Neural Information
Processing Systems

NIPS 2016

CONFERENCE BOOK

BARCELONA

SPAIN

NIPS 2016
CONFERENCE AT A GLANCE

Neural Information Processing Systems

MONDAY DECEMBER 5TH

Tutorial 1	8:30 - 10:30 am
Coffee break	10:30 - 11:00 am
Tutorial 2	11:00 - 1:00 pm
Lunch on your own.	1:00 - 2:30 pm
Tutorial 3	2:30 - 4:30 pm
Coffee break	4:30 - 5:00 pm
Opening Remarks	5:00 - 5:30 pm
Invited talk, Yann LeCun (Facebook & NYU)	5:30 - 6:20 pm
Reception and Posters	6:30 - 9:30 pm

TUESDAY DECEMBER 6TH

Drew Purves (DeepMind)	9:00 - 9:50 am
Award talk	9:50 - 10:10 am
Coffee break	10:10 - 10:40 am
Parallel tracks: **Clustering/Graphical Models**	10:40 - 12:20 pm
Lunch on your own and poster viewing	12:20 - 3:00 pm
Saket Navlakha (Salk Institute)	3:00 - 3:50 pm
Coffee break	3:50 - 4:20 pm
Parallel tracks: **Deep Learning/Theory**	4:20 - 6:00 pm
Poster session & Demonstrations	6:00 - 9:30 pm

WEDNESDAY DECEMBER 7TH

Kyle Cranmer (NYU)	9:00 - 9:50 am
Award talk	9:50 - 10:10 am
Coffee break	10:10 - 10:40 am
Parallel tracks: **Algorithms/Applications**	10:40 - 12:20pm
Lunch on your own and poster viewing	12:20 - 3:00 pm
Marc Raibert (Boston Dynamics)	3:00 - 3:50 pm
Coffee Break	3:50 - 4:20 pm
Parallel tracks: **Deep Learning /Optimization**	4:20 - 6:00 pm
Poster session & Demonstrations	6:00 - 9:30 pm

THURSDAY DECEMBER 8TH

Irina Rish (IBM)	9:00 - 9:50 am
Susan Holmes (Stanford)	9:50 - 10:40 am
Coffee break	10:40 - 11:10 am
Parallel tracks:	
Interpretable Models/Neuroscience/Cognitive	11:10 - 12:20 pm
Lunch on your own	12:20 - 2:00 pm
Symposia	2:00 - 4:00 pm
coffee break	4:00 - 4:30 pm
Symposia	4:30 - 6:30 pm
Light dinner	6:30 - 7:30 pm
Symposia	7:30 - 9:30 pm

FRIDAY & SATURDAY DECEMBER 9TH & 10TH

Workshop Sessions	8 am - 6:30 pm
Check workshop schedules for actual start times	
Coffee break	10:30 – 11:00 am
Coffee break	3:00 – 3:30 pm

Contents

ORGANIZING COMMITTEE

General Chairs: Daniel D Lee (U. of Pennsylvania)
Masashi Sugiyama (U. of Tokyo)
Program Chairs: Ulrike von Luxburg (U. of Tübingen)
Isabelle Guyon (Clopinet)
Tutorials Chair: Joelle Pineau (McGill U.)
Hanna Wallach (Microsoft)

Workshop Chairs: Ralf Herbrich (Amazon)
Demonstration Chair: Raia Hadsell (DeepMind)
Publications Chair and Electronic Proceedings Chair
Roman Garnett (Washington U. St. Louis)
Program Managers: Krikamol Muandet (Mahidol U. and MPI)
Behzad Tabibian (MPI)

NIPS FOUNDATION OFFICERS & BOARD MEMBERS

PRESIDENT
Terrence Sejnowski, The Salk Institute

TREASURER
Marian Stewart Bartlett, Apple

SECRETARY
Michael Mozer, UC Boulder

EXECUTIVE DIRECTOR
Mary Ellen Perry, The Salk Institute

LEGAL ADVISOR
David Kirkpatrick

IT DIRECTOR
Lee Campbell, The Salk Institute

EXECUTIVE BOARD
Zoubin Ghahramani, Univ. of Cambridge
Corinna Cortes, Google Research
Léon Bottou, Microsoft Research
Chris J.C. Burges, Microsoft Research
Neil D. Lawerence, Univ. of Sheffield
Fernando Pereira, Google Research
Max Welling, Univ. of Amsterdam

ADVISORY BOARD
Peter Bartlett, Queensland Univ., UC Berkley
Sue Becker, McMaster Univ., Ontario, Canada
Yoshua Bengio, Univ. of Montreal, Canada
Jack Cowan, Univ. of Chicago
Thomas G. Dietterich, Oregon State Univ.
Stephen Hanson, Rutgers Univ.
Michael I. Jordan, Univ. of California, Berkeley
Michael Kearns, Univ. of Pennsylvania
Scott Kirkpatrick, Hebrew Univ., Jerusalem
Daphne Koller, Stanford Univ.
John Lafferty, Univ. of Chicago
Todd K. Leen, Oregon Health & Sciences Univ.
Richard Lippmann, MIT
Bartlett Mel, Univ. of Southern California
John Moody, UC Berkeley & Portland
John C. Platt, Google
Gerald Tesauro, IBM Watson Labs
Sebastian Thrun, Stanford Univ.
Dave Touretzky, Carnegie Mellon Univ.
Lawrence Saul, Univ. of California, San Diego
Bernhard Schölkopf, MP, Tübingen/Stuttgart
Dale Schuurmans, Univ. of Alberta, Canada
John Shawe-Taylor, Univ. College London
Sara A. Solla, Northwestern Univ. Med. School
Yair Weiss, Hebrew Univ. of Jerusalem
Chris Williams, Univ. of Edinburgh
Rich Zemel, Univ. of Toronto

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 Director
Mary Ellen Perry - Executive Dir.
Terrance Gaines - Administrator
Susan Perry - Volunteer Mgr.
Jen Perry - Administrator
Ramona Marchand - Administrator

PROGRAM COMMITTEE

Emmanuel Abbe (Princeton Univ.)
Alekh Agarwal (Microsoft)
Anima Anandkumar (UC Irvine)
Chloé-Agathe Azencott (MINES ParisTech)
Shai Ben-David (Univ. Waterloo)
Alina Beygelzimer (Yahoo Research)
Jeff Bilmes (Univ. of Washington (Seattle)
Gilles Blanchard (Univ. of Potsdam)
Matthew Blaschko (KU Leuven)
Tamara Broderick (MIT)
Sebastien Bubeck (Princeton)
Alexandra Carpentier (Univ. Potsdam)
Miguel Carreira-Perpinan (UC Merced)
Kamalika Chaudhuri (UC San Diego)
Gal Chechik (Google (Bar-Ilan Univ.)
Kyunghyun Cho (New York Univ.)
Aaron Courville (Univ. of Montreal)
Koby Crammer (Technion)
Florence d'Alché-Buc (Telecom Paris Tech)
Arnak Dalalyan (ENSAE ParisTech)
Marc Deisenroth (Imperial College London)
Francesco Dinuzzo (Amazon)
Finale Doshi-Velez (Harvard)
Ran El-Yaniv (Technion)
Hugo Jair Escalante (INAOE)
Sergio Escalera (Univ. of Barcelona)
Maryam Fazel (Univ. of Washington)
Aasa Feragen (Univ. of Copenhagen)
Rob Fergus (New York Univ.)
Xiaoli Fern (Oregon State Univ.)
Francois Fleuret (Idiap Research Institute)
Surya Ganguli (Stanford)
Peter Gehler (Univ. of Tübingen)
Claudio Gentile (DiSTA (Universita dell'Insubria)
Lise Getoor (UC Santa Cruz)
Mark Girolami (Imperial College London)
Amir Globerson (Tel Aviv Univ.)
Yoav Goldberg (Bar Ilan Univ.)
Manuel Gomez (Max Planck Institute)
Yves Grandvalet (Univ. of Compiègne & CNRS)
Moritz Grosse-Wentrup (MPI)
Zaid Harchaoui (Univ. of Washington)
Moritz Hardt (Google)
Matthias Hein (Saarland Univ.)
Philipp Hennig (MPI IS Tübingen)
Frank Hutter (Univ. of Freiburg)
Prateek Jain (Microsoft Research)
Navdeep Jaitly (Google Brain)
Stefanie Jegelka (MIT)
Samuel Kaski (Aalto Univ.)
Koray Kavukcuoglu (DeepMind)
Jens Kober (TU Delft)
Samory Kpotufe (Princeton Univ.)
Sanjiv Kumar (Google Research)
James Kwok (Hong Kong Univ.)
Simon Lacoste-Julien (U. of Montreal)
Christoph Lampert (IST Austria)
Hugo Larochelle (Twitter)
Francois Laviolette (L'Université Laval)
Honglak Lee (Univ. of Michigan)
Christoph Lippert (Human Longevity)
Po-Ling Loh (UW-Madison)
Phil Long (Sentient Technologies)
Jakob Macke (Caesar Bonn)
Julien Mairal (Inria)
Shie Mannor (Technion)
Marina Meila (Univ. of Washington)
Claire Monteleoni (George) Washington Univ.)
Remi Munos (DeepMind)
Guillaume Obozinski (Ecole Paris)
Cheng Soon Ong (Data61 and ANU)
Francesco Orabona (Stony Brook U.)
Fernando Perez-Cruz (Universidad)
Carlos III de Madrid (Bell Labs (Nokia))
Jonathan Pillow (Princeton Univ.)
Doina Precup (McGill Montreal)
Alain Rakotomamonjy (Univ. of Rouen)
Manuel Rodriguez (Max Planck Inst.)
Rómer Rosales (Linkedin)
Lorenzo Rosasco (U. of Genova (MIT)
Sivan Sabato (Ben-Gurion Univ.)
Mehreen Saeed (FAST (Univ of CES)
Ruslan Salakhutdinov (CMU)
Purnamrita Sarkar (Univ. T. Austin)
Fei Sha (USC)
Ohad Shamir Weizmann (Inst of Science)
Jonathon Shlens (Google Brain)
David Sontag (New York Univ.)
Suvrit Sra (MIT)
Karthik Sridharan (Cornell Univ.)
Bharath Sriperumbudur (Pennsylvania State Univ.)
Erik Sudderth (Brown Univ.)
Csaba Szepesvari (Univ. of Alberta)
Graham Taylor (Univ. of Guelph)
Ambuj Tewari (Univ. of Michigan)
Ruth Urner (MPI Tübingen)
Benjamin Van Roy (Stanford)
Jean-Philippe Vert (MINES ParisTech)
Bob Williamson (Data61 and ANU)
Jennifer Wortman (Vaughan) Microsoft Research)
Lin Xiao (Microsoft Research)
Kun Zhang (CMU)

NIPS gratefully acknowledges the generosity of those individuals and organizations who have provided financial support for the NIPS 2016 conference. Their financial support enables us to sponsor student travel and participation, general support to host the conference, and the volunteers who assist during NIPS.

PLATINUM SPONSORS

American International Group, Inc. (AIG)'s vision is to become our clients' most valued insurer. For the past 100 years, we have been a leading international insurance organisation serving customers in more than 100 countries and jurisdictions. AIG serves commercial, institutional, and individual customers through one of the most extensive worldwide property-casualty networks of any insurer. At AIG, we believe that harnessing the power of machine learning and deep learning techniques is essential to go beyond merely generating new insights from data but also to systematically enhance individual human judgement in real business contexts. If you are also feeling passionate about being a catalyst for evidence-based decision making across the world, let's connect!

Apple revolutionized personal technology with the introduction of the Macintosh in 1984. Today, Apple leads the world in innovation with iPhone, iPad, the Mac & Apple Watch. Apple's three software platforms—iOS, OS X & watchOS—provide seamless experiences across all Apple devices & empower people with breakthrough services including the App Store, Apple Music, Apple Pay & 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.

The Audi Group, with its brands Audi, Ducati and Lamborghini, is one of the most successful manufacturers of automobiles and motorcycles in the premium segment. It is present in more than 100 markets worldwide and produces at 16 locations in twelve countries. In the second half of 2016, the production of the Audi Q5 will start in San José Chiapa (Mexico). 100-percent subsidiaries of AUDI AG include quattro GmbH (Neckarsulm), Automobili Lamborghini S.p.A. (Sant'Agata Bolognese, Italy) and Ducati Motor Holding S.p.A. (Bologna, Italy).

Citadel is 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.

DCVC is a venture capital fund that invests in entrepreneurs applying deep compute, big data and IT infrastructure technologies to transform giant industries. DCVC and its principals have backed brilliant people changing global-scale businesses for 20+ years, helping create billions of dollars of wealth for these entrepreneurs while also making the world a markedly better place.

At DeepMind, our mission is to solve intelligence and then use that to make the world a better place. Our motivation in all we do is to maximise the positive and transformative impact of AI. We believe that AI should ultimately belong to the world, in order to benefit the many and not the few, and we steadfastly research, publish and implement our work to that end.

Didi Chuxing is the world's largest comprehensive one-stop mobile transportation platform. The company offers a full range of mobile technology-based transportation options for close to 300 million users across over 400 Chinese cities, including taxi hailing, private car hailing, Hitch, Chauffeur, DiDi Bus, DiDi Test Drive, and DiDi Enterprise Solutions. In August 2016, Didi Chuxing acquired Uber China. Didi Chuxing is also growing in global markets. In particular, in the United States, Didi provides Private car hailing services in more than 200 cities via our strategic alliance with Lyft.

Facebook's mission is to give people the power to share and make the world more open and connected-this requires constant innovation. At Facebook, we believe the most interesting research questions are derived from real world problems. Working on cutting edge research with a practical focus, we push product boundaries while finding new ways to collaborate with the academic community.

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.

Intel, the world leader in silicon innovation, develops technologies, products and initiatives to continually advance how people work and live. Intel's innovations in cloud computing, data center, IoT, & PC solutions are powering the smart and connected digital world. Learn more about Intel's vision for the future of artificial intelligence at www.intel.com/ai.

At KLA-Tencor, we research, develop, and manufacture the world's most advanced inspection and measurement equipment for the semiconductor industry. We enable the digital age by pushing the boundaries of optics, sensors, image processing, machine learning and computing technologies, creating systems capable of finding Nano-scale defects at 50 GB/second data rates. If you are passionate in driving R&D in advanced deep learning, 3D sensor fusion, Bayesian & Physics based Machine Learning, advanced Neural & HPC architectures, then KLA-Tencor is the place for you.

PLATINUM SPONSORS

At Microsoft, we aim to empower every person and every organization on the planet to achieve more. We care deeply about having a global perspective and making a difference in lives and organizations in all corners of the planet. This involves playing a small part in the most fundamental of human activities: Creating tools that enable each of us along our journey to become something more. Our mission is grounded in both the world in which we live and the future we strive to create. Today, we live in a mobile-first, cloud-first world, and we aim to enable our customers to thrive in this world.

Tencent, Inc. is China's largest and most used Internet service portal. Tencent's mission to enhance the quality of human life through Internet services. Presently, Tencent provides social platforms and digital content services under the "Connection" Strategy. Tencent's leading Internet platforms in China – QQ (QQ Instant Messenger), Weixin/WeChat, QQ.com, QQ Games, Qzone, and Tenpay – have brought together China's largest Internet community, to meet the various needs of Internet users including communication, information, entertainment, financial services and others.

Winton is a British-based global investment management and data technology company. We believe the best approach to investing is the application of the scientific method. Combining statistics and mathematical modelling with cutting-edge technology, we create and evolve intelligent systems to invest in global financial markets, on behalf of our clients around the world. Winton was established in 1997 by David Harding, a pioneer in the development of investment systems who has founded two of the world's most successful investment management firms.

GOLD SPONSORS

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.

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.

Amazon.com strives to be Earth's most customer-centric company where people can find and discover virtually anything they want to buy online. The world's brightest technology minds come to Amazon.com to research and develop technology that improves the lives of shoppers, sellers and developers.

AtlaSense is an A.I. platform for modern legal departments who want to take control of their data. Whether for eDiscovery, Contract Gathering or Records Management, AtlaSense virtually finds information anywhere and automatically classifies your files based on your preferences.

Baidu, Inc. is the leading Chinese language Internet search provider. As a technology-based media company, Baidu provides the best and most equitable way for people to find they're looking for. In addition to serving Internet search users, Baidu provides an effective platform for businesses to reach potential customers. Baidu's ADSs trade on the NASDAQ Global Select Market under the symbol "BIDU"

Criteo Research is pioneering innovations in computational advertising. As the center of scientific excellence in the company, Criteo Research delivers both fundamental and applied scientific leadership through published research, product innovations and new technologies powering the company's products.

IBM Research embraces Grand Challenges like Deep Blue and Watson, and is continually extending Watson's Cognitive capabilities to enable real-world transformations throughout various businesses. It is home to 3000+ researchers including 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 awakened the world to computer graphics when it invented the GPU in 1999. Researchers utilize GPUs to advance the frontiers of science with high performance computing. Industry and academia leverage them for deep learning to make groundbreaking improvements across a variety of applications including image classification, video analytics and speech recognition. www.nvidia.co.uk

Founded in 2007 by leading machine learning scientists, The Voleon Group designs, develops, and implements advanced technology for investment management. We are committed to solving large-scale financial prediction problems with statistical machine learning.

SILVER SPONSORS

Qihoo 360 Technology Co. Ltd. is a leading Internet platform company in China as measured by our active user base of 496 million active Internet users. Recognizing security as a fundamental need of all Internet and mobile users, Qihoo 360 built a large user base by offering comprehensive, effective, cloud-based and user-friendly Internet and mobile security products. Qihoo 360 is one of top three Internet companies measured by user base.

Automotive Safety Technologies GmbH was founded in 2009 in Gaimersheim/Ingolstadt. Our areas of competency cover the full spectrum of development for integrated safety systems. From function and software development to required simulation and testing competency through to tool and process development in the area of integrated safety – from a single source.

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. Our over 4,800 technologists work to define, architect, build and deploy complete systems and solutions that anticipate and fulfill our clients' needs and market demands.

Best known as the world's number one automotive supplier, Bosch also has a broad product portfolio in industrial technology and consumer goods. In Corporate Research, researchers work on technological breakthroughs such as in software development and autonomous systems. In this way, new ideas are constantly taking shape improving existing products, while opening up entirely new lines of business.

Cubist Systematic Strategies, our systematic investing business, 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. The organization is structured to strongly support investment professionals in their research efforts.

The D. E. Shaw group is a global investment and technology development firm with more than 38 billion in investment capital as of July 1, 2016, and offices in North America, Europe, and Asia. Since our founding in 1988, our firm has earned an international reputation for successful investing based on innovation, careful risk management, and the quality and depth of our staff.

eBay Inc. (NASDAQ: EBAY) is a global commerce leader including the Marketplace, StubHub and Classifieds platforms. We connect millions of buyers and sellers around the world, empowering people and creating opportunity through Connected Commerce. Founded in 1995 in San Jose CA, eBay is one of the world's largest and most vibrant marketplaces for discovering great value and unique selection. In 2015, eBay enabled 82 billion of gross merchandise volume.

FeatureX is a well-funded machine learning startup next to MIT, and we're hiring. We analyze data to extract time-series features related to economic activity, and then use statistical machine learning to build real-time models at various scales - from global macroeconomic activity down to the performance of a single company. One key dataset is satellite imagery, so we're deep into computer vision. Sound interesting? Visit our booth at NIPS or apply at www.featurex.ai.

G-Research researches investment strategies to predict price returns in financial markets. We develop the research and execution platform to deploy these ideas in markets globally. Our Machine Learning team develop techniques and apply them to seek patterns in large, dirty and noisy data sets. Their codebase is at the forefront of machine learning, pushing its boundaries to new and exciting areas!

Hutchin Hill Capital manages approximately 3.4 billion with a global staff of 170. We are a multi-strategy manager focused on liquid investments in fundamental & quantitative strategies. We seek to generate attractive, risk-adjusted returns with zero beta and low correlation to traditional risk assets through investments in four distinct core strategies: equities, credit, macro and quantitative.

Jump Trading is a leading research-focused trading firm that combines sophisticated quantitative research, best-in-class technology, and an entrepreneurial culture across offices in Chicago, New York, London and Singapore. We foster intellectual curiosity and learning so employees can leverage petaflops of computing power and petabytes of data identify trends in global markets across asset classes.

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.

Start Your AI Startup In Canada! Announcing NextAI, a program and innovation hub in Toronto, Canada for individuals and teams who will solve major problems using AI. We select the most promising entrepreneurs and innovators and provide them with access to leading AI tools and scientists, founder development, capital and exposure to the corporate network to turn their ideas into reality.

SILVER SPONSORS

Thirty years ago, Optiver started business as a single trader on the floor of the Amsterdam's options exchange. Today we are at the forefront of trading and technology, employing over 950 Optiverians from over 40 nationalities. We stick to what we're good at: making markets in a wide range of financial products. At Optiver we solve puzzles together over breakfast and spend our time using state-of-the-art data-science, quantitative models and technological systems to improve our Trading. Innovation through quantification is what gets us going.

Qualcomm Technologies, Inc., a wholly-owned subsidiary of Qualcomm Incorporated, operates, along with its subsidiaries, substantially all of Qualcomm's engineering, research and development functions, and substantially all of its products and services businesses, including its semiconductor business, QCT.

Renaissance Technologies is a quantitative hedge fund management company founded by James Simons in 1982 and located in East Setauket, NY. Renaissance has 300 employees, 90 of whom have PhDs in mathematics, physics, statistics, or computer science. The firm's trading is based on models developed through the application of machine learning to massive quantities of financial data.

Rosetta Analytics is a newly formed investment management firm committed to using new data sources and new computational methods to uncover actionable investment signals. Through collaboration and co-investment with our clients, these signals are used to create customized but scalable investment strategies.

As market leader in enterprise application software, SAP (NYSE: SAP) helps companies of all sizes and industries run better. From back office to boardroom, warehouse to storefront, desktop to mobile device – SAP empowers people and organizations to work together more efficiently and use business insight more effectively to stay ahead of the competition.

Sentient has created the largest and most powerful distributed intelligent system in the world. We use this platform to create products which will transform large sectors of the economy, by harnessing massive data sets to solve their most complex problems. Our artificial intelligence (AI) can identify and answer critical questions in new and groundbreaking ways, and act autonomously, while empowering people and businesses to make smarter decisions. Learn more at: sentient.ai.

We imagine breakthroughs in investment management, insurance and related fields by pushing the boundaries of what open source and proprietary technology can do. In the process, we work to help real people. Through our investors, we support the retirements of millions around the world and help fund breakthrough research, education and a wide range of charities and foundations.

As the innovation hub of United Technologies Corp., United Technologies Research Center and its Ireland subsidiary, United Technologies Research Centre Ireland, Ltd., work to develop game-changing technologies and capabilities across the company and collaborate with external research organizations, universities and government agencies globally to push the boundaries of science and technology.

BRONZE SPONSORS

BenevolentAI is a British technology company harnessing the power of AI to enhance and accelerate scientific discovery by turning the world's highly fragmented scientific research data into new insight and usable knowledge that benefits society. Simply put, the company is bringing people and technology together to revolutionise the process of scientific discovery.

Beijing Institute of Big Data Research (BIBDR) is a new institution jointly sponsored by the Peking University and government of Beijing. It is the first big data institution in China that combines education, research, entrepreneurship and government service. Our mission is to developing educational programs for data science in China and as a platform for nurturing new enterprises in big data.

Cheetah Mobile (NYSE:CMCM) is a leading mobile application developer, the #2 largest internet and mobile security corporation in China and the #3 global non-gaming developer on Google Play, with over 2.3 billion downloads times worldwide and 634 million mobile Monthly Active Users.

Datatang is a global data provider. We are the trusted partner to many of the most influential corporations and institutions in the world. At Datatang, we believe we could connect the dots of data by providing ample opportunities for data exchange with diversified resources and ability to work toward new solutions together with our clients.

Disney Research's objective is to drive value across The Walt Disney Company by injecting scientific & technological innovation. Our world-class research seeks to invent and transfer the most compelling technologies enabling the company to differentiate its content, services, and products. Disney Research combines the best of academia and industry, by doing both basic and application-driven research.

Invenia Labs develops machine learning techniques to solve some of the world's most complex forecasting and decision problems. Located in Cambridge (UK), Invenia's current mission is to improve the efficiency of electrical grids, helping to reduce pollution and fight the global climate crisis.

Maluuba's research powers a new era of artificial intelligence. We are driven by the single purpose of building great experiences powered by natural language processing. Our Montreal lab is one of the world's leading research centres, led by a team of scientists focused on natural language and deep learning. Maluuba's technology is used in over 50 million devices and experiences around the world.

BRONZE SPONSORS

Nokia is a global leader in the technologies that connect people and things. Powered by the pioneering work of Nokia Bell Labs, our research and innovation division, and Nokia Technologies, we are at the forefront of creating and licensing the technologies that are increasingly at the heart of our connected lives.

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.

Palantir Technologies builds software platforms that help human experts perform powerful, collaborative analysis of data at scale. Palantir's software is deployed at public institutions, private enterprises, and in the non-profit sector to address the challenges of responsibly making sense of complex, diverse data.

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.

PDT Partners is a top quantitative hedge fund where world class researchers analyze rich data to develop and deploy model-driven algorithmic trading strategies. We offer a strong track record of hiring, challenging and retaining scientists interested in conducting research where the lab is the financial markets.

QuantumBlack is an advanced analytics firm operating at the intersection of strategy, technology & design to improve performance outcomes for organisations. Starting life in the world of F1, looking at problems of race strategy & engineering productivity, QuantumBlack codified their analytical approach into a platform & a process called NERVE which has now been successfully deployed across a variety of industries.

RBC Research is the research arm of the Royal Bank of Canada. The team's mandate is to advance the state of the art in financial technologies by conducting research in machine learning and computer vision. RBC is Canada's largest financial institution with over 80,000 employees across Banking, Insurance, Wealth Management, Investor & Treasury Services, as well as Capital Markets.

Recursion Pharma is 25 people generating biological data as fast as the biggest bio research groups in the world. We've won 2M in NIH grants, and this fall closed a 15M series A led by Lux Capital. We're using cutting edge microscopes to turn human cellular experiments into 100s of TBs of rich biological data, and ML to seek treatments for 100s of diseases as fast as possible. Join us!

Schibsted Media Group is a global media company driven by new technology and employing 6,900 people in 30 countries to support 200 million users. We are driving new technology across online marketplaces and newspapers with data-driven products that leverage machine learning, are pioneering digital media houses, and have a portfolio of flourishing digital companies.

SigOpt is the optimization platform that amplifies your research, providing an ensemble the state of the art in Bayesian optimization research via a simple API and web interface. SigOpt takes any research pipeline and tunes it, right in place. Our cloud-based platform is used by globally recognized leaders in the insurance, credit card, algorithmic trading and consumer packaged goods industries.

Telefónica I+D, the research and development company of the Telefónica Group, was founded in 1988 and its mission is to contribute to the Group's competitiveness through technological innovation. With this aim, the company applies new ideas, concepts and practices in addition to developing products and advanced services.

Yandex is one of the largest internet companies in Europe, operating Russia's most popular search engine. We provide user-centric products and services based on the latest innovations in information retrieval, machine learning and machine intelligence to a worldwide customer audience on all digital platforms and devices. Headquartered in Moscow, we have development and sales offices in 17 locations across nine countries.

EXHIBITORS

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.

From The President

Bienvenido,

The 2016 NIPS Conference and Workshops in Barcelona will host a record number of participants, topping the record number last year in Montréal by over 1800!

	Year	Registrations
Barcelona	2016	5680
Montreal	2015	3800
Montreal	2014	2478
Lake Tahoe	2013	1902
Lake Tahoe	2012	1610
Granada	2011	1394
Vancouver	2010	1301

We had to cap the registration for the first time because of limited space in the Convention Center. Since 2013 the number of registered participants has been growing exponentially and we have outgrown the venues that were planned several years ago. The likelihood of a cap was announced in September when registration opened, but many who in the past registered late were caught by surprise. We apologize to all who we had to turn away. Register early and register often.

For the first time the main NIPS Conference will have two tracks. Researchers from many fields come to NIPS and our goal is to provide a common meeting ground for all.

Unlike most large conferences that are multitrack, NIPS has maintained a single track to keep the meeting from fragmenting. Along with more participants NIPS has had more submissions and with two tracks we were able to expand the number of oral talks and still have time for breaks.

The Posner Lecture this year will be given by Yann LeCun and the Breiman Lecture by Susan Holmes. Ed Posner founded NIPS in 1987 and Leo Breiman bridged the statics and machine learning communities.

The Symposium track between the end of the conference and the beginning of the workshops was popular last year in Montreal and this year three exciting symposia are on our program.

The NIPS Workshops attract as many attendees as the NIPS Conference and this year there will be 50 sessions covering a broad range of topics.

Despite its rapid growth, NIPS has maintained the highest standards and this year our acceptance rate was 24%. The technical program includes 7 invited talks and 569 accepted papers, selected from a total of 2403 submissions considered by the program committee.

Papers presented at the conference will appear in "Advances in Neural Information Processing 29," edited by Ulrike von Luxburg, Isabelle Guyon, Daniel D. Lee, Masashi Sugiyama and Roman Garnett.

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

Long Beach, California 2017, Dec 4 - 9

Montreal Canada 2018, Dec 3 - 8

AREA MAP

GENERAL INFORMATION

REGISTRATION DESK, LEVEL P0

Sunday, December 4: 12:00pm – 8:00pm
Monday – Friday: 7:00 am – 6:00 pm
Saturday, December 10 : 7:00 am - 12:00pm

OPENING RECEPTION AND POSTER SESSION

Monday, December 5 starting at 6:30 pm

Coffee breaks and Food service will be in many locations
P1 with the exhibitors
P2 in the Banquet Hall with more exhibitors.

Please see the maps on the next page

CLOSING RECEPTION

Saturday, December 10 at 7:00 pm

POSTER SESSIONS LEVEL P0 AREA 5+ 6 + 7 + 8

Monday, Dec. 5 - 6:00 pm – 9:30 pm
Tuesday, Dec. 6 - 6:00 pm – 9:30 pm
Wednesday, Dec. 7 6:00 pm – 9:30 pm

- No pins or special tape will be provided
- Take down your poster at 9:30 pm

WORKSHOP LOCATIONS

- CCIB (P1, P2, M1)
- Hilton Diagonal Del Mar (Ballrooms)
- AC Barcelona hotel

WIFI

SSID: NIPS
Password: conference

MOBILE APP

Step 1: Download and install the Whova app from App Store (for iPhones) or Google Play (for Android phones).
Step 2: Sign up in the app using the email address you registered with.
Step 3: You're all set.

Now you will be able to:
- View the event agenda and plan your schedule.
- If you set up your own profile, you can send in-app messages and exchange contact information
- Receive update notifications from organizers.
- Access agenda, maps, and directions.

After downloading, sign up on Whova with the email address that you used to RSVP for our event, or sign up using your social media accounts. If you are asked to enter an invitation code to join the event, please use the following invitation code: **nips**

CHARGING TABLES

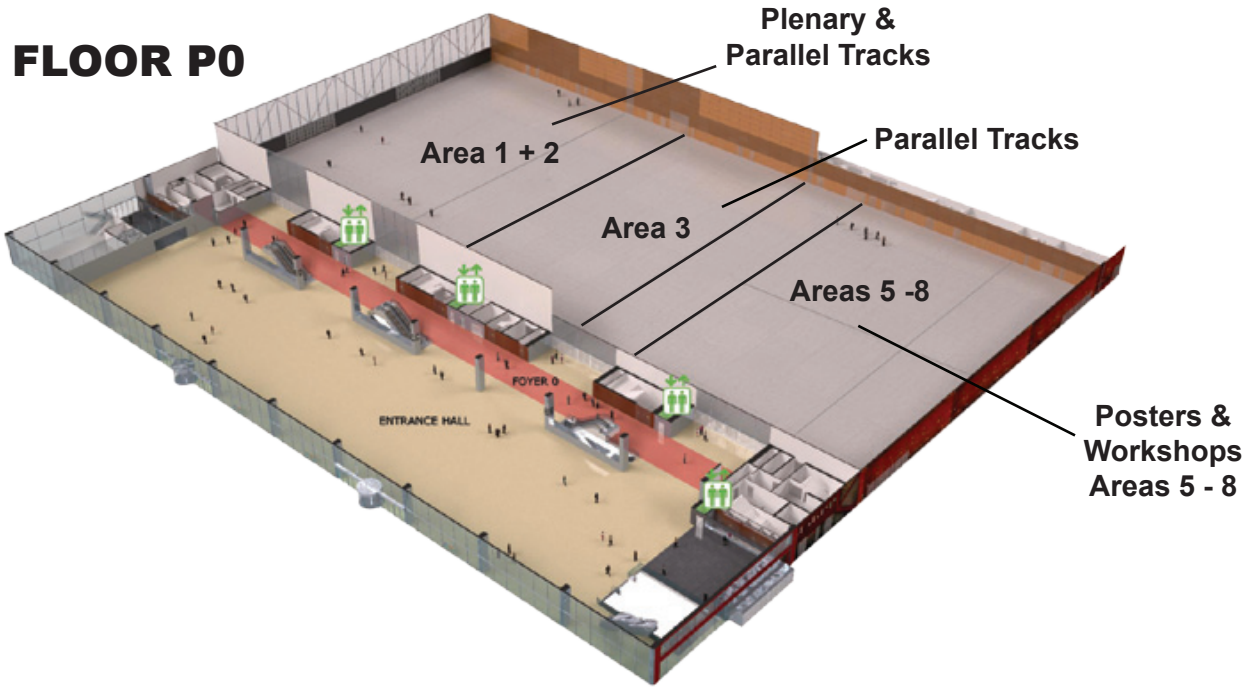
P1 Rooms 113 - 117
P2 Banquet Hall

EXHIBITOR ROOMS

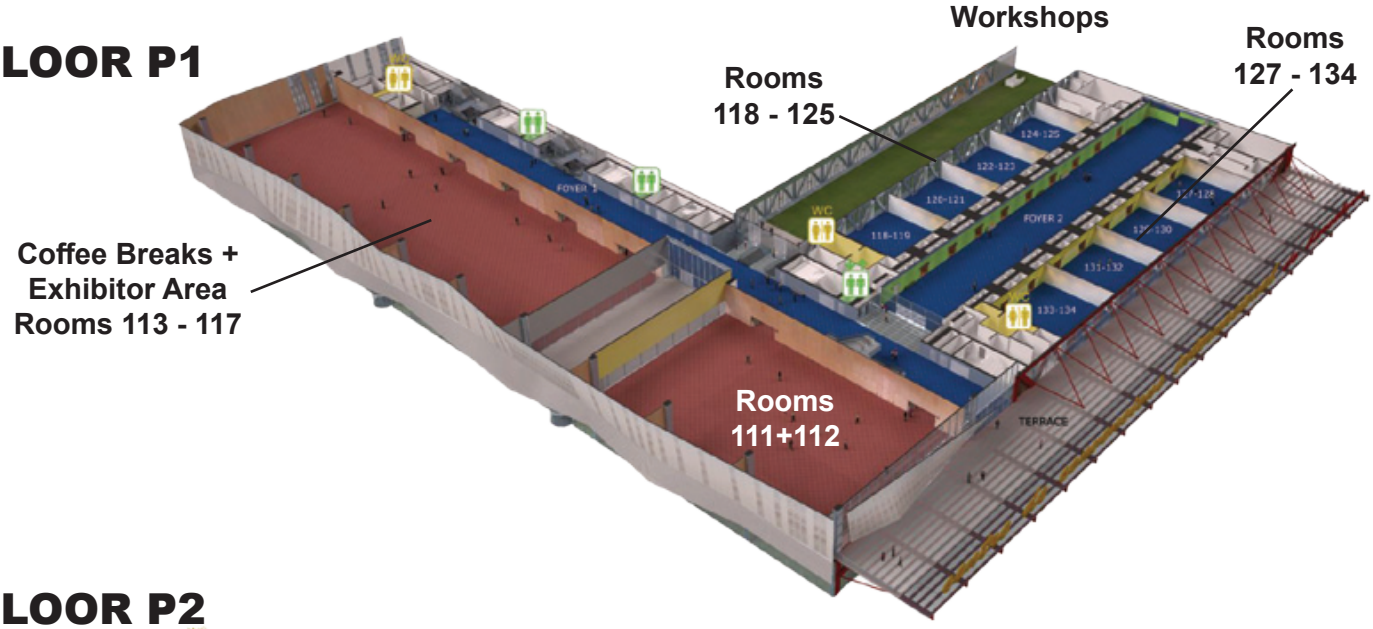
P1 (rooms 113 - 117)
P2 Banquet Hall

CONFERENCE MAP

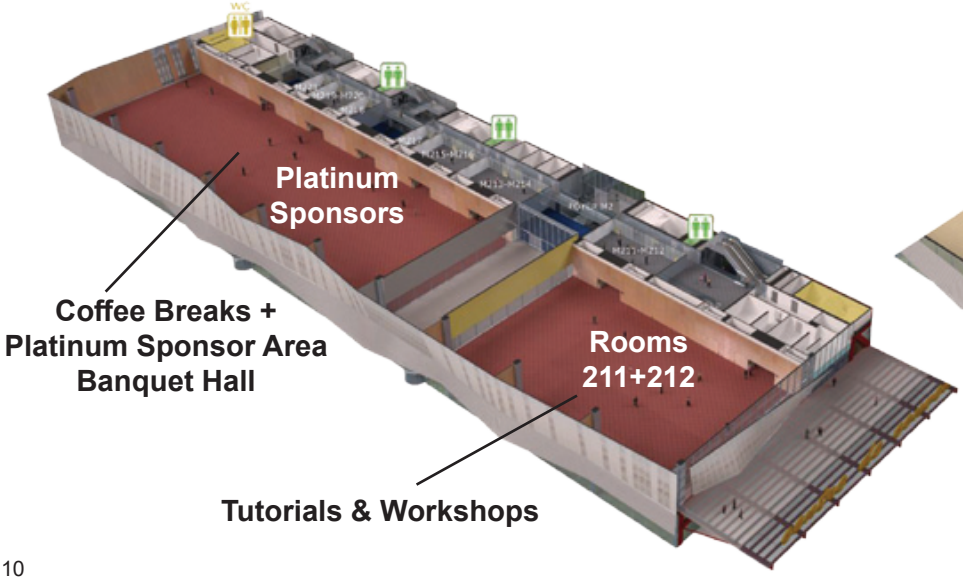
FLOOR P0



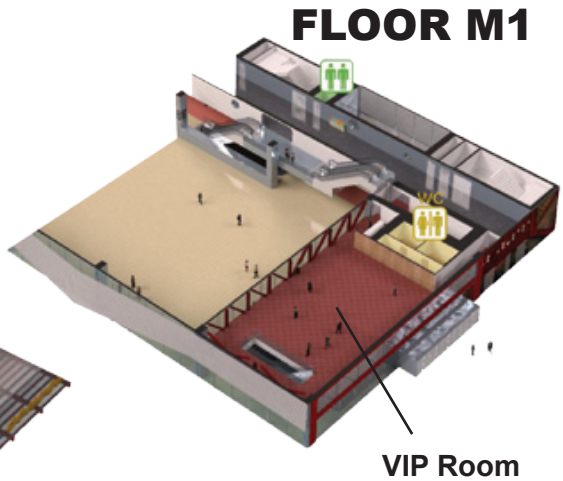
FLOOR P1



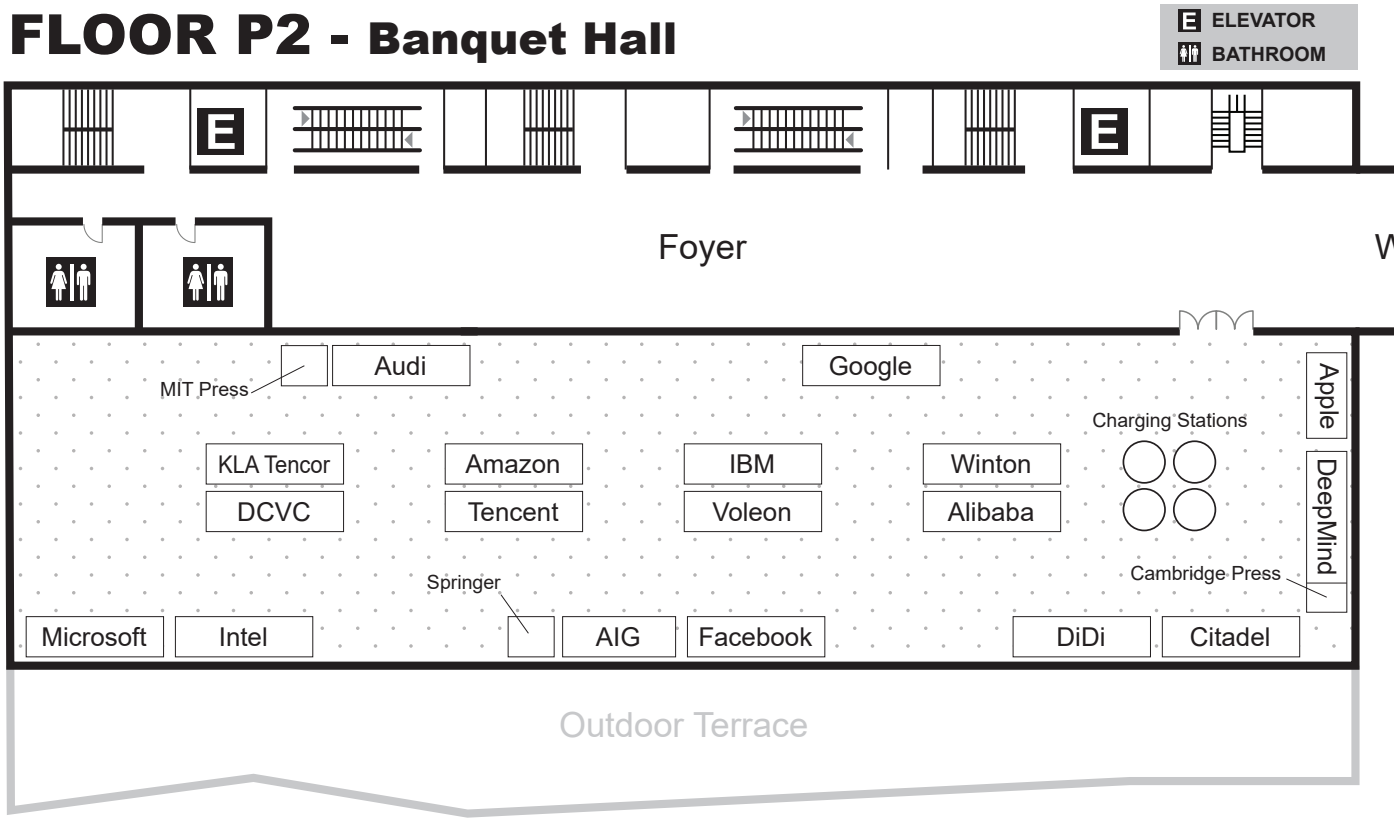
FLOOR P2



FLOOR M1

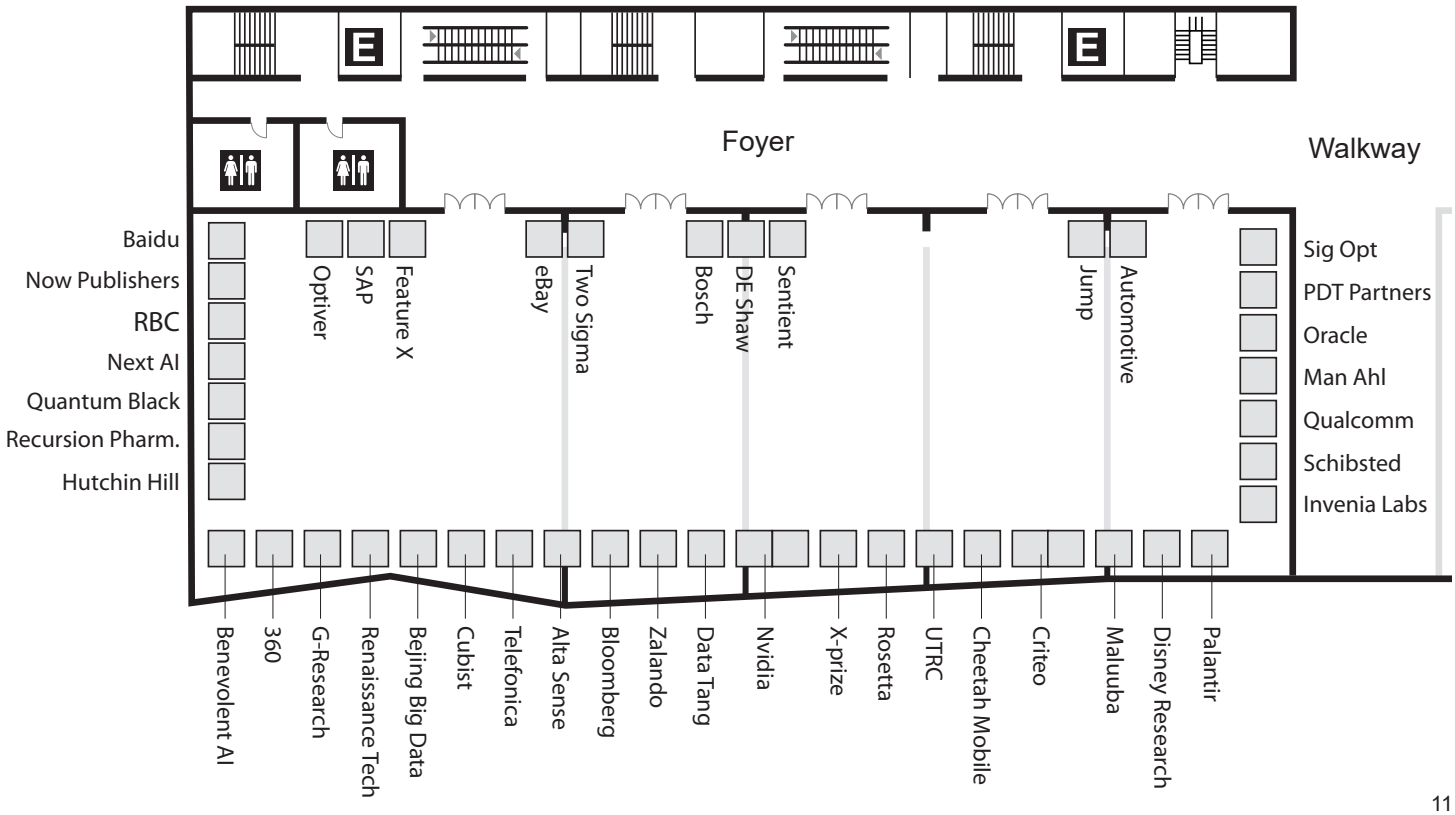


FLOOR P2 - Banquet Hall



COFFEE BREAKS, FOOD AND BEVERAGES WILL BE SERVED ON THESE FLOORS.
CHARGING TABLES ALSO AVAILABLE

FLOOR P1 - Rooms 113 - 117



MONDAY TUTORIALS

TIME & DESCRIPTION	LOCATION
8:30 am - 10:30 am - Tutorial Sessions	
Variational Inference: Foundations and Modern Methods David Blei · Shakir Mohamed · Rajesh Ranganath	Area 1 + 2
Crowdsourcing: Beyond Label Generation Jennifer Wortman Vaughan	Area 3
Deep Reinforcement Learning Through Policy Optimization Pieter Abbeel · John Schulman	Rooms 211 + 212
Coffee break - Level P1, P2 10:30 am - 11 am	
11:00 am -- 1:00 pm - Tutorial Sessions	
Nuts and Bolts of Building AI systems using Deep Learning Andrew Y Ng	Area 1 + 2
Natural Language Processing for Computational Social Science Cristian Danescu-Niculescu-Mizil · Lillian J Lee	Area 3
Theory and Algorithms for Forecasting Non-Stationary Time Series Vitaly Kuznetsov · Mehryar Mohri	Rooms 211 + 212
1 pm - 2:30 pm - Lunch Break (On Your Own)	
2:30 pm - 4:30 pm - Tutorial Sessions	
Generative Adversarial Networks Ian Goodfellow	Area 1 + 2
ML Foundations and Methods for Precision Medicine and Healthcare Suchi Saria · Peter Schulam	Area 3
Large-Scale Optimization: Beyond Stochastic Gradient Descent and Convexity Suvrit Sra · Francis Bach	Rooms 211 + 212
Coffee break - Level P1, P2 4:30 pm - 5 pm	
5:30 pm - 6:20 pm -	
Invited Talk: Posner Lecture - Predictive Learning Yann LeCun	
6:30 pm - 9:30 pm Opening Reception & Posters	P1 & P2

Variational Inference: Foundations and Modern Methods

Area 1 + 2



David Blei (Columbia Univ.)
Shakir Mohamed (DeepMind)
Rajesh Ranganath (Princeton Univ.)

