncheng Li raingomm@gmail.com
Ji Liu jliu@cs.rochester.edu
University of Rochester

The asynchronous parallel implementations of stochastic gradient (SG) have been broadly used in solving deep neural network and received many successes in practice recently. However, existing theories cannot explain their convergence and speedup properties, mainly due to the nonconvexity of most deep learning formulations and the asynchronous parallel mechanism. To fill the gaps in theory and provide theoretical supports, this paper studies two asynchronous parallel implementations of SG: one is on the computer network and the other is on the shared memory system. We establish an ergodic convergence rate O(1/ K−−√) for both algorithms and prove that the linear speedup is achievable if the number of workers is bounded by K−−√ (K is the total number of iterations). Our results generalize and improve existing analysis for convex minimization.

64 Solving Random Quadratic Systems of Equations Is Nearly as Easy as Solving Linear Systems

Yuxin Chen yxchen@stanford.edu
Emmanuel Candes candes@stanford.edu
Stanford University

This paper is concerned with finding a solution x to a quadratic system of equations $y_i = |< a_i, x >|^2$, i = 1, 2, ..., m. We prove that it is possible to solve unstructured quadratic systems in n variables exactly from O(n) equations in linear time, that is, in time proportional to reading and evaluating the data. This is accomplished by a novel procedure, which starting from an initial guess given by a spectral initialization procedure, attempts to minimize a non-convex objective. The proposed algorithm distinguishes from prior approaches by regularizing the initialization and descent procedures in an adaptive fashion, which discard terms bearing too much influence on the initial estimate or search directions. These careful selection rules---which effectively serve as a variance reduction scheme---provide a tighter initial guess, more robust descent directions, and thus enhanced practical performance. Further, this procedure also achieves a near-optimal statistical accuracy in the presence of noise. Finally, we demonstrate empirically that the computational cost of our algorithm is about four times that of solving a least-squares problem of the same size.

65 Distributed Submodular Cover: Succinctly Summarizing Massive Data

Baharan Mirzasoleiman baharanm@student.ethz.ch
Andreas Krause krausea@ethz.ch
ETHZ
Amin Karbasi amin.karbasi@yale.edu
Yale
Ashwinkumar Badanidiyuru ashwinkumarbv@google.com
Google Research

How can one find a subset, ideally as small as possible, that well represents a massive dataset? I.e., its corresponding utility, measured according to a suitable utility function, should be comparable to that of the whole dataset. In this paper, we formalize this challenge as a submodular cover problem. Here, the utility is assumed to exhibit submodularity, a natural diminishing returns condition prevalent in many data summarization applications. The classical greedy algorithm is known to provide solutions with logarithmic approximation guarantees compared to the optimum solution. However, this sequential, centralized approach is impractical for truly large-scale problems. In this work, we develop the first distributed algorithm–DISCOVER–for submodular set cover that is easily implementable using MapReduce-style computations. We theoretically analyze our approach, and present approximation guarantees for the solutions returned by DISCOVER. We also study a natural trade-off between the communication cost and the number of rounds required to obtain such a solution. In our extensive experiments, we demonstrate the effectiveness of our approach on several applications, including active set selection, exemplar based clustering, and vertex cover on tens of millions of data points using Spark.

66 Parallel Correlation Clustering on Big Graphs

Xinghao Pan xinghao@berkeley.edu
Dimitris Papailiopoulos dimitrisp@berkeley.edu
Benjamin Recht brecht@berkeley.edu
Kannan Ramchandran kannanr@eecs.berkeley.edu
Michael Jordan jordan@berkeley.edu
UC Berkeley

Given a similarity graph between items, correlation clustering (CC) groups similar items together and dissimilar ones apart. One of the most popular CC algorithms is KwikCluster: an algorithm that serially clusters neighborhoods of vertices, and obtains a 3-approximation ratio. Unfortunately, in practice KwikCluster requires a large number of clustering rounds, a potential bottleneck for large graphs.We present C4 and ClusterWild!, two algorithms for parallel correlation clustering that run in a polylogarithmic number of rounds, and provably achieve nearly linear speedups. C4 uses concurrency control to enforce serializability of a parallel clustering process, and guarantees a 3-approximation ratio. ClusterWild! is a coordination free algorithm that abandons consistency for the benefit of better scaling; this leads to a provably small loss in the 3 approximation ratio.We provide extensive experimental results for both algorithms, where we outperform the state of the art, both in terms of clustering accuracy and running time. We show that our algorithms can cluster billion-edge graphs in under 5 seconds on 32 cores, while achieving a 15x speedup.

67 Fast Bidirectional Probability Estimation in Markov Models

Sid Banerjee siddhartha.banerjee@gmail.com
Cornell University
Peter Lofgren plofgren@stanford.edu
Stanford University

We develop a new bidirectional algorithm for estimating Markov chain multi-step transition probabilities: given a Markov chain, we want to estimate the probability of hitting a given target state in ℓ steps after starting from a given source distribution. Given the target state t, we use a (reverse) local power iteration to construct a random variable with the desired mean but having small variance -- this can then be sampled efficiently by a Monte Carlo algorithm. This method extends to any Markov chain on a discrete (finite or countable) state-space, and can be extended to compute functions of multi-step transition probabilities such as PageRank, graph diffusions, hitting/return times, etc. Surprisingly, we also show that in `sparse' Markov Chains -- wherein the number of transitions between states is comparable to the number of states -- the running time of our algorithm for a uniform-random target node is orderwise smaller than Monte Carlo and power iteration based algorithms; in particular, our method can estimate a probability p using only $O(1/p\sqrt{})$ running time.

68 Evaluating the statistical significance of biclusters

Jason D Lee jdl17@stanford.edu
Yuekai Sun yuekai@gmail.com
Jonathan E Taylor jonathan.taylor@stanford.edu
Stanford University

Biclustering (also known as submatrix localization) is a problem of high practical relevance in exploratory analysis of high-dimensional data. We develop a framework for performing statistical inference on biclusters found by score-based

algorithms. Since the bicluster was selected in a data dependent manner by a biclustering or localization algorithm, this is a form of selective inference. Our framework gives exact (non-asymptotic) confidence intervals and p-values for the significance of the selected biclusters. Further, we generalize our approach to obtain exact inference for Gaussian statistics.

69 Regularization Path of Cross-Validation Error Lower Bounds

Atsushi Shibagaki shibagaki.a.mllab.nit@gmail.com
Yoshiki Suzuki suzuki.mllab.nit@gmail.com
Masayuki Karasuyama karasuyama@nitech.ac.jp
Ichiro Takeuchi takeuchi.ichiro@nitech.ac.jp
Nagoya Institute of Technology

Careful tuning of a regularization parameter is indispensable in many machine learning tasks because it has a significant impact on generalization performances.Nevertheless, current practice of regularization parameter tuning is more of an art than a science, e.g., it is hard to tell how many grid-points would be needed in cross-validation (CV) for obtaining a solution with sufficiently small CV error.In this paper we propose a novel framework for computing a lower bound of the CV errors as a function of the regularization parameter, which we call regularization path of CV error lower bounds.The proposed framework can be used for providing a theoretical approximation guarantee on a set of solutions in the sense that how far the CV error of the current best solution could be away from best possible CV error in the entire range of the regularization parameters.We demonstrate through numerical experiments that a theoretically guaranteed a choice of regularization parameter in the above sense is possible with reasonable computational costs.

70 Sampling from Probabilistic Submodular Models

Alkis Gotovos alkisg@inf.ethz.ch
Hamed Hassani hamed@inf.ethz.ch
Andreas Krause krausea@ethz.ch
ETH Zurich

Submodular and supermodular functions have found wide applicability in machine learning, capturing notions such as diversity and regularity, respectively. These notions have deep consequences for optimization, and the problem of (approximately) optimizing submodular functions has received much attention. However, beyond optimization, these notions allow specifying expressive probabilistic models that can be used to quantify predictive uncertainty via marginal inference. Prominent, well-studied special cases include Ising models and determinantal point processes, but the general class of log-submodular and log-supermodular models is much richer and little studied. In this paper, we investigate the use of Markov chain Monte Carlo sampling to perform approximate inference in general log-submodular and log-supermodular models. In particular, we consider a simple Gibbs sampling procedure, and establish two sufficient conditions, the first guaranteeing polynomial-time, and the second fast (O(nlogn)) mixing. We also evaluate the efficiency of the Gibbs sampler on three examples of such models, and compare against a recently proposed variational approach.

71 Submodular Hamming Metrics

Jennifer Gillenwater jengi@uw.edu
Rishabh K Iyer rkiyer@u.washington.edu
Bethany Lusch herwaldt@uw.edu
Rahul Kidambi rkidambi@uw.edu
Jeff A Bilmes bilmes@ee.washington.edu
University of Washington, Seattle

We show that there is a largely unexplored class of functions (positive polymatroids) that can define proper discrete metrics (over pairs of binary vectors) and that are fairly tractable to optimize over. By exploiting the properties of submodularity, we are able to give hardness results and approximation algorithms for optimizing over such metrics. Additionally, we demonstrate empirically the effectiveness of these metrics and associated algorithms by applying them to the task of generating diverse k-best lists.

72 Extending Gossip Algorithms to Distributed Estimation of U-statistics

Igor Colin igor.colin@telecom-paristech.fr
Aurélien Bellet aurelien.bellet@telecom-paristech.fr
Joseph Salmon joseph.salmon@telecom-paristech.fr
Stéphan Clémençon stephan.clemencon@telecom-paristech.fr
Telecom ParisTech

Efficient and robust algorithms for decentralized estimation in networks are essential to many distributed systems. Whereas distributed estimation of sample mean statistics has been the subject of a good deal of attention, computation of U-statistics, relying on more expensive averaging over pairs of observations, is a less investigated area. Yet, such data functionals are essential to describe global properties of a statistical population, with important examples including Area Under the Curve, empirical variance, Gini mean difference and within-cluster point scatter. This paper proposes new synchronous and asynchronous randomized gossip algorithms which simultaneously propagate data across the network and maintain local estimates of the U-statistic of interest. We establish convergence rate bounds of O(1/t) and O(log t / t) for the synchronous and asynchronous cases respectively, where t is the number of iterations, with explicit data and network dependent terms. Beyond favorable comparisons in terms of rate analysis, numerical experiments provide empirical evidence the proposed algorithms surpasses the previously introduced approach.

73 Newton-Stein Method: A Second Order Method for GLMs via Stein's Lemma

Murat A. Erdogdu erdogdu@stanford.edu
Stanford University

We consider the problem of efficiently computing the maximum likelihood estimator in Generalized Linear Models (GLMs)when the number of observations is much larger than the number of coefficients (n > > p > > 1). In this regime, optimization algorithms can immensely benefit fromapproximate second order information. We propose an alternative way of constructing the curvature information by formulatingit as an estimation problem and applying a "Stein-type lemma", which allows further improvements through sub-sampling andeigenvalue thresholding.Our algorithm enjoys fast convergence rates, resembling that of second order methods, with modest per-iteration cost. We provide its convergence analysis for the general case where the rows of the design matrix are samples from a sub-gaussian distribution.We show that the convergence has two phases, aquadratic phase followed by a linear phase. Finally,we empirically demonstrate that our algorithm achieves the highest accuracy for any fixed amount of time compared to a wide variety of fast algorithms on several datasets.

74 Collaboratively Learning Preferences from Ordinal Data

Sewoong Oh swoh@illinois.edu
Kiran K Thekumparampil thekump2@illinois.edu
UIUC
Jiaming Xu xjmoffside@gmail.com

In applications such as recommendation systems and revenue management, it is important to predict preferences on items that have not been seen by a user or predict outcomes of comparisons among those that have never been compared. A popular discrete choice model of multinomial logit model captures the structure of the hidden preferences with a low-rank matrix. In order to predict the preferences, we want to learn the underlying model from noisy observations of the low-rank matrix, collected as revealed preferences in various forms of ordinal data. A natural approach to learn such a model is to solve a convex relaxation of nuclear norm minimization. We present the convex relaxation approach in two contexts of interest: collaborative ranking and bundled choice modeling. In both cases, we show that the convex relaxation is minimax optimal. We prove an upper bound on the resulting error with finite samples, and provide a matching information-theoretic lower bound.

75 SGD Algorithms based on Incomplete U-statistics: Large-Scale Minimization of Empirical Risk

Guillaume Papa guillaume.papa@telecom-paristech.fr
Stéphan Clémençon
stephan.clemencon@telecom-paristech.fr
Aurélien Bellet aurelien.bellet@telecom-paristech.fr
Telecom ParisTech

In many learning problems, ranging from clustering to ranking through metric learning, empirical estimates of the risk functional consist of an average over tuples of observations. In this paper, we argue that in the large-scale setting, the inductive principle of stochastic approximation for risk minimization should be implemented in a very specific manner. Precisely, the gradient estimates should be obtained by sampling tuples of data points with replacement (incomplete U-statistics) instead of sampling data points without replacement (complete U-statistics based on subsamples). We develop a theoretical framework accounting for the considerable impact of this strategy on the generalization ability of the prediction model returned by the Stochastic Gradient Descent (SGD) algorithm. It reveals that the method we promote achieves a much better trade-off between statistical accuracy and computational cost. Beyond the rate bound analysis, numerical experiments on AUC maximization and metric learning provide strong empirical evidence of the superiority of the proposed approach.

76 Alternating Minimization for Regression Problems with Vector-valued Outputs

Prateek Jain prajain@microsoft.com
Microsoft Research
Ambuj Tewari tewaria@umich.edu
University of Michigan

In regression problems involving vector-valued outputs (or equivalently, multiple responses), it is well known that the maximum likelihood estimator (MLE), which takes noise covariance structure into account, can be significantly more accurate than the ordinary least squares (OLS) estimator. However, existing literature compares OLS and MLE in terms of their asymptotic, not finite sample, guarantees. More crucially, computing the MLE in general requires solving a non-convex optimization problem and is not known to be efficiently solvable. We provide finite sample upper and lower bounds on the estimation error of OLS and MLE, in two popular models: a) Pooled model, b) Seemingly Unrelated Regression (SUR) model. We provide precise instances where the MLE is significantly more accurate than OLS. Furthermore, for both models, we show that the output of a computationally efficient alternating minimization procedure enjoys the same performance guarantee as MLE, up to universal constants. Finally, we show that for high-dimensional settings as well, the alternating minimization procedure leads to significantly more accurate solutions than the corresponding OLS solutions but with error bound that depends only logarithmically on the data dimensionality.

77 On Variance Reduction in Stochastic Gradient Descent and its Asynchronous Variants

Sashank J. Reddi sjakkamr@cs.cmu.edu
Ahmed Hefny ahefny@cs.cmu.edu
Barnabas Poczos bapoczos@cs.cmu.edu
Alex J Smola alex@smola.org
Carnegie Mellon University
Suvrit Sra suvrit@mit.edu
MIT

We study optimization algorithms based on variance reduction for stochastic gradientdescent (SGD). Remarkable recent progress has been made in this directionthrough development of algorithms like SAG, SVRG, SAGA. These algorithmshave been shown to outperform SGD, both theoretically and empirically. However,asynchronous versions of these algorithms—a crucial requirement for modernlarge-scale applications—have not been studied. We bridge this gap by presentinga unifying framework that captures many variance reduction techniques. Subsequently, we propose an asynchronous algorithm grounded in our framework,with fast convergence rates. An important consequence of our general approachis that it yields asynchronous versions of variance reduction algorithms such asSVRG, SAGA as a byproduct. Our method achieves near linear speedup in sparsesettings common to machine learning. We demonstrate the empirical performanceof our method through a concrete realization of asynchronous SVRG.

78 Subset Selection by Pareto Optimization

Chao Qian qianc@lamda.nju.edu.cn
Yang Yu yuy@lamda.nju.edu.cn
Zhi-Hua Zhou zhouzh@nju.edu.cn
Nanjing University

Selecting the optimal subset from a large set of variables is a fundamental problem in various learning tasks such as feature selection, sparse regression, dictionary learning, etc. In this paper, we propose the POSS approach which employs evolutionary Pareto optimization to find a small-sized subset with good performance. We prove that for sparse regression, POSS is able to achieve the best-so-far theoretically guaranteed approximation performance efficiently. Particularly, for the \emph{Exponential Decay} subclass, POSS is proven to achieve an optimal solution.

Empirical study verifies the theoretical results, and exhibits the superior performance of POSS to greedy and convex relaxation methods.

79 Interpolating Convex and Non-Convex Tensor Decompositions via the Subspace Norm

Qinqing Zheng qinqing@cs.uchicago.edu
University of Chicago
Ryota Tomioka tomioka@ttic.edu
Toyota Technological Institute at Chicago

We consider the problem of recovering a low-rank tensor from its noisy observation. Previous work has shown a recovery guarantee with signal to noise ratio O(n\ceilK/2/2) for recovering a Kth order rank one tensor of size n×···×n by recursive unfolding. In this paper, we first improve this bound to O(nK/4) by a much simpler approach, but with a more careful analysis. Then we propose a new norm called the \textit{subspace} norm, which is based on the Kronecker products of factors obtained by the proposed simple estimator. The imposed Kronecker structure allows us to show a nearly ideal O(n√+HK−1−−−−−√) bound, in which the parameter H controls the blend from the non-convex estimator to mode-wise nuclear norm minimization. Furthermore, we empirically demonstrate that the subspace norm achieves the nearly ideal denoising performance even with H=O(1).

80 Minimum Weight Perfect Matching via Blossom Belief Propagation

Sung-Soo Ahn sungsoo.ahn@kaist.ac.kr
Sejun Park sejun.park@kaist.ac.kr
Jinwoo Shin jinwoos@kaist.ac.kr
KAIST
Misha Chertkov chertkov@lanl.gov

Max-product Belief Propagation (BP) is a popular message-passing algorithm for computing a Maximum-A-Posteriori (MAP) assignment over a distribution represented by a Graphical Model (GM). It has been shown that BP can solve a number of combinatorial optimization problems including minimum weight matching, shortest path, network flow and vertex cover under the following common assumption: the respective Linear Programming (LP) relaxation is tight, i.e., no integrality gap is present. However, when LP shows an integrality gap, no model has been known which can be solved systematically via sequential applications of BP. In this paper, we develop the first such algorithm, coined Blossom-BP, for solving the minimum weight matching problem over arbitrary graphs. Each step of the sequential algorithm requires applying BP over a modified graph constructed by contractions and expansions of blossoms, i.e., odd sets of vertices. Our scheme guarantees termination in O(n^2) of BP runs, where n is the number of vertices in the original graph. In essence, the Blossom-BP offers a distributed version of the celebrated Edmonds' Blossom algorithm by jumping at once over many sub-steps with a single BP. Moreover, our result provides an interpretation of the Edmonds' algorithm as a sequence of LPs.

81 b-bit Marginal Regression

Martin Slawski martin.slawski@rutgers.edu
Ping Li pingli@stat.rutgers.edu
Rugters University

We consider the problem of sparse signal recovery from m linear measurements quantized to b bits. b-bit Marginal Regression is proposed as recovery algorithm. We study the question of choosing b in the setting of a given budget of bits B=m·b and derive a single easy-to-compute expression characterizing the trade-off between m and b. The choice b=1 turns out to be optimal for estimating the unit vector corresponding to the signal for any level of additive Gaussian noise before quantization as well as for adversarial noise. For b≥2, we show that Lloyd-Max quantization constitutes an optimal quantization scheme and that the norm of the signal canbe estimated consistenly by maximum likelihood.

82 LASSO with Non-linear Measurements is Equivalent to One With Linear Measurements

CHRISTOS THRAMPOULIDIS cthrampo@caltech.edu
Ehsan Abbasi eabbasi@caltech.edu
Babak Hassibi hassibi@caltech.edu
Caltech

Consider estimating an unknown, but structured (e.g. sparse, low-rank, etc.), signal $x0\in\mathrm{Rn}$ from a vector $y\in\mathrm{Rm}$ of measurements of the form yi=gi(aTix0), where the ai's are the rows of a known measurement matrix A, and, g is a (potentially unknown) nonlinear and random link-function. Such measurement functions could arise in applications where the measurement device has nonlinearities and uncertainties. It could also arise by design, e.g., gi(x)=sign(x+zi), corresponds to noisy 1-bit quantized measurements. Motivated by the classical work of Brillinger, and more recent work of Plan and Vershynin, we estimate x0 via solving the Generalized-LASSO, i.e., $x\hat{}=\arg\min x\|y-Ax0\|2+\lambda f(x)$ for some regularization parameter λ>0 and some (typically non-smooth) convex regularizer f that promotes the structure of x0, e.g. ℓ1-norm, nuclear-norm. While this approach seems to naively ignore the nonlinear function g, both Brillinger and Plan and Vershynin have shown that, when the entries of A are iid standard normal, this is a good estimator of x0 up to a constant of proportionality μ, which only depends on g. In this work, we considerably strengthen these results by obtaining explicit expressions for $\|x\hat{}-\mu x0\|2$, for the regularized Generalized-LASSO, that are asymptotically precise when m and n grow large. A main result is that the estimation performance of the Generalized LASSO with non-linear measurements is asymptotically the same as one whose measurements are linear yi=μaTix0+σzi, with μ=E[γg(γ)] and σ2=E[(g(γ)−μγ)2], and, γ standard normal. The derived expressions on the estimation performance are the first-known precise results in this context. One interesting consequence of our result is that the optimal quantizer of the measurements that minimizes the estimation error of the LASSO is the celebrated Lloyd-Max quantizer.

83 Randomized Block Krylov Methods for Stronger and Faster Approximate Singular Value Decomposition

Cameron Musco cnmusco@mit.edu
Christopher Musco cpmusco@mit.edu
Massachusetts Institute of Technology

Since being analyzed by Rokhlin, Szlam, and Tygert \ cite{RokhlinTygertPCA} and popularized by Halko, Martinsson, and Tropp \cite{Halko:2011}, randomized Simultaneous Power Iteration has become the method of choice for approximate singular value decomposition. It is more accurate than simpler sketching algorithms, yet still converges quickly for \emph{any} matrix, independently of singular value gaps. After O~(1/ϵ) iterations, it gives a low-rank approximation within (1+ϵ) of optimal for spectral norm error.We give the first provable runtime improvement on Simultaneous Iteration. A simple randomized block Krylov method, closely related to the classic Block Lanczos algorithm, gives the same guarantees in just O~(1/ϵ√) iterations and performs substantially better experimentally. Despite their long history, our analysis is the first of a Krylov subspace method that does not depend on singular value gaps, which are unreliable in practice.Furthermore, while it is a simple accuracy benchmark, even (1+ϵ) error for spectral norm low-rank approximation does not imply that an algorithm returns high quality principal components, a major issue for data applications. We address this problem for the first time by showing that both block Krylov methods and Simultaneous Iteration give nearly optimal PCA for any matrix. This result further justifies their strength over non-iterative sketching methods.

84 On the Pseudo-Dimension of Nearly Optimal Auctions

Jamie H Morgenstern jamiemmt.cs@gmail.com
University of Pennsylvania
Tim Roughgarden tim@cs.stanford.edu
Stanford University

This paper develops a general approach, rooted in statistical learning theory, to learning an approximately revenue-maximizing auction from data. We introduce t-level auctions to interpolate between simple auctions, such as welfare maximization with reserve prices, and optimal auctions, thereby balancing the competing demands of expressivity and simplicity. We prove that such auctions have small representation error, in the sense that for every product distribution F over bidders' valuations, there exists a t-level auction with small t and expected revenue close to optimal. We show that the set of t-level auctions has modest pseudo-dimension (for polynomial t) and therefore leads to small learning error. One consequence of our results is that, in arbitrary single-parameter settings, one can learn a mechanism with expected revenue arbitrarily close to optimal from a polynomial number of samples.

85 Closed-form Estimators for High-dimensional Generalized Linear Models

Eunho Yang yangeh@gmail.com
Aurelie C Lozano aclozano@us.ibm.com
IBM Research
Pradeep K Ravikumar pradeepr@cs.utexas.edu
University of Texas at Austin

We propose a class of closed-form estimators for GLMs under high-dimensional sampling regimes. Our class of estimators is based on deriving closed-form variants of the vanilla unregularized MLE but which are (a) well-defined even under high-dimensional settings, and (b) available in closed-form. We then perform thresholding operations on this MLE variant to obtain our class of estimators. We derive a unified statistical analysis of our class of estimators, and show that it enjoys strong statistical guarantees in both parameter error as well as variable selection, that surprisingly match those of the more complex regularized GLM MLEs, even while our closed-form estimators are computationally much simpler. We derive instantiations of our class of closed-form estimators, as well as corollaries of our general theorem, for the special cases of logistic, exponential and Poisson regression models. We corroborate the surprising statistical and computational performance of our class of estimators via extensive simulations.

86 Fast, Provable Algorithms for Isotonic Regression in all l_p-norms

Rasmus Kyng rjkyng@gmail.com
Yale University
Anup Rao raoanupb@gmail.com
School of Computer Science, Georgia Tech
Sushant Sachdeva sachdevasushant@gmail.com
Yale University

Given a directed acyclic graph G, and a set of values y on the vertices, the Isotonic Regression of y is a vector x that respects the partial order described by G, and minimizes $\|x-y\|$, for a specified norm. This paper gives improved algorithms for computing the Isotonic Regression for all weighted ℓp-norms with rigorous performance guarantees. Our algorithms are quite practical, and their variants can be implemented to run fast in practice.

87 Semi-proximal Mirror-Prox for Nonsmooth Composite Minimization

Niao He nhe6@gatech.edu
Georgia Institute of Technology
Zaid Harchaoui zaid.harchaoui@inria.fr
Inria

We propose a new first-order optimisation algorithm to solve high-dimensional non-smooth composite minimisation problems. Typical examples of such problems have an objective that decomposes into a non-smooth empirical risk part and a non-smooth regularisation penalty.The proposed algorithm, called \ spmp, leverages the Fenchel-type representation of one part of the objective while handling the other part of the objective via linear minimization over the domain. The algorithm stands in contrast with more classical proximal gradient algorithms with smoothing, which require the computation of proximal operators at each iteration and can therefore be impractical for high-

dimensional problems. We establish the theoretical convergence rate of \spmp, which exhibits the optimal complexity bounds, i.e. O(1/ϵ2), for the number of calls to linear minimization oracle. We present promising experimental results showing the interest of the approachin comparison to competing methods.

88 Competitive Distribution Estimation: Why is Good-Turing Good

Alon Orlitsky alon@ucsd.edu
Ananda Suresh s.theertha@gmail.com
University of California, San Diego

Estimating distributions over large alphabets is a fundamental tenet of machine learning. Yet no estimator is known to estimate all distributions well. For example, add-constant estimators are nearly min-max optimal, but perform poorly in practice, while practical estimators such as Jelinek-Mercer, absolute discounting, and Good-Turing, are not known to be near optimal for essentially any distribution.We provide the first uniform optimality proof for any distribution estimator. We show that a variant of Good-Turing estimators is nearly best for all distributions in two competitive ways. First it estimates every distribution nearly as well as the best estimator designed with prior knowledge of the distribution up to a permutation. Second, it estimates every distribution nearly as well as the best estimator designed with prior knowledge of the exact distribution but restricted, as all natural estimators, to assign the same probability to all symbols appearing the same number of times. Specifically, we show that for both comparisons, the KL divergence of the Good-Turing variant is always within O~(min(k/n,1/n√)) of the best estimator. Conversely, any estimator must have a KL divergence ≥Ω~(min(k/n,1/n2/3)) over the best estimator for the first comparison, and ≥Ω~(min(k/n,1/n√)) for the second.

89 A Universal Primal-Dual Convex Optimization Framework

Alp Yurtsever alp.yurtsever@epfl.ch
Volkan Cevher volkan.cevher@epfl.ch
EPFL
Quoc Tran Dinh quoctd@email.unc.edu
UNC, North Carolina

We propose a new primal-dual algorithmic framework for a prototypical constrained convex optimization template. The algorithmic instances of our framework are universal since they can automatically adapt to the unknown Holder continuity properties within the template. They are also guaranteed to have optimal convergence rates in the objective residual and the feasibility gap for each smoothness level. In contrast to existing primal-dual algorithms, our framework avoids the proximity operator of the objective function altogether. We instead leverage computationally cheaper, Fenchel-type operators, which are the main workhorses of the generalized conditional gradient (GCG)-type methods. In contrast to the GCG-type methods, our framework does not require the objective function to be differentiable, and can also process additional general linear inclusion constraints. Our analysis technique unifies Nesterov's universal gradient methods and GCG-type methods to address the more broadly applicable primal-dual setting.

90 Sample Complexity of Episodic Fixed-Horizon Reinforcement Learning

Christoph Dann cdann@cdann.net
Emma Brunskill ebrun@cs.cmu.edu
Carnegie Mellon University

Recent reinforcement learning research has made significant progress in understanding learning in discounted infinite-horizon Markov decision processes (MDPs) by deriving tight sample complexity bounds. However, in many real-world applications, an interactive learning agent operates for a fixed or bounded period of time, for example tutoring students for exams or handling customer service requests. Such scenarios can often be better treated as episodic fixed-horizon MDPs, for which only loose bounds on the sample complexity exist. A natural notion of sample complexity in this setting is the number of episodes required to guarantee a certain performance with high probability (PAC guarantee). In this paper, we derive an upper PAC bound O~(\numS2\numAH2ϵ2ln1δ) and a lower PAC bound Ω~(\numS\numAH2ϵ2ln1δ+c) that match up to log-terms and an additional linear dependency on the number of states \numS. The lower bound is the first of its kind for this setting and our upper bound leverages Bernstein's inequality to improve on previous bounds for finite-horizon MDPs which have a time-horizon dependency of at least H4.

91 Private Graphon Estimation for Sparse Graphs

Christian Borgs christian.borgs@microsoft.com
Jennifer Chayes jchayes@microsoft.com
Microsoft Research
Adam Smith asmith@cse.psu.edu
Pennsylvania State University

We design algorithms for fitting a high-dimensional statistical model to a large, sparse network without revealing sensitive information of individual members. Given a sparse input graph G, our algorithms output a node-differentially-private nonparametric block model approximation. By node-differentially-private, we mean that our output hides the insertion or removal of a vertex and all its adjacent edges. If G is an instance of the network obtained from a generative nonparametric model defined in terms of a graphon W, our model guarantees consistency, in the sense that as the number of vertices tends to infinity, the output of our algorithm converges to W in an appropriate version of the L2 norm. In particular, this means we can estimate the sizes of all multi-way cuts in G. Our results hold as long as W is bounded, the average degree of G grows at least like the log of the number of vertices, and the number of blocks goes to infinity at an appropriate rate. We give explicit error bounds in terms of the parameters of the model; in several settings, our bounds improve on or match known nonprivate results.

92 HONOR: Hybrid Optimization for NOn-convex Regularized problems

Pinghua Gong gongp@umich.edu
Jieping Ye jpye@umich.edu
University of Michigan

Recent years have witnessed the superiority of non-convex sparse learning formulations over their convex counterparts in both theory and practice. However, due to the non-convexity and non-smoothness of the regularizer, how to efficiently solve the non-convex optimization problem for large-scale data is still quite challenging. In this paper, we propose an efficient \underline{H}ybrid \underline{O}ptimization algorithm for \underline{NO}n-convex \underline{R}egularized problems (HONOR). Specifically, we develop a hybrid scheme which effectively integrates a Quasi-Newton (QN) step and a Gradient Descent (GD) step. Our contributions are as follows: (1) HONOR incorporates the second-order information to greatly speed up the convergence, while it avoids solving a regularized quadratic programming and only involves matrix-vector multiplications without explicitly forming the inverse Hessian matrix. (2) We establish a rigorous convergence analysis for HONOR, which shows that convergence is guaranteed even for non-convex problems, while it is typically challenging to analyze the convergence for non-convex problems. (3) We conduct empirical studies on large-scale data sets and results demonstrate that HONOR converges significantly faster than state-of-the-art algorithms.

93 A Convergent Gradient Descent Algorithm for Rank Minimization and Semidefinite Programming from Random Linear Measurements

Qinqing Zheng qinqing@cs.uchicago.edu
John Lafferty lafferty@galton.uchicago.edu
University of Chicago

We propose a simple, scalable, and fast gradient descent algorithm to optimize a nonconvex objective for the rank minimization problem and a closely related family of semidefinite programs. With O(r3κ2nlogn) random measurements of a positive semidefinite n×n matrix of rank r and condition number κ, our method is guaranteed to converge linearly to the global optimum.

94 Super-Resolution Off the Grid

Qingqing Huang qqh2011@gmail.com
MIT
Sham Kakade sham@cs.washington.edu
University of Washington

Super-resolution is the problem of recovering a superposition of point sources using bandlimited measurements, which may be corrupted with noise. This signal processing problem arises in numerous imaging problems, ranging from astronomy to biology to spectroscopy, where it is common to take (coarse) Fourier measurements of an object. Of particular interest is in obtaining estimation procedures which are robust to noise, with the following desirable statistical and computational properties: we seek to use coarse Fourier measurements (bounded by some \emph{cutoff frequency}); we hope to take a (quantifiably) small number of measurements; we desire our algorithm to run quickly. Suppose we have k point sources in d dimensions, where the points are separated by at least Δ from each other (in Euclidean distance). This work provides an algorithm with the following favorable guarantees:1. The algorithm uses Fourier measurements, whose frequencies are bounded by O(1/Δ) (up to log factors). Previous algorithms require a \emph{cutoff frequency} which may be as large as Ω(d√/Δ).2. The number of measurements taken by and the computational complexity of our algorithm are bounded by a polynomial in both the number of points k and the dimension d, with \emph{no} dependence on the separation Δ. In contrast, previous algorithms depended inverse polynomially on the minimal separation and exponentially on the dimension for both of these quantities.Our estimation procedure itself is simple: we take random bandlimited measurements (as opposed to taking an exponential number of measurements on the hyper-grid). Furthermore, our analysis and algorithm are elementary (based on concentration bounds of sampling and singular value decomposition).

95 Optimal Rates for Random Fourier Features

Bharath Sriperumbudur bharathsv.ucsd@gmail.com
The Pennsylvania State University
Zoltan Szabo zoltan.szabo@gatsby.ucl.ac.uk
Gatsby Unit, UCL

Kernel methods represent one of the most powerful tools in machine learning to tackle problems expressed in terms of function values and derivatives due to their capability to represent and model complex relations. While these methods show good versatility, they are computationally intensive and have poor scalability to large data as they require operations on Gram matrices. In order to mitigate this serious computational limitation, recently randomized constructions have been proposed in the literature, which allow the application of fast linear algorithms. Random Fourier features (RFF) are among the most popular and widely applied constructions: they provide an easily computable, low-dimensional feature representation for shift-invariant kernels. Despite the popularity of RFFs, very little is understood theoretically about their approximation quality. In this paper, we provide a detailed finite-sample theoretical analysis about the approximation quality of RFFs by (i) establishing optimal (in terms of the RFF dimension, and growing set size) performance guarantees in uniform norm, and (ii) presenting guarantees in L^r (1 ≤ r < ∞) norms. We also propose an RFF approximation to derivatives of a kernel with a theoretical study on its approximation quality.

96 Combinatorial Bandits Revisited

Richard Combes richard.combes@supelec.fr
Supelec
Mohammad Sadegh Talebi Mazraeh Shahi mstms@kth.se
KTH Royal Inst. of Technology
Alexandre Proutiere alepro@kth.se

marc lelarge marc.lelarge@ens.fr
INRIA - ENS

This paper investigates stochastic and adversarial combinatorial multi-armed banditproblems. In the stochastic setting under semi-bandit feedback, we derivea problem-specific regret lower bound, and analyze its scaling with the dimensionof the decision space. We propose ESCB, an algorithm that efficiently exploitsthe structure of the problem and provide a finite-time analysis of its regret.ESCB has better performance guarantees than existing algorithms, and significantlyoutperforms these

algorithms in practice. In the adversarial setting underbandit feedback, we propose COMBEXP, and algorithm with the same regret scalingas state-of-the-art algorithms, but with lower computational complexity forsome combinatorial problems.

97 Fast Convergence of Regularized Learning in Games

Vasilis Syrgkanis vasy@microsoft.com
Alekh Agarwal alekha@microsoft.com
Robert Schapire schapire@microsoft.com
MIcrosoft Research
Haipeng Luo haipengl@cs.princeton.edu
Princeton University

We show that natural classes of regularized learning algorithms with a form of recency bias achieve faster convergence rates to approximate efficiency and to coarse correlated equilibria in multiplayer normal form games. When each player in a game uses an algorithm from our class, their individual regret decays at O(T−3/4), while the sum of utilities converges to an approximate optimum at O(T−1)--an improvement upon the worst case O(T−1/2) rates. We show a black-box reduction for any algorithm in the class to achieve O(T−1/2) rates against an adversary, while maintaining the faster rates against algorithms in the class. Our results extend those of Rakhlin and Shridharan~\cite{Rakhlin2013} and Daskalakis et al.~\cite{Daskalakis2014}, who only analyzed two-player zero-sum games for specific algorithms.

98 On Elicitation Complexity

Rafael Frongillo raf@cs.berkeley.edu
CU Boulder
Ian Kash iankash@microsoft.com
Microsoft

Elicitation is the study of statistics or properties which are computable via empirical risk minimization. While several recent papers have approached the general question of which properties are elicitable, we suggest that this is the wrong question---all properties are elicitable by first eliciting the entire distribution or data set, and thus the important question is how elicitable. Specifically, what is the minimum number of regression parameters needed to compute the property?Building on previous work, we introduce a new notion of elicitation complexity and lay the foundations for a calculus of elicitation. We establish several general results and techniques for proving upper and lower bounds on elicitation complexity. These results provide tight bounds for eliciting the Bayes risk of any loss, a large class of properties which includes spectral risk measures and several new properties of interest.

99 Online Learning with Adversarial Delays

Kent Quanrud quanrud2@illinois.edu
Daniel Khashabi khashab2@illinois.edu
UIUC

We study the performance of standard online learning algorithms when the feedback is delayed by an adversary. We show that \texttt{online-gradient-descent} and \texttt{follow-the-perturbed-leader} achieve regret O(D−−√) in the delayed setting, where D is the sum of delays of each round's feedback. This bound collapses to an optimal O(T√) bound in the usual setting of no delays (where D=T). Our main contribution is to show that standard algorithms for online learning already have simple regret bounds in the most general setting of delayed feedback, making adjustments to the analysis and not to the algorithms themselves. Our results help affirm and clarify the success of recent algorithms in optimization and machine learning that operate in a delayed feedback model.

100 Structured Estimation with Atomic Norms: General Bounds and Applications

Sheng Chen shengc@cs.umn.edu
Arindam Banerjee banerjee@cs.umn.edu
University of Minnesota

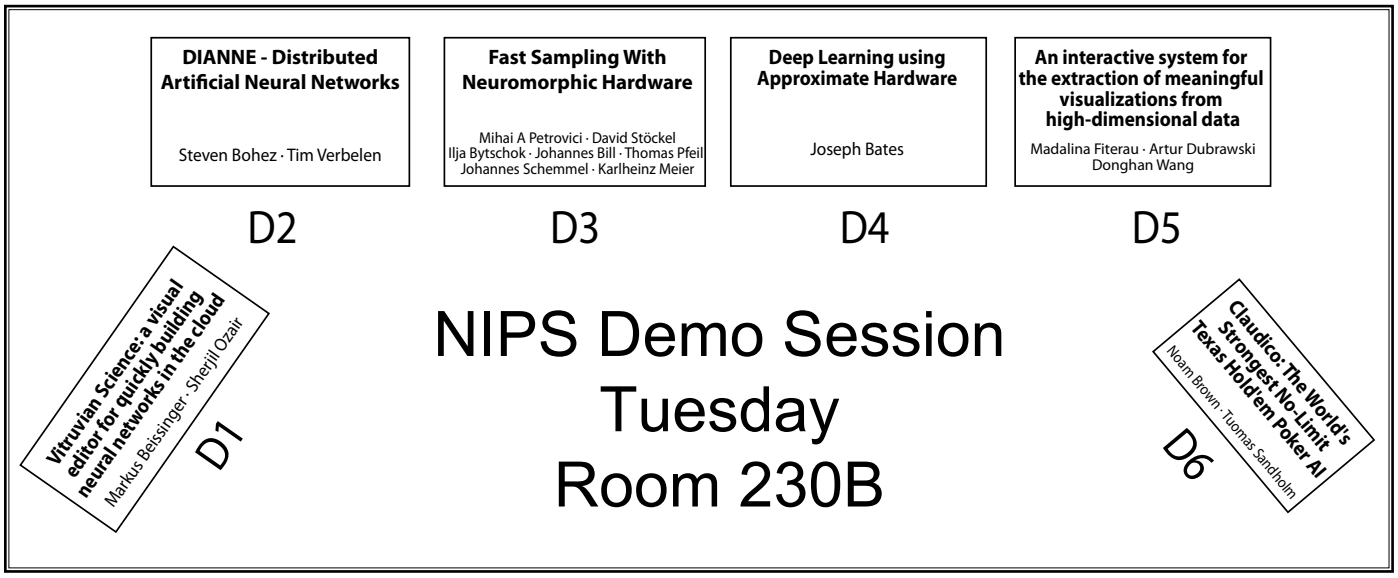
For structured estimation problems with atomic norms, existing literature expresses sample complexity and estimation error bounds in terms of certain geometric measures, in particular Gaussian width of the unit norm ball, Gaussian width of a spherical cap determined by an error cone, and a restricted norm compatibility constant. However, given an atomic norm, these geometric measures are usually difficult to characterize or bound. In this paper, we present general bounds for these geometric measures, which only require simple information regarding the atomic norm under consideration, and we establish tightness of the bounds. We show applications of our analysis to certain atomic norms, including the recently proposed k-support norm, for which existing analysis is incomplete.

101 Subsampled Power Iteration: a Unified Algorithm for Block Models and Planted CSP's

Vitaly Feldman
Will Perkins
Santosh Vempala

We present an algorithm for recovering planted solutions in two well-known models, the stochastic block model and planted constraint satisfaction problems (CSP), via a common generalization in terms of random bipartite graphs. Our algorithm matches up to a constant factor the best-known bounds for the number of edges (or constraints) needed for perfect recovery and its running time is linear in the number of edges used. The time complexity is significantly better than both spectral and SDP-based approaches.The main contribution of the algorithm is in the case of unequal sizes in the bipartition that arises in our reduction from the planted CSP. Here our algorithm succeeds at a significantly lower density than the spectral approaches, surpassing a barrier based on the spectral norm of a random matrix.Other significant features of the algorithm and analysis include (i) the critical use of power iteration with subsampling, which might be of independent interest; its analysis requires keeping track of multiple norms of an evolving solution (ii) the algorithm can be implemented statistically, i.e., with very limited access to the input distribution (iii) the algorithm is extremely simple to implement and runs in linear time, and thus is practical even for very large instances.

DEMONSTRATIONS ABSTRACTS



D1 Vitruvian Science: a visual editor for quickly building neural networks in the cloud

Markus Beissinger mbeissinger@gmail.com
Vitruvian Science
Sherjil Ozair sherjilozair@gmail.com
Indian Institute of Technology Delhi

We present a visual editor to make prototyping deep learning systems akin to building with Legos. This system allows you to visually depict architecture layouts (as you would on a whiteboard or scrap of paper) and compiles them to Theano functions running on the cloud. Our goal is to speed up development and debugging for novel deep learning architectures.

D2 DIANNE - Distributed Artificial Neural Networks

Steven Bohez steven.bohez@intec.ugent.be
Tim Verbelen tim.verbelen@intec.ugent.be
Ghent University - iMinds

In this demo users will be able to build neural networks using a drag & drop interface. Different neural network building blocks such as Convolutional, Linear, ReLu, Sigmoid, ... can be configured and connected to form a neural network. These networks can then be connected one of the supported Datasets (i.e. MNIST, CIFAR, ImageNet, ...) to evaluate the network accuracy, visualise the output for specific samples, or train a (small) network (for example in case of MNIST for immediate results). The users can also couple sensors and actuators to a neural network input/output, for example triggering a lamp based on classified camera input. The UI also allows to deploy the neural network building blocks on various different single-board platforms such as Raspberry Pi, Intel Edison and Nvidia Jetson and the difference in response time can be experienced.

D3 Fast sampling with neuromorphic hardware

Mihai A Petrovici mpedro@kip.uni-heidelberg.de
David Stöckel david.stoeckel@kip.uni-heidelberg.de
Ilja Bytschok ibyt@kip.uni-heidelberg.de
Johannes Bill bill.scientific@gmail.com
Thomas Pfeil thomas.pfeil@kip.uni-heidelberg.de
Johannes Schemmel schemmel@kip.uni-heidelberg.de
Karlheinz Meier meierk@kip.uni-heidelberg.de
Heidelberg University

Many of the problems that a typical brain is required to solve are of essentially Bayesian nature. How - or if at all - a Bayesian algorithm is embedded in the structure and dynamics of cortical neural networks remains the subject of considerable debate. Experimental evidence of a “Bayesian brain” has been steadily growing, as has the collection of models that have been put forward as possible candidates for a neuronal implementation of Bayesian inference. The study of generative and discriminative models for various types of data is further bolstered by advances in fields that lie outside of neuroscience, but have obvious conceptual connections, such as machine learning and AI. Recent theoretical studies point towards a link between these models and biological neural networks. Whether biological or not, all of these models are only as efficient as the physical substrate that they are embedded in allows them to be. Power efficiency as well as execution speed are of increasing importance as both the network models and their associated learning algorithms become large and/or complex. In our demo session, we will show the first implementation of neural sampling in mixed-signal, low-power and highly accelerated neuromorphic hardware. We will discuss the network structure and mechanisms by which we were able to counteract the imperfections that are inherent to the hardware manufacturing process and provide an interactive interface for users to manipulate the emulation parameters, in particular those of the target probability distribution from which the neuromorphic chip samples. The corresponding parameters of the on-chip spiking neural network can then either be calculated analytically or trained with simple learning rules. More advanced users which are familiar with spiking network simulators will be able to make use of the full versatility of the hardware substrate and program their own network models for neuromorphic emulation.

D4 Deep Learning using Approximate Hardware

Joseph Bates jalias1@singularcomputing.com
Singular Computing LLC

We demo deep learning algorithms doing real-time vision on an “approximate computer”. The machine is a prototype for embedded applications that provides 10x the compute per watt of a modern GPU. The technology allows a single (non-prototype) chip to contain 256,000 cores and to compute using 50x less energy that an equivalent GPU. A rack could hold 50 million cores, and accelerate deep learning training and other applications. The technology has been funded by DARPA (U.S.) and tested at MIT CSAIL, Carnegie Mellon, and elsewhere.

D5 An interactive system for the extraction of meaningful visualizations from high-dimensional data

Madalina Fiterau mfiterau@cs.cmu.edu
Stanford University
Artur Dubrawski awd@cs.cmu.edu
Donghan Wang donghanw@cs.cmu.edu
Carnegie Mellon University

We demonstrate our novel techniques for building ensembles of low-dimensional projections that facilitate data understanding and visualization by human users, given a learning task such as classification or regression. Our system trains user-friendly models, called Informative Projection Ensembles (IPEs). Such ensembles comprise of a set of compact submodels that ensure compliance with stringent user-specified requirement on model size and complexity, in order to allow visualization of the extracted patterns from data. IPEs handle data in a query-specific manner, each sample being assigned to a specialized Informative Projection, with data being automatically partitioned during learning. Through this setup, the models attain high performance while maintaining the transparency and simplicity of low-dimensional classifiers and regressors. In this demo, we illustrate how Informative Projection Ensembles were of great use in practical applications. Moreover, we allow users the possibility to train their own models in real time, specifying such settings as the number of submodels, the dimensionality of the subspaces, costs associated with features as well as the type of base classifier or regressor to be used. Users are also able to see the decision-support system in action, performing classification, regression or clustering on batches of test data. The process of handling test data is also transparent, with the system highlighting the selected submodel, and how the queries are assigned labels/values by the submodel itself. Users can give feedback to the system in terms of the assigned outputs, and they will be able to perform pairwise comparisons of the trained models. We encourage participants to bring their own data to analyze. Users have the possibility of saving the outcome of the analysis, for their own datasets or non-proprietary ones. The system supports the csv format for data and xml for the models.

D6 Claudico: The World’s Strongest No-Limit Texas Hold’em Poker AI

Noam Brown noamb@cmu.edu
Tuomas Sandholm sandholm@cs.cmu.edu
Carnegie Mellon University

Claudico is the world’s strongest poker AI for two-player no-limit Texas Hold’em. Earlier this year, Claudico faced off against four of the top human poker players in the world in the Brains vs. AI man-machine competition for 80,000 hands of poker. Claudico is based on an earlier agent, Tartanian7, which won the latest Annual Computer Poker Competition, defeating all other agents with statistical significance. This is the first poker AI to play at a competitive level with the very best humans in the world in no-limit Texas Hold’em, the most popular poker variant in the world. It is the result of over one million core hours, computed on the Blacklight supercomputer.

WEDNESDAY SESSIONS

OUR SPONSORS

Google

Microsoft

Alibaba Group 阿里巴巴集团

Ketchum Trading, LLC

amazon.com

Apple

Baidu Research

CITADEL

facebook

IBM Research

NVIDIA

The Voleon Group

Artificial Intelligence
www.elsevier.com/locate/artint

Bloomberg

twitter

AdRoll

Analog Devices | Lyric Labs

CenturyLink Business

criteo

Cubist Systematic Strategies

deep genomics

D E Shaw & Co

NETFLIX

Disney Research

ebay

imagia

MaluubA

Man AHL

ORACLE

Panasonic

PDT PARTNERS

SONY

The Alan Turing Institute

TOYOTA

United Technologies Research Center

Vatic Labs

TWO SIGMA

YAHOO! LABS

WINTON

Adobe

ORAL SESSION

SESSION 5: - 9:00 – 10:10 AM

INVITED TALK: BREIMAN LECTURE
Post-selection Inference for Forward Stepwise Regression, Lasso and other Adaptive Statistical procedures

Robert Tibshirani tibs@stanford.edu
Standford University

In this talk I will present new inference tools for adaptive statistical procedures. These tools provide p-values and confidence intervals that have correct “post-selection” properties: they account for the selection that has already been carried out on the same data. I discuss application of these ideas to a wide variety of problems including Forward Stepwise Regression, Lasso, PCA, and graphical models. I will also discuss computational issues and software for implementation of these ideas.

This talk represents work (some joint) with many people including Jonathan Taylor, Richard Lockhart, Ryan Tibshirani, Will Fithian, Jason Lee, Dennis Sun, Yuekai Sun and Yunjin Choi.

Learning Theory and Algorithms for Forecasting Non-stationary Time Series

Vitaly Kuznetsov vitaly@cims.nyu.edu
Courant Institute
Mehryar Mohri mohri@cs.nyu.edu
Courant Institute and Google

We present data-dependent learning bounds for the general scenario of non-stationary non-mixing stochastic processes. Our learning guarantees are expressed in terms of the notion of sequential complexity and a discrepancy measure that can be estimated from data under some mild assumptions. We use our learning bounds to devise new algorithms for non-stationary time series forecasting for which we report some preliminary experimental results.

SPOTLIGHT SESSION

SESSION 5: 10:10 – 10:40 AM

- **Empirical Localization of Homogeneous Divergences on Discrete Sample Spaces**
  Takashi Takenouchi, Future University Hakodate
  Takafumi Kanamori, Nagoya University

- **Multi-Layer Feature Reduction for Tree Structured Group Lasso via Hierarchical Projection**
  Jie Wang, University of Michigan-Ann Arbor
  Jieping Ye, University of Michigan

- **Optimal Testing for Properties of Distributions**
  Jayadev Acharya, Massachusetts Institute of Technology
  Constantinos Daskalakis, MIT
  Gautam C Kamath, MIT

- **Market Scoring Rules Act As Opinion Pools For Risk-Averse Agents**
  Mithun Chakraborty, Washington Univ. in St. Louis
  Sanmay Das, Washington University in St. Louis

- **Information-theoretic lower bounds for convex optimization with erroneous oracles**
  Yaron Singer, Harvard University
  Jan Vondrak, IBM Research

- **Bandit Smooth Convex Optimization: Improving the Bias-Variance Tradeoff**
  Ofer Dekel, Microsoft Research
  Ronen Eldan
  Tomer Koren, Technion

- **Accelerated Mirror Descent in Continuous and Discrete Time**
  Walid Krichene, UC Berkeley
  Alexandre Bayen, UC Berkeley
  Peter L Bartlett, UC Berkeley

- **Adaptive Online Learning**
  Dylan J Foster, Cornell University
  Alexander Rakhlin, UPenn
  Karthik Sridharan, Cornell

ORAL SESSION

SESSION 6: 11:10 – 11:50 AM

Deep Visual Analogy-Making

Scott E Reed reedscot@umich.edu
Yi Zhang yeezhang@umich.edu
Yuting Zhang yutingzh@umich.edu
Honglak Lee honglak@eecs.umich.edu
University of Michigan

In addition to identifying the content within a single image, relating images and generating related images are critical tasks for image understanding. Recently, deep convolutional networks have yielded breakthroughs in producing image labels, annotations and captions, but have only just begun to be used for producing high-quality image outputs. In this paper we develop a novel deep network trained end-to-end to perform visual analogy making, which is the task of transforming a query image according to an example pair of related images. Solving this problem requires both accurately recognizing a visual relationship and generating a transformed query image accordingly. Inspired by recent advances in language modeling, we propose to solve visual analogies by learning to map images to a neural embedding in which analogical reasoning is simple, such as by vector subtraction and addition. In experiments, our model effectively models visual analogies on several datasets: 2D shapes, animated video game sprites, and 3D car models.

End-To-End Memory Networks

Sainbayar Sukhbaatar sainbar@cs.nyu.edu
New York University
arthur szlam aszlam@fb.com
Jason Weston jase@fb.com
Rob Fergus robfergus@fb.com
Facebook AI Research

We introduce a neural network with a recurrent attention model over a possibly large external memory. The architecture is a form of Memory Network (Weston et al., 2015) but unlike the model in that work, it is trained end-to-end, and hence requires significantly less supervision during training, making it more generally applicable in realistic settings. It can also be seen as an extension of RNNsearch to the case where multiple computational steps (hops) are performed per output symbol. The flexibility of the model allows us to apply it to tasks as diverse as (synthetic) question answering and to language modeling. For the former our approach is competitive with Memory Networks, but with less supervision. For the latter, on the Penn TreeBank and Text8 datasets our approach demonstrates comparable performance to RNNs and LSTMs. In both cases we show that the key concept of multiple computational hops yields improved results.

SPOTLIGHT SESSION

SESSION 6: 11:50 AM – 12:00 PM

- **Attention-Based Models for Speech Recognition**
  Jan K Chorowski, University of Wroclaw
  Dzmitry Bahdanau, Jacobs University, Germany
  Dmitriy Serdyuk, Université de Montréal
  Kyunghyun Cho, NYU
  Yoshua Bengio, U. Montreal

- **Where are they looking?**
  Adria Recasens, MIT
  Aditya Khosla, MIT
  Carl Vondrick, MIT
  Antonio Torralba, MIT

- **Semi-supervised Convolutional Neural Networks for Text Categorization via Region Embedding**
  Rie Johnson, RJ Research Consuulting
  Tong Zhang, Rutgers

- **Training Very Deep Networks**
  Rupesh K Srivastava, IDSIA
  Klaus Greff, IDSIA
  Juergen Schmidhuber

- **Deep Convolutional Inverse Graphics Network**
  Tejas D Kulkarni, MIT
  Will Whitney, MIT
  Pushmeet Kohli, Microsoft Research
  Josh Tenenbaum, MIT

- **Learning to Segment Object Candidates**
  Pedro O Pinheiro, EPFL
  Ronan Collobert, Facebook
  Piotr Dollar, Facebook AI Research

- **The Return of the Gating Network: Combining Generative Models and Discriminative Training in Natural Image Priors**
  Dan Rosenbaum, The Hebrew University
  Yair Weiss, Hebrew University

- **Spatial Transformer Networks**
  Max Jaderberg, Google DeepMind
  Karen Simonyan, Google DeepMind
  Andrew Zisserman, Google DeepMind
  koray kavukcuoglu, Google DeepMind

ORAL SESSION

SESSION 7: 2:00 – 3:30 PM

Invited Talk:
Diagnosis and Therapy of Psychiatric Disorders Based on Brain Dynamics

Mitsuo Kawato kawato@atr.jp
Advanced Telecommunication Research Inst.

Arthur Winfree was one of the pioneers who postulated that several diseases are actually disorders of dynamics of biological systems. Following this path, many now believe that psychiatric diseases are disorders of brain dynamics. Combination of noninvasive brain measurement techniques, brain decoding and neurofeedback, and machine learning algorithms opened up a revolutionary pathway to quantitative diagnosis and therapy of neuropsychiatric disorders.

A Reduced-Dimension fMRI Shared Response Model

Po-Hsuan (Cameron) Chen
pohsuan@princeton.edu
Janice Chen janice@princeton.edu
Yaara Yeshurun yaara@princeton.edu
Uri Hasson hasson@princeton.edu
Peter J Ramadge ramadge@princeton.edu
Princeton University
James Haxby james.v.haxby@dartmouth.edu
Dartmouth

We develop a shared response model for aggregating multi-subject fMRI data that accounts for different functional topographies among anatomical aligned datasets. Multi-subject data is critical for evaluating the generality and validity of findings across subjects, and its effective utilization helps improve analysis sensitivity. Our model demonstrates improved sensitivity in identifying a shared response for a variety of datasets and anatomical brain regions of interest. Furthermore, by removing the identified shared response, it allows improved detection of group differences. The ability to identify what is shared and what is not shared, opens the model to a wide range of multi-subject fMRI studies.

Attractor Network Dynamics Enable Preplay and Rapid Path Planning in Maze–like Environments

Dane S Corneil dane.corneil@epfl.ch
Wulfram Gerstner wulfram.gerstner@epfl.ch
EPFL

Rodents navigating in a well-known environment can rapidly learn and revisit observed reward locations, often after a single trial. While the mechanism for rapid path planning is unknown, the CA3 region in the hippocampus plays an important role, and emerging evidence suggests that place cell activity during hippocampal “preplay” periods may trace out future goal-directed trajectories. Here, we show how a particular mapping of space allows for the immediate generation of trajectories between arbitrary start and goal locations in an environment, based only on the mapped representation of the goal. We show that this representation can be implemented in a neural attractor network model, resulting in bump--like activity profiles resembling those of the CA3 region of hippocampus. Neurons tend to locally excite neurons with similar place field centers, while inhibiting other neurons with distant place field centers, such that stable bumps of activity can form at arbitrary locations in the environment. The network is initialized to represent a point in the environment, then weakly stimulated with an input corresponding to an arbitrary goal location. We show that the resulting activity can be interpreted as a gradient ascent on the value function induced by a reward at the goal location. Indeed, in networks with large place fields, we show that the network properties cause the bump to move smoothly from its initial location to the goal, around obstacles or walls. Our results illustrate that an attractor network with hippocampal-like attributes may be important for rapid path planning.

SPOTLIGHT SESSION

SESSION 7: 3:30 – 4:00 PM

- **Inferring Algorithmic Patterns with Stack-Augmented Recurrent Nets**
  Armand Joulin, Facebook AI research
  Tomas Mikolov, Facebook AI Research

- **Decoupled Deep Neural Network for Semi-supervised Semantic Segmentation**
  Seunghoon Hong, POSTECH
  Hyeonwoo Noh, POSTECH
  Bohyung Han, POSTECH

- **Action-Conditional Video Prediction using Deep Networks in Atari Games**
  Junhyuk Oh, University of Michigan
  Xiaoxiao Guo, University of Michigan
  Honglak Lee, U. Michigan
  Richard L Lewis, University of Michigan
  Satinder Singh, University of Michigan

- **On-the-Job Learning with Bayesian Decision Theory**
  Keenon Werling, Stanford University
  Arun Tejasvi Chaganty, Stanford
  Percy S Liang, Stanford University
  Chris Manning, Stanford University

- **Learning Wake-Sleep Recurrent Attention Models**
  Jimmy Ba, University of Toronto
  Russ R Salakhutdinov, University of Toronto
  Roger B Grosse, University of Toronto
  Brendan J Frey, U. Toronto

- **Backpropagation for Energy-Efficient Neuromorphic Computing**
  Steve K Esser, IBM Research-Almaden
  Rathinakumar Appuswamy, IBM Research-Almaden
  Paul Merolla, IBM Research-Almaden
  John Arthur, IBM Research-Almaden
  Dharmendra S Modha, IBM Research-Almaden

- **A Tractable Approximation to Optimal Point Process Filtering: Application to Neural Encoding**
  Yuval Harel, Technion
  Ron Meir, Technion
  Manfred Opper, TU Berlin

- **Color Constancy by Learning to Predict Chromaticity from Luminance**
  Ayan Chakrabarti, TTI Chicago

ORAL SESSION

SESSION 8: 4:30 – 5:40 PM

INVITED TALK: Computational Principles for Deep Neuronal Architectures

Haim Sompolinsky
haim@fiz.huji.ac.il
Hebrew University and Harvard University

Recent progress in machine applications of deep neural networks have highlighted the need for a theoretical understanding of the capacity and limitations of these architectures. I will review our understanding of sensory processing in such architectures in the context of the hierarchies of processing stages observed in many brain systems. I will also address the possible roles of recurrent and top - down connections, which are prominent features of brain information processing circuits.

Efficient Exact Gradient Update for training Deep Networks with Very Large Sparse Targets

Pascal Vincent vincentp@iro.umontreal.ca
Alexandre de Brébisson alexandre.de.brebisson@umontreal.ca
Xavier Bouthillier xavier.bouthillier@umontreal.ca
Universit de Montréal

An important class of problems involves training deep neural networks with sparse prediction targets of very high dimension D. These occur naturally in e.g. neural language models or the learning of word-embeddings, often posed as predicting the probability of next words among a vocabulary of size D (e.g. 200,000). Computing the equally large, but typically non-sparse D-dimensional output vector from a last hidden layer of reasonable dimension d (e.g. 500) incurs a prohibitive O(Dd) computational cost for each example, as does updating the D×d output weight matrix and computing the gradient needed for backpropagation to

previous layers. While efficient handling of large sparse network inputs is trivial, this case of large sparse targets is not, and has thus so far been sidestepped with approximate alternatives such as hierarchical softmax or sampling-based approximations during training. In this work we develop an original algorithmic approach that, for a family of loss functions that includes squared error and spherical softmax, can compute the exact loss, gradient update for the output weights, and gradient for backpropagation, all in O(d2) per example instead of O(Dd), remarkably without ever computing the D-dimensional output. The proposed algorithm yields a speedup of D4d, i.e. two orders of magnitude for typical sizes, for that critical part of the computations that often dominates the training time in this kind of network architecture.

SPOTLIGHT SESSION

SESSION 8: 5:40 – 6:05 PM

- **Pointer Networks**
  Oriol Vinyals, Google
  Meire Fortunato
  Navdeep Jaitly, Google

- **Precision-Recall-Gain Curves: PR Analysis Done Right**
  Peter Flach, University of Bristol
  Meelis Kull, University of Bristol

- **NEXT: A System for Real-World Development, Evaluation, and Application of Active Learning**
  Kevin G Jamieson, University of Wisconsin
  Lalit Jain, University of Wisconsin
  Chris Fernandez, University of Wisconsin
  Nicholas J. Glattard, University of Wisconsin
  Rob Nowak, University of Wisconsin

- **Structured Transforms for Small-Footprint Deep Learning**
  Vikas Sindhwani, Google
  Tara Sainath, Google
  Sanjiv Kumar, Google

- **Equilibrated adaptive learning rates for non-convex optimization**
  Yann Dauphin, Facebook AI Research
  Harm de Vries
  Yoshua Bengio, U. Montreal

DEMONSTRATIONS

Demonstrations, 7:00 – 11:59 PM

D7 CodaLab Worksheets for Reproducible, Executable Papers
Percy S Liang · Evelyne Viegas

D8 The pMMF multiresolution matrix factorization library
Risi Kondor · Pramod Mudrakarta · Nedelina Teneva

D9 Interactive Incremental Question Answering
Jordan L Boyd-Graber · Mohit Iyyer

D10 Scaling up visual search for product recommendation
Kevin Jing

D11 Accelerated Deep Learning on GPUs: From Large Scale Training to Embedded Deployment
Allison Gray · Julie Bernauer

D12 Data-Driven Speech Animation
Yisong Yue · Iain Matthews

POSTER SESSION

POSTERS 7:00 – 11:59 PM

1 Deep Visual Analogy-Making
Scott E Reed · Yi Zhang · Yuting Zhang · Honglak Lee

2 Where are they looking?
Adria Recasens · Aditya Khosla · Carl Vondrick · Antonio Torralba

3 Spatial Transformer Networks
Max Jaderberg · Karen Simonyan · Andrew Zisserman · koray kavukcuoglu

4 Training Very Deep Networks
Rupesh K Srivastava · Klaus Greff · Juergen Schmidhuber

5 Attention-Based Models for Speech Recognition
Jan K Chorowski · Dzmitry Bahdanau · Dmitriy Serdyuk · Kyunghyun Cho · Yoshua Bengio

6 Deep Convolutional Inverse Graphics Network
Tejas D Kulkarni · Will Whitney · Pushmeet Kohli · Josh Tenenbaum

7 End-To-End Memory Networks
Sainbayar Sukhbaatar · arthur szlam · Jason Weston · Rob Fergus

8 Learning to Segment Object Candidates
Pedro O Pinheiro · Ronan Collobert · Piotr Dollar

9 Inferring Algorithmic Patterns with Stack-Augmented Recurrent Nets
Armand Joulin · Tomas Mikolov

10 Attractor Network Dynamics Enable Preplay and Rapid Path Planning in Maze–like Environments
Dane S Corneil · Wulfram Gerstner

11 Semi-supervised Convolutional Neural Networks for Text Categorization via Region Embedding
Rie Johnson · Tong Zhang

12 The Return of the Gating Network: Combining Generative Models and Discriminative Training in Natural Image Priors
Dan Rosenbaum · Yair Weiss

13 Backpropagation for Energy-Efficient Neuromorphic Computing
Steve K Esser · Rathinakumar Appuswamy · Paul Merolla · John Arthur · Dharmendra S Modha

14 Learning Wake-Sleep Recurrent Attention Models
Jimmy Ba · Russ R Salakhutdinov · Roger B Grosse · Brendan J Frey

15 On-the-Job Learning with Bayesian Decision Theory
Keenon Werling · Arun Tejasvi Chaganty · Percy S Liang · Chris Manning

16 Color Constancy by Learning to Predict Chromaticity from Luminance
Ayan Chakrabarti

17 Decoupled Deep Neural Network for Semi-supervised Semantic Segmentation
Seunghoon Hong · Hyeonwoo Noh · Bohyung Han

18 Action-Conditional Video Prediction using Deep Networks in Atari Games
Junhyuk Oh · Xiaoxiao Guo · Honglak Lee · Richard L Lewis · Satinder Singh

19 Bayesian Active Model Selection with an Application to Automated Audiometry
Jacob Gardner · Luiz Gustavo Sant Anna Malkomes Muniz · Roman Garnett · Kilian Q Weinberger · Dennis Barbour · John Cunningham

20 Efficient and Robust Automated Machine Learning
Matthias Feurer · Aaron Klein · Katharina Eggensperger · Jost Springenberg · Manuel Blum · Frank Hutter

21 A Framework for Individualizing Predictions of Disease Trajectories by Exploiting Multi-Resolution Structure
Peter Schulam · Suchi Saria

22 Pointer Networks
Oriol Vinyals · Meire Fortunato · Navdeep Jaitly

23 A Reduced-Dimension fMRI Shared Response Model
Po-Hsuan (Cameron) Chen · Janice Chen · Yaara Yeshurun · Uri Hasson · James Haxby · Peter J Ramadge

24 Efficient Exact Gradient Update for training Deep Networks with Very Large Sparse Targets
Pascal Vincent · Alexandre de Brébisson · Xavier Bouthillier

25 Precision-Recall-Gain Curves: PR Analysis Done Right
Peter Flach · Meelis Kull

26 A Tractable Approximation to Optimal Point Process Filtering: Application to Neural Encoding
Yuval Harel · Ron Meir · Manfred Opper

27 Equilibrated adaptive learning rates for non-convex optimization
Yann Dauphin · Harm de Vries · Yoshua Bengio

28 NEXT: A System for Real-World Development, Evaluation, and Application of Active Learning
Kevin G Jamieson · Lalit Jain · Chris Fernandez · Nicholas J. Glattard · Rob Nowak

29 Gaussian Process Random Fields
Dave Moore · Stuart J Russell

30 MCMC for Variationally Sparse Gaussian Processes
James J Hensman · Alexander G Matthews · Maurizio Filippone · Zoubin Ghahramani

31 Streaming, Distributed Variational Inference for Bayesian Nonparametrics
Trevor Campbell · Julian Straub · John W Fisher III · Jonathan P How

32 Fixed-Length Poisson MRF: Adding Dependencies to the Multinomial
David I Inouye · Pradeep K Ravikumar · Inderjit S Dhillon

33 Human Memory Search as Initial-Visit Emitting Random Walk
Kwang-Sung Jun · Jerry Zhu · Timothy T Rogers · Zhuoran Yang · ming yuan

34 Structured Transforms for Small-Footprint Deep Learning
Vikas Sindhwani · Tara Sainath · Sanjiv Kumar

35 Spectral Learning of Large Structured HMMs for Comparative Epigenomics
Chicheng Zhang · Jimin Song · Kamalika Chaudhuri · Kevin Chen

36 A Structural Smoothing Framework For Robust Graph Comparison
Pinar Yanardag Delul · S.V.N. Vishwanathan

37 Optimization Monte Carlo: Efficient and Embarrassingly Parallel Likelihood-Free Inference
Ted Meeds · Max Welling

38 Inverse Reinforcement Learning with Locally Consistent Reward Functions
Quoc Phong Nguyen · Bryan Kian Hsiang Low · Patrick Jaillet

39 Consistent Multilabel Classification
Sanmi Koyejo · Nagarajan Natarajan · Pradeep K Ravikumar · Inderjit S Dhillon

40 Is Approval Voting Optimal Given Approval Votes?
Ariel D Procaccia · Nisarg Shah

41 A Normative Theory of Adaptive Dimensionality Reduction in Neural Networks
Cengiz Pehlevan · Dmitri Chklovskii

45 Efficient Non-greedy Optimization of Decision Trees
Mohammad Norouzi · Maxwell Collins · Matthew Johnson · David J Fleet · Pushmeet Kohli

46 Statistical Topological Data Analysis - A Kernel Perspective
Roland Kwitt · Stefan Huber · Marc Niethammer · Weili Lin · Ulrich Bauer

44 Variational Consensus Monte Carlo
Maxim Rabinovich · Elaine Angelino · Michael I Jordan

45 Softstar: Heuristic-Guided Probabilistic Inference
Mathew Monfort · Brenden M Lake · Brian Ziebart · Patrick Lucey · Josh Tenenbaum

46 Gradient-free Hamiltonian Monte Carlo with Efficient Kernel Exponential Families
Heiko Strathmann · Dino Sejdinovic · Samuel Livingstone · Zoltan Szabo · Arthur Gretton

47 A Complete Recipe for Stochastic Gradient MCMC
Yi-An Ma · Tianqi Chen · Emily Fox

48 Barrier Frank-Wolfe for Marginal Inference
Rahul G Krishnan · Simon Lacoste-Julien · David Sontag

49 Practical and Optimal LSH for Angular Distance
Alexandr Andoni · Piotr Indyk · Thijs Laarhoven · Ilya Razenshteyn · Ludwig Schmidt

50 Principal Differences Analysis: Interpretable Characterization of Differences between Distributions
Jonas W Mueller · Tommi Jaakkola

51 Kullback-Leibler Proximal Variational Inference
Emtiyaz E Khan · Pierre Baque · François Fleuret · Pascal Fua

52 Learning Large-Scale Poisson DAG Models based on OverDispersion Scoring
Gunwoong Park · Garvesh Raskutti

53 Streaming Min-max Hypergraph Partitioning
Dan Alistarh · Jennifer Iglesias · Milan Vojnovic

54 Efficient Output Kernel Learning for Multiple Tasks
Pratik Jawanpuria · Maksim Lapin · Matthias Hein · Bernt Schiele

55 Gradient Estimation Using Stochastic Computation Graphs
John Schulman · Nicolas Heess · Theophane Weber · Pieter Abbeel

56 Lifted Inference Rules With Constraints
Happy Mittal · Anuj Mahajan · Vibhav G Gogate · Parag Singla

57 Sparse PCA via Bipartite Matchings
megas asteris · Dimitris Papailiopoulos · Tasos Kyrillidis · Alex G Dimakis

58 Empirical Localization of Homogeneous Divergences on Discrete Sample Spaces
Takashi Takenouchi · Takafumi Kanamori

59 Weighted Theta Functions and Embeddings with Applications to Max-Cut, Clustering and Summarization
Fredrik D Johansson · Ankani Chattoraj · Chiranjib Bhattacharyya · Devdatt Dubhashi

60 Online Rank Elicitation for Plackett-Luce: A Dueling Bandits Approach
Balázs Szörényi · Róbert Busa-Fekete · Adil Paul · Eyke Hüllermeier

61 Segregated Graphs and Marginals of Chain Graph Models
Ilya Shpitser

62 Approximating Sparse PCA from Incomplete Data
ABHISEK KUNDU · Petros Drineas · Malik Magdon-Ismail

63 Multi-Layer Feature Reduction for Tree Structured Group Lasso via Hierarchical Projection
Jie Wang · Jieping Ye

64 Recovering Communities in the General Stochastic Block Model Without Knowing the Parameters
Emmanuel Abbe · Colin Sandon

65 Maximum Likelihood Learning With Arbitrary Treewidth via Fast-Mixing Parameter Sets
Justin Domke

66 Testing Closeness With Unequal Sized Samples
Bhaswar Bhattacharya · Greg Valiant

67 Learning Causal Graphs with Small Interventions
Karthikeyan Shanmugam · Murat Kocaoglu · Alex G Dimakis · Sriram Vishwanath

68 Regret-Based Pruning in Extensive-Form Games
Noam Brown · Tuomas Sandholm

69 Nonparametric von Mises Estimators for Entropies, Divergences and Mutual Informations
Kirthevasan Kandasamy · Akshay Krishnamurthy · Barnabas Poczos · Larry Wasserman · james m robins

70 Bounding errors of Expectation-Propagation
Guillaume P Dehaene · Simon Barthelmé

71 Market Scoring Rules Act As Opinion Pools For Risk-Averse Agents
Mithun Chakraborty · Sanmay Das

72 Local Smoothness in Variance Reduced Optimization
Daniel Vainsencher · Han Liu · Tong Zhang

73 High Dimensional EM Algorithm: Statistical Optimization and Asymptotic Normality
Zhaoran Wang · Quanquan Gu · Yang Ning · Han Liu

74 Associative Memory via a Sparse Recovery Model
Arya Mazumdar · Ankit Singh Rawat

75 Matrix Completion Under Monotonic Single Index Models
Ravi Ganti · Laura Balzano · Rebecca Willett

76 Sparse Linear Programming via Primal and Dual Augmented Coordinate Descent
Ian En-Hsu Yen · Kai Zhong · Cho-Jui Hsieh · Pradeep K Ravikumar · Inderjit S Dhillon

77 Convergence rates of sub-sampled Newton methods
Murat A. Erdogdu · Andrea Montanari

78 Variance Reduced Stochastic Gradient Descent with Neighbors
Thomas Hofmann · Aurelien Lucchi · Simon Lacoste-Julien · Brian McWilliams

79 Non-convex Statistical Optimization for Sparse Tensor Graphical Model
Wei Sun · Zhaoran Wang · Han Liu · Guang Cheng

80 Convergence Rates of Active Learning for Maximum Likelihood Estimation
Kamalika Chaudhuri · Sham Kakade · Praneeth Netrapalli · Sujay Sanghavi

81 When are Kalman-Filter Restless Bandits Indexable?
Christopher R Dance · Tomi Silander

82 Policy Gradient for Coherent Risk Measures
Aviv Tamar · Yinlam Chow · Mohammad Ghavamzadeh · Shie Mannor

83 A Dual Augmented Block Minimization Framework for Learning with Limited Memory
Ian En-Hsu Yen · Shan-Wei Lin · Shou-De Lin

84 On the Global Linear Convergence of Frank-Wolfe Optimization Variants
Simon Lacoste-Julien · Martin Jaggi

85 Quartz: Randomized Dual Coordinate Ascent with Arbitrary Sampling
Zheng Qu · Peter Richtarik · Tong Zhang

86 A Generalization of Submodular Cover via the Diminishing Return Property on the Integer Lattice
Tasuku Soma · Yuichi Yoshida

87 A Universal Catalyst for First-Order Optimization
Hongzhou Lin · Julien Mairal · Zaid Harchaoui

88 Fast and Memory Optimal Low-Rank Matrix Approximation
Se-Young Yun · marc lelarge · Alexandre Proutiere

89 Stochastic Online Greedy Learning with Semi-bandit Feedbacks
Tian Lin · Jian Li · Wei Chen

90 Linear Multi-Resource Allocation with Semi-Bandit Feedback
Tor Lattimore · Koby Crammer · Csaba Szepesvari

91 Exactness of Approximate MAP Inference in Continuous MRFs
Nicholas Ruozzi

92 On the consistency theory of high dimensional variable screening
Xiangyu Wang · Chenlei Leng · David B Dunson

93 Finite-Time Analysis of Projected Langevin Monte Carlo
Sebastien Bubeck · Ronen Eldan · Joseph Lehec

94 Optimal Testing for Properties of Distributions
Jayadev Acharya · Constantinos Daskalakis · Gautam C Kamath

95 Learning Theory and Algorithms for Forecasting Non-stationary Time Series
Vitaly Kuznetsov · Mehryar Mohri

96 Accelerated Mirror Descent in Continuous and Discrete Time
Walid Krichene · Alexandre Bayen · Peter L Bartlett

97 Information-theoretic lower bounds for convex optimization with erroneous oracles
Yaron Singer · Jan Vondrak

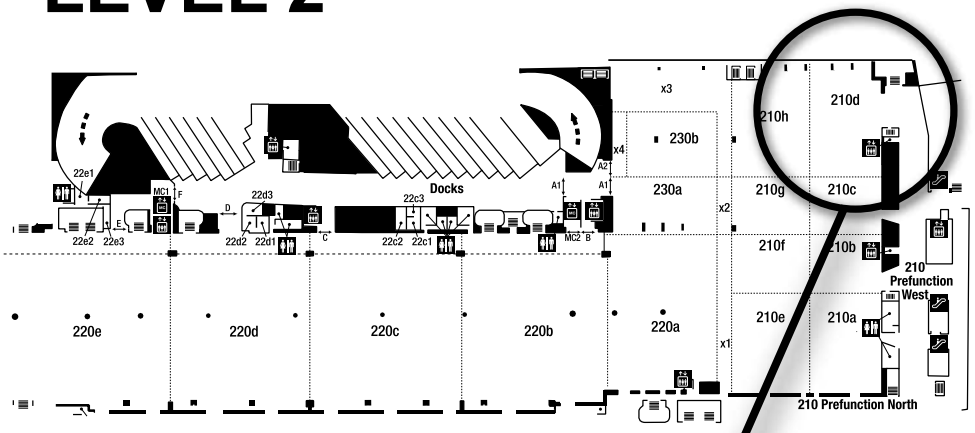
98 Bandit Smooth Convex Optimization: Improving the Bias-Variance Tradeoff
Ofer Dekel · Ronen Eldan · Tomer Koren

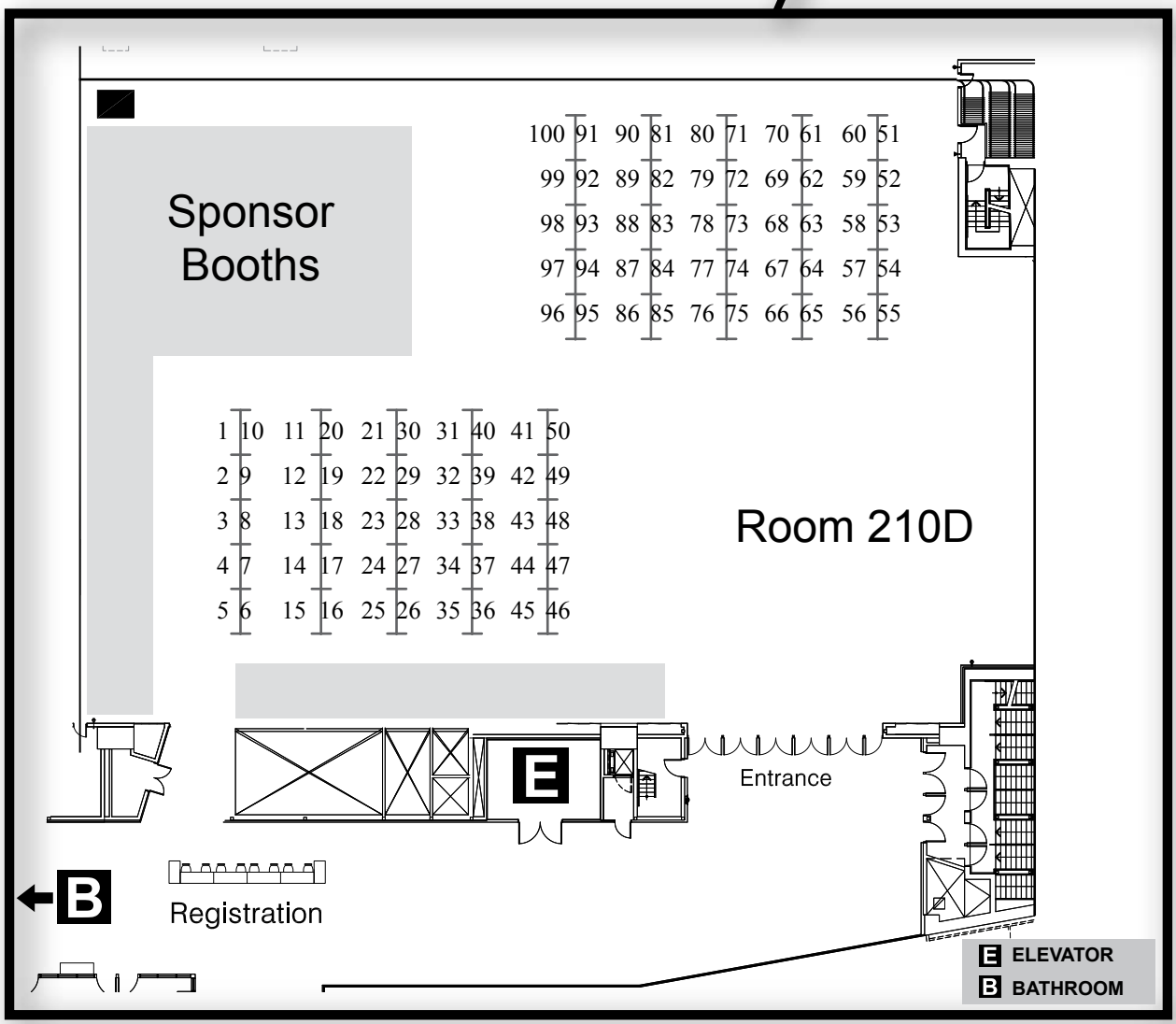
99 Beyond Sub-Gaussian Measurements: High-Dimensional Structured Estimation with Sub-Exponential Designs
Vidyashankar Sivakumar · Arindam Banerjee · Pradeep K Ravikumar

100 Adaptive Online Learning
Dylan J Foster · Alexander Rakhlin · Karthik Sridharan

WEDNESDAY POSTER FLOORPLAN

LEVEL 2





1 Deep Visual Analogy-Making

Scott E Reed reedscot@umich.edu
Yi Zhang yeezhang@umich.edu
Yuting Zhang yutingzh@umich.edu
Honglak Lee honglak@eecs.umich.edu
University of Michigan

In addition to identifying the content within a single image, relating images and generating related images are critical tasks for image understanding. Recently, deep convolutional networks have yielded breakthroughs in producing image labels, annotations and captions, but have only just begun to be used for producing high-quality image outputs. In this paper we develop a novel deep network trained end-to-end to perform visual analogy making, which is the task of transforming a query image according to an example pair of related images. Solving this problem requires both accurately recognizing a visual relationship and generating a transformed query image accordingly. Inspired by recent advances in language modeling, we propose to solve visual analogies by learning to map images to a neural embedding in which analogical reasoning is simple, such as by vector subtraction and addition. In experiments, our model effectively models visual analogies on several datasets: 2D shapes, animated video game sprites, and 3D car models.

2 Where are they looking?

Adria Recasens recasens@mit.edu
Aditya Khosla khosla@mit.edu
Carl Vondrick vondrick@mit.edu
Antonio Torralba torralba@mit.edu
MIT

Humans have the remarkable ability to follow the gaze of other people to identify what they are looking at. Following eye gaze, or gaze-following, is an important ability that allows us to understand what other people are thinking, the actions they are performing, and even predict what they might do next. Despite the importance of this topic, this problem has only been studied in limited scenarios within the computer vision community. In this paper, we propose a deep neural network-based approach for gaze-following and a new benchmark dataset for thorough evaluation. Given an image and the location of a head, our approach follows the gaze of the person and identifies the object being looked at. After training, the network is able to discover how to extract head pose and gaze orientation, and to select objects in the scene that are in the predicted line of sight and likely to be looked at (such as televisions, balls and food). The quantitative evaluation shows that our approach produces reliable results, even when viewing only the back of the head. While our method outperforms several baseline approaches, we are still far from reaching human performance at this task. Overall, we believe that this is a challenging and important task that deserves more attention from the community.

3 Spatial Transformer Networks

Max Jaderberg jaderberg@google.com
Karen Simonyan simonyan@google.com
Andrew Zisserman zisserman@google.com
koray kavukcuoglu korayk@google.com
Google DeepMind

Convolutional Neural Networks define an exceptionallypowerful class of model, but are still limited by the lack of abilityto be spatially invariant to the input data in a computationally and parameterefficient manner. In this work we introduce a new learnable module, theSpatial Transformer, which explicitly allows the spatial manipulation ofdata within the network. This differentiable module can be insertedinto existing convolutional architectures, giving neural networks the ability toactively spatially transform feature maps, conditional on the feature map itself,without any extra training supervision or modification to the optimisation process. We show that the useof spatial transformers results in models which learn invariance to translation,scale, rotation and more generic warping, resulting in state-of-the-artperformance on several benchmarks, and for a numberof classes of transformations.

4 Training Very Deep Networks

Rupesh K Srivastava rupesh@idsia.ch
Klaus Greff klaus@idsia.ch
Juergen Schmidhuber juergen@idsia.ch
IDSIA

Theoretical and empirical evidence indicates that the depth of neural networks is crucial for their success. However, training becomes more difficult as depth increases, and training of very deep networks remains an open problem. Here we introduce a new architecture designed to overcome this. Our so-called highway networks allow for unimpeded information flow across many layers on information highways. They are inspired by Long Short-Term Memory recurrent networks and use adaptive gating units to regulate the information flow. Even with hundreds of layers, highway networks can learn directly through simple gradient descent. This allows for studying extremely deep and efficient architectures.

5 Attention-Based Models for Speech Recognition

Jan K Chorowski jan.chorowski@ii.uni.wroc.pl
University of Wroclaw
Dzmitry Bahdanau d.bahdanau@jacobs-university.de
Jacobs University, Germany
Dmitriy Serdyuk serdyuk@iro.umontreal.ca
Université de Montréal
Kyunghyun Cho kyunghyun.cho@nyu.edu
NYU
Yoshua Bengio yoshua.bengio@gmail.com
U. Montreal

Recurrent sequence generators conditioned on input data through an attention mechanism have recently shown very good performance on a range of tasks including machine translation, handwriting synthesis and image caption generation. We extend the attention-mechanism with features needed for speech recognition. We show that while an adaptation of the model used for machine translation reaches a competitive 18.6\% phoneme error rate (PER) on the TIMIT phoneme recognition task, it can only be applied to utterances which are roughly as long as the ones it was trained on. We offer a qualitative explanation of this failure and propose a novel and generic method of adding location-awareness to the attention mechanism to alleviate this issue. The new method yields a model that is robust to long inputs and achieves 18\% PER in single utterances and 20\% in 10-times longer (repeated) utterances. Finally, we propose a change to the attention mechanism that prevents it from concentrating too much on single frames, which further reduces PER to 17.6\% level.

6 Deep Convolutional Inverse Graphics Network

Tejas D Kulkarni tejasdkulkarni@gmail.com
Will Whitney wwhitney@mit.edu
Josh Tenenbaum jbt@mit.edu
MIT
Pushmeet Kohli pkohli@microsoft.com
Microsoft Research

This paper presents the Deep Convolution Inverse Graphics Network (DC-IGN), a model that learns an interpretable representation of images. This representation is disentangled with respect to transformations such as out-of-plane rotations and lighting variations. The DC-IGN model is composed of multiple layers of convolution and de-convolution operators and is trained using the Stochastic Gradient Variational Bayes (SGVB) algorithm. We propose a training procedure to encourage neurons in the graphics code layer to represent a specific transformation (e.g. pose or light). Given a single input image, our model can generate new images of the same object with variations in pose and lighting. We present qualitative and quantitative results of the model's efficacy at learning a 3D rendering engine.