One of the core problems of modern statistics and machine learning is to approximate difficult-to-compute probability distributions. This problem is especially important in probabilistic modeling, which frames all inference about unknown quantities as a calculation about a conditional distribution.

In this tutorial we review and discuss variational inference (VI), a method a that approximates probability distributions through optimization. VI has been used in myriad applications in machine learning and tends to be faster than more traditional methods, such as Markov chain Monte Carlo sampling. Brought into machine learning in the 1990s, recent advances and easier implementation have renewed interest and application of this class of methods. This tutorial aims to provide both an introduction to VI with a modern view of the field, and an overview of the role that probabilistic inference plays in many of the central areas of machine learning.

The tutorial has three parts. First, we provide a broad review of variational inference from several perspectives. This part serves as an introduction (or review) of its central concepts. Second, we develop and connect some of the pivotal tools for VI that have been developed in the last few years, tools like Monte Carlo gradient estimation, black box variational inference, stochastic approximation, and variational auto-encoders. These methods have lead to a resurgence of research and applications of VI. Finally, we discuss some of the unsolved problems in VI and point to promising research directions.

Learning objectives:

- Gain a well-grounded understanding of modern advances in variational inference
- Understand how to implement basic versions for a wide class of models
- Understand connections and different names used in other related research areas
- Understand important problems in variational inference research

Target audience:

- Machine learning researchers across all level of experience from first year grad students to other more experienced researchers
- Targeted at those who want to understand recent advances in variational inference
- Basic understanding of probability is sufficient

Crowdsourcing: Beyond Label Generation

Area 3

Jennifer Wortman Vaughan (Microsoft Research)

This tutorial will showcase some of the most innovative uses of crowdsourcing that have emerged in the past few years. While some have clear and immediate benefits to machine learning, we will also discuss examples in which crowdsourcing has allowed researchers to answer exciting questions in psychology, economics, and other fields.

We will discuss best practices for crowdsourcing (such as how and why to maintain a positive relationship with crowdworkers) and available crowdsourcing tools. We will survey recent research examining the effect of incentives on crowdworker performance. Time permitting, we will also touch on recent ethnographic research studying the community of crowdworkers and/or delve into the ethical implications of crowdsourcing.

Despite the inclusion of best practices and tools, this tutorial should not be viewed as a prescriptive guide for applying existing techniques. The goals of the tutorial are to inspire you to find novel ways of using crowdsourcing in your own research and to provide you with the resources you need to avoid common pitfalls when you do.

Target audience:
This tutorial is open to anyone who wants to learn more about cutting edge research in crowdsourcing. No assumptions will be made about the audience's familiarity with either crowdsourcing or specific machine learning techniques. Anyone who is curious is welcome to attend!

As the tutorial approaches, more information will be available on the tutorial website: http://www.jennwv.com/projects/crowdtutorial.html

Deep Reinforcement Learning Through Policy Optimization

Rooms 211 + 212


Pieter Abbeel (UC Berkley)
John Schulman (Open AI)

Deep Reinforcement Learning (Deep RL) has seen several breakthroughs in recent years. In this tutorial we will focus on recent advances in Deep RL through policy gradient methods and actor critic methods. These methods have shown significant success in a wide range of domains, including continuous-action domains such as manipulation, locomotion, and flight. They have also achieved the state of the art in discrete action domains such as Atari. Fundamentally, there are two types of gradient calculations: likelihood ratio gradients (aka score function gradients) and path derivative gradients (aka perturbation analysis gradients). We will teach policy gradient methods of each type, connect with Actor-Critic methods (which learn both a value function and a policy), and cover a generalized view of the computation of gradients of expectations through Stochastic Computation Graphs.

Learning Objectives:
The objective is to provide attendees with a good understanding of foundations as well as recent advances in policy gradient methods and actor critic methods. Approaches that will be taught: Likelihood Ratio Policy Gradient (REINFORCE), Natural Policy Gradient, Trust Region Policy Optimization, Generalized Advantage Estimation, Asynchronous Advantage Actor Critic, Path Derivative Policy Gradients, Deterministic Policy Gradient, Stochastic Value Gradients, Guided Policy Search. As well as a generalized view of the computation of gradients of expectations through Stochastic Computation Graphs.

Target Audience:
Machine learning researchers. RL background not assumed, but some prior familiarity with the basic concepts could be helpful. Good resource: Sutton and Barto Chapters 3 & 4 (http://webdocs.cs.ualberta.ca/~sutton/book/the-book.html).

Monday Tutorial Session - 11 AM - 1 PM

Nuts and Bolts of Building AI systems using Deep Learning

Area 1 & 2

Andrew Y Ng (Stanford University)

How do you get deep learning to work in your business, product, or scientific study? The rise of highly scalable deep learning techniques is changing how you can best approach AI problems. This includes how you define your train/dev/test split, how you organize your data, how you should think through your search among promising model architectures, and even how you might develop new AI-enabled products. In this tutorial, you'll learn about the emerging best practices in this nascent area. You'll come away able to better organize your and your team's work when developing deep learning applications.

Learning objectives:

- Understand best-practices for applying deep learning in your organization, whether improving existing applications or creating brand new ones.
- Be able to organize and help prioritize your team's work using principles suited to the deep learning era; understand how these practices have changed relative to previous machine learning eras.
- Able to apply error analysis and other debugging techniques suited to deep learning systems.
- Gain a systematic process for selecting among architectures and data for your machine learning tasks.

Target audience:
Attendees should have basic knowledge of machine learning (such as supervised learning). Prior knowledge in deep learning is helpful but not required.

Natural Language Processing for Computational Social Science

Area 3

Cristian Danescu-Niculescu-Mizil (Cornell University)
Lillian J Lee (Cornell University)

More and more of life is now manifested online, and many of the digital traces that are left by human activity are increasingly recorded in natural-language format. This tutorial will examine the opportunities for natural language processing (NLP) to contribute to computational social science, facilitating our understanding of how humans interact with others at both grand and intimate scales.

Learning objectives:

- Influence and persuasion: Can language choices affect whether a political ad is successful, a social-media post gets more re-shares, or a get-out-the-vote campaign will work?
- Language as a reflection of social processes: can we detect status differences, or more broadly, the roles people take in online communities? How does language define collective identity, or signal imminent departure from a community?
- Group success: can language cues help us predict whether a group will cohere or fracture? Or whether a betrayal is forthcoming? Or whether a team will succeed at its task?
- Understand important problems in variational inference research

Target audience:
Unrestricted

Theory and Algorithms for Forecasting Non-Stationary Time Series

Rooms 211 + 212

Vitaly Kuznetsov (Google)
Mehryar Mohri (Courant Institute, Google)

Time series appear in a variety of key real-world applications such as signal processing, including audio and video processing; the analysis of natural phenomena such as local weather, global temperature, and earthquakes; the study of economic variables such as stock values, sales amounts, energy demand; and many other areas. But, while time series forecasting is critical for many applications, it has received little attention in the ML community in recent years, probably due to a lack of familiarity with time series and the fact that standard i.i.d. learning concepts and tools are not readily applicable in that scenario.

This tutorial precisely addresses these and many other related questions. It provides theoretical and algorithmic tools for research related to time series and for designing new solutions. We first present a concise introduction to time series, including basic concepts, common challenges and standard models. Next, we discuss important statistical learning tools and results developed in recent years and show how they are useful for deriving guarantees and designing algorithms both in stationary and non-stationary scenarios. Finally, we show how the online learning framework can be leveraged to derive algorithms that tackle important and notoriously difficult problems including model selection and ensemble methods.

Learning objectives:

- familiarization with basic time series concepts
- introduction to statistical learning theory and algorithms for stationary and non-stationary time series
- introduction to model selection and ensemble methods for time series via online learning

Target audience:
This tutorial is targeted for a very general ML audience and should be accessible to most machine learning researchers and practitioners. We will introduce all the necessary tools from scratch and of course make slides and other detailed tutorial documents available.

Monday Tutorial Session - 2:30 - 4:30 PM

Session Chair: Tamara Broderick

Generative Adversarial Networks

Area 1 & 2

Ian Goodfellow (OpenAI)

Generative adversarial networks (GANs) are a recently introduced class of generative models, designed to produce realistic samples. This tutorial is intended to be accessible to an audience who has no experience with GANs, and should prepare the audience to make original research contributions applying GANs or improving the core GAN algorithms. GANs are universal approximators of probability distributions. Such models generally have an intractable log-likelihood gradient, and require approximations such as Markov chain Monte Carlo or variational lower bounds to make learning feasible. GANs avoid using either of these classes of approximations. The learning process consists of a game between two adversaries: a generator network that attempts to produce realistic samples, and a discriminator network that attempts to identify whether samples originated from the training data or from the generative model. At the Nash equilibrium of this game, the generator network reproduces the data distribution exactly, and the discriminator network cannot distinguish samples from the model from training data. Both networks can be trained using stochastic gradient descent with exact gradients computed by maximum likelihood.

Topics include: - An introduction to the basics of GANs. - A review of work applying GANs to large image generation. - Extending the GAN framework to approximate maximum likelihood, rather than minimizing the Jensen-Shannon divergence. - Improved model architectures that yield better learning in GANs. - Semi-supervised learning with GANs. - Research frontiers, including guaranteeing convergence of the GAN game. - Other applications of adversarial learning, such as domain adaptation and privacy.

Learning objectives:

- To explain the fundamentals of how GANs work to someone who has not heard of them previously
- To bring the audience up to date on image generation applications of GANs
- To prepare the audience to make original contributions to generative modeling research

Target audience:

People who are interested in generative modeling. Both people who do not have prior knowledge of GANs and people who do should find something worthwhile, but the first part of the tutorial will be less interesting to people who have prior knowledge of GANs

---

ML Foundations and Methods for Precision Medicine and Healthcare

Area 3

Suchi Saria (Johns Hopkins University)
Peter Schulam (Johns Hopkins University)

Electronic health records and high throughput measurement technologies are changing the practice of healthcare to become more algorithmic and data-driven. This offers an exciting opportunity for machine learning to impact healthcare.

The aim of this tutorial is to introduce you to the most important challenges and techniques for developing "personalized decision-making" tools in medicine. We will also cover example data sources and describe ongoing national initiatives that provide a way for you to get involved.

Learning objectives:

- Become familiar with important (computational) problems in precision medicine and individualized health care
- Get introduced to state-of-the-art approaches
- Hear about relevant datasets (and potential funding sources).

Target audience:

The majority of this tutorial will be targeted at an audience with basic machine learning knowledge. No background in medicine or health care is needed. We will make our slides and any relevant documents accessible after the talk.

Large-Scale Optimization: Beyond Stochastic Gradient Descent and Convexity

Rooms 211 + 212

Suvrit Sra (MIT)
Francis Bach (INRIA)

Stochastic optimization lies at the heart of machine learning, and its cornerstone is stochastic gradient descent (SGD), a staple introduced over 60 years ago! Recent years have, however, brought an exciting new development: variance reduction (VR) for stochastic methods. These VR methods excel in settings where more than one pass through the training data is allowed, achieving convergence faster than SGD, in theory as well as practice. These speedups underline the huge surge of interest in VR methods; by now a large body of work has emerged, while new results appear regularly! This tutorial brings to the wider machine learning audience the key principles behind VR methods, by positioning them vis-à-vis SGD. Moreover, the tutorial takes a step beyond convexity and covers research-edge results for non-convex problems too, while outlining key points and as yet open challenges.

Learning objectives:

Introduce fast stochastic methods to the wider ML audience to go beyond a 60-year-old algorithm (SGD) – Provide a guiding light through this fast moving area, to unify, and simplify its presentation, outline common pitfalls, and to demystify its capabilities – Raise awareness about open challenges in the area, and thereby spur future research.

Target audience:

- Graduate students (masters as well as PhD stream)
- ML researchers in academia and industry who are not experts in stochastic optimization
- Practitioners who want to widen their repertoire of tools

Posner Lecture - Monday @ 5:30pm

Predictive Learning

Area 1 & 2

Yann LeCun (Facebook, New York University)

Deep learning has been at the root of significant progress in many application areas, such as computer perception and natural language processing. But almost all of these systems currently use supervised learning with human-curated labels. The challenge of the next several years is to let machines learn from raw, unlabeled data, such as images, videos and text. Intelligent systems today do not possess "common sense", which humans and animals acquire by observing the world, acting in it, and understanding the physical constraints of it. I will argue that allowing machine to learn predictive models of the world is key to significant progress in artificial intelligence, and a necessary component of model-based planning and reinforcement learning. The main technical difficulty is that the world is only partially predictable. A general formulation of unsupervised learning that deals with partial predictability will be presented. The formulation connects many well-known approaches to unsupervised learning, as well as new and exciting ones such as adversarial training.

Yann LeCun is Director of AI Research at Facebook, and Silver Professor of Data Science, Computer Science, Neural Science, and Electrical Engineering at New York University. He received the Electrical Engineer Diploma from ESIEE, Paris in 1983, and a PhD in Computer Science from Université Pierre et Marie Curie (Paris) in 1987. After a postdoc at the University of Toronto, he joined AT&T Bell Laboratories in Holmdel, NJ in 1988. He became head of the Image Processing Research Department at AT&T Labs-Research in 1996, and joined NYU as a professor in 2003, after a brief period as a Fellow of the NEC Research Institute in Princeton. From 2012 to 2014 he directed NYU's initiative in data science and became the founding director of the NYU Center for Data Science. He was named Director of AI Research at Facebook in late 2013 and retains a part-time position on the NYU faculty. His current interests include AI, machine learning, computer perception, mobile robotics, and computational neuroscience. He has published over 180 technical papers and book chapters on these topics as well as on neural networks, handwriting recognition, image processing and compression, and on dedicated circuits for computer perception.

Monday Poster Session

#1 **Improved Dropout for Shallow and Deep Learning**
Zhe Li, Boqing Gong, Tianbao Yang

#2 **Communication-Optimal Distributed Clustering**
Jiecao Chen, He Sun, David Woodruff, Qin Zhang

#3 **On Robustness of Kernel Clustering**
Bowei Yan, Purnamrita Sarkar

#4 **Combinatorial semi-bandit with known covariance**
Rémy Degenne, Vianney Perchet

#5 **A posteriori error bounds for joint matrix decomposition problems**
Nicolo Colombo, Nikos Vlassis

#6 **Object based Scene Representations using Fisher Scores of Local Subspace Projections**
Mandar D Dixit, Nuno Vasconcelos

#7 **MoCap-guided Data Augmentation for 3D Pose Estimation in the Wild**
Gregory Rogez, Cordelia Schmid

#8 **Regret of Queueing Bandits**
Subhashini Krishnasamy, Rajat Sen, Ramesh Johari, Sanjay Shakkottai

#9 **Efficient Nonparametric Smoothness Estimation**
Shashank Singh, SimonDu S Du, Barnabas Poczos

#10 **Completely random measures for modelling block-structured sparse networks**
Tue Herlau, Mikkel N. N Schmidt, Morten Mørup

#11 **DISCO Nets : DISsimilarity COefficients Networks**
Diane Bouchacourt, Pawan K Mudigonda, Sebastian Nowozin

#12 **An Architecture for Deep, Hierarchical Generative Models**
Philip Bachman

#13 **A Multi-Batch L-BFGS Method for Machine Learning**
Albert S Berahas, Jorge Nocedal, Martin Takac

#14 **Higher-Order Factorization Machines**
Mathieu Blondel, Akinori Fujino, Naonori Ueda, Masakazu Ishihata

#15 **A Bio-inspired Redundant Sensing Architecture**
Anh Tuan Nguyen, Jian Xu, Zhi Yang

#16 **Learning Supervised PageRank with Gradient-Based and Gradient-Free Optimization Methods**
Lev Bogolubsky, Pavel Dvurechenskii, Alexander Gasnikov, Gleb Gusev, Yurii Nesterov, Andrei M Raigorodskii, altsoph Tikhonov, Maksim Zhukovskii

#17 **Linear Relaxations for Finding Diverse Elements in Metric Spaces**
Aditya Bhaskara, Mehrdad Ghadiri, Vahab Mirrokni, Ola Svensson

#18 **Stochastic Optimization for Large-scale Optimal Transport**
Aude Genevay, Marco Cuturi, Gabriel Peyré, Francis Bach

#19 **Threshold Bandits, With and Without Censored Feedback**
Jacob D Abernethy, Kareem Amin, Ruihao Zhu

#20 **Mistake Bounds for Binary Matrix Completion**
Mark Herbster, Stephen Pasteris, Massimiliano Pontil

#21 **Learning Sound Representations from Unlabeled Video**
Yusuf Aytar, Carl Vondrick, Antonio Torralba

#22 **Doubly Convolutional Neural Networks**
Shuangfei Zhai, Yu Cheng, Zhongfei (Mark) Zhang

#23 **Maximizing Influence in an Ising Network: A Mean-Field Optimal Solution**
Christopher Lynn, Daniel D Lee

#24 **Learning from Rational Behavior: Predicting Solutions to Unknown Linear Programs**
Shahin Jabbari, Ryan M Rogers, Aaron Roth, Steen Wu

#25 **Fairness in Learning: Classic and Contextual Bandits**
Matthew Joseph, Michael Kearns, Jamie H Morgenstern, Aaron Roth

Monday Poster Session

#26 **A Powerful Generative Model Using Random Weights for the Deep Image Representation**
Kun He, Yan Wang, John Hopcroft

#27 **Improved Error Bounds for Tree Representations of Metric Spaces**
Samir Chowdhury, Facundo Mémoli, Zane T Smith

#28 **Adaptive optimal training of animal behavior**
Ji Hyun Bak, Jung Choi, Ilana Witten, Jonathan W Pillow

#29 **PAC-Bayesian Theory Meets Bayesian Inference**
Pascal Germain, Francis Bach, Alexandre Lacoste, Simon Lacoste-Julien

#30 **Nearly Isometric Embedding by Relaxation**
James McQueen, Marina Meila, Dominique Joncas

#31 **Graph Clustering: Block-models and model free results**
Yali Wan, Marina Meila

#32 **Learning Transferrable Representations for Unsupervised Domain Adaptation**
Ozan Sener, Hyun Oh Song, Ashutosh Saxena, Silvio Savarese

#33 **Measuring Neural Net Robustness with Constraints**
Osbert Bastani, Yani Ioannou, Leonidas Lampropoulos, Dimitrios Vytiniotis, Aditya Nori, Antonio Criminisi

#34 **Forward models at Purkinje synapses facilitate cerebellar anticipatory control**
Ivan Herreros, Xerxes Arsiwalla, Paul Verschure

#35 **Estimating Nonlinear Neural Response Functions using GP Priors and Kronecker Methods**
Cristina Savin, Gasper Tkacik

#36 **A Bayesian method for reducing bias in neural representational similarity analysis**
Mingbo Cai, Nicolas W Schuck, Jonathan W Pillow, Yael Niv

#37 **Learning to Communicate with Deep Multi-Agent Reinforcement Learning**
Jakob Foerster, Yannis M. Assael, Nando de Freitas, Shimon Whiteson

#38 **Total Variation Classes Beyond 1d: Minimax Rates, and the Limitations of Linear Smoothers**
Veeru Sadhanala, Yu-Xiang Wang, Ryan J Tibshirani

#39 **Exponential Family Embeddings**
Maja Rudolph, Francisco J. R. Ruiz, Stephan Mandt, David Blei

#40 **k^*-Nearest Neighbors: From Global to Local**
Oren Anava, Kfir Levy

#41 **Reward Augmented Maximum Likelihood for Neural Structured Prediction**
Mohammad Norouzi, Samy Bengio, ZF Chen, Navdeep Jaitly, Mike Schuster, Yonghui Wu, Dale Schuurmans

#42 **A Probabilistic Model of Social Decision Making based on Reward Maximization**
Koosha Khalvati, Seongmin A. Park, Jean-Claude Dreher, Rajesh P Rao

#43 **Active Learning with Oracle Epiphany**
T.K. Huang, Lihong Li, Ara Vartanian, Saleema Amershi, Jerry Zhu

#44 **On Regularizing Rademacher Observation Losses**
Richard Nock

#45 **A Non-generative Framework and Convex Relaxations for Unsupervised Learning**
Elad Hazan, Tengyu Ma

#46 **Learning Tree Structured Potential Games**
Vikas Garg, Tommi Jaakkola

#47 **Equality of Opportunity in Supervised Learning**
Moritz Hardt, Eric Price, Nati Srebro

#48 **Interaction Networks for Learning about Objects, Relations and Physics**
Peter Battaglia, Razvan Pascanu, Matthew Lai, Danilo Jimenez Rezende, koray kavukcuoglu

#49 **beta-risk: a New Surrogate Risk for Learning from Weakly Labeled Data**
Valentina Zantedeschi, Rémi Emonet, Marc Sebban

#50 **Binarized Neural Networks**
Itay Hubara, Matthieu Courbariaux, Daniel Soudry, Ran El-Yaniv, Yoshua Bengio

#51 **Regularization With Stochastic Transformations and Perturbations for Deep Semi-Supervised Learning**
Mehdi Sajjadi, Mehran Javanmardi, Tolga Tasdizen

#52 **Generating Images with Perceptual Similarity Metrics based on Deep Networks**
Alexey Dosovitskiy, Thomas Brox

#53 **Exploiting Tradeoffs for Exact Recovery in Heterogeneous Stochastic Block Models**
Amin Jalali, Qiyang Han, Ioana Dumitriu, Maryam Fazel

#54 **Tensor Switching Networks**
Kenyon Tsai, Andrew M Saxe, David Cox

#55 **Finite-Dimensional BFRY Priors and Variational Bayesian Inference for Power Law Models**
Juho Lee, Lancelot F James, Seungjin Choi

#56 **Temporal Regularized Matrix Factorization for High-dimensional Time Series Prediction**
Hsiang-Fu (Rofu) Yu, Nikhil Rao, Inderjit S Dhillon

#57 **Composing graphical models with neural networks for structured representations and fast inference**
Matthew Johnson, David Duvenaud, Alex Wiltschko, Ryan P Adams, Sandeep R Datta

#58 **Contextual-MDPs for PAC Reinforcement Learning with Rich Observations**
Akshay Krishnamurthy, Alekh Agarwal, John Langford

#59 **Algorithms and matching lower bounds for approximately-convex optimization**
Andrej Risteski, Yuanzhi Li

#60 **Fast Stochastic Methods for Nonsmooth Nonconvex Optimization**
Sashank J. Reddi, Suvrit Sra, Barnabas Poczos, Alex J Smola

#61 **A Simple Practical Accelerated Method for Finite Sums**
Aaron Defazio

#62 **Unsupervised Learning for Physical Interaction through Video Prediction**
Chelsea Finn, Ian Goodfellow, Sergey Levine

#63 **Threshold Learning for Optimal Decision Making**
Nathan F Lepora

Monday Poster Session

#64 **Collaborative Recurrent Autoencoder: Recommend while Learning to Fill in the Blanks**
Hao Wang, Xingjian SHI, Dit-Yan Yeung

#65 **Finding significant combinations of features in the presence of categorical covariates**
Laetitia Papaxanthos, Felipe Llinares-Lopez, Dean Bodenham, Karsten Borgwardt

#66 **Synthesizing the preferred inputs for neurons in neural networks via deep generator networks**
Anh Nguyen, Alexey Dosovitskiy, Jason Yosinski, Thomas Brox, Jeff Clune

#67 **Learning Infinite RBMs with Frank-Wolfe**
Wei Ping, Qiang Liu, Alexander Ihler

#68 **Sorting out typicality with the inverse moment matrix SOS polynomial**
Edouard Pauwels, Jean B Lasserre

#69 **Improving PAC Exploration Using the Median of Means**
Jason Pazis, Ron E Parr, Jonathan P How

#70 **Reconstructing Parameters of Spreading Models from Partial Observations**
Andrey Lokhov

#71 **Dynamic Filter Networks**
Xu Jia, Bert De Brabandere, Tinne Tuytelaars, Luc V Gool

#72 **Long-Term Trajectory Planning Using Hierarchical Memory Networks**
Stephan Zheng, Yisong Yue, Patrick Lucey

#73 **Cooperative Inverse Reinforcement Learning**
Dylan Hadfield-Menell, Stuart J Russell, Pieter Abbeel, Anca Dragan

#74 **Encode, Review, and Decode: Reviewer Module for Caption Generation**
Zhilin Yang, Ye Yuan, Yuexin Wu, William W Cohen, Russ Salakhutdinov

#75 **Gradient-based Sampling: An Adaptive Importance Sampling for Least-squares**
Rong Zhu

#76 **Robust k-means: a Theoretical Revisit**
ALEX GEORGOGIANNIS

#77 **Boosting with Abstention**
Corinna Cortes, Giulia DeSalvo, Mehryar Mohri

#78 **Estimating the class prior and posterior from noisy positives and unlabeled data**
Shantanu J Jain, Martha White, Pedja Radivojac

#79 **Bootstrap Model Aggregation for Distributed Statistical Learning**
JUN HAN, Qiang Liu

#80 **Noise-Tolerant Life-Long Matrix Completion via Adaptive Sampling**
Maria-Florina Balcan, Hongyang Zhang

#81 **FPNN: Field Probing Neural Networks for 3D Data**
Yangyan Li, pirk Pirk, Hao Su, Charles R Qi, Leonidas J Guibas

#82 **Causal meets Submodular: Subset Selection with Directed Information**
Yuxun Zhou, Costas Spanos

#83 **Improving Variational Autoencoders with Inverse Autoregressive Flow**
Diederik P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, Max Welling

#84 **Adaptive Smoothed Online Multi-Task Learning**
Keerthiram Murugesan, Hanxiao Liu, Jaime Carbonell, Yiming Yang

#85 **The Limits of Learning with Missing Data**
Brian Bullins, Elad Hazan, Tomer Koren

#86 **Safe Exploration in Finite Markov Decision Processes with Gaussian Processes**
Matteo Turchetta, Felix Berkenkamp, Andreas Krause

#87 **Sparse Support Recovery with Non-smooth Loss Functions**
Kévin Degraux, Gabriel Peyré, Jalal Fadili, Laurent Jacques

#88 **Crowdsourced Clustering: Querying Edges vs Triangles**
Ramya Korlakai Vinayak, Babak Hassibi

#89 **Dual Decomposed Learning with Factorwise Oracle for Structural SVM of Large Output Domain**
Ian Yen, Xiangru Huang, Kai Zhong, Ruohan Zhang, Pradeep K Ravikumar, Inderjit S Dhillon

#90 **Sampling for Bayesian Program Learning**
Kevin Ellis, Armando Solar-Lezama, Josh Tenenbaum

#91 **Multiple-Play Bandits in the Position-Based Model**
Paul Lagrée, Claire Vernade, Olivier Cappe

#92 **Image Restoration Using Very Deep Convolutional Encoder-Decoder Networks with Symmetric Skip Connections**
Xiaojiao Mao, Chunhua Shen, Yu-Bin Yang

#93 **Optimistic Bandit Convex Optimization**
Scott Yang, Mehryar Mohri

#94 **Computing and maximizing influence in linear threshold and triggering models**
Justin T Khim, Varun Jog, Po-Ling Loh

#95 **Clustering with Bregman Divergences: an Asymptotic Analysis**
Chaoyue Liu, Mikhail Belkin

#96 **Community Detection on Evolving Graphs**
LEONARDI Leonardi, Aris Anagnostopoulos, Jakub Łącki, Silvio Lattanzi, Mohammad Mahdian

#97 **Dueling Bandits: Beyond Condorcet Winners to General Tournament Solutions**
Siddartha Y. Ramamohan, Arun Rajkumar, Shivani Agarwal

#98 **Learning a Metric Embedding for Face Recognition using the Multibatch Method**
Oren Tadmor, Tal Rosenwein, Shai Shalev-Shwartz, Yonatan Wexler, Amnon Shashua

#99 **Convergence guarantees for kernel-based quadrature rules in misspecified settings**
Motonobu Kanagawa, Bharath K. Sriperumbudur, Kenji Fukumizu

#100 **Stochastic Variational Deep Kernel Learning**
Andrew G Wilson, Zhiting Hu, Russ Salakhutdinov, Eric P Xing

#101 **Deep Submodular Functions**
Brian W Dolhansky, Jeff A Bilmes

#102 **Scaled Least Squares Estimator for GLMs in Large-Scale Problems**
Murat A Erdogdu, Lee H Dicker, Mohsen Bayati

#103 **Matrix Completion and Clustering in Self-Expressive Models**
Ehsan Elhamifar

#104 **Stochastic Three-Composite Convex Minimization**
Alp Yurtsever, Bang Cong Vu, Volkan Cevher

#105 **Tree-Structured Reinforcement Learning for Sequential Object Localization**
Zequn Jie, Xiaodan Liang, Jiashi Feng, Xiaojie Jin, Wen Lu, Shuicheng Yan

#106 **The non-convex Burer-Monteiro approach works on smooth semidefinite programs**
Nicolas Boumal, Vlad Voroninski, Afonso Bandeira

#107 **Neurons Equipped with Intrinsic Plasticity Learn Stimulus Intensity Statistics**
Travis Monk, Cristina Savin, Jörg Lücke

#108 **Greedy Feature Construction**
Dino Oglic, Thomas Gärtner

#109 **Dynamic Mode Decomposition with Reproducing Kernels for Koopman Spectral Analysis**
Yoshinobu Kawahara

#110 **Learning the Number of Neurons in Deep Networks**
Jose M Alvarez, Mathieu Salzmann

#111 **Strategic Attentive Writer for Learning Macro-Actions**
Alexander Vezhnevets, Volodymyr Mnih, Simon Osindero, Alex Graves, Oriol Vinyals, John Agapiou, koray kavukcuoglu

#112 **Active Learning from Imperfect Labelers**
Songbai Yan, Kamalika Chaudhuri, Tara Javidi

#113 **Probabilistic Linear Multistep Methods**
Onur Teymur, Kostas Zygalakis, Ben Calderhead

#114 **More Supervision, Less Computation: Statistical-Computational Tradeoffs in Weakly Supervised Learning**
Xinyang Yi, Zhaoran Wang, Zhuoran Yang, Constantine Caramanis, Han Liu

#115 **Mutual information for symmetric rank-one matrix estimation: A proof of the replica formula**
jean barbier, Mohamad Dia, Nicolas Macris, Florent Krzakala, Thibault Lesieur, Lenka Zdeborová

#116 **Coin Betting and Parameter-Free Online Learning**
Francesco Orabona, David Pal

#117 **Normalized Spectral Map Synchronization**
Yanyao Shen, Qixing Huang, Nati Srebro, Sujay Sanghavi

#118 **On Explore-Then-Commit strategies**
Aurelien Garivier, Tor Lattimore, Emilie Kaufmann

#119 **Learning Kernels with Random Features**
Aman Sinha, John C Duchi

#120 **Robustness of classifiers: from adversarial to random noise**
Alhussein Fawzi, Seyed-Mohsen Moosavi-Dezfooli, Pascal Frossard

#121 **Adaptive Skills Adaptive Partitions (ASAP)**
Daniel J Mankowitz, Timothy A Mann, Shie Mannor

#122 **Gaussian Process Bandit Optimisation with Multi-fidelity Evaluations**
Kirthevasan Kandasamy, Gautam Dasarathy, Junier B Oliva, Jeff Schneider, Barnabas Poczos

#123 **Flexible Models for Microclustering with Applications to Entity Resolution**
Brenda Betancourt, Giacomo Zanella, Jeff Miller, Hanna Wallach, Abbas Zaidi, Rebecca Steorts

#124 **Stochastic Gradient Richardson-Romberg Markov Chain Monte Carlo**
Alain Durmus, Umut Simsekli, Eric Moulines, Roland Badeau, Gaël RICHARD

#125 **Online and Differentially-Private Tensor Decomposition**
Yining Wang, Anima Anandkumar

#126 **Maximal Sparsity with Deep Networks?**
Bo Xin, Yizhou Wang, Wen Gao, David Wipf

#127 **Efficient High-Order Interaction-Aware Feature Selection Based on Conditional Mutual Information**
Alexander Shishkin, Anastasia Bezzubtseva, Alexey Drutsa, Ilia Shishkov, kglad Gladkikh, Gleb Gusev, Pavel Serdyukov

#128 **Geometric Dirichlet Means Algorithm for Topic Inference**
Mikhail Yurochkin, Long Nguyen

#129 **Interaction Screening: Efficient and Sample-Optimal Learning of Ising Models**
Marc Vuffray, Sidhant Misra, Andrey Lokhov, Michael Chertkov

#130 **Multi-armed Bandits: Competing with Optimal Sequences**
Zohar Karnin, Oren Anava

#131 **Catching heuristics are optimal control policies**
Boris Belousov, Gerhard Neumann, Constantin A Rothkopf, Jan R Peters

#132 **Fast stochastic optimization on Riemannian manifolds**
Hongyi Zhang, Sashank J. Reddi, Suvrit Sra

#133 **A Comprehensive Linear Speedup Analysis for Asynchronous Stochastic Parallel Optimization from Zeroth-Order to First-Order**
Xiangru Lian, Huan Zhang, Cho-Jui Hsieh, Yijun Huang, Ji Liu

#134 **Stochastic Gradient MCMC with Stale Gradients**
Changyou Chen, Nan Ding, Chunyuan Li, Yizhe Zhang, Lawrence Carin

#135 **Disentangling factors of variation in deep representation using adversarial training**
Michael F Mathieu, Zhizhen Zhao, Aditya Ramesh, Pablo Sprechmann, Yann LeCun

#136 **Consistent Kernel Mean Estimation for Functions of Random Variables**
Adam Scibior, Carl-Johann Simon-Gabriel, Ilya Tolstikhin, Prof. Bernhard Schölkopf

#137 **DECOrrelated feature space partitioning for distributed sparse regression**
Xiangyu Wang, David B Dunson, Chenlei Leng

#138 **Coupled Generative Adversarial Networks**
Ming-Yu Liu, Oncel Tuzel

Monday Poster Session

#139 **Matching Networks for One Shot Learning**
Oriol Vinyals, Charles Blundell, Timothy Lillicrap, koray kavukcuoglu, Daan Wierstra

#140 **Distributed Flexible Nonlinear Tensor Factorization**
Shandian Zhe, Kai Zhang, Pengyuan Wang, Kuang-chih Lee, Zenglin Xu, Alan Qi, Zoubin Ghahramani

#141 **Tracking the Best Expert in Non-stationary Stochastic Environments**
Chen-Yu Wei, Yi-Te Hong, Chi-Jen Lu

#142 **Deep Alternative Neural Networks: Exploring Contexts as Early as Possible for Action Recognition**
Jinzhuo Wang, Wenmin Wang, xiongtao Chen, Ronggang Wang, Wen Gao

#143 **Learning Parametric Sparse Models for Image Super-Resolution**
Yongbo Li, Weisheng Dong, Xuemei Xie, GUANGMING Shi, Xin Li, Donglai Xu

#144 **Kernel Observers: Systems-Theoretic Modeling and Inference of Spatiotemporally Evolving Processes**
Hassan A Kingravi, Harshal R Maske, Girish Chowdhary

#145 **Learning brain regions via large-scale online structured sparse dictionary learning**
Elvis DOHMATOB, Arthur Mensch, Gael Varoquaux, Bertrand Thirion

#146 **Scaling Factorial Hidden Markov Models: Stochastic Variational Inference without Messages**
Yin Cheng Ng, Pawel M Chilinski, Ricardo Silva

#147 **A Bandit Framework for Strategic Regression**
Yang Liu, Yiling Chen

#148 **Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering**
Michaël Defferrard, Xavier Bresson, Pierre Vandergheynst

#149 **Variational Gradient Descent: A General Purpose Bayesian Inference Algorithm**
Qiang Liu, Dilin Wang

#150 **Deep Learning Models of the Retinal Response to Natural Scenes**
Lane McIntosh, Niru Maheswaranathan, Aran Nayebi, Surya Ganguli, Stephen Baccus

#151 **Safe and Efficient Off-Policy Reinforcement Learning**
Remi Munos, Tom Stepleton, Anna Harutyunyan, Marc Bellemare

#152 **Yggdrasil: An Optimized System for Training Deep Decision Trees at Scale**
Firas Abuzaid, Joseph K Bradley, Feynman T Liang, Andrew Feng, Lee Yang, Matei Zaharia, Ameet S Talwalkar

#153 **Sample Complexity of Automated Mechanism Design**
Maria-Florina Balcan, Tuomas Sandholm, Ellen Vitercik

#154 **Deep Exploration via Bootstrapped DQN**
Ian Osband, Charles Blundell, Alexander Pritzel, Benjamin Van Roy

#155 **Search Improves Label for Active Learning**
Alina Beygelzimer, Daniel Hsu, John Langford, Chicheng Zhang

#156 **Efficient and Robust Spiking Neural Circuit for Navigation Inspired by Echolocating Bats**
Bipin Rajendran, Pulkit Tandon, Yash H Malviya

#157 **Theoretical Comparisons of Positive-Unlabeled Learning against Positive-Negative Learning**
Gang Niu, Marthinus Christoffel du Plessis, Tomoya Sakai, Yao Ma, Masashi Sugiyama

#158 **Quantized Random Projections and Non-Linear Estimation of Cosine Similarity**
Ping Li, Michael Mitzenmacher, Martin Slawski

#159 **CNNpack: Packing Convolutional Neural Networks in the Frequency Domain**
Yunhe Wang, Chang Xu, Shan You, Dacheng Tao, Chao Xu

#160 **Verification Based Solution for Structured MAB Problems**
Zohar Karnin

#161 **Neurally-Guided Procedural Models: Amortized Inference for Procedural Graphics Programs using Neural Networks**
Daniel Ritchie, Anna Thomas, Pat Hanrahan, Noah Goodman

#162 **Edge-Exchangeable Graphs and Sparsity**
Diana Cai, Trevor Campbell, Tamara Broderick

#163 **Learning and Forecasting Opinion Dynamics in Social Networks**
Abir De, Isabel Valera, Niloy Ganguly, Sourangshu Bhattacharya, Manuel Gomez Rodriguez

#164 **Probing the Compositionality of Intuitive Functions**
Eric Schulz, Josh Tenenbaum, David Duvenaud, Maarten Speekenbrink, Samuel J Gershman

#165 **Learning shape correspondence with anisotropic convolutional neural networks**
Davide Boscaini, Jonathan Masci, Emanuele Rodolà, Michael Bronstein

#166 **Improved Techniques for Training GANs**
Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, Xi Chen

#167 **Automated scalable segmentation of neurons from multispectral images**
Uygar Sümbül, Douglas Roossien, Dawen Cai, John Cunningham, Liam Paninski

#168 **Optimal Cluster Recovery in the Labeled Stochastic Block Model**
Se-Young Yun, Alexandre Proutiere

#169 **Phased Exploration with Greedy Exploitation in Stochastic Combinatorial Partial Monitoring Games**
Sougata Chaudhuri, Ambuj Tewari

#170 **Dual Space Gradient Descent for Online Learning**
Trung Le, Tu Nguyen, Vu Nguyen, Dinh Phung

#171 **Data Programming: Creating Large Training Sets, Quickly**
Alexander J Ratner, Christopher M De Sa, Sen Wu, Daniel Selsam, Christopher Ré

#172 **Near-Optimal Smoothing of Structured Conditional Probability Matrices**
Moein Falahatgar, Mesrob I Ohannessian, Alon Orlitsky

#173 **An urn model for majority voting in classification ensembles**
Victor Soto, Alberto Suárez, Gonzalo Martinez-Muñoz

#1 Improved Dropout for Shallow and Deep Learning

Zhe Li (The Univ. of Iowa)
Boqing Gong (Univ. of Central Florida)
Tianbao Yang (Univ. of Iowa)

Dropout has been witnessed with great success in training deep neural networks by independently zeroing out the outputs of neurons at random. It has also received a surge of interest for shallow learning, e.g., logistic regression. However, the independent sampling for dropout could be suboptimal for the sake of convergence. In this paper, we propose to use multinomial sampling for dropout, i.e., sampling features or neurons according to a multinomial distribution with different probabilities for different features/neurons. To exhibit the optimal dropout probabilities, we analyze the shallow learning with multinomial dropout and establish the risk bound for stochastic optimization. By minimizing a sampling dependent factor in the risk bound, we obtain a distribution-dependent dropout with sampling probabilities dependent on the second order statistics of the data distribution. To tackle the issue of evolving distribution of neurons in deep learning, we propose an efficient adaptive dropout (named \textbf{evolutional dropout}) that computes the sampling probabilities on-the-fly from a mini-batch of examples. Empirical studies on several benchmark datasets demonstrate that the proposed dropouts achieve not only much faster convergence and but also a smaller testing error than the standard dropout. For example, on the CIFAR-100 data, the evolutional dropout achieves relative improvements over 10\% on the prediction performance and over 50\% on the convergence speed compared to the standard dropout.

#2 Communication-Optimal Distributed Clustering

Jiecao Chen (Indiana Univ. Bloomington)
He Sun (The Univ. of Bristol)
David Woodruff
Qin Zhang

Clustering large datasets is a fundamental problem with a number of applications in machine learning. Data is often collected on different sites and clustering needs to be performed in a distributed manner with low communication. We would like the quality of the clustering in the distributed setting to match that in the centralized setting for which all the data resides on a single server. In this work, we study both graph and geometric clustering problems in two distributed models: (1) a point-to-point model, and (2) a model with a broadcast channel. We give protocols in both models which we show are nearly optimal by proving almost matching communication lower bounds. Our work highlights the surprising power of a broadcast channel for clustering problems; roughly speaking, to cluster n points or n vertices in a graph distributed across s servers, for a worst-case partitioning the communication complexity in a point-to-point model is ns, while in the broadcast model it is n + s. We implement our algorithms and demonstrate this phenomenon on real life datasets, showing that our algorithms are also very efficient in practice.

#3 On Robustness of Kernel Clustering

Bowei Yan (Univ. of Texas at Austin)
Purnamrita Sarkar (U.C. Berkeley)

Clustering is one of the most important unsupervised problems in machine learning and statistics. Among many existing algorithms, kernel \km has drawn much research attention due to its ability to find non-linear cluster boundaries and inherent simplicity. There are two main approaches for kernel k-means: SVD of the kernel matrix and convex relaxations. Despite the attention kernel clustering has received both from theoretical and applied quarters, not much is known about robustness of the methods. In this paper we first introduce a semidefinite programming relaxation for the kernel clustering problem, then prove that under a suitable model specification, both the K-SVD and SDP approaches are consistent in the limit, albeit SDP is strongly consistent, i.e. achieves exact recovery, whereas K-SVD is weakly consistent, i.e. the fraction of misclassified nodes vanish.

#4 Combinatorial semi-bandit with known covariance

Rémy Degenne (Université Paris Diderot)
Vianney Perchet (Ensae & Criteo Labs)

The combinatorial stochastic semi-bandit problem is an extension of the classical multi-armed bandit problem in which an algorithm pulls more than one arm at each stage and the rewards of all pulled arms are revealed. One difference with the single arm variant is that the dependency structure of the arms is crucial. Previous works on this setting either used a worst-case approach or imposed independence of the arms. We introduce a way to quantify the dependency structure of the problem and design an algorithm that adapts to it. The algorithm is based on linear regression and the analysis uses techniques from the linear bandit literature. By comparing its performance to a new lower bound, we prove that it is optimal, up to a poly-logarithmic factor in the number of arms pulled.

#5 A posteriori error bounds for joint matrix decomposition problems

Nicolo Colombo (Univ of Luxembourg)
Nikos Vlassis (Adobe Research)

Joint matrix decomposition problems appear frequently in statistics and engineering, with notable applications in learning latent variable models and tensor factorization. Joint triangularization of nearly diagonalizable matrices can be performed via orthogonal matrices, and the related nonconvex optimization problem is more tractable than alternative approaches. We carry out a perturbation analysis of the noisy joint matrix triangularization problem, and we derive a bound on the distance between any feasible approximate triangularizer and its noise-free counterpart. The bound is an a posteriori one, in the sense that it is based on quantities that are observed (input matrices and functions thereof), and moreover it is oblivious to the algorithm used and/or the properties of the feasible solution (e.g., proximity to critical points) that are typical of existing bounds in the literature. To our knowledge, this is the first a posteriori bound for joint matrix decomposition. We discuss possible applications and we compare with analogous a priori bounds for the same problem.

#6 Object based Scene Representations using Fisher Scores of Local Subspace Projections

Mandar D Dixit (UC San Diego)
Nuno Vasconcelos

Several works have shown that deep CNN classifiers can be easily transferred across datasets, e.g. the transfer of a CNN trained to recognize objects on ImageNET to an object detector on Pascal VOC. Less clear, however, is the ability of CNNs to transfer knowledge {\it across\/} tasks. A common example of such transfer is the problem of scene classification that should leverage localized object detections to recognize holistic visual concepts. While this problem is currently addressed with Fisher vector representations, these are now shown ineffective for the high-dimensional and highly non-linear features extracted by modern CNNs. It is argued that this is mostly due to the reliance on a model, the Gaussian mixture of diagonal covariances, which has a very limited ability to capture the second order statistics of CNN features. This problem is addressed by the adoption of a better model, the mixture of factor analyzers (MFA), which approximates the non-linear data manifold by a collection of local subspaces. The Fisher score with respect to the MFA (MFA-FS) is derived and proposed as an image representation for holistic image classifiers. Extensive experiments show that the MFA-FS has state of the art performance for object-to-scene transfer and this transfer actually **outperforms** the training of a scene CNN from a large scene dataset. The two representations are also shown to be **complementary** in the sense that their combination outperforms each of the representations by itself. When combined, they produce a state of the art scene classifier.

#7 MoCap-guided Data Augmentation for 3D Pose Estimation in the Wild

Gregory Rogez (Inria)
Cordelia Schmid

In this paper, we address the problem of 3D human pose understanding in the wild. The lack of large enough training datasets has prevented the development of end-to-end architectures based on deep Convolutionnal Neural Networks (CNN). Such methods require millions of training images that are very difficult to collect and annotate with accurate 3D poses. We propose a solution to generate a large set of photorealistic synthetic images of humans with 3D pose annotations. At the heart of our approach is our image-based synthesis engine that artificially augments a dataset of real images with 2D annotations of body poses using a library of 3D Motion Capture (MoCap) data. These synthetic images are then used to train an end-to-end CNN for full-body 3D pose estimation. We cluster the training data into a large number K of pose classes and tackle pose estimation as a K-way classification problem. Such approach is tractable only with very large training datasets such as ours. Our method outperforms state-of-the-art results in terms of 3D pose estimation in controlled environments and show promising results for in-the-wild images.

#8 Regret of Queueing Bandits

Subhashini Krishnasamy (The Univ. of Texas at Austin)
Rajat Sen (The Univ. of Texas at Austin)
Ramesh Johari
Sanjay Shakkottai (The Univ. of Texas at Aus)

We consider a variant of the multiarmed bandit problem where jobs queue for service, and service rates of different servers may be unknown. We study algorithms that minimize queue-regret: the (expected) difference between the queue-lengths obtained by the algorithm, and those obtained by a genie-aided matching algorithm that knows exact service rates. A naive view of this problem would suggest that queue-regret should grow logarithmically: since queue-regret cannot be larger than classical regret, results for the standard MAB problem give algorithms that ensure queue-regret increases no more than logarithmically in time. Our paper shows surprisingly more complex behavior. In particular, the naive intuition is correct as long as the bandit algorithm's queues have relatively long regenerative cycles: in this case queue-regret is similar to cumulative regret, and scales (essentially) logarithmically. However, we show that this "early stage" of the queueing bandit eventually gives way to a "late stage", where the optimal queue-regret scaling is O(1/t). We demonstrate an algorithm that (order-wise) achieves this asymptotic queue-regret, and also exhibits close to optimal switching time from the early stage to the late stage.

#9 Efficient Nonparametric Smoothness Estimation

Shashank Singh (Carnegie Mellon Univ.)
SimonDu S Du (Carnegie Mellon Univ.)
Barnabas Poczos

Sobolev quantities (norms, inner products, and distances) of probability density functions are important in the theory of nonparametric statistics, but have rarely been used in practice, partly due to a lack of practical estimators. They also include, as special cases, L^2 quantities which are used in many applications. We propose and analyze a family of estimators for Sobolev quantities of unknown probability density functions. We bound the bias and variance of our estimators over finite samples, finding that they are generally minimax rate-optimal. Our estimators are significantly more computationally tractable than previous estimators, and exhibit a statistical/computational trade-off allowing them to adapt to computational constraints. We also draw theoretical connections to recent work on fast two-sample testing. Finally, we empirically validate our estimators on synthetic data.

#10 Completely random measures for modelling block-structured sparse networks

Tue Herlau
Mikkel N. N Schmidt (DTU)
Morten Mørup (Technical Univ. of Denmark)

Statistical methods for network data often parameterize the edge-probability by attributing latent traits to the vertices such as block structure and assume exchangeability in the sense of the Aldous-Hoover representation theorem. These assumptions are however incompatible with traits found in real-world network such as a power-law degree-distribution. Recently Caron & Fox (2014) proposed the use of a different notion of exchangeability due to Kallenberg (2005) and obtained a network model which models edge-inhomogeneity such as power-law degree-distribution while retaining desirable statistical properties. However, this model does not capture latent vertex traits such as block-structure. In this work we re-introduce the use of block-structure for network models obeying Kallenberg's notion of exchangeability and thereby obtain a collapsed model which admits the inference of block-structure and edge inhomogeneity. We derive a simple expression for the likelihood and an efficient sampling method. The obtained model is not significantly more difficult to implement than existing approaches to block-modelling and performs well on real network datasets.

#11 DISCO Nets : DISsimilarity COefficients Networks

Diane Bouchacourt (Univ. of Oxford)
Pawan K Mudigonda (Univ. of Oxford)
Sebastian Nowozin

We present a new type of probabilistic model which we call DISsimilarity COefficient Networks (DISCO Nets). DISCO Nets allow us to efficiently sample from a posterior distribution parametrised by a neural network. During training, DISCO Nets are learned by minimising the dissimilarity coefficient between the true distribution and the estimated distribution. This allows us to tailor the training to the loss related to the task at hand. We empirically show that (i) by modeling uncertainty on the output value, DISCO Nets outperform equivalent non-probabilistic predictive networks and (ii) DISCO Nets accurately model the uncertainty of the output, outperforming existing probabilistic models based on deep neural networks.

#12 An Architecture for Deep, Hierarchical Generative Models

Philip Bachman

We present an architecture which makes it easy to train deep, directed generative models with many layers of latent variables. We facilitate learning by providing deterministic connections between latent variables and the generated output, and by providing a richer set of connections between computations for inference and generation, which enables more effective communication of information throughout the model during training. Our approach permits end-to-end training of models with 10+ hierarchical layers of latent variables. We present experiments showing that our approach achieves state of the art performance on standard image modelling benchmarks, can expose latent class structure in the absence of label information, and can provide convincing imputations of occluded regions in natural images.

#13 A Multi-Batch L-BFGS Method for Machine Learning

Albert S Berahas (Northwestern Univ.)
Jorge Nocedal (Northwestern Univ.)
Martin Takac (Lehigh Univ.)

The question of how to parallelize the stochastic gradient descent (SGD) method has received much attention in the literature. In this paper, we focus instead on batch methods that use a sizeable fraction of the training set at each iteration to facilitate parallelism, and that employ second-order information. In order to improve the learning process, we follow a multi-batch approach in which the batch changes at each iteration. This inherently gives the algorithm a stochastic flavor that can cause instability in L-BFGS, a popular batch method in machine learning. These difficulties arise because L-BFGS employs gradient differences to update the Hessian approximations; when these gradients are computed using different data points the process can be unstable. This paper shows how to perform stable quasi-Newton updating in the multi-batch setting, illustrates the behavior of the algorithm in a distributed computing platform, and studies its convergence properties for both the convex and nonconvex cases.

#14 Higher-Order Factorization Machines

Mathieu Blondel (NTT)
Akinori Fujino (NTT)
Naonori Ueda
Masakazu Ishihata (Hokkaido Univ.)

Factorization machines (FMs) are a supervised learning approach that can use second-order feature combinations even when the data is very high-dimensional. Unfortunately, despite increasing interest in FMs, there exists to date no efficient training algorithm for higher-order FMs (HOFMs). In this paper, we present the first generic yet efficient algorithms for training arbitrary-order HOFMs. We also present new variants of HOFMs with shared parameters, which greatly reduce model size and prediction times while maintaining similar accuracy. We demonstrate the proposed approaches on four different link prediction tasks.

#15 A Bio-inspired Redundant Sensing Architecture

Anh Tuan Nguyen (Univ. of Minnesota)
Jian Xu (Univ. of Minnesota)
Zhi Yang (Univ. of Minnesota)

Sensing is the process of deriving signals from the environment that allows artificial systems to interact with the physical world. The Shannon theorem specifies the maximum rate at which information can be acquired. However, this upper bound is hard to achieve in many man-made systems. The biological visual systems, on the other hand, have highly efficient signal representation and processing mechanisms that allow precise sensing. In this work, we argue that redundancy is one of the critical characteristics for such superior performance. We show architectural advantages by utilizing redundant sensing, including correction of mismatch error and significant precision enhancement. For a proof-of-concept demonstration, we have designed a heuristic-based analog-to-digital converter - a zero-dimensional quantizer. Through Monte Carlo simulation with the error probabilistic distribution as a priori, the performance approaching the Shannon limit is feasible. In actual measurements without knowing the error distribution, we observe at least 2-bit extra precision. The results may also help explain biological processes including the dominance of binocular vision, the functional roles of the fixational eye movements, and the structural mechanisms allowing hyperacuity.

#16 Learning Supervised PageRank with Gradient-Based and Gradient-Free Optimization Methods

Lev Bogolubsky
Pavel Dvurechenskii (Weierstrass Institute for Appl)
Alexander Gasnikov
Gleb Gusev (Yandex LLC)
Yurii Nesterov
Andrei M Raigorodskii
altsoph Tikhonov
Maksim Zhukovskii

In this paper, we consider a non-convex loss-minimization problem of learning Supervised PageRank models, which can account for some properties not considered by classical approaches such as the classical PageRank model. We propose gradient-based and random gradient-free methods to solve this problem. Our algorithms are based on the concept of an inexact oracle and unlike the state state-of-the-art gradient-based method we manage to provide theoretically the convergence rate guarantees for both of them. Finally, we apply proposed optimization algorithms to the web page ranking problem and compare proposed and state-of-the-art algorithms in terms of the considered loss function.

#17 Linear Relaxations for Finding Diverse Elements in Metric Spaces

Aditya Bhaskara (Univ. of Utah)
Mehrdad Ghadiri (Sharif Univ. of Technolog)
Vahab Mirrokni (Google)
Ola Svensson (EPFL)

Choosing a diverse subset of a large collection of points in a metric space is a fundamental problem, with applications in feature selection, recommender systems, web search, data summarization, etc. Various notions of diversity have been proposed, tailored to different applications. The general algorithmic goal is to find a subset of points that maximize diversity, while obeying a cardinality (or more generally, matroid) constraint. The goal of this paper is to develop a novel linear programming (LP) framework that allows us to design approximation algorithms for such problems. We study an objective known as {\em sum-min} diversity, which is known to be effective in many applications, and give the first constant factor approximation algorithm. Our LP framework allows us to easily incorporate additional constraints, as well as secondary objectives. We also prove a hardness result for two natural diversity objectives, under the so-called {\em planted clique} assumption. Finally, we study the empirical performance of our algorithm on several standard datasets. We first study the approximation quality of the algorithm by comparing with the LP objective. Then, we compare the quality of the solutions produced by our method with other popular diversity maximization algorithms.

#18 Stochastic Optimization for Large-scale Optimal Transport

Aude Genevay (Université Paris Dauphine)
Marco Cuturi
Gabriel Peyré
Francis Bach

Optimal transport (OT) defines a powerful framework to compare probability distributions in a geometrically faithful way. However, the practical impact of OT is still limited because of its computational burden. We propose a new class of stochastic optimization algorithms to cope with large-scale problems routinely encountered in machine learning applications. These methods are able to manipulate arbitrary distributions (either discrete or continuous) by simply requiring to be able to draw samples from them, which is the typical setup in high-dimensional learning problems. This alleviates the need to discretize these densities, while giving access to provably convergent methods that output the correct distance without discretization error. These algorithms rely on two main ideas: (a) the dual OT problem can be re-cast as the maximization of an expectation; (b) entropic regularization of the primal OT problem results in a smooth dual optimization optimization which can be addressed with algorithms that have a provably faster convergence. We instantiate these ideas in three different computational setups: (i) when comparing a discrete distribution to another, we show that incremental stochastic optimization schemes can beat the current state of the art finite dimensional OT solver (Sinkhorn's algorithm) ; (ii) when comparing a discrete distribution to a continuous density, a re-formulation (semi-discrete) of the dual program is amenable to averaged stochastic gradient descent, leading to better performance than approximately solving the problem by discretization ; (iii) when dealing with two continuous densities, we propose a stochastic gradient descent over a reproducing kernel Hilbert space (RKHS). This is currently the only known method to solve this problem, and is more efficient than discretizing beforehand the two densities. We backup these claims on a set of discrete, semi-discrete and continuous benchmark problems.

#19 Threshold Bandits, With and Without Censored Feedback

Jacob D Abernethy
Kareem Amin
Ruihao Zhu (MIT)

We consider the Threshold Bandit} setting, a variant of the classical multi-armed bandit problem in which the reward on each round depends on a piece of side information known as a threshold value}. The learner selects one of k actions (arms), this action generates a random sample from a fixed distribution, and the action then receives a unit payoff in the event that this sample exceeds the threshold value. We consider two versions of this problem, the **uncensored and censored** case, that determine whether the sample is always observed or only when the threshold is not met. Using new tools to understand the popular UCB algorithm, we show that the uncensored case is essentially no more difficult than the classical multi-armed bandit setting. Finally we show that the censored case exhibits more challenges, but we give guarantees in the event that the sequence of threshold values is generated optimistically.

#20 Mistake Bounds for Binary Matrix Completion

Mark Herbster
Stephen Pasteris (UCL)
Massimiliano Pontil

We study the problem of completing a binary matrix in an online learning setting. On each trial we predict a matrix entry and then receive the true entry. We propose a Matrix Exponentiated Gradient algorithm[1] to solve this problem. We provide a mistake bound for the algorithm, which scales with the {\em margin complexity} [2,3] of the underlying matrix. The bound suggests an interpretation where each row of the matrix is a prediction task over a finite set of objects, the columns. Using this we show that the algorithm makes a number of mistakes which is comparable up to a logarithmic factor to the number of mistakes made by the Kernel Perceptron with an optimal kernel in hindsight. We discuss applications of the algorithm to predicting as well as the best biclustering and to the problem of predicting the labeling of a graph without knowing the graph in advance.

#21 Learning Sound Representations from Unlabeled Video

Yusuf Aytar (MIT)
Carl Vondrick (MIT)
Antonio Torralba

In this paper we propose to learn semantically rich natural sound representations using big sound data collected in the wild. Harnessing from the natural synchronization between vision and sound, we learn an acoustic representation from a large collection of (2M videos) unlabeled videos. Unlabeled video has the advantage that it can be economically acquired at massive scales, yet contains useful signals about natural sound. We propose a student-teacher training procedure which transfers this discriminative visual knowledge from well established visual models (e.g. ImageNet and Places2 CNNs) into sound domain using unlabeled video as a bridge. Our sound representation yields significant performance improvements over the state-of-the-art results on standard benchmarks for acoustic scene/object classification.

#22 Doubly Convolutional Neural Networks

Shuangfei Zhai (Binghamton Univ.)
Yu Cheng (IBM Research)
Zhongfei (Mark) Zhang (Binghamton Univ.)

Building large models with parameter sharing accounts for most of the success of deep convolutional neural networks (CNNs). In this paper, we propose doubly convolutional neural networks (DCNNs), which significantly improve the performance of CNNs by further exploring this idea. In stead of allocating a set of convolutional filters that are independently learned, a DCNN maintains groups of filters where filters within each group are translated versions of each other. Practically, a DCNN can be easily implemented by a two-step convolution procedure, which is supported by most modern deep learning libraries. We perform extensive experiments on three image classification benchmarks: CIFAR-10, CIFAR-100 and ImageNet, and show that DCNNs consistently outperform other competing architectures, with a margin. We have also verified that replacing a convolutional layer with a doubly convolutional layer at any depth of a CNN can improve its performance. Moreover, various design choices of DCNNs are demonstrated, which shows that DCNN can serve the dual purpose of building more accurate models and/or reducing the memory footprint without sacrificing the accuracy.

#23 Maximizing Influence in an Ising Network: A Mean-Field Optimal Solution

Christopher Lynn (Univ. of Pennsylvania)
Daniel D Lee (Univ. of Pennsylvania)

The problem of influence maximization in social networks has typically been studied in the context of contagion models and irreversible processes. In this paper, we consider an alternate model that treats individual opinions as spins in an Ising network at dynamic equilibrium. We formalize the Ising influence maximization (IIM) problem, which has a physical interpretation as the maximization of the magnetization given a budget of external magnetic field. Under the mean-field (MF) approximation, we develop a number of sufficient conditions for when the problem is convex and exactly solvable, and we provide a gradient ascent algorithm that efficiently achieves an ϵ-approximation to the optimal solution. We show that optimal strategies exhibit a phase transition from focusing influence on high-degree individuals at high interaction strengths to spreading influence among low-degree individuals at low interaction strengths. We also establish a number of novel results about the structure of steady-states in the ferromagnetic MF Ising model on general graphs, which are of independent interest.

#24 Learning from Rational Behavior: Predicting Solutions to Unknown Linear Programs

Shahin Jabbari (Univ. of Pennsylvania)
Ryan M Rogers (Univ. of Pennsylvania)
Aaron Roth
Steven Z. Wu (Univ. of Pennsylvania)

We define and study the problem of predicting the solution to a linear program (LP) given only partial information about its objective and constraints. This generalizes the problem of learning to predict the purchasing behavior of a rational agent who has an unknown objective function, that has been studied under the name "Learning from Revealed Preferences". We give mistake bound learning algorithms in two settings: in the first, the objective of the LP is known to the learner but there is an arbitrary, fixed set of constraints which are unknown. Each example is defined by an additional known constraint and the goal of the learner is to predict the optimal solution of the LP given the union of the known and unknown constraints. This models the problem of predicting the behavior of a rational agent whose goals are known, but whose resources are unknown. In the second setting, the objective of the LP is unknown, and changing in a controlled way. The constraints of the LP may also change every day, but are known. An example is given by a set of constraints and partial information about the objective, and the task of the learner is again to predict the optimal solution of the partially known LP.

#25 Fairness in Learning: Classic and Contextual Bandits

Matthew Joseph (Univ. of Pennsylvania)
Michael Kearns
Jamie H Morgenstern (Univ. of Pennsylvania)
Aaron Roth

We introduce the study of fairness in contextual multi-armed bandit problems. Our fairness definition can be interpreted as demanding that given a pool of applicants (say, for college admission or mortgages), a worse applicant is never favored over a better one, despite a learning algorithm's uncertainty over the true payoffs. Our main results prove a tight connection between fairness and the KWIK (Knows What It Knows) learning model: a KWIK algorithm for a class of functions can be transformed into a provably fair contextual bandit algorithm, and conversely any fair contextual bandit algorithm can be transformed into a KWIK learning algorithm. This tight connection allows us to provide a provably fair algorithm for the linear contextual bandit problem with a polynomial dependence on the dimension, and to show (for a different class of functions) a worst-case exponential gap in regret between fair and non-fair learning algorithms.

#26 A Powerful Generative Model Using Random Weights for the Deep Image Representation

Kun He (Huazhong Univ. of Science and Technology)
Yan Wang (HUAZHONG Univ. OF SCIENCE)
John Hopcroft (Cornell Univ.)

To what extent is the success of deep visualization due to the training? Could we do deep visualization using untrained, random weight networks? To address this issue, we explore new and powerful generative models for three popular deep visualization tasks using untrained, random weight convolutional neural networks. First we invert representations in feature spaces and reconstruct images from white noise inputs. The reconstruction quality is statistically higher than that of the same method applied on well trained networks with the same architecture. Next we synthesize textures using scaled correlations of representations in multiple layers and our results are almost indistinguishable with the original natural texture and the synthesized textures based on the trained network. Third, by recasting the content of an image in the style of various artworks, we create artistic images with high perceptual quality, highly competitive to the prior work of Gatys et al. on pretrained networks. To our knowledge this is the first demonstration of image representations using untrained deep neural networks. Our work provides a new and fascinating tool to study the representation of deep network architecture and sheds light on new understandings on deep visualization. It may possibly lead to a way to compare network architectures without training.

#27 Improved Error Bounds for Tree Representations of Metric Spaces

Samir Chowdhury (The Ohio State Univ.)
Facundo Mémoli
Zane T Smith

Estimating optimal phylogenetic trees or hierarchical clustering trees from metric data is an important problem in evolutionary biology and data analysis. Intuitively, the goodness-of-fit of a metric space to a tree depends on its inherent treeness, as well as other metric properties such as intrinsic dimension. Existing algorithms for embedding metric spaces into tree metrics provide distortion bounds depending on cardinality. Because cardinality is a simple property of any set, we argue that such bounds do not fully capture the rich structure endowed by the metric. We consider an embedding of a metric space into a tree proposed by Gromov. By proving a stability result, we obtain an improved additive distortion bound depending only on the hyperbolicity and doubling dimension of the metric. We observe that Gromov's method is dual to the well-known single linkage hierarchical clustering (SLHC) method. By means of this duality, we are able to transport our results to the setting of SLHC, where such additive distortion bounds were previously unknown.

#28 Adaptive optimal training of animal behavior

Ji Hyun Bak (Princeton Univ.)
Jung Choi
Ilana Witten
Jonathan W Pillow

Neuroscience experiments often require training animals to perform tasks designed to elicit various sensory, cognitive, and motor behaviors. Training typically involves a series of gradual adjustments of stimulus conditions and rewards in order to bring about learning. However, training protocols are usually hand-designed, relying on a combination of intuition, guesswork, and trial-and-error, and often require weeks or months to achieve a desired level of task performance. Here we combine ideas from reinforcement learning and optimal experimental design to formulate methods for adaptive optimal training of animal behavior. Our work addresses two intriguing problems at once: first, it seeks to infer the learning rules underlying an animal's behavioral changes during training; second, it seeks to exploit these rules to select stimuli that will maximize the rate of learning toward a desired objective. We develop and test these methods using data collected from rats during training on an auditory discrimination task. We show that we can accurately infer the parameters of a policy-gradient-based learning algorithm that describes how the animal's internal model of the task evolves over the course of training. We then formulate a theory for optimal training, which involves selecting sequences of stimuli that will drive the animal's internal policy toward a desired location in the parameter space. Simulations show that our adaptive training method can achieve a substantial speedup over standard training methods. These results will hold broad theoretical interest for researchers in reinforcement learning, and offer immense practical benefits to neuroscientists tasked with training animals.

#29 PAC-Bayesian Theory Meets Bayesian Inference

Pascal Germain
Francis Bach
Alexandre Lacoste
Simon Lacoste-Julien (INRIA)

We exhibit a strong link between frequentist PAC-Bayesian bounds and the Bayesian marginal likelihood. That is, for the negative log-likelihood loss function, we show that the minimization of PAC-Bayesian generalization bounds maximizes the Bayesian marginal likelihood. This provides an alternative explanation to the Bayesian Occam's razor criteria, under the assumption that the data is generated by a i.i.d. distribution. Moreover, as the negative log-likelihood is an unbounded loss function, we motivate and propose a PAC-Bayesian theorem tailored for the sub-Gamma loss family, and we show that our approach is sound on classical Bayesian linear regression tasks.

#30 Nearly Isometric Embedding by Relaxation

James McQueen (Univ. of Washington)
Marina Meila (Univ. of Washington)
Dominique Joncas (Google)

Many manifold learning algorithms aim to create embeddings with low or no distortion (i.e. isometric). If the data has intrinsic dimension d, it is often impossible to obtain an isometric embedding in d dimensions, but possible in s > d dimensions. Yet, most geometry preserving algorithms cannot do the latter. This paper proposes an embedding algorithm that overcomes this problem. The algorithm directly computes, for any data embedding Y, a distortion loss(Y), and iteratively updates Y in order to decrease it. The distortion measure we propose is based on the push-forward Riemannian metric associated with the coordinates Y. The experiments confirm the superiority of our algorithm in obtaining low distortion embeddings.

#31 Graph Clustering: Block-models and model free results

Yali Wan (Univ. of Washington)
Marina Meila (Univ. of Washington)

Clustering graphs under the Stochastic Block Model (SBM) and extensions are well studied. Guarantees of correctness exist under the assumption that the data is sampled from a model. In this paper, we propose a framework, in which we obtain "correctness" guarantees without assuming the data comes from a model. The guarantees we obtain depend instead on the statistics of the data that can be checked. We also show that this framework ties in with the existing model-based framework, and that we can exploit results in model-based recovery, as well as strengthen the results existing in that area of research.

#32 Learning Transferrable Representations for Unsupervised Domain Adaptation

Ozan Sener (Cornell Univ.)
Hyun Oh Song (Google Research)
Ashutosh Saxena (Brain of Things)
Silvio Savarese (Stanford Univ.)

Supervised learning with large scale labelled datasets and deep layered models has caused a paradigm shift in diverse areas in learning and recognition. However, this approach still suffers from generalization issues under the presence of a domain shift between the training and the test data distribution. Since unsupervised domain adaptation algorithms directly address this domain shift

problem between a labelled source dataset and an unlabelled target dataset, recent papers have shown promising results by fine-tuning the networks with domain adaptation loss functions which try to align the mismatch between the training and testing data distributions. Nevertheless, these recent deep learning based domain adaptation approaches still suffer from issues such as high sensitivity to the gradient reversal hyperparameters and overfitting during the fine-tuning stage. In this paper, we propose a unified deep learning framework where the representation, cross domain transformation, and target label inference are all jointly optimized in an end-to-end fashion for unsupervised domain adaptation. Our experiments show that the proposed method significantly outperforms state-of-the-art algorithms in both object recognition and digit classification experiments by a large margin. We will make our learned models as well as the source code available immediately upon acceptance.

#33 Measuring Neural Net Robustness with Constraints

Osbert Bastani (Stanford Univ.)
Yani Ioannou (Univ. of Cambridge)
Leonidas Lampropoulos (Univ. of Pennsylvania)
Dimitrios Vytiniotis (Microsoft Research)
Aditya Nori (Microsoft Research)
Antonio Criminisi

Despite having high accuracy, neural nets have been shown to be susceptible to adversarial examples, where a small perturbation to an input can cause it to become mislabeled. We propose metrics for measuring the robustness of a neural net and devise a novel algorithm for approximating these metrics based on an encoding of robustness as a linear program. We show how our metrics can be used to evaluate the robustness of deep neural nets with experiments on the MNIST and CIFAR-10 datasets. Our algorithm generates more informative estimates of robustness metrics compared to estimates based on existing algorithms. Furthermore, we show how existing approaches to improving robustness "overfit" to adversarial examples generated using a specific algorithm. Finally, we show that our techniques can be used to additionally improve neural net robustness both according to the metrics that we propose, but also according to previously proposed metrics.

#34 Forward models at Purkinje synapses facilitate cerebellar anticipatory control

Ivan Herreros (Universitat Pompeu Fabra)
Xerxes Arsiwalla
Paul Verschure

How does our motor system solve the problem of anticipatory control in spite of a wide spectrum of response dynamics from different musculo-skeletal systems, transport delays and response latencies throughout the central nervous system? To a great extend, our highly-skilled motor responses are a result of a reactive feedback system, originating in the brain-stem and spinal cord, combined with a feed-forward anticipatory system, that is adaptively fine-tuned by sensory experience and originates in the cerebellum. In this work, we design an anticipatory adaptive motor control architecture, based on cerebellar anatomy. For synapses of cerebellar Purkinje cells, we derive a novel synaptic learning rule that involves an eligibility trace and converges to the optimal solution. The existence of eligibility trace provides a mechanism beyond co-incidence detection in that it convolves a history of prior inputs from parallel fibers with the error signals coming from climbing fibers. Our solution implies that Purkinje cell synapses should generate eligibility traces using a forward model of the system being controlled. From an engineering perspective, our model provides a general-purpose anticipatory control architecture equipped with a learning rule that exploits the dynamics of the closed-loop system.

#35 Estimating Nonlinear Neural Response Functions using GP Priors and Kronecker Methods

Cristina Savin (IST Austria)
Gasper Tkacik (Institute of Science and Technology Austria)

Jointly characterizing neural responses in terms of several external variables promises novel insights into circuit computation, but remains computationally prohibitive in practice. Here we use gaussian process (GP) priors and exploit recent advances in fast GP inference and learning, based on Kronecker methods, to efficiently estimate multidimensional nonlinear tuning functions. Our estimators require considerably less data than traditional methods and further provide principled uncertainty estimates. We apply these tools to hippocampal recordings during open field exploration and use them to characterize the joint dependence of CA1 responses on the position of the animal and several other variables, including the animal's speed, direction of motion, and network oscillations. Our results provide an unprecedentedly detailed quantification of the tuning of hippocampal neurons. The model's generality suggests that our approach can be used to estimate neural response properties in other cortical regions.

#36 A Bayesian method for reducing bias in neural representational similarity analysis

Mingbo Cai (Princeton Univ.)
Nicolas W Schuck (Princeton Neuroscience Institute)
Jonathan W Pillow
Yael Niv

In neuroscience, the similarity matrix of neural activity patterns in response to different sensory stimuli or under different cognitive states reflects the structure of neural representational space. Existing methods derive point estimations of neural activity patterns from noisy neural imaging data, and the similarity is calculated from these point estimations. We show that this approach translates structured noise from estimated patterns into spurious bias structure in the resulting similarity matrix, which is especially severe when signal-to-noise ratio is low and experimental conditions cannot be fully randomized in a cognitive task. We propose an alternative Bayesian framework for computing representational similarity in which we treat the covariance structure of neural activity patterns as a hyper-parameter in a generative model of the neural data, and directly estimate this covariance structure from imaging data while marginalizing over the unknown activity patterns. Converting the estimated covariance structure into a correlation matrix offers an unbiased estimate of neural representational similarity. Our method can also simultaneously estimate a signal-to-noise map that informs where the learned representational structure is supported more strongly, and the learned covariance matrix can be used as a structured prior to constrain Bayesian estimation of neural activity patterns.

#37 Learning to Communicate with Deep Multi-Agent Reinforcement Learning

Jakob Foerster (Univ. of Oxford)
Yannis M. Assael (Univ. of Oxford)
Nando de Freitas (Univ. of Oxford)
Shimon Whiteson

We consider the problem of multiple agents sensing and acting in environments with the goal of maximising their shared utility. In these environments, agents must learn communication protocols in order to share information that is needed to solve the tasks. By embracing deep neural networks, we are able to demonstrate end-to-end learning of protocols in complex environments inspired by communication riddles and multi-agent computer vision problems with partial observability. We propose two approaches for learning in these domains: Reinforced Inter-Agent Learning (RIAL) and Differentiable Inter-Agent Learning (DIAL). The former uses deep Q-learning, while the latter exploits the fact that, during learning, agents can propagate error derivatives through (noisy) communication channels. Hence, this approach uses centralised learning but decentralised execution. Our experiments introduce new environments for studying the learning of communication protocols and present a set of engineering innovations that are essential for success in these domains.

#38 Total Variation Classes Beyond 1d: Minimax Rates, and the Limitations of Linear Smoothers

Veeru Sadhanala (Carnegie Mellon Univ.)
Yu-Xiang Wang (Carnegie Mellon Univ.)
Ryan J Tibshirani

We consider the problem of estimating a function defined over n locations on a d-dimensional grid (having all side lengths equal to $n^{1/d}$). When the function is constrained to have discrete total variation bounded by C_n, we derive the minimax optimal (squared) ℓ_2 estimation error rate, parametrized by n, C_n. Total variation denoising, also known as the fused lasso, is seen to be rate optimal. Several simpler estimators exist, such as Laplacian smoothing and Laplacian eigenmaps. A natural question is: can these simpler estimators perform just as well? We prove that these estimators, and more broadly all estimators given by linear transformations of the input data, are suboptimal over the class of functions with bounded variation. This extends fundamental findings of Donoho and Jonestone (1998) on 1-dimensional total variation spaces to higher dimensions. The implication is that the computationally simpler methods cannot be used for such sophisticated denoising tasks, without sacrificing statistical accuracy. We also derive minimax rates for discrete Sobolev spaces over d-dimensional grids, which are, in some sense, smaller than the total variation function spaces. Indeed, these are small enough spaces that linear estimators can be optimal---and a few well-known ones are, such as Laplacian smoothing and Laplacian eigenmaps, as we show. Lastly, we investigate the adaptivity of the total variation denoiser to these smaller Sobolev function spaces.

#39 Exponential Family Embeddings

Maja Rudolph (Columbia Univ.)
Francisco J. R. Ruiz
Stephan Mandt (Disney Research)
David Blei

Word embeddings are a powerful approach to capturing semantic similarity among terms in a vocabulary. In this paper, we develop exponential family embeddings, which extends the idea of word embeddings to other types of high-dimensional data. As examples, we studied several types of data: neural data with real-valued observations, count data from a market basket analysis, and ratings data from a movie recommendation system. The main idea is that each observation is modeled conditioned on a set of latent embeddings and other observations, called the context, where the way the context is defined depends on the problem. In language the context is the surrounding words; in neuroscience the context is close-by neurons; in market basket data the context is other items in the shopping cart. Each instance of an embedding defines the context, the exponential family of conditional distributions, and how the embedding vectors are shared across data. We infer the embeddings with stochastic gradient descent, with an algorithm that connects closely to generalized linear models. On all three of our applications—neural activity of zebrafish, users' shopping behavior, and movie ratings—we found that exponential family embedding models are more effective than other dimension reduction methods. They better reconstruct held-out data and find interesting qualitative structure.

#40 k^*-Nearest Neighbors: From Global to Local

Oren Anava (Technion)
Kfir Levy (Technion)

The weighted k-nearest neighbors algorithm is one of the most fundamental non-parametric methods in pattern recognition and machine learning. The question of setting the optimal number of neighbors as well as the optimal weights has received much attention throughout the years, nevertheless this problem seems to have remained unsettled. In this paper we offer a simple approach to locally weighted regression/classification, where we make the bias-variance tradeoff explicit. Our formulation enables us to phrase a notion of optimal weights, and to efficiently find these weights as well as the optimal number of neighbors efficiently and adaptively, for each data point whose value we wish to estimate}. The applicability of our approach is demonstrated on several datasets, showing superior performance over standard locally weighted methods.

#41 Reward Augmented Maximum Likelihood for Neural Structured Prediction

Mohammad Norouzi
Samy Bengio
ZF Chen
Navdeep Jaitly
Mike Schuster
Yonghui Wu
Dale Schuurmans

A key problem in structured output prediction is enabling direct optimization of the task reward function that matters for test evaluation. This paper presents a simple and computationally efficient method that incorporates task reward into maximum likelihood training. We establish a connection between maximum likelihood and regularized expected reward, showing that they are approximately equivalent in the vicinity of the optimal solution. Then we show how maximum likelihood can be generalized by optimizing the conditional probability of auxiliary outputs that are sampled proportional to their exponentiated scaled rewards. We apply this framework to optimize edit distance in the output space, by sampling from edited targets. Experiments on speech recognition and machine translation for neural sequence to sequence models show notable improvements over maximum likelihood baseline by simply sampling from target output augmentations.

#42 A Probabilistic Model of Social Decision Making based on Reward Maximization

Koosha Khalvati (Univ. of Washington)
Seongmin A. Park (Cognitive Neuroscience Center)
Jean-Claude Dreher (Centre de Neurosciences Cognitives)
Rajesh P Rao (Univ. of Washington)

A fundamental problem in cognitive neuroscience is how humans make decisions, act, and behave in relation to other humans. Here we adopt the hypothesis that when we are in an interactive social setting, our brain performs Bayesian inference of the intentions and cooperativeness of others using probabilistic representations. We employ the framework of partially observable Markov decision processes (POMDPs) to model human decision making in a social context, focusing specifically on the volunteer's dilemma in a version of the classic Public Goods Game. We show that the POMDP model explains both the behavior of subjects as well as neural activity recorded using fMRI during the game. The decisions of subjects can be modeled across all trials using two interpretable parameters. Furthermore, the expected reward predicted by the model for each subject was correlated with the activation of brain areas related to reward expectation in social interactions. Our results suggest a probabilistic basis for human social decision making within the framework of expected reward maximization.

#43 Active Learning with Oracle Epiphany

T.K. Huang (Uber Advanced Technologies Center)
Lihong Li (Microsoft Research)
Ara Vartanian (Univ. of Wisconsin-Madison)
Saleema Amershi (Microsoft)
Jerry Zhu

We present theoretical analysis of active learning with more realistic interactions with human oracles. Previous empirical studies have shown oracles abstaining on difficult queries until accumulating enough information to make label decisions. We formalize this phenomenon with an "oracle epiphany model." We then analyze active learning query complexity under such oracles for both the realizable and the agnostic cases. Our analysis shows that active learning is possible with oracle epiphany, but incurs an additional cost depending on when the epiphany happens. Our results suggest new, principled active learning approaches with realistic oracles.

#44 On Regularizing Rademacher Observation Losses

Richard Nock (Data61 and ANU)

It has recently been shown that supervised learning linear classifiers with two of the most popular losses, the logistic and square loss, is equivalent to optimizing an equivalent loss over sufficient statistics about the class: Rademacher observations (rados). It has also been shown that learning over rados brings solutions to two prominent problems for which the state of the art of learning from examples is comparatively inferior and in fact less convenient: (i) protecting and learning from private examples, (ii) learning from distributed datasets without entity resolution. {B is repetita placent}: the two proofs of equivalence are different and rely on specific properties of the corresponding losses, so whether these can be unified and generalized inevitably comes to mind. This is our first contribution: we show how they can be fit into the same theory for the equivalence between example and rado losses. As a second contribution, we show that the generalization unveils a surprising new connection to regularized learning, and in particular a sufficient condition under which regularizing the loss over examples is equivalent to regularizing the rados (i.e. the data) in the equivalent rado loss, in such a way that an efficient algorithm for one regularized rado loss may be as efficient when changing the regularizer. This is our third contribution: we give a formal boosting algorithm for the regularized exponential rado-loss which boost with any of the ridge, lasso, slope, l_\infty, or elastic net regularizer, using the same master routine for all. Because the regularized exponential rado-loss is the equivalent of the regularized logistic loss over examples we obtain the first efficient proxy to the minimization of the regularized logistic loss over examples using such a wide spectrum of regularizers. Experiments display that regularization significantly improves rado-based learning and compares favourably with example-based learning.

#45 A Non-generative Framework and Convex Relaxations for Unsupervised Learning

Elad Hazan
Tengyu Ma (Princeton Univ.)

We give a novel formal theoretical framework for unsupervised learning with two distinctive characteristics. First, it does not assume any generative model and based on a worst-case performance metric. Second, it is comparative, namely performance is measured with respect to a given hypothesis class. This allows to avoid known computational hardness results and algorithms based on convex relaxations. We show how several families of unsupervised learning models, which were previously only analyzed under probabilistic assumptions and are otherwise provably intractable, can be efficiently learned in our framework by convex optimization.

#46 Learning Tree Structured Potential Games

Vikas Garg (MIT)
Tommi Jaakkola

Many real phenomena, including behaviors, involve strategic interactions that can be learned from data. We focus on learning tree structured potential games where equilibria are represented by local maxima of an underlying potential function. We cast the learning problem within a max margin setting and show that the problem is NP-hard even when the strategic interactions form a tree. We develop a variant of dual decomposition to estimate the underlying game and demonstrate with synthetic and real decision/voting data that the game theoretic perspective (carving out local maxima) enables meaningful recovery.

#47 Equality of Opportunity in Supervised Learning

Moritz Hardt (Google Brain)
Eric Price
Nati Srebro

We propose a statistical notion of fairness, grounded in the principle of equality of opportunity. Our notion applies to essentially all supervised learning tasks, while overcoming conceptual problems that plagued previous algorithmic fairness constraints. We present an easily interpretable and effective framework for achieving our fairness notion based on a geometric post-processing step without the need of changing a possibly complicated training pipeline. Despite its simplicity, we prove a strong optimality principle for our post-processing approach. Finally, we conduct an in-depth analysis of our approach to real FICO credit score data, showing that our approach yields substantially higher utility than previous methods.

#48 Interaction Networks for Learning about Objects, Relations and Physics

Peter Battaglia (Google DeepMind)
Razvan Pascanu
Matthew Lai (Google DeepMind)
Danilo Jimenez Rezende
koray kavukcuoglu (Google DeepMind)

Reasoning about objects, relations, and physics is central to human intelligence, and a key goal of artificial intelligence. Here we introduce the interaction network, a model which can reason about how objects in complex systems interact, supporting dynamical predictions, as well as inferences about the abstract properties of the system. Our model takes graphs as input, performs object- and relation-centric reasoning in a way that is analogous to a simulation, and is implemented using deep neural networks. We evaluate its ability to reason about several challenging physical domains: n-body problems, rigid-body collision, and non-rigid dynamics. Our results show it can be trained to accurately simulate the physical trajectories of dozens of objects over thousands of time steps, estimate abstract quantities such as energy, and generalize automatically to systems with different numbers and configurations of objects and relations. Our interaction network implementation is the first general-purpose, learnable physics engine, and a powerful general framework for reasoning about object and relations in a wide variety of complex real-world domains.

#49 beta-risk: a New Surrogate Risk for Learning from Weakly Labeled Data

Valentina Zantedeschi (UJM Saint-Etienne)
Rémi Emonet
Marc Sebban

During the past few years, the machine learning community has paid attention to developping new methods for learning from weakly labeled data. This field covers different settings like semi-supervised learning, learning with label proportions, multi-instance learning, noise-tolerant learning, etc. This paper presents a generic framework to deal with these weakly labeled scenarios. We introduce the beta-risk as a generalized formulation of the standard empirical risk based on surrogate margin-based loss functions. This risk allows us to express the reliability on the labels and to derive different kinds of learning algorithms. We specifically focus on SVMs and propose a soft margin beta-svm algorithm which behaves better that the state of the art.

#50 Binarized Neural Networks

Itay Hubara (Technion)
Matthieu Courbariaux (Université de Montréal)
Daniel Soudry (Columbia Univ.)
Ran El-Yaniv (Technion)
Yoshua Bengio (Université de Montréal)

We introduce a method to train Binarized Neural Networks (BNNs) - neural networks with binary weights and activations at run-time. At train-time the binary weights and activations are used for computing the parameter gradients. During the forward pass, BNNs drastically reduce memory size and accesses, and replace most arithmetic operations with bit-wise operations, which is expected to substantially improve power efficiency. To validate the effectiveness of BNNs, we conducted two sets of experiments on the Torch7 and Theano frameworks. On both, BNNs achieved nearly state-of-the-art results over the MNIST, CIFAR-10 and SVHN datasets. We also report our preliminary results on the challenging ImageNet dataset. Last but not least, we wrote a binary matrix multiplication GPU kernel with which it is possible to run our MNIST BNN 7 times faster than with an unoptimized GPU kernel, without suffering any loss in classification accuracy. The code for training and running our BNNs is available on-line.

#51 Regularization With Stochastic Transformations and Perturbations for Deep Semi-Supervised Learning

Mehdi Sajjadi (Univ. of Utah)
Mehran Javanmardi (Univ. of Utah)
Tolga Tasdizen (Univ. of Utah)

Effective convolutional neural networks are trained on large sets of labeled data. However, creating large labeled datasets is a very costly and time-consuming task. Semi-supervised learning uses unlabeled data to train a model with higher accuracy when there is a limited set of labeled data available. In this paper, we consider the problem of semi-supervised learning with convolutional neural networks. Techniques such as randomized data augmentation, dropout and random max-pooling provide better generalization and stability for classifiers that are trained using gradient descent. Multiple passes of an individual sample through the network might lead to different predictions due to the non-deterministic behavior of these techniques. We propose an unsupervised loss function that takes advantage of the stochastic nature of these methods and minimizes the difference between the predictions of multiple passes of a training sample through the network. We evaluate the proposed method on several benchmark datasets.

#52 Generating Images with Perceptual Similarity Metrics based on Deep Networks

Alexey Dosovitskiy (Univ. of Freiburg)
Thomas Brox (Univ. of Freiburg)

We propose a class of loss functions, which we call deep perceptual similarity metrics (DeePSiM), allowing to generate sharp high resolution images from compressed abstract representations. Instead of computing distances in the image space, we compute distances between image features extracted by deep neural networks. This metric reflects perceptual similarity of images much better and, thus, leads to better results. We demonstrate two examples of use cases of the proposed loss: (1) networks that invert the AlexNet convolutional network; (2) a modified version of a variational autoencoder that generates realistic high-resolution random images.

#53 Exploiting Tradeoffs for Exact Recovery in Heterogeneous Stochastic Block Models

Amin Jalali (Univ. of Washington)
Qiyang Han (Univ. of Washington)
Ioana Dumitriu (Univ. of Washington)
Maryam Fazel (Univ. of Washington)

The Stochastic Block Model (SBM) is a widely used random graph model for networks with communities. Despite the recent burst of interest in community detection under the SBM from statistical and computational points of view, there are still gaps in understanding the fundamental limits of recovery. In this paper, we consider the SBM in its full generality, where there is no restriction on the number and sizes of communities or how they grow with the number of nodes, as well as on the connectivity probabilities inside or across communities. For such stochastic block models, we provide guarantees for exact recovery via a semidefinite program as well as upper and lower bounds on SBM parameters for exact recoverability. Our results exploit the tradeoffs among the various parameters of heterogenous SBM and provide recovery guarantees for many new interesting SBM configurations.

#54 Tensor Switching Networks

Kenyon Tsai (Harvard Univ.)
Andrew M Saxe
David Cox

Many high-performing deep learning algorithms have been built using relatively simple neural nonlinearities, such as the rectified linear unit (ReLU). Here, we ask if it is possible to take advantage of the special structure of the ReLU to improve the training of neural networks. We present a novel neural network architecture, the Tensor Switching Network (TS Net), which shares with ReLU the property of being linear once it is known which hidden units will be active for a specific example. We develop several algorithms that exploit this linearity: first, we derive kernels that are equivalent to multilayer Tensor Switching networks with infinitely wide hidden layers. Next, we propose a simple algorithm that requires only a single pass through a dataset, and finally, we develop a method suitable for representation learning in the TS Net. Our experimental results show that the TS Net performs surprisingly well in comparison to standard ReLU networks in the regime of very few, or even just one, passes over the training data, suggesting potential applications in online learning and big data.

#55 Finite-Dimensional BFRY Priors and Variational Bayesian Inference for Power Law Models

Juho Lee (POSTECH)
Lancelot F James (HKUST)
Seungjin Choi (POSTECH)

Bayesian nonparametric methods based on the Dirichlet process (DP), gamma process and beta process, have proven effective in capturing aspects of various datasets arising in machine learning. However, it is now recognized that such processes have their limitations in terms of the ability to capture power law behavior. As such there is now considerable interest in models based on the Stable Processs (SP), Generalized Gamma process (GGP) and Stable-beta process (SBP). These models present new challenges in terms of practical statistical implementation. In analogy to tractable processes such as the finite-dimensional Dirichlet process, we describe a class of random processes, we call iid finite-dimensional BFRY processes, that enables one to begin to develop efficient posterior inference algorithms such as variational Bayes that readily scale to massive datasets. For illustrative purposes, we describe a simple variational Bayes algorithm for normalized SP mixture models, and demonstrate its usefulness with experiments on synthetic and real-world datasets.

#56 Temporal Regularized Matrix Factorization for High-dimensional Time Series Prediction

Hsiang-Fu (Rofu) Yu (Univ. of Texas at Austin)
Nikhil Rao
Inderjit S Dhillon

Time series prediction problems are becoming increasingly high-dimensional in modern applications, such as climatology and demand forecasting. For example, in the latter problem, the number of items for which demand needs to be forecast might be as large as 50,000. In addition, the data is generally noisy and full of missing values. Thus, modern applications require methods that are highly scalable, and can deal with noisy data in terms of corruptions or missing values. However, classical time series methods usually fall short of handling these issues. In this paper, we present a temporal regularized matrix factorization (TRMF) framework which supports data-driven temporal learning and forecasting. We develop novel regularization schemes and use scalable matrix factorization methods that are eminently suited for high-dimensional time series data that has many missing values. Our proposed TRMF is highly general, and subsumes many existing approaches for time series analysis. We make interesting connections to graph regularization methods in the context of learning the dependencies in an autoregressive framework. Experimental results show the superiority of TRMF in terms of scalability and prediction quality. In particular, TRMF is two orders of magnitude faster than other methods on a problem of dimension 50,000, and generate better forecasts on real-world datasets such as Wal-mart E-commerce datasets.

#57 Composing graphical models with neural networks for structured representations and fast inference

Matthew Johnson
David Duvenaud
Alex Wiltschko (Harvard Univ. and Twitter)
Ryan P Adams
Sandeep R Datta (Harvard Medical School)

We propose a general modeling and inference framework that composes probabilistic graphical models with deep learning methods and combines their respective strengths. Our model family augments graphical structure in latent variables with neural network observation models. For inference, we extend variational autoencoders to use graphical model approximating distributions, paired with recognition networks that output conjugate potentials. All components of these models are learned simultaneously with a single objective, giving a scalable algorithm that leverages stochastic variational inference, natural gradients, graphical model message passing, and the reparameterization trick. We illustrate this framework with several example models and an application to mouse behavioral phenotyping.

#58 Contextual-MDPs for PAC Reinforcement Learning with Rich Observations

Akshay Krishnamurthy
Alekh Agarwal (Microsoft)
John Langford

We propose and study a new tractable model for reinforcement learning with rich observations called Contextual-MDPs, generalizing contextual bandits to sequential decision making. These models require an agent to take actions based on observations (features) with the goal of achieving long-term performance competitive with a large set of policies. To avoid barriers to sample-efficient learning associated with large observation spaces and general POMDPs, Contextual-MDPs can be summarized by a small number of hidden states and long-term rewards are predictable by a reactive function class. In this setting, we design a new reinforcement learning algorithm that engages in global exploration and analyze its sample complexity. We prove that the algorithm learns near optimal behavior after a number of episodes that is polynomial in all relevant parameters, logarithmic in the number of policies, and independent of the size of the observation space. This represents an exponential improvement over all existing alternative approaches and provides theoretical justification for reinforcement learning with function approximation.