7 End-To-End Memory Networks

Sainbayar Sukhbaatar sainbar@cs.nyu.edu
New York University
arthur szlam aszlam@fb.com
Jason Weston jase@fb.com
Rob Fergus robfergus@fb.com
Facebook AI Research

We introduce a neural network with a recurrent attention model over a possibly large external memory. The architecture is a form of Memory Network (Weston et al., 2015) but unlike the model in that work, it is trained end-to-end, and hence requires significantly less supervision during training, making it more generally applicable in realistic settings. It can also be seen as an extension of RNNsearch to the case where multiple computational steps (hops) are performed per output symbol. The flexibility of the model allows us to apply it to tasks as diverse as (synthetic) question answering and to language modeling. For the former our approach is competitive with Memory Networks, but with less supervision. For the latter, on the Penn TreeBank and Text8 datasets our approach demonstrates comparable performance to RNNs and LSTMs. In both cases we show that the key concept of multiple computational hops yields improved results.

8 Learning to Segment Object Candidates

Pedro O Pinheiro pedro@opinheiro.com
EPFL
Ronan Collobert ronan@collobert.com
Piotr Dollar pdollar@gmail.com
Facebook AI Research

Recent object detection systems rely on two critical steps: (1) a set of object proposals is predicted as efficiently as possible, and (2) this set of candidate proposals is then passed to an object classifier. Such approaches have been shown they can be fast, while achieving the state of the art in detection performance. In this paper, we propose a new way to generate object proposals, introducing an approach based on a discriminative convolutional network. Our model is trained jointly with two objectives: given an image patch, the first part of the system outputs a class-agnostic segmentation mask, while the second part of the system outputs the likelihood of the patch being centered on a full object. At test time, the model is efficiently applied on the whole test image and generates a set of segmentation masks, each of them being assigned with a corresponding object likelihood score. We show that our model yields significant improvements over state-of-the-art object proposal algorithms. In particular, compared to previous approaches, our model obtains substantially higher object recall using fewer proposals. We also show that our model is able to generalize to unseen categories it has not seen during training. Unlike all previous approaches for generating object masks, we do not rely on edges, superpixels, or any other form of low-level segmentation.

9 Inferring Algorithmic Patterns with Stack-Augmented Recurrent Nets

Armand Joulin ajoulin@fb.com
Tomas Mikolov tmikolov@fb.com
Facebook AI Research

Despite the recent achievements in machine learning, we are still very far from achieving real artificial intelligence. In this paper, we discuss the limitations of standard deep learning approaches and show that some of these limitations can be overcome by learning how to grow the complexity of a model in a structured way. Specifically, we study the simplest sequence prediction problems that are beyond the scope of what is learnable with standard recurrent networks, algorithmically generated sequences which can only be learned by models which have the capacity to count and to memorize sequences. We show that some basic algorithms can be learned from sequential data using a recurrent network associated with a trainable memory.

10 Attractor Network Dynamics Enable Preplay and Rapid Path Planning in Maze–like Environments

Dane S Corneil dane.corneil@epfl.ch
Wulfram Gerstner wulfram.gerstner@epfl.ch
EPFL

Rodents navigating in a well-known environment can rapidly learn and revisit observed reward locations, often after a single trial. While the mechanism for rapid path planning is unknown, the CA3 region in the hippocampus plays an important role, and emerging evidence suggests that place cell activity during hippocampal "preplay" periods may trace out future goal-directed trajectories. Here, we show how a particular mapping of space allows for the immediate generation of trajectories between arbitrary start and goal locations in an environment, based only on the mapped representation of the goal. We show that this representation can be implemented in a neural attractor network model, resulting in bump--like activity profiles resembling those of the CA3 region of hippocampus. Neurons tend to locally excite neurons with similar place field centers, while inhibiting other neurons with distant place field centers, such that stable bumps of activity can form at arbitrary locations in the environment. The network is initialized to represent a point in the environment, then weakly stimulated with an input corresponding to an arbitrary goal location. We show that the resulting activity can be interpreted as a gradient ascent on the value function induced by a reward at the goal location. Indeed, in networks with large place fields, we show that the network properties cause the bump to move smoothly from its initial location to the goal, around obstacles or walls. Our results illustrate that an attractor network with hippocampal-like attributes may be important for rapid path planning.

11 Semi-supervised Convolutional Neural Networks for Text Categorization via Region Embedding

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

This paper presents a new semi-supervised framework with convolutional neural networks (CNNs) for text categorization. Unlike the previous approaches that rely on word embeddings, our method learns embeddings of small text regions from unlabeled data for integration into a supervised CNN. The proposed scheme for embedding learning is based on the idea of two-view semi-supervised learning, which is intended to be useful for the task of interest even though the training is done on unlabeled data. Our models achieve better results than previous approaches on sentiment classification and topic classification tasks.

12 The Return of the Gating Network: Combining Generative Models and Discriminative Training in Natural Image Priors

Dan Rosenbaum danrsm@cs.huji.ac.il
Yair Weiss yweiss@cs.huji.ac.il
Hebrew University

In recent years, approaches based on machine learning have achieved state-of-the-art performance on image restoration problems. Successful approaches include both generative models of natural images as well as discriminative training of deep neural networks. Discriminative training of feed forward architectures allows explicit control over the computational cost of performing restoration and therefore often leads to better performance at the same cost at run time. In contrast, generative models have the advantage that they can be trained once and then adapted to any image restoration task by a simple use of Bayes' rule. In this paper we show how to combine the strengths of both approaches by training a discriminative, feed-forward architecture to predict the state of latent variables in a generative model of natural images. We apply this idea to the very successful Gaussian Mixture Model (GMM) model of natural images. We show that it is possible to achieve comparable performance as the original GMM model but with two orders of magnitude improvement in run time while maintaining the advantage of generative models.

13 Backpropagation for Energy-Efficient Neuromorphic Computing

Steve K Esser sesser@us.ibm.com
Rathinakumar Appuswamy rappusw@us.ibm.com
Paul Merolla pameroll@us.ibm.com
John Arthur arthurjo@us.ibm.com
Dharmendra S Modha dmodha@us.ibm.com
IBM Research-Almaden

Solving real world problems with embedded neural networks requires both training algorithms that achieve high performance and compatible hardware that runs in real time while remaining energy efficient. For the former, deep learning using backpropagation has recently achieved a string of successes across many domains and datasets. For the latter, neuromorphic chips that run spiking neural networks have recently achieved unprecedented energy efficiency. To bring these two advances together, we must first resolve the incompatibility between backpropagation, which uses continuous-output neurons and synaptic weights, and neuromorphic designs, which employ spiking neurons and discrete synapses. Our approach is to treat spikes and discrete synapses as continuous probabilities, which allows training the network using standard backpropagation. The trained network naturally maps to neuromorphic hardware by sampling the probabilities to create one or more networks, which are merged using ensemble averaging. To demonstrate, we trained a sparsely connected network that runs on the TrueNorth chip using the MNIST dataset. With a high performance network (ensemble of 64), we achieve 99.42% accuracy at 121μJ per image, and with a high efficiency network (ensemble of 1) we achieve 92.7% accuracy at 0.408μJ per image.

14 Learning Wake-Sleep Recurrent Attention Models

Jimmy Ba jimmy@psi.utoronto.ca
Russ R Salakhutdinov rsalakhu@cs.toronto.edu
Roger B Grosse rgrosse@cs.toronto.edu
Brendan J Frey frey@psi.toronto.edu
University of Toronto

Despite their success, convolutional neural networks are computationally expensive because they must examine all image locations. Stochastic attention-based models have been shown to improve computational efficiency at test time, but they remain difficult to train because of intractable posterior inference and high variance in the stochastic gradient estimates. Borrowing techniques from the literature on training deep generative models, we present the Wake-Sleep Recurrent Attention Model, a method for training stochastic attention networks which improves posterior inference and which reduces the variability in the stochastic gradients. We show that our method can greatly speed up the training time for stochastic attention networks in the domains of image classification and caption generation.

15 On-the-Job Learning with Bayesian Decision Theory

Keenon Werling keenon@stanford.edu
Arun Tejasvi Chaganty chaganty@cs.stanford.edu
Percy S Liang pliang@cs.stanford.edu
Chris Manning manning@cs.stanford.edu
Stanford University

How can we deploy a high-accuracy system starting with zero training examples? We consider an "on-the-job" setting, where as inputs arrive, we use crowdsourcing to resolve uncertainty where needed and output our prediction when confident. As the model improves over time, the reliance on crowdsourcing queries decreases. We cast our setting as a stochastic game based on Bayesian decision theory, which allows us to balance latency, cost, and accuracy objectives in a principled way. Computing the optimal policy is intractable, so we develop an approximation based on Monte Carlo Tree Search. We tested our approach across three datasets---named-entity recognition, sentiment classification,and image classification. On the NER task we obtained a 6-7 fold reduction in cost compared to full human annotation. We also achieve a 17% F1 improvement over having a single human label the whole set, and a 28% F1 improvement over online learning.

16 Color Constancy by Learning to Predict Chromaticity from Luminance

Ayan Chakrabarti ayanc@ttic.edu
TTI Chicago

Color constancy is the recovery of true surface color from observed color, and requires estimating the chromaticity of scene illumination to correct for the bias it induces. In this paper, we show that the per-pixel color statistics of natural scenes---without any spatial or semantic context---can by themselves be a powerful cue for color constancy. Specifically, we describe an illuminant estimation method that is built around a "classifier" for identifying the true chromaticity of a pixel given its luminance (absolute brightness across color channels). During inference, each pixel's observed color restricts its true chromaticity to those values that can be explained by one of a candidate set of illuminants, and applying the classifier over these values yields a distribution over the corresponding illuminants. A global estimate for the scene illuminant is computed through a simple aggregation of these distributions across all pixels. We begin by simply defining the luminance-to-chromaticity classifier by computing empirical histograms over discretized chromaticity and luminance values from a training set of natural images. These histograms reflect a preference for hues corresponding to smooth reflectance functions, and for achromatic colors in brighter pixels. Despite its simplicity, the resulting estimation algorithm outperforms current state-of-the-art color constancy methods. Next, we propose a method to learn the luminance-to-chromaticity classifier "end-to-end". Using stochastic gradient descent, we set chromaticity-luminance likelihoods to minimize errors in the final scene illuminant estimates on a training set. This leads to further improvements in accuracy, most significantly in the tail of the error distribution.

17 Decoupled Deep Neural Network for Semi-supervised Semantic Segmentation

Seunghoon Hong maga33@postech.ac.kr
Hyeonwoo Noh hyeonwoonoh_@postech.ac.kr
Bohyung Han bhhan@postech.ac.kr
POSTECH

We propose a novel deep neural network architecture for semi-supervised semantic segmentation using heterogeneous annotations. Contrary to existing approaches posing semantic segmentation as region-based classification, our algorithm decouples classification and segmentation, and learns a separate network for each task. In this architecture, labels associated with an image are identified by classification network, and binary segmentation is subsequently performed for each identified label by segmentation network. The decoupled architecture enables us to learn classification and segmentation networks separately based on the training data with image-level and pixel-wise class labels, respectively. It facilitates to reduce search space for segmentation effectively by exploiting class-specific activation maps obtained from bridging layers. Our algorithm shows outstanding performance compared to other semi-supervised approaches even with much less training images with strong annotations in PASCAL VOC dataset.

18 Action-Conditional Video Prediction using Deep Networks in Atari Games

Junhyuk Oh junhyuk@umich.edu
Xiaoxiao Guo guoxiao@umich.edu
Honglak Lee honglak@eecs.umich.edu
Richard L Lewis rickl@umich.edu
Satinder Singh baveja@umich.edu
University of Michigan

Motivated by vision-based reinforcement learning (RL) problems, in particular Atari games from the recent benchmark Aracade Learning Environment (ALE), we consider spatio-temporal prediction problems where future (image-)frames are dependent on control variables or actions as well as previous frames. While not composed of natural scenes, frames in Atari games are high-dimensional in size, can involve tens of objects with one or more objects being controlled by the actions directly and many other objects being influenced indirectly, can involve entry and departure of objects, and can involve deep partial observability. We propose and evaluate two deep neural network architectures that consist of encoding, action-conditional transformation, and decoding layers based on convolutional neural networks and recurrent neural networks. Experimental results show that the proposed architectures are able to generate visually-realistic frames that are also useful for control over approximately 100-step action-conditional futures in some games. To the best of our knowledge, this paper is the first to make and evaluate long-term predictions on high-dimensional video conditioned by control inputs.

19 Bayesian Active Model Selection with an Application to Automated Audiometry

Jacob Gardner jrg365@cornell.edu
Kilian Q Weinberger kqw4@cornell.edu
Cornell University
Gustavo Malkomes luizgustavo@wustl.edu
Roman Garnett garnett@wustl.edu
Dennis Barbour dbarbour@wustl.edu
Washington University in St. Louis
John Cunningham jpc2181@columbia.edu
University of Columbia

We introduce a novel information-theoretic approach for active model selection and demonstrate its effectiveness in a real-world application. Although our method works with arbitrary models, we focus on actively learning the appropriate structure for Gaussian process (GP) models with arbitrary observation likelihoods. We then apply this framework to rapid screening for noise-induced hearing loss (NIHL), a widespread and preventible disability if diagnosed early. We construct a GP model for pure-tone audiometric responses for patients with NIHL. Using this and a previously published model for healthy responses, the proposed method is shown to be capable of diagnosing the presence or absence of NIHL with drastically fewer samples than existing approaches. Further, the method is extremely fast and enables the hearing-loss diagnosis to be performed in real time.

20 Efficient and Robust Automated Machine Learning

Matthias Feurer feurerm@cs.uni-freiburg.de
Aaron Klein kleinaa@cs.uni-freiburg.de
Katharina Eggensperger eggenspk@cs.uni-freiburg.de
Jost Springenberg springj@cs.uni-freiburg.de
Manuel Blum mblum@cs.uni-freiburg.de
Frank Hutter fh@cs.uni-freiburg.de
University of Freiburg

The success of machine learning in a broad range of applications has led to an ever-growing demand for machine learning systems that can be used off the shelf by non-experts. To be effective in practice, such systems need to automatically choose a good algorithm and feature preprocessing approach for a new dataset at hand, and also set their respective hyperparameters. Recent work has startedto tackle this automated machine learning (AutoML) problem with the help of efficient Bayesian optimization methods. We substantially improve upon these methods by taking into account past performance on similar datasets, and by constructing ensembles from the models evaluated during the optimization. We also introduce a robust new AutoML system based on scikit-learn (using 16 classifiers,14 feature processing methods, and 3 data preprocessing methods, giving rise to a structured hypothesis space with 132 hyperparameters). This system, which we dub auto-sklearn, won the first phase of the ongoing ChaLearn AutoML challenge, and our own comprehensive analysis on over 100 diverse datasets shows that it substantially outperforms the previous state of the art in AutoML. We also demonstrate the clear performance gains due to each of our contributions and derive insights into the effectiveness of the individual components of auto-sklearn.

21 A Framework for Individualizing Predictions of Disease Trajectories by Exploiting Multi-Resolution Structure

Peter Schulam pschulam@gmail.com
Suchi Saria ssaria@cs.jhu.edu
Johns Hopkins University

For many complex diseases, there is a wide variety of ways in which an individual can manifest the disease. The challenge of personalized medicine is to develop tools that can accurately predict the trajectory of an individual's disease. We represent an individual's disease trajectory as a continuous-valued continuous-time function describing the severity of the disease over time. We propose a hierarchical latent variable model that shares statistical strength across observations at different resolutions--the population, subpopulation and the individual level. We describe an algorithm for learning population and subpopulation parameters offline, and an online procedure for dynamically learning individual-specific parameters. Finally, we validate our model on the task of predicting the course of interstitial lung disease, one of the leading causes of death among patients with the autoimmune disease scleroderma. We compare our approach against strong baselines and demonstrate significant improvements in predictive accuracy.

22 Pointer Networks

Oriol Vinyals vinyals@google.com
Meire Fortunato meirefortunato@berkeley.edu
Navdeep Jaitly ndjaitly@google.com
Google

We introduce a new neural architecture to learn the conditional probability of an output sequence with elements that arediscrete tokens corresponding to positions in an input sequence. Such problems cannot be trivially addressed by existent approaches such as sequence-to-sequence and Neural Turing Machines,because the number of target classes in eachstep of the output depends on the length of the input, which is variable. Problems such as sorting variable sized sequences, and various combinatorialoptimization problems belong to this class. Our model solvesthe problem of variable size output dictionaries using a recently proposedmechanism of neural attention. It differs from the previous attentionattempts in that, instead of using attention to blend hidden units of anencoder to a context vector at each decoder step, it uses attention asa pointer to select a member of the input sequence as the output. We call this architecture a Pointer Net (Ptr-Net).We show Ptr-Nets can be used to learn approximate solutions to threechallenging geometric problems -- finding planar convex hulls, computingDelaunay triangulations, and the planar Travelling Salesman Problem-- using training examples alone. Ptr-Nets not only improve oversequence-to-sequence with input attention, butalso allow us to generalize to variable size output dictionaries.We show that the learnt models generalize beyond the maximum lengthsthey were trained on. We hope our results on these taskswill encourage a broader exploration of neural learning for discreteproblems.

23 A Reduced-Dimension fMRI Shared Response Model

Po-Hsuan (Cameron) Chen pohsuan@princeton.edu
Janice Chen janice@princeton.edu
Yaara Yeshurun yaara@princeton.edu
Uri Hasson hasson@princeton.edu
Peter J Ramadge ramadge@princeton.edu
Princeton University
James Haxby james.v.haxby@dartmouth.edu
Dartmouth

We develop a shared response model for aggregating multi-subject fMRI data that accounts for different functional topographies among anatomical aligned datasets. Multi-subject data is critical for evaluating the generality and validity of findings across subjects, and its effective utilization helps improve analysis sensitivity. Our model demonstrates improved sensitivity in identifying a shared response for a variety of datasets and anatomical brain regions of interest. Furthermore, by removing the identified shared response, it allows improved detection of group differences. The ability to identify what is shared and what is not shared, opens the model to a wide range of multi-subject fMRI studies.

24 Efficient Exact Gradient Update for training Deep Networks with Very Large Sparse Targets

Pascal Vincent vincentp@iro.umontreal.ca
Alexandre de Brébisson alexandre.de.brebisson@umontreal.ca
Xavier Bouthillier xavier.bouthillier@umontreal.ca
Universit de Montréal

An important class of problems involves training deep neural networks with sparse prediction targets of very high dimension D. These occur naturally in e.g. neural language models or the learning of word-embeddings, often posed as predicting the probability of next words among a vocabulary of size D (e.g. 200,000). Computing the equally large, but typically non-sparse D-dimensional output vector from a last hidden layer of reasonable dimension d (e.g. 500) incurs a prohibitive O(Dd) computational cost for each example, as does updating the D×d output weight matrix and computing the gradient needed for backpropagation to previous layers. While efficient handling of large sparse network inputs is trivial, this case of large sparse targets is not, and has thus so far been sidestepped with approximate alternatives such as hierarchical softmax or sampling-based approximations during training. In this work we develop an original algorithmic approach that, for a family of loss functions that includes squared error and spherical softmax, can compute the exact loss, gradient update for the output weights, and gradient for backpropagation, all in O(d2) per example instead of O(Dd), remarkably without ever computing the D-dimensional output. The proposed algorithm yields a speedup of D4d, i.e. two orders of magnitude for typical sizes, for that critical part of the computations that often dominates the training time in this kind of network architecture.

25 Precision-Recall-Gain Curves: PR Analysis Done Right

Peter Flach peter.flach@bristol.ac.uk
Meelis Kull meelis.kull@bristol.ac.uk
University of Bristol

Precision-Recall analysis abounds in applications of binary classification where true negatives do not add value and hence should not affect assessment of the classifier's performance. Perhaps inspired by the many advantages of receiver operating characteristic (ROC) curves and the area under such curves for accuracy-based performance assessment, many researchers have taken to report Precision-Recall (PR) curves and associated areas as performance metric. We demonstrate in this paper that this practice is fraught with difficulties, mainly because of incoherent scale assumptions -- e.g., the area under a PR curve takes the arithmetic mean of precision values whereas the Fβ score applies the harmonic mean. We show how to fix this by plotting PR curves in a different coordinate system, and demonstrate that the new Precision-Recall-Gain curves inherit all key advantages of ROC curves. In particular, the area under Precision-Recall-Gain curves conveys an expected F1 score, and the convex hull of a Precision-Recall-Gain curve allows us to calibrate the classifier's scores so as to determine, for each operating point on the convex hull, the interval of β values for which the point optimises Fβ. We demonstrate experimentally that the area under traditional PR curves can easily favour models with lower expected F1 score than others, and so the use of Precision-Recall-Gain curves will result in better model selection.

26 A Tractable Approximation to Optimal Point Process Filtering: Application to Neural Encoding

Yuval Harel yharel@tx.technion.ac.il
Ron Meir rmeir@ee.technion.ac.il
Technion
Manfred Opper opperm@cs.tu-berlin.de
TU Berlin

The process of dynamic state estimation (filtering) based on point process observations is in general intractable. Numerical sampling techniques are often practically useful, but lead to limited conceptual insight about optimal encoding/decoding strategies, which are of significant relevance to Computational Neuroscience. We develop an analytically tractable Bayesian approximation to optimal filtering based on point process observations, which allows us to introduce distributional assumptions about sensory cell properties, that greatly facilitates the analysis of optimal encoding in situations deviating from common assumptions of uniform coding. The analytic framework leads to insights which are difficult to obtain from numerical algorithms, and is consistent with experiments about the distribution of tuning curve centers. Interestingly, we find that the information gained from the absence of spikes may be crucial to performance.

27 Equilibrated adaptive learning rates for non-convex optimization

Yann Dauphin yann-nicolas.dauphin@umontreal.ca
Facebook AI Research
Harm de Vries mail@harmdevries.com
Yoshua Bengio yoshua.bengio@gmail.com
U. Montreal

Parameter-specific adaptive learning rate methods are computationally efficient ways to reduce the ill-conditioning problems encountered when training large deep networks. Following recent work that strongly suggests that most of thecritical points encountered when training such networks are saddle points, we find how considering the presence of negative eigenvalues of the Hessian could help us design better suited adaptive learning rate schemes. We show that the popular Jacobi preconditioner has undesirable behavior in the presence of both positive and negative curvature, and present theoretical and empirical evidence that the so-called equilibration preconditioner is comparatively better suited to non-convex problems. We introduce a novel adaptive learning rate scheme, called ESGD, based on the equilibration preconditioner. Our experiments demonstrate that both schemes yield very similar step directions but that ESGD sometimes surpasses RMSProp in terms of convergence speed, always clearly improving over plain stochastic gradient descent.

28 NEXT: A System for Real-World Development, Evaluation, and Application of Active Learning

Kevin G Jamieson kgjamieson@wisc.edu
Lalit Jain jain@math.wisc.edu
Chris Fernandez crfernandez@wisc.edu
Nicholas J. Glattard glattard@wisc.edu
Rob Nowak nowak@ece.wisc.edu
University of Wisconsin

Active learning methods automatically adapt data collection by selecting the most informative samples in order to accelerate machine learning. Because of this, real-world testing and comparing

active learning algorithms requires collecting new datasets (adaptively), rather than simply applying algorithms to benchmark datasets, as is the norm in (passive) machine learning research.To facilitate the development, testing and deployment of active learning for real applications, we have built an open-source software system for large-scale active learning research and experimentation. The system, called NEXT, provides a unique platform for real-world, reproducible active learning research. This paper details the challenges of building the system and demonstrates its capabilities with several experiments. The results show how experimentation can help expose strengths and weaknesses of active learning algorithms, in sometimes unexpected and enlightening ways.

29 Gaussian Process Random Fields

Dave Moore dmoore@cs.berkeley.edu
Stuart J Russell russell@cs.berkeley.edu
UC Berkeley

Gaussian processes have been successful in both supervised andunsupervised machine learning tasks, but their cubic complexity has constrained practical applications. We introduce a new approximation for large-scale Gaussian processes, the Gaussian Process Random Field (GPRF), in which local GPs are coupled via pairwise potentials. The GPRF likelihood is a simple, tractable, and parallelizeable approximation to the full GP marginal likelihood, enabling hyperparameter selection and latent variable modeling on large datasets.

30 MCMC for Variationally Sparse Gaussian Processes

James J Hensman james.hensman@gmail.com
The University of Sheffield
Alexander G Matthews am554@cam.ac.uk
Zoubin Ghahramani zoubin@eng.cam.ac.uk
University of Cambridge
Maurizio Filippone maurizio.filippone@eurecom.fr
EURECOM

Gaussian process (GP) models form a core part of probabilistic machine learning. Considerable research effort has been made into attacking three issues with GP models: how to compute efficiently when the number of data is large; how to approximate the posterior when the likelihood is not Gaussian and how to estimate covariance function parameter posteriors. This paper simultaneously addresses these, using a variational approximation to the posterior which is sparse in sup- port of the function but otherwise free-form. The result is a Hybrid Monte-Carlo sampling scheme which allows for a non-Gaussian approximation over the function values and covariance parameters simultaneously, with efficient computations based on inducing-point sparse GPs.

31 Streaming, Distributed Variational Inference for Bayesian Nonparametrics

Trevor Campbell tdjc@mit.edu
Julian Straub jstraub@csail.mit.edu
John W Fisher III fisher@csail.mit.edu
Jonathan P How jhow@mit.edu
MIT

This paper presents a methodology for creating streaming, distributed inference algorithms for Bayesian nonparametric (BNP) models. In the proposed framework, processing nodes receive a sequence of data minibatches, compute a variational posterior for each, and make asynchronous streaming updates to a central model. In contrast to previous algorithms, the proposed framework is truly streaming, distributed, asynchronous, learning-rate-free, and truncation-free. The key challenge in developing the framework, arising from fact that BNP models do not impose an inherent ordering on their components, is finding the correspondence between minibatch and central BNP posterior components before performing each update. To address this, the paper develops a combinatorial optimization problem over component correspondences, and provides an efficient solution technique. The paper concludes with an application of the methodology to the DP mixture model, with experimental results demonstrating its practical scalability and performance.

32 Fixed-Length Poisson MRF: Adding Dependencies to the Multinomial

David I Inouye dinouye@cs.utexas.edu
Pradeep K Ravikumar pradeepr@cs.utexas.edu
Inderjit S Dhillon inderjit@cs.utexas.edu
University of Texas at Austin

We propose a novel distribution that generalizes the Multinomial distribution to enable dependencies between the dimensions. Our novel distribution is based on the parametric form of the Poisson MRF model (Yang et al. 2012) but is fundamentally different because of the domain restriction to a fixed-length vector like in a Multinomial where the number of trials is fixed or known. Thus, we propose the Fixed-Length Poisson MRF (LPMRF) distribution. We develop methods to estimate the likelihood and log partition function (i.e. the log normalizing constant), which was not previously possible with the Poisson MRF model. In addition, we create mixture and topic models that use LPMRF as a base distribution and discuss the similarities and differences with previous topic models such as the recently proposed Admixture of Poisson MRFs (Inouye et al. 2014). Finally, we show the effectiveness of our LPMRF distribution over Multinomial models by evaluating the test set perplexity on a dataset of abstracts. Qualitatively, we show that the positive dependencies discovered by LPMRF are interesting and intuitive.

33 Human Memory Search as Initial-Visit Emitting Random Walk

Kwang-Sung Jun deltakam@cs.wisc.edu
Jerry Zhu jerryzhu@cs.wisc.edu
Timothy T Rogers ttrogers@wisc.edu
ming yuan myuan@stat.wisc.edu
University of Wisconsin - Madison
Zhuoran Yang yzr11@mails.tsinghua.edu.cn
Tsinghua University

Imagine a random walk that outputs a state only when visiting it for the first time. The observed output is therefore a repeat-censored version of the underlying walk, and consists of a permutation of the states or a prefix of it. We call this model initial-visit emitting random walk (INVITE). Prior work has shown that the random walks with such a repeat-censoring mechanism explain well human behavior in memory search tasks, which is of great interest in both the study of human cognition and various clinical applications. However, parameter estimation in INVITE is challenging, because naive likelihood computation by marginalizing over infinitely many hidden random walk trajectories is intractable. In this paper, we propose the first efficient maximum likelihood estimate (MLE) for INVITE by decomposing the censored output into a series of absorbing random walks. We also prove theoretical properties of the MLE including identifiability and consistency. We show that INVITE outperforms several existing methods on real-world human response data from memory search tasks.

34 Structured Transforms for Small-Footprint Deep Learning

Vikas Sindhwani vikas.sindhwani@gmail.com
Tara Sainath tsainath@google.com
Sanjiv Kumar sanjivk@google.com
Google

We consider the task of building compact deep learning pipelines suitable for deployment on storage and power constrained mobile devices. We propose a unified framework to learn a broad family of structured parameter matrices that are characterized by the notion of low displacement rank. Our structured transforms admit fast function and gradient evaluation, and span a rich range of parameter sharing configurations whose statistical modeling capacity can be explicitly tuned along a continuum from structured to unstructured. Experimental results show that these transforms can dramatically accelerate inference and forward/backward passes during training, and offer superior accuracy-compactness-speed tradeoffs in comparison to a number of existing techniques. In keyword spotting applications in mobile speech recognition, our methods are much more effective than standard linear low-rank bottleneck layers and nearly retain the performance of state of the art models, while providing more than 3.5-fold compression.

35 Spectral Learning of Large Structured HMMs for Comparative Epigenomics

Chicheng Zhang chz038@cs.ucsd.edu
Kamalika Chaudhuri kamalika@cs.ucsd.edu
UC San Diego
Jimin Song song@dls.rutgers.edu
Kevin Chen kcchen@dls.rutgers.edu
Rutgers

We develop a latent variable model and an efficient spectral algorithm motivated by the recent emergence of very large data sets of chromatin marks from multiple human cell types. A natural model for chromatin data in one cell type is a Hidden Markov Model (HMM); we model the relationship between multiple cell types by connecting their hidden states by a fixed tree of known structure. The main challenge with learning parameters of such models is that iterative methods such as EM are very slow, while naive spectral methods result in time and space complexity exponential in the number of cell types. We exploit properties of the tree structure of the hidden states to provide spectral algorithms that are more computationally efficient for current biological datasets. We provide sample complexity bounds for our algorithm and evaluate it experimentally on biological data from nine human cell types. Finally, we show that beyond our specific model, some of our algorithmic ideas can be applied to other graphical models.

36 A Structural Smoothing Framework For Robust Graph Comparison

Pinar Yanardag ypinar@purdue.edu
Purdue
S.V.N. Vishwanathan vishy@ucsc.edu
UCSC

In this paper, we propose a general smoothing framework for graph kernels that takes \textit{structural similarity} into account, and apply it to derive smoothed versions of three popular graph kernels namely graphlet kernels, Weisfeiler-Lehman subtree kernels, and shortest-path kernels. Our framework is inspired by state-of-the-art smoothing techniques used in natural language processing (NLP). However, unlike NLP applications which primarily deal with strings, we show how one can apply smoothing to a richer class of inter-dependent sub-structures which naturally arise in graphs. Moreover, we discuss extensions of the Pitman-Yor process that can be adapted to smooth subgraph distributions thereby leading to novel graph kernels. Our kernels are able to tackle the diagonal dominance problem, while respecting the structural similarity between sub-structures in graph kernels, especially under the presence of edge or label noise. Experimental evaluation shows that not only do our kernels outperform the unsmoothed variants, but also achieve statistically significant improvements in classification accuracy over several other graph kernels that have been recently proposed in literature. Our kernels are competitive in terms of runtime, and offer a viable option for practitioners.

37 Optimization Monte Carlo: Efficient and Embarrassingly Parallel Likelihood-Free Inference

Ted Meeds tmeeds@gmail.com
Max Welling welling.max@gmail.com
University of Amsterdam

We describe an embarrassingly parallel, anytime Monte Carlo method for likelihood-free models. The algorithm starts with the view that the stochasticity of the pseudo-samples generated by the simulator can be controlled externally by a vector of random numbers u, in such a way that the outcome, knowing u, is deterministic. For each instantiation of u we run an optimization procedure to minimize the distance between summary statistics of the simulator and the data. After reweighing these samples using the prior and the Jacobian (accounting for the change of volume in transforming from the space of summary statistics to the space of parameters) we show that this weighted ensemble represents a Monte Carlo estimate of the posterior distribution. The procedure can be run embarrassingly parallel (each node handling one sample) and anytime (by allocating resources to the worst performing sample). The procedure is validated on six experiments.

38 Inverse Reinforcement Learning with Locally Consistent Reward Functions

Quoc Phong Nguyen qphong@comp.nus.edu.sg
Bryan Kian Hsiang Low lowkh@comp.nus.edu.sg
National University of Singapore
Patrick Jaillet jaillet@mit.edu
Massachusetts Institute of Technology

Existing inverse reinforcement learning (IRL) algorithms have assumed each expert's demonstrated trajectory to be produced by only a single reward function. This paper presents a novel generalization of the IRL problem that allows each trajectory to be generated by multiple locally consistent reward functions, hence catering to more realistic and complex experts' behaviors. Solving our generalized IRL problem thus involves not only learning these reward functions but also the stochastic transitions between them at any state (i.e., including unvisited states). By representing our IRL problem with a probabilistic graphical model, an expectation-maximization (EM) algorithm can be devised to iteratively learn the reward functions and stochastic transitions between them that jointly improve the likelihood of the expert's demonstrated trajectories. As a result, the most likely partition of a trajectory into segments that are generated from different locally consistent reward functions selected by EM can be derived. Empirical evaluation on synthetic and real-world datasets shows that our IRL algorithm outperforms the state-of-the-art EM clustering with maximum likelihood IRL, which is, interestingly, a reduced variant of our approach.

39 Consistent Multilabel Classification

Sanmi Koyejo sanmi@stanford.edu
Stanford University
Nagarajan Natarajan naga86@cs.utexas.edu
Pradeep K Ravikumar pradeepr@cs.utexas.edu
Inderjit S Dhillon inderjit@cs.utexas.edu
University of Texas at Austin

Multilabel classification is becoming increasingly relevant due to modern applications such as image tagging and text document categorization. While there is now a rich understanding of learning with respect to general performance metrics in the traditional settings of binary and multiclass classification, less is known for multilabel classification beyond a few special cases. In this paper, we take a step towards establishing results at the level of generality currently enjoyed only in the traditional settings. In particular, we propose a framework for defining multilabel performance metrics on a finite sample as well as the corresponding population utility with respect to the underlying data distribution. We then provide a simple characterization of the optimal classifier for a large family of multilabel performance metrics. We also propose a consistent plug-in estimation algorithm that is efficient as well as theoretically consistent with respect to the underlying multilabel metric. Results on synthetic and benchmark datasets are supportive of our theoretical findings.

40 Is Approval Voting Optimal Given Approval Votes?

Ariel D Procaccia arielpro@cs.cmu.edu
Nisarg Shah nkshah@cs.cmu.edu
Carnegie Mellon University

Some crowdsourcing platforms ask workers to express their opinions by approving a set of k good alternatives. It seems that the only reasonable way to aggregate these k-approval votes is the approval voting rule, which simply counts the number of times each alternative was approved. We challenge this assertion by proposing a probabilistic framework of noisy voting, and asking whether approval voting yields an alternative that is most likely to be the best alternative, given k-approval votes. While the answer is generally positive, our theoretical and empirical results call attention to situations where approval voting is suboptimal.

41 A Normative Theory of Adaptive Dimensionality Reduction in Neural Networks

Cengiz Pehlevan cpehlevan@simonsfoundation.org
Dmitri Chklovskii dchklovskii@simonsfoundation.org
Simons Foundation

To make sense of the world our brains must analyze high-dimensional datasets streamed by our sensory organs. Because such analysis begins with dimensionality reduction, modelling early sensory processing requires biologically plausible online dimensionality reduction algorithms. Recently, we derived such an algorithm, termed similarity matching, from a Multidimensional Scaling (MDS) objective function. However, in the existing algorithm, the number of output dimensions is set a priori by the number of output neurons and cannot be changed. Because the number of informative dimensions in sensory inputs is variable there is a need for adaptive dimensionality reduction. Here, we derive biologically plausible dimensionality reduction algorithms which adapt the number of output dimensions to the eigenspectrum of the input covariance matrix. We formulate three objective functions which, in the offline setting, are optimized by the projections of the input dataset onto its principal subspace scaled by the eigenvalues of the output covariance matrix. In turn, the output eigenvalues are computed as i) soft-thresholded, ii) hard-thresholded, iii) equalized thresholded eigenvalues of the input covariance matrix. In the online setting, we derive the three corresponding adaptive algorithms and map them onto the dynamics of neuronal activity in networks with biologically plausible local learning rules. Remarkably, in the last two networks, neurons are divided into two classes which we identify with principal neurons and interneurons in biological circuits.

42 Efficient Non-greedy Optimization of Decision Trees

Mohammad Norouzi norouzi@cs.toronto.edu
David J Fleet f leet@cs.toronto.edu
University of Toronto
Maxwell Collins mcollins@cs.wisc.edu
UW-Madison
Matthew A Johnson matjoh@microsoft.com
Pushmeet Kohli pkohli@microsoft.com
Microsoft Research

Decision trees and randomized forests are widely used in computer vision and machine learning. Standard algorithms for decision tree induction select the split functions one node at a time according to some splitting criteria. This greedy procedure may lead to suboptimal trees. In this paper, we present an algorithm for optimizing the split functions at all levels of the tree, jointly with the leaf parameters, based on a global objective. We show that the problem of finding optimal linear combination splits for decision trees is an instance of structured prediction with latent variables, and we formulate a convex-concave upper bound on the tree's empirical loss. Computing the gradient of the proposed surrogate objective with respect to each exemplar is O(d2) (where d is the tree depth) and thus training deep trees is feasible. The use of stochastic gradient descent for optimization enables effective training with large datasets. Experiments on several classification benchmarks demonstrate that our non-greedy decision trees outperform standard greedy axis-aligned trees.

43 Statistical Topological Data Analysis - A Kernel Perspective

Roland Kwitt rkwitt@gmx.at
University of Salzburg
Stefan Huber shuber@sthu.org
IST Austria
Marc Niethammer mn@cs.unc.edu
Weili Lin weili_lin@med.unc.edu
UNC Chapel Hill
Ulrich Bauer mail@ulrich-bauer.org
TU Munich

We consider the problem of statistical computations with persistence diagrams, a summary representation of topological features in data. These diagrams encode persistent homology, a widely used invariant in topological data analysis. While several avenues towards a statistical treatment of the diagrams have been explored recently, we follow an alternative route that is motivated by the success of methods based on the embedding of probability measures into reproducing kernel Hilbert spaces. In fact, a positive definite kernel on persistence diagrams has recently been proposed, connecting persistent homology to popular kernel-based learning techniques such as support vector machines. However, important properties of that kernel which would enable a principled use in the context of probability measure embeddings remain to be explored. Our contribution is to close this gap by proving universality of a variant of the original kernel, and to demonstrate its effective use in two-sample hypothesis testing on synthetic as well as real-world data.

44 Variational Consensus Monte Carlo

Maxim Rabinovich rabinovich@eecs.berkeley.edu
Michael I Jordan jordan@cs.berkeley.edu
UC Berkeley
Elaine Angelino elaine@eecs.berkeley.edu
Harvard

Practitioners of Bayesian statistics have long depended on Markov chain Monte Carlo (MCMC) to obtain samples from intractable posterior distributions. Unfortunately, MCMC algorithms are typically serial, and do not scale to the large datasets typical of modern machine learning. The recently proposed consensus Monte Carlo algorithm removes this limitation by partitioning the data and drawing samples conditional on each partition in parallel (Scott et al, 2013). A fixed aggregation function then combines these samples, yielding approximate posterior samples. We introduce variational consensus Monte Carlo (VCMC), a variational Bayes algorithm that optimizes over aggregation functions to obtain samples from a distribution that better approximates the target. The resulting objective contains an intractable entropy term; we therefore derive a relaxation of the objective and show that the relaxed problem is blockwise concave under mild conditions. We illustrate the advantages of our algorithm on three inference tasks from the literature, demonstrating both the superior quality of the posterior approximation and the moderate overhead of the optimization step. Our algorithm achieves a relative error reduction (measured against serial MCMC) of up to 39% compared to consensus Monte Carlo on the task of estimating 300-dimensional probit regression parameter expectations; similarly, it achieves an error reduction of 92% on the task of estimating cluster comembership probabilities in a Gaussian mixture model with 8 components in 8 dimensions. Furthermore, these gains come at moderate cost compared to the runtime of serial MCMC, achieving near-ideal speedup in some instances.

45 Softstar: Heuristic-Guided Probabilistic Inference

Mathew Monfort mmonfo2@uic.edu
University of Illinois at Chicago
Brenden M Lake brenden@mit.edu
Josh Tenenbaum jbt@mit.edu
MIT
Brian Ziebart bziebart@uic.edu
University of Illinois at Chicago
Patrick Lucey patrick.lucey@disneyresearch.com
Disney Research Pittsburgh

Recent machine learning methods for sequential behavior prediction estimate the motives of behavior rather than the behavior itself. This higher-level abstraction improves generalization in different prediction settings, but computing predictions often becomes intractable in large decision spaces. We propose the Softstar algorithm, a softened heuristic-guided search technique for the maximum entropy inverse optimal control model of sequential behavior. This approach supports probabilistic search with bounded approximation error at a significantly reduced computational cost when compared to sampling based methods. We present the algorithm, analyze approximation guarantees, and compare performance with simulation-based inference on two distinct complex decision tasks.

46 Gradient-free Hamiltonian Monte Carlo with Efficient Kernel Exponential Families

Heiko Strathmann heiko.strathmann@gmail.com
Samuel Livingstone samuel.livingstone@ucl.ac.uk
Zoltan Szabo zoltan.szabo@gatsby.ucl.ac.uk
Arthur Gretton arthur.gretton@gmail.com
University College London
Dino Sejdinovic dino.sejdinovic@gmail.com
University of Oxford

We propose Kernel Hamiltonian Monte Carlo (KMC), a gradient-free adaptive MCMC algorithm based on Hamiltonian Monte Carlo (HMC). On target densities where HMC is unavailable due to intractable gradients, KMC adaptively learns the target's gradient structure by fitting an exponential family model in a Reproducing Kernel Hilbert Space. Computational costs are reduced by two novel efficient approximations to this gradient. While being asymptotically exact, KMC mimics HMC in terms of sampling efficiency and offers substantial mixing improvements to state-of-the-art gradient free samplers. We support our claims with experimental studies on both toy and real-world applications, including Approximate Bayesian Computation and exact-approximate MCMC.

47 A Complete Recipe for Stochastic Gradient MCMC

Yi-An Ma yianma@u.washington.edu
Tianqi Chen tqchen@cs.washington.edu
Emily Fox ebfox@uw.edu
University of Washington

Many recent Markov chain Monte Carlo (MCMC) samplers leverage stochastic dynamics with state adaptation to define a Markov transition kernel that efficiently explores a target distribution. In tandem, a focus has been on devising scalable MCMC algorithms via data subsampling and using stochastic gradients in the stochastic dynamic simulations. However, such stochastic gradient MCMC methods have used simple stochastic dynamics, or required significant physical intuition to modify the dynamical system to account for the stochastic gradient noise. In this paper, we provide a general recipe for constructing MCMC samplers---including stochastic gradient versions---based on continuous Markov processes specified via two matrices. We constructively prove that the framework is complete. That is, any continuous Markov process that provides samples from the target distribution can be written in our framework. We demonstrate the utility of our recipe by trivially "reinventing" previously proposed stochastic gradient MCMC samplers, and in proposing a new state-adaptive sampler: stochastic gradient Riemann Hamiltonian Monte Carlo (SGRHMC). Our experiments on simulated data and a streaming Wikipedia analysis demonstrate that the proposed sampler inherits the benefits of Riemann HMC, with the scalability of stochastic gradient methods.

48 Barrier Frank-Wolfe for Marginal Inference

Rahul G Krishnan rahul@cs.nyu.edu
David Sontag dsontag@cs.nyu.edu
New York University
Simon Lacoste-Julien simon.lacoste-julien@ens.fr
INRIA

We introduce a globally-convergent algorithm for optimizing the tree-reweighted (TRW) variational objective over the marginal polytope. The algorithm is based on the conditional gradient method (Frank-Wolfe) and moves pseudomarginals within the marginal polytope through repeated maximum a posteriori (MAP) calls. This modular structure enables us to leverage black-box MAP solvers (both exact and approximate) for variational inference, and obtains more accurate results than tree-reweighted algorithms that optimize over the local consistency relaxation. Theoretically, we bound the sub-optimality for the proposed algorithm despite the TRW objective having unbounded gradients at the boundary of the marginal polytope. Empirically, we demonstrate the increased quality of results found by tightening the relaxation over the marginal polytope as well as the spanning tree polytope on synthetic and real-world instances.

49 Practical and Optimal LSH for Angular Distance

Alexandr Andoni andoni@mit.edu
Columbia
Piotr Indyk indyk@mit.edu
Ilya Razenshteyn ilyaraz@mit.edu
Ludwig Schmidt ludwigs@mit.edu
MIT
Thijs Laarhoven mail@thijs.com
TU/e

We show the existence of a Locality-Sensitive Hashing (LSH) family for the angular distance that yields an approximate Near Neighbor Search algorithm with the asymptotically optimal running time exponent. Unlike earlier algorithms with this property (e.g., Spherical LSH (Andoni-Indyk-Nguyen-Razenshteyn 2014) (Andoni-Razenshteyn 2015)), our algorithm is also practical, improving upon the well-studied hyperplane LSH (Charikar 2002) in practice. We also introduce a multiprobe version of this algorithm, and conduct experimental evaluation on real and synthetic data sets.We complement the above positive results with a fine-grained lower bound for the quality of any LSH family for angular distance. Our lower bound implies that the above LSH family exhibits a trade-off between evaluation time and quality that is close to optimal for a natural class of LSH functions.

50 Principal Differences Analysis: Interpretable Characterization of Differences between Distributions

Jonas W Mueller jonasmueller@csail.mit.edu
Tommi Jaakkola tommi@csail.mit.edu
MIT

We introduce principal differences analysis for analyzing differences between high-dimensional distributions. The method operates by finding the projection that maximizes the Wasserstein divergence between the resulting univariate populations. Relying on the Cramer-Wold device, it requires no assumptions about the form of the underlying distributions, nor the nature of their inter-class differences. A sparse variant of the method is introduced to identify features responsible for the differences. We provide algorithms for both the original minimax formulation as well as its semidefinite relaxation. In addition to deriving some convergence results, we illustrate how the approach may be applied to identify differences between cell populations in the somatosensory cortex and hippocampus as manifested by single cell RNA-seq. Our broader framework extends beyond the specific choice of Wasserstein divergence.

51 Kullback-Leibler Proximal Variational Inference

Emtiyaz E Khan emtiyaz@gmail.com
Pierre Baque pierre.baque@epfl.ch
Pascal Fua pascal.fua@epfl.ch
EPFL
François Fleuret francois.fleuret@idiap.ch
Idiap Research Institute

We propose a new variational inference method based on the Kullback-Leibler (KL) proximal term. We make two contributions towards improving efficiency of variational inference. Firstly, we derive a KL proximal-point algorithm and show its equivalence to gradient descent with natural gradient in stochastic variational inference. Secondly, we use the proximal framework to derive efficient variational algorithms for non-conjugate models. We propose a splitting procedure to separate non-conjugate terms from conjugate ones. We then linearize the non-conjugate terms and show that the resulting subproblem admits a closed-form solution. Overall, our approach converts a non-conjugate model to subproblems that involve inference in well-known conjugate models. We apply our method to many models and derive generalizations for non-conjugate exponential family. Applications to real-world datasets show that our proposed algorithms are easy to implement, fast to converge, perform well, and reduce computations.

52 Learning Large-Scale Poisson DAG Models based on OverDispersion Scoring

Gunwoong Park parkg@stat.wisc.edu
Garvesh Raskutti raskutti@cs.wisc.edu
University of Wisconsin, Madison

In this paper, we address the question of identifiability and learning algorithms for large-scale Poisson Directed Acyclic Graphical (DAG) models. We define general Poisson DAG models as models where each node is a Poisson random variable with rate parameter depending on the values of the parents in the underlying DAG. First, we prove that Poisson DAG models are identifiable from observational data, and present a polynomial-time algorithm that learns the Poisson DAG model under suitable regularity conditions. The main idea behind our algorithm is based on overdispersion, in that variables that are conditionally Poisson are overdispersed relative to variables that are marginally Poisson. Exploiting overdispersion allows us to learn the causal ordering and then use ideas from learning large-scale regression models to reduce computational complexity. We provide both theoretical guarantees and simulation results for both small and large-scale DAGs to validate the success of our algorithm.

53 Streaming Min-max Hypergraph Partitioning

Dan Alistarh dan.alistarh@microsoft.com
Milan Vojnovic milanv@microsoft.com
Microsoft Research
Jennifer Iglesias jiglesia@andrew.cmu.edu
Carnegie Mellon University

In many applications, the data is of rich structure that can be represented by a hypergraph, where the data items are represented by vertices and the associations among items are represented by hyperedges. Equivalently, we are given an input bipartite graph with two types of vertices: items, and associations (which we refer to as topics). We consider the problem of partitioning the set of items into a given number of parts such that the maximum number of topics covered by a part of the partition is minimized. This is a natural clustering problem, with various applications, e.g. partitioning of a set of information objects such as documents, images, and videos, and load balancing in the context of computation platforms.In this paper, we focus on the streaming computation model for this problem, in which items arrive online one at a time and each item must be assigned irrevocably to a part of the partition at its arrival time. Motivated by scalability requirements, we focus on the class of streaming computation algorithms with memory limited to be at most linear in the number of the parts of the partition. We show that a greedy assignment strategy is able to recover a hidden co-clustering of items under a natural set of recovery conditions. We also report results of an extensive empirical evaluation, which demonstrate that this greedy strategy yields superior performance when compared with alternative approaches.

54 Efficient Output Kernel Learning for Multiple Tasks

Pratik Jawanpuria pratik.iitb@gmail.com
Matthias Hein hein@cs.uni-sb.de
Saarland University
Maksim Lapin mlapin@mpi-inf.mpg.de
Bernt Schiele schiele@mpi-inf.mpg.de
Max Planck Institute for Informatics

The paradigm of multi-task learning is that one can achieve better generalization by learning tasks jointly and thus exploiting the similarity between the tasks rather than learning them independently of each other. While previously, the relationship between tasks had to be user-defined, in the form of an output kernel, recent approaches jointly learn the tasks and the output kernel. As the output kernel is a positive semidefinite matrix, the resulting optimization problems are not scalable in the number of tasks as an eigendecomposition is required in each step. Using the theory of positive semidefinite kernels we show in this paper that for a certain class of regularizers on the output kernel, the constraint of being positive semidefinite can be dropped as it is automatically satisfied for the relaxed problem. This leads to an unconstrained dual problem which can be solved efficiently. Experiments on several multi-task and multi-class data sets illustrates the efficacy of our approach in terms of computational efficiency as well as generalization performance.

55 Gradient Estimation Using Stochastic Computation Graphs

John Schulman john.d.schulman@gmail.com
UC Berkeley / Google
Nicolas Heess heess@google.com
Theophane Weber theophane@google.com
Google DeepMind
Pieter Abbeel pabbeel@cs.berkeley.edu
UC Berkeley

In a variety of problems originating in supervised, unsupervised, and reinforcement learning, the loss function is defined by an expectation over a collection of random variables, which might be part of a probabilistic model or the external world. Estimating the gradient of this loss function, using samples, lies at the core of gradient-based learning algorithms for these problems. We introduce the formalism of stochastic computation graphs--directed acyclic graphs that include both deterministic functions and conditional probability distributions and describe how to easily and automatically derive an unbiased estimator of the loss function's gradient. The resulting algorithm for computing the gradient estimator is a simple modification of the standard backpropagation algorithm. The generic scheme we propose unifies estimators derived in variety of prior work, along with variance-reduction techniques therein. It could assist researchers in developing intricate models involving a combination of stochastic and deterministic operations, enabling, for example, attention, memory, and control actions.

56 Lifted Inference Rules With Constraints

Happy Mittal happy.mittal@cse.iitd.ac.in
Anuj Mahajan anujmahajan.iitd@gmail.com
Parag Singla parags@cse.iitd.ac.in
Indian Institute of Technology
Vibhav G Gogate vgogate@hlt.utdallas.edu
UT Dallas

Lifted inference rules exploit symmetries for fast reasoning in statistical rela-tional models. Computational complexity of these rules is highly dependent onthe choice of the constraint language they operate on and therefore coming upwith the right kind of representation is critical to the success and wider adoptionof lifted inference methods. In this paper, we propose a new constraint language,called setineq, which allows subset, equality and inequality constraints, to repre-sent substitutions over the variables in the theory. Our constraint formulation isstrictly more expressive than existing representations, yet easy to operate on. Wereformulate the three main lifting rules: decomposer, generalized binomial andthe recently proposed single occurrence for MAP inference, to work with our con-straint representation. Experiments on benchmark MLNs for exact and samplingbased inference demonstrate the effectiveness of our approach over several otherexisting techniques.

57 Sparse PCA via Bipartite Matchings

Megasthenis Asteris megas@utexas.edu
Tasos Kyrillidis anastasios@utexas.edu
Alex G Dimakis dimakis@austin.utexas.edu
University of Texas at Austin
Dimitris Papailiopoulos dimitrisp@berkeley.edu
UC Berkeley

We consider the following multi-component sparse PCA problem:given a set of data points, we seek to extract a small number of sparse components with \emph{disjoint} supports that jointly capture the maximum possible variance.Such components can be computed one by one, repeatedly solving the single-component problem and deflating the input data matrix, but this greedy procedure is suboptimal.We present a novel algorithm for sparse PCA that jointly optimizes multiple disjoint components. The extracted features capture variance that lies within a multiplicative factor arbitrarily close to 1 from the optimal.Our algorithm is combinatorial and computes the desired components by solving multiple instances of the bipartite maximum weight matching problem.Its complexity grows as a low order polynomial in the ambient dimension of the input data, but exponentially in its rank.However, it can be effectively applied on a low-dimensional sketch of the input data.We evaluate our algorithm on real datasets and empirically demonstrate that in many cases it outperforms existing, deflation-based approaches.

58 Empirical Localization of Homogeneous Divergences on Discrete Sample Spaces

Takashi Takenouchi ttakashi@fun.ac.jp
Future University Hakodate
Takafumi Kanamori kanamori@is.nagoya-u.ac.jp
Nagoya University

In this paper, we propose a novel parameter estimator for probabilistic models on discrete space. The proposed estimator is derived from minimization of homogeneous divergence and can be constructed without calculation of the normalization constant, which is frequently infeasible for models in the discrete space. We investigate statistical properties of the proposed estimator such as consistency and asymptotic normality, and reveal a relationship with the alpha-divergence. Small experiments show that the proposed estimator attains comparable performance to the MLE with drastically lower computational cost.

59 Weighted Theta Functions and Embeddings with Applications to Max-Cut, Clustering and Summarization

Fredrik D Johansson frejohk@chalmers.se
Ankani Chattoraj ankanichattoraj91@gmail.com
Devdatt Dubhashi dubhashi@chalmers.se
Chalmers University, Sweden
Chiranjib Bhattacharyya chiru@csa.iisc.ernet.in
Indian Institute of Science

We introduce a unifying generalization of the Lovász theta function, and the associated geometric embedding, for graphs with weights on both nodes and edges. We show how it can be computed exactly by semidefinite programming, and how to approximate it using SVM computations. We show how the theta function can be interpreted as a measure of diversity in graphs and use this idea, and the graph embedding in algorithms for Max-Cut, correlation clustering and document summarization, all of which are well represented as problems on weighted graphs.

60 Online Rank Elicitation for Plackett-Luce: A Dueling Bandits Approach

Balázs Szörényi szorenyi@inf.u-szeged.hu
The Technion / University of Szeged
Róbert Busa-Fekete busarobi@gmail.com
Adil Paul adil.paul@upb.de
UPB
Eyke Hüllermeier eyke@upb.de
Marburguniversity

We study the problem of online rank elicitation, assuming that rankings of a set of alternatives obey the Plackett-Luce distribution. Following the setting of the dueling bandits problem, the learner is allowed to query pairwise comparisons between alternatives, i.e., to sample pairwise marginals of the distribution in an online fashion. Using this information, the learner seeks to reliably predict the most probable ranking (or top-alternative). Our approach is based on constructing a surrogate probability distribution over rankings based on a sorting procedure, for which the pairwise marginals provably coincide with the marginals of the Plackett-Luce distribution. In addition to a formal performance and complexity analysis, we present first experimental studies.

61 Segregated Graphs and Marginals of Chain Graph Models

Ilya Shpitser ilyas@cs.jhu.edu
Johns Hopkins University

Bayesian networks are a popular representation of asymmetric (for example causal) relationships between random variables. Markov random fields (MRFs) are a complementary model of symmetric relationships used in computer vision, spatial modeling, and social and gene expression networks. A chain graph model under the Lauritzen-Wermuth-Frydenberg interpretation (hereafter a chain graph model) generalizes both Bayesian networks and MRFs, and can represent asymmetric and symmetric relationships together. As in other graphical models, the set of marginals from distributions in a chain graph model induced by the presence of hidden variables forms a complex model. One recent approach to the study of marginal graphical models is to consider a well-behaved supermodel. Such a supermodel of marginals of Bayesian networks, defined only by conditional independences, and termed the ordinary Markov model, was studied at length in (Evans and Richardson, 2014).In this paper, we show that special mixed graphs which we call segregated graphs can be associated, via a Markov property, with supermodels of a marginal of chain graphs defined only by conditional independences. Special features of segregated graphs imply the existence of a very natural factorization for these supermodels, and imply many existing results on the chain graph model, and ordinary Markov model carry over. Our results suggest that segregated graphs define an analogue of the ordinary Markov model for marginals of chain graph models.

62 Approximating Sparse PCA from Incomplete Data

ABHISEK KUNDU abhisekkundu@gmail.com
Petros Drineas drinep@cs.rpi.edu
Malik Magdon-Ismail magdon@cs.rpi.edu
Rensselaer Polytechnic Institute

We study how well one can recover sparse principal componentsof a data matrix using a sketch formed from a few of its elements. We show that for a wide class of optimization problems,if the sketch is close (in the spectral norm) to the original datamatrix, then one can recover a near optimal solution to the optimizationproblem by using the sketch. In particular, we use this approach toobtain sparse principal components and show that for data points in {n} dimensions, {O(e-2 ~k\max{m,n}) elements gives an e-additive approximation to the sparse PCA problem(k~} is the stable rank of the data matrix).We demonstrate our algorithms extensivelyon image, text, biological and financial data. The results show that not only are we able to recover the sparse PCAs from the incomplete data, but by using our sparse sketch, the running timedrops by a factor of five or more.

63 Multi-Layer Feature Reduction for Tree Structured Group Lasso via Hierarchical Projection

Jie Wang jwangumi@umich.edu
Jieping Ye jpye@umich.edu
University of Michigan

Tree structured group Lasso (TGL) is a powerful technique in uncovering the tree structured sparsity over the features, where each node encodes a group of features. It has been applied successfully in many real-world applications. However, with extremely large feature dimensions, solving TGL remains a significant challenge due to its highly complicated regularizer. In this paper, we propose a novel Multi-Layer Feature reduction method (MLFre) to quickly identify the inactive nodes (the groups of features with zero coefficients in the solution) hierarchically in a top-down fashion, which are guaranteed to be irrelevant to the response. Thus, we can remove the detected nodes from the optimization without sacrificing accuracy. The major challenge in developing such testing rules is due to the overlaps between the parents and their children nodes. By a novel hierarchical projection algorithm, MLFre is able to test the nodes independently from any of their ancestor nodes. Moreover, we can integrate MLFre---that has a low computational cost---with any existing solvers. Experiments on both synthetic and real data sets demonstrate that the speed-up gained by MLFre can be orders of magnitude.

64 Recovering Communities in the General Stochastic Block Model Without Knowing the Parameters

Emmanuel Abbe eabbe@princeton.edu
Colin Sandon sandon@princeton.edu
Princeton University

The stochastic block model (SBM) has recently generated significant research activity due to new threshold phenomena. However, most developments rely on the knowledge of the model parameters, or at least on the number of communities. This paper introduces efficient algorithms that do not require such knowledge and yet achieve the optimal information-theoretic tradeoffs identified in [AS15]. In the constant degree regime, an algorithm

is developed that requires only a lower-bound on the relative sizes of the communities and achieves the optimal accuracy scaling for large degrees. In the logarithmic degree regime, this is further enhanced into a fully agnostic algorithm that simultaneously learns the model parameters, achieves the optimal CH-limit, and runs in quasi-linear time. These provide the first algorithms affording efficiency, universality and information-theoretic optimality for strong and weak consistency in the general SBM with linear size communities.

65 Maximum Likelihood Learning With Arbitrary Treewidth via Fast-Mixing Parameter Sets

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

Inference is typically intractable in high-treewidth undirected graphical models, making maximum likelihood learning a challenge. One way to overcome this is to restrict parameters to a tractable set, most typically the set of tree-structured parameters. This paper explores an alternative notion of a tractable set, namely a set of “fast-mixing parameters” where Markov chain Monte Carlo (MCMC) inference can be guaranteed to quickly converge to the stationary distribution. While it is common in practice to approximate the likelihood gradient using samples obtained from MCMC, such procedures lack theoretical guarantees. This paper proves that for any exponential family with bounded sufficient statistics, (not just graphical models) when parameters are constrained to a fast-mixing set, gradient descent with gradients approximated by sampling will approximate the maximum likelihood solution inside the set with high-probability. When unregularized, to find a solution epsilon-accurate in log-likelihood requires a total amount of effort cubic in 1/epsilon, disregarding logarithmic factors. When ridge-regularized, strong convexity allows a solution epsilon-accurate in parameter distance with an effort quadratic in 1/epsilon. Both of these provide of a fully-polynomial time randomized approximation scheme.

66 Testing Closeness With Unequal Sized Samples

Bhaswar Bhattacharya bhaswar@stanford.edu
Greg Valiant valiant@stanford.edu
Stanford University

We consider the problem of testing whether two unequal-sized samples were drawn from identical distributions, versus distributions that differ significantly. Specifically, given a target error parameter \eps>0, m1 independent draws from an unknown distribution p with discrete support, and m2 draws from an unknown distribution q of discrete support, we describe a test for distinguishing the case that p=q from the case that ||p−q||1≥\eps. If p and q are supported on at most n elements, then our test is successful with high probability provided m1≥n2/3/ε4/3 and m2=Ω(max{nm√1ε2,n√ε2}). We show that this tradeoff is information theoretically optimal throughout this range, in the dependencies on all parameters, n,m1, and \eps, to constant factors. As a consequence, we obtain an algorithm for estimating the mixing time of a Markov chain on n states up to a logn factor that uses O~(n3/2τmix) queries to a ``next node'' oracle. The core of our testing algorithm is a relatively simple statistic that seems to perform well in practice, both on synthetic data and on natural language data. We believe that this statistic might prove to be a useful primitive within larger machine learning and natural language processing systems.

67 Learning Causal Graphs with Small Interventions

Karthikeyan Shanmugam karthiksh@utexas.edu
Murat Kocaoglu mkocaoglu@utexas.edu
Alex G Dimakis dimakis@austin.utexas.edu
Sriram Vishwanath sriram@ece.utexas.edu
UT Austin

We consider the problem of learning causal networks with interventions, when each intervention is limited in size. The objective is to minimize the number of experiments to discover the causal directions of all the edges in a causal graph. Previous work has focused on the use of separating systems for complete graphs for this task. We prove that any deterministic adaptive algorithm needs to be a separating system in order to learn complete graphs in the worst case. In addition, we present a novel separating system construction, whose size is close to optimal and is arguably simpler than previous work in combinatorics. We also develop a novel information theoretic lower bound on the number of interventions that applies in full generality, including for randomized adaptive learning algorithms. For general chordal graphs, we derive worst case lower bounds on the number of interventions. Building on observations about induced trees, we give a new deterministic adaptive algorithm to learn directions on any chordal skeleton completely. In the worst case, our achievable scheme is an alpha-approximation algorithm where alpha is the independence number of the graph. We also show that there exist graph classes for which the sufficient number of experiments is close to the lower bound. In the other extreme, there are graph classes for which the required number of experiments is multiplicatively alpha away from our lower bound. In simulations, our algorithm almost always performs very close to the lower bound, while the approach based on separating systems for complete graphs is significantly worse for random chordal graphs.

68 Regret-Based Pruning in Extensive-Form Games

Noam Brown noamb@cmu.edu
Tuomas Sandholm sandholm@cs.cmu.edu
Carnegie Mellon University

Counterfactual Regret Minimization (CFR) is a leading algorithm used to solve large zero-sum imperfect-information games. It is an iterative algorithm for finding a Nash equilibrium based on self play. It repeatedly traverses the game tree, updating regrets at each information set. We introduce an improvement to CFR that prunes any path of play in the tree, and its descendants, that has negative regret. It revisits that sequence at the earliest subsequent CFR iteration where the regret could have become positive, had that path been explored on every iteration. The new algorithm maintains CFR’s convergence guarantees while making iterations significantly faster---even if previously known pruning techniques are used in the comparison. This improvement carries over to CFR+, a recent variant of CFR. Experiments show an order of magnitude speed improvement, and the relative speed improvement increases with the size of the game.

69 Nonparametric von Mises Estimators for Entropies, Divergences and Mutual Informations

Kirthevasan Kandasamy kandasamy@cs.cmu.edu
Akshay Krishnamurthy akshaykr@cs.cmu.edu
Barnabas Poczos bapoczos@cs.cmu.edu
Larry Wasserman larry@stat.cmu.edu
Carnegie Mellon University
james m robins robins@hsph.harvard.edu
Harvard University

We propose and analyse estimators for statistical functionals of one or moredistributions under nonparametric assumptions. Our estimators are derived from the von Mises expansion andare based on the theory of influence functions, which appearin the semiparametric statistics literature.We show that estimators based either on data-splitting or a leave-one-out techniqueenjoy fast rates of convergence and other favorable theoretical properties.We apply this framework to derive estimators for several popular informationtheoretic quantities, and via empirical evaluation, show the advantage of thisapproach over existing estimators.

70 Bounding errors of Expectation-Propagation

Guillaume P Dehaene guillaume.dehaene@gmail.com
University of Geneva
Simon Barthelmé
simon.barthelme@gipsa-lab.grenoble-inp.fr
Gipsa-lab CNRS

Expectation Propagation is a very popular algorithm for variational inference, but comes with few theoretical guarantees. In this article, we prove that the approximation errors made by EP can be bounded. Our bounds have an asymptotic interpretation in the number n of datapoints, which allows us to study EP's convergence with respect to the true posterior. In particular, we show that EP converges at a rate of O(n−2) for the mean, up to an order of magnitude faster than the traditional Gaussian approximation at the mode. We also give similar asymptotic expansions for moments of order 2 to 4, as well as excess Kullback-Leibler cost (defined as the additional KL cost incurred by using EP rather than the ideal Gaussian approximation). All these expansions highlight the superior convergence properties of EP. Our approach for deriving those results is likely applicable to many similar approximate inference methods. In addition, we introduce bounds on the moments of log-concave distributions that may be of independent interest.

71 Market Scoring Rules Act As Opinion Pools For Risk-Averse Agents

Mithun Chakraborty mithunchakraborty@wustl.edu
Sanmay Das sanmay@wustl.edu
Washington University in St. Louis

A market scoring rule (MSR) – a popular tool for designing algorithmic prediction markets – is an incentive-compatible mechanism for the aggregation of probabilistic beliefs from myopic risk-neutral agents. In this paper, we add to a growing body of research aimed at understanding the precise manner in which the price process induced by a MSR incorporates private information from agents who deviate from the assumption of risk-neutrality. We first establish that, for a myopic trading agent with a risk-averse utility function, a MSR satisfying mild regularity conditions elicits the agent's risk-neutral probability conditional on the latest market state rather than her true subjective probability. Hence, we show that a MSR under these conditions effectively behaves like a more traditional method of belief aggregation, namely an opinion pool, for agents' true probabilities. In particular, the logarithmic market scoring rule acts as a logarithmic pool for constant absolute risk aversion utility agents, and as a linear pool for an atypical budget-constrained agent utility with decreasing absolute risk aversion. We also point out the interpretation of a market maker under these conditions as a Bayesian learner even when agent beliefs are static.

72 Local Smoothness in Variance Reduced Optimization

Daniel Vainsencher daniel.vainsencher@princeton.edu
Han Liu hanliu@princeton.edu
Princeton University
Tong Zhang tzhang@stat.rutgers.edu
Rutgers

We propose a family of non-uniform sampling strategies to speed up a class of stochastic optimization algorithms with linear convergence including Stochastic Variance Reduced Gradient (SVRG) and Stochastic Dual Coordinate Ascent(SDCA). For a large family of penalized empirical risk minimization problems, our methods exploit local smoothness of the loss functions near the optimum. Our bounds are at least as good as previously available bounds, and for some problems significantly better. Empirically, we provide thorough numerical results to back up our theory. Additionally we present algorithms exploiting local smoothness in more aggressive ways, which perform even better in practice.

73 High Dimensional EM Algorithm: Statistical Optimization and Asymptotic Normality

Zhaoran Wang zhaoran@princeton.edu
Quanquan Gu qg5w@virginia.edu
University of Virginia
Yang Ning yning@princeton.edu
Han Liu hanliu@princeton.edu
Princeton University

We provide a general theory of the expectation-maximization (EM) algorithm for inferring high dimensional latent variable models. In particular, we make two contributions: (i) For parameter estimation, we propose a novel high dimensional EM algorithm which naturally incorporates sparsity structure into parameter estimation. With an appropriate initialization, this algorithm converges at a geometric rate and attains an estimator with the (near-) optimal statistical rate of convergence. (ii) Based on the obtained estimator, we propose a new inferential procedure for testing hypotheses for low dimensional components of high dimensional parameters. For a broad family of statistical models, our framework establishes the first computationally feasible approach for optimal estimation and asymptotic inference in high dimensions. Our theory is supported by thorough numerical results.

74 Associative Memory via a Sparse Recovery Model

Arya Mazumdar arya@umn.edu
University of Minnesota -- Twin Cities
Ankit Singh Rawat asrawat@andrew.cmu.edu
Carnegie Mellon University

An associative memory is a structure learned from a dataset M of vectors (signals) in a way such that, given a noisy version of one of the vectors as input, the nearest valid vector from M (nearest neighbor) is provided as output, preferably via a fast iterative algorithm. Traditionally, binary (or q-ary) Hopfield neural networks are used to model the above structure. In this paper, for the first time, we propose a model of associative memory based on sparse recovery of signals. Our basic premise is simple. For a dataset, we learn a set of linear constraints that every vector in the dataset must satisfy. Provided these linear constraints possess some special properties, it is possible to cast the task of finding nearest neighbor as a sparse recovery problem. Assuming generic random models for the dataset, we show that it is possible to store super-polynomial or exponential number of n-length vectors in a neural network of size O(n). Furthermore, given a noisy version of one of the stored vectors corrupted in near-linear number of coordinates, the vector can be correctly recalled using a neurally feasible algorithm.

75 Matrix Completion Under Monotonic Single Index Models

Ravi Ganti gmravi2003@gmail.com
Rebecca Willett willett@discovery.wisc.edu
University of Wisconsin
Laura Balzano girasole@umich.edu
University of Michigan-Ann Arbor

Most recent results in matrix completion assume that the matrix under consideration is low-rank or that the columns are in a union of low-rank subspaces. In real-world settings, however, the linear structure underlying these models is distorted by a (typically unknown) nonlinear transformation. This paper addresses the challenge of matrix completion in the face of such nonlinearities. Give a few observations of a matrix, that is obtained by applying a Lipschitz, monotonic function to a low rank matrix, our task is to estimate the remaining unobserved entries. We propose a novel matrix completion method which alternates between low-rank matrix estimation and monotonic function estimation to estimate the missing matrix elements. Mean squared error bounds provide insight into how well the matrix can be estimated based on the size, rank of the matrix and properties of the nonlinear transformation. Empirical results on synthetic and real-world datasets demonstrate the competitiveness of the proposed approach.

76 Sparse Linear Programming via Primal and Dual Augmented Coordinate Descent

Ian En-Hsu Yen a061105@gmail.com
Kai Zhong zhongkai@ices.utexas.edu
Pradeep K Ravikumar pradeepr@cs.utexas.edu
Inderjit S Dhillon inderjit@cs.utexas.edu
University of Texas at Austin
Cho-Jui Hsieh cjhsieh@cs.utexas.edu
UC Davis

Over the past decades, Linear Programming (LP) has been widely used in different areas and considered as one of the mature technologies in numerical optimization. However, the complexity offered by state-of-the-art algorithms (i.e. interior-point method and primal, dual simplex methods) is still unsatisfactory for many problems in machine learning due to the expensive complexity w.r.t. number of variables and constraints. In this paper, we investigate a general LP method based on the combination of Augmented Lagrangian and Coordinate Descent (AL-CD). As we show, the proposed method achieves ϵ suboptimality with a complexity of O(nnz(A)(log(1/ϵ))2), where nnz(A) is the number of non-zeros in the m×n constraint matrix A, and in practice, one can further reduce cost of each iterate to the order of non-zeros in columns corresponding to the active variables through an active-set strategy. The algorithm thus yields a tractable alternative to standard LP methods for large-scale problems with nnz(A)$\ll$mn and sparse (primal or dual) solution. We conduct experiments on large-scale LP instances from ℓ1-regularized multi-class SVM, Sparse Inverse Covariance Estimation, and Nonnegative Matrix Factorization, where the proposed approach finds solution of 10−3 precision orders of magnitude faster than state-of-the-art implementations of interior-point and simplex methods.

77 Convergence rates of sub-sampled Newton methods

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

We consider the problem of minimizing a sum of n functions over a convex parameter set \C$\subset$\realsp where $\mathrm{n} \gg \mathrm{p} \gg 1$. In this regime, algorithms which utilize sub-sampling techniques are known to be effective.In this paper, we use sub-sampling techniques together with low-rank approximation to design a new randomized batch algorithm which possesses comparable convergence rate to Newton's method, yet has much smaller per-iteration cost. The proposed algorithm is robust in terms of starting point and step size, and enjoys a composite convergence rate, namely, quadratic convergence at start and linear convergence when the iterate is close to the minimizer. We develop its theoretical analysis which also allows us to select near-optimal algorithm parameters. Our theoretical results can be used to obtain convergence rates of previously proposed sub-sampling based algorithms as well. We demonstrate how our results apply to well-known machine learning problems.Lastly, we evaluate the performance of our algorithm on several datasets under various scenarios.

78 Variance Reduced Stochastic Gradient Descent with Neighbors

Thomas Hofmann thomas.hofmann@inf.ethz.ch
Aurelien Lucchi aurelien.lucchi@inf.ethz.ch
Brian McWilliams brian.mcwilliams@inf.ethz.ch
ETH Zurich
Simon Lacoste-Julien simon.lacoste-julien@ens.fr
INRIA

Stochastic Gradient Descent (SGD) is a workhorse in machine learning, yet it is also known to be slow relative to steepest descent. Recently, variance reduction techniques such as SVRG and SAGA have been proposed to overcome this weakness. With asymptotically vanishing variance, a constant step size can be maintained, resulting in geometric convergence rates. However, these methods are either based on occasional computations of full gradients at pivot points (SVRG), or on keeping per data point corrections in memory (SAGA). This has the disadvantage that one cannot employ these methods in a streaming setting and that speed-ups relative to SGD may need a certain number of epochs in order to materialize. This paper investigates a new class of algorithms that can exploit neighborhood structure in the training data to share and re-use information about past stochastic gradients across data points. While not meant to be offering advantages in an asymptotic setting, there are significant benefits in the transient optimization phase, in particular in a streaming or single-epoch setting. We investigate this family of algorithms in a thorough analysis and show supporting experimental results. As a side-product we provide a simple and unified proof technique for a broad class of variance reduction algorithms.

79 Non-convex Statistical Optimization for Sparse Tensor Graphical Model

Wei Sun sunweisurrey8@gmail.com
Yahoo Labs
Zhaoran Wang zhaoran@princeton.edu
Han Liu hanliu@princeton.edu
Princeton University
Guang Cheng chengg@purdue.edu
Purdue University

We consider the estimation of sparse graphical models that characterize the dependency structure of high-dimensional tensor-valued data. To facilitate the estimation of the precision matrix corresponding to each way of the tensor, we assume the data follow a tensor normal distribution whose covariance has a Kronecker product structure. The penalized maximum likelihood estimation of this model involves minimizing a non-convex objective function. In spite of the non-convexity of this estimation problem, we prove that an alternating minimization algorithm, which iteratively estimates each sparse precision matrix while fixing the others, attains an estimator with the optimal statistical rate of convergence as well as consistent graph recovery. Notably, such an estimator achieves estimation consistency with only one tensor sample, which is unobserved in previous work. Our theoretical results are backed by thorough numerical studies.

80 Convergence Rates of Active Learning for Maximum Likelihood Estimation

Kamalika Chaudhuri kamalika@cs.ucsd.edu
UCSD
Sham Kakade sham@cs.washington.edu
University of Washington
Praneeth Netrapalli praneethn@gmail.com
Microsoft Research
Sujay Sanghavi sanghavi@mail.utexas.edu
UTexas-Austin

An active learner is given a class of models, a large set of unlabeled examples, and the ability to interactively query labels of a subset of these examples; the goal of the learner is to learn a model in the class that fits the data well. Previous theoretical work has rigorously characterized label complexity of active learning, but most of this work has focused on the PAC or the agnostic PAC model. In this paper, we shift our attention to a more general setting -- maximum likelihood estimation. Provided certain conditions hold on the model class, we provide a two-stage active learning algorithm for this problem. The conditions we require are fairly general, and cover the widely popular class of Generalized Linear Models, which in turn, include models for binary and multiclass classification, regression, and conditional random fields. We provide an upper bound on the label requirement of our algorithm, and a lower bound that matches it up to lower order terms. Our analysis shows that unlike binary classification in the realizable case, just a single extraround of interaction is sufficient to achieve near-optimal performance in maximum likelihood estimation. On the empirical side, the recent work in (Gu et al. 2012) and (Gu et al. 2014) (on active linear and logistic regression) shows the promise of this approach.

81 When are Kalman-Filter Restless Bandits Indexable?

Christopher R Dance dance@xrce.xerox.com
Tomi Silander tomi.silander@xrce.xerox.com
Xerox Research Centre Europe

We study the restless bandit associated with an extremely simple scalar Kalman filter model in discrete time. Under certain assumptions, we prove that the problem is {\it indexable} in the sense that the {\it Whittle index} is a non-decreasing function of the relevant belief state. In spite of the long history of this problem, this appears to be the first such proof. We use results about {\it Schur-convexity} and {\it mechanical words}, which are particularbinary strings intimately related to {\it palindromes}.

82 Policy Gradient for Coherent Risk Measures

Aviv Tamar avivt@tx.technion.ac.il
UC Berkeley
Yinlam Chow yldick.chow@gmail.com
Stanford
Mohammad Ghavamzadeh
mohammad.ghavamzadeh@inria.fr
Adobe Research & INRIA
Shie Mannor shie@ee.technion.ac.il
Technion

Several authors have recently developed risk-sensitive policy gradient methods that augment the standard expected cost minimization problem with a measure of variability in cost. These

studies have focused on specific risk-measures, such as the variance or conditional value at risk (CVaR). In this work, we extend the policy gradient method to the whole class of coherent risk measures, which is widely accepted in finance and operations research, among other fields. We consider both static and time-consistent dynamic risk measures. For static risk measures, our approach is in the spirit of policy gradient algorithms and combines a standard sampling approach with convex programming. For dynamic risk measures, our approach is actor-critic style and involves explicit approximation of value function. Most importantly, our contribution presents a unified approach to risk-sensitive reinforcement learning that generalizes and extends previous results.

83 A Dual Augmented Block Minimization Framework for Learning with Limited Memory

Ian En-Hsu Yen a061105@gmail.com
University of Texas at Austin
Shan-Wei Lin skylark6802@gmail.com
Shou-De Lin sdlin@csie.ntu.edu.tw
National Taiwan University

In past few years, several techniques have been proposed for training of linear Support Vector Machine (SVM) in limited-memory setting, where a dual block-coordinate descent (dual-BCD) method was used to balance cost spent on I/O and computation. In this paper, we consider the more general setting of regularized \emph{Empirical Risk Minimization (ERM)} when data cannot fit into memory. In particular, we generalize the existing block minimization framework based on strong duality and \emph{Augmented Lagrangian} technique to achieve global convergence for ERM with arbitrary convex loss function and regularizer. The block minimization framework is flexible in the sense that, given a solver working under sufficient memory, one can integrate it with the framework to obtain a solver globally convergent under limited-memory condition. We conduct experiments on L1-regularized classification and regression problems to corroborate our convergence theory and compare the proposed framework to algorithms adopted from online and distributed settings, which shows superiority of the proposed approach on data of size ten times larger than the memory capacity.

84 On the Global Linear Convergence of Frank-Wolfe Optimization Variants

Simon Lacoste-Julien simon.lacoste-julien@ens.fr
INRIA
Martin Jaggi jaggi@inf.ethz.ch
ETH Zurich

The Frank-Wolfe (FW) optimization algorithm has lately regained popularity thanks in particular to its ability to nicely handle the structured constraints appearing in machine learning applications. However, its convergence rate is known to be slow (sublinear) when the solution lies at the boundary. A simple less-known fix is to add the possibility to take `away steps' during optimization, an operation that importantly does not require a feasibility oracle. In this paper, we highlight and clarify several variants of the Frank-Wolfe optimization algorithm that has been successfully applied in practice: FW with away steps, pairwise FW, fully-corrective FW and Wolfe's minimum norm point algorithm, and prove for the first time that they all enjoy global linear convergence under a weaker condition than strong convexity. The constant in the convergence rate has an elegant interpretation as the product of the (classical) condition number of the function with a novel geometric quantity that plays the role of the `condition number' of the constraint set. We provide pointers to where these algorithms have made a difference in practice, in particular with the flow polytope, the marginal polytope and the base polytope for submodular optimization.

85 Quartz: Randomized Dual Coordinate Ascent with Arbitrary Sampling

Zheng Qu zhengqu@maths.hku.hk
University of Hong Kong
Peter Richtarik peter.richtarik@ed.ac.uk
University of Edinburgh
Tong Zhang tzhang@stat.rutgers.edu
Rutgers

We study the problem of minimizing the average of a large number of smooth convex functions penalized with a strongly convex regularizer. We propose and analyze a novel primal-dual method (Quartz) which at every iteration samples and updates a random subset of the dual variables, chosen according to an arbitrary distribution. In contrast to typical analysis, we directly bound the decrease of the primal-dual error (in expectation), without the need to first analyze the dual error. Depending on the choice of the sampling, we obtain efficient serial and mini-batch variants of the method. In the serial case, our bounds match the best known bounds for SDCA (both with uniform and importance sampling). With standard mini-batching, our bounds predict initial data-independent speedup as well as additional data-driven speedup which depends on spectral and sparsity properties of the data.

86 A Generalization of Submodular Cover via the Diminishing Return Property on the Integer Lattice

Tasuku Soma tasuku_soma@mist.i.u-tokyo.ac.jp
University of Tokyo
Yuichi Yoshida yyoshida@nii.ac.jp
National Institute of Informatics

We consider a generalization of the submodular cover problem based on the concept of diminishing return property on the integer lattice. We are motivated by real scenarios in machine learning that cannot be captured by (traditional) submodular set functions. We show that the generalized submodular cover problem can be applied to various problems and devise a bicriteria approximation algorithm. Our algorithm is guaranteed to output a log-factor approximate solution that satisfies the constraints with the desired accuracy. The running time of our algorithm is roughly O(nlog(nr) logr), where n is the size of the ground set and r is the maximum value of a coordinate. The dependency on r is exponentially better than the naive reduction algorithms. Several experiments on real and artificial datasets demonstrate that the solution quality of our algorithm is comparable to naive algorithms, while the running time is several orders of magnitude faster.

87 A Universal Catalyst for First-Order Optimization

Hongzhou Lin hongzhou.lin@inria.fr
Julien Mairal julien.mairal@m4x.org
Zaid Harchaoui zaid.harchaoui@inria.fr
INRIA

We introduce a generic scheme for accelerating first-order optimization methods in the sense of Nesterov, which builds upon a new analysis of the accelerated proximal point algorithm. Our approach consists of minimizing a convex objective by approximately solving a sequence of well-chosen auxiliary problems, leading to faster convergence. This strategy applies to a large class of algorithms, including gradient descent, block coordinate descent, SAG, SAGA, SDCA, SVRG, Finito/MISO, and their proximal variants. For all of these methods, we provide acceleration and explicit support for non-strongly convex objectives. In addition to theoretical speed-up, we also show that acceleration is useful in practice, especially for ill-conditioned problems where we measure significant improvements.

88 Fast and Memory Optimal Low-Rank Matrix Approximation

Se-Young Yun seyoung.yun@inria.fr
marc lelarge marc.lelarge@ens.fr
INRIA
Alexandre Proutiere alepro@kth.se

In this paper, we revisit the problem of constructing a near-optimal rank k approximation of a matrix M∈[0,1]m×n under the streaming data model where the columns of M are revealed sequentially. We present SLA (Streaming Low-rank Approximation), an algorithm that is asymptotically accurate, when ksk+1(M)=o(mn−−−√) where sk+1(M) is the (k+1)-th largest singular value of M. This means that its average mean-square error converges to 0 as m and n grow large (i.e., $||\mathbb{M}$^(k)−M(k)$||2\mathrm{F}$=o(mn) with high probability, where M^(k) and M(k) denote the output of SLA and the optimal rank k approximation of M, respectively). Our algorithm makes one pass on the data if the columns of M are revealed in a random order, and two passes if the columns of M arrive in an arbitrary order. To reduce its memory footprint and complexity, SLA uses random sparsification, and samples each entry of M with a small probability δ. In turn, SLA is memory optimal as its required memory space scales as k(m+n), the dimension of its output. Furthermore, SLA is computationally efficient as it runs in O(δkmn) time (a constant number of operations is made for each observed entry of M), which can be as small as O(klog(m)4n) for an appropriate choice of δ and if n≥m.

89 Stochastic Online Greedy Learning with Semi-bandit Feedbacks

Tian Lin lint10@mails.tsinghua.edu.cn
Jian Li lijian83@mail.tsinghua.edu.cn
Tsinghua University
Wei Chen weic@microsoft.com
Microsoft.com

The greedy algorithm is extensively studied in the field of combinatorial optimization for decades. In this paper, we address the online learning problem when the input to the greedy algorithm is stochastic with unknown parameters that have to be learned over time. We first propose the greedy regret and ϵ-quasi greedy regret as learning metrics comparing with the performance of offline greedy algorithm. We then propose two online greedy learning algorithms with semi-bandit feedbacks, which use multi-armed bandit and pure exploration bandit policies at each level of greedy learning, one for each of the regret metrics respectively. Both algorithms achieve O(logT) problem-dependent regret bound (T being the time horizon) for a general class of combinatorial structures and reward functions that allow greedy solutions. We further show that the bound is tight in T and other problem instance parameters.

90 Linear Multi-Resource Allocation with Semi-Bandit Feedback

Tor Lattimore tor.lattimore@gmail.com
Csaba Szepesvari szepesva@ualberta.ca
University of Alberta
Koby Crammer koby@ee.technion.ac.il
Technion

We study an idealised sequential resource allocation problem. In each time step the learner chooses an allocation of several resource types between a number of tasks. Assigning more resources to a task increases the probability that it is completed. The problem is challenging because the alignment of the tasks to the resource types is unknown and the feedback is noisy. Our main contribution is the new setting and an algorithm with nearly-optimal regret analysis. Along the way we draw connections to the problem of minimising regret for stochastic linear bandits with heteroscedastic noise. We also present some new results for stochastic linear bandits on the hypercube that significantly outperforms existing work, especially in the sparse case.

91 Exactness of Approximate MAP Inference in Continuous MRFs

Nicholas Ruozzi nicholas.ruozzi@utdallas.edu
UTDallas

Computing the MAP assignment in graphical models is generally intractable. As a result, for discrete graphical models, the MAP problem is often approximated using linear programming relaxations. Much research has focused on characterizing when these LP relaxations are tight, and while they are relatively well-understood in the discrete case, only a few results are known for their continuous analog. In this work, we use graph covers to provide necessary and sufficient conditions for continuous MAP relaxations to be tight. We use this characterization to give simple proofs that the relaxation is tight for log-concave decomposable and log-supermodular decomposable models. We conclude by exploring the relationship between these two seemingly distinct classes of functions and providing specific conditions under which the MAP relaxation can and cannot be tight.

92 On the consistency theory of high dimensional variable screening

Xiangyu Wang xw56@stat.duke.edu
David B Dunson dunson@stat.duke.edu
Duke University
Chenlei Leng c.leng@warwick.ac.uk

Variable screening is a fast dimension reduction technique for assisting high dimensional feature selection. As a preselection method, it selects a moderate size subset of candidate variables for further refining via feature selection to produce the final model. The performance of variable screening depends on both computational efficiency and the ability to dramatically reduce the

number of variables without discarding the important ones. When the data dimension p is substantially larger than the sample size n, variable screening becomes crucial as 1) Faster feature selection algorithms are needed; 2) Conditions guaranteeing selection consistency might fail to hold.This article studies a class of linear screening methods and establishes consistency theory for this special class. In particular, we prove the restricted diagonally dominant (RDD) condition is a necessary and sufficient condition for strong screening consistency. As concrete examples, we show two screening methods SIS and HOLP are both strong screening consistent (subject to additional constraints) with large probability if n>O((ρs+σ/τ)2logp) under random designs. In addition, we relate the RDD condition to the irrepresentable condition, and highlight limitations of SIS.