#59 Algorithms and matching lower bounds for approximately-convex optimization

Andrej Risteski (Princeton Univ.)
Yuanzhi Li (Princeton Univ.)

In recent years, a rapidly increasing number of applications in practice requires solving non-convex objectives, like training neural networks, learning graphical models, maximum likelihood estimation etc. Though simple heuristics such as gradient descent with very few modifications tend to work well, theoretical understanding is very weak. We consider possibly the most natural class of non-convex functions where one could hope to obtain provable guarantees: functions that are ``approximately convex", i.e. functions tf: *Real*^{d} -->*Real* for which there exists a convex function f such that for all x, |\tf(x) - f(x)| ≤ *errnoise* for a fixed value*errnoise*. We then want to minimize*tf*, i.e. output a point*tx* such that \tf(\tx) \le \min_{x} \tf(x) + \err. It is quite natural to conjecture that for fixed \err, the problem gets harder for larger \errnoise, however, the exact dependency of \ err and \errnoise was not known. In this paper, we provide the first essentially tight} characterization of the rate of \errnoise as a function of \err: we identify a function T: \Real^{+} \to \Real^{+} such that when \errnoise = O(T(\err)), we can give an algorithm that outputs a point \tx such that \tf(\tx) \le \min_{x} \tf(x) + \err within time \ poly\left(d, \frac{1}{\err}\right). On the other hand, when \errnoise = Ω(T(\err)), we also prove an information theoretic} lower bound that any algorithm that outputs such a \tx must use super polynomial} number of evaluations of \tf.

#60 Fast Stochastic Methods for Nonsmooth Nonconvex Optimization

Sashank J. Reddi (Carnegie Mellon Univ.)
Suvrit Sra (MIT)
Barnabas Poczos
Alex J Smola

We analyze stochastic algorithms for optimizing nonconvex, nonsmooth finite-sum problems, where the nonconvex part is smooth and the nonsmooth part is convex. Surprisingly, unlike the smooth case, our knowledge of this fundamental problem is very limited. For example, it is not known whether the proximal stochastic gradient method with constant minibatch converges to a stationary point. To tackle this issue, we develop fast stochastic algorithms that provably converge to a stationary point for constant minibatches. Furthermore, using a variant of these algorithms, we show provably faster convergence than batch proximal gradient descent. Our results are based on the recent variance reduction techniques for convex optimization, with a novel analysis for nonconvex and nonsmooth functions. Finally, we prove global linear convergence rate for an interesting subclass of nonsmooth nonconvex functions, that subsumes several recent works.

#61 A Simple Practical Accelerated Method for Finite Sums

Aaron Defazio (Ambiata)

We describe a novel optimization method for finite sums (such as empirical risk minimization problems) building on the recently introduced SAGA method. Our method achieves an accelerated convergence rate on strongly convex smooth problems matching the optimal possible rate. Our method has only one parameter (a step size), and is radically simpler than other accelerated methods for finite sums. Additionally it can be applied when the terms are non-smooth, yielding a method applicable in many areas where operator splitting methods would traditionally be applied.

#62 Unsupervised Learning for Physical Interaction through Video Prediction

Chelsea Finn (Google)
Ian Goodfellow
Sergey Levine (Univ. of Washington)

A core challenge for an agent learning to interact with the world is to predict how its actions affect objects in its environment. Many existing methods for learning the dynamics of physical interactions require labeled object information. However, to scale real-world interaction learning to a variety of scenes and objects, acquiring labeled data becomes increasingly impractical. To learn about physical object motion without labels, we develop an action-conditioned video prediction model that explicitly models pixel motion, by predicting a distribution over pixel motion from previous frames. Because our model explicitly predicts motion, it is partially invariant to object appearance, enabling it to generalize to previously unseen objects. To explore video prediction for real-world interactive agents, we also introduce a dataset of 50,000 robot interactions involving pushing motions, including a test set with novel objects. In this dataset, accurate prediction of videos conditioned on the robot's future actions amounts to learning a "visual imagination" of different futures based on different courses of action. Our experiments show that our proposed method not only produces more accurate video predictions, but also more accurately predicts object motion, when compared to prior methods.

#63 Threshold Learning for Optimal Decision Making

Nathan F Lepora (Univ. of Bristol)

Decision making under uncertainty is commonly modeled as a process of competitive stochastic evidence accumulation to threshold. Then an animal should learn these thresholds to optimize their decision costs, e.g. from delays and mistakes. The problem of how to learn the thresholds has been neglected in the literature, and is non-trivial because these sampled costs are stochastic. We show that the learning problem coincides with a continuous-armed bandit over the thresholds, by re-casting Wald's many-trial cost function for optimality of the drift-diffusion model/SPRT as a single-trial reward function. We then compare two learning algorithms: (i) a standard learning rule derived from Williams' REINFORCE algorithm for neural networks; (ii) Bayesian optimization that models the reward function with a Gaussian Process and acquires samples accordingly. Bayesian optimization converges in fewer trials (~10^2) than REINFORCE (~10^3) but is slower computationally with greater variance. Open questions remain on improving the performance of the methods and the mechanism that animals use.

#64 Collaborative Recurrent Autoencoder: Recommend while Learning to Fill in the Blanks

Hao Wang (HKUST)
Xingjian SHI
Dit-Yan Yeung

Hybrid methods that utilize both content and rating information are commonly used in many recommender systems. However, most of them use either handcrafted features or the bag-of-words representation as a surrogate for the content information but they are neither effective nor natural enough. To address this problem,

we develop a collaborative recurrent autoencoder (CRAE) which is a denoising recurrent autoencoder (DRAE) that models the generation of content sequences in the collaborative filtering (CF) setting. The model generalizes recent advances in recurrent deep learning from i.i.d. input to non-i.i.d. (CF-based) input and provides a new denoising scheme along with a novel learnable pooling scheme for the recurrent autoencoder. To do this, we first develop a hierarchical Bayesian model for the DRAE and then generalize it to the CF setting. The synergy between denoising and CF enables CRAE to make accurate recommendations while learning to fill in the blanks in sequences. Experiments on real-world datasets from different domains (CiteULike and Netflix) show that, by jointly modeling the order-aware generation of sequences for the content information and performing CF for the ratings, CRAE is able to significantly outperform the state of the art on both the recommendation task based on ratings and the sequence generation task based on content information.

#65 Finding significant combinations of features in the presence of categorical covariates

Laetitia Papaxanthos (ETH Zurich)
Felipe Llinares-Lopez (ETH Zurich)
Dean Bodenham (ETH Zurich)
Karsten Borgwardt

In high-dimensional settings, where the number of features p is typically much larger than the number of samples n, methods which can systematically examine arbitrary combinations of features, a potentially huge $2p$-dimensional space, have recently begun to be explored. However, none of the current methods is able to assess the association between feature combinations and a target variable while conditioning on a covariate, in order to correct for potential confounding effects. We propose FACS, a significant descriminative itemset mining algorithm which conditions on categorical covariates and only scales as $\mathrm{O}(\mathrm{k}\ \log\ \mathrm{k})$, where k is the number of states of the categorical covariate. Based on the Cochran-Mantel, FACS demonstrates superior speed and statistical power on simulated and real-world datasets compared to the state of the art, opening the door to numerous applications in biomedicine.

#66 Synthesizing the preferred inputs for neurons in neural networks via deep generator networks

Anh Nguyen (Univ. of Wyoming)
Alexey Dosovitskiy (Univ. of Freiburg)
Jason Yosinski (Cornell)
Thomas Brox (Univ. of Freiburg)
Jeff Clune

Deep neural networks (DNNs) have demonstrated state-of-the-art results on many pattern recognition tasks, especially vision classification problems. Understanding the inner workings of such computational brains is both fascinating basic science that is interesting in its own right---similar to why we study the human brain---and will enable researchers to further improve DNNs. One path to understanding how a neural network functions internally is to study what each of its neurons has learned to detect. One such method is called activation maximization, which synthesizes an input (e.g. an image) that highly activates a neuron. Here we dramatically improve the qualitative state of the art of activation maximization by harnessing a powerful, learned prior: a deep generator network. The algorithm (1) generates qualitatively state-of-the-art synthetic images that look almost real, (2) reveals the features learned by each neuron in an interpretable way, (3) generalizes well to new datasets and somewhat well to different network architectures without requiring the prior to be relearned, and (4) can be considered as a high-quality generative method (in this case, by generating novel, creative, interesting, recognizable images).

#67 Learning Infinite RBMs with Frank-Wolfe

Wei Ping (UC Irvine)
Qiang Liu
Alexander Ihler

In this work, we propose an infinite restricted Boltzmann machine~(RBM), whose maximum likelihood estimation~(MLE) corresponds to a constrained convex optimization. We consider the Frank-Wolfe algorithm to solve the program, which provides a sparse solution that can be interpreted as inserting a hidden unit at each iteration, so that the optimization process takes the form of a sequence of finite models of increasing complexity. As a side benefit, this can be used to easily and efficiently identify an appropriate number of hidden units during the optimization. The resulting model can also be used as an initialization for typical state-of-the-art RBM training algorithms such as contrastive divergence, leading to models with consistently higher test likelihood than random initialization.

#68 Sorting out typicality with the inverse moment matrix SOS polynomial

Edouard Pauwels
Jean B Lasserre (LAAS-CNRS)

We study a surprising phenomenon related to the representation of a cloud of data points using polynomials. We start with the previously unnoticed empirical observation that, given a collection (a cloud) of data points, the sublevel sets of a certain distinguished polynomial capture the shape of the cloud very accurately. This distinguished polynomial is a sum-of-squares (SOS) derived in a simple manner from the inverse of the empirical moment matrix. In fact, this SOS polynomial is directly related to orthogonal polynomials and the Christoffel function. This allows to generalize and interpret extremality properties of orthogonal polynomials and to provide a mathematical rationale for the observed phenomenon. Among diverse potential applications, we illustrate the relevance of our results on a network intrusion detection task for which we obtain performances similar to existing dedicated methods reported in the literature.

#69 Improving PAC Exploration Using the Median of Means

Jason Pazis (MIT)
Ron E Parr
Jonathan P How (MIT)

We present the first application of the median of means in a PAC-optimal exploration algorithm for MDPs. Our use of the median of means allows us to significantly reduce the dependence of our bounds on the range of values that the value function can take, while introducing a dependence on the (potentially much smaller) variance of the Bellman operator. In addition, our algorithm is the first algorithm with PAC bounds that can be applied to MDPs with unbounded rewards.

#70 Reconstructing Parameters of Spreading Models from Partial Observations
Andrey Lokhov (Los Alamos National Laboratory)

Spreading processes are often modelled as a stochastic dynamics occurring on top of a given network with edge weights corresponding to the transmission probabilities. Knowledge of veracious transmission probabilities is essential for prediction, optimization, and control of diffusion dynamics. Unfortunately, in most cases the transmission rates are unknown and need to be reconstructed from the spreading data. Moreover, in realistic settings it is impossible to monitor the state of each node at every time, and thus the data is highly incomplete. We introduce an efficient dynamic message-passing algorithm, which is able to reconstruct parameters of the spreading model given only partial information on the activation times of nodes in the network. The method is generalizable to a large class of dynamic models, as well to the case of temporal graphs.

#71 Dynamic Filter Networks
Xu Jia (KU Leuven)
Bert De Brabandere
Tinne Tuytelaars (KU Leuven)
Luc V Gool (ETH Zürich)

In a traditional convolutional layer, the learned filters stay fixed after training. In contrast, we introduce a new framework, the Dynamic Filter Network, where filters are generated dynamically conditioned on an input. We show that this architecture is a powerful one, with increased flexibility thanks to its adaptive nature, yet without an excessive increase in the number of model parameters. A wide variety of filtering operation can be learned this way, including local spatial transformations, but also others like selective (de)blurring or adaptive feature extraction. Moreover, multiple such layers can be combined, e.g. in a recurrent architecture. We demonstrate the effectiveness of the dynamic filter network on the tasks of video and stereo prediction, and reach state-of-the-art performance on the moving MNIST dataset with a much smaller model. By visualizing the learned filters, we illustrate that the network has picked up flow information by only looking at unlabelled training data. This suggests that the network can be used to pretrain networks for various supervised tasks in an unsupervised way, like optical flow and depth estimation.

#72 Long-Term Trajectory Planning Using Hierarchical Memory Networks
Stephan Zheng (Caltech)
Yisong Yue
Patrick Lucey (Stats)

We study the problem of learning to plan spatiotemporal trajectories over long time horizons using expert demonstrations. For instance, in sports, agents often choose action sequences with long-term goals in mind, such as achieving a certain strategic position. Conventional policy learning approaches, such as those based on Markov decision processes, generally fail at learning cohesive long-term behavior in such high-dimensional state spaces, and are only effective when myopic planning leads to the desired behavior. The key difficulty is that conventional approaches are ``shallow" planners that only learn a single state-action policy. We instead propose to learn a hierarchical planner that automatically reasons about both long-term and short-term goals, which we instantiate as a hierarchical deep memory network. We showcase our approach in a case study on learning to imitate demonstrated basketball trajectories, and show that it generates significantly more realistic trajectories compared to non-hierarchical baselines as judged by professional sports analysts.

#73 Cooperative Inverse Reinforcement Learning
Dylan Hadfield-Menell (UC Berkeley)
Stuart J Russell (UC Berkeley)
Pieter Abbeel (OpenAI / UC Berkeley / Gradescope)
Anca Dragan

For an autonomous system to be helpful to humans and to pose no unwarranted risks, it needs to align its values with those of the humans in its environment in such a way that its actions contribute to the maximization of value for the humans. We propose a formal definition of the value alignment problem as cooperative inverse reinforcement learning (CIRL). A CIRL problem is a cooperative, partial- information game with two agents, human and robot; both are rewarded according to the human's reward function, but the robot does not initially know what this is. In contrast to classical IRL, where the human is assumed to act optimally in isolation, optimal CIRL solutions produce behaviors such as active teaching, active learning, and communicative actions that are more effective in achieving value alignment. We show that computing optimal joint policies in CIRL games can be reduced to solving a POMDP, prove that optimality in isolation is suboptimal in CIRL, and derive an approximate CIRL algorithm.

#74 Encode, Review, and Decode: Reviewer Module for Caption Generation
Zhilin Yang (Carnegie Mellon Univ.)
Ye Yuan (Carnegie Mellon Univ.)
Yuexin Wu (Carnegie Mellon Univ.)
William W Cohen (Carnegie Mellon Univ.)
Russ Salakhutdinov (Univ. of Toronto)

We propose a novel module, the reviewer module, to improve the encoder-decoder learning framework. The reviewer module is generic, and can be plugged into an existing encoder-decoder model. The reviewer module performs a number of review steps with attention mechanism on the encoder hidden states, and outputs a fact vector after each review step; the fact vectors are used as the input of the attention mechanism in the decoder. We show that the conventional encoder-decoders are a special case of our framework. Empirically, we show that our framework can improve over state-of-the-art encoder-decoder systems on the tasks of image captioning and source code captioning.

#75 Gradient-based Sampling: An Adaptive Importance Sampling for Least-squares
Rong Zhu (Chinese Academy of Sciences)

In modern data analysis, random sampling is an efficient and widely-used strategy to overcome the computational difficulties brought by large sample size. In previous studies, researchers conducted random sampling which is according to the input data but independent on the response variable, however the response variable may also be informative for sampling. In this paper we propose an adaptive sampling called the gradient-based sampling which is dependent on both the input data and the output for fast solving of least-square (LS) problems. We draw the data points by random sampling from the full data according to their gradient values. This sampling is computationally saving, since the running time of computing the sampling probabilities is reduced to $\mathrm{O}(\mathrm{nd})$ where n is the full sample size and d is the dimension of the input. Theoretically, we establish an error bound analysis of the general importance sampling with respect to LS solution from full data. The result establishes an improved performance of the use of our gradient-based sampling. Synthetic and real data sets are used to empirically argue that the gradient-based sampling has obvious advantage over existing sampling methods from two aspects of statistical efficiency and computational saving.

#76 Robust k-means: a Theoretical Revisit
ALEX GEORGOGIANNIS (TECHNICAL Univ. OF CRETE)

Over the last years, many variations of the quadratic k-means clustering procedure have been proposed, all aiming to robustify the performance of the algorithm in the presence of outliers. In general terms, two main approaches have been developed: one based on penalized regularization methods, and one based on trimming functions. In this article, we give a theoretical analysis of the robustness and consistency properties of a variant of the classical quadratic k-means algorithm, the robust k-means, which borrows ideas from outlier detection in the regression setting. We show that two outliers in a dataset are enough to breakdown this clustering procedure. However, if we focus on ``well-structured'' datasets then robust k-means can recover the cluster structure in spite of the outliers. Finally, we show that with slight modifications, the most general non-asymptotic results for consistency of quadratic k-means still remain valid for this robust variant.

#77 Boosting with Abstention
Corinna Cortes
Giulia DeSalvo
Mehryar Mohri

We present a new boosting algorithm for binary classification with abstention which, at each round, simultaneously selects a pair of functions, a base predictor and a base abstention function. We introduce convex surrogate losses that we prove to be calibrated with respect to the Bayes solution and derive theoretical learning guarantees in terms of the Rademacher complexity of the resulting ensemble functions. From these data-dependent bounds, we derive a regularized boosting with abstention (BA) algorithm, that is based on projected coordinate descent applied to our convex surrogate losses. We provide its convergence guarantees along with a linear-time weak-learning algorithm for learning abstention stumps. We also report the results of several experiments comparing BA to two different confidence-based algorithms, which suggest that BA provides a significant improvement in practice.

#78 Estimating the class prior and posterior from noisy positives and unlabeled data
Shantanu J Jain (Indiana Univ.)
Martha White
Pedja Radivojac

We develop a classification algorithm for estimating posterior distributions from positive-unlabeled data, that is robust to noise in the positive labels and effective for high-dimensional data. In recent years, several algorithms have been proposed to learn from positive-unlabeled data; however, many of these contributions remain theoretical, performing poorly on real high-dimensional data that is typically contaminated with noise. We build on this previous work to develop two practical classification algorithms that explicitly model the noise in the positive labels and utilize univariate transforms built on discriminative classifiers. We prove that these univariate transforms preserve the class prior, enabling estimation in the univariate space and avoiding kernel density estimation for high-dimensional data. The theoretical development and both parametric and nonparametric algorithms proposed here constitutes an important step towards wide-spread use of robust classification algorithms for positive-unlabeled data.

#79 Bootstrap Model Aggregation for Distributed Statistical Learning
JUN HAN (Dartmouth College)
Qiang Liu

In distributed, or privacy-preserving learning, we are often given a set of probabilistic models estimated from different local repositories, and asked to combine them into a single model that gives efficient statistical estimation. A simple method is to linearly average the parameters of the local models, which, however, tends to be degenerate or not applicable on non-convex models, or models with different parameter dimensions. One more practical strategy is to generate bootstrap samples from the local models, and then learn a joint model based on the combined bootstrap set. Unfortunately, the bootstrap procedure introduces additional noise and can significantly deteriorate the performance. In this work, we propose two variance reduction methods to correct the bootstrap noise, including a weighted M-estimator that is both statistically efficient and practically powerful. Both theoretical and empirical analysis is provided to demonstrate our methods.

#80 Noise-Tolerant Life-Long Matrix Completion via Adaptive Sampling
Maria-Florina Balcan
Hongyang Zhang (CMU)

We study the problem of recovering an incomplete m\times n matrix of rank r with columns arriving online over time. This is known as the problem of life-long matrix completion, and is widely applied to recommendation system, computer vision, system identification, etc. The challenge is to design provable algorithms tolerant to a large amount of noise, with small sample complexity. In this work, we give algorithms achieving strong guarantee under two realistic noise models. In bounded deterministic noise, an adversary can add any bounded yet unstructured noise to each column. For this problem, we present an algorithm that returns a matrix of a small error, with sample complexity almost as small as the best prior results in the noiseless case. For sparse random noise, where the corrupted columns are sparse and drawn randomly, we give an algorithm that exactly recovers an μ0-incoherent matrix by probability at least 1−δ with sample complexity as small as Ω(μ0rnlog(r/δ)). This result advances the state-of-the-art work and matches the lower bound in a worst case. We also study the scenario where the hidden matrix lies on a mixture of subspaces and show that the sample complexity can be even smaller. Experiments verify our theories.

#81 FPNN: Field Probing Neural Networks for 3D Data

Yangyan Li (Stanford Univ.)
pirk Pirk (Stanford Univ.)
Hao Su (Stanford Univ.)
Charles R Qi (Stanford Univ.)
Leonidas J Guibas (Stanford Univ.)

Building discriminative representations for 3D data has been an important task in computer graphics and computer vision research. Convolutional Neural Networks (CNNs) have shown to operate on 2D images with great success for a variety of tasks. Lifting convolution operators to 3D (3DCNNs) seems like a plausible and promising next step. Unfortunately, the computational complexity of 3D CNNs grows cubically with respect to voxel resolution. Moreover, since most 3D geometry representations are boundary based, occupied regions do not increase proportionately with the size of the discretization, resulting in wasted computation. In this work, we represent 3D spaces as volumetric fields, and propose a novel design that employs field probing filters to efficiently extract features from them. Each field probing filter is a set of probing points --- sensors that perceive the space. Our learning algorithm optimizes not only the weights associated with the probing points, but also their locations, which deforms the shape of the probing filters and adaptively distributes them in 3D space. The optimized probing points sense the 3D space ``intelligently", rather than operating blindly over the entire domain. We show that field probing is significantly more efficient than 3DCNNs, while providing state-of-the-art performance, on classification tasks for 3D object recognition benchmark datasets.

#82 Causal meets Submodular: Subset Selection with Directed Information

Yuxun Zhou (UC Berkeley)
Costas J Spanos

We study causal subset selection with Directed Information as the causality measure. Two typical tasks, source detection and causal covariate selection, are correspondingly formulated into cardinality constrained directed information maximizations. To attack the NP-hard problems with greedy heuristics, we show that the first problem is submodular while not necessarily monotonic. And the second one is ``nearly" submodular. To substantiate the idea of approximate submodularity, we introduce a novel quantity, namely submodularity index (SmI), for general set functions. Moreover, we show that based on SmI, random greedy algorithm has performance guarantee for the maximization of possibly non-monotonic and non-submodular functions, justifying its usage for a much broader class of problems. We evaluate the theoretical results with several case studies, and also illustrate the application of the subset selection to causal structure learning.

#83 Improving Variational Autoencoders with Inverse Autoregressive Flow

Diederik P Kingma
Tim Salimans
Rafal Jozefowicz (OpenAI)
Xi Chen (UC Berkeley and OpenAI)
Ilya Sutskever
Max Welling

We propose a simple and scalable method for improving the flexibility of variational inference through a transformation with autoregressive neural networks. Autoregressive neural networks, such as RNNs and the PixelCNN, are very powerful models; however, ancestral sampling in such networks is a sequential operation, therefore unappealing for direct use as approximate posteriors in variational inference on parallel hardware such as GPUs. We find that by inverting autoregressive neural networks we can obtain equally powerful data transformations that can often be computed in parallel. We show that such data transformations, inverse autoregressive flows (IAF), can be used to transform a simple distribution over the latent variables into a much more flexible distribution, while still allowing us to compute the resulting variables' probability density function. The method is simple to implement, can be made arbitrarily flexible, and (in contrast with previous work) is naturally applicable to latent variables that are organized in multidimensional tensors, such as 2D grids or time series. The method is applied to a novel deep architecture of variational auto-encoders. In experiments we demonstrate that autoregressive flow leads to significant performance gains when applied to variational autoencoders for natural images.

#84 Adaptive Smoothed Online Multi-Task Learning

Keerthiram Murugesan (Carnegie Mellon Univ.)
Hanxiao Liu (Carnegie Mellon Univ.)
Jaime Carbonell (CMU)
Yiming Yang (CMU)

This paper addresses the challenge of jointly learning both the per-task model parameters and the inter-task relationships in a multi-task online learning setting. The proposed algorithm features probabilistic interpretation, efficient updating rules and flexible modulation on whether learners focus on their specific tasks or on jointly address all tasks. The paper also proves a sub-linear regret bound as compared to the best linear predictor in hindsight. Experiments over three multitask learning benchmark datasets show advantageous performance of the proposed approach over several state-of-the-art online multi-task learning baselines.

#85 The Limits of Learning with Missing Data

Brian Bullins (Princeton Univ.)
Elad Hazan
Tomer Koren (Technion---Israel Inst. of Technology)

We study regression and classification in a setting where the learning algorithm is allowed to access only a limited number of attributes per example, known as the limited attribute observation model. In this well-studied model, we provide the first lower bounds giving a limit on the precision attainable by any algorithm for several variants of regression, notably linear regression with the absolute loss and the squared loss, as well as for classification with the hinge loss. We complement these lower bounds with a general purpose algorithm that gives an upper bound on the achievable precision limit in the setting of learning with missing data.

#86 Safe Exploration in Finite Markov Decision Processes with Gaussian Processes

Matteo Turchetta (ETH Zurich)
Felix Berkenkamp (ETH Zurich)
Andreas Krause

In classical reinforcement learning, when exploring an environment, agents accept arbitrary short term loss for long term gain. This is infeasible for safety critical applications, such as robotics, where even a single unsafe action may cause system failure. In this paper, we address the problem of safely exploring finite Markov decision processes (MDP). We define safety in terms of an, a priori unknown, safety constraint that depends on states and actions. We aim to explore the MDP under this constraint, assuming that the unknown

function satisfies regularity conditions expressed via a Gaussian process prior. We develop a novel algorithm for this task and prove that it is able to completely explore the safely reachable part of the MDP without violating the safety constraint. To achieve this, it cautiously explores safe states and actions in order to gain statistical confidence about the safety of unvisited state-action pairs from noisy observations collected while navigating the environment. Moreover, the algorithm explicitly considers reachability when exploring the MDP, ensuring that it does not get stuck in any state with no safe way out. We demonstrate our method on digital terrain models for the task of exploring an unknown map with a rover.

#87 Sparse Support Recovery with Non-smooth Loss Functions

Kévin Degraux (Université catholique de Louva)
Gabriel Peyré
Jalal Fadili
Laurent Jacques (Université catholique de Louvain)

In this paper, we study the support recovery guarantees of underdetermined sparse regression using the $l1$-norm as a regularizer and a non-smooth loss function for data fidelity. More precisely, we focus in detail on the cases of $l1$ and $\ell\infty$ losses, and contrast them with the usual ℓ_2 loss. While these losses are routinely used to account for either sparse ($\ell1$ loss) or uniform ($\ell\infty$ loss) noise models, a theoretical analysis of their performance is still lacking. In this article, we extend the existing theory from the smooth $\ell_2$case to these non-smooth cases. We derive a sharp condition which ensures that the support of the vector to recover is stable to small additive noise in the observations, as long as the loss constraint size is tuned proportionally to the noise level. A distinctive feature of our theory is that it also explains what happens when the support is unstable. While the support is not anymore stable, we identify an ``extended support'' (which corresponds to the equicorrelation set), and show that this extended support is stable to small additive noise. To exemplify the usefulness of our theory, we give a detailed numerical analysis of the support stability/instability of compressed sensing recovery with these different losses. This highlights different parameter regimes, ranging from total support stability to progressively increasing support instability.

#88 Crowdsourced Clustering: Querying Edges vs Triangles

Ramya Korlakai Vinayak (Caltech)
Babak Hassibi (Caltech)

We consider the task of clustering items using answers from non-expert crowd workers. In such cases, the workers are often not able to label the items directly, however, it is reasonable to assume that they can compare items and judge whether they are similar or not. An important question is what queries to make, and we compare two types: random edge queries, where a pair of items is revealed, and random triangles, where a triple is. Since it is far too expensive to query all possible edges and/or triangles, we need to work with partial observations subject to a fixed query budget constraint. When a generative model for the data is available (and we consider a few of these) we determine the cost of a query by its entropy; when such models do not exist we use the average response time per query of the workers as a surrogate for the cost. In addition to theoretical justification, through several simulations and experiments on two real data sets on Amazon Mechanical Turk, we empirically demonstrate that, for a fixed budget, triangle queries uniformly outperform edge queries. Even though, in contrast to edge queries, triangle queries reveal dependent edges, they provide more reliable edges and, for a fixed budget, many more of them. We also provide a sufficient condition on the number of observations, edge densities inside and outside the clusters and the minimum cluster size required for the exact recovery of the true adjacency matrix via triangle queries using a convex optimization-based clustering algorithm.

#89 Dual Decomposed Learning with Factorwise Oracle for Structural SVM of Large Output Domain

Ian Yen (Univ. of Texas at Austin)
Xiangru Huang (Univ. of Texas at Austin)
Kai Zhong (Univ. of Texas at Austin)
Ruohan Zhang (Univ. of Texas at Austin)
Pradeep K Ravikumar
Inderjit S Dhillon

Many applications of machine learning involve structured output with large domain, where learning of structured predictor is prohibitive due to repetitive calls to expensive inference oracle. In this work, we show that, by decomposing training of Structural Support Vector Machine (SVM) into a series of multiclass SVM problems connected through messages, one can replace expensive structured oracle with Factorwise Maximization Oracle (FMO) that allows efficient implementation of complexity sublinear to the factor domain. A Greedy Direction Method of Multiplier (GDMM) algorithm is proposed to exploit sparsity of messages which guarantees ϵ sub-optimality after O(1/ϵ) passes of FMO calls. We conduct experiments on chain-structured problems and fully-connected problems of large output domains. The proposed approach is orders-of-magnitude faster than the state-of-the-art training algorithms for Structural SVM.

#90 Sampling for Bayesian Program Learning

Kevin Ellis (MIT)
Armando Solar-Lezama (MIT)
Josh Tenenbaum

Towards learning programs from data, we introduce the problem of sampling programs from posterior distributions conditioned on that data. Within this setting, we propose an algorithm that uses a symbolic solver to efficiently sample programs. The proposal combines constraint-based program synthesis with sampling via random parity constraints. We give theoretical guarantees on how well the samples approximate the true posterior, and have empirical results showing the algorithm is efficient in practice, evaluating our approach on 22 program learning problems in the domains of text editing and computer-aided programming.

#91 Multiple-Play Bandits in the Position-Based Model

Paul Lagrée (Université Paris Sud)
Claire Vernade (Université Paris Saclay)
Olivier Cappe

Sequentially learning to place items in multi-position displays or lists is a task that can be cast into the multiple-play semi-bandit setting. However, a major concern in this context is when the system cannot decide whether the user feedback for each item is actually exploitable. Indeed, much of the content may have been simply ignored by the user. The present work proposes to exploit available information regarding the display position bias under the so-called Position-based click model (PBM). We first discuss how this model differs from the Cascade model and its variants considered in several recent works on multiple-play bandits. We then provide a novel regret lower bound for this model as well as computationally efficient algorithms that display good empirical and theoretical performance.

#92 Image Restoration Using Very Deep Convolutional Encoder-Decoder Networks with Symmetric Skip Connections

Xiaojiao Mao (Nanjing Univ.)
Chunhua Shen
Yu-Bin Yang

In this paper, we propose a very deep fully convolutional encoding-decoding framework for image restoration such as denoising and super-resolution. The network is composed of multiple layers of convolution and de-convolution operators, learning end-to-end mappings from corrupted images to the original ones. The convolutional layers act as the feature extractor, which capture the abstraction of image contents while eliminating noises/corruptions. De-convolutional layers are then used to recover the image details. We propose to symmetrically link convolutional and de convolutional layers with skip-layer connections, with which the training converges much faster and attains a higher quality local optimum. First, The skip connections allow the signal to be back-propagated to bottom layers directly, and thus tackles the problem of gradient vanishing, making training deep networks easier and achieving restoration performance gains consequently. Second, these skip connections pass image details from convolutional layers to de-convolutional layers, which is beneficial in recovering the original image. Significantly, with the large capacity, we can handle different levels of noises using a single model. Experimental results show that our network achieves better performance than all previously reported state-of-the-art methods.

#93 Optimistic Bandit Convex Optimization

Scott Yang (New York Univ.)
Mehryar Mohri

We introduce the general and powerful scheme of predicting information re-use in optimization algorithms. This allows us to devise a computationally efficient algorithm for bandit convex optimization with new state-of-the-art guarantees for both Lipschitz loss functions and loss functions with Lipschitz gradients. This is the first algorithm admitting both a polynomial time complexity and a regret that is polynomial in the dimension of the action space that improves upon the original regret bound for Lipschitz loss functions, achieving a regret of $\mathrm{O}\tilde{}(\mathrm{T}^{11/16}\mathrm{d}^{3/8})$. Our algorithm further improves upon the best existing polynomial-in-dimension bound (both computationally and in terms of regret) for loss functions with Lipschitz gradients, achieving a regret of $\mathrm{O}\tilde{}(\mathrm{T}8/13\mathrm{d}5/3)$.

#94 Computing and maximizing influence in linear threshold and triggering models

Justin T Khim (Univ. of Pennsylvania)
Varun Jog
Po-Ling Loh (Berkeley)

We establish upper and lower bounds for the influence of a set of nodes in certain types of contagion models. We derive two sets of bounds, the first designed for linear threshold models, and the second more broadly applicable to a general class of triggering models, which subsumes the popular independent cascade models, as well. We quantify the gap between our upper and lower bounds in the case of the linear threshold model and illustrate the gains of our upper bounds for independent cascade models in relation to existing results. Importantly, our lower bounds are monotonic and submodular, implying that a greedy algorithm for influence maximization is guaranteed to produce a maximizer within a $(1-1/\mathrm{e})$-factor of the truth. Although the problem of exact influence computation is NP-hard in general, our bounds may be evaluated efficiently. This leads to an attractive, highly scalable algorithm for influence maximization with rigorous theoretical guarantees.

#95 Clustering with Bregman Divergences: an Asymptotic Analysis

Chaoyue Liu (The Ohio State Univ.)
Mikhail Belkin

Clustering, in particular k-means clustering, is a central topic in data analysis. Clustering with Bregman divergences is a recently proposed generalization of k-means clustering which has already been widely used in applications. In this paper we analyze theoretical properties of Bregman clustering when the number of the clusters k is large. We establish quantization rates and describe the limiting distribution of the centers as k\to \infty, extending well-known results for k-means clustering.

#96 Community Detection on Evolving Graphs

LEONARDI Leonardi (Sapienza Univ. of Rome)
Aris Anagnostopoulos (Sapienza Univ. of Rome)
Jakub Łącki (Sapienza Univ. of Rome)
Silvio Lattanzi (Google)
Mohammad Mahdian (Google Research)

Clustering is a fundamental step in many information retrieval and data mining applications. Detecting clusters in graphs is also a key tool for finding the community structure in social and behavioral networks. In many of these applications, the input graph evolves over time in a continuous and decentralized manner, and to maintain a good clustering, the algorithm needs to repeatedly probe the graph. Furthermore, there are often limitations on the frequency of such probes, either imposed explicitly by the online platform (e.g., in the case of crawling proprietary social networks like twitter) or implicitly due to resource limitations (e.g., in the case of crawling the web). In this paper, we study a model of clustering on evolving graphs that captures this aspect of the problem. Our model is based on the classical stochastic block model, which has been used to rigorously assess the quality of various static clustering methods. In our model, the algorithm is supposed to reconstruct the planted clustering, given the ability to query for small pieces of local information about the graph, at a limited rate. We design and analyze clustering algorithms that work in this model, and show asymptotically tight upper and lower bounds on their accuracy.

#97 Dueling Bandits: Beyond Condorcet Winners to General Tournament Solutions

Siddartha Y. Ramamohan (Indian Institute of Science)
Arun Rajkumar
Shivani Agarwal (Radcliffe Institute)

Recent work on deriving O(logT) anytime regret bounds for stochastic dueling bandit problems has considered mostly Condorcet winners, which do not always exist, and more recently, winners defined by the Copeland set, which do always exist. In this work, we consider a broad notion of winners defined by tournament solutions in social choice theory, which include the Copeland set as a special case but also include several other notions of winners such as the top cycle, uncovered set, and Banks set, and which, like the Copeland set, always exist. We develop a family of UCB-style dueling bandit algorithms for such general tournament solutions, and show O(logT) anytime regret bounds for them. Experiments confirm the ability of our algorithms to achieve low regret relative to the target winning set of interest.

#98 Learning a Metric Embedding for Face Recognition using the Multibatch Method

Oren Tadmor (OrCam)
Tal Rosenwein (Orcam)
Shai Shalev-Shwartz (OrCam)
Yonatan Wexler (OrCam)
Amnon Shashua (OrCam)

This work is motivated by the engineering task of achieving a near state-of-the-art face recognition on a minimal computing budget running on an embedded system. Our main technical contribution centers around a novel training method, called Multibatch, for similarity learning, i.e., for the task of generating an invariant ``face signature'' through training pairs of ``same'' and ``not-same'' face images. The Multibatch method first generates signatures for a mini-batch of k face images and then constructs an unbiased estimate of the full gradient by relying on all k^2–k pairs from the mini-batch. We prove that the variance of the Multibatch estimator is bounded by $O(1/k^2)$, under some mild conditions. In contrast, the standard gradient estimator that relies on random k/2 pairs has a variance of order 1/k. The smaller variance of the Multibatch estimator significantly speeds up the convergence rate of stochastic gradient descent. Using the Multibatch method we train a deep convolutional neural network that achieves an accuracy of 98.2\% on the LFW benchmark, while its prediction runtime takes only 30msec on a single ARM Cortex A9 core. Furthermore, the entire training process took only 12 hours on a single Titan X GPU.

#99 Convergence guarantees for kernel-based quadrature rules in misspecified settings

Motonobu Kanagawa
Bharath K. Sriperumbudur
Kenji Fukumizu

Kernel-based quadrature rules are powerful tools for numerical integration which yield convergence rates much faster than usual Mote Carlo methods. These rules are constructed based on the assumption that the integrand has a certain degree of smoothness, and this assumption is expressed as that the integrand belongs to a certain reproducing kernel Hilbert space (RKHS). However, in practice such an assumption can be violated, and no general theory has been established for the convergence in such misspecified cases. In this paper, we prove that kernel quadrature rules can be consistent even when an integrand does not belong to an assumed RKHS, i.e., when the integrand is less smooth than assumed. We derive convergence rates that depend on the (unknown) smoothness of the integrand, where the degree of smoothness is expressed via powers of RKHSs or via Sobolev spaces.

#100 Stochastic Variational Deep Kernel Learning

Andrew G Wilson (Carnegie Mellon Univ.)
Zhiting Hu (Carnegie Mellon Univ.)
Russ Salakhutdinov (Univ. of Toronto)
Eric P Xing (Carnegie Mellon Univ.)

Deep kernel learning combines the non-parametric flexibility of kernel methods with the inductive biases of deep learning architectures. We propose a novel deep kernel learning model and stochastic variational inference procedure which generalizes deep kernel learning approaches to enable classification, multi-task learning, additive covariance structures, and stochastic gradient training. Specifically, we apply additive base kernels to subsets of output features from deep neural architectures, and jointly learn the parameters of the base kernels and deep network through a Gaussian process marginal likelihood objective. Within this framework, we derive an efficient form of stochastic variational inference which leverages local kernel interpolation, inducing points, and structure exploiting algebra. We show improved performance over stand alone deep networks, SVMs, and state of the art scalable Gaussian processes on several classification benchmarks, including an airline delay dataset containing 6 million training points.

#101 Deep Submodular Functions

Brian W Dolhansky (Univ. of Washington)
Jeff A Bilmes (Univ. of Washington)

We propose and study a new class of submodular functions called deep submodular functions (DSFs). We define DSFs and situate them within the broader context of classes of submodular functions, in relationship both to various matroid ranks, and sums of concave composed with modular functions (SCMs). Notably, we find that DSFs constitute a strictly broader class than SCMs, thus motivating their use, but that they do not comprise all submodular functions. Interestingly, some DSFs can be seen as special cases of certain deep neural networks (DNNs), hence the name. Finally, we show how to learn DSFs in a max-margin framework, and successfully apply this to both synthetic and real-world data instances.

#102 Scaled Least Squares Estimator for GLMs in Large-Scale Problems

Murat A Erdogdu (Stanford Univ.)
Lee H Dicker (Rutgers Univ. and Amazon)
Mohsen Bayati

We study the problem of efficiently estimating the coefficients of generalized linear models (GLMs) in the large-scale setting where the number of observations n is much larger than the number of predictors p, i.e. $n \gg p \gg 1$. We show that in GLMs with random (not necessarily Gaussian) design, the GLM coefficients are approximately proportional to the corresponding ordinary least squares (OLS) coefficients. Using this relation, we design an algorithm that achieves the same accuracy as the maximum likelihood estimator (MLE) through iterations that attain up to a cubic convergence rate, and that are cheaper than any batch optimization algorithm by at least a factor of O(p). We provide theoretical guarantees for our algorithm, and analyze the convergence behavior in terms of data dimensions. Finally, we demonstrate the performance of our algorithm through extensive numerical studies on large-scale datasets, and show that it achieves the highest performance compared to several other widely used optimization methods for computing the MLEs in GLMs.

#103 Matrix Completion and Clustering in Self-Expressive Models

Ehsan Elhamifar

We propose efficient algorithms for simultaneous clustering and completion of incomplete high-dimensional data that lie in a union of low-dimensional subspaces. We cast the problem as finding a completion of the data matrix so that each point can be reconstructed as a linear or affine combination of a few data points. Since the problem is NP-hard, we propose a lifting framework and reformulate the problem as a group-sparse recovery of each incomplete data point in a dictionary built using incomplete data, subject to rank-one constraints. To solve the problem efficiently, we propose a rank pursuit algorithm and a convex relaxation. The solution of our algorithms recover missing entries and provides a similarity matrix for clustering. Our algorithms can deal with both low-rank and high-rank

matrices, does not suffer from initialization, does not need to know dimensions of subspaces and can work with a small number of data points. By extensive experiments on synthetic data and real problems of video motion segmentation and completion of motion capture data, we show that when the data matrix is low-rank, our algorithm performs on par with or better than low-rank matrix completion methods, while for high-rank data matrices, our method significantly outperforms existing algorithms.

#104 Stochastic Three-Composite Convex Minimization

Alp Yurtsever (EPFL)
Bang Cong Vu
Volkan Cevher

We propose a stochastic optimization method for the convex minimization of the sum of three convex functions, one of which has Lipschitz continuous gradient as well as restricted strong convexity. Our approach is most suitable in the setting where it is computationally advantageous to process smooth term in the decomposition with it stochastic gradient estimate and the other two functions separately with their proximal operators, such as doubly regularized empirical risk minimization problems. We prove the convergence characterization of the proposed algorithm in expectation under the standard assumptions for the stochastic gradient estimate of the smooth term. Our method operates in the primal space and can be considered as a stochastic extension of the three-operator splitting method. Finally, numerical evidence supports the effectiveness of our method in real-word problems.

#105 Tree-Structured Reinforcement Learning for Sequential Object Localization

Zequn Jie (National Univ of Singapore)
Xiaodan Liang (Sun Yat-sen Univ.)
Jiashi Feng (National Univ. of Singapo)
Xiaojie Jin (NUS)
Wen Lu (National Univ of Singapore)
Shuicheng Yan

Existing object proposal algorithms usually search for possible object regions over multiple locations and scales separately, which ignore the interdependency among different objects and deviate from the human perception procedure. To incorporate global interdependency between objects into object localization, we propose an effective Tree-structured Reinforcement Learning (Tree-RL) approach to sequentially search for objects by fully exploiting both the current observation and historical search paths. The Tree-RL approach learns multiple searching policies through maximizing the long-term reward that reflects localization accuracies over all the objects. Starting with taking the entire image as a proposal, the Tree-RL approach allows the agent to sequentially discover multiple objects via a tree-structured traversing scheme. Allowing multiple near-optimal policies, Tree-RL offers more diversity in search paths and is able to find multiple objects with a single feed-forward pass. Therefore, Tree-RL can better cover different objects with various scales which is quite appealing in the context of object proposal. Experiments on PASCAL VOC 2007 and 2012 validate the effectiveness of the Tree-RL, which can achieve comparable recalls with current object proposal algorithms via much fewer candidate windows as well as better detection mAP than Faster R-CNN (ResNet-101).

#106 The non-convex Burer-Monteiro approach works on smooth semidefinite programs

Nicolas Boumal
Vlad Voroninski (MIT)
Afonso Bandeira

Computational tasks arising in machine learning are often naturally framed as optimization problems. When these are not tractable, it is common to resort to convex relaxations instead. Semidefinite programs (SDP's) are particularly popular in that context. SDP's can be solved in polynomial time, but scalability can be an issue. Over a decade ago, Burer and Monteiro proposed to solve SDP's with few equality constraints by solving rank-restricted and non-convex versions of them instead. Some theory supports the empirical success of their approach on special SDP's, but it has remained unclear why local optimization methods seemed to converge to global optima so reliably on the non-convex formulation. In this paper, we consider a class of SDP's which includes applications such as max-cut, community detection in the stochastic block model, robust PCA, phase retrieval and synchronization of rotations. We show that the low-rank Burer-Monteiro formulation of SDP's in that class almost never has any spurious local optima.

#107 Neurons Equipped with Intrinsic Plasticity Learn Stimulus Intensity Statistics

Travis Monk (Univ. of Oldenburg)
Cristina Savin (IST Austria)
Jörg Lücke

Experience constantly shapes neural circuits through a variety of plasticity mechanisms. While the functional roles of some plasticity mechanisms are well-understood, it remains unclear how changes in neural excitability contribute to learning. Here, we develop a normative interpretation of intrinsic plasticity (IP) as a key component of unsupervised learning. We introduce a novel generative mixture model that accounts for the statistics of stimulus intensities, and we derive a neural circuit that learns the classes and intensities of its inputs. Our analytical results show that inference and learning for our generative model can be achieved in a neural circuit if feature-sensitive neurons are equipped with a specific form of IP. Numerical experiments verify our analytical derivations and show robust behavior for artificial and natural stimuli. Our results link IP to nontrivial input statistics, in particular the statistics of stimulus intensity for classes to which a neuron is sensitive. More generally, the model paves the way for a novel class of clustering and classification algorithms that are robust to gain variations.

#108 Greedy Feature Construction

Dino Oglic (Univ. of Bonn)
Thomas Gärtner (The Univ. of Nottingham)

We present an effective method for supervised feature construction. The main goal of the approach is to exploit the scalability of existing algorithms for training linear models while overcoming their low capacity on input features. To embed the data into a space with a high capacity set of linear hypotheses, we take a greedy approach and construct features by fitting residuals. We show that the empirical squared error residual fitting is consistent and provide a convergence rate for our constructive procedure. To show that the method can construct a high capacity feature space, we make a connection to shift-invariant reproducing kernel Hilbert spaces and show that it can approximate any bounded function from these spaces. The effectiveness of the approach is evaluated empirically by training a linear ridge regression model in the constructed feature space and the empirical evidence indicates the superior performance of our approach over the competing methods.

#109 Dynamic Mode Decomposition with Reproducing Kernels for Koopman Spectral Analysis

Yoshinobu Kawahara (Osaka Univ.)

A spectral analysis of the Koopman operator, which is an infinite dimensional linear operator on an observable, gives a (modal) description of the global behavior of a nonlinear dynamical system without any explicit prior knowledge of its governing equations. In this paper, we consider a spectral analysis of the Koopman operator in a reproducing kernel Hilbert space (RKHS). We propose a modal decomposition algorithm to perform the analysis using finite-length data sequences generated from a nonlinear system. The algorithm is in essence reduced to the calculation of a set of orthogonal bases for the Krylov matrix in RKHS and the eigendecomposition of the projection of the Koopman operator onto the subspace spanned by the bases. The algorithm returns a decomposition of the dynamics into a finite number of modes, and thus it can be thought of as a feature extraction procedure for a nonlinear dynamical system. Therefore, we further consider applications in machine learning using extracted features with the presented analysis. We illustrate the method on the applications using synthetic and real-world data.