93 Finite-Time Analysis of Projected Langevin Monte Carlo

Sebastien Bubeck sbubeck@princeton.edu
MSR
Ronen Eldan roneneldan@gmail.com
Joseph Lehec joseph.lehec@gmail.com

We analyze the projected Langevin Monte Carlo (LMC) algorithm, a close cousin of projected Stochastic Gradient Descent (SGD). We show that LMC allows to sample in polynomial time from a posterior distribution restricted to a convex body and with concave log-likelihood. This gives the first Markov chain to sample from a log-concave distribution with a first-order oracle, as the existing chains with provable guarantees (lattice walk, ball walk and hit-and-run) require a zeroth-order oracle. Our proof uses elementary concepts from stochastic calculus which could be useful more generally to understand SGD and its variants.

94 Optimal Testing for Properties of Distributions

Jayadev Acharya jayadev@csail.mit.edu
Constantinos Daskalakis costis@csail.mit.edu
Gautam C Kamath g@csail.mit.edu
Massachusetts Institute of Technology

Given samples from an unknown distribution, p, is it possible to distinguish whether p belongs to some class of distributions C versus p being far from every distribution in C? This fundamental question has receivedtremendous attention in Statistics, albeit focusing onasymptotic analysis, as well as in Computer Science, wherethe emphasis has been on small sample size and computationalcomplexity. Nevertheless, even for basic classes ofdistributions such as monotone, log-concave, unimodal, and monotone hazard rate, the optimal sample complexity is unknown.We provide a general approach via which we obtain sample-optimal and computationally efficient testers for all these distribution families. At the core of our approach is an algorithm which solves the following problem:Given samplesfrom an unknown distribution p, and a known distribution q, are p and q close in Chi^2-distance, or far in total variation distance?The optimality of all testers is established by providing matching lower bounds. Finally, a necessary building block for our tester and important byproduct of our work are the first known computationally efficient proper learners for discretelog-concave and monotone hazard rate distributions. We exhibit the efficacy of our testers via experimental analysis.

95 Learning Theory and Algorithms for Forecasting Non-stationary Time Series

Vitaly Kuznetsov vitaly@cims.nyu.edu
Mehryar Mohri mohri@cs.nyu.edu
Courant Institute and Google

We present data-dependent learning bounds for the general scenario of non-stationary non-mixing stochastic processes. Our learning guarantees are expressed in terms of the notion of sequential complexity and a discrepancy measure that can be estimated from data under some mild assumptions. We use our learning bounds to devise new algorithms for non-stationary time series forecasting for which we report some preliminary experimental results.

96 Accelerated Mirror Descent in Continuous and Discrete Time

Walid Krichene walid@eecs.berkeley.edu
Alexandre Bayen bayen@berkeley.edu
Peter L Bartlett bartlett@cs.berkeley.edu
UC Berkeley

We study accelerated mirror descent dynamics in continuous and discrete time. Combining the original continuous-time motivation of mirror descent with a recent ODE interpretation of Nesterov's accelerated method, we propose a family of continuous-time descent dynamics for convex functions with Lipschitz gradients, such that the solution trajectories are guaranteed to converge to the optimum at a O(1/t2) rate. We then show that a large family of first-order accelerated methods can be obtained as a discretization of the ODE, and these methods converge at a O(1/k2) rate. This connection between accelerated mirror descent and the ODE provides an intuitive approach to the design and analysis of accelerated first-order algorithms.

97 Information-theoretic lower bounds for convex optimization with erroneous oracles

Yaron Singer yaron@seas.harvard.edu
Harvard University
Jan Vondrak jvondrak@gmail.com
IBM Research

We consider the problem of optimizing convex and concave functions with access to an erroneous zeroth-order oracle. In particular, for a given function x→f(x) we consider optimization when one is given access to absolute error oracles that return values in [f(x) - \epsilon,f(x)+\epsilon] or relative error oracles that return value in [(1+\epsilon)f(x), (1 +\epsilon)f (x)], for small fixed \epsilon larger than 0. We show stark information theoretic impossibility results for minimizing convex functions and maximizing concave functions over polytopes in this model.

98 Bandit Smooth Convex Optimization: Improving the Bias-Variance Tradeoff

Ofer Dekel oferd@microsoft.com
Microsoft Research
Ronen Eldan roneneldan@gmail.com
Tomer Koren tomerk@technion.ac.il
Technion

Bandit convex optimization is one of the fundamental problems in the field of online learning. The best algorithm for the general bandit convex optimization problem guarantees a regret of O˜(T5/6), while the best known lower bound is Ω(T1/2). Many attemptshave been made to bridge the huge gap between these bounds. A particularly interesting special case of this problem assumes that the loss functions are smooth. In this case, the best known algorithm guarantees a regret of O˜(T2/3). The current state of the problem has given rise to two competing conjectures: some believe that there exists an algorithm that matches the known lower bound, while others believe that the lower bound can be improved to Ω(T2/3). We present an efficient algorithm for the bandit smooth convex optimization problem that guarantees a regret of O˜(T5/8). Our result rules out a Ω(T2/3) lower bound and takes a significant step towards the resolution of this open problem.

99 Beyond Sub-Gaussian Measurements: High-Dimensional Structured Estimation with Sub-Exponential Designs

Vidyashankar Sivakumar sivak017@umn.edu
Arindam Banerjee banerjee@cs.umn.edu
University of Minnesota
Pradeep K Ravikumar pradeepr@cs.utexas.edu
University of Texas at Austin

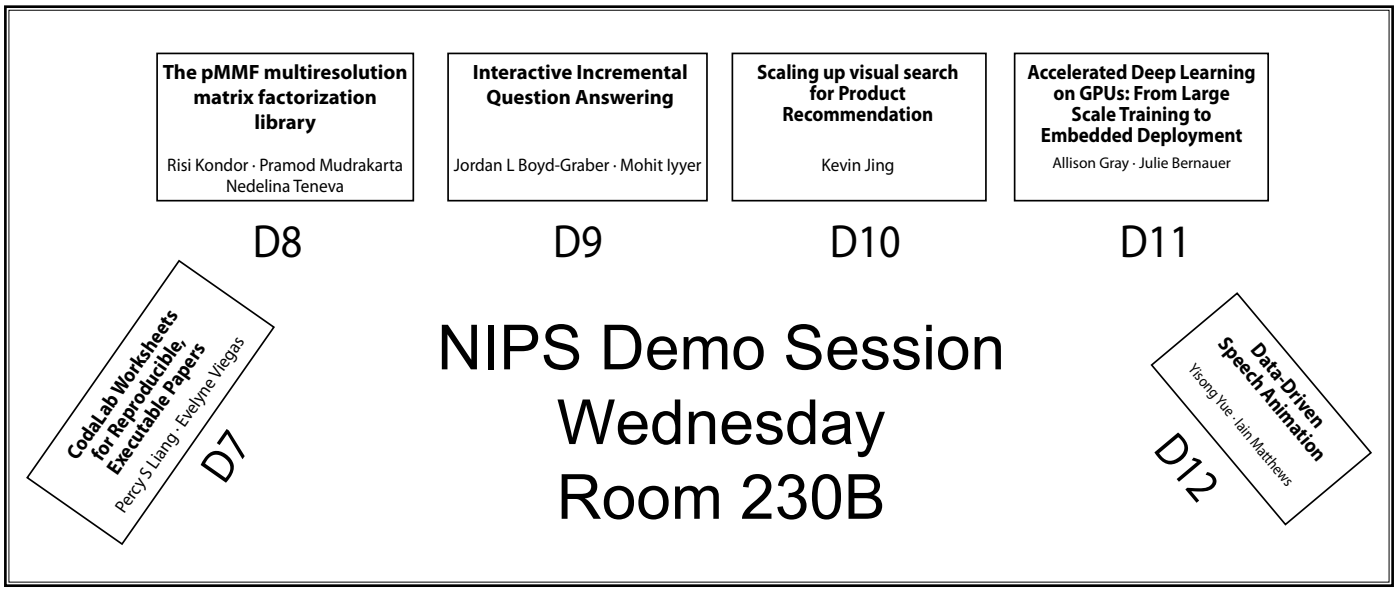
We consider the problem of high-dimensional structured estimation with norm-regularized estimators, such as Lasso, when the design matrix and noise are sub-exponential.Existing results only consider sub-Gaussian designs and noise, and both the sample complexity and non-asymptotic estimation error have been shown to depend on the Gaussian width of suitable sets. In contrast, for the sub-exponential setting, we show that the sample complexity and the estimation error will depend on the exponential width of the corresponding sets, and the analysis holds for any norm. Further, using generic chaining, we show that the exponential width for any set will be at most logp−−−−√ times the Gaussian width of the set, yielding Gaussian width based results even for the sub-exponential case. Further, for certain popular estimators, viz Lasso and Group Lasso, using a VC-dimension based analysis, we show that the sample complexity will in fact be the same order as Gaussian designs. Our results are the first in the sub-exponential setting, and are readily applicable to special sub-exponential families such as log-concave distributions.

100 Adaptive Online Learning

Dylan J Foster djfoster@cs.cornell.edu
Karthik Sridharan sridharan@cs.cornell.edu
Cornell University
Alexander Rakhlin rakhlin@wharton.upenn.edu
UPenn

We propose a general framework for studying adaptive regret bounds in the online learning framework, including model selection bounds and data-dependent bounds. Given a data- or model- dependent bound we ask, “Does there exist some algorithm achieving this bound?” We show that modifications to recently introduced sequential complexity measures can be used to answer this question by providing sufficient conditions under which adaptive rates can be achieved. In particular, an adaptive rate induces a set of so-called offset complexity measures, and obtaining small upper bounds on these quantities is sufficient to demonstrate achievability. A cornerstone of our analysis technique is the use of one-sided tail inequalities to bounding suprema of offset random processes.Our framework recovers and improves a wide variety of adaptive bounds including quantile bounds, second-order data-dependent bounds, and small-loss bounds. In addition, we derive a new type of adaptive bound for online linear optimization based on the spectral norm, as well as a new online PAC-Bayes theorem that holds for countably infinite sets.

DEMONSTRATIONS ABSTRACTS



D7 CodaLab Worksheets for Reproducible, Executable Papers

Percy S Liang pliang@cs.stanford.edu
Stanford University
Evelyne Viegas evelynev@microsoft.com
Microsoft Research

We are interested in solving two infrastructural problems in data-centric fields such as machine learning: First, an inordinate amount of time is spent on preprocessing datasets, getting other people's code to run, writing evaluation/visualization scripts, with much of this effort duplicated across different research groups. Second, a only static set of final results are ever published, leaving it up to the reader to guess how the various methods would fare in unreported scenarios. We present CodaLab Worksheets, a new platform which aims to tackle these two problems by creating an online community around sharing and executing immutable components called bundles, thereby streamlining the research process.

D8 The pMMF multiresolution matrix factorization library

Risi Kondor risi@uchicago.edu
Pramod Kaushik Mudrakarta pramodkm@uchicago.edu
Nedelina Teneva nteneva@uchicago.edu
The University of Chicago

pMMF is a high performance, general purpose, parallel C++ library for computing Multiresolution Matrix Factorizations at massive scales. In addition to its C++ API, pMMF can also be accessed from Matlab or through the command line. We demonstrate pMMF with an interactive visualization, showing what it does to graphs/networks.

D9 Interactive Incremental Question Answering

Jordan L Boyd-Graber jbg@umiacs.umd.edu
University of Colorado
Mohit Iyyer m.iyyer@gmail.com
University of Maryland, College Park

We present a machine learning system that plays a trivia game called "quiz bowl", in which questions are incrementally revealed to players. This task requires players (both human and computer) to decide not only what answer to guess but also when to interrupt the moderator ("buzz in") with that guess. Our system uses a recently-introduced deep learning model called the deep averaging network (or DAN) to generate a set of candidate guesses given a partially-revealed question. For each candidate guess, features are generated from language models and fed to a classifier that decides whether to buzz in with that guess. Previous demonstrations have shown that our system can compete with skilled human players.

D10 Scaling up visual search for product recommendation

Kevin Jing jing@pinterest.com
Pinterest

We demonstrate that, with the availability of distributed computation platforms such as Amazon Web Services and open-source tools such as Caffe, it is possible for a small engineering team to build, launch and maintain a cost-effective, large-scale visual search system. This demo showcases an interactive visual search system with more than 1 billion Pinterest image in its index built by a team of 2 engineers. By sharing our implementation details and learnings from launching a commercial visual search engine from scratch, we hope visual search becomes more widely incorporated into today's commercial applications.

D11 Accelerated Deep Learning on GPUs: From Large Scale Training to Embedded Deployment

Allison Gray agray@nvidia.com
Julie Bernauer jbernauer@nvidia.com
NVIDIA

Time is critical when developing and training the the right network on real-world datasets. GPUs offer a great computational advantage for training deep networks. Newly developed hardware resources, such as the DIGITS DevBox, can greatly reduce this development time. Using multiple GPUs at the same time allows both multi-GPU training and parallel network model tuning to accelerate training. The four Titan X GPUs in the DIGITS DevBox allow rapid development in a single deskside appliance. Libraries are key to maximizing performance when exploiting the power of the GPU. CUDA libraries, including the NVIDIA cuDNN deep learning library, are also used to accelerate training. cuDNN is an optimized library comprised of functions commonly used to create artificial neural networks. This library includes activation functions, convolutions, and Fourier transforms, with calls catered to processing neural network data. This library is designed with deep neural network frameworks in mind and is easily used with popular open source frameworks such as Torch, Caffe, and Theano. A trained deep neural network learns its features through an iterative process, convolving abstracted features to discern between the objects it has been trained to identify. NVIDIA DIGITS allows researchers to quickly and easily visualize the layers of trained networks. This novel visualization tool can be used with any GPU accelerated platform and works with popular frameworks like Caffe and Torch. Working with popular frameworks enables easy deployment of trained networks to other platforms, such as embedded platforms. This demonstration will show how easy it is to quickly develop a trained network using multiple GPUs on the DIGITS DevBox with the DIGITS software and deploy it to the newly released embedded platform for classification on a mobile deployment scenario.

D12 Data-Driven Speech Animation

Yisong Yue yyue@caltech.edu
Caltech
Iain Matthews iainm@disneyresearch.com
Disney Research Pittsburgh

Speech animation is an extremely tedious task, where the animation artist must manually animate the face to match the spoken audio. The often prohibitive cost of speech animation has limited the types of animations that are feasible, including localization to different languages. In this demo, we will showcase a new machine learning approach for automated speech animation. Given audio or phonetic input, our approach predicts the lower-face configurations of an animated character to match the input. In our demo, you can speak to our system, and our system will automatically animate an animated character to lip sync to your speech. The technical details can be found in a recent KDD paper titled "A Decision Tree Framework for Spatiotemporal Sequence Prediction" by Taehwan Kim, Yisong Yue, Sarah Taylor, and Iain Matthews.

THURSDAY SESSIONS

OUR SPONSORS

Google

Microsoft

Alibaba Group 阿里巴巴集团

Ketchum Trading, LLC

amazon.com

Apple

Baidu Research

CITADEL

facebook

IBM Research

NVIDIA

The Voleon Group

Artificial Intelligence
www.elsevier.com/locate/artint

Bloomberg

twitter

AdRoll

Analog Devices | Lyric Labs

CenturyLink Business

criteo

Cubist Systematic Strategies

deep genomics

D E Shaw & Co

NETFLIX

Disney Research

ebay

imagia

MaluubA

Man AHL

ORACLE

Panasonic

PDT PARTNERS

SONY

The Alan Turing Institute

TOYOTA

United Technologies Research Center

Vatic Labs

TWO SIGMA

YAHOO! LABS

WINTON

Adobe

ORAL SESSION

SESSION 9: 9:00 - 10:10 AM

INVITED TALK: POSNER LECTURE
Learning with Intelligent Teacher: Similarity Control and Knowledge Transfer

Vladimir Vapnik vladimir.vapnik@gmail.com
Columbia University and Facebook AI Research

In the talk, I will introduce a model of learning with Intelligent Teacher. In this model, Intelligent Teacher supplies (some) training examples (x_i, y_i), i=1,..., ℓ, $x_i \in X$,$y_i \in \{-1,1\}$ with additional (privileged) information) $\{x_i^* \in X^*$ forming training triplets (x_i,x_i^*, y_i), i = 1,...,ℓ. Privileged information is available only for training examples and *not available for test examples* Using privileged information it is required to find a better training processes (that use less examples or more accurate with the same number of examples) than the classical ones.

In this lecture, I will present two additional mechanisms that exist in learning with Intelligent Teacher

- The mechanism to control Student's concept of examples similarity and
- The mechanism to transfer knowledge that can be obtained in space of privileged information to the desired space of decision rules.

Privileged information exists for many inference problem and Student-Teacher interaction can be considered as the basic element of intelligent behavior.

Less is More: Nyström Computational Regularization

Alessandro Rudi ale_rudi@mit.edu
Raffaello Camoriano raffaello.camoriano@iit.it
IIT - UNIGE
Lorenzo Rosasco lrosasco@mit.edu
University of Genova

We study Nyström type subsampling approaches to large scale kernel methods, and prove learning bounds in the statistical learning setting, where random sampling and high probability estimates are considered. In particular, we prove that these approaches can achieve optimal learning bounds, provided the subsampling level is suitably chosen. These results suggest a simple incremental variant of Nyström kernel ridge regression, where the subsampling level controls at the same time regularization and computations. Extensive experimental analysis shows that the considered approach achieves state of the art performances on benchmark large scale datasets.

SPOTLIGHT SESSION

SESSION 9: 10:10 – 10:40 AM

- **Logarithmic Time Online Multiclass prediction**
  Anna E Choromanska · John Langford
- **Collaborative Filtering with Graph Information: Consistency and Scalable Methods**
  Nikhil Rao · Hsiang-Fu Yu · Inderjit S Dhillon · Pradeep K Ravikumar
- **Efficient and Parsimonious Agnostic Active Learning**
  T.-K. Huang · Alekh Agarwal · Daniel J Hsu · John Langford · Robert Schapire
- **Matrix Completion with Noisy Side Information**
  Kai-Yang Chiang · Cho-Jui Hsieh · Inderjit S Dhillon
- **Learning with Symmetric Label Noise: The Importance of Being Unhinged**
  Brendan van Rooyen · Aditya Menon · Robert Williamson
- **Scalable Semi-Supervised Aggregation of Classifiers**
  Akshay Balsubramani · Yoav Freund
- **Spherical Random Features for Polynomial Kernels**
  Jeffrey Pennington · · Sanjiv Kumar
- **Fast and Guaranteed Tensor Decomposition via Sketching**
  Yining Wang · Hsiao-Yu Tung · Alex J Smola · Anima Anandkumar

POSTER SESSION

POSTERS 1:00 – 3:00 PM

1 **Teaching Machines to Read and Comprehend**
Karl Moritz Hermann · Tomas Kocisky · Edward Grefenstette · Lasse Espeholt · Will Kay · Mustafa Suleyman · Phil Blunsom

2 **Saliency, Scale and Information: Towards a Unifying Theory**
Shafin Rahman · Neil Bruce

3 **Semi-supervised Learning with Ladder Networks**
Antti Rasmus · Mathias Berglund · Mikko Honkala · Harri Valpola · Tapani Raiko

4 **Enforcing balance allows local supervised learning in spiking recurrent networks**
Ralph Bourdoukan · Sophie Denève

5 **Semi-supervised Sequence Learning**
Andrew M Dai · Quoc V Le

6 **Skip-Thought Vectors**
Ryan Kiros · Yukun Zhu · Russ R Salakhutdinov · Richard Zemel · Raquel Urtasun · Antonio Torralba · Sanja Fidler

7 **Learning to Linearize Under Uncertainty**
Ross Goroshin · Michael F Mathieu · Yann LeCun

8 **Synaptic Sampling: A Bayesian Approach to Neural Network Plasticity and Rewiring**
David Kappel · Stefan Habenschuss · Robert Legenstein · Wolfgang Maass

9 **Natural Neural Networks**
Guillaume Desjardins · Karen Simonyan · Razvan Pascanu · koray kavukcuoglu

10 Convolutional Networks on Graphs for Learning Molecular Fingerprints
David K Duvenaud · Dougal Maclaurin · Jorge Iparraguirre · Rafael Bombarell · Timothy Hirzel · Alan Aspuru-Guzik · Ryan P Adams

11 Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting
Xingjian SHI · Zhourong Chen · Hao Wang · Dit-Yan Yeung · Wai-kin Wong · Wang-chun WOO

12 Scheduled Sampling for Sequence Prediction with Recurrent Neural Networks
Samy Bengio · Oriol Vinyals · Navdeep Jaitly · Noam Shazeer

13 Mind the Gap: A Generative Approach to Interpretable Feature Selection and Extraction
Been Kim · Julie A Shah · Finale Doshi-Velez

14 Max-Margin Deep Generative Models
Chongxuan Li · Jun Zhu · Tianlin Shi · Bo Zhang

15 Cross-Domain Matching for Bag-of-Words Data via Kernel Embeddings of Latent Distributions
Yuya Yoshikawa · Tomoharu Iwata · Hiroshi Sawada · Takeshi Yamada

16 A Gaussian Process Model of Quasar Spectral Energy Distributions
Andrew Miller · Albert Wu · Jeff Regier · Jon McAuliffe · Dustin Lang · Mr. Prabhat · David Schlegel · Ryan P Adams

17 Neural Adaptive Sequential Monte Carlo
Shixiang Gu · Zoubin Ghahramani · Richard E Turner

18 Convolutional spike-triggered covariance analysis for neural subunit models
Anqi Wu · Memming Park · Jonathan W Pillow

19 Rectified Factor Networks
Djork-Arné Clevert · Andreas Mayr · Thomas Unterthiner · Sepp Hochreiter

20 Embed to Control: A Locally Linear Latent Dynamics Model for Control from Raw Images
Manuel Watter · Jost Springenberg · Joschka Boedecker · Martin Riedmiller

21 Bayesian dark knowledge
Anoop Korattikara Balan · Vivek Rathod · Kevin P Murphy · Max Welling

22 GP Kernels for Cross-Spectrum Analysis
Kyle R Ulrich · David E Carlson · Kafui Dzirasa · Lawrence Carin

23 End-to-end Learning of LDA by Mirror-Descent Back Propagation over a Deep Architecture
Jianshu Chen · Ji He · Yelong Shen · Lin Xiao · Xiaodong He · Jianfeng Gao · Xinying Song · Li Deng

24 Particle Gibbs for Infinite Hidden Markov Models
Nilesh Tripuraneni · Shixiang Gu · Hong Ge · Zoubin Ghahramani

25 Sparse Local Embeddings for Extreme Multi-label Classification
Kush Bhatia · Himanshu Jain · Puru Kar · Manik Varma · Prateek Jain

26 Robust Spectral Inference for Joint Stochastic Matrix Factorization
Moontae Lee · David Bindel · David Mimno

27 Space-Time Local Embeddings
Ke Sun · Jun Wang · Alexandros Kalousis · Stephane Marchand-Maillet

28 A fast, universal algorithm to learn parametric nonlinear embeddings
Miguel A. Carreira-Perpinan · Max Vladymyrov

29 Bayesian Manifold Learning: the Locally Linear Latent Variable Model (LL-LVM)
Mijung Park · Wittawat Jitkrittum · Ahmad Qamar · Zoltan Szabo · Lars Buesing · Maneesh Sahani

30 Local Causal Discovery of Direct Causes and Effects
Tian Gao · Qiang Ji

31 Discriminative Robust Transformation Learning
Jiaji Huang · Qiang Qiu · Guillermo Sapiro · Robert Calderbank

32 Max-Margin Majority Voting for Learning from Crowds
TIAN TIAN · Jun Zhu

33 M-Best-Diverse Labelings for Submodular Energies and Beyond
Alexander Kirillov · Dmytro Shlezinger · Dmitry P Vetrov · Carsten Rother · Bogdan Savchynskyy

34 Covariance-Controlled Adaptive Langevin Thermostat for Large-Scale Bayesian Sampling
Xiaocheng Shang · Zhanxing Zhu · Benedict Leimkuhler · Amos J Storkey

35 Time-Sensitive Recommendation From Recurrent User Activities
Nan Du · Yichen Wang · Niao He · Jimeng Sun · Le Song

36 Parallel Recursive Best-First AND/OR Search for Exact MAP Inference in Graphical Models
Akihiro Kishimoto · Radu Marinescu · Adi Botea

37 Logarithmic Time Online Multiclass prediction
Anna E Choromanska · John Langford

38 Scalable Semi-Supervised Aggregation of Classifiers
Akshay Balsubramani · Yoav Freund

39 Bounding the Cost of Search-Based Lifted Inference
David B Smith · Vibhav G Gogate

40 Efficient Learning by Directed Acyclic Graph For Resource Constrained Prediction
Joseph Wang · Kirill Trapeznikov · Venkatesh Saligrama

41 **Estimating Jaccard Index with Missing Observations: A Matrix Calibration Approach**
Wenye Li

42 **Sample Efficient Path Integral Control under Uncertainty**
Yunpeng Pan · Evangelos Theodorou - Michail Kontitsis

43 **Efficient Thompson Sampling for Online Matrix-Factorization Recommendation**
Jaya Kawale · Hung H Bui · Branislav Kveton · Long Tran-Thanh · Sanjay Chawla

44 **Parallelizing MCMC with Random Partition Trees**
Xiangyu Wang · Richard Guo · Katherine Heller · David B Dunson

45 **Fast Lifted MAP Inference via Partitioning**
Somdeb Sarkhel · Parag Singla · Vibhav G Gogate

46 **Active Learning from Weak and Strong Labelers**
Chicheng Zhang · Kamalika Chaudhuri

47 **Fast and Guaranteed Tensor Decomposition via Sketching**
Yining Wang · Hsiao-Yu Tung · Alex J Smola · Anima Anandkumar

48 **Spherical Random Features for Polynomial Kernels**
Jeffrey Pennington · · Sanjiv Kumar

49 **Learnability of Influence in Networks**
Hari Harikrishna · David C Parkes · Yaron Singer

50 **A Pseudo-Euclidean Iteration for Optimal Recovery in Noisy ICA**
James R Voss · Mikhail Belkin · Luis Rademacher

51 **Differentially private subspace clustering**
Yining Wang · Yu-Xiang Wang · Aarti Singh

52 **Compressive spectral embedding: sidestepping the SVD**
Dinesh Ramasamy · Upamanyu Madhow

53 **Generalization in Adaptive Data Analysis and Holdout Reuse**
Cynthia Dwork · Vitaly Feldman · Moritz Hardt · Toni Pitassi · Omer Reingold · Aaron Roth

54 **Online F-Measure Optimization**
Róbert Busa-Fekete · Balázs Szörényi · Krzysztof Dembczynski · Eyke Hüllermeier

55 **Matrix Completion with Noisy Side Information**
Kai-Yang Chiang · Cho-Jui Hsieh · Inderjit S Dhillon

56 **A Market Framework for Eliciting Private Data**
Bo Waggoner · Rafael M Frongillo · Jacob D Abernethy

57 **Optimal Ridge Detection using Coverage Risk**
Yen-Chi Chen · Christopher Genovese · Shirley Ho · Larry Wasserman

58 **Fast Distributed k-Center Clustering with Outliers on Massive Data**
Luiz Gustavo Sant Anna Malkomes Muniz · Matt J Kusner · Wenlin Chen · Kilian Q Weinberger · Benjamin Moseley

59 **Orthogonal NMF through Subspace Exploration**
megas asteris · Dimitris Papailiopoulos · Alex G Dimakis

60 **Fast Classification Rates for High-dimensional Gaussian Generative Models**
Tianyang Li · Adarsh Prasad · Pradeep K Ravikumar

61 **Efficient and Parsimonious Agnostic Active Learning**
T.-K. Huang · Alekh Agarwal · Daniel J Hsu · John Langford · Robert Schapire

62 **Collaborative Filtering with Graph Information: Consistency and Scalable Methods**
Nikhil Rao · Hsiang-Fu Yu · Inderjit S Dhillon · Pradeep K Ravikumar

63 **Less is More: Nyström Computational Regularization**
Alessandro Rudi · Raffaello Camoriano · Lorenzo Rosasco

64 **Predtron: A Family of Online Algorithms for General Prediction Problems**
Prateek Jain · Nagarajan Natarajan · Ambuj Tewari

65 **On the Optimality of Classifier Chain for Multi-label Classification**
Weiwei Liu · Ivor Tsang

66 **Smooth Interactive Submodular Set Cover**
Bryan D He · Yisong Yue

67 **Tractable Bayesian Network Structure Learning with Bounded Vertex Cover Number**
Janne H Korhonen · Pekka Parviainen

68 **Secure Multi-party Differential Privacy**
Peter Kairouz · Sewoong Oh · Pramod Viswanath

69 **Adaptive Stochastic Optimization: From Sets to Paths**
Zhan Wei Lim · David Hsu · Wee Sun Lee

70 **Learning structured densities via infinite dimensional exponential families**
Siqi Sun · mladen kolar · Jinbo Xu

71 **Lifelong Learning with Non-i.i.d. Tasks**
Anastasia Pentina · Christoph H Lampert

72 **Learning with Symmetric Label Noise: The Importance of Being Unhinged**
Brendan van Rooyen · Aditya Menon · Robert Williamson

73 **Algorithms with Logarithmic or Sublinear Regret for Constrained Contextual Bandits**
Huasen Wu · R. Srikant · Xin Liu · Chong Jiang

74 **From random walks to distances on unweighted graphs**
Tatsunori Hashimoto · Yi Sun · Tommi Jaakkola

75 Robust Regression via Hard Thresholding
Kush Bhatia · Prateek Jain · Puru Kar

76 Column Selection via Adaptive Sampling
Saurabh Paul · Malik Magdon-Ismail · Petros Drineas

77 Multi-class SVMs: From Tighter Data-Dependent Generalization Bounds to Novel Algorithms
Yunwen Lei · Urun Dogan · Alexander Binder · Marius Kloft

78 Optimal Linear Estimation under Unknown Nonlinear Transform
Xinyang Yi · Zhaoran Wang · Constantine Caramanis · Han Liu

79 Risk-Sensitive and Robust Decision-Making: a CVaR Optimization Approach
Yinlam Chow · Aviv Tamar · Shie Mannor · Marco Pavone

80 Learning with Incremental Iterative Regularization
Lorenzo Rosasco · Silvia Villa

81 No-Regret Learning in Repeated Bayesian Games
Jason Hartline · Vasilis Syrgkanis · Eva Tardos

82 Sparse and Low-Rank Tensor Decomposition
Parikshit Shah · Nikhil Rao · Gongguo Tang

83 Analysis of Robust PCA via Local Incoherence
Huishuai Zhang · Yi Zhou · Yingbin Liang

84 Algorithmic Stability and Uniform Generalization
Ibrahim M Alabdulmohsin

85 Mixing Time Estimation in Reversible Markov Chains from a Single Sample Path
Daniel J Hsu · Aryeh Kontorovich · Csaba Szepesvari

86 Efficient Compressive Phase Retrieval with Constrained Sensing Vectors
Sohail Bahmani · Justin Romberg

87 Unified View of Matrix Completion under General Structural Constraints
Suriya Gunasekar · Arindam Banerjee · Joydeep Ghosh

88 Copeland Dueling Bandits
Masrour Zoghi · Zohar S Karnin · Shimon Whiteson · Maarten de Rijke

89 Regret Lower Bound and Optimal Algorithm in Finite Stochastic Partial Monitoring
Junpei Komiyama · Junya Honda · Hiroshi Nakagawa

90 Online Learning for Adversaries with Memory: Price of Past Mistakes
Oren Anava · Elad Hazan · Shie Mannor

91 Revenue Optimization against Strategic Buyers
Mehryar Mohri · Andres Munoz

92 On Top-k Selection in Multi-Armed Bandits and Hidden Bipartite Graphs
Wei Cao · Jian Li · Yufei Tao · Zhize Li

93 Improved Iteration Complexity Bounds of Cyclic Block Coordinate Descent for Convex Problems
Ruoyu Sun · Mingyi Hong

94 Cornering Stationary and Restless Mixing Bandits with Remix-UCB
Julien Audiffren · Liva Ralaivola

95 Fighting Bandits with a New Kind of Smoothness
Jacob D Abernethy · Chansoo Lee · Ambuj Tewari

96 Asynchronous stochastic approximation: the noise is in the noise and SGD don't care
John C Duchi · Sorathan Chaturapruek · Chris Ré

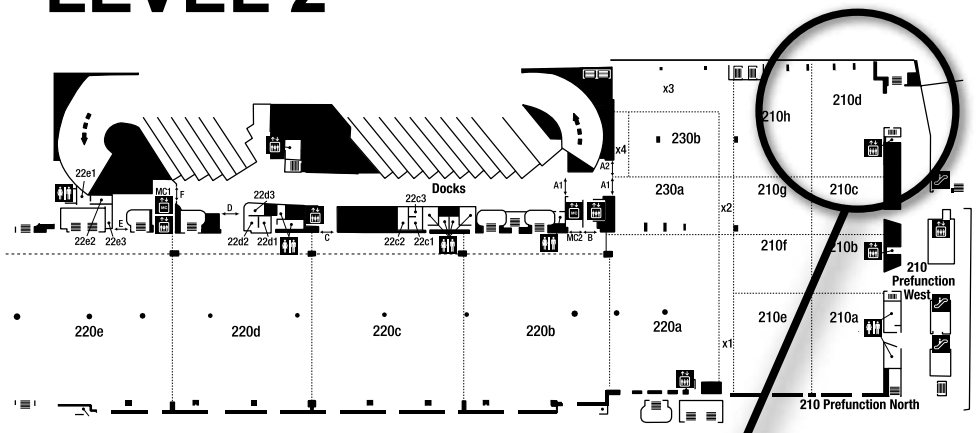
97 The Pareto Regret Frontier for Bandits
Tor Lattimore

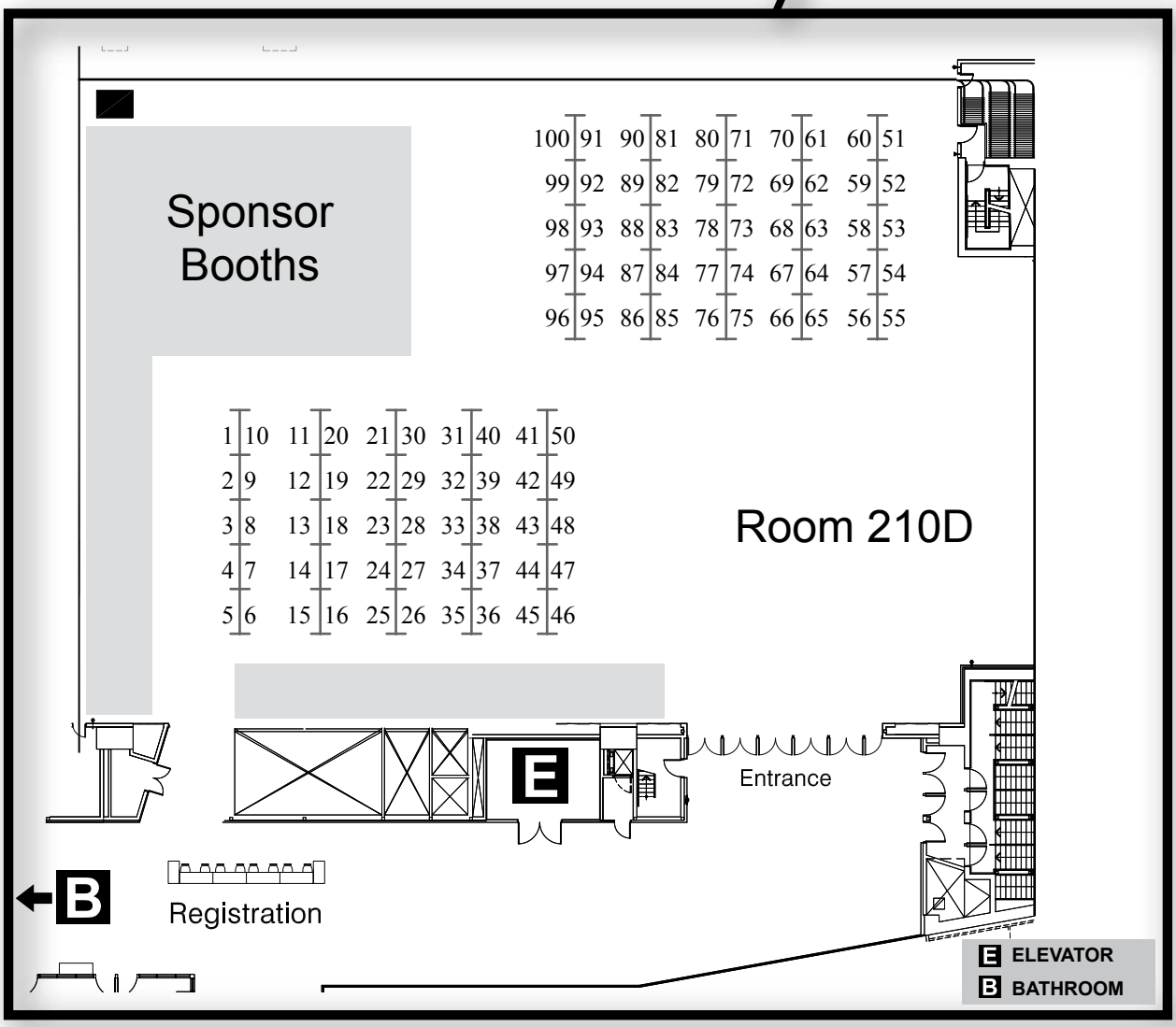
98 Online Learning with Gaussian Payoffs and Side Observations
Yifan Wu · Gyorgy Andras · Csaba Szepesvari

99 Fast Rates for Exp-concave Empirical Risk Minimization
Tomer Koren · Kfir Levy

100 Adaptive Low-Complexity Sequential Inference for Dirichlet Process Mixture Models
Theodoros Tsiligkaridis · Theodoros Tsiligkaridis · Keith Forsythe

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


Sponsor
Booths
100 91 90 81 80 71 70 61 60 51
99 92 89 82 79 72 69 62 59 52
98 93 88 83 78 73 68 63 58 53
97 94 87 84 77 74 67 64 57 54
96 95 86 85 76 75 66 65 56 55
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
E
Entrance
B
Registration
E ELEVATOR
B BATHROOM

1 Teaching Machines to Read and Comprehend

Karl Moritz Hermann kmh@google.com
Lasse Espeholt lespeholt@google.com
Will Kay wkay@google.com
Mustafa Suleyman mustafasul@google.com
Phil Blunsom pblunsom@google.com
Edward Grefenstette etg@google.com
Google DeepMind
Tomas Kocisky tomas@kocisky.eu
Oxford University

Teaching machines to read natural language documents remains an elusive challenge. Machine reading systems can be tested on their ability to answer questions posed on the contents of documents that they have seen, but until now large scale training and test datasets have been missing for this type of evaluation. In this work we define a new methodology that resolves this bottleneck and provides large scale supervised reading comprehension data. This allows us to develop a class of attention based deep neural networks that learn to read real documents and answer complex questions with minimal prior knowledge of language structure.

2 Saliency, Scale and Information: Towards a Unifying Theory

Shafin Rahman shafin109@gmail.com
Neil Bruce bruce@cs.umanitoba.ca
University of Manitoba

In this paper we present a definition for visual saliency grounded in information theory. This proposal is shown to relate to a variety of classic research contributions in scale-space theory, interest point detection, bilateral filtering, and to existing models of visual saliency. Based on the proposed definition of visual saliency, we demonstrate results competitive with the state-of-the art for both prediction of human fixations, and segmentation of salient objects. We also characterize different properties of this model including robustness to image transformations, and extension to a wide range of other data types with 3D mesh models serving as an example. Finally, we relate this proposal more generally to the role of saliency computation in visual information processing and draw connections to putative mechanisms for saliency computation in human vision.

3 Semi-supervised Learning with Ladder Networks

Antti Rasmus antti.rasmus@aalto.fi
Harri Valpola harri.valpola@iki.fi
The Curious AI Company
Mathias Berglund mathias.berglund@aalto.fi
Tapani Raiko tapani.raiko@aalto.fi
Aalto University
Mikko Honkala mikko.honkala@nokia.com
Nokia Labs

We combine supervised learning with simultaneous unsupervised learning tasks. The proposed model is trained to simultaneously minimize the sum of supervised and unsupervised cost functions by back-propagation, avoiding the need for layer-wise pretraining. Our work builds on top of the Ladder network proposed by Valpola 2015, which we extend by combining the model with supervised. We show that the resulting model reaches state-of-the-art performance in various tasks: MNIST and CIFAR-10 classification in a semi-supervised setting and permutation invariant MNIST in both semi-supervised and full-labels setting.

4 Enforcing balance allows local supervised learning in spiking recurrent networks

Ralph Bourdoukan ralph.bourdoukan@gmail.com
Sophie Denève sophie.deneve@ens.fr
GNT, Ecole Normale Superieure

To predict sensory inputs or control motor trajectories, the brain must constantly learn temporal dynamics based on error feedback. However, it remains unclear how such supervised learning is implemented in biological neural networks. Learning in recurrent spiking networks is notoriously difficult because local changes in connectivity may have an unpredictable effect on the global dynamics. The most commonly used learning rules, such as temporal back-propagation, are not local and thus not biologically plausible. Furthermore, reproducing the Poisson-like statistics of neural responses requires the use of networks with balanced excitation and inhibition. Such balance is easily destroyed during learning. Using a top-down approach, we show how networks of integrate-and-fire neurons can learn arbitrary linear dynamical systems by feeding back their error as a feed-forward input. The network uses two types of recurrent connections: fast and slow. The fast connections learn to balance excitation and inhibition using a voltage-based plasticity rule. The slow connections are trained to minimize the error feedback using a current-based Hebbian learning rule. Importantly, the balance maintained by fast connections is crucial to ensure that global error signals are available locally in each neuron, in turn resulting in a local learning rule for the slow connections. This demonstrates that spiking networks can learn complex dynamics using purely local learning rules, using E/I balance as the key rather than an additional constraint. The resulting network implements a predictive coding scheme with minimal dimensions and activity in order to implement a given function.

5 Semi-supervised Sequence Learning

Andrew M Dai adai@google.com
Quoc V Le qvl@google.com
Google

We consider the problem of learning long range dependencies with application to document classification. Despite their power in utilizing word ordering, recurrent networks have rarely been used for these applications due to optimization difficulties. In our experiments, we find that it is possible to use long short term memory recurrent networks for document classification with careful tuning. We also find that a simple pretraining step can improve and stabilize the training of recurrent networks. A basic pretraining method is to use a recurrent language model as an initialization of the supervised network. A better pretraining method is a sequence autoencoder, which uses a recurrent network to read a variable length sequence into a vector that can reconstruct the original sequence. The parameters obtained from the pretraining step can then be used as a starting point for other supervised training models. The pretraining methods help because their architectures are designed such that gradients can flow in shorter directions than in other supervised settings. Our experiments show that long short term memory recurrent networks when pretrained with a recurrent language model or a sequence autoencoder become more stable to train. Another important result is that additional unlabeled data in pretraining improves the generalization ability of recurrent networks. With pretraining, we were able to achieve strong performance in many document classification tasks.

6 Skip-Thought Vectors

Ryan Kiros rkiros@cs.toronto.edu
Yukun Zhu yukun@cs.toronto.edu
Russ R Salakhutdinov rsalakhu@cs.toronto.edu
Richard Zemel zemel@cs.toronto.edu
Raquel Urtasun urtasun@cs.toronto.edu
Sanja Fidler fidler@cs.toronto.edu
University of Toronto
Antonio Torralba torralba@mit.edu
MIT

We describe an approach for unsupervised learning of a generic, distributed sentence encoder. Using the continuity of text from books, we train an encoder-decoder model that tries to reconstruct the surrounding sentences of an encoded passage. Sentences that share semantic and syntactic properties are thus mapped to similar vector representations. We next introduce a simple vocabulary expansion method to encode words that were not seen as part of training, allowing us to expand our vocabulary to a million words. After training our model, we extract and evaluate our vectors with linear models on 8 tasks: semantic relatedness, paraphrase detection, image-sentence ranking, question-type classification and 4 benchmark sentiment and subjectivity datasets. The end result is an off-the-shelf encoder that can produce highly generic sentence representations that are robust and perform well in practice. We will make our encoder publicly available.

7 Learning to Linearize Under Uncertainty

Ross Goroshin goroshin@cs.nyu.edu
Michael F Mathieu mathieu@cs.nyu.edu
Yann LeCun yann@cs.nyu.edu
New York University

Training deep feature hierarchies to solve supervised learning tasks has achieving state of the art performance on many problems in computer vision. However, a principled way in which to train such hierarchies in the unsupervised setting has remained elusive. In this work we suggest a new architecture and loss for training deep feature hierarchies that linearize the transformations observed in unlabelednatural video sequences. This is done by training a generative model to predict video frames. We also address the problem of inherent uncertainty in prediction by introducing a latent variables that are non-deterministic functions of the input into the network architecture.