#110 Learning the Number of Neurons in Deep Networks

Jose M Alvarez (NICTA)
Mathieu Salzmann (EPFL)

Nowadays, the number of layers and of neurons in each layer of a deep network are typically set manually. While very deep and wide networks have proven effective in general, they come at a high memory and computation cost, thus making them impractical for constrained platforms. These networks, however, are known to have many redundant parameters, and could thus, in principle, be replaced by more compact architectures. In this paper, we introduce an approach to automatically determining the number of neurons in each layer of a deep network during learning. To this end, we propose to make use of a group sparsity regularizer on the parameters of the network, where each group is defined to act on a single neuron. Starting from an overcomplete network, we show that our approach can reduce the number of parameters by up to 72\% while retaining or even improving the network accuracy.

#111 Strategic Attentive Writer for Learning Macro-Actions

Alexander Vezhnevets (Google DeepMind)
Volodymyr Mnih
Simon Osindero (Google DeepMind)
Alex Graves
Oriol Vinyals
John Agapiou
koray kavukcuoglu (Google DeepMind)

We present a novel deep recurrent neural network architecture that learns to build implicit plans in an end-to-end manner by purely interacting with an environment in reinforcement learning setting. The network builds an internal plan, which is continuously updated upon observation of the next input from the environment. It can also partition this internal representation into contiguous sub- sequences by learning for how long the plan can be committed to -- i.e. followed without re-planing. Combining these properties, the proposed model, dubbed STRategic Attentive Writer (STRAW) can learn high-level, temporally abstracted macro- actions of varying lengths that are solely learnt from data without any prior information. These macro-actions enable both structured exploration and economic computation. We experimentally demonstrate that STRAW delivers strong improvements on several ATARI games by employing temporally extended planning strategies (e.g. Ms. Pacman and Frostbite). It is at the same time a general algorithm that can be applied on any sequence data. To that end, we also show that when trained on text prediction task, STRAW naturally predicts frequent n-grams (instead of macro- actions), demonstrating the generality of the approach.

#112 Active Learning from Imperfect Labelers

Songbai Yan (Univ. of California)
Kamalika Chaudhuri (Univ. of California)
Tara Javidi (Univ. of California)

This paper studies the problem of active learning where the labeler can not only return an incorrect label but also abstain from labeling. Different noise and abstention models of the labeler are considered and the amount of queries to the labeler required to learn a good classifier is analyzed. An adaptive algorithm which automatically requests less queries with a more informative labeler is provided and is proved to have nearly optimal query complexity. The analysis also shows the gains of allowing a labeler to abstain from labeling by quantifying the reduction in the number of queries.

#113 Probabilistic Linear Multistep Methods

Onur Teymur (Imperial College London)
Kostas Zygalakis
Ben Calderhead

We present a derivation of Adams family LMMs starting from a Gaussian process framework. In the limit, this formulation coincides with classic (deterministic) methods, which have been used as higher-order IVP solvers for over a century. Furthermore, the natural probabilistic framework provided by the Gaussian process formulation allows a probabilistic version of these methods to be derived, in the spirit of a number of other simpler probabilistic ODE solvers presented in recent literature [citations]. The improved efficiency arising from the multistep approach comes at very little additional computational cost. One shortcoming with previous work in this area is the apparent arbitrariness of the scale of the introduced `noise' (though some conditions on the relationship between noise and integrator step-size are proven in [citations]). The Gaussian process approach rectifies this by careful choice of covariance function -- the scale of the probabilistic noise can be made to exactly match the theoretical Local Truncation Error from the deterministic method. We show that the integrator possesses the correct theoretical convergence properties.

#114 More Supervision, Less Computation: Statistical-Computational Tradeoffs in Weakly Supervised Learning

Xinyang Yi (UT Austin)
Zhaoran Wang (Princeton Univ.)
Zhuoran Yang (Princeton Univ.)
Constantine Caramanis
Han Liu

We study the weakly supervised binary classification problem where the labels are randomly flipped or missing with probability 1-α. Although there exist numerous algorithms, it remains theoretically unexplored how the statistical accuracy and computational efficiency of these algorithms depend on the degree of supervision, which is quantified by α. In this paper, we characterize the effect of α by exploring both computational and information-theoretic boundaries, namely, the minimax-optimal statistical accuracies achievable for tractable algorithms and by all algorithms, correspondingly. For small α, there exists a wide gap between these two boundaries, which represents the computational price of achieving the information-theoretic boundary due to the lack of supervision. Interestingly, this gap narrows as α increases. In other words, having more supervision, i.e., more correct labels, not only improves the statistical accuracy as one may expect, but more importantly, enhances the computational efficiency.

#115 Mutual information for symmetric rank-one matrix estimation: A proof of the replica formula

jean barbier (EPFL)
Mohamad Dia (EPFL)
Nicolas Macris (EPFL)
Florent Krzakala
Thibault Lesieur (IPHT Saclay)
Lenka Zdeborová

Factorizing low-rank matrices has many applications in machine learning and statistics. For probabilistic models in the Bayes optimal setting, a general expression for the mutual information has been proposed using heuristic statistical physics computations, and proven in few specific cases. Here, we show how to rigorously prove the conjectured formula for the symmetric rank-one case. This allows to express the minimal mean-square-error and to characterize the detectability phase transitions in a large set of estimation problems ranging from community detection to sparse PCA. We also show that for a large set of parameters, an iterative algorithm called approximate message-passing is Bayes optimal. There exists, however, a gap between what currently known polynomial algorithms can do and what is expected information theoretically. Additionally, the proof technique has an interest of its own and exploits three essential ingredients: the interpolation method introduced in statistical physics by Guerra, the analysis of the approximate message-passing algorithm and the theory of spatial coupling and threshold saturation in coding. Our approach is generic and applicable to other open problems in statistical estimation where heuristic statistical physics predictions are available.

#116 Coin Betting and Parameter-Free Online Learning

Francesco Orabona (Yahoo Research)
David Pal

In the recent years, a number of parameter-free algorithms for online linear optimization over Hilbert spaces and for learning with expert advice have been developed. These algorithms achieve optimal regret bounds that depend on the unknown competitors, without having to tune the learning rates with oracle choices. We present a new intuitive framework to design parameter-free algorithms based on a reduction to betting on outcomes of an adversarial coin. We instantiate it using a betting algorithm based on the Krichevsky-Trofimov estimator. The resulting algorithms are simple, with no parameters to be tuned, and they improve or match previous results in terms of regret guarantee and per-round complexity.

#117 Normalized Spectral Map Synchronization

Yanyao Shen (UT Austin)
Qixing Huang (Toyota Technological Institute at Chicago)
Nati Srebro
Sujay Sanghavi

The algorithmic advancement of synchronizing maps is important in order to solve a wide range of practice problems with possible large-scale dataset. In this paper, we provide theoretical justifications for spectral techniques for the map synchronization problem, i.e., it takes as input a collection of objects and noisy maps estimated between pairs of objects, and outputs clean maps between all pairs of objects. We show that a simple normalized spectral method that projects the blocks of the top eigenvectors of a data matrix to the map space leads to surprisingly good results. As the noise is modelled naturally as random permutation matrix, this algorithm NormSpecSync leads to competing theoretical guarantees as state-of-the-art convex optimization techniques, yet it is much more efficient. We demonstrate the usefulness of our algorithm in a couple of applications, where it is optimal in both complexity and exactness among existing methods.

#118 On Explore-Then-Commit strategies

Aurelien Garivier
Tor Lattimore
Emilie Kaufmann

We study the problem of minimising regret in two-armed bandit problems with Gaussian rewards. Our objective is to use this simple setting to illustrate that strategies based on an exploration phase (up to a stopping time) followed by exploitation are necessarily suboptimal. The results hold regardless of whether or not the difference in means between the two arms is known. Besides the main message, we also refine existing deviation inequalities, which allow us to design fully sequential strategies with finite-time regret guarantees that are (a) asymptotically optimal as the horizon grows and (b) order-optimal in the minimax sense. Furthermore we provide empirical evidence that the theory also holds in practice and discuss extensions to non-gaussian and multiple-armed case.

#119 Learning Kernels with Random Features

Aman Sinha (Stanford Univ.)
John C Duchi

Randomized features provide a computationally efficient way to approximate kernel machines in machine learning tasks. However, such methods require a user-defined kernel as input. In this paper, we extend the randomized-feature approach to the task of learning a kernel (via its associated random features). Specifically, we present an efficient optimization problem that learns a kernel in a supervised manner. We prove the consistency of the estimated kernel as well as generalization bounds for the class of estimators induced by the optimized kernel, and we experimentally evaluate our technique on numerous datasets. Our approach is efficient and highly scalable, and we attain competitive results with a fraction of the training cost of other techniques.

#120 Robustness of classifiers: from adversarial to random noise

Alhussein Fawzi
Seyed-Mohsen Moosavi-Dezfooli (EPFL)
Pascal Frossard (EPFL)

Several recent works have shown that state-of-the-art classifiers are vulnerable to worst-case (i.e., adversarial) perturbations of the datapoints. On the other hand, it has been empirically observed that these same classifiers are relatively robust to random noise. In this paper, we propose to study a semi-random noise regime that generalizes both the random and worst-case noise regimes. We establish precise theoretical bounds on the robustness of classifiers in this general regime, which depend on the curvature of the classifier's decision boundary. Our results are, as far as we know, the first rigorous analysis on the robustness of nonlinear classifiers in this general noise regime based on the curvature. Our bounds confirm and quantify the empirical observations that classifiers satisfying curvature constraints are robust to random noise. Moreover, we quantify the robustness of classifiers in terms of the subspace dimension in the semi-random noise regime, and show that our bounds remarkably interpolate between the worst-case and random noise regimes. We perform experiments and show that the derived bounds provide very accurate estimates when applied to various state-of-the-art deep neural networks and datasets. This result suggests bounds on the curvature of the classifiers' decision boundaries that we support experimentally, and offers insights onto the geometry of high dimensional classification problems.

#121 Adaptive Skills Adaptive Partitions (ASAP)

Daniel J Mankowitz (Technion)
Timothy A Mann (Google DeepMind)
Shie Mannor (Technion)

We introduce the Adaptive Skills, Adaptive Partitions (ASAP) framework that (1) learns skills (i.e., temporally extended actions or options) as well as (2) where to apply them. We believe that both (1) and (2) are necessary for a truly general skill learning framework, which is a key building block needed to scale up to lifelong learning agents. The ASAP framework is also able to solve related new tasks simply by adapting where it applies its existing learned skills. We prove that ASAP converges to a local optimum under natural conditions. Finally, our experimental results, which include a RoboCup domain, demonstrate the ability of ASAP to learn where to reuse skills as well as solve multiple tasks with considerably less experience than solving each task from scratch.

#122 Gaussian Process Bandit Optimisation with Multi-fidelity Evaluations

Kirthevasan Kandasamy (CMU)
Gautam Dasarathy (Carnegie Mellon Univ.)
Junier B Oliva
Jeff Schneider (CMU)
Barnabas Poczos

In many scientific and engineering applications, we are tasked with the optimisation of an expensive to evaluate black box function . Traditional methods for this problem assume just the availability of this single function. However, in many cases, cheap approximations to may be obtainable. For example, the expensive real world behaviour of a robot can be approximated by a cheap computer simulation. We can use these approximations to eliminate low function value regions and use the expensive evaluations to in a small but promising region and speedily identify the optimum. We formalise this task as a multi-fidelity} bandit problem where the target function and its approximations are sampled from a Gaussian process. We develop \mfgpucb, a novel method based on upper confidence bound techniques for this settting. In our theoretical analysis we demonstrate that it exhibits precisely the above behaviour, and achieves better regret than strategies which ignore multi-fidelity information. \mfgpucbs outperforms such naive strategies and other multi-fidelity methods on several synthetic and real experiments.

#123 Flexible Models for Microclustering with Applications to Entity Resolution

Brenda Betancourt (Duke Univ.)
Giacomo Zanella (The Univ. of Warick)
Jeff Miller (Duke Univ.)
Hanna Wallach (Microsoft Research)
Abbas Zaidi (Duke Univ.)
Rebecca Steorts (Duke Univ.)

Most generative models for clustering implicitly assume that the number of data points in each cluster grows linearly with the total number of data points. Finite mixture models, Dirichlet process mixture models, and Pitman--Yor process mixture models make this assumption, as do all other infinitely exchangeable clustering models. However, for some tasks, this assumption is inappropriate. For example, when performing entity resolution, the size of each cluster is often unrelated to the size of the data set. Consequently, each cluster contains a negligible fraction of the total number of data points. Such tasks therefore require models that yield clusters whose sizes grow sublinearly with the size of the data set. We address this requirement by defining the microclustering property} and introducing two new models that exhibit this property. We compare the two proposed models to several commonly used clustering models using four data sets.

#124 Stochastic Gradient Richardson-Romberg Markov Chain Monte Carlo

Alain Durmus (Telecom ParisTech)
Umut Simsekli
Eric Moulines (Ecole Polytechnique)
Roland Badeau (Telecom ParisTech)
Gaël RICHARD (Telecom ParisTech)

Stochastic Gradient Markov Chain Monte Carlo (SG-MCMC) algorithms have become increasingly popular for Bayesian inference in large-scale applications. Even though these methods have proved useful in several scenarios, their performance is often limited by their bias. In this study, we propose a novel sampling algorithm that aims to reduce the bias of SG-MCMC while keeping the variance at a reasonable level. Our approach is based on a numerical sequence acceleration method, namely the Richardson-Romberg extrapolation, which simply boils down to running almost the same SG-MCMC algorithm twice in parallel with different step sizes. We illustrate our framework on the popular Stochastic Gradient Langevin Dynamics (SGLD) algorithm and propose a novel SG-MCMC algorithm referred to as Stochastic Gradient Richardson-Romberg Langevin Dynamics (SGRRLD). We provide formal theoretical analysis and show that SGRRLD is asymptotically consistent, satisfies a central limit theorem, and its non-asymptotic bias and the mean squared-error can be bounded. Our results show that SGRRLD attains higher rates of convergence than SGLD in both finite-time and asymptotically, and it achieves the theoretical accuracy of the methods that are based on higher-order integrators. We support our findings using both synthetic and real data experiments.

#125 Online and Differentially-Private Tensor Decomposition

Yining Wang (Carnegie Mellon Univ.)
Anima Anandkumar (UC Irvine)

Tensor decomposition is positioned to be a pervasive tool in the era of big data. In this paper, we resolve many of the key algorithmic questions regarding robustness, memory efficiency, and differential privacy of tensor decomposition. We propose simple variants of the tensor power method which enjoy these strong properties. We propose the first streaming method with a linear memory requirement. Moreover, we present a noise calibrated tensor power method with efficient privacy guarantees. At the heart of all these guarantees lies a careful perturbation analysis derived in this paper which improves up on the existing results significantly.

#126 Maximal Sparsity with Deep Networks?

Bo Xin (Peking Univ.)
Yizhou Wang (Peking Univ.)
Wen Gao (peking Univ.)
David Wipf

The iterations of many sparse estimation algorithms are comprised of a fixed linear filter cascaded with a thresholding nonlinearity, which collectively resemble a typical neural network layer. Consequently, a lengthy sequence of algorithm iterations can be viewed as a deep network with shared, hand-crafted layer weights. It is therefore quite natural to examine the degree to which a learned network model might act as a viable surrogate for traditional sparse estimation in domains where ample training data is available. While the possibility of a reduced computational budget is readily apparent when a ceiling is imposed on the number of layers, our work primarily focuses on estimation accuracy. In particular, it is well-known that when a signal dictionary has coherent columns, as quantified by a large RIP constant, then most tractable iterative algorithms are unable to find maximally sparse representations. In contrast, we demonstrate both theoretically and empirically the potential for a trained deep network to recover minimal ℓ_0-norm representations in regimes where existing methods fail. The resulting system is deployed on a practical photometric stereo estimation problem, where the goal is to remove sparse outliers that can disrupt the estimation of surface normals from a 3D scene.

#127 Efficient High-Order Interaction-Aware Feature Selection Based on Conditional Mutual Information

Alexander Shishkin (Yandex)
Anastasia Bezzubtseva (Yandex)
Alexey Drutsa (Yandex)
Ilia Shishkov (Yandex)
kglad Gladkikh (Yandex)
Gleb Gusev (Yandex LLC)
Pavel Serdyukov (Yandex)

This study introduces a novel feature selection approach CMICOT, which is a further evolution of filter methods with sequential forward selection (SFS) whose scoring functions are based on conditional mutual information (MI). We state and study a novel saddle point (max-min) optimization problem to build a scoring function that is able to identify joint interactions between several (up to t\in\mathbb{N}) features. In this way, our method fills in the underexplored gap of SFS-based techniques with high-order (t > 3) dependencies in MI. The use of such MI raises the associated costs of increasing computational complexity and demands a larger number of data instances required for accurate MI estimation. We mitigate these costs by means of a greedy approximation and utilization of binary representatives that make our technique able to be effectively used. The superiority of our approach is demonstrated by comparison with recently proposed interaction-aware filters and several interaction-agnostic state-of-the-art ones on ten publicly available benchmark datasets.

#128 Geometric Dirichlet Means Algorithm for Topic Inference

Mikhail Yurochkin (Univ. of Michigan)
Long Nguyen

We propose a geometric algorithm for topic learning and inference that is built on the convex geometry of topics arising from the Latent Dirichlet Allocation (LDA) model and its nonparametric extensions. To this end we study the optimization of a geometric loss function, which is a surrogate to the LDA's likelihood. Our method involves a fast optimization based weighted clustering procedure augmented with geometric corrections, which overcomes the computational and statistical inefficiencies encountered by other techniques based on Gibbs sampling and variational inference, while achieving the accuracy comparable to that of a Gibbs sampler. The topic estimates produced by our method are shown to be statistically consistent under some conditions. The algorithm is evaluated with extensive experiments on simulated and real data.

#129 Interaction Screening: Efficient and Sample-Optimal Learning of Ising Models

Marc Vuffray (Los Alamos National Laboratory)
Sidhant Misra (Los Alamos National Laboratory)
Andrey Lokhov (Los Alamos National Laboratory)
Michael Chertkov (Los Alamos National Laboratory)

We consider the problem of learning the underlying graph of an unknown Ising model on p spins from a collection of i.i.d. samples generated from the model. We suggest a new estimator that is computationally efficient and requires a number of samples that is near-optimal with respect to previously established information-theoretic lower-bound. Our statistical estimator has a physical interpretation in terms of "interaction screening". The estimator is consistent and is efficiently implemented using convex optimization. We prove that with appropriate regularization, the estimator recovers the underlying graph using a number of samples that is logarithmic in the system size p and exponential in the maximum coupling-intensity and maximum node-degree.

#130 Multi-armed Bandits: Competing with Optimal Sequences

Zohar Karnin
Oren Anava (Technion)

We consider sequential decision making problem in the adversarial setting, where regret is measured with respect to the optimal sequence of actions} and the feedback adheres the bandit setting. It is well-known that obtaining sublinear regret in this setting is impossible in general, which arises the question of when can we do better than linear regret?} Previous works show that when the environment is guaranteed to vary slowly and furthermore we are given prior knowledge regarding its variation} (i.e., a limit on the

amount of changes suffered by the environment), then this task is feasible. The caveat however is that such prior knowledge is not likely to be available in practice, which causes the obtained regret bounds to be somewhat irrelevant. Our main result is a regret guarantee that scales with the variation} parameter of the environment, without requiring any prior knowledge about it whatsoever. By that, we also resolve an open problem posted by [Gur, Zeevi and Besbes, NIPS' 14]. An important key component in our result is a statistical test for identifying non-stationarity in a sequence of independent random variables. This test either identifies non-stationarity or upper-bounds the absolute deviation of the corresponding sequence of mean values in terms of its total variation. This test is interesting on its own right and has the potential to be found useful in additional settings.

#131 Catching heuristics are optimal control policies

Boris Belousov (TU Darmstadt)
Gerhard Neumann
Constantin A Rothkopf
Jan R Peters

Two seemingly contradictory theories attempt to explain how humans move to intercept an airborne ball. One theory posits that humans predict the ball trajectory to optimally plan future actions; the other claims that, instead of performing such complicated computations, humans employ heuristics to reactively choose appropriate actions based on immediate visual feedback. In this paper, we show that interception strategies appearing to be heuristics can be understood as computational solutions to the optimal control problem faced by a ball-catching agent acting under uncertainty. Modeling catching as a continuous partially observable Markov decision process and employing stochastic optimal control theory, we discover that the four main heuristics described in the literature are optimal solutions if the catcher has sufficient time to continuously visually track the ball. Specifically, by varying model parameters such as noise, time to ground contact, and perceptual latency, we show that different strategies arise under different circumstances. The catcher's policy switches between generating reactive and predictive behavior based on the ratio of system to observation noise and the ratio between reaction time and task duration. Thus, we provide a rational account of human ball-catching behavior and a unifying explanation for seemingly contradictory theories of target interception on the basis of stochastic optimal control.

#132 Fast stochastic optimization on Riemannian manifolds

Hongyi Zhang (MIT)
Sashank J. Reddi (Carnegie Mellon Univ.)
Suvrit Sra (MIT)

We study optimization of finite sums of geodesically} smooth functions on Riemannian manifolds. Although variance reduction techniques for optimizing finite-sum problems have witnessed a huge surge of interest in recent years, all existing work is limited to vector space problems. We introduce Riemannian SVRG}, a new variance reduced Riemannian optimization method. We analyze this method for both geodesically smooth convex} and nonconvex} functions. Our analysis reveals that Riemannian SVRG comes with advantages of the usual SVRG method, but with factors depending on manifold curvature that influence its convergence. To the best of our knowledge, ours is the first fast} stochastic Riemannian method. Moreover, our work offers the first non-asymptotic complexity analysis for nonconvex Riemannian optimization (even for the batch setting). Our results have several implications; for instance, they offer a Riemannian perspective on variance reduced PCA, which promises a short, transparent convergence analysis.

#133 A Comprehensive Linear Speedup Analysis for Asynchronous Stochastic Parallel Optimization from Zeroth-Order to First-Order

Xiangru Lian (Univ. of Rochester)
Huan Zhang
Cho-Jui Hsieh
Yijun Huang
Ji Liu

Asynchronous parallel optimization received substantial successes and extensive attention recently. One of core theoretical questions is how much speedup (or benefit) the asynchronous parallelization can bring to us. This paper provides a comprehensive and generic analysis to study the speedup property for a broad range of asynchronous parallel stochastic algorithms from the zeroth order to the first order methods. Our result recovers or improves existing analysis on special cases, provides more insights for understanding the asynchronous parallel behaviors, and suggests a novel asynchronous parallel zeroth order method for the first time. Our experiments provide novel applications of the proposed asynchronous parallel zeroth order method on hyper parameter tuning and model blending problems.

#134 Stochastic Gradient MCMC with Stale Gradients

Changyou Chen
Nan Ding (Google)
Chunyuan Li (Duke)
Yizhe Zhang (Duke Univ.)
Lawrence Carin

Stochastic gradient MCMC (SG-MCMC) has played an important role in large-scale Bayesian learning, with well-developed theoretical convergence properties. In such applications of SG-MCMC, it is becoming increasingly popular to employ distributed systems, where stochastic gradients are computed based on some outdated parameters, yielding what are termed stale gradients. While stale gradients could be directly used in SG-MCMC, their impact on convergence properties has not been well studied. In this paper we develop theory to show that while the bias and MSE of an SG-MCMC algorithm depend on the staleness of stochastic gradients, its estimation variance (relative to the expected estimate, based on a prescribed number of samples) is independent of it. In a simple Bayesian distributed system with SG-MCMC, where stale gradients are computed asynchronously by a set of workers, our theory indicates a linear speedup on the decrease of the estimation variance w.r.t.\! the number of workers. Experiments on synthetic data and deep neural networks validate our theory, demonstrating the effectiveness and scalability of SG-MCMC with stale gradients.

#135 Disentangling factors of variation in deep representation using adversarial training

Michael F Mathieu (NYU)
Junbo Zhao (NYU)
Aditya Ramesh (NYU)
Pablo Sprechmann
Yann LeCun (NYU)

We propose a deep generative model for learning to distill the hidden factors of variation within a set of labeled observations into two complementary codes. One code describes the factors of variation relevant to solving a specified task. The other code describes the remaining factors of variation that are irrelevant to solving this task. The only available source of supervision during the training process comes from our ability to distinguish among different observations belonging to the same category. Concrete examples include multiple images of the same object from different viewpoints, or multiple speech samples from the same speaker. In both of these instances, the factors of variation irrelevant to classification are implicitly expressed by intra-class variabilities, such as the relative position of an object in an image, or the linguistic content of an utterance. Most existing approaches for solving this problem rely heavily on having access to pairs of observations only sharing a single factor of variation, e.g. different objects observed in the exact same conditions. This assumption is often not encountered in realistic settings where data acquisition is not controlled and labels for the uninformative components are not available. In this work, we propose to overcome this limitation by augmenting deep convolutional autoencoders with a form of adversarial training. Both factors of variation are implicitly captured in the organization of the learned embedding space, and can be used for solving single-image analogies. Experimental results on synthetic and real datasets show that the proposed method is capable of disentangling the influences of style and content factors using a flexible representation, as well as generalizing to unseen styles or content classes.

#136 Consistent Kernel Mean Estimation for Functions of Random Variables

Adam Scibior (Univ. of Cambridge)
Carl-Johann Simon-Gabriel (MPI Tuebingen)
Ilya Tolstikhin
Prof. Bernhard Schölkopf

We provide a theoretical foundation for non-parametrically estimating functions of random variables using kernel mean embeddings. We show that for any continuous function f, consistent estimators of the mean embedding of a random variable X lead to consistent estimators of the mean embedding of f(X). For Gaussian kernels and sufficiently smooth functions we also provide rates of convergence. Our results also apply for functions of multiple random variables. If the variables are dependent, we require an estimator of the mean embedding of their joint distribution as a starting point; if they are independent, it is sufficient to have separate mean embeddings of their marginal distributions. In either case, our results cover both mean embeddings expressed based on i.i.d. samples as well as reduced set expansions in terms of dependent expansion points. The latter serves as a justification for using such expansions to limit memory resources when we use the approach as a basis for probabilistic programming.

#137 DECOrrelated feature space partitioning for distributed sparse regression

Xiangyu Wang (Duke Univ.)
David B Dunson (Duke Univ.)
Chenlei Leng (Univ. of Warwick)

Fitting statistical models is computationally challenging when the sample size or the dimension of the dataset is huge. An attractive approach for down-scaling the problem size is to first partition the dataset into subsets and then fit using distributed algorithms. The dataset can be partitioned either horizontally (in the sample space) or vertically (in the feature space). While the majority of the literature focuses on sample space partitioning, feature space partitioning is more effective when p >> n. Existing methods for partitioning features, however, are either vulnerable to high correlations or inefficient in reducing the model dimension. In this paper, we solve these problems through a new embarrassingly parallel framework named DECO for distributed variable selection and parameter estimation. In DECO, variables are first partitioned and allocated to m distributed workers. The decorrelated subset data within each worker are then fitted via any algorithm designed for high-dimensional problems. We show that by incorporating the decorrelation step, DECO can achieve consistent variable selection and parameter estimation on each subset with (almost) no assumptions. In addition, the convergence rate is nearly minimax optimal for both sparse and weakly sparse models and does NOT depend on the partition number m. Extensive numerical experiments are provided to illustrate the performance of the new framework.

#138 Coupled Generative Adversarial Networks

Ming-Yu Liu (MERL)
Oncel Tuzel (Mitsubishi Electric Research Labs (MERL))

We propose the coupled generative adversarial nets (CoGAN) framework for generating pairs of corresponding images in two different domains. The framework consists of a pair of generative adversarial nets, each responsible for generating images in one domain. We show that by enforcing a simple weight-sharing constraint, the CoGAN learns to generate pairs of corresponding images without existence of any pairs of corresponding images in the two domains in the training set. In other words, the CoGAN learns a joint distribution of images in the two domains from images drawn separately from the marginal distributions of the individual domains. This is in contrast to the existing multi-modal generative models, which require corresponding images for training. We apply the CoGAN to several pair image generation tasks. For each task, the CoGAN learns to generate convincing pairs of corresponding images. We further demonstrate the applications of the CoGAN framework for the domain adaptation and cross-domain image generation tasks.

#139 Matching Networks for One Shot Learning

Oriol Vinyals
Charles Blundell (DeepMind)
Timothy Lillicrap (Google DeepMind)
koray kavukcuoglu (Google DeepMind)
Daan Wierstra (Google DeepMind)

Learning from a few examples remains a key challenge in machine learning. Despite recent advances in important domains such as vision and language, the standard supervised deep learning paradigm does not offer a satisfactory solution for learning new concepts rapidly from little data. In this work, we employ ideas from metric learning based on deep neural features and from recent advances that augment neural

networks with external memories. Our framework learns a network that maps a small labelled support set and an unlabelled example to its label, obviating the need for fine-tuning to adapt to new class types. We then define one-shot learning problems on vision (using Omniglot, ImageNet) and language tasks. Our algorithm improves one-shot accuracy on ImageNet from 82.2% to 87.8% and from 88% accuracy to 95% accuracy on Omniglot compared to competing approaches. We also demonstrate the usefulness of the same model on language modeling by introducing a one-shot task on the Penn Treebank.

#140 Distributed Flexible Nonlinear Tensor Factorization

Shandian Zhe (Purdue Univ.)
Kai Zhang (NEC Labs America)
Pengyuan Wang (Yahoo! Research)
Kuang-chih Lee
Zenglin Xu
Alan Qi (Ant financial service group)
Zoubin Ghahramani

Tensor factorization is a powerful tool to analyse multi-way data. Compared with traditional multi-linear methods, nonlinear tensor factorization models are capable of capturing more complex relationships in the data; however, they are computationally quite expensive and may suffer severe learning bias in case of extreme data sparsity. To overcome these limitations, in this paper we propose a distributed, flexible nonlinear tensor factorization model. Our model can effectively avoid the expensive computations and structural restrictions of the Kronecker-product in existing TGP formulations, allowing an arbitrary subset of tensorial entries to be selected to contribute to the training. At the same time, we derive a tractable and tight variational evidence lower bound (ELBO) that enables highly decoupled, parallel computations and high-quality inference. Based on the new bound, we develop a distributed inference algorithm in the MapReduce framework, which is key-value-free and can fully exploit the memory cache mechanism in fast MapReduce systems such as SPARK. Experimental results fully demonstrate the advantages of our method over several state-of-the-art approaches, in terms of both predictive performance and computational efficiency. Moreover, our approach shows a promising potential in the application of Click-Through-Rate (CTR) prediction for online advertising.

#141 Tracking the Best Expert in Non-stationary Stochastic Environments

Chen-Yu Wei (Academia Sinica)
Yi-Te Hong (Academia Sinica)
Chi-Jen Lu (Academia Sinica)

We study the dynamic regret of sequential prediction in the non-stationary stochastic environment. We introduce a new parameter Λ, which measures the total statistical variance of the loss distributions over T rounds of the process, and study how this amount affects the achievable regret. We investigate the interaction between Λ and Γ, which counts the number of times the distributions change, as well as Λ and V, which measures how far the distributions deviates over time. One striking result we find is that even when Γ, V, and Λ are all restricted to constant, the regret lower bound in the bandit setting still grows with T. The other highlight is that in the full-information setting, a constant regret becomes achievable with constant Γ and Λ, as it can be made independent of T, while with constant V and Λ, the regret still has a T^{1/3} dependency. We not only propose algorithms with upper bound guarantee, but prove their matching lower bounds as well.

#142 Deep Alternative Neural Networks: Exploring Contexts as Early as Possible for Action Recognition

Jinzhuo Wang (PKU)
Wenmin Wang (peking Univ.)
xiongtao Chen (peking Univ.)
Ronggang Wang (peking Univ.)
Wen Gao (peking Univ.)

Contexts are crucial for action recognition in video. Existing methods often mine contexts after extracting hierarchical local features and focus on their high-order encodings. This paper instead explores contexts as early as possible and leverages their evolutions for action recognition. In particular, we introduce a novel architecture called deep alternative neural network (DANN) stacking alternative layers. Each alternative layer consists of a volumetric convolutional layer followed by a recurrent layer. The former acts as local feature learner while the latter is responsible for collecting context information. Compared with feed-forward neural networks, DANN learns local features and their contexts from the very beginning of the architecture. This setting helps to preserve context evolutions which we show are essential to recognize similar actions and improve performance. Besides, we present an adaptive method to determine the temporal size for network input based on the density of optical flow energy, and develop a volumetric pyramid pooling layer to deal with input video clips of arbitrary sizes. We demonstrate the effectiveness of DANN architecture on two standard benchmarks HMDB51 and UCF101 and report competitive or superior results compared to the state-of-the-art results.

#143 Learning Parametric Sparse Models for Image Super-Resolution

Yongbo Li (Xidian Univ.)
Weisheng Dong (Xidian Univ.)
Xuemei Xie (Xidian Univ.)
GUANGMING Shi (Xidian Univ.)
Xin Li (WVU)
Donglai Xu (Teesside Univ.)

Learning accurate prior knowledge of natural images is of great importance for single image super-resolution (SR). Existing SR methods either learn the prior from the low/high-resolution patch pairs or estimate the prior models from the input low-resolution (LR) image. Specifically, high-frequency details are learned in the former methods. Though effective, they are heuristic and have limitations in dealing with blurred LR images; while the latter suffers from the limitations of frequency aliasing. In this paper, we propose to combine those two lines of ideas for image super-resolution. More specifically, the parametric sparse prior of the desirable high-resolution (HR) image patches are learned from both the input low-resolution (LR) image and a training image dataset. With the learned sparse priors, the sparse codes and thus the HR image patches can be accurately recovered by solving a sparse coding problem. Experimental results show that the proposed SR method outperforms existing state-of-the-art methods in terms of both subjective and objective image qualities.

#144 Kernel Observers: Systems-Theoretic Modeling and Inference of Spatiotemporally Evolving Processes

Hassan A Kingravi (Pindrop Security Services)
Harshal R Maske (UIUC)
Girish Chowdhary (UIUC)

We consider the problem of estimating the latent state of a spatiotemporally evolving continuous function using very few sensor measurements.We show that layering a dynamical systems prior over temporal evolution of weights of a kernel model is a valid approach to spatiotemporal modeling that does not necessarily require the design of complex nonstationary kernels. Furthermore, we show that such a predictive model can be utilized to determine sensing locations that guarantee that the hidden state of the phenomena can be recovered with very few measurements. We provide sufficient conditions on the number and spatial location of samples required to guarantee state recovery, and provide a lower bound on the minimum number of samples required to robustly infer the hidden states. Our approach outperforms existing methods in numerical experiments.

#145 Learning brain regions via large-scale online structured sparse dictionary learning

Elvis DOHMATOB (Inria)
Arthur Mensch (inria)
Gael Varoquaux
Bertrand Thirion

In neuro-imaging, inter-subject variability is often handled as a statistical residual and discarded. Yet there is evidence that it displays structure and contains important information. Uni-variate models are ineffective both computationally and statistically due to the large number of voxels compared to the number of subjects. We propose a multi-variate online dictionary-learning / matrix-factorization method for obtaining decompositions with structured and sparse components (aka atoms). Sparsity is to be understood in the usual sense: the atoms contain mostly zeros. This is imposed via an L1 penalty on the atoms. By "structured", we mean that the atoms are piece-wise smooth and compact, making up blobs, as opposed to scattered patterns of activation. We propose to use a Sobolev (Laplacian) penalty to impose this type of structure. Combining the two penalties, we therefore obtain decompositions which are closer to what is known of the brain. This non-trivially extends the online dictionary-learning / matrix-factorization work of Mairal et al. (2010), at the price of only a factor of 2 or 3 on the overall running time. Just like the reference Mairal et al. method, the online nature of our proposed algorithm allows it to scale to arbitrarily sized datasets. To complement the theoretical results, we also present and discuss comparative numerical experiments on brain data.

#146 Scaling Factorial Hidden Markov Models: Stochastic Variational Inference without Messages

Yin Cheng Ng (Univ. College London)
Pawel M Chilinski (Univ. College London)
Ricardo Silva (Univ. College London)

Factorial Hidden Markov Models (FHMMs) are powerful models for sequential data but they do not scale well with long sequences. We propose a scalable inference and learning algorithm for FHMMs that draws on ideas from the stochastic variational inference, neural network and copula literatures. Unlike existing approaches, the proposed algorithm requires no message passing procedure among latent variables and can be distributed to a network of computers to speed up learning. Our experiments corroborate that the proposed algorithm does not introduce further approximation bias compared to the proven structured mean-field algorithm, and achieves better performance with long sequences and large FHMMs.

#147 A Bandit Framework for Strategic Regression

Yang Liu (Harvard Univ.)
Yiling Chen

We consider a learner's problem of acquiring data dynamically for training a regression model, where the training data are collected from strategic data sources. A fundamental challenge is to incentivize data holders to exert effort to improve the quality of their reported data, despite that the quality is not directly verifiable by the learner. In this work, we study a dynamic data acquisition process where data holders can contribute multiple times. Using a bandit framework, we leverage on the long-term incentive of future job opportunities to incentivize high-quality contributions. We propose a Strategic Regression-Upper Confidence Bound (SR-UCB) framework,\com{might be better to spell out UCB when it's first used.} an UCB-style index combined with a simple payment rule, where the index of a worker approximates the quality of his past contributions and is used by the learner to determine whether the worker receives future work. For linear regression and certain family of non-linear regression problems, we show that SR-UCB enables a $O(\sqrt{\log T}/T)$-Bayesian Nash Equilibrium (BNE) where each worker exerting a target effort level that the learner has chosen, with T being the number of data acquisition stages. The SR-UCB framework also has some other desirable properties: (1) The indexes can be updated in an online fashion (hence computationally light). (2) A slight variant, namely PSR-UCB, is able to preserve (O(√log T), O(1/T^2))-differential privacy for workers' data, with only a small compromise on incentives (achieving $O(\sqrt{\log^3 T}/\sqrt{T})$-BNE).

#148 Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering

Michaël Defferrard (EPFL)
Xavier Bresson
Pierre Vandergheynst (EPFL)

Convolutional neural networks have greatly improved state-of-the-art performances in computer vision and speech analysis tasks, due to its high ability to extract multiple levels of representations of data. In this work, we are interested in generalizing convolutional neural networks from low-dimensional regular grids, where image, video and speech are represented, to high-dimensional irregular domains, such as social networks, biological graphs like gene regulatory and brain connectivity networks, telecommunication networks, or words' embedding. We present a formulation of convolutional neural networks on graphs in the context of the emerging field of signal processing on graphs, which provides the necessary mathematical background and efficient numerical schemes to design fast localized convolutional filters on graphs. Numerical experiments on MNIST and 20NEWS demonstrate the ability of the system to learn local stationary features on graphs, as long as the graph is well-constructed.

#149 Variational Gradient Descent: A General Purpose Bayesian Inference Algorithm

Qiang Liu
Dilin Wang (Dartmouth College)

We propose a general purpose variational inference algorithm that forms a natural counterpart of gradient descent for optimization. Our method iteratively transports a set of particles to match with the target

distribution, by applying a form of functional gradient descent that minimizes the KL divergence. Comprehensive empirical comparisons with existing methods are performed on real world models, including a Bayesian neural network on which we are competitive with the state-of-art methods. The derivation of our method is based on a new theoretical result that connects KL divergence with a recently proposed kernelized Stein discrepancy, which is of independent interest.

#150 Deep Learning Models of the Retinal Response to Natural Scenes

Lane McIntosh (Stanford Univ.)
Niru Maheswaranathan (Stanford Univ.)
Aran Nayebi (Stanford Univ.)
Surya Ganguli (Stanford)
Stephen Baccus (Stanford Univ.)

A central challenge in sensory neuroscience is to understand neural computations and circuit mechanisms that underlie the encoding of ethologically relevant, natural stimuli. In multilayered neural circuits, nonlinear processes such as synaptic transmission and spiking dynamics present a significant obstacle to the creation of accurate computational models of responses to natural stimuli. Here we demonstrate that deep convolutional neural networks (CNNs) capture retinal responses to natural scenes nearly to within the variability of a cell's response, and are markedly more accurate than linear-nonlinear (LN) models. Moreover, we find two additional surprising properties of CNNs: they are less susceptible to overfitting than their LN counterparts when trained on small amounts of data, and generalize better when tested on stimuli drawn from a different distribution (e.g. between natural scenes and white noise). An examination of the learned parameters of CNNs indicates that a richer set of feature maps is necessary for predicting the responses of natural scenes compared to white noise. Visualizing the activity of the internal units in the network reveals that CNNs generate temporally precise responses from slowly varying inputs using feedforward inhibition, similar to known retinal mechanisms. Finally, by injecting latent noise sources in the intermediate layers during training, these models were able to capture both the precision and aspects of the uncertainty of the neural response. These methods can be readily generalized to other sensory modalities and stimulus ensembles. Overall, this work demonstrates that CNNs not only accurately capture sensory circuit responses to natural scenes, but also can yield information about the circuit's internal structure and function.

#151 Safe and Efficient Off-Policy Reinforcement Learning

Remi Munos (Google DeepMind)
Tom Stepleton (Google DeepMind)
Anna Harutyunyan (Vrije Universiteit Brussel)
Marc Bellemare (Google DeepMind)

In this work, we take a fresh look at some old and new algorithms for off-policy, return-based reinforcement learning. Expressing these in a common form, we derive a novel algorithm, Retrace(lambda), with three desired properties: (1) low variance; (2) safety, as it safely uses samples collected from any behaviour policy, whatever its degree of "off-policyness"; and (3) efficiency, as it makes the best use of samples collected from near on-policy behaviour policies. We analyse the contractive nature of the related operator under both policy evaluation and control settings and derive online sample-based algorithms. To our knowledge, this is the first return-based off-policy control algorithm converging a.s. to Q* that does not require the GLIE assumption (Greedy in the Limit with Infinite Exploration). As a corollary, we prove the convergence of Watkins' Q(lambda), which was still an open problem. We illustrate the benefits of Retrace(lambda) on a standard suite of Atari 2600 games.

#152 Yggdrasil: An Optimized System for Training Deep Decision Trees at Scale

Firas Abuzaid (MIT)
Joseph K Bradley (Databricks)
Feynman T Liang (Cambridge Univ. Engineering Department)
Andrew Feng (Yahoo!)
Lee Yang (Yahoo!)
Matei Zaharia (MIT)
Ameet S Talwalkar

Deep distributed decision trees and tree ensembles have grown in importance due to the increasing need to model high-dimensional data. However, PLANET, the standard distributed tree learning algorithm implemented in systems such as XGBoost and Spark MLlib, scales poorly as data dimensionality and tree depths grow. We present Yggdrasil, a new distributed tree learning method that outperforms existing methods by up to 24x. Unlike PLANET, Yggdrasil is based on vertical partitioning of the data (i.e., partitioning by feature), along with a set of optimized data structures that reduce both the CPU and communication cost of training. In particular, Yggdrasil (1) trains on compressed data without decompression; (2) introduces efficient data structures for training on uncompressed data; and and (3) minimizes communication between nodes by using sparse bitvectors. Moreover, while PLANET approximates split points through feature binning, Yggdrasil does not require binning, and we theoretically characterize the impact of this approximation. Yggdrasil is already in production use at a large Web company, and we show that it achieves order-of-magnitude speedups on high-dimensional datasets.

#153 Sample Complexity of Automated Mechanism Design

Maria-Florina Balcan
Tuomas Sandholm (Carnegie Mellon Univ.)
Ellen Vitercik (Carnegie Mellon Univ.)

The design of revenue-maximizing combinatorial auctions, i.e. multi-item auctions over bundles of goods, is one of the most fundamental problems in computational economics, unsolved even for two bidders and two items for sale. In the traditional economic models, it is assumed that the bidders' valuations are drawn from an underlying distribution and that the auction designer has perfect knowledge of this distribution. Despite this strong and oftentimes unrealistic assumption, it is remarkable that the revenue-maximizing combinatorial auction remains unknown. In recent years, automated mechanism design has emerged as one of the most practical and promising approaches to designing high-revenue combinatorial auctions. The most scalable automated mechanism design algorithms take as input samples from the bidders' valuation distribution and then search for a high-revenue auction in a rich auction class. In this work, we provide the first sample complexity analysis for the standard hierarchy of deterministic combinatorial auction classes used in automated mechanism design. In particular, we provide tight sample complexity bounds on the number of samples needed to guarantee that the empirical revenue of the designed mechanism on the samples is close to its expected revenue on the underlying, unknown distribution over bidder valuations, for each of the auction classes in the hierarchy. In addition to helping set automated mechanism design on firm foundations, our results also push the boundaries of learning theory. In particular, the hypothesis functions used in our contexts are defined through multi-stage combinatorial optimization procedures, rather than simple decision boundaries, as are common in machine learning.

#154 Deep Exploration via Bootstrapped DQN

Ian Osband (DeepMind)
Charles Blundell (DeepMind)
Alexander Pritzel
Benjamin Van Roy

Efficient exploration remains a major challenge for reinforcement learning (RL). Common dithering strategies for exploration, such as epsilon-greedy, do not carry out temporally-extended (or deep) exploration; this can lead to exponentially larger data requirements. However, most algorithms for statistically efficient RL are not computationally tractable in complex environments. Randomized value functions offer a promising approach to efficient exploration with generalization, but existing algorithms are not compatible with nonlinearly parameterized value functions. As a first step towards addressing such contexts we develop bootstrapped DQN. We demonstrate that bootstrapped DQN can combine deep exploration with deep neural networks for exponentially faster learning than any dithering strategy. In the Arcade Learning Environment bootstrapped DQN substantially improves learning speed and cumulative performance across most games.

#155 Search Improves Label for Active Learning

Alina Beygelzimer (Yahoo Inc)
Daniel Hsu
John Langford
Chicheng Zhang (UCSD)

We investigate active learning with access to two distinct oracles: LABEL (which is standard) and SEARCH (which is not). The SEARCH oracle models the situation where a human searches a database to seed or counterexample an existing solution. SEARCH is stronger than LABEL while being natural to implement in many situations. We show that an algorithm using both oracles can provide exponentially large problem-dependent improvements over LABEL alone.

#156 Efficient and Robust Spiking Neural Circuit for Navigation Inspired by Echolocating Bats

Bipin Rajendran (NJIT)
Pulkit Tandon (IIT Bombay)
Yash H Malviya (IIT Bombay)

We demonstrate a spiking neural circuit for azimuth angle detection inspired by the echolocation circuits of the Horseshoe bat Rhinolophus ferrumequinum and utilize it to devise a model for navigation and target tracking, capturing several key aspects of information transmission in biology. Our network, using only a simple local-information based sensor implementing the cardioid angular gain function, operates at biological spike rate of 10 Hz. The network tracks large angular targets (60 degrees) within 1 sec with a 10% RMS error. We study the navigational ability of our model for foraging and target localization tasks in a forest of obstacles and show that our network requires less than 200X spike-triggered decisions, while suffering only a 1% loss in performance compared to a proportional-integral-derivative controller, in the presence of 50% additive noise. Superior performance can be obtained at a higher average spike rate of 100 Hz and 1000 Hz, but even the accelerated networks requires 20X and 10X lesser decisions respectively, demonstrating the superior computational efficiency of bio-inspired information processing systems.