8 Synaptic Sampling: A Bayesian Approach to Neural Network Plasticity and Rewiring

David Kappel david@igi.tugraz.at
Graz University of Technology
Stefan Habenschuss habenschuss@igi.tugraz.at
Robert Legenstein legi@igi.tugraz.at
Wolfgang Maass maass@igi.tugraz.at

We reexamine in this article the conceptual and mathematical framework for understanding the organization of plasticity in spiking neural networks. We propose that inherent stochasticity enables synaptic plasticity to carry out probabilistic inference by sampling from a posterior distribution of synaptic parameters. This view provides a viable alternative to existing models that propose convergence of synaptic weights to maximum likelihood parameters. It explains how priors on weight distributions and connection probabilities can be merged optimally with learned experience. In simulations we show that our model for synaptic plasticity allows spiking neural networks to compensate continuously for unforeseen disturbances. Furthermore it provides a normative mathematical framework to better understand the permanent variability and rewiring observed in brain networks.

9 Natural Neural Networks

Guillaume Desjardins gdesjardins@google.com
Karen Simonyan simonyan@google.com
Razvan Pascanu razp@google.com
koray kavukcuoglu korayk@google.com
Google DeepMind

We introduce Natural Neural Networks a novel family of algorithms that speed up convergence by adapting their internal representation during training to improve conditioning of the Fisher matrix. In particular, we show a specific example that employs a simple and efficient reparametrization of the neural network weights by implicitly whitening the representation obtained at each layer, while preserving the feed-forward computation of the network. Such networks can be trained efficiently via the proposed Projected Natural Gradient Descent algorithm (PNGD), which amortizes the cost of these reparametrizations over many parameter updates and is closely related to the Mirror Descent online learning algorithm. We highlight the benefits of our method on both unsupervised and supervised learning tasks, and showcase its scalability by training on the large-scale ImageNet Challenge dataset.

10 Convolutional Networks on Graphs for Learning Molecular Fingerprints

David K Duvenaud dduvenaud@seas.harvard.edu
Dougal Maclaurin maclaurin@physics.harvard.edu
Jorge Iparraguirre jorgeag@chemistry.harvard.edu
Rafael Bombarell rgbombarelli@chemistry.harvard.edu
Timothy Hirzel hirzel@chemistry.harvard.edu
Alan Aspuru-Guzik aspuru@chemistry.harvard.edu
Ryan P Adams rpa@seas.harvard.edu
Harvard University

Predicting properties of molecules requires functions that take graphs as inputs.Molecular graphs are usually preprocessed using hash-based functions to produce fixed-size fingerprint vectors, which are used as features for making predictions.We introduce a convolutional neural network that operates directly on graphs, allowing end-to-end learning of the feature pipeline. This architecture generalizes standard molecular fingerprints.We show that these data-driven features are more interpretable, and have better predictive performance on a variety of tasks.

11 Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting

Xingjian SHI xshiab@cse.ust.hk
Zhourong Chen zchenbb@cse.ust.hk
Hao Wang hwangaz@cse.ust.hk
Dit-Yan Yeung dyyeung@cse.ust.hk
Wai-kin Wong wkwong@hko.gov.hk
Wang-chun WOO wcwoo@hko.gov.hk
HKUST

The goal of precipitation nowcasting is to predict the future rainfall intensity in a local region over a relatively short period of time. Very few previous studies have examined this crucial and challenging weather forecasting problem from the machine learning perspective. In this paper, we formulate precipitation nowcasting as a spatiotemporal sequence forecasting problem in which both the input and the prediction target are spatiotemporal sequences. By extending the fully connected LSTM (FC-LSTM) to have convolutional structures in both the input-to-state and state-to-state transitions, we propose the convolutional LSTM (ConvLSTM) and use it to build an end-to-end trainable model for the precipitation nowcasting problem. Experiments show that our ConvLSTM network captures spatiotemporal correlations better and consistently outperforms FC-LSTM and the state-of-the-art operational ROVER algorithm for precipitation nowcasting.

12 Scheduled Sampling for Sequence Prediction with Recurrent Neural Networks

Samy Bengio bengio@google.com
Oriol Vinyals vinyals@google.com
Navdeep Jaitly ndjaitly@google.com
Noam Shazeer noam@google.com
Google Research

Recurrent Neural Networks can be trained to produce sequences of tokens given some input, as exemplified by recent results in machine translation and image captioning. The current approach to training them consists of maximizing the likelihood of each token in the sequence given the current (recurrent) state and the previous token. At inference, the unknown previous token is then replaced by a token generated by the model itself. This discrepancy between training and inference can yield errors that can accumulate quickly along the generated sequence. We propose a curriculum learning strategy to gently change the training process from a fully guided scheme using the true previous token, towards a less guided scheme which mostly uses the generated token instead. Experiments on several sequence prediction tasks show that this approach yields significant improvements. Moreover, it was used successfully in our winning bid to the MSCOCO image captioning challenge, 2015.

13 Mind the Gap: A Generative Approach to Interpretable Feature Selection and Extraction

Been Kim beenkim@csail.mit.edu
Julie A Shah julie_a_shah@csail.mit.edu
MIT
Finale Doshi-Velez finale@seas.harvard.edu
Harvard

We present the Mind the Gap Model (MGM), an approach for interpretablefeature extraction and selection. By placing interpretabilitycriteria directly into the model, we allow for the model to bothoptimize parameters related to interpretability and to directly report an interpretable set of results from the final posterior parameter values. MGM extracts distinguishing features on real-worlddatasets of animal features, recipes ingredients, and diseaseco-occurrence. It also performs as well as or better than relatedapproaches on clustering benchmarks.

14 Max-Margin Deep Generative Models

Chongxuan Li licx14@mails.tsinghua.edu.cn
Jun Zhu dcszj@mail.tsinghua.edu.cn
Tianlin Shi stl501@gmail.com
Bo Zhang dcszb@mail.tsinghua.edu.cn
Tsinghua University

Deep generative models (DGMs) are effective on learning multilayered representations of complex data and performing inference of input data by exploring the generative ability. However, little work has been done on examining or empowering the discriminative ability of DGMs on making accurate predictions. This paper presents max-margin deep generative models (mmDGMs), which explore the strongly discriminative principle of max-margin learning to improve the discriminative power of DGMs, while retaining the generative capability. We develop an efficient doubly stochastic subgradient algorithm for the piecewise linear objective. Empirical results on MNIST and SVHN datasets demonstrate that (1) max-margin learning can significantly improve the prediction performance of DGMs and meanwhile retain the generative ability; and (2) mmDGMs are competitive to the state-of-the-art fully discriminative networks by employing deep convolutional neural networks (CNNs) as both recognition and generative models.

15 Cross-Domain Matching for Bag-of-Words Data via Kernel Embeddings of Latent Distributions

Yuya Yoshikawa yoshikawa.yuya.yl9@is.naist.jp
NAIST
Tomoharu Iwata iwata.tomoharu@lab.ntt.co.jp
Hiroshi Sawada sawada.hiroshi@lab.ntt.co.jp
Takeshi Yamada yamada.tak@lab.ntt.co.jp
Nippon Telegraph and Telephone Corporation

We propose a kernel-based method for finding matching between instances across different domains, such as multi-lingual documents and images with annotations. Each instance is assumed to be represented as a set of features, e.g., a bag-of-words representation for documents. The major difficulty in finding cross-domain relationships is that the similarity between instances in different domains cannot be directly measured. To overcome this difficulty, the proposed method embeds all the features of different domains in a shared latent space, and regards each instance as a distribution of its own features in the shared latent space.To represent the distributions efficiently and nonparametrically, we employ the framework of the kernel embeddings of distributions. The embedding is estimated so as to minimize the distance between distributions of paired instances while keeping unpaired instances apart. In our experiments, we show that the proposed method can achieve high performance on finding correspondence between multi-lingual Wikipedia articles, between documents and tags, and between images and tags.

16 A Gaussian Process Model of Quasar Spectral Energy Distributions

Andrew Miller acm@seas.harvard.edu
Albert Wu awu@college.harvard.edu
Ryan P Adams rpa@seas.harvard.edu
Harvard
Jeff Regier jeff@stat.berkeley.edu
Jon McAuliffe jon@stat.berkeley.edu
Berkeley
Dustin Lang dstn@cmu.edu
CMU
Mr. Prabhat prabhat@lbl.gov
David Schlegel djschlegel@lbl.gov
LBL/NERSC

We propose a method for combining two sources of astronomical data, spectroscopy and photometry, which carry information about sources of light (e.g., stars, galaxies, and quasars) at extremely different spectral resolutions. Our model treats the spectral energy distribution (SED) of the radiation from a source as a latent variable, hierarchically generating both photometric and spectroscopic observations. We place a flexible, nonparametric prior over the SED of a light source that admits a physically interpretable decomposition, and allows us to tractably perform inference. We use our model to predict the distribution of the redshift of a quasar from five-band (low spectral resolution) photometric data, the so called ``photo-z'' problem. Our method shows that tools from machine learning and Bayesian statistics allow us to leverage multiple resolutions of information to make accurate predictions with well-characterized uncertainties.

17 Neural Adaptive Sequential Monte Carlo

Shixiang Gu sg717@cam.ac.uk
University of Cambridge / Max Planck Institute
Zoubin Ghahramani zoubin@eng.cam.ac.uk
Richard E Turner ret26@cam.ac.uk
University of Cambridge

Sequential Monte Carlo (SMC), or particle filtering, is a popular class of methods for sampling from an intractable target distribution using a sequence of simpler intermediate distributions. Like other importance sampling-based methods, performance is critically dependent on the proposal distribution: a bad proposal can lead to arbitrarily inaccurate estimates of the target distribution. This paper presents a new method for automatically adapting the proposal using an approximation of the Kullback-Leibler divergence between the true posterior and the proposal distribution. The method is very flexible, applicable to any parameterized proposal distribution and it supports online and batch variants. We use the new framework to adapt powerful proposal distributions with rich parameterizations based upon neural networks leading to Neural Adaptive Sequential Monte Carlo (NASMC). Experiments indicate that NASMC significantly improves inference in a non-linear state space model outperforming adaptive proposal methods including the Extended Kalman and Unscented Particle Filters. Experiments also indicate that improved inference translates into improved parameter learning when NASMC is used as a subroutine of Particle Marginal Metropolis Hastings. Finally we show that NASMC is able to train a neural network-based deep recurrent generative model achieving results that compete with the state-of-the-art for polymorphic music modelling. NASMC can be seen as bridging the gap between adaptive SMC methods and the recent work in scalable, black-box variational inference.

18 Convolutional spike-triggered covariance analysis for neural subunit models

Anqi Wu anqiw@princeton.edu
Jonathan W Pillow pillow@princeton.edu
Princeton University
Memming Park memming.park@stonybrook.edu
Stony Brook University

Subunit models provide a powerful yet parsimonious description of neural spike responses to complex stimuli. They can be expressed by a cascade of two linear-nonlinear (LN) stages, with the first linear stage defined by convolution with one or more filters. Recent interest in such models has surged due to their biological plausibility and accuracy for characterizing early sensory responses. However, fitting subunit models poses a difficult computational challenge due to the expense of evaluating the log-likelihood and the ubiquity of local optima. Here we address this problem by forging a theoretical connection between spike-triggered covariance analysis and nonlinear subunit models. Specifically, we show that a ''convolutional'' decomposition of the spike-triggered average (STA) and covariance (STC) provides an asymptotically efficient estimator for the subunit model under certain technical conditions. We also prove the identifiability of such convolutional decomposition under mild assumptions. Our moment-based methods outperform highly regularized versions of the GQM on neural data from macaque primary visual cortex, and achieves nearly the same prediction performance as the full maximum-likelihood estimator, yet with substantially lower cost.

19 Rectified Factor Networks

Djork-Arné Clevert okko@clevert.de
Andreas Mayr mayr@bioinf.jku.at
Thomas Unterthiner unterthiner@bioinf.jku.at
Sepp Hochreiter hochreit@bioinf.jku.at
Johannes Kepler University Linz

We propose rectified factor networks (RFNs) to efficiently construct very sparse, non-linear, high-dimensional representations of the input. RFN models identify rare and small events, have a low interference between code units, have a small reconstruction error, and explain the data covariance structure. RFN learning is a generalized alternating minimization algorithm derived from the posterior regularization method which enforces non-negative and normalized posterior means. We proof convergence and correctness of the RFN learning algorithm.On benchmarks, RFNs are compared to other unsupervised methods like autoencoders, RBMs, factor analysis, ICA, and PCA. In contrast to previous sparse coding methods, RFNs yield sparser codes, capture the data's covariance structure more precisely, and have a significantly smaller reconstruction error. We test RFNs as pretraining technique of deep networks on different vision datasets, where RFNs were superior to RBMs and autoencoders. On gene expression data from two pharmaceutical drug discovery studies, RFNs detected small and rare gene modules that revealed highly relevant new biological insights which were so far missed by other unsupervised methods.

20 Embed to Control: A Locally Linear Latent Dynamics Model for Control from Raw Images

Manuel Watter watterm@cs.uni-freiburg.de
Jost Springenberg springj@cs.uni-freiburg.de
Joschka Boedecker jboedeck@cs.uni-freiburg.de
University of Freiburg
Martin Riedmiller riedmiller@google.com
Google DeepMind

We introduce Embed to Control (E2C), a method for model learning and control of non-linear dynamical systems from raw pixel images. E2C consists of a deep generative model, belonging to the family of variational autoencoders, that learns to generate image trajectories from a latent space in which the dynamics is constrained to be locally linear. Our model is derived directly from an optimal control formulation in latent space, supports long-term prediction of image sequences and exhibits strong performance on a variety of complex control problems.

21 Bayesian dark knowledge

Anoop Korattikara Balan kbanoop@google.com
Vivek Rathod rathodv@google.com
Kevin P Murphy kpmurphy@google.com
Google
Max Welling m.welling@uva.nl

We consider the problem of Bayesian parameter estimation for deep neural networks, which is important in problem settings where we may have little data, and/ or where we need accurate posterior predictive densities p(y|x, D), e.g., for applications involving bandits or active learning. One simple approach to this is to use online Monte Carlo methods, such as SGLD (stochastic gradient Langevin dynamics). Unfortunately, such a method needs to store many copies of the parameters (which wastes memory), and needs to make predictions using many versions of the model (which wastes time).We describe a method for "distilling" a Monte Carlo approximation to the posterior predictive density into a more compact form, namely a single deep neural network. We compare to two very recent approaches to Bayesian neural networks, namely an approach based on expectation propagation [HLA15] and an approach based on variational Bayes [BCKW15]. Our method performs better than both ofthese, is much simpler to implement, and uses less computation at test time.

22 GP Kernels for Cross-Spectrum Analysis

Kyle R Ulrich kyle.ulrich@duke.edu
David E Carlson david.carlson@duke.edu
Kafui Dzirasakafui. dzirasa@duke.edu
Lawrence Carin lcarin@duke.edu
Duke University

Multi-output Gaussian processes provide a convenient framework for multi-task problems. An illustrative and motivating example of a multi-task problem is multi-region electrophysiological time-series data, where experimentalists are interested in both power and phase coherence between channels. Recently, the spectral mixture (SM) kernel was proposed to model the spectral density of a single task in a Gaussian process framework. In this paper, we develop a novel covariance kernel for multiple outputs, called the cross-spectral mixture (CSM) kernel. This new, flexible kernel represents both the power and phase relationship between multiple observation channels. We demonstrate the expressive capabilities of the CSM kernel through implementation of a Bayesian hidden Markov model, where the emission distribution is a multi-output Gaussian process with a CSM covariance kernel. Results are presented for measured multi-region electrophysiological data.

23 End-to-end Learning of LDA by Mirror-Descent Back Propagation over a Deep Architecture

Jianshu Chen jianshuc@microsoft.com
Yelong Shen yeshen@microsoft.com
Lin Xiao lin.xiao@microsoft.com
Xiaodong He xiaohe@microsoft.com
Jianfeng Gao jfgao@microsoft.com
Xinying Song xinson@microsoft.com
Li Deng deng@microsoft.com
Microsoft Research, Redmond, W
Ji He jvking@uw.edu
University Washington

We develop a fully discriminative learning approach for supervised Latent Dirichlet Allocation (LDA) model (i.e., BP-sLDA), which maximizes the posterior probability of the prediction variable given the input document. Different from traditional variational learning or Gibbs sampling approaches, the proposed learning method applies (i) the mirror descent algorithm for maximum a posterior inference and (ii) back propagation with stochastic gradient/mirror descent for model parameter estimation, leading to scalable and end-to-end discriminative learning of the model. As a byproduct, we also apply this technique to develop a new learning method for the traditional unsupervised LDA model (i.e., BP-LDA). Experimental results on three real-world regression and classification tasks show that the proposed methods significantly outperform the previous supervised topic models, neural networks, and is on par with deep neural networks.

24 Particle Gibbs for Infinite Hidden Markov Models

Nilesh Tripuraneni nt357@cam.ac.uk
Zoubin Ghahramani zoubin@eng.cam.ac.uk
Cambridge University
Shixiang Gu sg717@cam.ac.uk
Cambridge University / MPI
Hong Ge hg344@cam.ac.uk

Infinite Hidden Markov Models (iHMM's) are an attractive, nonparametric generalization of the classical Hidden Markov Model which can automatically infer the number of hidden states in the system. However, due to the infinite-dimensional nature of transition dynamics, performing inference in the iHMM is difficult. In this paper, we present an infinite-state Particle Gibbs (PG) algorithm to resample state trajectories for the iHMM. The proposed algorithm uses an efficient proposal optimized for iHMMs, and leverages ancestor sampling to improve the mixing of the standard PG algorithm. Our algorithm demonstrates significant convergence improvements on synthetic and real world data sets.

25 Sparse Local Embeddings for Extreme Multi-label Classification

Kush Bhatia kushbhatia03@gmail.com
Manik Varma manik@microsoft.com
Prateek Jain prajain@microsoft.com
Microsoft Research
Himanshu Jain himanshu.j689@gmail.com
Puru Kar purushot@cse.iitk.ac.in
Indian Institute of Technology Kanpur

The objective in extreme multi-label learning is to train a classifier that can automatically tag a novel data point with the most relevant subset of labels from an extremely large label set. Embedding based approaches make training and prediction tractable by assuming that the training label matrix is low-rank and hence the effective number of labels can be reduced by projecting the high dimensional label vectors onto a low dimensional linear subspace. Still, leading embedding approaches have been unable to deliver high prediction accuracies or scale to large problems as the low rank assumption is violated in most real world applications.This paper develops the SLEEC classifier to address both limitations. The main technical contribution in SLEEC is a formulation for learning a small ensemble of local distance preserving embeddings which can accurately predict infrequently occurring (tail) labels. This allows SLEEC to break free of the traditional low-rank assumption and boost classification accuracy by learning embeddings which preserve pairwise distances between only the nearest label vectors. We conducted extensive experiments on several real-world as well as benchmark data sets and compare our method against state-of-the-art methods for extreme multi-label classification. Experiments reveal that SLEEC can make significantly more accurate predictions then the state-of-the-art methods including both embeddings (by as much as 35%) as well as trees (by as much as 6%). SLEEC can also scale efficiently to data sets with a million labels which are beyond the pale of leading embedding methods.

26 Robust Spectral Inference for Joint Stochastic Matrix Factorization

Moontae Lee moontae@cs.cornell.edu
David Bindel bindel@cs.cornell.edu
David Mimno mimno@cornell.edu
Cornell University

Spectral inference provides fast algorithms and provable optimality for latent topic analysis. But for real data these algorithms require additional ad-hoc heuristics, and even then often produce unusable results. We explain this poor performance by casting the problem of topic inference in the framework of Joint Stochastic Matrix Factorization (JSMF) and showing that previous methods violate the theoretical conditions necessary for a good solution to exist. We then propose a novel rectification method that learns high quality topics and their interactions even on small, noisy data. This method achieves results comparable to probabilistic techniques in several domains while maintaining scalability and provable optimality.

27 Space-Time Local Embeddings

Ke Sun sunk.edu@gmail.com
Stephane Marchand-Maillet
stephane.marchand-maillet@unige.ch
University of Geneva
Jun Wang jwang1@expedia.com
Expedia, Geneva
Alexandros Kalousis alexandros.kalousis@hesge.ch

Space-time is a profound concept in physics. This concept was shown to be useful for dimensionality reduction. We present basic definitions with interesting counter-intuitions. We give theoretical propositions to show that the space-time is a more powerful representation than an Euclidean space. We apply this concept to manifold learning for preserving local information. Empirical results on non-metric datasets show that more information can be preserved in a space-time.

28 A fast, universal algorithm to learn parametric nonlinear embeddings

Miguel A. Carreira-Perpinan
mcarreira-perpinan@ucmerced.edu
UC Merced
Max Vladymyrov maxv@yahoo-inc.com
Yahoo

Nonlinear embedding algorithms such as stochastic neighbor embedding do dimensionality reduction by optimizing an objective function involving similarities between pairs of input patterns. The result is a low-dimensional projection of each input pattern. A common way to define an out-of-sample mapping is to optimize the objective directly over a parametric mapping of the inputs, such as a neural net. This can be done using the chain rule and a nonlinear optimizer, but is very slow, because the objective involves a quadratic number of terms each dependent on the entire mapping's parameters. Using the method of auxiliary coordinates, we derive a training algorithm that works by alternating steps that train an auxiliary embedding with steps that train the mapping. This has two advantages: 1) The algorithm is universal in that a specific learning algorithm for any choice of embedding and mapping can be constructed by simply reusing existing algorithms for the embedding and for the mapping. A user can then try possible mappings and embeddings with less effort. 2) The algorithm is fast, and it can reuse N-body methods developed for nonlinear embeddings, yielding linear-time iterations.

29 Bayesian Manifold Learning: the Locally Linear Latent Variable Model (LL-LVM)

Mijung Park mijung@gatsby.ucl.ac.uk
Wittawat Jitkrittum wittawatj@gmail.com
Ahmad Qamar atqamar@gmail.com
Zoltan Szabo zoltan.szabo@gatsby.ucl.ac.uk
Lars Buesing lbuesing@gmail.com
Maneesh Sahani maneesh@gatsby.ucl.ac.uk
Gatsby Unit, UCL

We introduce the Locally Linear Latent Variable Model (LL-LVM), a probabilistic model for non-linear manifold discovery that describes a joint distribution over observations, their manifold coordinates and locally linear maps conditioned on a set of neighbourhood relationships. The model allows straightforward variational optimisation of the posterior distribution on coordinates and locally linear maps from the latent space to the observation space given the data. Thus, the LL-LVM encapsulates the local-geometry preserving intuitions that underlie non-probabilistic methods such as locally linear embedding (LLE). Its probabilistic semantics make it easy to evaluate the quality of hypothesised neighbourhood relationships, select the intrinsic dimensionality of the manifold, construct out-of-sample extensions and to combine the manifold model with additional probabilistic models that capture the structure of coordinates within the manifold.

30 Local Causal Discovery of Direct Causes and Effects

Tian Gao gaot@rpi.edu
Qiang Ji qji@ecse.rpi.edu
Rensselaer Polytechnic Institute

We focus on the discovery and identification of direct causes and effects of a target variable in a causal network. State-of-the-art algorithms generally need to find the global causal structures in the form of complete partial directed acyclic graphs in order to identify the direct causes and effects of a target variable. While these algorithms are effective, it is often unnecessary and wasteful to find the global structures when we are only interested in one target variable (such as class labels). We propose a new local causal discovery algorithm, called Causal Markov Blanket (CMB), to identify the direct causes and effects of a target variable based on Markov Blanket Discovery. CMB is designed to conduct causal discovery among multiple variables, but focuses only on finding causal relationships between a specific target variable and other variables. Under standard assumptions, we show both theoretically and experimentally that the proposed local causal discovery algorithm can obtain the comparable identification accuracy as global methods but significantly improve their efficiency, often by more than one order of magnitude.

31 Discriminative Robust Transformation Learning

Jiaji Huang jiaji.huang@duke.edu
Qiang Qiu qiang.qiu@duke.edu
Guillermo Sapiro guillermo.sapiro@duke.edu
Robert Calderbank robert.calderbank@duke.edu
Duke University

This paper proposes a framework for learning features that are robust to data variation, which is particularly important when only a limited number of trainingsamples are available. The framework makes it possible to tradeoff the discriminative value of learned features against the generalization error of the learning algorithm. Robustness is achieved by encouraging the transform that maps data to features to be a local isometry. This geometric property is shown to improve (K, \epsilon)-robustness, thereby providing theoretical justification for reductions in generalization error observed in experiments. The proposed optimization frameworkis used to train standard learning algorithms such as deep neural networks. Experimental results obtained on benchmark datasets, such as labeled faces in the wild,demonstrate the value of being able to balance discrimination and robustness.

32 Max-Margin Majority Voting for Learning from Crowds

TIAN TIAN rossowhite@163.com
Jun Zhu dcszj@mail.tsinghua.edu.cn
Tsinghua University

Learning-from-crowds aims to design proper aggregation strategies to infer the unknown true labels from the noisy labels provided by ordinary web workers. This paper presents max-margin majority voting (M3V) to improve the discriminative ability of majority voting and further presents a Bayesian generalization to incorporate the flexibility of generative methods on modeling noisy observations with worker confusion matrices. We formulate the joint learning as a regularized Bayesian inference problem, where the posterior regularization is derived by maximizing the margin between the aggregated score of a potential true label and that of any alternative label. Our Bayesian model naturally covers the Dawid-Skene estimator and M3V. Empirical results demonstrate that our methods are competitive, often achieving better results than state-of-the-art estimators.

33 M-Best-Diverse Labelings for Submodular Energies and Beyond

Alexander Kirillov alexander.kirillov@tu-dresden.de
Dmytro Shlezinger dmytro.shlezinger@tu-dresden.de
Carsten Rother carsten.rother@tu-dresden.de
Bogdan Savchynskyy bogdan.savchynskyy@tu-dresden.de
TU Dresden
Dmitry P Vetrov vetrovd@yandex.ru
Skoltech, Moscow

We consider the problem of finding M best diverse solutions of energy minimization problems for graphical models. Contrary to the sequential method of Batra et al., which greedily finds one solution after another, we infer all M solutions jointly. It was shown recently that such jointly inferred labelings not only have smaller total energy but also qualitatively outperform the sequentially obtained ones. The only obstacle for using this new technique is the complexity of the corresponding inference problem, since it is considerably slower algorithm than the method of Batra et al. In this work we show that the joint inference of M best diverse solutions can be formulated as a submodular energy minimization if the original MAP-inference problem is submodular, hence fast inference techniques can be used. In addition to the theoretical results we provide practical algorithms that outperform the current state-of-the art and can be used in both submodular and non-submodular case.

34 Covariance-Controlled Adaptive Langevin Thermostat for Large-Scale Bayesian Sampling

Xiaocheng Shang x.shang@ed.ac.uk
Zhanxing Zhu zhanxing.zhu@ed.ac.uk
Benedict Leimkuhler b.leimkuhler@ed.ac.uk
Amos J Storkey a.storkey@ed.ac.uk
University of Edinburgh

Using Monte Carlo sampling for Bayesian posterior inference is a common approach in machine learning. Often the Markov Chain Monte Carlo procedures that are used are discrete-time analogues of associated stochastic differential equations (SDEs). These SDEs are guaranteed to have the required posterior distribution as the invariant distribution. One area of current research asks how to utilise the computational benefits of stochastic gradient methods in this setting. Existing techniques rely on estimating the variance or covariance of the subsampling error, and assume constant variance. We propose a covariance-controlled adaptive Langevin thermostat that can effectively dissipate parameter-dependent noise while maintaining a desired target distribution. This method achieves a substantial speedup over popular alternative schemes for large-scale machine learning applications.

35 Time-Sensitive Recommendation From Recurrent User Activities

Nan Du dunan@gatech.edu
Yichen Wang yichen.wang@gatech.edu
Niao He nhe6@gatech.edu
Jimeng Sun jsun@cc.gatech.edu
Le Song lsong@cc.gatech.edu
Georgia Institute of Technology

Recommender systems, by providing personalized suggestions, are becoming increasingly important for modern web-service providers to improve the engagement of users. However, most recommendation algorithms do not explicitly take into account the temporal behavior and the recurrent activities of users. Two central but less explored questions are \emph{how to recommend the most desirable item at the right moment}, and \emph{how to predict the next returning time of users to our services}.In this paper, we address these questions by building a previously under explored connection between self-exciting point processes and low-rank models to capture the recurrent temporal usage pattern of a large collection of user and item pairs. Furthermore, we develop a new optimization algorithm that maintains $O(1/\epsilon)$ convergence rate, scales up to problems of millions of user-item pairs and thousands of millions of temporal events, and achieves superb predictive performance of solving the two time-sensitive questions compared to other state-of-the-arts on both synthetic and real datasets. Finally, we point out that our formulation can be readily generalized to incorporate other extra context information of users, such as the spatial and textual features, in addition to time.

36 Parallel Recursive Best-First AND/OR Search for Exact MAP Inference in Graphical Models

Akihiro Kishimoto akihirok@ie.ibm.com
Radu Marinescu radu.marinescu@ie.ibm.com
Adi Botea adibotea@ie.ibm.com
IBM Research

The paper presents and evaluates the power of parallel search for exact MAP inference in graphical models. We introduce a new parallel shared-memory recursive best-first AND/OR search algorithm, called SPRBFAOO, that explores the search space in a best-first manner while operating with restricted memory. Our experiments show that SPRBFAOO is often superior to the current state-of-the-art sequential AND/OR search approaches, leading to considerable speed-ups (up to 7-fold with 12 threads), especially on hard problem instances.

37 Logarithmic Time Online Multiclass prediction

Anna E Choromanska achoroma@cims.nyu.edu
Courant Institute, NYU
John Langford jcl@microsoft.com
Microsoft Research New York

We study the problem of multiclass classification with an extremely large number of classes (k), with the goal of obtaining train and test time complexity logarithmic in the number of classes. We develop top-down tree construction approaches for constructing logarithmic depth trees. On the theoretical front, we formulate a new objective function, which is optimized at each node of the tree and creates dynamic partitions of the data which are both pure (in terms of class labels) and balanced. We demonstrate that under favorable conditions, we can construct logarithmic depth trees that have leaves with low label entropy. However, the objective function at the nodes is challenging to optimize computationally. We address the empirical problem with a new online decision tree construction procedure. Experiments demonstrate that this online algorithm quickly achieves improvement in test error compared to more common logarithmic training time approaches, which makes it a plausible method in computationally constrained large-k applications.

38 Scalable Semi-Supervised Aggregation of Classifiers

Akshay Balsubramani abalsubr@cs.ucsd.edu
Yoav Freund yfreund@cs.ucsd.edu
UC San Diego

We present and empirically evaluate an efficient algorithm that learns to aggregate the predictions of an ensemble of binary classifiers. The algorithm uses the structure of the ensemble predictions on unlabeled data to yield significant performance improvements. It does this without making assumptions on the structure or origin of the ensemble, without parameters, and as scalably as linear learning. We empirically demonstrate these performance gains with random forests.

39 Bounding the Cost of Search-Based Lifted Inference

David B Smith dbs014200@utdallas.edu
Vibhav G Gogate vgogate@hlt.utdallas.edu
University of Texas at Dallas

Recently, there has been growing interest in systematic search-based and importance sampling-based lifted inference algorithms for statistical relational models (SRMs). These lifted algorithms achieve significant complexity reductions over their propositional counterparts by using lifting rules that leverage symmetries in the relational representation. One drawback of these algorithms is that they use an inference-blind representation of the search space, which makes it difficult to efficiently pre-compute tight upper bounds on the exact cost of inference without running the algorithm to completion. In this paper, we present a principled approach to address this problem. We introduce a lifted analogue of the propositional And/Or search space framework, which we call a lifted And/Or schematic. Given a schematic-based representation of an SRM, we show how to efficiently compute a tight upper bound on the time and space cost of exact inference from a current assignment and the remaining schematic. We show how our bounding method can be used within a lifted importance sampling algorithm, in order to perform effective Rao-Blackwellisation, and demonstrate experimentally that the Rao-Blackwellised version of the algorithm yields more accurate estimates on several real-world datasets.

40 Efficient Learning by Directed Acyclic Graph For Resource Constrained Prediction

Joseph Wang joewang@bu.edu
Venkatesh Saligrama srv@bu.edu
Boston University
Kirill Trapeznikov kirill.trapeznikov@stresearch.com
STR

We study the problem of reducing test-time acquisition costs in classification systems. Our goal is to learn decision rules that adaptively select sensors for each example as necessary to make a confident prediction. We model our system as a directed acyclic graph (DAG) where internal nodes correspond to sensor subsets and decision functions at each node choose whether to acquire a new sensor or classify using the available measurements. This problem can be naturally posed as an empirical risk minimization over training data. Rather than jointly optimizing such a highly coupled and non-convex problem over all decision nodes, we propose an efficient algorithm motivated by dynamic programming. We learn node policies in the DAG by reducing the global objective to a series of cost sensitive learning problems. Our approach is computationally efficient and has proven guarantees of convergence to the optimal system for a fixed architecture. In addition, we present an extension to map other budgeted learning problems with large number of sensors to our DAG architecture and demonstrate empirical performance exceeding state-of-the-art algorithms for data composed of both few and many sensors.

41 Estimating Jaccard Index with Missing Observations: A Matrix Calibration Approach

Wenye Li wyli@ipm.edu.mo
Macao Polytechnic Institute

The Jaccard index is a standard statistics for comparing the pairwise similarity between data samples. We investigated the problem of estimating a Jaccard index matrix when the samples have missing observations. Our proposed method is to firstly approximate a Jaccard index matrix with the incomplete data, and next calibrate the matrix to satisfy the requirement of positive semi-definiteness. The calibration problem is convex and the optimal solution to it can be found effectively by a simple yet effective alternating projection algorithm. Compared with conventional imputation approaches that try to replace the missing observations with substituted values and then calculate the similarity matrix, our method has a strong advantage in that the calibrated matrix is guaranteed to be closer to the unknown true matrix in Frobenius norm than the un-calibrated matrix (except in special cases they are identical). Our method reported practical improvement in empirical evaluations. The improvement is especially significant when there are highly missing observations.

42 Sample Efficient Path Integral Control under Uncertainty

Yunpeng Pan ypan37@gatech.edu
Evangelos Theodorou
evangelos.theodorou@ae.gatech.edu
Michail Kontitsis kontitsis@gatech.edu
Georgia Institute of Technolog

We present a data-driven optimal control framework that can be viewed as a generalization of the path integral (PI) control approach. We find iterative feedback control laws without parameterization based on probabilistic representation of learned dynamics model. The proposed algorithm operates in a forward-backward manner which differentiate from other PI-related methods that perform forward sampling to find optimal controls. Our method uses significantly less samples to find optimal controls compared to other approaches within the PI control family that relies on extensive sampling from given dynamics models or trials on physical systems in model-free fashions. In addition, the learned controllers can be generalized to new tasks without re-sampling based on the compositionality theory for the linearly-solvable optimal control framework.We provide experimental results on three different systems and comparisons with state-of-the-art model-based methods to demonstrate the efficiency and generalizability of the proposed framework.

43 Efficient Thompson Sampling for Online Matrix-Factorization Recommendation

Jaya Kawale kawale@adobe.com
Hung H Bui hubui@adobe.com
Branislav Kveton kveton@adobe.com
Adobe Research
Long Tran-Thanh ltt08r@ecs.soton.ac.uk
University of Southampton
Sanjay Chawla sanjay.chawla@sydney.edu.au
Qatar Computing Research, University of Sydney

Matrix factorization (MF) collaborative filtering is an effective and widely used method in recommendation systems. However,

the problem of finding an optimal trade-off between exploration and exploitation (otherwise known as the bandit problem), a crucial problem in collaborative filtering from cold-start, has not been previously addressed.In this paper, we present a novel algorithm for online MF recommendation that automatically combines finding the most relevantitems with exploring new or less-recommended items.Our approach, called Particle Thompson Sampling for Matrix-Factorization, is based on the general Thompson sampling framework, but augmented with a novel efficient online Bayesian probabilistic matrix factorization method based on the Rao-Blackwellized particle filter.Extensive experiments in collaborative filtering using several real-world datasets demonstrate that our proposed algorithm significantly outperforms the current state-of-the-arts.

44 Parallelizing MCMC with Random Partition Trees

Xiangyu Wang xw56@stat.duke.edu
Richard Guo guo@cs.duke.edu
Katherine Heller kheller@stat.duke.edu
David B Dunson dunson@stat.duke.edu
Duke University

The modern scale of data has brought new challenges to Bayesian inference. In particular, conventional MCMC algorithms are computationally very expensive for large data sets. A promising approach to solve this problem is embarrassingly parallel MCMC (EP-MCMC), which first partitions the data into multiple subsets and runs independent sampling algorithms on each subset. The subset posterior draws are then aggregated via some combining rules to obtain the final approximation. Existing EP-MCMC algorithms are limited by approximation accuracy and difficulty in resampling. In this article, we propose a new EP-MCMC algorithm PART that solves these problems. The new algorithm applies random partition trees to combine the subset posterior draws, which is distribution-free, easy to resample from and can adapt to multiple scales. We provide theoretical justification and extensive experiments illustrating empirical performance.

45 Fast Lifted MAP Inference via Partitioning

Somdeb Sarkhel sxs104721@utdallas.edu
Vibhav G Gogate vgogate@hlt.utdallas.edu
University of Texas at Dallas
Parag Singla parags@cse.iitd.ac.in
Indian Institute of Technology

Recently, there has been growing interest in lifting MAP inference algorithms for Markov logic networks (MLNs). A key advantage of these lifted algorithms is that they have much smaller computational complexity than propositional algorithms when symmetries are present in the MLN and these symmetries can be detected using lifted inference rules. Unfortunately, lifted inference rules are sound but not complete and can often miss many symmetries. This is problematic because when symmetries cannot be exploited, lifted inference algorithms ground the MLN, and search for solutions in the much larger propositional space. In this paper, we present a novel approach, which cleverly introduces new symmetries at the time of grounding. Our main idea is to partition the ground atoms and force the inference algorithm to treat all atoms in each part as indistinguishable. We show that by systematically and carefully refining (and growing) the partitions, we can build advanced any-time and any-space MAP inference algorithms. Our experiments on several real-world datasets clearly show that our new algorithm is superior to previous approaches and often finds useful symmetries in the search space that existing lifted inference rules are unable to detect.

46 Active Learning from Weak and Strong Labelers

Chicheng Zhang chz038@cs.ucsd.edu
Kamalika Chaudhuri kamalika@cs.ucsd.edu
UC San Diego

An active learner is given a hypothesis class, a large set of unlabeled examples and the ability to interactively query labels to an oracle of a subset of these examples; the goal of the learner is to learn a hypothesis in the class that fits the data well by making as few label queries as possible.This work addresses active learning with labels obtained from strong and weak labelers, where in addition to the standard active learning setting, we have an extra weak labeler which may occasionally provide incorrect labels. An example is learning to classify medical images where either expensive labels may be obtained from a physician (oracle or strong labeler), or cheaper but occasionally incorrect labels may be obtained from a medical resident (weak labeler). Our goal is to learn a classifier with low error on data labeled by the oracle, while using the weak labeler to reduce the number of label queries made to this labeler. We provide an active learning algorithm for this setting, establish its statistical consistency, and analyze its label complexity to characterize when it can provide label savings over using the strong labeler alone.

47 Fast and Guaranteed Tensor Decomposition via Sketching

Yining Wang yiningwa@cs.cmu.edu
Hsiao-Yu Tung htung@cs.cmu.edu
Alex J Smola alex@smola.org
Carnegie Mellon University
Anima Anandkumar a.anandkumar@uci.edu
UC Irvine

Tensor CANDECOMP/PARAFAC (CP) decomposition has wide applications in statistical learning of latent variable models and in data mining. In this paper, we propose fast and randomized tensor CP decomposition algorithms based on sketching. We build on the idea of count sketches, but introduce many novel ideas which are unique to tensors. We develop novel methods for randomized com- putation of tensor contractions via FFTs, without explicitly forming the tensors. Such tensor contractions are encountered in decomposition methods such as ten- sor power iterations and alternating least squares. We also design novel colliding hashes for symmetric tensors to further save time in computing the sketches. We then combine these sketching ideas with existing whitening and tensor power iter- ative techniques to obtain the fastest algorithm on both sparse and dense tensors. The quality of approximation under our method does not depend on properties such as sparsity, uniformity of elements, etc. We apply the method for topic mod- eling and obtain competitive results.

48 Spherical Random Features for Polynomial Kernels

Jeffrey Pennington jpennin@google.com
Felix Yu felixyu@google.com
Sanjiv Kumar sanjivk@google.com
Google Research

Compact explicit feature maps provide a practical framework to scale kernel methods to large-scale learning, but deriving such maps for many types of kernels remains a challenging open problem. Among the commonly used kernels for nonlinear classification are polynomial kernels, for which low approximation error has thus far necessitated explicit feature maps of large dimensionality, especially for higher-order polynomials. Meanwhile, because polynomial kernels are unbounded, they are frequently applied to data that has been normalized to unit l2 norm. The question we address in this work is: if we know a priori that data is so normalized, can we devise a more compact map? We show that a putative affirmative answer to this question based on Random Fourier Features is impossible in this setting, and introduce a new approximation paradigm, Spherical Random Fourier (SRF) features, which circumvents these issues and delivers a compact approximation to polynomial kernels for data on the unit sphere. Compared to prior work, SRF features are less rank-deficient, more compact, and achieve better kernel approximation, especially for higher-order polynomials. The resulting predictions have lower variance and typically yield better classification accuracy.

49 Learnability of Influence in Networks

Harikrishna Narasimhan hnarasimhan@g.harvard.edu
David C Parkes parkes@eecs.harvard.edu
Yaron Singer yaron@seas.harvard.edu
Harvard University

We establish PAC learnability of influence functions for three common influence models, namely, the Linear Threshold (LT), Independent Cascade (IC) and Voter models, and present concrete sample complexity results in each case. Our results for the LT model are based on interesting connections with VC-dimension of neural networks; those for the IC model are based an interpretation of the influence function as an expectation over random draw of a subgraph; and those for the Voter model are based on a reduction to linear regression. We show these results for the case in which cascades are only partially observed and we do not see the time steps in which a node has been infected. We also provide efficient polynomial time learning algorithms for the case in which the observed cascades contain the time steps in which nodes are influenced.

50 A Pseudo-Euclidean Iteration for Optimal Recovery in Noisy ICA

James R Voss vossj@cse.ohio-state.edu
Mikhail Belkin mbelkin@cse.ohio-state.edu
Luis Rademacher lrademac@cse.ohio-state.edu
Ohio State University

Independent Component Analysis (ICA) is a popular model for blind signal separation. The ICA model assumes that a number of independent source signals are linearly mixed to form the observed signals. We propose a new algorithm, PEGI (for pseudo-Euclidean Gradient Iteration), for provable model recovery for ICA with Gaussian noise. The main technical innovation of the algorithm is to use a fixed point iteration in a pseudo-Euclidean (indefinite “inner product”) space. The use of this indefinite “inner product” resolves technical issues common to several existing algorithms for noisy ICA. This leads to an algorithm which is conceptually simple, efficient and accurate in testing.Our second contribution is combining PEGI with the analysis of objectives for optimal recovery in the noisy ICA model. It has been observed that the direct approach of demixing with the inverse of the mixing matrix is suboptimal for signal recovery in terms of the natural Signal to Interference plus Noise Ratio (SINR) criterion. There have been several partial solutions proposed in the ICA literature. It turns out that any solution to the mixing matrix reconstruction problem can be used to construct an SINR-optimal ICA demixing, despite the fact that SINR itself cannot be computed from data. That allows us to obtain a practical and provably SINR-optimal recovery method for ICA with arbitrary Gaussian noise.

51 Differentially private subspace clustering

Yining Wang yiningwa@cs.cmu.edu
Yu-Xiang Wang yuxiangw@cs.cmu.edu
Aarti Singh aartisingh@cmu.edu
Carnegie Mellon University

Subspace clustering is an unsupervised learning problem that aims at grouping data points into multiple ``clusters'' so that data points in a single cluster lie approximately on a low-dimensional linear subspace. It is originally motivated by 3D motion segmentation in computer vision, but has recently been generically applied to a wide range of statistical machine learning problems, which often involves sensitive datasets about human subjects. This raises a dire concern for data privacy. In this work, we build on the framework of ``differential privacy'' and present two provably private subspace clustering algorithms. We demonstrate via both theory and experiments that one of the presented methods enjoys formal privacy and utility guarantees; the other one asymptotically preserves differential privacy while having good performance in practice. Along the course of the proof, we also obtain two new provable guarantees for the agnostic subspace clustering and the graph connectivity problem which might be of independent interests.