#157 Theoretical Comparisons of Positive-Unlabeled Learning against Positive-Negative Learning

Gang Niu (Univ. of Tokyo)
Marthinus Christoffel du Plessis
Tomoya Sakai
Yao Ma
Masashi Sugiyama (RIKEN / Univ. of Tokyo)

In PU learning, a binary classifier is trained from positive (P) and unlabeled (U) data without negative (N) data. Although N data is missing, it sometimes outperforms PN learning (i.e., ordinary supervised learning). Hitherto, neither theoretical nor experimental analysis has been given to explain this phenomenon. In this paper, we theoretically compare PU (and NU) learning against PN learning based on the upper bounds of estimation errors. We find simple conditions when PU and NU learning are likely to outperform PN learning, and we prove that, in terms of the upper bounds, either PU or NU learning (depending on the class-prior probability and the sizes of P and N data) given infinite U data will improve on PN learning. Our theoretical findings well agree with the experimental results on artificial and benchmark data even when the experimental setup does not match the theoretical assumptions exactly.

#158 Quantized Random Projections and Non-Linear Estimation of Cosine Similarity

Ping Li
Michael Mitzenmacher (Harvard Univ.)
Martin Slawski

Random projections constitute a simple, yet effective technique for dimensionality reduction with applications in learning and search problems. In the present paper, we consider the problem of estimating cosine similarities when the projected data undergo scalar quantization to b bits. We here argue that the maximum likelihood estimator (MLE) is a principled approach to deal with the non-linearity resulting from quantization, and subsequently study its computational and statistical properties. A specific focus is on the on the trade-off between bit depth and the number of projections given a fixed budget of bits for storage or transmission. Along the way, we also touch upon the existence of a qualitative counterpart to the Johnson-Lindenstrauss lemma in the presence of quantization.

#159 CNNpack: Packing Convolutional Neural Networks in the Frequency Domain

Yunhe Wang (Peking Univ.)
Chang Xu (Peking Univ.)
Shan You
Dacheng Tao
Chao Xu

Deep convolutional neural networks (CNNs) are successfully used in a number of applications. However, their storage and computational requirements have largely prevented their widespread use on mobile devices. Here we present an effective CNN compression approach in the frequency domain, which focuses not only on smaller weights but on all the weights and their underlying connections. By treating convolutional filters as images, we decompose their representations in the frequency domain as common parts (i.e., cluster centers) shared by other similar filters and their individual private parts (i.e., individual

residuals). A large number of low-energy frequency coefficients in both parts can be discarded to produce high compression without significantly compromising accuracy. We relax the computational burden of convolution operations in CNNs by linearly combining the convolution responses of discrete cosine transform (DCT) bases. The compression and speed-up ratios of the proposed algorithm are thoroughly analyzed and evaluated on benchmark image datasets to demonstrate its superiority over state-of-the-art methods.

#160 Verification Based Solution for Structured MAB Problems

Zohar Karnin

We consider the problem of finding the best arm in a stochastic Mutli-armed Bandit (MAB) game and propose a general framework based on verification that applies to multiple well-motivated generalizations of the classic MAB problem. In these generalizations, additional structure is known in advance, causing the task of verifying the optimality of a candidate to be easier than discovering the best arm. Our results are focused on the scenario where the failure probability δ must be very low; we essentially show that in this high confidence regime, identifying the best arm is as easy as the task of verification. We demonstrate the effectiveness of our framework by applying it, and improving the state-of-the art results in the problems of: Linear bandits, Dueling bandits with the Condorcet assumption, Copeland dueling bandits, Unimodal bandits and Graphical bandits.

#161 Neurally-Guided Procedural Models: Amortized Inference for Procedural Graphics Programs using Neural Networks

Daniel Ritchie (Stanford Univ.)
Anna Thomas (Stanford Univ.)
Pat Hanrahan (Stanford Univ.)
Noah Goodman

Probabilistic inference algorithms such as Sequential Monte Carlo (SMC) provide powerful tools for constraining procedural models in computer graphics, but they require many samples to produce desirable results. In this paper, we show how to create procedural models which learn how to satisfy constraints. We augment procedural models with neural networks which control how the model makes random choices based on the output it has generated thus far. We call such models neurally-guided procedural models. As a pre-computation, we train these models to maximize the likelihood of example outputs generated via SMC. They are then used as efficient SMC importance samplers, generating high-quality results with very few samples. We evaluate our method on L-system-like models with image-based constraints. Given a desired quality threshold, neurally-guided models can generate satisfactory results up to 10x faster than unguided models.

#162 Edge-Exchangeable Graphs and Sparsity

Diana Cai (Univ. of Chicago)
Trevor Campbell (MIT)
Tamara Broderick (MIT)

Many popular network models rely on the assumption of (vertex) exchangeability, in which the distribution of the graph is invariant to relabelings of the vertices. However, the Aldous-Hoover theorem guarantees that these graphs are dense or empty with probability one, whereas many real-world graphs are sparse. We present an alternative notion of exchangeability for random graphs, which we call edge exchangeability, in which the distribution of a graph sequence is invariant to the order of the edges. We demonstrate that edge-exchangeable models, unlike models that are traditionally vertex exchangeable, can exhibit sparsity. To do so, we outline a general framework for graph generative models; by contrast to the pioneering work of Caron and Fox (2014), models within our framework are stationary across steps of the graph sequence. In particular, our model grows the graph by instantiating more latent atoms of a single random measure as the dataset size increases, rather than adding new atoms to the measure.

#163 Learning and Forecasting Opinion Dynamics in Social Networks

Abir De (IIT Kharagpur)
Isabel Valera
Niloy Ganguly (IIT Kharagpur)
Sourangshu Bhattacharya (IIT Kharagpur)
Manuel Gomez Rodriguez (MPI-SWS)

Social media and social networking sites have become a global pinboard for exposition and discussion of news, topics, and ideas, where social media users often update their opinions about a particular topic by learning from the opinions shared by their friends. In this context, can we learn a data-driven model of opinion dynamics that is able to accurately forecast users' opinions? In this paper, we introduce SLANT, a probabilistic modeling framework of opinion dynamics, which represents users' opinions over time by means of marked jump diffusion stochastic differential equations, and allows for efficient model simulation and parameter estimation from historical fine grained event data. We then leverage our framework to derive a set of efficient predictive formulas for opinion forecasting and identify conditions under which opinions converge to a steady state. Experiments on data gathered from Twitter show that our model provides a good fit to the data and our formulas achieve more accurate forecasting than alternatives.

#164 Probing the Compositionality of Intuitive Functions

Eric Schulz (Univ. College London)
Josh Tenenbaum
David Duvenaud
Maarten Speekenbrink (Univ. College London)
Samuel J Gershman

How do people learn about complex functional structure? Taking inspiration from other areas of cognitive science, we propose that this is accomplished by harnessing compositionality: complex structure is decomposed into simpler building blocks. We formalize this idea within the framework of Bayesian regression using a grammar over Gaussian process kernels. We show that participants prefer compositional over non-compositional function extrapolations, that samples from the human prior over functions are best described by a compositional model, and that people perceive compositional functions as more predictable than their non-compositional but otherwise similar counterparts. We argue that the compositional nature of intuitive functions is consistent with broad principles of human cognition.

#165 Learning shape correspondence with anisotropic convolutional neural networks

Davide Boscaini (Univ. of Lugano)
Jonathan Masci
Emanuele Rodolà (Univ. of Lugano)
Michael Bronstein (Univ. of Lugano)

Convolutional neural networks have achieved extraordinary results in many computer vision and pattern recognition applications; however, their adoption in the computer graphics and geometry processing communities is limited due to the non-Euclidean structure of their data. In this paper, we propose Anisotropic Convolutional Neural Network (ACNN), a generalization of classical CNNs to non-Euclidean domains, where classical convolutions are replaced by projections over a set of oriented anisotropic diffusion kernels. We use ACNNs to effectively learn intrinsic dense correspondences between deformable shapes, a fundamental problem in geometry processing, arising in a wide variety of applications. We tested ACNNs performance in very challenging settings, achieving state-of-the-art results on some of the most difficult recent correspondence benchmarks.

#166 Improved Techniques for Training GANs

Tim Salimans
Ian Goodfellow (OpenAI)
Wojciech Zaremba (OpenAI)
Vicki Cheung (OpenAI)
Alec Radford (OpenAI)
Xi Chen (UC Berkeley and OpenAI)

We present a variety of new architectural features and training procedures that we apply to the generative adversarial networks (GANs) framework. We focus on two applications of GANs: semi-supervised learning, and the generation of images that humans find visually realistic. Unlike most work on generative models, our primary goal is not to train a model that assigns high likelihood to test data, nor do we require the model to be able to learn well without using any labels. Using our new techniques, we achieve state-of-the-art results in semi-supervised classification on MNIST, CIFAR-10 and SVHN. The generated images are of high quality as confirmed by a visual Turing test: our model generates MNIST samples that humans cannot distinguish from real data, and CIFAR-10 samples that yield a human error rate of 21.3%. We also present ImageNet samples with unprecedented resolution and show that our methods enable the model to learn recognizable features of ImageNet classes.

#167 Automated scalable segmentation of neurons from multispectral images

Uygar Sümbül (Columbia Univ.)
Douglas Roossien (Univ. of Michigan)
Dawen Cai (Univ. of Michigan)
John Cunningham
Liam Paninski

Reconstruction of neuroanatomy is a fundamental problem in neuroscience. Stochastic expression of colors in individual cells is a promising tool, although its use in the nervous system has been limited due to various sources of variability in expression. Moreover, the intermingled anatomy of neuronal trees is challenging for existing segmentation algorithms. Here, we propose a method to automate the segmentation of neurons in such (potentially pseudo-colored) images. The method uses spatio-color relations between the voxels, reduces the problem size by four orders of magnitude before the final segmentation, and is scalable. To quantify performance and gain insight, we generate simulated images, where the noise level and characteristics, the density of expression, and the number of fluorophore types are variable. Our segmentations achieve adjusted Rand scores around 0.75 on simulated multispectral images of retinal ganglion cells with realistic expression densities and neuron counts. We also present segmentation results of an actual Brainbow tissue obtained from the mouse hippocampus, which reveals many of the dendritic segments.

#168 Optimal Cluster Recovery in the Labeled Stochastic Block Model

Se-Young Yun (Los Alamos National Laboratory)
Alexandre Proutiere

We consider the problem of community detection or clustering in the labeled Stochastic Block Model (LSBM) with a finite number k of clusters of sizes linearly growing with the global population of items n. Every pair of items is labeled independently at random, and label ℓ appears with probability p(i,j,ℓ) between two items in clusters indexed by i and j, respectively. The objective is to reconstruct the clusters from the observation of these random labels. Clustering under the SBM and their extensions has attracted much attention recently. Most existing work aimed at characterizing the set of parameters such that it is possible to infer clusters either positively correlated with the true clusters, or with a vanishing proportion of misclassified items, or exactly matching the true clusters. We find the set of parameters such that there exists a clustering algorithm with at most s misclassified items in average under the general LSBM and for any s=o(n), which solves one open problem raised in \cite{abbe2015community}. We further develop an algorithm, based on simple spectral methods, that achieves this fundamental performance limit within O(n \mbox{polylog}(n)) computations and without the a-priori knowledge of the model parameters.

#169 Phased Exploration with Greedy Exploitation in Stochastic Combinatorial Partial Monitoring Games

Sougata Chaudhuri (Univ. of Michigan)
Ambuj Tewari (Univ. of Michigan)

Partial monitoring games are repeated games where the learner receives feedback that might be different from adversary's move or even the reward gained by the learner. Recently, a general model of combinatorial partial monitoring (CPM) games was proposed \ cite{lincombinatorial2014}, where the learner's action space can be exponentially large and adversary samples its moves from a bounded, continuous space, according to a fixed distribution. The paper gave a confidence bound based algorithm (GCB) that achieves O(T^{2/3} log T) distribution independent and O(log T) distribution dependent regret bounds. The implementation of their algorithm depends on two separate offline oracles and the distribution dependent regret additionally requires existence of a unique optimal action for the learner. Adopting their CPM model, our first contribution is a Phased Exploration with Greedy Exploitation (PEGE) algorithmic framework for the problem. Different algorithms within the framework achieve O(T^{2/3}√{log T}) distribution independent and O(log^2 T) distribution dependent regret respectively. Crucially, our framework needs only the simpler ``argmax" oracle from GCB and the distribution dependent regret does not require existence of a unique optimal action. Our second contribution is another algorithm, PEGE2, which combines gap estimation with a PEGE algorithm, to achieve an O(log T) regret bound, matching the GCB guarantee but removing the dependence on size of the learner's action space. However, like GCB, PEGE2 requires access to both offline oracles and the existence of a unique optimal action. Finally, we discuss how our algorithm can be efficiently applied to a CPM problem of practical interest: namely, online ranking with feedback at the top.

#170 Dual Space Gradient Descent for Online Learning

Trung Le (Univ. of Pedagogy Ho Chi Minh city)
Tu Nguyen (Deakin Univ.)
Vu Nguyen (Deakin Univ.)
Dinh Phung (Deakin Univ.)

One crucial goal in kernel online learning is to bound the model size. Common approaches employ budget maintenance procedures to restrict the model sizes using removal, projection, or merging strategies. Although projection and merging, in the literature, are known to be the most effective strategies, they demand extensive computation whilst removal strategy fails to retain information of the removed vectors. An alternative way to address the model size problem is to apply random features to approximate the kernel function. This allows the model to be maintained directly in the random feature space, hence effectively resolve the curse of kernelization. However, this approach still suffers from a serious shortcoming as it needs to use a high dimensional random feature space to achieve a sufficiently accurate kernel approximation. Consequently, it leads to a significant increase in the computational cost. To address all of these aforementioned challenges, we present in this paper the Dual Space Gradient Descent (DualSGD), a novel framework that utilizes random features as an auxiliary space to maintain information from data points removed during budget maintenance. Consequently, our approach permits the budget to be maintained in a simple, direct and elegant way while simultaneously mitigating the impact of the dimensionality issue on learning performance. We further provide convergence analysis and extensively conduct experiments on five real-world datasets to demonstrate the predictive performance and scalability of our proposed method in comparison with the state-of-the-art baselines.

#171 Data Programming: Creating Large Training Sets, Quickly

Alexander J Ratner (Stanford Univ.)
Christopher M De Sa (Stanford Univ.)
Sen Wu (Stanford Univ.)
Daniel Selsam (Stanford)
Christopher Ré (Stanford Univ.)

Large labeled training sets are the critical building blocks of supervised learning methods and are key enablers of deep learning techniques. For some applications, creating labeled training sets is the most time-consuming and expensive part of applying machine learning. We therefore propose a paradigm for the programmatic creation of training sets called data programming in which users provide a set of labeling functions, which are programs that heuristically label large subsets of data points, albeit noisily. By viewing these labeling functions as implicitly describing a generative model for this noise, we show that we can recover the parameters of this model to "denoise" the training set. Then, we show how to modify a discriminative loss function to make it noise-aware. We demonstrate our method over a range of discriminative models including logistic regression and LSTMs. We establish theoretically that we can recover the parameters of these generative models in a handful of settings. Experimentally, on the 2014 TAC-KBP relation extraction challenge, we show that data programming would have obtained a winning score, and also show that applying data programming to an LSTM model leads to a TAC-KBP score almost 6 F1 points over a supervised LSTM baseline (and into second place in the competition). Additionally, in initial user studies we observed that data programming may be an easier way to create machine learning models for non-experts.

#172 Near-Optimal Smoothing of Structured C onditional Probability Matrices

Moein Falahatgar (UCSD)
Mesrob I Ohannessian
Alon Orlitsky

When learning multiple related probability distributions, it is imperative to use the underlying relationship or structure in order not to dilute the data. This paper considers the case of low-rank structure, a paradigm that has been widely successful in various learning problems. Motivated by language modeling, and in particular bigram models, the performance measure adopted is an appropriately averaged pairwise KL-risk. This choice makes smoothing, that is the careful handling of low-probability categories, paramount. We provide an iterative algorithm that extends classical non-negative matrix factorization to naturally incorporate smoothing. We then give sample complexity bounds for this simple algorithm, and show that it is within a small factor of the optimal sample complexity.

#173 An urn model for majority voting in classification ensembles

Victor Soto (Columbia Univ.)
Alberto Suárez
Gonzalo Martinez-Muñoz

In this work we analyze the class prediction by majority voting in parallel randomized ensembles as an urn model. For a given test instance, the ensemble can be viewed as an urn of marbles of different colors. A marble represents an individual classifier. Its color represents the class label prediction of the corresponding classifier. The sequential querying of classifiers in the ensemble can be seen as draws without replacement from the urn. An analysis of this classical urn model based on the hypergeometric distribution makes it possible estimate the confidence on the outcome of majority voting when only a fraction of the individual predictions are known. These estimates can be used to speed up the prediction by the ensemble. Specifically, the aggregation of votes can be halted when the confidence in the final prediction is sufficiently high. If one assumes a uniform prior for the distribution of possible votes the analysis is shown to be equivalent to a previous one based on Dirichlet distributions. The advantage of the current approach is that prior knowledge on the possible vote outcomes can be readily incorporated in a Bayesian framework. We show how incorporating this type of problem-specific knowledge into the statistical analysis of majority voting leads to faster classification by the ensemble and permits to easily estimate the expected average speed-up beforehand.

TUESDAY SESSIONS

9:00 - 9:50 am - INVITED TALK:
Intelligent Biosphere - Drew Purves Area 1 + 2

9:50 - 10:10 am - AWARD TALK:
Value Iteration Networks Area 1 + 2
Aviv Tamar, Sergey Levine, Pieter Abbeel, YI WU, Garrett Thomas

10:10 - 10:40 am - Coffee Break - P1 & P2

10:40 - 12:20 pm - Track 1 - Clustering Area 3

- **Graphons, mergeons, and so on!**
  Justin Eldridge, Mikhail Belkin, Yusu Wang
- **Hierarchical Clustering via Spreading Metrics**
  Aurko Roy, Sebastian Pokutta
- **Clustering with Same-Cluster Queries**
  Hassan Ashtiani, Shrinu Kushagra, Shai Ben-David
- **Unsupervised Feature Extraction by Time-Contrastive Learning and Nonlinear CA**
  Aapo Hyvarinen, Hiroshi Morioka
- **Fast and Provably Good Seedings for k-Means**
  Olivier Bachem, Mario Lucic, Hamed Hassani, Andreas Krause

10:40 - 12:20 pm - Track 2 - Graphical Models Area 1 + 2

- **Tractable Operations for Arithmetic Circuits of Probabilistic Models**
  Yujia Shen, Arthur Choi, Adnan Darwiche
- **Testing for Differences in Gaussian Graphical Models: Applications to Brain Connectivity**
  Eugene Belilovsky, Gaël Varoquaux, Matthew B Blaschko
- **SDP Relaxation with Randomized Rounding for Energy Disaggregation**
  Kiarash Shaloudegi, András György, Csaba Szepesvari, Wilsun Xu
- **Bayesian Intermittent Demand Forecasting for Large Inventories**
  Matthias W Seeger, David Salinas, Valentin Flunkert
- **Synthesis of MCMC and Belief Propagation**
  Sung-Soo Ahn, Michael Chertkov, Jinwoo Shin

12:20 - 3:00 pm - LUNCH ON YOUR OWN

3:00 - 3:50 pm - INVITED TALK:
Engineering Principles From Stable and Developing Brains
Saket Navlakha Area 1 + 2

3:50 - 4:20 pm - Coffee Break - P1 & P2

4:20 - 5:40 pm - Track 1 - Deep Learning Area 1 + 2

- **Deep Learning for Predicting Human Strategic Behavior**
  Jason S Hartford, James R Wright, Kevin Leyton-Brown
- **Using Fast Weights to Attend to the Recent Past**
  Jimmy Ba, Geoffrey E Hinton, Volodymyr Mnih, Joel Z Leibo, Catalin Ionescu
- **Sequential Neural Models with Stochastic Layers**
  Marco Fraccaro, Søren Kaae Sønderby, Ulrich Paquet, Ole Winther
- **Phased LSTM: Accelerating Recurrent Network Training for Long or Event-based Sequences**
  Daniel Neil, Michael Pfeiffer, Shih-Chii Liu

4:20 - 5:40 pm - Track 2 - Theory Area 3

- **Supervised learning through the lens of compression**
  Ofir David, Shay Moran, Amir Yehudayoff
- **MetaGrad: Multiple Learning Rates in Online Learning**
  Tim van Erven, Wouter M Koolen
- **Blazing the trails before beating the path: Sample-efficient Monte-Carlo planning**
  Jean-Bastien Grill, Michal Valko, Remi Munos
- **Global Analysis of Expectation Maximization for Mixtures of Two Gaussians**
  Ji Xu, Daniel Hsu,

Tuesday, Dec 6th, 9 - 9:50 am

Intelligent Biosphere
Area 1 & 2

Drew Purves (DeepMind)

The biosphere is a stupendously complex and poorly understood system, which we depend on for our survival, and which we are attacking on every front. Worrying. But what has that got to do with machine learning and AI? I will explain how the complexity and stability of the entire biosphere depend on, and select for, the intelligence of the individual organisms that comprise it; why simulations of ecological tasks in naturalistic environments could be an important test bed for Artificial General Intelligence, AGI; how new technology and machine learning are already giving us a deeper understanding of life on Earth; and why AGI is needed to maintain the biosphere in a state that is compatible with the continued existence of human civilization.

A teenage interest in the emergent dynamics of self-interested, evolving, interacting agents, sparked by the Artificial Life movement, was Drew's route into studying real ecology at Cambridge, York, and Princeton. Throughout, his focus was on developing realistic simulation models of ecological processes, something that he was able to scale up hugely during his 8 years as head of the Computational Ecology and Environmental Science group (CEES) at Microsoft Research, which developed many such models, at spatiotemporal scales from millimetres to global, seconds to centuries. CEES built the first fully data-constrained model of the global carbon cycle, and The Madingley Model, which simulates the key ecological interactions among nearly all macroorganisms on Earth. From a technical perspective, CEES specialized in Bayesian approaches to constraining esoteric nonlinear ecological models to heterogeneous data, developing new methods and software tools to facilitate such an approach, from algorithms such as Filzbach, to geotemporal software such as FetchClimate. In November 2015, after 20 years devoted to ecological research, Purves changed tack to join DeepMind's mission to create General Artificial Intelligence.

Tuesday, Dec 6th, 9:50 - 10:10 am

Award Talk: Value Iteration Networks
Area 1 & 2

Aviv Tamar (UC Berkely)
Sergey Levine (UC Berkely)
Pieter Abbeel (OpenAI / UC Berkeley / Gradescope)
YI WU (UC Berkely)
Garrett Thomas (UC Berkely)

We introduce the value iteration network (VIN): a fully differentiable neural network with a `planning module' embedded within. VINs can learn to plan, and are suitable for predicting outcomes that involve planning-based reasoning, such as policies for reinforcement learning. Key to our approach is a novel differentiable approximation of the value-iteration algorithm, which can be represented as a convolutional neural network, and trained end-to-end using standard backpropagation. We evaluate VIN based policies on discrete and continuous path-planning domains, and on a natural-language based search task. We show that by learning an explicit planning computation, VIN policies generalize better to new, unseen domains.

Tuesday, Dec 6th, 3 - 3:50 pm

Engineering Principles From Stable and Developing Brains
Area 1 & 2

Saket Navlakha
(The Salk Institute for Biological Studies)

Robust, efficient, and low-cost networks are advantageous in both biological and engineered systems. First, I will describe a joint computational-experimental approach to explore how neural networks in the brain form during development. I will discuss how the brain uses a very uncommon and surprising strategy to build networks and how this idea can be used to enhance the design and function of energy-efficient distributed networks. Second, I will describe how two fundamental plasticity rules (LTP and LTD) help neural networks approach desirable synaptic weight distributions in a gradient-descent-like manner. I will derive connections between different experimentally-derived forms of these rules and distributed algorithms commonly used to regulate traffic flow on the Internet. Our work is motivated by the study of "algorithms in nature".

Saket Navlakha is an assistant professor at the Salk Institute for Biological Studies.He received an A.A. from Simon's Rock College in 2002, a B.S. from Cornell University in 2005, and a Ph.D. in computer science from the University of Maryland College Park in 2010. He was a post-doctoral researcher in the Machine Learning Department at Carnegie Mellon University from 2011-2014. His research interests include the design of algorithms for understanding large biological networks and the study of algorithms in nature.

Track 1 - 10:40 am - 12:20 pm - Area 3 Clustering

Graphons, mergeons, and so on!

Justin Eldridge (The Ohio State University)
Mikhail Belkin
Yusu Wang (The Ohio State University)

In this work we develop a theory of hierarchical clustering for graphs. Our modelling assumption is that graphs are sampled from a graphon, which is a powerful and general model for generating graphs and analyzing large networks. Graphons are a far richer class of graph models than stochastic blockmodels, the primary setting for recent progress in the statistical theory of graph clustering. We define what it means for an algorithm to produce the ``correct'' clustering, give sufficient conditions in which a method is statistically consistent, and provide an explicit algorithm satisfying these properties.

Hierarchical Clustering via Spreading Metrics

Aurko Roy (Georgia Tech)
Sebastian Pokutta (GeorgiaTech)

We study the cost function for hierarchical clusterings introduced by [Dasgupta, 2015] where hierarchies are treated as first-class objects rather than deriving their cost from projections into flat clusters. It was also shown in [Dasgupta, 2015] that a top-down algorithm returns a hierarchical clustering of cost at most \(O\left(\alpha_n log n\ right)\) times the cost of the optimal hierarchical clustering, where \(\alpha_n\) is the approximation ratio of the Sparsest Cut subroutine used. Thus using the best known approximation algorithm for Sparsest Cut due to Arora-Rao-Vazirani, the top down algorithm returns a hierarchical clustering of cost at most \(O\left(log^{3/2} n\ right)\) times the cost of the optimal solution. We improve this by giving an \(O(log{n})\)-approximation algorithm for this problem. Our main technical ingredients are a combinatorial characterization of ultrametrics induced by this cost function, deriving an Integer Linear Programming (ILP) formulation for this family of ultrametrics, and showing how to iteratively round an LP relaxation of this formulation by using the idea of {sphere growing} which has been extensively used in the context of graph partitioning. We also prove that our algorithm returns an \(O(log{n})\)-approximate hierarchical clustering for a generalization of this cost function also studied in [Dasgupta, 2015]. Experiments show that the hierarchies found by using the ILP formulation as well as our rounding algorithm often have better projections into flat clusters than the standard linkage based algorithms. We conclude with an inapproximability result for this problem, namely that no polynomial sized LP or SDP can be used to obtain a constant factor approximation for this problem.

Clustering with Same-Cluster Queries

Hassan Ashtiani (University of Waterloo)
Shrinu Kushagra (University of Waterloo)
Shai Ben-David (U. Waterloo)

We propose a framework for Semi-Supervised Active Clustering framework (SSAC), where the learner is allowed to interact with a domain expert, asking whether two given instances belong to the same cluster or not. We study the query and computational complexity of clustering in this framework. We consider a setting where the expert conforms to a center-based clustering with a notion of margin. We show that there is a trade off between computational complexity and query complexity; We prove that for the case of k-means clustering (i.e., when the expert conforms to a solution of k-means), having access to relatively few such queries allows efficient solutions to otherwise NP hard problems. In particular, we provide a probabilistic polynomial-time (BPP) algorithm for clustering in this setting that asks $\mathrm{O}(\mathrm{k}^2\mathrm{log}\mathrm{k} + \mathrm{klog}\ \mathrm{n})$ same-cluster queries and runs with time complexity $\mathrm{O}(\mathrm{klog}\ \mathrm{n})$ (where k is the number of clusters and n is the number of instances). The success of the algorithm is guaranteed for data satisfying the margin condition under which, without queries, we show that the problem is NP hard. We also prove a lower bound on the number of queries needed to have a computationally efficient clustering algorithm in this setting.

Unsupervised Feature Extraction by Time-Contrastive Learning and Nonlinear ICA

Aapo Hyvarinen
Hiroshi Morioka (University of Helsinki)

Nonlinear independent component analysis (ICA) provides an appealing framework for unsupervised feature learning, but the models proposed so far are not identifiable. Here, we first propose a new intuitive principle of unsupervised deep learning from time series which uses the nonstationary structure of the data. Our learning principle, time-contrastive learning (TCL), finds a representation which allows optimal discrimination of time segments (windows). Surprisingly, we show how TCL can be related to a nonlinear ICA model, when ICA is redefined to include temporal nonstationarities. In particular, we show that TCL combined with linear ICA estimates the nonlinear ICA model up to point-wise transformations of the sources, and this solution is unique --- thus providing the first identifiability result for nonlinear ICA which is rigorous, constructive, as well as very general.

Fast and Provably Good Seedings for k-Means

Olivier Bachem (ETH Zurich)
Mario Lucic (ETH Zurich)
Hamed Hassani (ETH Zurich)
Andreas Krause

Seeding - the task of finding initial cluster centers - is critical in obtaining high-quality clusterings for k-Means. However, k-means++ seeding, the state of the art algorithm, does not scale well to massive datasets as it is inherently sequential and requires k full passes through the data. It was recently shown that Markov chain Monte Carlo sampling can be used to efficiently approximate the seeding step of k-means++. However, this result requires assumptions on the data generating distribution. We propose a simple yet fast seeding algorithm that produces *provably* good clusterings even *without assumptions* on the data. Our analysis shows that the algorithm allows for a favourable trade-off between solution quality and computational cost, speeding up k-means++ seeding by up to several orders of magnitude. We validate our theoretical results in extensive experiments on a variety of real-world data sets.

Track 2 - 10:40 am - 12:20 pm- Area 1 + 2 Graphical Models

Tractable Operations for Arithmetic Circuits of Probabilistic Models

Yujia Shen
Arthur Choi
Adnan Darwiche

We consider tractable representations of probability distributions and the polytime operations they support. In particular, we consider a recently proposed arithmetic circuit representation, the Probabilistic Sentential Decision Diagram (PSDD). We show that PSDD supports a polytime multiplication operator, while they do not support a polytime operator for summing-out variables. A polytime multiplication operator make PSDDs suitable for a broader class of applications compared to arithmetic circuits, which do not in general support multiplication. As one example, we show that PSDD multiplication leads to a very simple but effective compilation algorithm for probabilistic graphical models: represent each model factor as a PSDD, and then multiply them.

Testing for Differences in Gaussian Graphical Models: Applications to Brain Connectivity

Eugene Belilovsky (CentraleSupelec)
Gaël Varoquaux
Matthew B Blaschko (KU Leuven)

Functional brain networks are well described and estimated from data with Gaussian Graphical Models (GGMs), e.g.\ using sparse inverse covariance estimators. Comparing functional connectivity of subjects in two populations calls for comparing these estimated GGMs. Our goal is to identify differences in GGMs known to have similar structure. We characterize the uncertainty of differences with confidence intervals obtained using a parametric distribution on parameters of a sparse estimator. Sparse penalties enable statistical guarantees and interpretable models even in high-dimensional and low-sample settings. Characterizing the distributions of sparse models is inherently challenging as the penalties produce a biased estimator. Recent work invokes the sparsity assumptions to effectively remove the bias from a sparse estimator such as the lasso. These distributions can be used to give confidence intervals on edges in GGMs, and by extension their differences. However, in the case of comparing GGMs, these estimators do not make use of any assumed joint structure among the GGMs. Inspired by priors from brain functional connectivity we derive the distribution of parameter differences under a joint penalty when parameters are known to be sparse in the difference. This leads us to introduce the debiased multi-task fused lasso, whose distribution can be characterized in an efficient manner. We then show how the debiased lasso and multi-task fused lasso can be used to obtain confidence intervals on edge differences in GGMs. We validate the techniques proposed on a set of synthetic examples as well as neuro-imaging dataset created for the study of autism.

SDP Relaxation with Randomized Rounding for Energy Disaggregation

Kiarash Shaloudegi
András György
Csaba Szepesvari (U. Alberta)
Wilsun Xu (University of Alberta)

We develop a scalable, computationally efficient method for the task of energy disaggregation for home appliance monitoring. In this problem the goal is to estimate the energy consumption of each appliance based on the total energy-consumption signal of a household. The current state of the art models the problem as inference in factorial HMMs, and finds an approximate solution to the resulting quadratic integer program via quadratic programming. Here we take a more principled approach, better suited to integer programming problems, and find an approximate optimum by combining convex semidefinite relaxations with randomized rounding, as well as with a scalable ADMM method that exploits the special structure of the resulting semidefinite program. Simulation results demonstrate the superiority of our methods both in synthetic and real-world datasets.

Bayesian Intermittent Demand Forecasting for Large Inventories

Matthias W Seeger (Amazon)
David Salinas (Amazon)
Valentin Flunkert (Amazon)

We present a scalable and robust Bayesian method for demand forecasting in the context of a large e-commerce platform, paying special attention to intermittent and bursty target statistics. Inference is approximated by the Newton-Raphson algorithm, reduced to linear-time Kalman smoothing, which allows us to operate on several orders of magnitude larger problems than previous related work. In a study on large real-world sales datasets, our method outperforms competing approaches on fast and medium moving items.

Synthesis of MCMC and Belief Propagation

Sung-Soo Ahn (KAIST)
Michael Chertkov (Los Alamos National Laboratory)
Jinwoo Shin (KAIST)

Markov Chain Monte Carlo (MCMC) and Belief Propagation (BP) are the most popular algorithms for computational inference in Graphical Models (GM). In principle, MCMC is an exact probabilistic method which, however, often suffers from exponentially slow mixing. In contrast, BP is a deterministic method, which is typically fast, empirically very successful, however in general lacking control of accuracy over loopy graphs. In this paper, we introduce MCMC algorithms correcting the approximation error of BP, i.e., we provide a way to compensate for BP errors via a consecutive BP-aware MCMC. Our framework is based on the Loop Calculus (LC) approach which allows to express the BP error as a sum of weighted generalized loops. Although the full series is computationally intractable, it is known that a truncated series, summing up all 2-regular loops, is computable in polynomial-time for planar pair-wise binary GMs and it also provides a highly accurate approximation empirically. Motivated by this, we, first, propose a polynomial-time approximation MCMC scheme for the truncated series of general (non-planar) pair-wise binary models. Our main idea here is to use the Worm algorithm, known to provide fast mixing in other (related) problems, and then design an appropriate rejection scheme to sample 2-regular loops. Furthermore, we also design an efficient rejection-free MCMC scheme for approximating the full series. The main novelty underlying our design is in utilizing the concept of cycle basis, which provides an efficient decomposition of the generalized loops. In essence, the proposed MCMC schemes run on transformed GM built upon the non-trivial BP solution, and our experiments show that this synthesis of BP and MCMC outperforms both direct MCMC and bare BP schemes.

Track 1 - 4:20 - 5:40 pm - Area 1 + 2
Deep Learning

Deep Learning for Predicting Human Strategic Behavior

Jason S Hartford (University of British Columbia)
James R Wright (University of British Columbia)
Kevin Leyton-Brown

Predicting the behavior of human participants in strategic settings is an important problem in many domains. Most existing work either assumes that participants are perfectly rational, or attempts to directly model each participant's cognitive processes based on insights from cognitive psychology and experimental economics. In this work, we present an alternative, a deep learning approach that automatically performs cognitive modeling without relying on such expert knowledge. We introduce a novel architecture that allows a single network to generalize across different input and output dimensions by using matrix units rather than scalar units, and show that its performance significantly outperforms that of the previous state of the art, which relies on expert-constructed features.

Using Fast Weights to Attend to the Recent Past

Jimmy Ba (University of Toronto)
Geoffrey E Hinton (Google)
Volodymyr Mnih
Joel Z Leibo (Google DeepMind)
Catalin Ionescu (Google)

Until recently, research on artificial neural networks was largely restricted to systems with only two types of variable: Neural activities that represent the current or recent input and weights that learn to capture regularities among inputs, outputs and payoffs. There is no good reason for this restriction. Synapses have dynamics at many different time-scales and this suggests that artificial neural networks might benefit from variables that change slower than activities but much faster than the standard weights. These ``fast weights'' can be used to store temporary memories of the recent past and they provide a neurally plausible way of implementing the type of attention to the past that has recently proven helpful in sequence-to-sequence models. By using fast weights we can avoid the need to store copies of neural activity patterns.

Sequential Neural Models with Stochastic Layers

Marco Fraccaro (DTU)
Søren Kaae Sønderby (KU)
Ulrich Paquet (DeepMind)
Ole Winther (DTU)

How can we efficiently propagate uncertainty in a latent state representation with recurrent neural networks? This paper introduces stochastic recurrent neural networks which glue a deterministic recurrent neural network and a state space model together to form a stochastic and sequential neural generative model. The clear separation of deterministic and stochastic layers allows a structured variational inference network to track the factorization of the model's posterior distribution. By retaining both the nonlinear recursive structure of a recurrent neural network and averaging over the uncertainty in a latent path, like a state space model, we improve the state of the art results on the Blizzard and TIMIT speech modeling data sets by a large margin, while achieving comparable performances to competing methods on polyphonic music modeling.

Phased LSTM: Accelerating Recurrent Network Training for Long or Event-based Sequences

Daniel Neil (Institute of Neuroinformatics)
Michael Pfeiffer (Institute of Neuroinformatics)
Shih-Chii Liu

Recurrent Neural Networks (RNNs) have become the state-of-the-art choice for extracting patterns from temporal sequences. Current RNN models are ill suited to process irregularly sampled data triggered by events generated in continuous time by sensors or other neurons. Such data can occur, for example, when the input comes from novel event-driven artificial sensors which generate sparse, asynchronous streams of events or from multiple conventional sensors with different update intervals. In this work, we introduce the Phased LSTM model, which extends the LSTM unit by adding a new time gate. This gate is controlled by a parametrized oscillation with a frequency range which require updates of the memory cell only during a small percentage of the cycle. Even with the sparse updates imposed by the oscillation, the Phased LSTM network achieves faster convergence than regular LSTMs on tasks which require learning of long sequences. The model naturally integrates inputs from sensors of arbitrary sampling rates, thereby opening new areas of investigation for processing asynchronous sensory events that carry timing information. It also greatly improves the performance of LSTMs in standard RNN applications, and does so with an order-of-magnitude fewer computes.

Track 2 - 4:20 - 5:40 pm - Area 3
Theory

Supervised learning through the lens of compression

Ofir David (Technion - Israel institute of technology)
Shay Moran (Technion - Israel institue of Technology)
Amir Yehudayoff (Technion - Israel institue of Technology)

This work continues the study of the relationship between sample compression schemes and statistical learning, which has been mostly investigated within the framework of binary classification. We first extend the investigation to multiclass categorization: we prove that in this case learnability is equivalent to compression of logarithmic sample size and that the uniform convergence property implies compression of constant size. We use the compressibility-learnability equivalence to show that (i) for multiclass categorization, PAC and agnostic PAC learnability are equivalent, and (ii) to derive a compactness theorem for learnability. We then consider supervised learning under general loss functions: we show that in this case, in order to maintain the compressibility-learnability equivalence, it is necessary to consider an approximate variant of compression. We use it to show that PAC and agnostic PAC are not equivalent, even when the loss function has only three values.

MetaGrad: Multiple Learning Rates in Online Learning

Tim van Erven
Wouter M Koolen

In online convex optimization it is well known that certain subclasses of objective functions are much easier than arbitrary convex functions. We are interested in designing adaptive methods that can automatically get fast rates in as many such subclasses as possible, without any manual tuning. Previous adaptive methods are able to interpolate between strongly convex and general convex functions. We present a new method, MetaGrad, that adapts to a much broader class of functions, including exp-concave and strongly convex functions, but also various types of stochastic and non-stochastic functions without any curvature. For instance, MetaGrad can achieve logarithmic regret on the unregularized hinge loss, even though it has no curvature, if the data come from a favourable probability distribution. MetaGrad's main feature is that it simultaneously considers multiple learning rates. Unlike all previous methods with provable regret guarantees, however, its learning rates are not monotonically decreasing over time and are not tuned based on a theoretically derived bound on the regret. Instead, they are weighted directly proportional to their empirical performance on the data using a tilted exponential weights master algorithm.

Blazing the trails before beating the path: Sample-efficient Monte-Carlo planning

Jean-Bastien Grill (Inria Lille - Nord Europe)
Michal Valko (Inria Lille - Nord Europe)
Remi Munos (Google DeepMind)

We study the sampling-based planning problem in Markov decision processes (MDPs) that we can access only through a generative model, usually referred to as Monte-Carlo planning. Our objective is to return a good estimate of the optimal value function at any state while minimizing the number of calls to the generative model, i.e. the sample complexity. We propose a new algorithm, TrailBlazer, able to handle MDPs with a finite or an infinite number of transitions from state-action to next states. TrailBlazer is an adaptive algorithm that exploits possible structures of the MDP by exploring only a subset of states reachable by following near-optimal policies. We provide bounds on its sample complexity that depend on a measure of the quantity of near-optimal states. The algorithm behavior can be considered as an extension of Monte-Carlo sampling (for estimating an expectation) to problems that alternate maximization (over actions) and expectation (over next states). Finally, another appealing feature of TrailBlazer is that it is simple to implement and computationally efficient.

Global Analysis of Expectation Maximization for Mixtures of Two Gaussians

Ji Xu (Columbia university)
Daniel Hsu
(Columbia University)

Expectation Maximization (EM) is among the most popular algorithms for estimating parameters of statistical models. However, EM, which is an iterative algorithm based on the maximum likelihood principle, is generally only guaranteed to find stationary points of the likelihood objective, and these points may be far from any maximizer. This article addresses this disconnect between the statistical principles behind EM and its algorithmic properties. Specifically, it provides a global analysis of EM for specific models in which the observations comprise an i.i.d. sample from a mixture of two Gaussians. This is achieved by (i) studying the sequence of parameters from idealized execution of EM in the infinite sample limit, and fully characterizing the limit points of the sequence in terms of the initial parameters; and then (ii) based on this convergence analysis, establishing statistical consistency (or lack thereof) for the actual sequence of parameters produced by EM.

Tuesday Poster Session

#1 **The Multi-fidelity Multi-armed Bandit**
Kirthevasan Kandasamy, Gautam Dasarathy, Barnabas Poczos, Jeff Schneider

#2 **Probabilistic Inference with Generating Functions for Poisson Latent Variable Models**
Kevin Winner, Dan Sheldon

#3 **Adaptive Maximization of Pointwise Submodular Functions With Budget Constraint**
Nguyen Cuong, Huan Xu

#4 **Machine Translation Through Learning From a Communication Game**
Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tieyan Liu, Wei-Ying Ma

#5 **Iterative Refinement of the Approximate Posterior for Directed Belief Networks**
devon Hjelm, Russ Salakhutdinov, Kyunghyun Cho, Nebojsa Jojic, Vince Calhoun, Junyoung Chung

#6 **Unsupervised Risk Estimation Using Only Conditional Independence Structure**
Jacob Steinhardt, Percy S Liang

#7 **Hierarchical Question-Image Co-Attention for Visual Question Answering**
Jiasen Lu, Jianwei Yang, Dhruv Batra, Devi Parikh

#8 **Bayesian Optimization with a Finite Budget: An Approximate Dynamic Programming Approach**
Remi Lam, Karen Willcox, David Wolpert

#9 **Learning to learn by gradient descent by gradient descent**
Marcin Andrychowicz, Misha Denil, Sergio Gómez, Matthew W Hoffman, David Pfau, Tom Schaul, Nando de Freitas

#10 **Computational and Statistical Tradeoffs in Learning to Rank**
Ashish Khetan, Sewoong Oh

#11 **Pairwise Choice Markov Chains**
Stephen Ragain, Johan Ugander

#12 **Incremental Learning for Variational Sparse Gaussian Process Regression**
Ching-An Cheng, Byron Boots

#13 **Combinatorial Multi-Armed Bandit with General Reward Functions**
Wei Chen, Wei Hu, Fu Li, Jian Li, Yu Liu, Pinyan Lu

#14 **Observational-Interventional Priors for Dose-Response Learning**
Ricardo Silva

#15 **On Graph Reconstruction via Empirical Risk Minimization: Fast Learning Rates and Scalability**
Guillaume Papa, Aurélien Bellet, Stephan Clémençon

#16 **DeepMath - Deep Sequence Models for Premise Selection**
Geoffrey Irving, Christian Szegedy, Alex A Alemi, Francois Chollet, Josef Urban

#17 **Efficient Second Order Online Learning by Sketching**
Haipeng Luo, Alekh Agarwal, Nicolò Cesa-Bianchi, John Langford

#18 **Gaussian Processes for Survival Analysis**
Tamara Fernandez, Nicolás Rivera, Yee Whye Teh

#19 **The Power of Optimization from Samples**
Eric Balkanski, Aviad Rubinstein, Yaron Singer

#20 **Global Optimality of Local Search for Low Rank Matrix Recovery**
Srinadh Bhojanapalli, Behnam Neyshabur, Nati Srebro

#21 **A state-space model of cross-region dynamic connectivity in MEG/EEG**
Ying Yang, Elissa Aminoff, Michael Tarr, Rob E Robert

#22 **Hypothesis Testing in Unsupervised Domain Adaptation with Applications in Neuroscience**
Hao Zhou, Vamsi K Ithapu, Sathya Narayanan Ravi, Vikas Singh, Grace Wahba, Sterling C Johnson

#23 **Bi-Objective Online Matching and Submodular Allocations**
Hossein Esfandiari, Nitish Korula, Vahab Mirrokni

#24 **A Constant-Factor Bi-Criteria Approximation Guarantee for k-means++**
Dennis Wei

#25 **Causal Bandits: Learning Good Interventions via Causal Inference**
Finnian Lattimore, Tor Lattimore, Mark Reid

#26 **Unsupervised Domain Adaptation with Residual Transfer Networks**
Mingsheng Long, Han Zhu, Jianmin Wang, Michael I Jordan

#27 **Data driven estimation of Laplace-Beltrami operator**
Frederic Chazal, Ilaria Giulini, Bertrand Michel

#28 **Fast Algorithms for Robust PCA via Gradient Descent**
Xinyang Yi, Dohyung Park, Yudong Chen, Constantine Caramanis

#29 **NESTT: A Nonconvex Primal-Dual Splitting Method for Distributed and Stochastic Optimization**
Davood Hajinezhad, Mingyi Hong, Tuo Zhao, Zhaoran Wang

#30 **Fundamental Limits of Budget-Fidelity Trade-off in Label Crowdsourcing**
Farshad Lahouti, Babak Hassibi

#31 **Supervised Learning with Tensor Networks**
Miles Stoudenmire, David J Schwab

#32 **Understanding Probabilistic Sparse Gaussian Process Approximations**
Matthias Bauer, Mark van der Wilk, Carl Edward Rasmussen

#33 **A Locally Adaptive Normal Distribution**
Georgios Arvanitidis, Lars K Hansen, Søren Hauberg

#34 **Anchor-Free Correlated Topic Modeling: Identifiability and Algorithm**
Kejun Huang, Xiao Fu, Nikos D. Sidiropoulos

#35 **Optimal Learning for Multi-pass Stochastic Gradient Methods**
Junhong Lin, Lorenzo Rosasco

#36 **Contextual semibandits via supervised learning oracles**
Akshay Krishnamurthy, Alekh Agarwal, Miro Dudik

#37 **One-vs-Each Approximation to Softmax for Scalable Estimation of Probabilities**
Michalis Titsias RC AUEB

#38 **Satisfying Real-world Goals with Dataset Constraints**
Gabe Goh, Andy Cotter, Maya Gupta, Michael P Friedlander

#39 **Blind Regression: Nonparametric Regression for Latent Variable Models via Collaborative Filtering**
Dogyoon Song, Christina E. Lee, Yihua Li, Devavrat Shah

#40 **Generative Adversarial Imitation Learning**
Jonathan Ho, Stefano Ermon

Tuesday Poster Session

#41 **Fast Active Set Methods for Online Spike Inference from Calcium Imaging**
Johannes Friedrich, Liam Paninski

#42 **Path-Normalized Optimization of Recurrent Neural Networks with ReLU Activations**
Behnam Neyshabur, Yuhuai Wu, Russ Salakhutdinov, Nati Srebro

#43 **Improved Regret Bounds for Oracle-Based Adversarial Contextual Bandits**
Vasilis Syrgkanis, Haipeng Luo, Akshay Krishnamurthy, Robert Schapire

#44 **Diffusion-Convolutional Neural Networks**
James Atwood

#45 **Faster Projection-free Convex Optimization over the Spectrahedron**
Dan Garber, Dan Garber

#46 **Structured Matrix Recovery via the Generalized Dantzig Selector**
Sheng Chen, Arindam Banerjee

#47 **Convex Two-Layer Modeling with Latent Structure**
Vignesh Ganapathiraman, Xinhua Zhang, Yaoliang Yu, Junfeng Wen

#48 **Finite-Sample Analysis of Fixed-k Nearest Neighbor Density Functionals Estimators**
Shashank Singh, Barnabas Poczos

#49 **Deep Learning Games**
Dale Schuurmans, Martin A Zinkevich

#50 **"Congruent" and "Opposite" Neurons: Sisters for Multisensory Integration and Segregation**
Wen-Hao Zhang, He Wang, K. Y. Michael Wong, Si Wu

#51 **Statistical Inference for Cluster Trees**
Jisu KIM, Yen-Chi Chen, Sivaraman Balakrishnan, Alessandro Rinaldo, Larry Wasserman

#52 **Minimizing Regret on Reflexive Banach Spaces and Nash Equilibria in Continuous Zero-Sum Games**
Maximilian Balandat, Walid Krichene, Claire Tomlin, Alexandre Bayen

#53 **A Neural Transducer**
Navdeep Jaitly, Quoc V Le, Oriol Vinyals, Ilya Sutskever, David Sussillo, Samy Bengio

#54 **Feature selection for classification of functional data using recursive maxima hunting**
José L. Torrecilla, Alberto Suárez

#55 **Homotopy Smoothing for Non-Smooth Problems with Lower Complexity than O(1/ϵ)**
Yi Xu, Yan Yan, Qihang Lin, Tianbao Yang

#56 **Nested Mini-Batch K-Means**
James Newling, François Fleuret

#57 **Density Estimation via Discrepancy Based Adaptive Sequential Partition**
Dangna Li, Kun Yang, Wing Hung Wong

#58 **Budgeted stream-based active learning via adaptive submodular maximization**
Kaito Fujii, Hisashi Kashima

#59 **Lifelong Learning with Weighted Majority Votes**
Anastasia Pentina, Ruth Urner

#60 **How Deep is the Feature Analysis underlying Rapid Visual Categorization?**
Sven2 Eberhardt, Jonah G Cader, Thomas Serre

#61 **Incremental Boosting Convolutional Neural Network for Facial Action Unit Recognition**
Shizhong Han, Zibo Meng, AHMED-SHEHAB KHAN, Yan Tong

#62 **Multivariate tests of association based on univariate tests**
Ruth Heller, Yair Heller

#63 **SURGE: Surface Regularized Geometry Estimation from a Single Image**
Peng Wang, Xiaohui Shen, Bryan Russell, Scott Cohen, Brian Price, Alan L Yuille

#64 **Memory-Efficient Backpropagation Through Time**
Audrunas Gruslys, Remi Munos, Ivo Danihelka, Marc Lanctot, Alex Graves

#65 **Scan Order in Gibbs Sampling: Models in Which it Matters and Bounds on How Much**
Bryan He, Christopher M De Sa, Ioannis Mitliagkas, Christopher Ré

#66 **2-Component Recurrent Neural Networks**
Xiang Li, Tao Qin, Jian Yang, Xiaolin Hu, Tieyan Liu

#67 **Direct Feedback Alignment Provides Learning in Deep Neural Networks**
Arild Nøkland

#68 **Variational Bayes on Monte Carlo Steroids**
Aditya Grover, Stefano Ermon

#69 **Agnostic Estimation for Misspecified Phase Retrieval Models**
Matey Neykov, Zhaoran Wang, Han Liu

#70 **Following the Leader and Fast Rates in Linear Prediction: Curved Constraint Sets and Other Regularities**
Ruitong Huang, Tor Lattimore, András György, Csaba Szepesvari

#71 **Combining Fully Convolutional and Recurrent Neural Networks for 3D Biomedical Image Segmentation**
Jianxu Chen, Lin Yang, Yizhe Zhang, Mark Alber, Danny Z Chen

#72 **The Product Cut**
Thomas Laurent, James von Brecht, Xavier Bresson, arthur szlam

#73 **Stochastic Gradient Methods for Distributionally Robust Optimization with f-divergences**
Hong Namkoong, John C Duchi

#74 **Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings**
Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, Adam T Kalai

#75 **Optimal spectral transportation with application to music transcription**
Rémi Flamary, Cédric Févotte, Nicolas Courty, Valentin Emiya

#76 **Combining Adversarial Guarantees and Stochastic Fast Rates in Online Learning**
Wouter M Koolen, Peter Grünwald, Tim van Erven

#77 **Towards Conceptual Compression**
Karol Gregor, Frederic Besse, Danilo Jimenez Rezende, Ivo Danihelka, Daan Wierstra

#78 **Can Peripheral Representations Improve Clutter Metrics on Complex Scenes?**
Arturo Deza, Miguel Eckstein

Tuesday Poster Session

#79 **GAP Safe Screening Rules for Sparse-Group Lasso**
Eugene Ndiaye, Olivier Fercoq, Alexandre Gramfort, Joseph Salmon

#80 **Learning Treewidth-Bounded Bayesian Networks with Thousands of Variables**
Mauro Scanagatta, Giorgio Corani, Cassio P de Campos, Marco Zaffalon

#81 **Ancestral Causal Inference**
Sara Magliacane, Tom Claassen, Joris M Mooij

#82 **Visual Question Answering with Question Representation Update**
Ruiyu Li, Jiaya Jia

#83 **Identification and Overidentification of Linear Structural Equation Models**
Bryant Chen

#84 **On Valid Optimal Assignment Kernels and Applications to Graph Classification**
Nils M. Kriege, Pierre-Louis Giscard, Richard Wilson

#85 **Constraints Based Convex Belief Propagation**
Y□aniv Tenzer, Alex Schwing, Kevin Gimpel, Tamir Hazan

#86 **Combinatorial Energy Learning for Image Segmentation**
Jeremy B Maitin-Shepard, Viren Jain, Michal Januszewski, Peter Li, Pieter Abbeel

#87 **A scalable end-to-end Gaussian process adapter for irregularly sampled time series classification**
Steve Li, Benjamin M Marlin

#88 **Stochastic Variance Reduction Methods for Saddle-Point Problems**
Balamurugan Palaniappan, Francis Bach

#89 **Dimensionality Reduction of Massive Sparse Datasets Using Coresets**
Dan Feldman, Mikhail Volkov, Daniela Rus

#90 **Efficient state-space modularization for planning: theory, behavioral and neural signatures**
Daniel McNamee, Daniel M Wolpert, Mate Lengyel

#91 **Adaptive Newton Method for Empirical Risk Minimization to Statistical Accuracy**
Aryan Mokhtari, Hadi Daneshmand, Aurelien Lucchi, Thomas Hofmann, Alejandro Ribeiro

#92 **RETAIN: Interpretable Predictive Model in Healthcare using Reverse Time Attention Mechanism**
Edward Choi, Mohammad Taha Bahadori, Jimeng Sun

#93 **Joint quantile regression in vector-valued RKHSs**
Maxime Sangnier, Olivier Fercoq, Florence d'Alché-Buc

#94 **Learnable Visual Markers**
Oleg Grinchuk, Vadim Lebedev, Victor Lempitsky

#95 **Exponential expressivity in deep neural networks through transient chaos**
Ben Poole, Subhaneil Lahiri, Maithreyi Raghu, Jascha Sohl-Dickstein, Surya Ganguli

#96 **On Multiplicative Integration with Recurrent Neural Networks**
Yuhuai Wu, Saizheng Zhang, Ying Zhang, Yoshua Bengio, Russ Salakhutdinov

#97 **Interpretable Nonlinear Dynamic Modeling of Neural Trajectories**
Yuan Zhao, Memming Park

#98 **Globally Optimal Training of Generalized Polynomial Neural Networks with Nonlinear Spectral Methods**
Antoine Gautier, Quynh N Nguyen, Matthias Hein

#99 **Linear Feature Encoding for Reinforcement Learning**
Zhao Song, Ron E Parr, Xuejun Liao, Lawrence Carin

#100 **Graphical Time Warping for Joint Alignment of Multiple Curves**
Yizhi Wang, David J Miller, Kira Poskanzer, Yue Wang, Lin Tian, Guoqiang Yu

#101 **Mixed Linear Regression with Multiple Components**
Kai Zhong, Prateek Jain, Inderjit S Dhillon

#102 **Statistical Inference for Pairwise Graphical Models Using Score Matching**
Ming Yu, Mladen Kolar, Varun Gupta

#103 **Hardness of Online Sleeping Combinatorial Optimization Problems**
Satyen Kale, Chansoo Lee, David Pal

#104 **An algorithm for L1 nearest neighbor search via monotonic embedding**
Xinan Wang, Sanjoy Dasgupta

#105 **On Local Maxima in the Population Likelihood of Gaussian Mixture Models: Structural Results & Algorithmic Consequences**
Chi Jin, Yuchen Zhang, Sivaraman Balakrishnan, Martin J Wainwright, Michael I Jordan

#106 **Learning User Perceived Clusters with Feature-Level Supervision**
Ting-Yu Cheng, Guiguan Lin, xinyang gong, Kang-Jun Liu, Shan-Hung Wu

#107 **InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets**
Xi Chen, Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, Pieter Abbeel

#108 **Neural universal discrete denoiser**
Taesup Moon, Seonwoo Min, Byunghan Lee, Sung R. Yoon

#109 **A primal-dual method for constrained consensus optimization**
Necdet Serhat Aybat, Erfan Yazdandoost Hamedani

#110 **Simple and Efficient Weighted Minwise Hashing**
Anshumali Shrivastava

#111 **Eliciting Categorical Data for Optimal Aggregation**
Chien-Ju Ho, Rafael Frongillo, Yiling Chen

#112 **Depth from a Single Image by Harmonizing Overcomplete Local Network Predictions**
Ayan Chakrabarti, Jingyu Shao, Greg Shakhnarovich

#113 **SEBOOST - Boosting Stochastic Learning Using Subspace Optimization Techniques**
Elad Richardson, Rom Herskovitz, Boris Ginsburg, Michael Zibulevsky

#114 **Reshaped Wirtinger Flow for Solving Quadratic Systems of Equations**
Huishuai Zhang, Yingbin Liang

#115 **Training and Evaluating Multimodal Word Embeddings with Large-scale Web Annotated Images**
Junhua Mao, Jiajing Xu, Kevin Jing, Alan L Yuille

Tuesday Poster Session

#116 **Online ICA: Understanding Global Dynamics of Nonconvex Optimization via Diffusion Processes**
Chris Junchi Li, Zhaoran Wang, Han Liu

#117 **Variational Information Maximizing Exploration**
Rein Houthooft, Xi Chen, Xi Chen, Yan Duan, John Schulman, Filip De Turck, Pieter Abbeel

#118 **Deconvolving Feedback Loops in Recommender Systems**
Ayan Sinha, David Gleich, Karthik Ramani

#119 **A Non-parametric Learning Method for Confidently Estimating Patient's Clinical State and Dynamics**
William Hoiles, Mihaela Van Der Schaar

#120 **Semiparametric Differential Graph Models**
Pan Xu, Quanquan Gu

#121 **A Non-convex One-Pass Framework for Generalized Factorization Machines and Rank-One Matrix Sensing**
Ming Lin, Jieping Ye

#122 **Sublinear Time Orthogonal Tensor Decomposition**
Zhao Song, David Woodruff, Huan Zhang

#123 **Achieving budget-optimality with adaptive schemes in crowdsourcing**
Ashish Khetan, Sewoong Oh

#124 **Joint Line Segmentation and Transcription for End-to-End Handwritten Paragraph Recognition**
Theodore Bluche

#125 **Human Decision-Making under Limited Time**
Pedro A Ortega, Alan A Stocker

#126 **Joint M-Best-Diverse Labelings as a Parametric Submodular Minimization**
Alexander Kirillov, Sasha Shekhovtsov, Carsten Rother, Bogdan Savchynskyy

#127 **Even Faster SVD Decomposition Yet Without Agonizing Pain**
Zeyuan Allen-Zhu, Yuanzhi Li

#128 **Fast and accurate spike sorting of high-channel count probes with KiloSort**
Marius Pachitariu, Nicholas A Steinmetz, Shabnam N Kadir, Matteo Carandini, Daniel D Harris

#129 **BBO-DPPs: Batched Bayesian Optimization via Determinantal Point Processes**
Tarun Kathuria, Amit Deshpande, Pushmeet Kohli

#130 **Stochastic Multiple Choice Learning for Training Diverse Deep Ensembles**
Stefan Lee, Senthil Purushwalkam Shiva Prakash, Michael Cogswell, Viresh Ranjan, David Crandall, Dhruv Batra

#131 **Optimal Sparse Linear Encoders and Sparse PCA**
Malik Magdon-Ismail, Christos Boutsidis

#132 **Using Social Dynamics to Make Individual Predictions: Variational Inference with Stochastic Kinetic Model**
Zhen Xu, Wen Dong, Sargur N Srihari

#133 **Learning Additive Exponential Family Graphical Models via \(ell_{2,1}\)-norm Regularized M-Estimation**
Xiaotong Yuan, Ping Li, Tong Zhang, Qingshan Liu, Guangcan Liu

#134 **Residual Networks are Exponential Ensembles of Relatively Shallow Networks**
Andreas Veit, Michael J Wilber, Serge Belongie

#135 **Full-Capacity Unitary Recurrent Neural Networks**
Scott Wisdom, Thomas Powers, John Hershey, Jonathan Le Roux, Les Atlas

#136 **Quantum Perceptron Models**
Ashish Kapoor, Nathan Wiebe, Krysta Svore

#137 **Mapping Estimation for Discrete Optimal Transport**
Michaël Perrot, Nicolas Courty, Rémi Flamary, Amaury Habrard

#138 **Stochastic Gradient Geodesic MCMC Methods**
Chang Liu, Jun Zhu, Yang Song

#139 **Variational Information Maximization for Feature Selection**
Shuyang Gao, Greg Ver Steeg, Aram Galstyan

#140 **A Minimax Approach to Supervised Learning**
Farzan Farnia, David Tse

#141 **Fast Distributed Submodular Cover: Public-Private Data Summarization**
Baharan Mirzasoleiman, Morteza Zadimoghaddam, Amin Karbasi

#142 **Domain Separation Networks**
Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, Dumitru Erhan

#143 **Multimodal Residual Learning for Visual QA**
Jin-Hwa Kim, Sang-Woo Lee, Donghyun Kwak, Min-Oh Heo, Jeonghee Kim, Jung-Woo Ha, Byoung-Tak Zhang

#144 **Optimizing affinity-based binary hashing using auxiliary coordinates**
Ramin Raziperchikolaei, Miguel A. Carreira-Perpinan

#145 **Coresets for Scalable Bayesian Logistic Regression**
Jonathan Huggins, Trevor Campbell, Tamara Broderick

#146 **The Parallel Knowledge Gradient Method for Batch Bayesian Optimization**
Jian Wu, Peter Frazier

#147 **Learning Multiagent Communication with Backpropagation**
Sainaa Sukhbaatar, arthur szlam, Rob Fergus

#148 **Optimal Binary Classifier Aggregation for General Losses**
Akshay Balsubramani, Yoav S Freund

#149 **The Generalized Reparameterization Gradient**
Francisco R Ruiz, Michalis Titsias RC AUEB, David Blei

#150 **Conditional Generative Moment-Matching Networks**
Yong Ren, Jun Zhu, Jialian Li, Yucen Luo

#151 **A Credit Assignment Compiler for Joint Prediction**
Kai-Wei Chang, He He, Stephane Ross, Hal Daume III, John Langford

#152 **Short-Dot: Computing Large Linear Transforms Distributedly Using Coded Short Dot Products**
Sanghamitra Dutta, Viveck Cadambe, Pulkit Grover

#153 **Spatio-Temporal Hilbert Maps for Continuous Occupancy Representation in Dynamic Environments**
Ransalu Senanayake, Lionel Ott, Simon O'Callaghan, Fabio Ramos

#154 **Learning HMMs with Nonparametric Emissions via Spectral Decompositions of Continuous Matrices**
Kirthevasan Kandasamy, Maruan Al-Shedivat, Eric P Xing

#155 **Integrator Nets**
Hakan Bilen, Andrea Vedaldi

#156 **Blind Attacks on Machine Learners**
Alex Beatson, Zhaoran Wang, Han Liu

Tuesday Poster Session

#157 **Optimistic Gittins Indices**
Eli Gutin, Vivek Farias

#158 **Sub-sampled Newton Methods with Non-uniform Sampling**
Peng Xu, Jiyan Yang, Farbod Roosta-Khorasani, Chris Ré, Michael W Mahoney

#159 **Learned Region Sparsity and Diversity Also Predicts Visual Attention**
Zijun Wei, Hossein Adeli, Minh Hoai, Greg Zelinsky, Dimitris Samaras

#160 **Adaptive Concentration Inequalities for Sequential Decision Problems**
Shengjia Zhao, Enze Zhou, Ashish Sabharwal, Stefano Ermon

#161 **Cooperative Graphical Models**
Josip Djolonga, Stefanie Jegelka, Sebastian Tschiatschek, Andreas Krause

#162 **Correlated-PCA: Principal Components' Analysis when Data and Noise are Correlated**
Namrata Vaswani, Han Guo

#163 **Hierarchical Object Representation for Open-Ended Object Category Learning and Recognition**
Hamidreza Kasaei

#164 **Optimal Tagging with Markov Chain Optimization**
Nir Rosenfeld, Amir Globerson

#165 **Bayesian optimization for automated model selection**
Gustavo Malkomes, Charles Schaff, Roman Garnett

#166 **Multi-view Anomaly Detection via Robust Probabilistic Latent Variable Models**
Tomoharu Iwata, Makoto Yamada

#167 **Inference by Reparameterization in Neural Population Codes**
Rajkumar Vasudeva Raju, Xaq Pitkow

#168 **Efficient Neural Codes under Metabolic Constraints**
Zhuo Wang, Xue-Xin Wei, Alan A Stocker, Daniel D Lee

#169 **Learning Deep Parsimonious Representations**
Renjie Liao, Alex Schwing, Richard Zemel, Raquel Urtasun

#170 **An equivalence between high dimensional Bayes optimal inference and M-estimation**
Madhu Advani, Surya Ganguli

#171 **Minimizing Quadratic Functions in Constant Time**
Kohei Hayashi, Yuichi Yoshida

#172 **Learning Structured Sparsity in Deep Neural Networks**
Wei Wen, Chunpeng Wu, Yandan Wang, Yiran Chen, Hai Li

#173 **Adversarial Multiclass Classification: A Risk Minimization Perspective**
Rizal Fathony, Anqi Liu, Kaiser Asif, Brian Ziebart

#174 **Unified Methods for Exploiting Piecewise Structure in Convex Optimization**
Tyler B Johnson, Carlos Guestrin

#175 **Fast and Provably Good Seedings for k-Means**
Olivier Bachem, Mario Lucic, Hamed Hassani, Andreas Krause

#176 **Testing for Differences in Gaussian Graphical Models: Applications to Brain Connectivity**
Eugene Belilovsky, Gaël Varoquaux, Matthew B Blaschko

#177 **Synthesis of MCMC and Belief Propagation**
Sung-Soo Ahn, Michael Chertkov, Jinwoo Shin

#178 **Value Iteration Networks**
Aviv Tamar, Sergey Levine, Pieter Abbeel, YI WU, Garrett Thomas

#179 **Sequential Neural Models with Stochastic Layers**
Marco Fraccaro, Søren Kaae Sønderby, Ulrich Paquet, Ole Winther

#180 **Graphons, mergeons, and so on!**
Justin Eldridge, Mikhail Belkin, Yusu Wang

#181 **Hierarchical Clustering via Spreading Metrics**
Aurko Roy, Sebastian Pokutta

#182 **Deep Learning for Predicting Human Strategic Behavior**
Jason S Hartford, James R Wright, Kevin Leyton-Brown

#183 **Global Analysis of Expectation Maximization for Mixtures of Two Gaussians**
Ji Xu, Daniel Hsu,

#184 **Supervised learning through the lens of compression**
Ofir David, Shay Moran, Amir Yehudayoff

#185 **Matrix Completion has No Spurious Local Minimum**
Rong Ge, Jason Lee, Tengyu Ma

#186 **Clustering with Same-Cluster Queries**
Hassan Ashtiani, Shrinu Kushagra, Shai Ben-David

#187 **MetaGrad: Multiple Learning Rates in Online Learning**
Tim van Erven, Wouter M Koolen

#188 **Unsupervised Feature Extraction by Time-Contrastive Learning and Nonlinear ICA**
Aapo Hyvarinen, Hiroshi Morioka

#189 **Phased LSTM: Accelerating Recurrent Network Training for Long or Event-based Sequences**
Daniel Neil, Michael Pfeiffer, Shih-Chii Liu

#190 **Tractable Operations for Arithmetic Circuits of Probabilistic Models**
Yujia Shen, Arthur Choi, Adnan Darwiche

#191 **Using Fast Weights to Attend to the Recent Past**
Jimmy Ba, Geoffrey E Hinton, Volodymyr Mnih, Joel Z Leibo, Catalin Ionescu

#192 **Bayesian Intermittent Demand Forecasting for Large Inventories**
Matthias W Seeger, David Salinas, Valentin Flunkert

#193 **Blazing the trails before beating the path: Sample-efficient Monte-Carlo planning**
Jean-Bastien Grill, Michal Valko, Remi Munos

#194 **SDP Relaxation with Randomized Rounding for Energy Disaggregation**
Kiarash Shaloudegi, András György, Csaba Szepesvari, Wilsun Xu

#195 **Markov Chain Sampling in Discrete Probabilistic Models with Constraints**
Chengtao Li, Suvrit Sra, Stefanie Jegelka

#196 **Unsupervised Learning of 3D Structure from Images**
Danilo Jimenez Rezende, Ali Eslami, Shakir Mohamed, Peter Battaglia, Max Jaderberg, Nicolas Heess

#197 **The Multiple Quantile Graphical Model**
Alnur Ali, J. Zico Kolter, Ryan J Tibshirani

#198 **Linear Contextual Bandits with Knapsacks**
Shipra Agrawal, Nikhil Devanur

#1 The Multi-fidelity Multi-armed Bandit
Kirthevasan Kandasamy (CMU)
Gautam Dasarathy (Carnegie Mellon Univ.)
Barnabas Poczos
Jeff Schneider (CMU)

We study a variant of the classical stochastic K-armed bandit problem where observing the outcome of each arm is expensive, but cheap approximations to this outcome are available. For example, in online advertising the performance of an ad can be approximated to varying degrees by displaying it for shorter time periods. We formalise this task as a {multi-fidelity} bandit problem, where, at each time step, the forecaster may choose to play an arm at any one of M fidelities and obtain a noisy reward. The highest fidelity (the desired outcome) expends cost \costM while any of the M-1 lower fidelities (the approximations) expends \costm < \costM. We develop \mfucb, a novel upper confidence bound procedure for this setting and prove that it naturally adapts to the sequence of available approximations and costs thus attaining better regret than naive strategies that ignore such approximations. For instance, in the above online advertising example, \mfucbs would use the lower fidelities to quickly eliminate suboptimal ads, while reserving the larger expensive experiments on a small set of promising candidates. We complement this result with a lower bound and show that \mfucbs is nearly optimal.

#2 Probabilistic Inference with Generating Functions for Poisson Latent Variable Models
Kevin Winner (UMass CICS)
Dan Sheldon

Graphical models with latent count variables arise in a number of fields. Standard exact inference techniques such as variable elimination and belief propagation do not apply to these models because the latent variables have countably infinite support. As a result, approximations such as truncation or MCMC are employed. We present the first exact inference algorithms for a class of models with latent count variables by developing a novel representation of countably infinite factors as probability generating functions, and then performing variable elimination with generating functions. Our approach is exact, runs in pseudo-polynomial time, and is much faster than existing approximate techniques. It leads to better parameter estimates for problems in population ecology by avoiding error introduced by approximate likelihood computations.

#3 Adaptive Maximization of Pointwise Submodular Functions With Budget Constraint
Nguyen Cuong (National Univ. of Singapore)
Huan Xu (NUS)

We study the worst-case adaptive optimization problem with budget constraint that is useful for modeling various practical applications in artificial intelligence and machine learning. We investigate the near-optimality of greedy algorithms for this problem with both modular and non-modular cost functions. In both cases, we prove that two simple greedy algorithms are not near-optimal but the best between them is near-optimal if the utility function satisfies pointwise submodularity and pointwise cost-sensitive submodularity respectively. This implies a combined algorithm that is near-optimal with respect to the optimal algorithm that uses half of the budget. We discuss applications of our theoretical results and also report experiments comparing the greedy algorithms on the active learning problem.

#4 Machine Translation Through Learning From a Communication Game
Di He (Microsoft)
Yingce Xia (USTC)
Tao Qin (Microsoft)
Liwei Wang
Nenghai Yu (USTC)
Tieyan Liu (Microsoft)
Wei-Ying Ma (Microsoft)

State-of-the-art machine translation (MT) systems are usually trained on aligned parallel corpora, which are limited in scale and costly to obtain in practice. Given that there exist almost unlimited monolingual data in the Web, in this work we study how to boost the performance of MT systems by leveraging monolingual data in two-language translation. We formulate the translation system as a two-player communication game: Player 1 only understands language A and sends a message in language A to Player 2 through a noisy channel, which is a translation model from language A to B. Player 2 only knows language B and sends the received message in language B to Player 1 through another noisy channel, which is a translation model from language B to A. By checking whether the received message is consistent with his/her original one, Player 1 gets to know the quality of the two channels and leverages this quality feedback to improve the two channels (the translation models). Similarly, Player 2 can send a message in B to Player 1, go through a symmetric process, and improve the two translation models. We call our learning framework communication-based machine translation(CMT). Note that in this communication game, the two players do not need aligned corpora and they can improve the two translation models through reward maximization. Distinguishing features of CMT include: (1) we jointly train two dual translation models in one framework, (2) we train translation models from unlabeled data through reinforcement feedback, (3) we develop a novel optimization method for this machine translation task using reinforcement learning. Experiments show that our proposed CMT works very well on both English\ leftrightarrowFrench translation and English\leftrightarrowChinese translation.

#5 Iterative Refinement of the Approximate Posterior for Directed Belief Networks
devon Hjelm (Univ. of New Mexico)
Russ Salakhutdinov (Univ. of Toronto)
Kyunghyun Cho (Univ. of Montreal)
Nebojsa Jojic (Microsoft Research)
Vince Calhoun (Mind Research Network)
Junyoung Chung (Univ. of Montreal)

Variational methods that rely on a recognition network to approximate the posterior of directed graphical models offer better inference and learning than previous methods. Recent advances that exploit the capacity and flexibility in this approach have greatly expanding on what kinds of models can be trained. However, as a proposal for the posterior, the recognition network has limited capacity, which can constrain the representational power of the generative model and increase the variance of Monte Carlo estimates. To address these issues, we introduce an iterative refinement procedure for improving the approximate posterior of the recognition network and show that training with the refined posterior is competitive with state-of-the-art methods. The advantages of refinement are further evident in an increased effective sample size, which implies a better fit to the true posterior.

#6 Unsupervised Risk Estimation Using Only Conditional Independence Structure
Jacob Steinhardt (Stanford Univ.)
Percy S Liang

We show how to estimate a model's test error from unlabeled data, on distributions very different from the training distribution, while assuming only that certain conditional independencies are preserved between train and test. We do not need to assume that the optimal predictor is the same between train and test, or that the true distribution lies in any parametric family. We can also efficiently differentiate the error estimate to perform unsupervised learning. Our technical tool is the method of moments, which allows us to exploit conditional independencies even in the absence of a specifiied parametric model. Our framework encompasses a large family of losses including the log and exponential loss, and extends to structured output settings such as hidden Markov models.

#7 Hierarchical Question-Image Co-Attention for Visual Question Answering
Jiasen Lu (Virginia Tech)
Jianwei Yang (Virginia Tech)
Dhruv Batra
Devi Parikh (Virginia Tech)

A number of recent works have proposed attention models for VQA that model spatial maps highlighting image regions relevant to answering the question. In this paper, we argue that in addition to modeling "where to look" or visual attention, it is equally important to model "what words to listen to" or question attention. We present a novel co-attention model for VQA that jointly reasons about image and question attention. In addition, our model reasons about the question (and consequently the image via the co-attention mechanism) in a hierarchical fashion via a novel 1-dimensional CNNs model. Our final model outperforms all reported methods, improving the state-of-the-art on the VQA dataset from 60.4% to 62.1%, and from 61.6% to 65.4% on the COCO-QA dataset.

#8 Bayesian Optimization with a Finite Budget: An Approximate Dynamic Programming Approach
Remi Lam (MIT)
Karen Willcox (MIT)
David Wolpert

Optimization of expensive objective functions has become an important part of many fields. Such optimization is often performed using a finite budget of evaluations. The optimal solution strategy for Bayesian optimization with finite budget can be formulated as a dynamic programming instance. This results in a complex problem with uncountable, dimension-increasing state space and an uncountable control space. To solve this problem, we invoke a classical technique of approximate dynamic programming: rollout. Within the rollout algorithm, we propose heuristics adapted to the Bayesian optimization setting. The performance of rollout is examined numerically and is shown to outperform several popular greedy Bayesian optimization algorithms.

#9 Learning to learn by gradient descent by gradient descent
Marcin Andrychowicz (Google Deepmind)
Misha Denil
Sergio Gómez (Google DeepMind)
Matthew W Hoffman (Google DeepMind)
David Pfau (Google DeepMind)
Tom Schaul
Nando de Freitas (Google)

The move from hand-designed features to learned features in machine learning has been wildly successful. In spite of this, optimization algorithms are still designed by hand. In this paper we show how the design of an optimization algorithm can be cast as a learning problem, allowing the algorithm to learn to exploit structure in the problems of interest in an automatic way. Our learned algorithms, implemented by LSTMs, outperform generic, hand-designed competitors on the tasks for which they are trained, and also generalize well to new tasks with similar structure. We demonstrate this on a number of tasks, including simple convex problems, training neural networks, and styling images with neural art.

#10 Computational and Statistical Tradeoffs in Learning to Rank
Ashish Khetan (Univ. of Illinois Urbana-)
Sewoong Oh

For massive and heterogeneous modern datasets, it is of fundamental interest to provide guarantees on the accuracy of estimation when computational resources are limited. In the application of learning to rank, we provide a hierarchy of rank-breaking mechanisms ordered by the complexity in thus generated sketch of the data. This allows the number of data points collected to be gracefully traded off against computational resources available, while guaranteeing the desired level of accuracy. Theoretical guarantees on the proposed generalized rank-breaking implicitly provide such trade-offs, which can be explicitly characterized under certain canonical scenarios on the structure of the data.

#11 Pairwise Choice Markov Chains
Stephen Ragain (Stanford Univ.)
Johan Ugander

As datasets capturing human choices grow in richness and scale, particularly in online domains, there is an increasing need for choice models flexible enough to handle data that violate traditional choice-theoretic axioms such as regularity, stochastic transitivity, or Luce's choice axiom. In this work we introduce the Pairwise Choice Markov Chain (PCMC) model of discrete choice, an inferentially tractable model that does not assume these traditional axioms while still satisfying the foundational axiom of uniform expansion, which can be viewed as a weaker version of Luce's axiom. We show that the PCMC model significantly outperforms the Multinomial Logit (MNL) model in prediction tasks on two empirical data sets known to exhibit violations of Luce's axiom. Our analysis also synthesizes several recent observations connecting the Multinomial Logit model and Markov chains; the PCMC model retains the Multinomial Logit model as a special case.

#12 Incremental Learning for Variational Sparse Gaussian Process Regression

Ching-An Cheng (Georgia Institute of Technolog)
Byron Boots

Recent work on scaling up Gaussian process regression (GPR) to large datasets has primarily focused on sparse GPR, which leverages a small set of basis functions to approximate the full Gaussian process during inference. However, the majority of these approaches are batch methods that operate on the entire training dataset at once, precluding the use of datasets that are streaming or too large to fit into memory. Although previous work has considered incrementally solving variational sparse GPR, most algorithms fail to update the basis functions and therefore perform suboptimally. We propose a novel incremental learning algorithm for variational sparse GPR based on stochastic mirror ascent of probability densities in reproducing kernel Hilbert space. This new formulation allows our algorithm to update basis functions online in accordance with the manifold structure of probability densities for fast convergence. We conduct several experiments and show that our proposed approach achieves better empirical performance in terms of prediction error than the recent state-of-the-art incremental solutions to variational sparse GPR.

#13 Combinatorial Multi-Armed Bandit with General Reward Functions

Wei Chen
Wei Hu (Princeton Univ.)
Fu Li (The Univ. of Texas at Austin)
Jian Li (Tsinghua Univ.)
Yu Liu (Tsinghua Univ.)
Pinyan Lu (Shanghai Univ. of Finance and Economics)

In this paper, we study the stochastic combinatorial multi-armed bandit (CMAB) framework that allows a general nonlinear reward function, whose expected value may not depend only on the means of the input random variables but possibly on the entire distributions of these variables. Our framework enables a much larger class of reward functions such as the \max function and nonlinear utility functions. Existing techniques relying on accurate estimations of the means of random variables, such as the upper confidence bound (UCB) technique, do not work directly on these functions. We propose a new algorithm called stochastically dominant confidence bound (SDCB), which estimates the distributions of underlying random variables and their stochastically dominant confidence bounds. We prove that if the underlying variables have known finite supports, SDCB can achieve O(log T) distribution-dependent regret and $O(\sqrt{T})$ distribution-independent regret, where T is the time horizon. For general arbitrary distributions, we further use a discretization technique and show an $O(\sqrt{T})$ regret bound. We apply our results to the K-MAX problem and the expected utility maximization problems. In particular, for K-MAX, we provide the first polynomial-time approximation scheme (PTAS) for its offline problem, and give the first $O(\sqrt{T})$ bound on the $(1-\epsilon)$-approximation regret of its online problem, for any $\epsilon > 0$.

#14 Observational-Interventional Priors for Dose-Response Learning

Ricardo Silva

Controlled interventions provide the most direct source of information for learning causal effects. In particular, a dose-response curve can be learned by varying the treatment level and observing the corresponding outcomes. However, interventions can be expensive and time-consuming. Observational data, where the treatment is not controlled by a known mechanism, is sometimes available. Under some strong assumptions, observational data allows for the estimation of dose-response curves. Estimating such curves nonparametrically is hard: sample sizes for controlled interventions may be small, while in the observational case a large number of measured confounders may need to be marginalized. In this paper, we introduce a hierarchical Gaussian process prior that constructs a distribution over the dose-response curve by learning from observational data, and reshapes the distribution with a nonparametric affine transform learned from controlled interventions. This function composition from different sources is shown to speed-up learning, which we demonstrate with a thorough sensitivity analysis and an application to modeling the effect of therapy on cognitive skills of premature infants.

#15 On Graph Reconstruction via Empirical Risk Minimization: Fast Learning Rates and Scalability

Guillaume Papa (Télécom ParisTech)
Aurélien Bellet
Stephan Clémençon

The problem of predicting connections between a set of data points finds many applications, in systems biology and social network analysis among others. This paper focuses on the \textit{graph reconstruction} problem, where the prediction rule is obtained by minimizing the average error over all n(n-1)/2 possible pairs of the n nodes of a training graph. Our first contribution is to derive learning rates of order O(log n / n) for this problem, significantly improving upon the slow rates of order $O(1/\sqrt{n})$ established in the seminal work of Biau & Bleakley (2006). Strikingly, these fast rates are universal, in contrast to similar results known for other statistical learning problems (e.g., classification, density level set estimation, ranking, clustering) which require strong assumptions on the distribution of the data. Motivated by applications to large graphs, our second contribution deals with the computational complexity of graph reconstruction. Specifically, we investigate to which extent the learning rates can be preserved when replacing the empirical reconstruction risk by a computationally cheaper Monte-Carlo version, obtained by sampling with replacement $B << n^2$ pairs of nodes. Finally, we illustrate our theoretical results by numerical experiments on synthetic and real graphs.

#16 DeepMath - Deep Sequence Models for Premise Selection

Geoffrey Irving
Christian Szegedy
Alex A Alemi (Google)
Francois Chollet
Josef Urban (Czech Technical Univ. in Prague)

We study the effectiveness of neural sequence models for premise selection in automated theorem proving, a key bottleneck for progress in formalized mathematics. We propose a two stage approach for this task that yields good results for the premise selection task on the Mizar corpus while avoiding the hand-engineered features of existing state-of-the-art models. To our knowledge, this is the first time deep learning has been applied to theorem proving.

#17 Efficient Second Order Online Learning by Sketching

Haipeng Luo (Princeton Univ.)
Alekh Agarwal (Microsoft)
Nicolò Cesa-Bianchi
John Langford

We propose Sketched Online Newton (SON), an online second order learning algorithm that enjoys substantially improved regret guarantees for ill-conditioned data. SON is an enhanced version of the Online Newton Step, which, via sketching techniques enjoys a running time linear in the dimension and sketch size. We further develop sparse forms of the sketching methods (such as Oja's rule), making the computation linear in the sparsity of features. Together, the algorithm eliminates all computational obstacles in previous second order online learning approaches.

#18 Gaussian Processes for Survival Analysis

Tamara Fernandez (Oxford)
Nicolás Rivera (King's College London)
Yee Whye Teh

We introduce a semi-parametric Bayesian model for survival analysis. The model is centred on a parametric baseline hazard, and uses a Gaussian process to model variations away from it nonparametrically, as well as dependence on covariates. As opposed to other methods in survival analysis, our framework does not impose unnecessary constraints in the hazard rate or in the survival function. Furthermore, our model handles left, right and interval censoring mechanisms common in survival analysis. We propose a MCMC algorithm to perform inference and an approximation scheme based on random Fourier features to make computations faster. We report experimental results on synthetic and real data, showing that our model performs better than competing models such as Cox proportional hazards, ANOVA-DDP and random survival forests.

#19 The Power of Optimization from Samples

Eric Balkanski (Harvard Univ.)
Aviad Rubinstein (UC Berkeley)
Yaron Singer

We consider the problem of optimization from samples of monotone submodular functions with bounded curvature. In numerous applications, the function optimized is not known a priori, but instead learned from data. What are the guarantees we have when optimizing functions from sampled data? In this paper we show that for any monotone submodular function with curvature c there is a (1 - c)/(1 + c - c^2) approximation algorithm for maximization under cardinality constraints when polynomially-many samples are drawn from the uniform distribution over feasible sets. Moreover, we show that this algorithm is optimal. That is, for any c < 1, there exists a submodular function with curvature c for which no algorithm can achieve a better approximation. The curvature assumption is crucial as for general monotone submodular functions no algorithm can obtain a constant-factor approximation for maximization under a cardinality constraint when observing polynomially-many samples drawn from any distribution over feasible sets, even when the function is statistically learnable.

#20 Global Optimality of Local Search for Low Rank Matrix Recovery

Srinadh Bhojanapalli (TTI Chicago)
Behnam Neyshabur (TTI-Chicago)
Nati Srebro

We show that there are no spurious local minima in the non-convex factorized parametrization of low-rank matrix recovery from incoherent linear measurements. With noisy measurements we show all local minima are very close to a global optimum. Together with a curvature bound at saddle points, this yields a polynomial time global convergence guarantee for stochastic gradient descent {\em from random initialization}.

#21 A state-space model of cross-region dynamic connectivity in MEG/EEG

Ying Yang (Carnegie Mellon Univ.)
Elissa Aminoff (Carnegie Mellon Univ.)
Michael Tarr (Carnegie Mellon Univ.)
Rob E Robert (Carnegie Mellon Univ.)

Cross-region dynamic connectivity, which describes spatio-temporal dependence of neural activity among multiple brain regions of interest (ROIs), can provide important information for understanding cognition. For estimating such connectivity, magnetoencephalography (MEG) and electroencephalogram (EEG) are well-suited tools because of their millisecond temporal resolution. However, localizing source activity in the brain requires solving an under-specified linear problem. In typical two-step approaches, researchers first solve the linear problem with general priors assuming independence across ROIs, and secondly quantify cross-region connectivity. In this work, we propose a one-step state-space model to improve estimation of dynamic connectivity. The model treats the mean activity in individual ROIs as the state variable, and describes non-stationary dynamic dependence across ROIs using time-varying auto-regression. Compared with a commonly used two-step method, which first obtains the minimum-norm estimates of source activity, and then fits the auto-regressive model, our state-space model yielded smaller estimation errors on simulated data where the model assumptions held. When applied on empirical MEG data from one participant in a scene-processing experiment, our state-space model also demonstrated intriguing preliminary results, indicating significant leading and lagged linear dependence between the early visual cortex and a higher-level scene-sensitive region, which suggests feed-forward and feedback information flow within the visual cortex during scene processing.

#22 Hypothesis Testing in Unsupervised Domain Adaptation with Applications in Neuroscience

Hao Zhou (Univ. of Wisconsin Madiso)
Vamsi K Ithapu (Univ. of Wisconsin Madison)
Sathya Narayanan Ravi (Univ. of Wisconsin Madiso)
Vikas Singh (UW Madison)
Grace Wahba (Univ. of Wisconsin Madison)
Sterling C Johnson (Univ. of Wisconsin Madison)

Consider samples from two different data sources \{\mathbf{x_s^i}\} \sim P_{\rm source} and \{\mathbf{x_t^i}\} \sim P_{\rm target}. We only observe their transformed versions h(\mathbf{x_s^i}) and g(\mathbf{x_t^i}), for some known function class h(\cdot) and g(\cdot). Our goal is to perform a statistical test checking if P_{\rm source} = P_{\rm target} while removing the distortions induced by the transformations. This problem is closely related to concepts underlying numerous domain adaptation algorithms, and in our case, is motivated by the need to combine clinical and imaging based biomarkers from multiple sites and/or batches, where this problem is fairly common and an impediment in the conduct of analyses with much larger sample sizes. We develop a framework that addresses this problem using ideas from hypothesis testing on the transformed measurements, where in the distortions need to be estimated {\it in tandem} with the testing. We derive a simple algorithm and study its convergence and consistency properties in detail, and we also provide lower-bound strategies based on recent work in continuous optimization. On a dataset of individuals at risk for neurological disease, we show that our results are competitive with alternative procedures that are twice as expensive and in some cases operationally infeasible to implement.

#23 Bi-Objective Online Matching and Submodular Allocations

Hossein Esfandiari (Univ. of Maryland)
Nitish Korula (Google Research)
Vahab Mirrokni (Google)

Online allocation problems have been widely studied due to their numerous prac-tical applications (particularly to Internet advertising), as well as considerable theoretical interest. The main challenge in such problems is making assignment decisions in the face of uncertainty about future input; effective algorithms need to predict which constraints are most likely to bind, and learn the balance between short-term gain and the value of long-term resource availability. In many important applications, the algorithm designer is faced with multiple objectives to optimize. While there has been considerable work on multi-objective offline optimization (when the entire input is known in advance), very little is known about the online case, particularly in the case of adversarial input. In this paper, we give the first results for bi-objective online submodular optimization, providing almost matching upper and lower bounds for allocating items to agents with two submodular value functions. We also study practically relevant special cases of this problem related to Internet advertising, and obtain improved results. All our algorithms are nearly best possible, as well as being efficient and easy to implement in practice.

#24 A Constant-Factor Bi-Criteria Approximation Guarantee for k-means++

Dennis Wei (IBM Research)

This paper studies the k-means++ algorithm for clustering as well as the class of Dℓ sampling algorithms to which k-means++ belongs. It is shown that for any constant factor $\beta > 1$, selecting β k cluster centers by Dℓ sampling yields a constant-factor approximation to the optimal clustering with k centers, in expectation and without conditions on the dataset. This result extends the previously known O(log k) guarantee for the case $\beta = 1$ to the constant-factor bi-criteria regime. It also improves upon an existing constant-factor bi-criteria result that holds only with constant probability.

#25 Causal Bandits: Learning Good Interventions via Causal Inference

Finnian Lattimore (Australian National Univ.)
Tor Lattimore
Mark Reid

We study the problem of using causal models to improve the rate at which good interventions can be learned online in a stochastic environment. Our formalism combines multi-arm bandits and causal inference to model a novel type of bandit feedback that is not exploited by existing approaches. We propose a new algorithm that exploits the causal feedback and prove a bound on its simple regret that is strictly better (in all quantities) than algorithms that do not use the additional causal information.

#26 Unsupervised Domain Adaptation with Residual Transfer Networks

Mingsheng Long (Tsinghua Univ.)
Han Zhu (Tsinghua Univ.)
Jianmin Wang (Tsinghua Univ.)
Michael I Jordan

The recent success of deep neural networks relies on massive amounts of labeled data. For a target task where labeled data is unavailable, domain adaptation can transfer a learner from a different source domain. In this paper, we propose a new approach to domain adaptation in deep networks that can simultaneously learn adaptive classifiers and transferable features from labeled data in the source domain and unlabeled data in the target domain. We relax a shared-classifier assumption made by previous methods and assume that the source classifier and target classifier differ by a residual function. We enable classifier adaptation by plugging several layers into the deep network to explicitly learn the residual function with reference to the target classifier. We embed features of multiple layers into reproducing kernel Hilbert spaces (RKHSs) and match feature distributions for feature adaptation. The adaptation behaviors can be achieved in most feed-forward models by extending them with new residual layers and loss functions, which can be trained efficiently using standard back-propagation. Empirical evidence shows that the new approach outperforms state of the art methods on standard domain adaptation benchmarks.

#27 Data driven estimation of Laplace-Beltrami operator

Frederic Chazal (INRIA)
Ilaria Giulini
Bertrand Michel

Approximations of Laplace-Beltrami operators on manifolds through graph Laplacians have become popular tools in data analysis and machine learning. These discretized operators usually depend on bandwidth parameters whose tuning remains a theoretical and practical problem. In this paper, we address this problem for the unormalized graph Laplacian by establishing an oracle inequality that opens the door to a well-founded data-driven procedure for the bandwidth selection. Our approach relies on recent results by Lacour and Massart (2015) on the so-called Lepski's method.

#28 Fast Algorithms for Robust PCA via Gradient Descent

Xinyang Yi (UT Austin)
Dohyung Park (Univ. of Texas at Austin)
Yudong Chen
Constantine Caramanis

We consider the problem of Robust PCA in the fully and partially observed settings. Without corruptions, this is the well-known matrix completion problem. From a statistical standpoint this problem has been recently well-studied, and conditions on when recovery is possible (how many observations do we need, how many corruptions can we tolerate) via polynomial-time algorithms is by now understood. This paper presents and analyzes a non-convex optimization approach that greatly reduces the computational complexity of the above problems, compared to the best available algorithms. In particular, in the fully observed case, with r denoting rank and d dimension, we reduce the complexity from O(r^2d^2log(1/ϵ)) to O(rd^2log(1/ϵ)) -- a big savings when the rank is big. For the partially observed case, we show the complexity of our algorithm is no more than O(r^4dlog(d)log(1/ϵ)). Not only is this the best-known run-time for a provable algorithm under partial observation, but in the setting where r is small compared to d, it also allows for near-linear-in-d run-time that can be exploited in the fully-observed case as well, by simply running our algorithm on a subset of the observations.

#29 NESTT: A Nonconvex Primal-Dual Splitting Method for Distributed and Stochastic Optimization

Davood Hajinezhad (Iowa State Univ.)
Mingyi Hong
Tuo Zhao (Johns Hopkins Univ.)
Zhaoran Wang (Princeton Univ.)