52 Compressive spectral embedding: sidestepping the SVD

Dinesh Ramasamy dineshr@ece.ucsb.edu
Upamanyu Madhow madhow@ece.ucsb.edu
UC Santa Barbara

Spectral embedding based on the Singular Value Decomposition (SVD) is a widely used “preprocessing” step in many learning tasks, typically leading to dimensionality reduction by projecting onto a number of dominant singular vectors and rescaling the coordinate axes (by a predefined function of the singular value). However, the number of such vectors required to capture problem structure grows with problem size, and even partial SVD computation becomes a bottleneck. In this paper, we propose a low-complexity it compressive spectral embedding algorithm, which employs random projections and finite order polynomial expansions to compute approximations to SVD-based embedding. For an m times n matrix with T non-zeros, its time complexity is O((T+m+n)log(m+n)), and the embedding dimension is O(log(m+n)), both of which are independent of the number of singular vectors whose effect we wish to capture. To the best of our knowledge, this is the first work to circumvent

this dependence on the number of singular vectors for general SVD-based embeddings. The key to sidestepping the SVD is the observation that, for downstream inference tasks such as clustering and classification, we are only interested in using the resulting embedding to evaluate pairwise similarity metrics derived from the euclidean norm, rather than capturing the effect of the underlying matrix on arbitrary vectors as a partial SVD tries to do. Our numerical results on network datasets demonstrate the efficacy of the proposed method, and motivate further exploration of its application to large-scale inference tasks.

53 Generalization in Adaptive Data Analysis and Holdout Reuse

Cynthia Dwork dwork@microsoft.com
Microsoft Research
Vitaly Feldman vitaly.edu@gmail.com
IBM Research - Almaden
Moritz Hardt m@mrtz.org
Google
Toni Pitassi toni@cs.toronto.edu
University of Toronto
Omer Reingold omer.reingold@gmail.com
Samsung Research
Aaron Roth aaroth@cis.upenn.edu
University of Pennsylvania

Overfitting is the bane of data analysts, even when data are plentiful. Formal approaches to understanding this problem focus on statistical inference and generalization of individual analysis procedures. Yet the practice of data analysis is an inherently interactive and adaptive process: new analyses and hypotheses are proposed after seeing the results of previous ones, parameters are tuned on the basis of obtained results, and datasets are shared and reused. An investigation of this gap has recently been initiated in (Dwork et al., 2014), who focused on the problem of estimating expectations of adaptively chosen functions.In this paper, we give a simple and practical method for reusing a holdout (or testing) set to validate the accuracy of hypotheses produced by a learning algorithm operating on a training set. Reusing a holdout set adaptively multiple times can easily lead to overfitting to the holdout set itself. We give an algorithm that enables the validation of a large number of adaptively chosen hypotheses, while provably avoiding overfitting. We illustrate the advantages of our algorithm over the standard use of the holdout set via a simple synthetic experiment.We also formalize and address the general problem of data reuse in adaptive data analysis. We show how the differential-privacy based approach in (Dwork et al., 2014) is applicable much more broadly to adaptive data analysis. We then show that a simple approach based on description length can also be used to give guarantees of statistical validity in adaptive settings. Finally, we demonstrate that these incomparable approaches can be unified via the notion of approximate max-information that we introduce. This, in particular, allows the preservation of statistical validity guarantees even when an analyst adaptively composes algorithms which have guarantees based on either of the two approaches.

54 Online F-Measure Optimization

Róbert Busa-Fekete busarobi@gmail.com
UPB
Balázs Szörényi szorenyi@inf.u-szeged.hu
The Technion / University of Szeged
Krzysztof Dembczynski
krzysztof.dembczynski@cs.put.poznan.pl
Poznan University of Technology
Eyke Hüllermeier eyke@upb.de
Marburg university

The F-measure is an important and commonly used performance metric for binary prediction tasks. By combining precision and recall into a single score, it avoids disadvantages of simple metrics like the error rate, especially in cases of imbalanced class distributions. The problem of optimizing the F-measure, that is, of developing learning algorithms that perform optimally in the sense of this measure, has recently been tackled by several authors. In this paper, we study the problem of F-measure maximization in the setting of online learning. We propose an efficient online algorithm and provide a formal analysis of its convergence properties. Moreover, first experimental results are presented, showing that our method performs well in practice.

55 Matrix Completion with Noisy Side Information

Kai-Yang Chiang kychiang@cs.utexas.edu
Inderjit S Dhillon inderjit@cs.utexas.edu
University of Texas at Austin
Cho-Jui Hsieh cjhsieh@cs.utexas.edu
UC Davis

We study matrix completion problem with side information. Side information has been considered in several matrix completion applications, and is generally shown to be useful empirically. Recently, Xu et al. studied the effect of side information for matrix completion under a theoretical viewpoint, showing that sample complexity can be significantly reduced given completely clean features. However, since in reality most given features are noisy or even weakly informative, how to develop a general model to handle general feature set, and how much the noisy features can help matrix recovery in theory, is still an important issue to investigate. In this paper, we propose a novel model that balances between features and observations simultaneously, enabling us to leverage feature information yet to be robust to feature noise. Moreover, we study the effectof general features in theory, and show that by using our model, the sample complexity can still be lower than matrix completion as long as features are sufficiently informative. This result provides a theoretical insight of usefulness for general side information. Finally, we consider synthetic data and two real applications - relationship prediction and semi-supervised clustering, showing that our model outperforms other methods for matrix completion with features both in theory and practice.

56 A Market Framework for Eliciting Private Data

Bo Waggoner bwaggoner@fas.harvard.edu
Harvard
Rafael Frongillo raf@cs.berkeley.edu
CU Boulder
Jacob D Abernethy jabernet@umich.edu
University of Michigan

We propose a mechanism for purchasing information from a sequence of participants.The participants may simply hold data points they wish to sell, or may have more sophisticated information; either way, they are incentivized to participate as long as they believe their data points are representative or their information will improve the mechanism's future prediction on a test set.The mechanism, which draws on the principles of prediction markets, has a bounded budget and minimizes generalization error for Bregman divergence loss functions.We then show how to modify this mechanism to preserve the privacy of participants' information: At any given time, the current prices and predictions of the mechanism reveal almost no information about any one participant, yet in total over all participants, information is accurately aggregated.

57 Optimal Ridge Detection using Coverage Risk

Yen-Chi Chen ga014528@gmail.com
Christopher Genovese genovese@stat.cmu.edu
Shirley Ho shirleyh@andrew.cmu.edu
Larry Wasserman larry@stat.cmu.edu
Carnegie Mellon University

We introduce the concept of coverage risk as an error measure for density ridge estimation.The coverage risk generalizes the mean integrated square error to set estimation.We propose two risk estimators for the coverage risk and we show that we can select tuning parameters by minimizing the estimated risk. We study the rate of convergence for coverage risk and prove consistency of the risk estimators.We apply our method to three simulated datasets and to cosmology data.In all the examples, the proposed method successfully recover the underlying density structure.

58 Fast Distributed k-Center Clustering with Outliers on Massive Data

Gustavo Malkomes luizgustavo@wustl.edu
Matt J Kusner mkusner@wustl.edu
Wenlin Chen wenlinchen@wustl.edu
Kilian Q Weinberger kilian@wustl.edu
Benjamin Moseley bmoseley@wustl.edu
Washington University in St Louis

Clustering large data is a fundamental problem with a vast number of applications. Due to the increasing size of data, practitioners interested in clustering have turned to distributed computation methods. In this work, we consider the widely used k-center clustering problem and its variant used to handle noisy data, k-center with outliers. In the noise-free setting we demonstrate how a previously-proposed distributed method is actually an O(1)-approximation algorithm, which accurately explains its strong empirical performance. Additionally, in the noisy setting, we develop a novel distributed algorithm that is also an O(1)-approximation. These algorithms are highly parallel and lend themselves to virtually any distributed computing framework. We compare both empirically against the best known noisy sequential clustering methods and show that both distributed algorithms are consistently close to their sequential versions. The algorithms are all one can hope for in distributed settings: they are fast, memory efficient and they match their sequential counterparts.

59 Orthogonal NMF through Subspace Exploration

Megasthenis Asteris megas@utexas.edu
Alex G Dimakis dimakis@austin.utexas.edu
University of Texas at Austin
Dimitris Papailiopoulos dimitrisp@berkeley.edu
UC Berkeley

Orthogonal Nonnegative Matrix Factorization {(ONMF)} aims to approximate a nonnegative matrix as the product of two k-dimensional nonnegative factors, one of which has orthonormal columns. It yields potentially useful data representations as superposition of disjoint parts, while it has been shown to work well for clustering tasks where traditional methods underperform. Existing algorithms rely mostly on heuristics, which despite their good empirical performance, lack provable performance guarantees.We present a new ONMF algorithm with provable approximation guarantees.For any constant dimension~k, we obtain an additive EPTAS without any assumptions on the input. Our algorithm relies on a novel approximation to the related Nonnegative Principal Component Analysis (NNPCA) problem; given an arbitrary data matrix, NNPCA seeks k nonnegative components that jointly capture most of the variance. Our NNPCA algorithm is of independent interest and generalizes previous work that could only obtain guarantees for a single component. We evaluate our algorithms on several real and synthetic datasets and show that their performance matches or outperforms the state of the art.

60 Fast Classification Rates for High-dimensional Gaussian Generative Models

Tianyang Li lty@cs.utexas.edu
Adarsh Prasad adarsh@cs.utexas.edu
Pradeep K Ravikumar pradeepr@cs.utexas.edu
University of Texas at Austin

We consider the problem of binary classification when the covariates conditioned on the each of the response values follow multivariate Gaussian distributions. We focus on the setting where the covariance matrices for the two conditional distributions are the same. The corresponding generative model classifier, derived via the Bayes rule, also called Linear Discriminant Analysis, has been shown to behave poorly in high-dimensional settings. We present a novel analysis of the classification error of any linear discriminant approach given conditional Gaussian models. This allows us to compare the generative model classifier, other recently proposed discriminative approaches that directly learn the discriminant function, and then finally logistic regression which is another classical discriminative model classifier. As we show, under a natural sparsity assumption, and letting s denote the sparsity of the Bayes classifier, p the number of covariates, and n the number of samples, the simple (ℓ1-regularized) logistic regression classifier achieves the fast misclassification error rates of O(slogpn), which is much better than the other approaches, which are either inconsistent under high-dimensional settings, or achieve a slower rate of O(slogpn−−−−√).

61 Efficient and Parsimonious Agnostic Active Learning

T.-K. Huang tkhuang@microsoft.com
Alekh Agarwal alekha@microsoft.com
John Langford jcl@microsoft.com
Robert Schapire schapire@microsoft.com
Microsoft Research
Daniel J Hsu danielhsu@gmail.com
Columbia University

We develop a new active learning algorithm for the streaming setting satisfying three important properties: 1) It provably works for any classifier representation and classification problem including those with severe noise. 2) It is efficiently implementable with an ERM oracle. 3) It is more aggressive than all previous approaches satisfying 1 and 2. To do this we create an algorithm based on a newly defined optimization problem and analyze it. We also conduct the first experimental analysis of all efficient agnostic active learning algorithms, discovering that this one is typically better across a wide variety of datasets and label complexities.

62 Collaborative Filtering with Graph Information: Consistency and Scalable Methods

Nikhil Rao nikhilr@cs.utexas.edu
Hsiang-Fu Yu rofuyu@cs.utexas.edu
Inderjit S Dhillon inderjit@cs.utexas.edu
Pradeep K Ravikumar pradeepr@cs.utexas.edu
University of Texas at Austin

Low rank matrix completion plays a fundamental role in collaborative filtering applications, the key idea being that the variables lie in a smaller subspace than the ambient space. Often, additional information about the variables is known, and it is reasonable to assume that incorporating this information will lead to better predictions. We tackle the problem of matrix completion when the variables are related to each other via a graph. We formulate and derive an efficient alternating minimization scheme that solves optimizations with over 15 million observations up to 2 orders of magnitude faster than SGD based methods. On the theoretical front, we show that such methods generalize weighted nuclear norm formulations, and derive statistical consistency guarantees. We validate our results on both real world and synthetic datasets.

63 Less is More: Nyström Computational Regularization

Alessandro Rudi ale_rudi@mit.edu
MIT
Raffaello Camoriano raffaello.camoriano@iit.it
IIT - UNIGE
Lorenzo Rosasco lrosasco@mit.edu
University of Genova

We study Nyström type subsampling approaches to large scale kernel methods, and prove learning bounds in the statistical learning setting, where random sampling and high probability estimates are considered. In particular, we prove that these approaches can achieve optimal learning bounds, provided the subsampling level is suitably chosen. These results suggest a simple incremental variant of Nyström kernel ridge regression, where the subsampling level controls at the same time regularization and computations. Extensive experimental analysis shows that the considered approach achieves state of the art performances on benchmark large scale datasets.

64 Predtron: A Family of Online Algorithms for General Prediction Problems

Prateek Jain prajain@microsoft.com
Microsoft Research
Nagarajan Natarajan naga86@cs.utexas.edu
UT Austin
Ambuj Tewari tewaria@umich.edu
University of Michigan

Modern prediction problems arising in multilabel learning and learning to rank pose unique challenges to the classical theory of supervised learning. These problems have large prediction and label spaces of a combinatorial nature and involve sophisticated loss functions. We offer a general framework to derive mistake driven online algorithms and associated loss bounds. The key ingredients in our framework are a general loss function, a general vector space representation of predictions, and a notion of margin with respect to a general norm. Our general algorithm, Predtron, yields the perceptron algorithm and its variants when instantiated on classic problems such as binary classification, multiclass classification, ordinal regression, and multilabel classification. For multilabel ranking and subset ranking, we derive novel algorithms, notions of margins, and loss bounds. A simulation study confirms the behavior predicted by our bounds and demonstrates the flexibility of the design choices in our framework.

65 On the Optimality of Classifier Chain for Multi-label Classification

Weiwei Liu liuweiwei863@gmail.com
Ivor Tsang ivor.tsang@uts.edu.au
University of Technology, Sydney

To capture the interdependencies between labels in multi-label classification problems, classifier chain (CC) tries to take the multiple labels of each instance into account under a deterministic high-order Markov Chain model. Since its performance is sensitive to the choice of label order, the key issue is how to determine the optimal label order for CC. In this work, we first generalize the CC model over a random label order. Then, we present a theoretical analysis of the generalization error for the proposed generalized model. Based on our results, we propose a dynamic programming based classifier chain (CC-DP) algorithm to search the globally optimal label order for CC and a greedy classifier chain (CC-Greedy) algorithm to find a locally optimal CC. Comprehensive experiments on a number of real-world multi-label data sets from various domains demonstrate that our proposed CC-DP algorithm outperforms state-of-the-art approaches and the CC-Greedy algorithm achieves comparable prediction performance with CC-DP.

66 Smooth Interactive Submodular Set Cover

Bryan D He bryanhe@stanford.edu
Stanford University
Yisong Yue yyue@caltech.edu
Caltech

Interactive submodular set cover is an interactive variant of submodular set cover over a hypothesis class of submodular functions, where the goal is to satisfy all sufficiently plausible submodular functions to a target threshold using as few (cost-weighted) actions as possible. It models settings where there is uncertainty regarding which submodular function to optimize. In this paper, we propose a new extension, which we call smooth interactive submodular set cover} that allows the target threshold to smoothly vary depending on the plausibility of each hypothesis. We present the first algorithm for this more general setting with theoretical guarantees on optimality. We further show how to extend our approach to deal with real-valued functions, which yields new theoretical results for real-valued submodular set cover for both the interactive and non-interactive settings.

67 Tractable Bayesian Network Structure Learning with Bounded Vertex Cover Number

Janne H Korhonen janne.h.korhonen@helsinki.fi
University of Helsinki
Pekka Parviainen pekka.parviainen@aalto.fi
Aalto University

Both learning and inference tasks on Bayesian networks are NP-hard in general. Bounded tree-width Bayesian networks have recently received a lot of attention as a way to circumvent this complexity issue; however, while inference on bounded tree-width networks is tractable, the learning problem remains NP-hard even for tree-width~2. In this paper, we propose bounded vertex cover number Bayesian networks as an alternative to bounded tree-width networks. In particular, we show that both inference and learning can be done in polynomial time for any fixed vertex cover number bound k, in contrast to the general and bounded tree-width cases; on the other hand, we also show that learning problem is W[1]-hard in parameter k. Furthermore, we give an alternative way to learn bounded vertex cover number Bayesian networks using integer linear programming (ILP), and show this is feasible in practice.

68 Secure Multi-party Differential Privacy

Peter Kairouz kairouz2@illinois.edu
Sewoong Oh swoh@illinois.edu
Pramod Viswanath pramodv@illinois.edu
UIUC

We study the problem of multi-party interactive function computation under differential privacy. In this setting, each party is interested in computing a function on its private bit and all the other parties' bits. The function to be computed can vary from one party to the other. Moreover, there could be a central observer who is interested in computing a separate function on all the parties' bits. Differential privacy ensures that there remains an uncertainty in any party's bit even when given the transcript of interactions and all other parties' bits. Performance at each party is measured via the accuracy of the function to be computed. We allow for an arbitrary cost metric to measure the distortion between the true and the computed function values. Our main result is the optimality of a simple non-interactive protocol: each party randomizes its bit (sufficiently) and shares the privatized version with the other parties. This optimality result is very general: it holds for all types of functions, heterogeneous privacy conditions on the parties, all types of cost metrics, and both average and worst-case (over the inputs) measures of accuracy.

69 Adaptive Stochastic Optimization: From Sets to Paths

Zhan Wei Lim zhanweiz@gmail.com
David Hsu dyhsu@comp.nus.edu.sg
Wee Sun Lee leews@comp.nus.edu.sg
National University of Singapore

Adaptive stochastic optimization optimizes an objective function adaptivelyunder uncertainty. Adaptive stochastic optimization plays a crucial role inplanning and learning under uncertainty, but is, unfortunately,computationally intractable in general. This paper introduces two conditionson the objective function, the marginal likelihood rate bound and the marginallikelihood bound, which enable efficient approximate solution of adaptivestochastic optimization. Several interesting classes of functions satisfythese conditions naturally, e.g., the version space reduction function forhypothesis learning. We describe Recursive Adaptive Coverage (RAC), a newadaptive stochastic optimization algorithm that exploits these conditions, andapply it to two planning tasks under uncertainty. In constrast to the earliersubmodular optimization approach, our algorithm applies to adaptive stochastic optimization algorithm over both sets and paths.

70 Learning structured densities via infinite dimensional exponential families

Siqi Sun siqi.sun@ttic.edu
Jinbo Xu jinbo.xu@gmail.com
Toyota Technological Institute at Chicago
mladen kolar mkolar@gmail.com
University of Chicago Booth School of Business

Learning the structure of a probabilistic graphical models is a well studied problem in the machine learning community due to its importance in many applications. Current approaches are mainly focused on learning the structure under restrictive parametric assumptions, which limits the applicability of these methods. In this paper, we study the problem of estimating the structure of a probabilistic graphical model without assuming a particular parametric model. We consider probabilities that are members of an infinite dimensional exponential family, which is parametrized by a reproducing kernel Hilbert space (RKHS) H and its kernel k. One difficulty in learning nonparametric densities is evaluation of the normalizing constant. In order to avoid this issue, our procedure minimizes the penalized score matching objective. We show how to efficiently minimize the proposed objective using existing group lasso solvers. Furthermore, we prove that our procedure recovers the graph structure with high-probability under mild conditions. Simulation studies illustrate ability of our procedure to recover the true graph structure without the knowledge of the data generating process.

71 Lifelong Learning with Non-i.i.d. Tasks

Anastasia Pentina apentina@ist.ac.at
Christoph H Lampert chl@ist.ac.at
IST Austria

In this work we aim at extending theoretical foundations of lifelong learning. Previous work analyzing this scenario is based on the assumption that the tasks are sampled i.i.d. from a task environment or limited to strongly constrained data distributions. Instead we study two scenarios when lifelong learning is possible, even though the observed tasks do not form an i.i.d. sample: first, when they are sampled from the same environment, but possibly with dependencies, and second, when the task environment is allowed to change over time. In the first case we prove a PAC-Bayesian theorem, which can be seen as a direct generalization of the analogous previous result for the i.i.d. case. For the second scenario we propose to learn an inductive bias in form of a transfer procedure. We present a generalization bound and show on a toy example how it can be used to identify a beneficial transfer algorithm.

72 Learning with Symmetric Label Noise: The Importance of Being Unhinged

Brendan van Rooyen brendan.vanrooyen@nicta.com.au
Aditya Menon aditya.menon@nicta.com.au
Robert Williamson bob.williamson@nicta.com.au
NICTA

Convex potential minimisation is the de facto approach to binary classification. However, Long and Servedio [2008] proved that under symmetric label noise (SLN), minimisation of any convex potential over a linear function class can result in classification performance equivalent to random guessing. This ostensibly shows that convex losses are not SLN-robust. In this paper, we propose a convex, classification-calibrated loss and prove that it is SLN-robust. The loss avoids the Long and Servedio [2008] result by virtue of being negatively unbounded. The loss is a modification of the hinge loss, where one does not clamp at zero; hence, we call it the unhinged loss. We show that the optimal unhinged solution is equivalent to that of a strongly regularised SVM, and is the limiting solution for any convex potential; this implies that strong l2 regularisation makes most standard learners SLN-robust. Experiments confirm the unhinged loss' SLN-robustness.

73 Algorithms with Logarithmic or Sublinear Regret for Constrained Contextual Bandits

Huasen Wu huasenwu@gmail.com
Xin Liu liu@cs.ucdavis.edu
University of California, Davis
R. Srikant rsrikant@illinois.edu
Chong Jiang jiang17@illinois.edu
University of Illinois at Urbana-Champaign

We study contextual bandits with budget and time constraints under discrete contexts, referred to as constrained contextual bandits. The time and budget constraints significantly complicate the exploration and exploitation tradeoff because they introduce complex coupling among contexts over time. To gain insight, we first study unit-cost systems with known context distribution. When the expected rewards are known, we develop an approximation of the oracle, referred to Adaptive-Linear-Programming(ALP), which achieves near-optimality and only requires the ordering of expected rewards. With these highly desirable features, we then combine ALP with the upper-confidence-bound (UCB) method in the general case where the expected rewards are unknown a priori. We show that the proposed UCB-ALP algorithm achieves logarithmic regret except in certain boundary cases.Further, we design algorithms and obtain similar regret analysis results for more general systems with unknown context distribution or heterogeneous costs. To the best of our knowledge, this is the first work that shows how to achieve logarithmic regret in constrained contextual bandits. Moreover, this work also sheds light on the study of computationally efficient algorithms for general constrained contextual bandits.

74 From random walks to distances on unweighted graphs

Tatsunori Hashimoto thashim@mit.edu
Yi Sun yisun@math.mit.edu
Tommi Jaakkola tommi@csail.mit.edu
MIT

Large unweighted directed graphs are commonly used to capture relations between entities. A fundamental problem in the analysis of such networks is to properly define the similarity or dissimilarity between any two vertices. Despite the significance of this problem, statistical characterization of the proposed metrics has been limited.We introduce and develop a class of techniques for analyzing random walks on graphs using stochastic calculus. Using these techniques we generalize results on the degeneracy of hitting times and analyze a metric based on the Laplace transformed hitting time (LTHT). The metric serves as a natural, provably well-behaved alternative to the expected hitting time. We establish a general correspondence between hitting times of the Brownian motion and analogous hitting times on the graph. We show that the LTHT is consistent with respect to the underlying metric of a geometric graph, preserves clustering tendency, and remains robust against random addition of non-geometric edges. Tests on simulated and real-world data show that the LTHT matches theoretical predictions and outperforms alternatives.

75 Robust Regression via Hard Thresholding

Kush Bhatia kushbhatia03@gmail.com
Prateek Jain prajain@microsoft.com
Microsoft Research
Puru Kar purushot@cse.iitk.ac.in
Indian Institute of Technology Kanpur

We study the problem of Robust Least Squares Regression (RLSR) where several response variables can be adversarially corrupted. More specifically, for a data matrix X \in \R^{p x n} and an underlying model w*, the response vector is generated as y = X'w* + b where b \in n is the corruption vector supported over at most C.n coordinates. Existing exact recovery results for RLSR focus solely on L1-penalty based convex formulations and impose relatively strict model assumptions such as requiring the corruptions b to be selected independently of X.In this work, we study a simple hard-thresholding algorithm called TORRENT which, under mild conditions on X, can recover w* exactly even if b corrupts the response variables in an adversarial manner, i.e. both the support and entries of b are selected adversarially after observing X and w*. Our results hold under deterministic assumptions which are satisfied if X is sampled from any sub-Gaussian distribution. Finally unlike existing results that apply only to a fixed w*, generated independently of X, our results are universal and hold for any w* \in \R^p.Next, we propose gradient descent-based extensions of TORRENT that can scale efficiently to large scale problems, such as high dimensional sparse recovery. and prove similar recovery guarantees for these extensions. Empirically we find TORRENT, and more so its extensions, offering significantly faster recovery than the state-of-the-art L1 solvers. For instance, even on moderate-sized datasets (with p = 50K) with around 40% corrupted responses, a variant of our proposed method called TORRENT-HYB is more than 20x faster than the best L1 solver.

76 Column Selection via Adaptive Sampling

Saurabh Paul saurabhpaul2006@gmail.com
Paypal Inc
Malik Magdon-Ismail magdon@cs.rpi.edu
Petros Drineas drinep@cs.rpi.edu
Rensselaer Polytechnic Institute

Selecting a good column (or row) subset of massive data matrices has found many applications in data analysis and machine learning. We propose a new adaptive sampling algorithm that can be used to improve any relative-error column selection algorithm. Our algorithm delivers a tighter theoretical bound on the approximation error which we also demonstrate empirically using two well known relative-error column subset selection algorithms. Our experimental results on synthetic and real-world data show that our algorithm outperforms non-adaptive sampling as well as prior adaptive sampling approaches.

77 Multi-class SVMs: From Tighter Data-Dependent Generalization Bounds to Novel Algorithms

Yunwen Lei yunwelei@cityu.edu.hk
City University of Hong Kong
Urun Dogan udogan@microsoft.com
Microsoft
Alexander Binder alexander_binder@sutd.edu.sg
TU Berlin & Singapore University
Marius Kloft kloft@hu-berlin.de
Humboldt University Berlin

This paper studies the generalization performance of multi-class classification algorithms, for which we obtain, for the first time, a data-dependent generalization error bound with a logarithmic dependence on the class size, substantially improving the state-of-the-art linear dependence in the existing data-dependent generalization analysis. The theoretical analysis motivates us to introduce a new multi-class classification machine based on lp-norm regularization, where the parameter p controls the complexity of the corresponding bounds. We derive an efficient optimization algorithm based on Fenchel duality theory. Benchmarks on several real-world datasets show that the proposed algorithm can achieve significant accuracy gains over the state of the art.

78 Optimal Linear Estimation under Unknown Nonlinear Transform

Xinyang Yi yixy@utexas.edu
Constantine Caramanis constantine@utexas.edu
UT Austin
Zhaoran Wang zhaoran@princeton.edu
Han Liu hanliu@princeton.edu
Princeton University

Linear regression studies the problem of estimating a model parameter $\beta* \in \R p$, from n observations $\{(yi,xi)\}ni=1$ from linear model $yi=\langle \xi,\beta* \rangle+\epsilon i$. We consider a significant generalization in which the relationship between $\langle xi,\beta* \rangle$ and yi is noisy, quantized to a single bit, potentially nonlinear, noninvertible, as well as unknown. This model is known as the single-index model in statistics, and, among other things, it represents a significant generalization of one-bit compressed sensing. We propose a novel spectral-based estimation procedure and show that we can recover $\beta*$ in settings (i.e., classes of link function f) where previous algorithms fail. In general, our algorithm requires only very mild restrictions on the (unknown) functional relationship between yi and $\langle xi,\beta* \rangle$. We also consider the high dimensional setting where $\beta*$ is sparse, and introduce a two-stage nonconvex framework that addresses estimation challenges in high dimensional regimes where $p \gg n$. For a broad class of link functions between $xi,\beta* \rangle$ and yi, we establish minimax lower bounds that demonstrate the optimality of our estimators in both the classical and high dimensional regimes.

79 Risk-Sensitive and Robust Decision-Making: a CVaR Optimization Approach

Yinlam Chow yldick.chow@gmail.com
Marco Pavone pavone@stanford.edu
Stanford
Aviv Tamar avivt@tx.technion.ac.il
UC Berkeley
Shie Mannor shie@ee.technion.ac.il
Technion

In this paper we address the problem of decision making within a Markov decision process (MDP) framework where risk and modeling errors are taken into account. Our approach is to minimize a risk-sensitive conditional-value-at-risk (CVaR) objective, as opposed to a standard risk-neutral expectation. We refer to such problem as CVaR MDP. Our first contribution is to show that a CVaR objective, besides capturing risk sensitivity, has an alternative interpretation as expected cost under worst-case modeling errors, for a given error budget. This result, which is of independent interest, motivates CVaR MDPs as a unifying

framework for risk-sensitive and robust decision making. Our second contribution is to present a value-iteration algorithm for CVaR MDPs, and analyze its convergence rate. To our knowledge, this is the first solution algorithm for CVaR MDPs that enjoys error guarantees. Finally, we present results from numerical experiments that corroborate our theoretical findings and show the practicality of our approach.

80 Learning with Incremental Iterative Regularization

Lorenzo Rosasco lrosasco@mit.edu
University of Genova
Silvia Villa silvia.villa@iit.it
IIT-MIT

Within a statistical learning setting, we propose and study an iterative regularization algorithm for least squares defined by an incremental gradient method. In particular, we show that, if all other parameters are fixed a priori, the number of passes over the data (epochs) acts as a regularization parameter, and prove strong universal consistency, i.e. almost sure convergence of the risk, as well as sharp finite sample bounds for the iterates. Our results are a step towards understanding the effect of multiple epochs in stochastic gradient techniques in machine learning and rely on integrating statistical and optimizationresults.

81 No-Regret Learning in Repeated Bayesian Games

Jason Hartline hartline@eecs.northwestern.edu
Northwestern University
Vasilis Syrgkanis vasy@microsoft.com
Microsoft Research
Eva Tardos eva@cs.cornell.edu
Cornell University

Recent price-of-anarchy analyses of games of complete information suggest that coarse correlated equilibria, which characterize outcomes resulting from no-regret learning dynamics, have near-optimal welfare. This work provides two main technical results that lift this conclusion to games of incomplete information, a.k.a., Bayesian games. First, near-optimal welfare in Bayesian games follows directly from the smoothness-based proof of near-optimal welfare in the same game when the private information is public. Second, no-regret learning dynamics converge to Bayesian coarse correlated equilibrium in these incomplete information games. These results are enabled by interpretation of a Bayesian game as a stochastic game of complete information.

82 Sparse and Low-Rank Tensor Decomposition

Parikshit Shah pshah@discovery.wisc.edu
Yahoo Labs
Nikhil Rao nikhilr@cs.utexas.edu
University of Texas at Austin
Gongguo Tang gtang@mines.edu
Colorado School of Mines

Motivated by the problem of robust factorization of a low-rank tensor, we study the question of sparse and low-rank tensor decomposition. We present an efficient computational algorithm that modifies Leurgans' algoirthm for tensor factorization. Our method relies on a reduction of the problem to sparse and low-rank matrix decomposition via the notion of tensor contraction. We use well-understood convex techniques for solving the reduced matrix sub-problem which then allows us to perform the full decomposition of the tensor. We delineate situations where the problem is recoverable and provide theoretical guarantees for our algorithm. We validate our algorithm with numerical experiments.

83 Analysis of Robust PCA via Local Incoherence

Huishuai Zhang hzhan23@syr.edu
Yi Zhou yzhou35@syr.edu
Yingbin Liang yliang06@syr.edu
Syracuse Univeristy

We investigate the robust PCA problem of decomposing an observed matrix into the sum of a low-rank and a sparse error matrices via convex programming Principal Component Pursuit (PCP). In contrast to previous studies that assume the support of the error matrix is generated by uniform Bernoulli sampling, we allow non-uniform sampling, i.e., entries of the low-rank matrix are corrupted by errors with unequal probabilities. We characterize conditions on error corruption of each individual entry based on the local incoherence of the low-rank matrix, under which correct matrix decomposition by PCP is guaranteed. Such a refined analysis of robust PCA captures how robust each entry of the low rank matrix combats error corruption. In order to deal with non-uniform error corruption, our technical proof introduces a new weighted norm and develops/exploits the concentration properties that such a norm satisfies.

84 Algorithmic Stability and Uniform Generalization

Ibrahim M Alabdulmohsin
ibrahim.alabdulmohsin@kaust.edu.sa
King Abdullah University of Science & Technology

One of the central questions in statistical learning theory is to determine the conditions under which agents can learn from experience. This includes the necessary and sufficient conditions for generalization from a given finite training set to new observations. In this paper, we prove that algorithmic stability in the inference process is equivalent to uniform generalization across all parametric loss functions. We provide various interpretations of this result. For instance, a relationship is proved between stability and data processing, which reveals that algorithmic stability can be improved by post-processing the inferred hypothesis or by augmenting training examples with artificial noise prior to learning. In addition, we establish a relationship between algorithmic stability and the size of the observation space, which provides a formal justification for dimensionality reduction methods. Finally, we connect algorithmic stability to the size of the hypothesis space, which recovers the classical PAC result that the size (complexity) of the hypothesis space should be controlled in order to improve algorithmic stability and improve generalization.

85 Mixing Time Estimation in Reversible Markov Chains from a Single Sample Path

Daniel J Hsu danielhsu@gmail.com
Columbia University
Aryeh Kontorovich karyeh@cs.bgu.ac.il
Ben Gurion University
Csaba Szepesvari szepesva@cs.ualberta.ca
University of Alberta

This article provides the first procedure for computing a fully data-dependent interval that traps the mixing time tmix of a finite reversible ergodic Markov chain at a prescribed confidence level. The interval is computed from a single finite-length sample path from the Markov chain, and does not require the knowledge of any parameters of the chain. This stands in contrast to previous approaches, which either only provide point estimates, or require a reset mechanism, or additional prior knowledge. The interval is constructed around the relaxation time trelax, which is strongly related to the mixing time, and the width of the interval converges to zero at a n√ rate, where n is the length of the sample path. Upper and lower bounds are given on the number of samples required to achieve constant-factor multiplicative accuracy. The lower bounds indicate that, unless further restrictions are placed on the chain, no procedure can achieve this accuracy level before seeing each state at least Ω(trelax) times on the average. Finally, future directions of research are identified.

86 Efficient Compressive Phase Retrieval with Constrained Sensing Vectors

Sohail Bahmani sohail.bahmani@ece.gatech.edu
Justin Romberg jrom@ece.gatech.edu
Georgia Institute of Technology

We propose a robust and efficient approach to the problem of compressive phase retrieval in which the goal is to reconstruct a sparse vector from the magnitude of a number of its linear measurements. The proposed framework relies on constrained sensing vectors and a two-stage reconstruction method that consists of two standard convex programs that are solved sequentially.In recent years, various methods are proposed for compressive phase retrieval, but they have suboptimal sample complexity or lack robustness guarantees. The main obstacle has been that there is no straightforward convex relaxations for the type of structure in the target. Given a set of underdetermined measurements, there is a standard framework for recovering a sparse matrix, and a standard framework for recovering a low-rank matrix. However, a general, efficient method for recovering a jointly sparse and low-rank matrix has remained elusive.Deviating from the models with generic measurements, in this paper we show that if the sensing vectors are chosen at random from an incoherent subspace, then the low-rank and sparse structures of the target signal can be effectively decoupled. We show that a recovery algorithm that consists of a low-rank recovery stage followed by a sparse recovery stage will produce an accurate estimate of the target when the number of measurements is O(klogdk), where k and d denote the sparsity level and the dimension of the input signal. We also evaluate the algorithm through numerical simulation.

87 Unified View of Matrix Completion under General Structural Constraints

Suriya Gunasekar suriya@utexas.edu
Joydeep Ghosh ghosh@ece.utexas.edu
UT Austin
Arindam Banerjee banerjee@cs.umn.edu
University of Minnesota

Matrix completion problems have been widely studied under special low dimensional structures such as low rank or structure induced by decomposable norms. In this paper, we present a unified analysis of matrix completion under general low--dimensional structural constraints induced by {\em any} norm. We consider two estimators for the general problem, and provide unified upper bounds on the sample complexity and the estimation error for such structured matrix completion. Our analysis relies on generic chaining, and we establish two intermediate results of independent interest: a certain partial complexity measure encountered in the analysis of matrix completion problems can be better understood and bounded in terms of Gaussian widths, and a form of Restricted Strong Convexity holds for matrix completion under general norm regularization.Further, we provide several non--trivial examples of structures included in our framework, notably including the recently proposed spectral k--support norm.

88 Copeland Dueling Bandits

Masrour Zoghi m.zoghi@uva.nl
Maarten de Rijke derijke@uva.nl
University of Amsterdam
Zohar S Karnin zkarnin@yahoo-inc.com
Shimon Whiteson shimon.whiteson@cs.ox.ac.uk
University of Oxford

A version of the dueling bandit problem is addressed in which a Condorcet winner may not exist. Two algorithms are proposed that instead seek to minimize regret with respect to the Copeland winner, which, unlike the Condorcet winner, is guaranteed to exist. The first, Copeland Confidence Bound (CCB), is designed for small numbers of arms, while the second, Scalable Copeland Bandits (SCB), works better for large-scale problems. We provide theoretical results bounding the regret accumulated by CCB and SCB, both substantially improving existing results. Such existing results either offer bounds of the form O(K log T) but require restrictive assumptions, or offer bounds of the form O(K^2 log T) without requiring such assumptions. Our results offer the best of both worlds: O(K log T) bounds without restrictive assumptions.

89 Regret Lower Bound and Optimal Algorithm in Finite Stochastic Partial Monitoring

Junpei Komiyama junpeikomiyama@gmail.com
Junya Honda honda@stat.t.u-tokyo.ac.jp
Hiroshi Nakagawa nakagawa@dl.itc.u-tokyo.ac.jp
The University of Tokyo

Partial monitoring is a general model for sequential learning with limited feedback formalized as a game between two players. In this game, the learner chooses an action and at the same time the opponent chooses an outcome, then the learner suffers a loss and receives a feedback signal. The goal of the learner is to minimize the total loss. In this paper, we study partial monitoring with finite actions and stochastic outcomes. We derive a logarithmic

distribution-dependent regret lower bound that defines the hardness of the problem. Inspired by the DMED algorithm (Honda and Takemura, 2010) for the multi-armed bandit problem, we propose PM-DMED, an algorithm that minimizes the distribution-dependent regret. PM-DMED significantly outperforms state-of-the-art algorithms in numerical experiments. To show the optimality of PM-DMED with respect to the regret bound, we slightly modify the algorithm by introducing a hinge function (PM-DMED-Hinge). Then, we derive an asymptotical optimal regret upper bound of PM-DMED-Hinge that matches the lower bound.

90 Online Learning for Adversaries with Memory: Price of Past Mistakes

Oren Anava oren.anava@gmail.com
Shie Mannor shie@ee.technion.ac.il
Technion
Elad Hazan ehazan@cs.princeton.edu
Princeton University

The framework of online learning with memory naturally captures learning problems with temporal effects, and was previously studied for the experts setting. In this work we extend the notion of learning with memory to the general Online Convex Optimization (OCO) framework, and present two algorithms that attain low regret. The first algorithm applies to Lipschitz continuous loss functions, obtaining optimal regret bounds for both convex and strongly convex losses. The second algorithm attains the optimal regret bounds and applies more broadly to convex losses without requiring Lipschitz continuity, yet is more complicated to implement. We complement the theoretic results with two applications: statistical arbitrage in finance, and multi-step ahead prediction in statistics.

91 Revenue Optimization against Strategic Buyers

Mehryar Mohri mohri@cs.nyu.edu
Andres Munoz amunoz88@gmail.com
Courant Institute of Mathematical Sciences

We present a revenue optimization algorithm for posted-price auctions when facing a buyer with random valuations who seeks to optimize his γ-discounted surplus. To analyze this problem, we introduce the notion of epsilon-strategic buyer, a more natural notion of strategic behavior than what has been used in the past. We improve upon the previous state-of-the-art and achieve an optimal regret bound in O(logT+1log(1/γ)) when the seller can offer prices from a finite set \cP and provide a regret bound in O˜(T√+T1/4log(1/γ)) when the buyer is offered prices from the interval [0,1].

92 On Top-k Selection in Multi-Armed Bandits and Hidden Bipartite Graphs

Wei Cao cao-w13@mails.tsinghua.edu.cn
Jian Li lijian83@mail.tsinghua.edu.cn
Zhize Li zz-li14@mails.tsinghua.edu.cn
Tsinghua University
Yufei Tao taoyf@cse.cuhk.edu.hk
CUHK

This paper discusses how to efficiently choose from n unknowndistributions the k ones whose means are the greatest by a certainmetric, up to a small relative error. We study the topic under twostandard settings---multi-armed bandits and hidden bipartitegraphs---which differ in the nature of the input distributions. In theformer setting, each distribution can be sampled (in the i.i.d.manner) an arbitrary number of times, whereas in the latter, eachdistribution is defined on a population of a finite size m (andhence, is fully revealed after m samples). For both settings, weprove lower bounds on the total number of samples needed, and proposeoptimal algorithms whose sample complexities match those lower bounds.

93 Improved Iteration Complexity Bounds of Cyclic Block Coordinate Descent for Convex Problems

Ruoyu Sun ruoyus@stanford.edu
Stanford university
Mingyi Hong mingyi@iastate.edu

The iteration complexity of the block-coordinate descent (BCD) type algorithm has been under extensive investigation. It was shown that for convex problemsthe classical cyclic BCD method achieves an O(1/r) complexity, where r is the iteration counter. However, such bounds are explicitly dependent on K (the number of variable blocks), and are at least K times worse than those of the gradient descent (GD) and proximal gradient (PG) methods. In this paper, we close such theoretical performance gap between BCD and GD/PG. First we show that for a family of quadratic nonsmooth problems, the complexity bounds for BCD and its popular variant Block Coordinate Proximal Gradient (BCPG) can match those ofthe GD/PG in terms of dependency on K. Second, we establish an improved iteration complexity bound of CGD (Coordinate Gradient Descent) for general convex problems which can match that of GD in certain scenarios. Our bounds are sharper than the known bounds as they are always at least K times worse than GD.

94 Cornering Stationary and Restless Mixing Bandits with Remix-UCB

Julien Audiffren julien.audiffren@ens-cachan.fr
CMLA, ENS Cachan
Liva Ralaivola liva.ralaivola@lif.univ-mrs.fr
Univesity of Marseille

We study the restless bandit problem where arms are associated with stationary φ-mixing processes and where rewards are therefore dependent: the question that arises from this setting is that of carfully recovering some independence by `ignoring' the values of some rewards. As we shall see, the bandit problem we tackle requires us to address the exploration/exploitation/independence trade-off, which we do by considering the idea of a {\em waiting arm} in the the new Remix-UCB algorithm, a generalization of Improved-UCB for the problem at hand, that we introduce. We provide a regret analysis for this bandit strategy; two noticeable features of Remix-UCB are that i) it reduces to the regular Improved-UCB when the φ-mixing coefficients are all 0, i.e. when the i.i.d scenario is recovered, and ii) when φ(n)=O(n−α), it is able to ensure a controlled regret of order \Ot(Δ(α−2)/α∗log1/αT), where Δ∗ encodes the distance between the best arm and the best suboptimal arm, even in the case when α<1, i.e. the case when the φ-mixing coefficients {\em are not} summable.

95 Fighting Bandits with a New Kind of Smoothness

Jacob D Abernethy jabernet@umich.edu
Chansoo Lee chansool@umich.edu
Ambuj Tewari tewaria@umich.edu
University of Michigan

We focus on the adversarial multi-armed bandit problem. The EXP3 algorithm of Auer et al. (2003) was shown to have a regret bound of O(TNlogN√), where T is the time horizon and N is the number of available actions (arms). More recently, Audibert and Bubeck (2009) improved the bound by a logarithmic factor via an entirely different method. In the present work, we provide a new set of analysis tools, using the notion of convex smoothing, to provide several novel algorithms with optimal guarantees. First we show that regularization via the Tsallis entropy matches the minimax rate of Audibert and Bubeck (2009) with an even tighter constant; it also fully generalizes EXP3. Second we show that a wide class of perturbation methods lead to near-optimal bandit algorithms as long as a simple condition on the perturbation distribution D is met: one needs that the hazard function of D remain bounded. The Gumbel, Weibull, Frechet, Pareto, and Gamma distributions all satisfy this key property; interestingly, the Gaussian and Uniform distributions do not.

96 Asynchronous stochastic approximation: the noise is in the noise and SGD don't care

John C Duchi jduchi@stanford.edu
Sorathan Chaturapruek sorathan@cs.stanford.edu
Chris Ré chrismre@cs.stanford.edu
Stanford University

We show that asymptotically, completely asynchronous stochastic gradientprocedures achieve optimal (even to constant factors) convergence rates forthe solution of convex optimization problems under nearly the sameconditions required for asymptotic optimality of standard stochasticgradient procedures. Roughly, the noise inherent to the stochasticapproximation scheme dominates any noise from asynchrony. We also giveempirical evidence demonstrating the strong performance of asynchronous,parallel stochastic optimization schemes. In short, we show that for manystochastic approximation problems, as Freddie Mercury sings in Queen's\emph{Bohemian Rhapsody}, ``Nothing really matters.''

97 The Pareto Regret Frontier for Bandits

Tor Lattimore tor.lattimore@gmail.com
University of Alberta

Given a multi-armed bandit problem it may be desirable to achieve a smaller-than-usual worst-case regret for some special actions. I show that the price for such unbalanced worst-case regret guarantees is rather high. Specifically, if an algorithm enjoys a worst-case regret of B with respect to some action, then there must exist another action for which the worst-case regret is at least Ω(nK/B), where n is the horizon and K the number of actions. I also give upper bounds in both the stochastic and adversarial settings showing that this result cannot be improved. For the stochastic case the pareto regret frontier is characterised exactly up to constant factors.

98 Online Learning with Gaussian Payoffs and Side Observations

Yifan Wu ywu12@ualberta.ca
Csaba Szepesvari szepesva@cs.ualberta.ca
University of Alberta
András György a.gyorgy@imperial.ac.uk
Imperial College London

We consider a sequential learning problem with Gaussian payoffs and side information: after selecting an action i, the learnerreceives information about the payoff of every action j in the form of Gaussian observations whose mean is the same as the mean payoff, but the variance depends on the pair (i,j) (and may be infinite). The setup allows a more refined information transfer from one action to another than previous partial monitoring setups, including the recently introduced graph-structured feedback case. For the first time in the literature, we provide non-asymptotic problem-dependent lower bounds on the regret of any algorithm, which recover existing asymptotic problem-dependent lower bounds and finite-time minimax lower bounds available in the literature. We also provide algorithms that achieve the problem-dependent lower bound (up to some universal constant factor) or the minimax lower bounds (up to logarithmic factors).

99 Fast Rates for Exp-concave Empirical Risk Minimization

Tomer Koren tomerk@technion.ac.il
Kfir Levy kfiryehud@gmail.com
Technion

We consider Empirical Risk Minimization (ERM) in the context of stochastic optimization with exp-concave and smooth losses---a general optimization framework that captures several important learning problems including linear and logistic regression, learning SVMs with the squared hinge-loss, portfolio selection and more. In this setting, we establish the first evidence that ERM is able to attain fast generalization rates, and show that the expected loss of the ERM solution in d dimensions converges to the optimal expected loss in a rate of d/n. This rate matches existing lower bounds up to constants and improves by a logn factor upon the state-of-the-art, which is only known to be attained by an online-to-batch conversion of computationally expensive online algorithms.

100 Adaptive Low-Complexity Sequential Inference for Dirichlet Process Mixture Models

Theodoros Tsiligkaridis ttsili@umich.edu
Theodoros Tsiligkaridis ttsili@ll.mit.edu
Keith Forsythe forsythe@ll.mit.edu
MIT Lincoln Laboratory

We develop a sequential low-complexity inference procedure for Dirichlet process mixtures of Gaussians for online clustering and parameter estimation when the number of clusters are unknown a-priori. We present an easily computable, closed form parametric expression for the conditional likelihood, in which hyperparameters are recursively updated as a function of the streaming data assuming conjugate priors. Motivated by large-sample asymptotics, we propose a noveladaptive low-complexity design for the Dirichlet process concentration parameter and show that the number of classes grow at most at a logarithmic rate. We further prove that in the large-sample limit, the conditional likelihood and datapredictive distribution become asymptotically Gaussian. We demonstrate through experiments on synthetic and real data sets that our approach is superior to otheronline state-of-the-art methods.

SYMPOSIA

Thursday December 10th: - 3:00 pm - 9:00 pm

Algorithms Among Us: the Societal Impacts of Machine Learning

Michael A Osborne mosb@robots.ox.ac.uk
U Oxford
Adrian Weller aw665@cam.ac.uk
University of Cambridge
Murray Shanahan m.shanahan@imperial.ac.uk
Imperial College London

Public interest in Machine Learning is mounting as the societal impacts of technologies derived from our community become evident. This symposium aims to turn the attention of ML researchers to the present and future consequences of our work, particularly in the areas of privacy, military robotics, employment and liability. These topics now deserve concerted attention to ensure the best interests of those both within and without ML: the community must engage with public discourse so as not to become the victim of it (as other fields have e.g. genetic engineering). The symposium will bring leaders within academic and industrial ML together with experts outside the field to debate the impacts of our algorithms and the possible responses we might adopt.

Brains, Minds, and Machines

Yoshua Bengio yoshua.bengio@umontreal.ca
Université of Montréal
Marc'Aurelio Ranzato ranzato@fb.com
Facebook
Honglak Lee honglak@eecs.umich.edu
UNIVERSITY Michigan
Max Welling welling.max@gmail.com
University of Amsterdam
Andrew Y Ng andrewng@baidu.com
Baidu Research

Deep Learning algorithms attempt to discover good representations, at multiple levels of abstraction. Deep Learning is a topic of broad interest, both to researchers who develop new algorithms and theories, as well as to the rapidly growing number of practitioners who apply these algorithms to a wider range of applications, from vision and speech processing, to natural language understanding, neuroscience, health, etc. Major conferences in these fields often dedicate several sessions to this topic, attesting the widespread interest of our community in this area of research.

There has been very rapid and impressive progress in this area in recent years, in terms of both algorithms and applications, but many challenges remain. This symposium aims at bringing together researchers in Deep Learning and related areas to discuss the new advances, the challenges we face, and to brainstorm about new solutions and directions.

Deep Learning Symposium

Gabriel Kreiman gabriel.kreiman@tch.harvard.edu
Harvard Medical School
Tomaso A Poggio tp@ai.mit.edu
Maximilian Nickel mnick@mit.edu
Massachusetts Institute of Technology

The science of today enables engineering solutions of tomorrow. In this symposium we will discuss state-of-the-art results in the scientific understanding of intelligence and how these results enable new approaches to replicate intelligence in engineered systems.

Understanding intelligence and the brain requires theories at different levels, ranging from the biophysics of single neurons to algorithms, computations, and a theory of learning. In this symposium, we aim to bring together researchers from machine learning and artificial intelligence, from neuroscience, and from cognitive science to present and discuss state-of-the-art research that is focused on understanding intelligence on these different levels.

Central questions of the symposium include how intelligence is grounded in computation, how these computations are implemented in neural systems, how intelligence can be described via unifying mathematical theories, and how we can build intelligent machines based on these principles. A particular focus of the symposium lies on how both models and algorithms can be guided by scientific concerns, incorporating constraints and findings from cognitive neuroscience, systems neuroscience, and cognitive development.

We believe that these topics, spanning the fields of artificial intelligence, neuroscience and cognitive science, lie at the core of the Conference on Neural Information Processing Systems and are of great interest to its general audience. Moreover, the accumulated knowledge and technology that is now in place has set the stage for rapid advances in these areas and in the creation of intelligent machines. We believe that this makes it an ideal time to hold this symposium at NIPS.

The list of speakers at the symposium include:

- Geoffrey Hinton (University of Toronto, Google)
- Tomaso Poggio (MIT)
- Christof Koch (Allen Institute for Brain Science)
- Joshua Tenenbaum (MIT)
- Demis Hassabis (Google DeepMind)
- Andrew Saxe (Stanford University)
- Surya Ganguli (Stanford University)

WORKSHOPS

Friday December 11th: - 8:30 am - 6:30 pm **Saturday December 12th: - 8:30 am - 6:30 pm**

- **Machine Learning and Interpretation in Neuroimaging**
  Room 515 A
- **Machine Learning For Healthcare (MLHC)**
  Room 510 DB
- **Feature Extraction: Modern Questions and Challenges**
  Room 513 EF
- **Nonparametric Methods for Large Scale Representation Learning**
  Room 511 C
- **Optimization for Machine Learning (OPT2015)**
  Room 510 AC
- **Statistical Methods for Understanding Neural Systems**
  Room 511 F
- **Modelling and inference for dynamics on complex interaction networks: joining up machine learning and statistical physics**
  Room 511 E
- **Advances in Approximate Bayesian Inference**
  Room 513 AB
- **Deep Reinforcement Learning**
  Room 513 CD
- **Bounded Optimality and Rational Metareasoning**
  Room 512 BF
- **Multimodal Machine Learning**
  Room 512 DH
- **ABC in Montreal**
  Room 511 A
- **Cognitive Computation: Integrating neural and symbolic approaches**
  Room 512 CG
- **Machine Learning for Spoken Language Understanding and Interactions**
  Room 511 B
- **Applying (machine) Learning to Experimental Physics (ALEPH) and "Flavours of Physics" challenge**
  Room 515 BC
- **Time Series Workshop**
  Room 514 BC
- **Learning Faster from Easy Data II**
  Room 511 D
- **Machine Learning for (e-)Commerce**
  Room 512 E
- **Probabilistic Integration**
  Room 512 A
- **Adaptive Data Analysis**
  Room 514 A
- **Extreme Classification 2015: Multi-class and Multi-label Learning in Extremely Large Label Spaces**
  Room 514 A
- **Bayesian Nonparametrics: The Next Generation**
  Room 515 BC
- **Bayesian Optimization: Scalability and Flexibility**
  Room 511 B
- **Challenges in Machine Learning (CiML 2015): "Open Innovation" and "Coopetitions"**
  Room 512 E
- **Quantum Machine Learning**
  Room 512 A
- **Transfer and Multi-Task Learning: Trends and New Perspectives**
  Room 514 BC
- **Machine Learning in Computational Biology**
  Room 510 DB
- **Learning and privacy with incomplete data and weak supervision**
  Room 512 DH
- **Networks in the Social and Information Sciences**
  Room 512 BF
- **Multiresolution methods for large-scale learning**
  Room 511 C
- **Scalable Monte Carlo Methods for Bayesian Analysis of Big Data**
  Room 513 AB
- **BigNeuro 2015: Making sense of big neural data**
  Room 511 E
- **Learning, Inference and Control of Multi-Agent Systems**
  Room 511 A
- **Machine Learning From and For Adaptive User Technologies: From Active Learning & Experimentation to Optimization & Personalization**
  Room 511 F
- **Reasoning, Attention, Memory (RAM) Workshop**
  Room 510 AC
- **Machine Learning Systems**
  Room 511 D
- **Non-convex Optimization for Machine Learning: Theory and Practice**
  Room 513 CD
- **Machine Learning & Interpretation in Neuroimaging**
  Room 515 A
- **Black box learning and inference**
  Room 513 EF
- **Cognitive Computation: Integrating neural and symbolic approaches**
  Room 512 CG

REVIEWERS

Yasin Abbasi-Yadkori
Jacob Abernethy
Jayadev Acharya
Margareta Ackerman
Ryan Adams
Tameem Adel
Alekh Agarwal
Shivani Agarwal
Edoardo Airoldi
Shotaro Akaho
Karteek Alahari
Morteza Alamgir
Suarez Alberto
Daniel Alexander
Alnur Ali
Alexandre Allauzen
Ethem Alpaydin
Marco Alvarez
Mauricio Alvarez
Carlos Alzate
Christopher Amato
Odalric Ambrym-Maillard
Kareem Amin
Massih-Reza Amini
Eyal Amir
Oren Anava
Charles Anderson
Bjoern Andres
Galen Andrew
Nick Andrews
Christophe Andrieu
David Andrzejewski
Elaine Angelino
Drago Anguelov
Ognjen Arandjelovic
Aleksandr Aravkin
Cedric Archambeau
Raman Arora
Thierry Artieres
John Ashburner
Ozlem Aslan
Hideki Asoh
Josh Attenberg
Hagai Attias
Joseph Austerweil
Paolo Avesani
Bernardo Avila Pires
Dubey Avinava
Yusuf Aytar
Javad Azimi
Martin Azizyan
Amadou Ba
Monica Babes-Vroman
Stephen Bach
Ashwinkumar Badanidiyuru
Bing Bai
Xiang Bai
Raphael Bailly
Krishnakumar Balasubramanian
Christopher Baldassano
Luca Baldassare
Borja Balle
Akshay Balsubramani
David Bamman
Arindam Banerjee
Bikramjit Banerjee
Mohit Bansal
Ying-Ze Bao
Yoseph Barash
Remi Bardenet
Elias Bareinboim
Andre Barreto
Jon Barron
Simon Barthelmé
Peter Bartlett
Sumanta Basu
Sumit Basu
Dhruv Batra
Peter Battaglia
Justin Bayer
Carlos Becker
Behrouz Behmardi
Mikhail Belkin
Peter Bell
Shai Ben-David
Yoshua Bengio
Philipp Berens
James Bergstra
Luc Berthouze
Michel Besserve
Badri Narayan Bhaskar
Chiranjib Bhattacharyya
Sharmodeep Bhattacharyya
Srinadh Bhojanapalli
Chetan Bhole
Jinbo Bi
Wei Bi
Manuele Bicego
Felix Biessmann
Battista Biggio
Zhao Bin
William Bishop
Matthew Blaschko
Mathieu Blondel
Charles Blundell
Liefeng Bo
Dean Bodenham
Sander Bohte
Danushka Bollegala
Edwin Bonilla
Byron Boots
Antoine Bordes
Jorg Bornschein
Reza Bosagh Zadeh
Guillaume Bouchard
Alexandre Bouchard-Cote
Abdeslam Boularias
YLan Boureau
Christos Boutsidis
Levi Boyles
Philemon Brakel
Steve Branson
Ulf Brefeld
Wieland Brendel
Xavier Bresson
John Bridle
Mark Briers
Tamara Broderick
Marcus Brubaker
Michael Brueckner
Joan Bruna
Sebastien Bubeck
Florian Buettner
Hung Bui
Yaroslav Bulatov
Yura Burda
Róbert Busa-Fekete
Lars Busing
Arunkumar Byravan
Deng Cai
Roberto Calandra
Cl?ment Calauz?nes
Trevor Campbell
William Campbell
Guillermo Canas
Stephane Canu
Liangliang Cao
Yongzhi Cao
Barbara Caputo
Constantine Caramanis
Jaime Carbonell
Alexandra Carpentier
Joao Carreira
Roberto Casarin
Robert Castelo
Gavin Cawley
Asli Celikyilmaz
Taylan Cemgil
Volkan Cevher
Hakan Cevikalp
Arun Chaganty
Kian Ming Chai
Brahim Chaib-draa
Ayan Chakrabarti
Antoni Chan
Laiwan CHAN
Allison Chaney
Kai-Wei Chang
Nicolas Chapados
Denis Charles
Laurent Charlin
Guillaume Charpiat
Kamalika Chaudhuri
Sougata Chaudhuri
Sourish Chaudhuri
Rama Chellappa
Bo Chen
Changyou Chen
Chao Chen
Chao-Yeh Chen
Jianhui Chen
Lin Chen
Ning Chen
Shang-Tse Chen
Tianqi Chen
Wenlin Chen
Xi Chen
Xianjie Chen
Yao-Nan Chen
Yen-Chi Chen
Yining Chen
Yudong Chen
Yutian Chen
Yuxin Chen
Weiwei Cheng
Michael Chertkov
William Cheung
chao-kai Chiang
David Chiang
Hai Leong Chieu
Arthur Choi
Dave Choi
Heeyoul Choi
Jaesik Choi
Jaegul Choo
Yinlam Chow
Girish Chowdhary
Konstantina Christakopoulou
Wei Chu
Kacper Chwialkowski
Moustapha Cisse
Tom Claassen
Stephan Clemencon
Mark Coates
Ruben Coen-cagli
Alon Cohen
Ronan Collobert
Richard Combes
Greg Corrado
Garrison Cottrell
Benjamin Cowley
Aron Culotta
James Cussens
William Dabney
Bo Dai
Zhenwen Dai
Arnak Dalalyan
Andreas Damianou
Christian Daniel
Ivo Danihelka
Christoph Dann
Gautam Dasarathy
Sanjoy Dasgupta
Emmanuel Dauce
Yann Dauphin
Ian Davidson
Jesse Davis
Peter Dayan
Tijl De Bie
Cassio de Campos
Teofilo de Campos
Yann-Michaël De Hauwere
Fernando de la Torre
Viriginia De Sa
Giulia De Salvo
Dennis Decoste
Olivier Delalleau
Krzysztof Dembczynski
Jia Deng
Xinwei Deng
Misha Denil
Ludovic Denoyer
Thomas Desautels
Guillaume Desjardins
Paramveer Dhillon
Amit Dhurandhar
Kostas Diamantaras
Lee Dicker
Tom Diethe
Thomas Dietterich
Laura Dietz
Alex Dimakis
Christos Dimitrakakis
Hu Ding
Nan Ding
Wei Ding
Josip Djolonga
Urun Dogan
Justin Domke
Frank Dondelinger
Jana Doppa
Finale Doshi-Velez
Arnaud Doucet
Lan Du
Nan Du
Lixin Duan
Artur Dubrawski
John Duchi
Haimonti Dutta
David Duvenaud
Jennifer Dy
Sandra Ebert
Alexander Ecker
David Eigen
Jacob Eisenstein
Carl Henrik Ek
Chaitanya Ekanadham
James Elder
Ehsan Elhamifar
Micha Elsner
Victor Elvira
Frank Emmert-Streib
Dominik Endres
Peter Englert
Tom Erez
Dumitru Erhan
Stefano Ermon
Tim van Erven
Ali Eslami
Vincent Etter
Ludger Evers
Georgios Exarchakis
Jalal Fadili
Bilal Fadlallah
Shobeir Fakhraei
Amir-massoud Farahmand
mehrdad Farajtabar
Jiashi FENG
Aasa Feragen
Raphaël Féraud
Olivier Fercoq
Rob Fergus
Xiaoli Fern
Carlos Fernandez-Granda
Vittorio Ferrari
Cedric Fevotte
Mario Figueiredo
Roman Filipovych
Maurizio Filippone
Marcelo Fiori
Jean Baptiste Fiot
Asja Fischer
Jozsef Fiser
Madalina Fiterau
Boris Flach
Peter Flach
Patrick Flaherty
Remi Flamary
Seth Flaxman
Francois Fleuret
Raphael Fonteneau
Florence Forbes
Jessica Forde
Blaz Fortuna
Dean Foster
Nick Foti
James Foulds
Rina Foygel Barber
Vojtech Franc
Eibe Frank
William Freeman
Oren Freifeld
Jes Frellsen
Abram Friesen
Roger Frigola
Mario Fritz
Rafael Frongillo
Pascal Frossard
Roy Frostig
Nicolo Fusi
Alona Fyshe
Victor Gabillon
Adrien Gaidon
Yarin Gal
Juergen Gall
Marcus Gallagher
Patrick Gallinari
Ravi Ganti
Jing Gao
Shenghua Gao
Tianshi Gao
Wei Gao
Xin Gao
Yuanjun Gao
Gilles Gasso
Jan Gasthaus
Romaric Gaudel
Rong Ge
Peter Gehler
Andreas Geiger
Matthieu Geist
Andrew Gelfand
Alborz Geramifard
Krzysztof Geras
Pascal Germain
Sean Gerrish
Samuel Gershman
Pierre Geurts
Mohammad Ghavamzadeh
Mohammad Gheshlaghi azar
Soumya Ghosh
Arnab Ghoshal
Bryan Gibson
Sébastien Giguère
Mark Girolami
Ross Girshick
Inmar Givoni
Tobias Glasmachers
Xavier Glorot
Nico Goernitz
Vibhav Gogate
Jacob Goldberger
Anna Goldenberg
Matthew Golub
Vicenç Gómez
Manuel Gomez Rodriguez
Mehmet Gönen
Boqing Gong
Pinghua Gong
Javier Gonzalez
Siddharth Gopal
Aditya Gopalan
Nakul Gopalan
Prem Gopalan
Raghuraman Gopalan
Matt Gormley
Dilan Gorur
Sergiu Goschin
Stephen Gould
Stephan Gouws
Amit Goyal
Navin Goyal
Agnieszka Grabska-Barwinska
Hans-Peter Graf
Ben Graham
Alexandre Gramfort
Kristen Grauman
Edouard Grave
Mihajlo Grbovic
Edward Grefenstette
Shane Griffith
Roger Grosse
Marek Grzes
Quanquan Gu
Shiyuan Gu
Arthur Guez
Sudipto Guha
Rajarshi Guhaniyogi
Vincent Guigue
Caglar Gulcehre
Fangjian Guo
Shengbo Guo
Yandong Guo
Yuhong Guo
Zhaohan Guo
Abhinav Gupta
Maya Gupta
Minh Ha Quang
Michael Habeck
Amaury Habrard
Hirotaka Hachiya
Raia Hadsell
Ben Haeffele
Stefan Haeusler
Gholamreza Haffari
Patrick Haffner
Assaf Hallak
Jonas Hallgren
Will Hamilton
Jihun Hamm
Fred Hamprecht
Fang Han
Shaobo Han
Lars Hansen
Satoshi Hara
Tatsuya Harada
Zaid Harchaoui
Moritz Hardt
Maayan Harel
Bharath Hariharan
Hamed Hassani
Kohei Hatano
Soren Hauberg
Kohei Hayashi
Tamir Hazan
Di He
He He
Jingrui He
Kaiming He
Xuming He
Creighton Heaukulani
Nicolas Heess
Chinmay Hegde
Hoda Heidari
Matthias Hein
Uri Heinemann
Markus Heinonen
Saied Hemati
James Henderson
Guillaume Hennequin
James Hensman
Karl Moritz Hermann
Michiel Hermans
Daniel Hernandez-Lobato
Jose Miguel Hernandez-Lobato
Alfred Hero
Aaron Hertzmann
Felix Hill
Hideitsu Hino
Jun-ichiro Hirayama
Chien-Ju Ho
Nhat Ho
Qirong Ho
Minh Hoai
Toby Hocking
Matt Hoffman
Jake Hofman
Thomas Hofmann
Steven Hoi
Sean Holden
Seunghoon Hong
Antti Honkela
Jean Honorio
Inbal Horev
Reshad Hosseini
Jonathan How
Cho-Jui Hsieh
Chun-Nan Hsu
Daniel Hsu
Xiaolin Hu
Yuening Hu
Bert Huang
Furong Huang
Heng Huang
Jim Huang
Junzhou Huang
Ruitong Huang
Shuai Huang
Tzu-Kuo Huang
Tzu-Kuo Huang
Manfred Huber
Eyke Huellermeier
Jonathan Huggins
Michael Hughes
Jonathan Hunt
Van Anh Huynh-Thu
Aapo Hyvarinen
Morteza Ibrahimi
Tsuyoshi Ide
Christian Igel
Alexander Ihler
Kazushi Ikeda
Shiro Ikeda
Giacomo Indiveri
Katsuhiko Ishiguro
Masakazu Ishihata
Rishabh Iyer
Shahin Jabbari
Laurent Jacob
Abigail Jacobs
Robert Jacobs
Max Jaderberg
Herbert Jaeger
Martin Jaggi
Ashesh Jain
Prateek Jain
Suyog Jain
Viren Jain
Ragesh Jaiswal
Navdeep Jaitly
Michael James
Varun Jampani
Jeremy Jancsary
Majid Janzamin
Ajay Jasra
Pratik Jawanpuria
Dinesh Jayaraman
Bruno Jedynak
Ming Ji
Qiang Ji
Shuiwang Ji
Jiaya Jia
Yangqing Jia
Zhaoyin Jia
Ke Jiang
Jiashun Jin
Wittawat Jitkrittum
Matthew Johnson
Vladimir Jojic
Galin Jones
Shalmali Joshi
Pooria Joulani
Frederic Jurie
Pasi Jyl?nki
Ata Kaban
Hachem Kadri
Leslie Kaelbling
Satyen Kale
Yuri Kalnishkan
Alexandros Kalousis
Hirokazu Kameoka
Ulugbek Kamilov
Varun Kanade
Motonobu Kanagawa
Takafumi Kanamori
Pallika Kanani
Joerg Kappes
Purushottam Kar
Theofanis Karaletsos
Nikos Karampatziakis
Masayuki Karasuyama
Yan Karklin
Dimitris Karlis
Zohar Karnin
Samuel Kaski
Itamar Katz
Yoshinobu Kawahara
Jaya Kawale
Yiping Ke
Mikaela Keller
Kristian Kersting
Charles Kervrann
Azadeh Khaleghi
Aditya Khosla
Tushar Khot
Bryan Kian Hsiang Low
Martin Kiefel
Dongho Kim
Eun-Sol Kim
Gunhee Kim
Hyunwoo Kim
Jiseob Kim
Kye-Hyeon Kim
Minje Kim
Myunghwan Kim
Saehoon Kim
Seungyeon Kim
Sun Kim
Taehwan Kim
YongDeok Kim
Young-Han Kim
Akisato Kimura
Irwin King
Katherine Kinnaird
Ryan Kiros
Sergey Kirshner
Akihiro Kishimoto
Jyrki Kivinen
Negar Kiyavash
Hedvig Kjellstr?m
Marius Kloft
Kei Kobayashi
Jens Kober
Jukka Kohonen
Iasonas Kokkinos
Mladen Kolar
Andrey Kolobov
Mamoru Komachi
George Konidaris
Arnd Konig
Piotr Koniusz
Palla Konstantina
Aryeh Kontorovich
Wouter Koolen
Anoop Korattikara
Wojciech Kotlowski
Adriana Kovashka
Matthieu Kowalski
Oluwasanmi Koyejo
Mark Kozdoba
Bartosz Krawczyk
Akshay Krishnamurthy
Dilip Krishnan
Balaji Krishnapuram
Oliver Kroemer
Florent Krzakala
Alp Kucukelbir
Alex Kulesza
Brian Kulis
Abhishek Kumar
M. Pawan Kumar
Sanjiv Kumar
Gautam Kunapuli
Anshul Kundaje
Matt Kusner
Finn Kuusisto
Vitaly Kuznetsov
Branislav Kveton
Suha Kwak
Theo Kypraios
Anastasios Kyrillidis
Prashanth LA
Simon Lacoste-Julien
Kevin Lai
Balaji Lakshminarayanan
Alex Lamb
Diane Lambert
Tian Lan
Yanyan Lan
Marc Lanctot
Loic Landrieu
Niels Landwehr
Tobias Lang
Ni Lao
Longin Jan Latecki
Tor Lattimore
Aurel Lazar
Alessandro Lazaric
Miguel Lazaro-Gredilla
Nevena Lazic
Hai-Son Le
Tam Le
Nicolas Le Roux
Chansoo Lee
Dongryeol Lee
Jason Lee
Joonseok Lee
Juho Lee
Kanghoon Lee
Moontae Lee
Sangkyun Lee
Seunghak Lee
Su-In Lee
Tai Sing Lee
Yong Jae Lee
Yuh-Jye Lee
Leonidas Lefakis
Robert Legenstein
Andreas Lehrmann
Huitian Lei
Jing Lei
Joel Leibo
Chenlei Leng
Ian Lenz
Vincent Lepetit
Guy Lever
Sergey Levine
John Lewis
Chun-Liang Li
Fuxin Li
Jing Li
Limin Li
Nan Li
Ping Li
Weixin Li
Wen Li
Wu-Jun Li
Yu-Feng Li
Yujia Li
Zhenguo Li
Wenzhao Lian
Dawen Liang
Yingyu Liang
Hank Liao
Thibaut Lienart
Timothy Lillicrap
Zhan Wei Lim
Binbin Lin
Scott Linderman
Fredrik Lindsten
Haibin Ling
Michael Littman
Ashok Litwin-Kumar
Bo Liu
Che-Yu Liu
Fei Liu
Guangcan LIU
Han Liu
Ji Liu
Jialiu Liu
Jie Liu
Jun Liu
Qiang Liu
Risheng Liu
Shih-Chii Liu
Wei Liu
Weiwei Liu
Wenyu Liu

REVIEWERS

Xianghang Liu
Xiaoming Liu
Yan Liu
Yashu Liu
Ying Liu
Zhandong Liu
Daniel Lizotte
Felipe Llinares Lopez
Po-Ling Loh
Maria Lomeli
Ben London
Marco Loog
David Lopez-Paz
Fabien Lotte
Xinghua Lou
Daniel Lowd
Wei Lu
Zhengdong Lu
Chris Lucas
Elliot Ludvig
Gediminas Luksys
Dijun Luo
Heng Luo
Michael Lyu
Shiqian Ma
Tengyu Ma
Yian Ma
Yifei Ma
Andrew Maas
Jakob Macke
Lester Mackey
Dougal Maclaurin
Chris Maddison
Sridhar Mahadevan
Vijay Mahadevan
Dhruv Mahajan
Mehrdad Mahdavi
A. Rupam Mahmood
Julien Mairal
Michael Maire
Alan Malek
Hiroshi Mamitsuka
Travis Mandel
Stephan Mandt
Daniel Mankowitz
Shie Mannor
Ioanna Manolopoulou
Vikash Mansinghka
Qi Mao
Jakub Marecek
Andre Marquand
James Martens
Georg Martius
Pekka Marttinen
Jonathan Masci
Tomoko Matsui
Shin Matsushima
Julian McAuley
Andrew McHutchon
James McInerney
Brian McWilliams
Ted Meeds
Chris Meek
Nishant Mehta
Shike Mei
Franziska Meier
Marina Meila
Ron Meir
Deyu Meng
Aditya Menon
Thomas Mensink
Ofer Meshi
Grégoire Mesnil
Jo?o Messias
Elad Mezuman
Andrew Miller
Jeff Miller
David Mimno
Martin Min
Tom Minka
Mehdi Mirza
Baharan Mirzasoleiman
Dunja Mladenic
Andriy Mnih
Hossein Mobahi
Daichi Mochihashi
Joseph Modayil
Abdel-rahman Mohamed
Mahesh Mohan
Gregoire Montavon
Guido Montufar
Juston Moore
Sebastian Moreno
Edward Moroshko
Dmitriy Morozov
Alessandro Moschitti
Youssef Mroueh
Krikamol Muandet
Andreas Mueller
Klaus-Robert Mueller
Sayan Mukherjee
Andres Munoz Medina
Robert Murphy
Iain Murray
Alejandro Murua
Pablo Muse
Sri Nagarajan
Vinod Nair
Hiroshi Nakagawa
Nils Napp
Shravan Narayanamurthy
Nagarajan Natarajan
Sriraam Natarajan
Saketha Nath
Willie Neiswanger
Blaine Nelson
Jelani Nelson
Shamim Nemati
Bernhard Nessler
Gergely Neu
Gerhard Neumann
Behnam Neyshabur
Huy Nguyen
Trung Nguyen
Viet-An Nguyen
Hannes Nickisch
Feiping NIE
Juan Carlos Niebles
Robert Nishihara
Yu Nishiyama
Gang Niu
William Noble
Richard Nock
Yung-Kyun Noh
David Nott
Rob Nowak
Thomas Nowotny
Tim Oates
Brendan O'Connor
Dino Oglic
Junier Oliva
Peder Olsen
Bruno Olshausen
Yew Soon Ong
Takshi Onoda
Francesco Orabona
Pedro Ortega
Mike Osborne
Simon Osindero
Takayuki Osogami
Anton Osokin
Deepti Pachauri
Jason Pacheco
Adam Packer
Benjamin Packer
Aline Paes
Brooks Paige
John Paisley
Ari Pakman
Christopher Pal
Kenneth Pao
Nicolas Papadakis
Tivadar Papai
Dimitris Papailiopoulos
George Papandreou
Ulrich Paquet
Alexandros Paraschos
Ankur Parikh
Il Memming Park
Mijung Park
Sunho Park
Charles Parker
Johannes Partzsch
Pekka Parviainen
Razvan Pascanu
alexandre Passos
Debdeep Pati
Giorgio Patrini
Genevieve Patterson
Edouard Pauwels
Vladimir Pavlovic
Klaus Pawelzik
Jason Pazis
Barak Pearlmutter
Tomi Peltola
Jaakko Peltonen
Jian Peng
Fernando Pereira
Julien Perez
Fernando Perez-Cruz
Gareth Peters
Jonas Peters
Marek Petrik
Gabriel Peyre
David Pfau
Nico Pfeifer
Joseph Pfeiffer
Dinh Phung
Massimo Piccardi
Olivier Pietquin
Natesh Pillai
Jonathan Pillow
Joelle Pineau
Bilal Piot
Matteo Pirotta
Hamed Pirsiavash
Patrick Pletscher
Eftychios Pnevmatikakis
Florian Pokorny
Jan Poland
Daniel Polani
Jean-Baptiste Poline
Pascal Poupart
Daniel Povey
Dennis Prangle
Doina Precup
Alexandre Proutiere
Jay Pujara
Guido Pusiol
Shaan Qamar
Yanjun Qi
Xingye Qiao
Tao Qin
Zengchang Qin
Qichao Que
Joaquin Quinonero-Candela
Maxim Rabinovich
Andrew Rabinovitch
Neil Rabinowitz
Vladan Radosavljevic
Ali Rahimi
Piyush Rai
Alexander Rakhlin
Liva Ralaivola
Deepak Ramachandran
Peter Ramadge
Subramanian Ramamoorthy
Karthik Raman
Vignesh Ramanathan
Harish Ramaswamy
Arti Ramesh
Jan Ramon
Fabio Ramos
Roland Ramsahai
Rajesh Ranganath
Shyam Rangapuram
Marc'Aurelio Ranzato
Magnus Rattray
Sujith Ravi
Balaraman Ravindran
Debajyoti Ray
Soumya Ray
Vikas Raykar
Benjamin Reddy
Scott Reed
Khaled Refaat
Roi Reichart
Thodoris Rekatsinas
Marcello Restelli
Danilo Rezende
Emile Richard
Jonas Richiardi
Sebastian Riedel
Irina Rish
Yaacov Ritov
St?phane Robin
Daniel Robinson
Erik Rodner
Karl Rohe
Marcus Rohrbach
Jaldert Rombouts
Teemu Roos
Romer Rosales
Nir Rosenfeld
Arun Ross
Fabrice Rossi
Afshin Rostamizadeh
Volker Roth
Constantin Rothkopf
Tim Roughgarden
Juho Rousu
Daniel Roy
Jean-Francis Roy
Nicholas Roy
Alessandro Rudi
Nicholas Ruozzi
Alexander Rush
Olga Russakovsky
Bryan Russell
Andreas Ruttor
Paul Ruvolo
Yunus Saatchi
Sivan Sabato
Ashish Sabharwal
Mrinmaya Sachan
Mohammad Amin Sadeghi
Peter Sadowski
Ankan Saha
Avishek Saha
Hiroto Saigo
Tara Sainath
Jun Sakuma
Venkatesh Saligrama
Mathieu Salzmann
Rajhans Samdani
Sujay Sanghavi
Maxime Sangnier
Guido Sanguinetti
Aswin Sankaranarayanan
Sriram Sankararaman
Scott Sanner
Vitor Santos Costa
George Saon
Simo Särkkä
Anand Sarwate
Richard Savage
Pierre-André Savalle
Cristina Savin
Christoph Sawade
Andrew Saxe
Stefan Schaal
Tom Schaul
Aaron Schein
Katya Scheinberg
Bernt Schiele
Tanya Schmah
J?rgen Schmidhuber
Mark Schmidt
Mikkel Schmidt
Uwe Schmidt
Jeff Schneider
Francois Schnitzler
Thomas Schoen
Hannes Schulz
Haim Schweitzer
Holger Schwenk
Alex Schwing
Erwan Scornet
D Sculley
Michele Sebag
Hanie Sedghi
Shahin Sefati
Sundararajan Sellamanickam
Bart Selman
Mike Seltzer
Joao Semedo
Srinivasan Sengamedu
Andrew Senior
Thomas Serre
Jun Sese
Simon Setzer
Eleni Sgouritsa
Ben Shababo
Patrick Shafto
Amar Shah
Mohak Shah
Nihar Shah
Bobak Shahriari
Greg Shaknarovich
Uri Shalit
Cosma Shalizi
Ohad Shamir
Chung-chieh Shan
Nataliya Shapovalova
Roshan Shariff
Viktoriia Sharmanska
Tatyana Sharpee
James Sharpnack
Abdul Saboor Sheikh
Samira Sheikhi
Daniel Sheldon
Christian Shelton
Jacquelyn Shelton
Bin Shen
Chunhua Shen
Li Shen
Lei Shi
Qinfeng Shi
Motoki Shiga
Nobuyuki Shimizu
HyunJung Shin
Jinwoo Shin
Shigeru Shinomoto
Pannaga Shivaswamy
Jon Shlens
Lavi Shpigelman
Suyash Shringarpure
Abhinav Shrivastava
Anshumali Shrivastava
Si Si
Leonid Sigal
Tomi Silander
Ricardo Silva
Khe Chai Sim
Karen Simonyan
Ozgur Simsek
Vikas Sindhwani
Sameer Singh
Vikas Singh
Adish Singla
Kaushik Sinha
Mathieu Sinn
Fabian Sinz
Scott Sisson
Martin Slawski
Kevin Small
Paris Smaragdis
Cristian Sminchisescu
Jasper Snoek
Anthony Man-Cho So
Richard Socher
Eric Sodomka
Jascha Sohl-Dickstein
Kihyuk Sohn
Kyung-Ah Sohn
Arno Solin
Justin Solomon
Mahdi Soltanolkotabi
Fritz Sommer
Diego Sona
Hyun Oh Song
Yangqiu Song
Daniel Soudry
Luciano Spinello
Pablo Sprechmann
Srikrishna Sridhar
Karthik Sridharan
Vivek Srikumar
Nitish Srivastava
Rupesh Srivastava
Sanvesh Srivastava
Bradly Stadie
Oliver Stegle
Daniel Steinberg
Florian Steinke
Ingo Steinwart
Philip Sterne
Bob Stine
Greg Stoddard
Dmitry Storcheus
Heiko Strathmann
Karl Stratos
Julian Straub
Andreas Stuhlmüller
Hang Su
Devika Subramanian
Erik Sudderth
Mahito Sugiyama
Uygar Sumbul
Dennis Sun
Min Sun
Siqi Sun
Yizhou Sun
Yuekai Sun
Matyas Sustik
Dougal Sutherland
Charles Sutton
Johan Suykens
Joe Suzuki
Adith Swaminathan
Kevin Swersky
Pawel Swietojanski
Zeeshan Syed
Adam Sykulski
Zoltan Szabo
Sandor Szedmak
Christian Szegedy
Arthur Szlam
Balazs Szorenyi
Yasuo Tabei
Nima Taghipour
Martin Takac
Akiko Takeda
Takashi Takenouchi
Koh Takeuchi
Eiji Takimoto
Partha Talukdar
Erik Talvitie
Ameet Talwalkar
Mingkui Tan
Cheng Tang
Kui Tang
Yichuan Tang
Alex Tank
Dacheng Tao
Danny Tarlow
Yuval Tassa
Nikolaj Tatti
Graham Taylor
Matthew Taylor
Matus Telgarsky
Lucas Theis
Georgios Theocharous
Evangelos Theodorou
Bertrand Thirion
Owen Thomas
Nicolas Thome
Zhou Tianyi
Robert Tillman
Jo-Anne Ting
Diego Tipaldi
Ivan Titov
Michalis Titsias
Sinisa Todorovic
Ilya Tolstikhin
Marc Tommasi
Hanghang Tong
Kari Torkkola
Antonio Torralba
Alexander Toshev
Ivana Tosic
Panos Toulis
Kristina Toutanova
Thomas Trappenberg
Volker Tresp
Felipe Trevizan
Manolis Tsakiris
Sebastian Tschiatschek
Konstantinos Tsianos
Evgeni Tsivtsivadze
Zhuowen Tu
Richard Turner
Ryan Turner
Hemant Tyagi
Stephen Tyree
Lyle Ungar
Benigno Uria
Daniel Urieli
Ruth Urner
Raquel Urtasun
Nicolas Usunier
Daniel Vainsencher
Michal Valko
Harri Valpola
Jan-Willem van de Meent
Ewout van den Berg
Guy Van den Broeck
Hado van Hasselt
Herke van Hoof
Hasta Vanchinathan
Bart Vandereycken
Robert Vandermeulen
Vincent Vanhoucke
Martijn Vanotterlo
Gael Varoquaux
Lav Varshney
Eleni Vasilaki
Sergei Vassilvitskii
Sandro Vega Pons
Shankar Vembu
Joel Veness
Suresh Venkatasubramanian
Dan Ventura
Deepak Venugopal
Nakul Verma
Paul Vernaza
Jean-Philippe Vert
Alexander Vezhnevets
Ricardo Vigario
Matthieu Vignes
Sudheendra Vijayanarasimhan
Mattias Villani
Brett Vinch
Vibhav Vineet
Oriol Vinyals
Joshua Vogelstein
Maksims Volkovs
Ulrike von Luxburg
Vincent Vu
Slobodan Vucetic
Marivate Vukosi
Ed Vul
Willem Waegeman
Stefan Wager
Niklas Wahlstroem
Christian Walder
Byron Wallace
Thomas Walsh
Chaohui Wang
Chong Wang
Hua Wang
Huahua Wang
Huayan Wang
Jack Wang
Jialei Wang
Jie Wang
JingDong Wang
Joseph Wang
Jun Wang
Jun Wang
Lei Wang
Li Wang
Lidan Wang
Liming Wang
Meihong Wang
Mengdi Wang
Shaojun Wang
Shusen Wang
Sida Wang
Wei Wang
Weiran Wang
Xiaogang Wang
Yang Wang
Yi Wang
Yining Wang
Yiwen Wang
Yu-Xiang Wang
Yuyang Wang
Zhaoran Wang
Zheng Wang
Zhuo Wang
Ziyu Wang
Zuoguan Wang
David Warde-Farley
Takashi Washio
Larry Wasserman
Kazuho Watanabe
Shinji Watanabe
Fabian Wauthier
Greg Wayne
Geoff Webb
Kai Wei
Lu Wei
Markus Weimer
David Weiss
Yair Weiss
Zheng Wen
Tim Weninger
Tomas Werner
Adam White
Martha White
Michael Wick
Jenna Wiens
Marco Wiering
Rebecca Willett
Chris Williams
Jason Williams
Andrew Wilson
Ole Winther
David Wipf
Jenn Wortman Vaughan
John Wright
Jianxin Wu
Lei Wu
Qiang Wu
Si Wu
Steven Wu
Teresa Wu
Yifan Wu
Yirong Wu
Lirong Xia
Jing Xiang
Shuo Xiang
Jianxiong Xiao
Lin Xiao
Yuan Xiaotong
Lexing Xie
Yu Xin
Bo Xiong
Chang Xu
Hongteng Xu
Jun Xu
Junming Xu
Li Xu
Linli Xu
Miao Xu
Min Xu
Minjie Xu
Zhixiang (Eddie) Xu
Oksana Yakhnenko
Makoto Yamada
Shuicheng Yan
Keiji Yanai
Pinar Yanardag
Haiqin Yang
Jaewon Yang
Jianchao YANG
Jimei Yang
Liu Yang
Michael Yang
Ming-Hsuan Yang
Scott Yang
Shuang Yang
Shulin Yang
Zhi Yang
Zhirong Yang
Zichao Yang
Angela Yao
Hengshuai Yao
Christopher Yau
Nan Ye
Dit Yan Yeung
Florian Yger
Jinfeng Yi
Xinyang Yi
Scott Yih
Junming Yin
Yiming Ying
Chang D. Yoo
Sungroh Yoon
Jason Yosinski
Chong You
Chun-Nam Yu
Felix Yu
Hsiang-Fu Yu
Yang Yu
Yaoliang Yu
Yisong Yue
Hyokun Yun
Jeong-Min Yun
François Yvon
Bianca Zadrozny
Stefanos Zafeiriou
Thorsten Zander
Giovanni Zappella
Mikhail Zaslavskiy
Aonan Zhang
Changshui ZHANG
Chicheng Zhang
Chiyuan Zhang
Haichao Zhang
Hao Helen Zhang
Jiji Zhang
Junping Zhang
Kai Zhang
Kun Zhang
Lei Zhang
Lijun Zhang
Min-Ling Zhang
Nevin Zhang
Peng Zhang
Ruiliang Zhang
Sen ZHANG
Shunan Zhang
Wen-hao Zhang
Xianxing Zhang
Yichuan Zhang
Yu Zhang
Yuchen Zhang
Zheng Zhang
Zhihua Zhang
Peilin Zhao
Tuo Zhao
Zheng Zhao
Shandian Zhe
Yi Zhen
Xun Zheng
Wenliang Zhong
Chunxiao Zhou
Feng Zhou
Guang-Tong Zhou
Jiayu Zhou
Joey Tianyi Zhou
Ke Zhou
Mingyuan Zhou
Shuheng Zhou
Xuezhong Zhou
Shenghuo Zhu
Jinfeng Zhuang
Hankz Hankui Zhuo
Brian Ziebart
Onno Zoeter
Daniel Zoran
Will Zou
Or Zuk
Alon Zweig

AUTHOR INDEX

AUTHOR INDEX

AUTHOR INDEX

AUTHOR INDEX

AUTHOR INDEX

AUTHOR INDEX