We study a stochastic and distributed algorithm for nonconvex problems whose objective consists a sum N nonconvex L_i/N-smooth functions, plus a nonsmooth regularizer. The proposed NonconvEx primal-dual SpliTTing (NESTT) algorithm splits the problem into N subproblems, and utilizes an augmented Lagrangian based primal-dual scheme to solve it in a distributed and stochastic manner. With a special non-uniform sampling, a version of NESTT achieves ϵ-stationary solution using \mathcal{O}((\sum_{i=1}^N√{L_i/N})^2/ϵ) gradient evaluations, which can be up to \mathcal{O}(N) times better than the (proximal) gradient descent methods. It also achieves Q-linear convergence rate for nonconvex ℓ_1 penalized quadratic problems with polyhedral constraints. Further, we reveal a fundamental connection between {\it primal-dual} based methods and a few {\it primal only} methods such as IAG/SAG/SAGA.

#30 Fundamental Limits of Budget-Fidelity Trade-off in Label Crowdsourcing

Farshad Lahouti (Caltech)
Babak Hassibi (Caltech)

Digital crowdsourcing (CS) is a modern approach to perform certain large projects using small contributions of a large crowd. In CS, a taskmaster typically breaks down the project into small batches of tasks and assigns them to so-called workers with imperfect skill levels. The crowdsourcer then collects and analyzes the results for inference and serving the purpose of the project. In this work, the CS problem, as a human-in-the-loop computation problem, is modeled and analyzed in an information theoretic rate-distortion framework. The purpose is to identify the ultimate fidelity that one can achieve by any form of query from the crowd and any decoding (inference) algorithm with a given budget. The results are established by a joint source channel (de)coding scheme, which represent the query scheme and inference, over parallel noisy channels, which model workers with imperfect skill levels. We also present and analyze a query scheme dubbed k-ary incidence coding and study optimized query pricing in this setting.

#31 Supervised Learning with Tensor Networks

Miles Stoudenmire (Univ of California Irvine)
David J Schwab (Northwestern Univ.)

Tensor networks are efficient representations of high-dimensional tensors which have been very successful for physics and mathematics applications. We demonstrate how algorithms for optimizing such networks can be adapted to supervised learning tasks by using matrix product states (tensor trains) to parameterize models for classifying images. For the MNIST data set we obtain less than 1% test set classification error. We discuss an interpretation of the additional structure imparted by the tensor network to the learned model.

#32 Understanding Probabilistic Sparse Gaussian Process Approximations

Matthias Bauer (Univ. of Cambridge)
Mark van der Wilk (Univ. of Cambridge)
Carl Edward Rasmussen (Univ. of Cambridge)

Good sparse approximations are essential for practical inference in Gaussian Processes as the computational cost of exact methods is prohibitive for large datasets. The Fully Independent Training Conditional (FITC) and the Variational Free Energy (VFE) approximations are two recent popular methods. Despite superficial similarities, these approximations have surprisingly different theoretical properties and behave differently in practice. We thoroughly investigate the two methods for regression both analytically and through illustrative examples, and draw conclusions to guide practical application.

#33 A Locally Adaptive Normal Distribution

Georgios Arvanitidis (DTU)
Lars K Hansen
Søren Hauberg

The multivariate normal density is a monotonic function of the distance to the mean, and its ellipsoidal shape is due to the underlying Euclidean metric. We suggest to replace this metric with a locally adaptive, smoothly changing (Riemannian) metric that favors regions of high local density. The resulting locally adaptive normal distribution (LAND) is a generalization of the normal distribution to the "manifold" setting, where data is assumed to lie near a potentially low-dimensional manifold embedded in R^D. The LAND is parametric, depending only on a mean and a covariance, and is the maximum entropy distribution under the given metric. The underlying metric is, however, non-parametric. We develop a maximum likelihood algorithm to infer the distribution parameters that relies on a combination of gradient descent and Monte Carlo integration. We further extend the LAND to mixture models, and provide the corresponding EM algorithm. We demonstrate the efficiency of the LAND to fit non-trivial probability distributions over both synthetic data, and EEG measurements of human sleep.

#34 Anchor-Free Correlated Topic Modeling: Identifiability and Algorithm

Kejun Huang (Univ. of Minnesota)
Xiao Fu (Univ. of Minnesota)
Nikos D. Sidiropoulos (Univ. of Minnesota)

In topic modeling, many algorithms that guarantee identifiability of the topics have been developed under the premise that there exist anchor words -- i.e., words that only appear (with positive probability) in one topic. Follow-up work has resorted to three or higher-order statistics of the data corpus to relax the anchor word assumption. Reliable estimates of higher-order statistics are hard to obtain, however, and the identification of topics under those models hinges on uncorrelatedness of the topics, which can be unrealistic. This paper revisits topic modeling based on second-order moments, and proposes an anchor-free topic mining framework. The proposed approach guarantees the identification of the topics under a much milder condition compared to the anchor-word assumption, thereby exhibiting much better robustness in practice. The associated algorithm only involves one eigen-decomposition and a few small linear programs. This makes it easy to implement and scale up to very large problem instances. Experiments using the TDT2 and Reuters-21578 corpus demonstrate that the proposed anchor-free approach exhibits very favorable performance (measured using coherence, similarity count, and clustering accuracy metrics) compared to the prior art.

#35 Optimal Learning for Multi-pass Stochastic Gradient Methods
Junhong Lin (Istituto Italiano di Tecnologia)
Lorenzo Rosasco

We analyze the learning properties of the stochastic gradient method when multiple passes over the data and mini-batches are allowed. In particular, we consider the square loss and show that for a universal step-size choice, the number of passes acts as a regularization parameter, and optimal finite sample bounds can be achieved by early-stopping. Moreover, we show that larger step-sizes are allowed when considering mini-batches. Our analysis is based on a unifying approach, encompassing both batch and stochastic gradient methods as special cases.

#36 Contextual semibandits via supervised learning oracles
Akshay Krishnamurthy
Alekh Agarwal (Microsoft)
Miro Dudik

We study an online decision making problem where on each round a learner chooses a list of items based on some side information, receives a scalar feedback value for each individual item, and a reward that is linearly related to this feedback. These problems, known as contextual semibandits, arise in crowd-sourcing, recommendation, and many other domains. This paper reduces contextual semibandits to supervised learning, so that we can leverage powerful supervised learning methods in this partial-feedback setting. Our first reduction, which applies when the mapping from feedback to reward is known, leads to a computationally efficient algorithm with a near-optimal regret guarantee. We show that this algorithm outperforms state-of-the-art approaches on real-world learning-to-rank datasets, demonstrating the advantage of oracle-based algorithms. We also develop and analyze a novel algorithm for the setting where the linear transformation is unknown.

#37 One-vs-Each Approximation to Softmax for Scalable Estimation of Probabilities
Michalis Titsias RC AUEB

The softmax representation of probabilities for categorical variables plays a prominent role in modern machine learning with numerous applications is areas such as large scale classification, neural language modeling and recommendation systems. However, softmax estimation is very expensive for large scale inference because of the high cost associated with computing the normalizing constant. Here, we introduce an efficient approximation to softmax probabilities which takes the form of a rigorous lower bound on the exact probability. This bound takes the form of a product over one-vs-one pairwise probabilities and it leads to scalable estimation based on stochastic optimization. It allows to perform doubly stochastic estimation by subsampling both training instances and class labels. We show that the new bound has interesting theoretical properties and we demonstrate its use in classification problems.

#38 Satisfying Real-world Goals with Dataset Constraints
Gabe Goh (UC Davis)
Andy Cotter
Maya Gupta
Michael P Friedlander (UC Davis)

The goal of minimizing misclassification error on a training set is often just one of several real-world goals that might be defined on different datasets. For example, one may require a classifier to also make positive predictions at some specified rate for some subpopulation (fairness), or to achieve a specified empirical recall. Other real-world goals include reducing churn with respect to a previously deployed model, or stabilizing online training. In this paper we propose handling multiple goals on multiple datasets by training with dataset constraints, using the ramp penalty to accurately quantify costs, and present an efficient algorithm to approximately optimize the resulting non-convex constrained optimization problem. Experiments on both benchmark and real-world industry datasets demonstrate the effectiveness of our approach.

#39 Blind Regression: Nonparametric Regression for Latent Variable Models via Collaborative Filtering
Dogyoon Song (MIT)
Christina E. Lee (MIT)
Yihua Li (MIT)
Devavrat Shah

We introduce the framework of blind regression motivated by matrix completion for recommendation systems: given n users, m movies, and a subset of user-movie ratings, the goal is to predict the unobserved user-movie ratings given the data, i.e., to complete the partially observed matrix. Following the framework of non-parametric statistics, we posit that user u and movie i have features x_1(u) and x_2(i) respectively, and their corresponding rating y(u,i) is a noisy measurement of f(x_1(u), x_2(i)) for some unknown function f. In contrast with classical regression, the features x = (x_1(u), x_2(i)) are not observed, making it challenging to apply standard regression methods to predict the unobserved ratings. Inspired by the classical Taylor's expansion for differentiable functions, we provide a prediction algorithm that is consistent for all Lipschitz functions. In fact, the analysis through our framework naturally leads to a variant of collaborative filtering, shedding insight into the widespread success of collaborative filtering in practice. Assuming each entry is sampled independently with probability at least \max(m^{-1/2+δ},n^{-1+δ}) with $\delta > 0$, we prove that the expected fraction of our estimates with error greater than ϵ is less than Γ^2 / ϵ^2 plus a polynomially decaying term, where Γ^2 is the variance of the additive entry-wise noise term. Experiments with the MovieLens and Netflix datasets suggest that our algorithm provides principled improvements over basic collaborative filtering and is competitive with matrix factorization methods. The algorithm and analysis naturally extend to higher order tensor completion by simply flattening the tensor into a matrix. We show that our simple and principled approach is competitive with respect to state-of-art Tensor completion algorithms when applied to image inpainting data.

#40 Generative Adversarial Imitation Learning
Jonathan Ho (Stanford)
Stefano Ermon

Consider learning a policy from example expert behavior, without interaction with expert or access to reinforcement signal. One approach is to recover the expert's cost function with inverse reinforcement learning, then extract a policy from that cost function with reinforcement learning. This approach is indirect and can be slow. We propose a new general framework for directly extracting a policy from data, as if it were obtained by reinforcement learning following inverse reinforcement learning. We show that a certain instantiation of our framework draws an analogy between imitation learning and generative adversarial networks, from which we derive a model-free imitation learning algorithm that obtains significant performance gains over existing model-free methods in imitating complex behaviors in large, high-dimensional environments.

#41 Fast Active Set Methods for Online Spike Inference from Calcium Imaging
Johannes Friedrich (Columbia Univ.)
Liam Paninski

Fluorescent calcium indicators are a popular means for observing the spiking activity of large neuronal populations. Unfortunately, extracting the spike train of each neuron from raw fluorescence calcium imaging data is a nontrivial problem. We present a fast online active set method to solve the sparse nonnegative deconvolution problem for spike inference. Importantly, the algorithm progresses through each time series sequentially from beginning to end, thus enabling real-time online spike inference during the imaging session. Our algorithm is inspired by the pool adjacent violators algorithm for isotonic regression and inherits its linear scaling but replaces the monotone function by an AR(p) process with nonnegative jumps to account for the calcium dynamics. We gain remarkable decreases in processing time by more than one order of magnitude compared to currently employed state of the art convex solvers relying on interior point methods. Our method can exploit warm starts, therefore optimizing the AR hyperparameters only requires a handful of passes through the data. The algorithm enables real-time simultaneous deconvolution of O(10^5) traces from whole-brain zebrafish data on a laptop.

#42 Path-Normalized Optimization of Recurrent Neural Networks with ReLU Activations
Behnam Neyshabur (TTI-Chicago)
Yuhuai Wu (Univ. of Toronto)
Russ Salakhutdinov (Univ. of Toronto)
Nati Srebro

We investigate the parameter-space geometry of recurrent neural networks (RNNs), and develop an adaptation of path-SGD optimization method, attuned to this geometry, that can learn plain RNNs with ReLU activations. On several datasets that require capturing long-term dependency structure, we show that path-SGD can significantly improve trainability of ReLU RNNs compared to RNNs trained with SGD, even with various recently suggested initialization schemes.

#43 Improved Regret Bounds for Oracle-Based Adversarial Contextual Bandits
Vasilis Syrgkanis
Haipeng Luo (Princeton Univ.)
Akshay Krishnamurthy
Robert Schapire

We give an oracle-based algorithm for the adversarial contextual bandit problem, where either contexts are drawn i.i.d. or the sequence of contexts is known a priori, but where the losses are picked adversarially. Our algorithm is computationally efficient, assuming access to an offline optimization oracle, and enjoys a regret of order O((KT)^(2/3)(log(N))^(1/3)), where K is the number of actions, T is the number of iterations and N is the number of baseline policies. Our algorithm is the first to break the O(T^(3/4)) barrier that is achieved by recently introduced algorithms. Breaking this barrier was left as a major open problem. Our algorithm is based on the recent relaxation based approach of (Rakhlin and Sridharan, ICML'16).

#44 Diffusion-Convolutional Neural Networks
James Atwood (UMass Amherst)

We present diffusion-convolutional neural networks (DCNNs), a new model for graph-structured data. Through the introduction of a diffusion-convolution operation, we show how diffusion-based representations can be learned from graph-structured data and used as an effective basis for node classification. DCNNs have several attractive qualities, including a latent representation for graphical data that is invariant under isomorphism, as well as polynomial-time prediction and learning that can be represented as tensor operations and efficiently implemented on the GPU. Through several experiments with real structured datasets, we demonstrate that DCNNs are able to outperform probabilistic relational models and kernel-on-graph methods at relational node classification tasks.

#45 Faster Projection-free Convex Optimization over the Spectrahedron
Dan Garber
Dan Garber

Minimizing a convex function over the spectrahedron, i.e., the set of all d\times d positive semidefinite matrices with unit trace, is an important optimization task with many applications in optimization, machine learning, and signal processing. It is also notoriously difficult to solve in large-scale since standard techniques require to compute expensive matrix decompositions. An alternative, is the conditional gradient method (aka Frank-Wolfe algorithm) that regained much interest in recent years, mostly due to its application to this specific setting. The key benefit of the CG method is that it avoids expensive matrix decompositions all together, and simply requires a single eigenvector computation per iteration, which is much more efficient. On the downside, the CG method, in general, converges with an inferior rate. The error for minimizing a β-smooth function after t iterations scales like β/t. This rate does not improve even if the function is also strongly convex. In this work we present a modification of the CG method tailored for the spectrahedron. The per-iteration complexity of the method is essentially identical to that of the standard CG method: only a single eigenvecor computation is required. For minimizing an α-strongly convex and β-smooth function, the \textit{expected} error of the method after t iterations is: O\left({\min\{\frac{β}{t} ,\left({\frac{β√{\rank(\X^*)}}{α^{1/4}t}}\right)^{4/3}, \left({\frac{β}{√{α}Λ_{\min}(\X^*)t}}\right)^{2}\}}\right). Beyond the significant improvement in convergence rate, it also follows that when the optimum is low-rank, our method provides better accuracy-rank tradeoff than the standard CG method. To the best of our knowledge, this is the first result that attains provably faster convergence rates for a CG variant for optimization over the spectrahedron. We also present encouraging preliminary empirical results.

#46 Structured Matrix Recovery via the Generalized Dantzig Selector
Sheng Chen (Univ. of Minnesota)
Arindam Banerjee

In recent years, structured matrix recovery problems have gained considerable attention for its real world applications, such as recommender systems and computer vision. Much of the existing work has focused on matrices with low-rank structure, and limited progress has been made on matrices with other types of structure. In this paper we present non-asymptotic analysis for estimation of generally structured matrices via the generalized Dantzig selector based on sub-Gaussian measurements. We show that the estimation error can always be succinctly expressed in terms of a few geometric measures such as Gaussian widths of suitable sets associated with the structure of the underlying true matrix. Further, we derive general bounds on these geometric measures for structures characterized by unitarily invariant norms, a large family covering most matrix norms of practical interest. Examples are provided to illustrate the utility of our theoretical development.

#47 Convex Two-Layer Modeling with Latent Structure

Vignesh Ganapathiraman (Univ. Of Illinois at Chicago)
Xinhua Zhang (UIC)
Yaoliang Yu
Junfeng Wen (UofA)

Unsupervised learning of structured predictors has been a long standing pursuit in machine learning. Recently a conditional random field auto-encoder has been proposed in a two-layer setting, allowing latent structured representation to be automatically inferred. Aside from being nonconvex, it also requires the demanding inference of normalization. In this paper, we develop a convex relaxation of two-layer conditional model which captures latent structure and estimates model parameters, jointly and optimally. We further expand its applicability by resorting to a weaker form of inference---maximum a-posteriori. The flexibility of the model is demonstrated on two structures based on total unimodularity---graph matching and linear chain. Experimental results confirm the promise of the method.

#48 Finite-Sample Analysis of Fixed-k Nearest Neighbor Density Functionals Estimators

Shashank Singh (Carnegie Mellon Univ.)
Barnabas Poczos

We provide finite-sample analysis of a general framework for using k-nearest neighbor statistics to estimate functionals of a nonparametric continuous probability density, including entropies, divergences, and mutual informations. Rather than plugging a consistent density estimate (which requires $k \to \infty$ as the sample size $n \to \infty$) into the functional of interest, the estimators we consider fix k and perform a bias correction. This is more efficient computationally, and, as we show in certain cases, statistically, leading to faster convergence rates. Our framework unifies several previous estimators, for most of which ours are the first finite sample guarantees.

#49 Deep Learning Games

Dale Schuurmans
Martin A Zinkevich (Google)

We investigate a reduction of supervised learning to game playing that reveals new connections and learning methods. For convex one-layer problems, we demonstrate an equivalence between global minimizers of the training problem and Nash equilibria in a simple game. We then show how the game can be extended to general acyclic neural networks with differentiable convex gates, establishing a bijection between the Nash equilibria and critical (or KKT) points of the deep learning problem. Based on these connections we investigate alternative learning methods, and find that regret matching can achieve competitive training performance while producing sparser models than current deep learning approaches.

#50 "Congruent" and "Opposite" Neurons: Sisters for Multisensory Integration and Segregation

Wen-Hao Zhang (Institute of Neuroscience)
He Wang (HKUST)
K. Y. Michael Wong (HKUST)
Si Wu

Experiments reveal that in the dorsal medial superior temporal (MSTd) and the ventral intraparietal (VIP) areas, where visual and vestibular cues are integrated to infer heading direction, there exist two types of neurons with comparable numbers. One is ``congruent'' cells, whose preferred heading directions are similar in response to visual and vestibular cues; and the other is ``opposite'' cells, whose preferred heading directions are nearly ``opposite'' (with an offset of 180 degree) in response to visual vs. vestibular cues. Congruent neurons are known to be responsible for cue integration, but the computational role of opposite neurons remains largely unknown. Here, we propose that opposite neurons may serve to encode the disparity information between cues necessary for multisensory segregation. We build a computational model composed of two reciprocally coupled modules, MSTd and VIP, and each module consists of a set of congruent and opposite neurons. In the model, congruent neurons in two modules are reciprocally connected with each other in the congruent manner, whereas opposite neurons are reciprocally connected in the opposite manner. Mimicking the experimental protocol, our model reproduces the characteristics of congruent and opposite neurons, and demonstrates that in each module, the sisters of congruent and opposite neurons can jointly achieve optimal multisensory information integration and segregation. This study sheds light on our understanding of how the brain implements optimal multisensory integration and segregation concurrently in a distributed manner.

#51 Statistical Inference for Cluster Trees

Jisu KIM (Carnegie Mellon Univ.)
Yen-Chi Chen (Carnegie Mellon Univ.)
Sivaraman Balakrishnan (Carnegie Mellon Univ.)
Alessandro Rinaldo (Carnegie Mellon Univ.)
Larry Wasserman (Carnegie Mellon Univ.)

A cluster tree provides an intuitive summary of a density function that reveals essential structure about the high-density clusters. The true cluster tree is estimated from a finite sample from an unknown true density. This paper addresses the basic question of quantifying our uncertainty by assessing the statistical significance of different features of an empirical cluster tree. We first study a variety of metrics that can be used to compare different trees, analyzing their properties and assessing their suitability for our inference task. We then propose methods to construct and summarize confidence sets for the unknown true cluster tree. We introduce a partial ordering on cluster trees which we use to prune some of the statistically insignificant features of the empirical tree, yielding interpretable and parsimonious cluster trees. Finally, we provide a variety of simulations to illustrate our proposed methods and furthermore demonstrate their utility in the analysis of a Graft-versus-Host Disease (GvHD) data set.

#52 Minimizing Regret on Reflexive Banach Spaces and Nash Equilibria in Continuous Zero-Sum Games

Maximilian Balandat (UC Berkeley)
Walid Krichene (UC Berkeley)
Claire Tomlin (UC Berkeley)
Alexandre Bayen (UC Berkeley)

We study a general adversarial online learning problem, in which we are given a decision set X' in a reflexive Banach space X and a sequence of reward vectors in the dual space of X. At each iteration, we choose an action from X', based on the observed sequence of previous rewards. Our goal is to minimize regret, defined as the gap between the realized reward and the reward of the best fixed action in hindsight. Using results from infinite dimensional convex analysis, we generalize the method of Dual Averaging (or Follow the Regularized Leader) to our setting and obtain upper bounds on the worst-case regret that generalize many previous results. Under the assumption of uniformly continuous rewards, we obtain explicit regret bounds in a setting where the decision set is the set of probability distributions on a compact metric space S whose Radon-Nikodym derivatives are elements of L^p(S) for some $p > 1$. Importantly, we make no convexity assumptions on either the set S or the reward functions. We also

prove a general lower bound on the worst-case regret for any online algorithm. We then apply these results to the problem of learning in repeated two-player zero-sum games on compact metric spaces. In doing so, we first prove that if both players play a Hannan-consistent strategy, then with probability 1 the empirical distributions of play weakly converge to the set of Nash equilibria of the game. We then show that, under mild assumptions, Dual Averaging on the (infinite-dimensional) space of probability distributions indeed achieves Hannan-consistency.

#53 A Neural Transducer

Navdeep Jaitly
Quoc V Le
Oriol Vinyals
Ilya Sutskever
David Sussillo (Google)
Samy Bengio

Sequence-to-sequence models have achieved impressive results on various tasks. However, they are unsuitable for tasks that require incremental predictions to be made as more data arrives or tasks that have long input sequences and output sequences. This is because they generate an output sequence conditioned on an entire input sequence. In this paper, we present a Neural Transducer that can make incremental predictions as more input arrives, without redoing the entire computation. Unlike sequence-to-sequence models, the Neural Transducer computes the next-step distribution conditioned on the partially observed input sequence and the partially generated sequence. At each time step, the transducer can decide to emit zero to many output symbols. The data can be processed using an encoder and presented as input to the transducer. The discrete decision to emit a symbol at every time step makes it difficult to learn with conventional backpropagation. It is however possible to train the transducer by using a dynamic programming algorithm to generate target discrete decisions. Our experiments show that the Neural Transducer works well in settings where it is required to produce output predictions as data come in. We also find that the Neural Transducer performs well for long sequences even when attention mechanisms are not used.

#54 Feature selection for classification of functional data using recursive maxima hunting

José L. Torrecilla (Universidad Autónoma de Madrid)
Alberto Suárez

Dimensionality reduction is one of the key issues in the design of machine learning methods for automatic induction from functional data. In this context, variable selection techniques are especially attractive because they facilitate the interpretation of the predictive models and can lead to performance improvements. In this work, we introduce recursive maxima hunting (RMH) for variable selection in functional data classification problems. The method, which is a recursive extension of maxima hunting (MH), performs variable selection by identifying the maxima of a relevance function that measures the strength of the correlation of the class label with the predictive functional variable. At each stage the information associated with the selected maximum is removed by subtracting the conditional expectation of the process. Extensive empirical evaluation shows that RMH has equivalent or higher predictive accuracy than standard dimensionality reduction techniques, such as PCA and PLS and functional feature selection methods (MH) in the problems investigated.

#55 Homotopy Smoothing for Non-Smooth Problems with Lower Complexity than O(1/ϵ)

Yi Xu (The Univ. of Iowa)
Yan Yan (Univ. of Technology Sydney)
Qihang Lin
Tianbao Yang (Univ. of Iowa)

n this paper, we develop a novel {\bf ho}moto{\bf p}y {\bf s}moothing (HOPS) algorithm for solving a family of non-smooth problems that is composed of a non-smooth term with an explicit max-structure and a smooth term or a simple non-smooth term whose proximal mapping is easy to compute. Such kind of non-smooth optimization problems arise in many applications, e.g., machine learning, image processing, statistics, cone programming, etc. The best known iteration complexity for solving such non-smooth optimization problems is O(1/ϵ) without any assumption on the strong convexity. In this work, we will show that the proposed HOPS achieved a lower iteration complexity of \tilde O(1/ϵ^{1-\theta}) with \theta\in(0,1] capturing the local sharpness of the objective function around the optimal solutions. To the best of our knowledge, this is the lowest iteration complexity achieved so far for the considered non-smooth optimization problems without strong convexity assumption. The HOPS algorithm uses Nesterov's smoothing trick and Nesterov's accelerated gradient method and runs in stages, which gradually decreases the smoothing parameter until it yields a sufficiently good approximation of the original function. Experimental results verify the effectiveness of HOPS in comparison with Nesterov's smoothing algorithm and the primal-dual style of first-order methods.

#56 Nested Mini-Batch K-Means

James Newling (Idiap Research Institute)
François Fleuret (Idiap Research Institute)

A new algorithm is proposed which accelerates the mini-batch k-means algorithm of Sculley (2010) by using the distance bounding approach of Elkan (2003). We argue that, when incorporating distance bounds into a mini-batch algorithm, already used data should preferentially be reused. To this end we propose using nested mini-batches, whereby data in a mini-batch at iteration t is automatically reused at iteration t+1. Using nested mini-batches presents two difficulties. The first is that unbalanced use of data can bias estimates, which we resolve by ensuring that each data sample contributes exactly once to centroids. The second is in choosing mini-batch sizes, which we address by balancing premature fine-tuning of centroids with redundancy induced slow-down. Experiments show that the resulting nmbatch algorithm is very effective, often arriving within 1\% of the empirical minimum 100 times earlier than the standard mini-batch algorithm.

#57 Density Estimation via Discrepancy Based Adaptive Sequential Partition

Dangna Li (Stanford Univ.)
Kun Yang (Google Inc)
Wing Hung Wong (Stanford Univ.)

Given iid observations from an unknown continuous distribution defined on some domain Ω, we propose a nonparametric method to learn a piecewise constant function to approximate the underlying probability density function. Our density estimate is a piecewise constant function defined on a binary partition of Ω. The key ingredient of the algorithm is to use discrepancy, a concept originates from Quasi Monte Carlo analysis, to control the partition process. The resulting algorithm is simple, efficient, and has provable convergence rate. We demonstrate empirically its efficiency as a density estimation method. We also show how it can be utilized to find good initializations for k-means.

#58 Budgeted stream-based active learning via adaptive submodular maximization

Kaito Fujii (Kyoto Univ.)
Hisashi Kashima (Kyoto Univ.)

Active learning enables us to reduce the annotation cost by adaptively selecting unlabeled instances to be labeled. For pool-based active learning, several effective methods with theoretical guarantees have been developed through maximizing some utility function satisfying adaptive submodularity. In contrast, there have been few methods for stream-based active learning based on adaptive submodularity. In this paper, we propose a new class of utility functions, policy-adaptive submodular functions, and prove this class includes many existing adaptive submodular functions appearing in real world problems. We provide a general framework based on policy-adaptive submodularity that makes it possible to convert existing pool-based methods to stream-based methods and give theoretical guarantees on their performance. In addition we empirically demonstrate their effectiveness comparing with existing heuristics on common benchmark datasets.

#59 Lifelong Learning with Weighted Majority Votes

Anastasia Pentina (IST Austria)
Ruth Urner (MPI Tuebingen)

Better understanding of the potential benefits of information transfer and representation learning is an important step towards the goal of building intelligent systems that are able to persist in the world and learn over time. In this work, we consider a setting where the learner encounters a stream of tasks but is able to retain only limited information from each encountered task, such as a learned predictor. In contrast to most previous works analyzing this scenario, we do not make any distributional assumptions on the task generating process. Instead, we formulate a complexity measure that captures the diversity of the observed tasks. We provide a lifelong learning algorithm with error guarantees for every observed task (rather than on average). We show sample complexity reductions in comparison to solving every task in isolation in terms of our task complexity measure. Further, our algorithmic framework can naturally be viewed as learning a representation from encountered tasks with a neural network.

#60 How Deep is the Feature Analysis underlying Rapid Visual Categorization?

Sven2 Eberhardt (Brown Univ.)
Jonah G Cader (Brown Univ.)
Thomas Serre

Rapid categorization paradigms have a long history in experimental psychology: Characterized by short presentation times and speedy behavioral responses, these tasks highlight the efficiency with which our visual system processes natural object categories. Previous studies have shown that feed-forward hierarchical models of the visual cortex provide a good fit to human visual decisions. At the same time, recent work in computer vision has demonstrated significant gains in object recognition accuracy with increasingly deep hierarchical architectures. But it is unclear how well these models account for human visual decisions and what they may reveal about the underlying brain processes. We have conducted a large-scale psychophysics study to assess the correlation between computational models and human participants on a rapid animal vs. non-animal categorization task. We considered visual representations of varying complexity by analyzing the output of different stages of processing in three state-of-the-art deep networks. We found that recognition accuracy increases with higher stages of visual processing (higher level stages indeed outperforming human participants on the same task) but that human decisions agree best with predictions from intermediate stages. Overall, these results suggest that human participants may rely on visual features of intermediate complexity and that the complexity of visual representations afforded by modern deep network models may exceed those used by human participants during rapid categorization.

#61 Incremental Boosting Convolutional Neural Network for Facial Action Unit Recognition

Shizhong Han (Univ. of South Carolina)
Zibo Meng (Univ. of South Carolina)
AHMED-SHEHAB KHAN (Univ. of South Carolina)
Yan Tong (Univ. of South Carolina)

Recognizing facial action units (AUs) from spontaneous facial expressions is still a challenging problem. Most recently, CNNs have shown promise on facial AU recognition. However, the learned CNNs are often overfitted and do not generalize well to unseen subject due to limited AU-coded training images. We proposed a novel Incremental Boosting CNN (IB-CNN) to integrate boosting into the CNN via an incremental boosting layer that selects discriminative neurons from the lower layer and is incrementally updated on successive mini-batches. In addition, a novel loss function that accounts for errors from both the incremental boosted classifier and individual weak classifiers was proposed to fine-tune the IB-CNN. Experimental results on two benchmark AU databases have demonstrated that the IB-CNN yields significant improvement over the traditional CNN and the one without incremental learning, as well as outperforming the state-of-the-art CNN-based methods in AU recognition. The improvement is more impressive for the AUs that have the lowest frequencies in the databases.

#62 Multivariate tests of association based on univariate tests

Ruth Heller (Tel-Aviv Univ.)
Yair Heller

For testing two vector random variables for independence, we propose testing whether the distance of one vector from an arbitrary center point is independent from the distance of the other vector from another arbitrary center point by a univariate test. We prove that under minimal assumptions, it is enough to have a consistent univariate independence test on the distances, to guarantee that the power to detect dependence between the random vectors increases to one with sample size. If the univariate test is distribution-free, the multivariate test will also be distribution-free. If we consider multiple center points and aggregate the center-specific univariate tests, the power may be further improved, and the resulting multivariate test may be distribution-free for specific aggregation methods (if the univariate test is distribution-free). We show that certain multivariate tests recently proposed in the literature can be viewed as instances of this general approach. Moreover, we show in experiments that novel tests constructed using our approach can have better power and computational time than competing approaches.

#63 SURGE: Surface Regularized Geometry Estimation from a Single Image

Peng Wang (UCLA)
Xiaohui Shen (Adobe Research)
Bryan Russell
Scott Cohen (Adobe Research)
Brian Price
Alan L Yuille

Estimating geometry from a single image, i.e. predicting a normal and depth for each pixel, is a fundamental yet challenging problem in computer vision. In this work, we augment the task by predicting additional 3D planar surface information, and incorporate such information in a principled way to generate better regularized geometry estimation. Specifically, the proposed framework is composed of two major components: multiple streams of convolutional neural networks (CNNs) that predict depth, normals, a binary plane map, and an edge map, followed by a dense conditional random field (DCRF) that fuses the four types of predictions, where depth and normals are made compatible and are regularized by the plane and edge information. Joint training of the two components are enabled where the errors from the DCRF are back-propagated through the normal and depth CNNs. In addition, we propose new planar-wise metrics to evaluate geometry consistency within planar surfaces, which are more tightly related to dependent 3D editing applications. Our approach gives about a 30\% relative improvement in planar consistency over previous state-of-the-art methods on the NYU v2 dataset.

#64 Memory-Efficient Backpropagation Through Time

Audrunas Gruslys (Google DeepMind)
Remi Munos (Google DeepMind)
Ivo Danihelka
Marc Lanctot (Google DeepMind)
Alex Graves

We propose a novel approach to reduce memory consumption of the backpropagation algorithm when training large recurrent neural networks (RNNs). Our approach uses dynamic programming to balance a trade-off between caching of intermediate results and re-computation. The algorithm is capable of fitting within almost any set memory budget while finding an optimal execution policy minimizing the computational cost. The algorithm has a worst-case asymptotic time complexity of O(m \cdot t^{1+\frac{1}{m}}) for a memory-scarce regime, where m is the number of hidden states for a long sequence length t > \frac{m^m}{m!}. For sequences where t < 10^5 and m < 10^3, the computational cost is bounded by 3 t^{1+\frac{1}{m}} forward operations. For sequences of length 1000, our algorithm saves 95\% of memory usage while using only one third more time per training iteration than the standard BPTT.

#65 Scan Order in Gibbs Sampling: Models in Which it Matters and Bounds on How Much

Bryan He (Stanford Univ.)
Christopher M De Sa (Stanford Univ.)
Ioannis Mitliagkas
Christopher Ré (Stanford Univ.)

Gibbs sampling is a Markov Chain Monte Carlo sampling technique that iteratively samples variables from their conditional distributions. There are two common scan orders for the variables: random scan and systematic scan. Due to the benefits of locality in hardware, systematic scan is commonly used, even though most statistical guarantees are only for random scan. While it has been conjectured that the mixing times of random scan and systematic scan do not differ by more than a logarithmic factor, we show by counterexample that this is not the case, and we prove under mild conditions that the mixing times do not differ by more than a polynomial factor. To do so, we introduce a method of augmenting the state space to study systematic scan using conductance.

#66 2-Component Recurrent Neural Networks

Xiang Li (NJUST)
Tao Qin (Microsoft)
Jian Yang
Xiaolin Hu
Tieyan Liu (Microsoft Research)

Recurrent neural networks (RNNs) have achieved state-of-the-art performances in many natural language processing tasks, such as language modeling and machine translation. However, when the vocabulary is large, the RNN model will become very big (e.g., possibly beyond the memory capacity of a GPU processor) and its training will become inefficient. In this work, we propose a novel technique to tackle this challenge. The key idea is to use 2-Component (2C) shared embedding for word representations. We allocate every word in the vocabulary into a table, each row of which is associated with a vector, and each column associated with another vector. Depending on its position in the table, a word is jointly represented by two components: a row vector and a column vector. Since words in the same row share the row vector and words in the same column share the column vector, we only need 2 √{|V|} unique vectors to represent a vocabulary with |V| words. In contrast, existing approaches require |V| unique vectors. Based on the 2C shared embedding, we design a new RNN algorithm (2C-RNN) and evaluate it using the language modeling task on several benchmark datasets. The results show that our algorithm significantly shrinks the model size and speeds up the training process, without sacrifice of accuracy (it achieves similar, if not better, perplexity as compared to state-of-the-art language models). Remarkably, on the One-Billion-Word benchmark Dataset, 2C-RNN achieves comparable perplexity to previous language models, whilst reducing the model size by a factor of 80-200, and speeding up the training process by a factor of 10.

#67 Direct Feedback Alignment Provides Learning in Deep Neural Networks

Arild Nøkland (None)

Artificial neural networks are most commonly trained with the back-propagation algorithm, where the gradient for learning is provided by back-propagating the error, layer by layer, from the output layer to the hidden layers. A recently discovered method called feedback-alignment shows that the weights used for propagating the error backward don't have to be symmetric with the weights used for propagation the activation forward. In fact, random feedback weights work evenly well, because the network learns how to make the feedback useful. In this work, the feedback alignment principle is used for training hidden layers more independently from the rest of the network, and from a zero initial condition. The error is propagated through fixed random feedback connections directly from the output layer to each hidden layer. This simple method is able to achieve zero training error even in convolutional networks and very deep networks, completely without error back-propagation. The method is a step towards biologically plausible machine learning because the error signal is almost local, and no symmetric or reciprocal weights are required. Experiments show that the test performance on MNIST and CIFAR is almost as good as those obtained with back-propagation for fully connected networks. If combined with dropout, the method achieves 1.45% error on the permutation invariant MNIST task.

#68 Variational Bayes on Monte Carlo Steroids

Aditya Grover (Stanford Univ.)
Stefano Ermon

Variational approaches are often used to approximate intractable posteriors or normalization constants in hierarchical latent variable models. While often effective in practice, it is known that the approximation error can be arbitrarily large. We propose a new class of bounds on the marginal likelihood of directed latent variable models. Our approach relies on random projections to simplify the posterior. In contrast with standard variational methods, our bounds are guaranteed to be tight with high probability. We provide a new approach for learning latent variables models based on optimizing our new bounds on the likelihood. We demonstrate empirically improvements on benchmark datasets in vision and language for sigmoid belief networks, where a neural network is used to approximate the posterior.

#69 Agnostic Estimation for Misspecified Phase Retrieval Models

Matey Neykov (Princeton Univ.)
Zhaoran Wang (Princeton Univ.)
Han Liu

The goal of noisy high-dimensional phase retrieval is to estimate an s-sparse parameter {β}^*\in R^d from n realizations of the model Y = (X^{\top} {β}^*)^2 + \varepsilon. Based on this model, we propose a significant semi-parametric generalization called \textit{misspecified phase retrieval} (MPR), in which Y = f(X^{\top}{β}^*, \varepsilon) with unknown f and \operatorname{Cov}(Y, (X^{\top}{β}^*)^2) > 0. For example, MPR encompasses Y = h(|X^{\top} {β}^*|) + \varepsilon with increasing h as a special case. Despite the generality of the MPR model, it eludes the reach of most existing semi-parametric estimators. In this paper, we propose an estimation procedure called AGENT (AGnostic EstimatioN for misspecified phase reTrieval), which consists of solving a cascade of two convex programs and provably recovers the direction of {β}^*. Furthermore, we prove that AGENT is minimax optimal over the class of MPR models. Interestingly, our minimax analysis characterizes the statistical price of misspecifying the link function in phase retrieval models. Our theory is backed up by thorough numerical results.

#70 Following the Leader and Fast Rates in Linear Prediction: Curved Constraint Sets and Other Regularities

Ruitong Huang (Univ. of Alberta)
Tor Lattimore
András György
Csaba Szepesvari (U. Alberta)

The follow the leader (FTL) algorithm, perhaps the simplest of all online learning algorithms, is known to perform well when the loss functions it is used on are positively curved. In this paper we ask whether there are other ``lucky" settings when FTL achieves sublinear, ``small" regret. In particular, we study the fundamental problem of linear prediction over a convex, non-empty compact domain. Amongst other results, we prove that the curvature of the boundary of the domain can act as if the losses were curved: In this case, we prove that as long as the mean of the loss vectors have positive lengths bounded away from zero, FTL enjoys a logarithmic growth rate of regret, while, e.g., for polyhedral domains and stochastic data it enjoys finite expected regret. Building on a previously known meta-algorithm, we also get an algorithm that simultaneously enjoys the worst-case guarantees and the bound available for FTL.

#71 Combining Fully Convolutional and Recurrent Neural Networks for 3D Biomedical Image Segmentation

Jianxu Chen (Univ. of Notre Dame)
Lin Yang (Univ. of Notre Dame)
Yizhe Zhang (Univ. of Notre Dame)
Mark Alber (Univ. of Notre Dame)
Danny Z Chen (Univ. of Notre Dame)

Segmentation of 3D images is a fundamental problem in biomedical image analysis. Deep learning (DL) approaches have achieved the state-of-the-art segmentation performance. To exploit the 3D contexts using neural networks, known DL segmentation methods, including 3D convolution, 2D convolution on the planes orthogonal to 2D slices, and LSTM in multiple directions, all suffer incompatibility with the highly anisotropic dimensions in common 3D biomedical images. In this paper, we propose a new DL framework for 3D image segmentation, based on a combination of a fully convolutional network (FCN) and a recurrent neural network (RNN), which are responsible for exploiting the intra-slice and inter-slice contexts, respectively. To our best knowledge, this is the first DL framework for 3D image segmentation that explicitly leverages 3D image anisotropism. Evaluating using a dataset from the ISBI Neuronal Structure Segmentation Challenge and in-house image stacks for 3D fungus segmentation, our approach achieves promising results, comparing to the known DL-based 3D segmentation approaches.

#72 The Product Cut

Thomas Laurent (Loyola Marymount Univ.)
James von Brecht (CSULB)
Xavier Bresson
arthur szlam

We introduce a theoretical and algorithmic framework for multi-way graph partitioning that relies on a multiplicative cut-based objective. We refer to this objective as the Product Cut. We provide a detailed investigation of the mathematical properties of this objective and an effective algorithm for its optimization. The proposed model has strong mathematical underpinnings, and the corresponding algorithm achieves state-of-the-art performance on benchmark data sets.

#73 Stochastic Gradient Methods for Distributionally Robust Optimization with f-divergences

Hong Namkoong (Stanford Univ.)
John C Duchi

We develop efficient solution methods for a robust empirical risk minimization problem designed to give calibrated confidence intervals on the performance of parameter vectors. Our methods apply to distributionally robust optimization probems proposed by Ben-Tal et al.\ that put more weight on observations inducing high loss via a worst-case approach over a non-parametric uncertainty set on the underlying data distribution. Our algorithm solves the resulting minimax problems with nearly the same computational cost of stochastic gradient descent through the use of several carefully designed data structures. More precisely, for a sample of size n, the per-iteration cost of our method scales as O(log n), which allows us to give statistical optimality certificates that distributionally robust optimization provides at little extra cost compared to empirical risk minimization and stochastic gradient methods.

#74 Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings

Tolga Bolukbasi (Boston Univ.)
Kai-Wei Chang
James Y Zou
Venkatesh Saligrama
Adam T Kalai (Microsoft Research)

In this paper, we study gender stereotypes in word embedding, a popular framework to represent text data. Even when trained on Google News articles, the resulting word embeddings exhibit a striking amount of gender bias. This raises concerns about their widespread adoption because applications can inadvertently amplify unwanted stereotypes. To systematically evaluate stereotypes, we created a novel gender analogy task and combined it with crowdsourcing to quantify the gender bias in a given embedding. Interestingly, we show that gender stereotypes are captured well by a low-dimensional subspace within the embedding. We develop algorithms that transform an input word embedding and remove gender stereotypes, such as the association between between the words receptionist and female, while maintaining desired associations such as between the words sister and female. Moreover we show that our algorithm preserves the utility of the embedding, as measured through standard analogy and similarity metrics.

#75 Optimal spectral transportation with application to music transcription

Rémi Flamary
Cédric Févotte (CNRS)
Nicolas Courty
Valentin Emiya (Aix-Marseille Univ.)

Many spectral unmixing methods rely on the non-negative decomposition of spectral data onto a dictionary of spectral templates. In particular, state-of-the-art music transcription systems decompose the spectrogram of the input signal onto a dictionary of representative note spectra. The typical measures of fit used to quantify the adequacy of the decomposition compare the data and template entries frequency-wise. As such, small displacements of energy from a frequency bin to another as well as variations of timber can disproportionally harm the fit. We address these issues by means of optimal transportation and propose a new measure of fit that treats the frequency distributions of energy holistically as opposed to frequency-wise. Building on the harmonic nature of sound, the new measure is invariant to shifts of energy to harmonically-related frequencies, as well as to small and local displacements of energy. Equipped with this new measure of fit, the dictionary of note templates can be considerably simplified to a set of Dirac vectors located at the target fundamental frequencies (musical pitch values). This in turns gives ground to a very fast and simple decomposition algorithm that achieves state-of-the-art performance on real musical data.

#76 Combining Adversarial Guarantees and Stochastic Fast Rates in Online Learning

Wouter M Koolen
Peter Grünwald (CWI)
Tim van Erven

We consider online learning algorithms that guarantee worst-case regret rates in adversarial environments (so they can be deployed safely and will perform robustly), yet adapt optimally to favorable stochastic environments (so they will perform well in a variety of settings of practical importance). We quantify the friendliness of stochastic environments by means of the well-known Bernstein (a.k.a. generalized Tsybakov margin) condition. For two recent algorithms (Squint for the Hedge setting and MetaGrad for online convex optimization) we show that the particular form of their data-dependent individual-sequence regret guarantees implies that they adapt automatically to the Bernstein parameters of the stochastic environment. We prove that these algorithms attain fast rates in their respective settings both in expectation and with high probability.

#77 Towards Conceptual Compression

Karol Gregor
Frederic Besse (Google DeepMind)
Danilo Jimenez Rezende
Ivo Danihelka
Daan Wierstra (Google DeepMind)

We introduce convolutional DRAW, a homogeneous deep generative model achieving state-of-the-art performance in latent variable image modeling. The algorithm naturally stratifies information into higher and lower level details, creating abstract features and as such addressing one of the fundamentally desired properties of representation learning. Furthermore, the hierarchical ordering of its latents creates the opportunity to store select global information about an image, yielding a high quality `conceptual compression' framework.

#78 Can Peripheral Representations Improve Clutter Metrics on Complex Scenes?

Arturo Deza (UCSB)
Miguel Eckstein (UCSB)

Previous studies have proposed image-based clutter measures that correlate with human search times and/or eye movements. However, most models do not take into account the fact that the effects of clutter interact with the foveated nature of the human visual system: visual clutter further from the fovea has an increasing detrimental influence on perception. Here, we introduce a new foveated clutter model to predict the detrimental effects in target search utilizing a forced fixation search task. We use Feature Congestion (Rosenholtz et al.) as our non foveated clutter model, and we stack a peripheral architecture on top of Feature Congestion for our foveated model. We introduce the Peripheral Integration Feature Congestion (PIFC) coefficient, as a fundamental ingredient of our model that modulates clutter as a non-linear gain contingent on eccentricity. We finally show that Foveated Feature Congestion (FFC) clutter scores ($r(44) = -0.82 \pm 0.04$, $p < 0.0001$) correlate better with target detection (hit rate) than regular Feature Congestion ($r(44) = -0.19 \pm 0.13$, $p = 0.0774$) in forced fixation search. Thus, our model allows us to enrich clutter perception research by computing fixation specific clutter maps. A toolbox for creating peripheral architectures: Piranhas: Peripheral Architectures for Natural, Hybrid and Artificial Systems will be made available.

#79 GAP Safe Screening Rules for Sparse-Group Lasso

Eugene Ndiaye (Télécom ParisTech)
Olivier Fercoq
Alexandre Gramfort
Joseph Salmon

In high dimensional settings, sparse structures are crucial for efficiency, both in term of memory, computation and performance. It is customary to consider ℓ_1 penalty to enforce sparsity in such scenarios. Sparsity enforcing methods, the Lasso being a canonical example, are popular candidates to address high dimension. For efficiency, they rely on tuning a parameter trading data fitting versus sparsity. For the Lasso theory to hold this tuning parameter should be proportional to the noise level, yet the latter is often unknown in practice. A possible remedy is to jointly optimize over the regression parameter as well as over the noise level. This has been considered under several names in the literature: Scaled-Lasso, Square-root Lasso, Concomitant Lasso estimation for instance, and could be of interest for confidence sets or uncertainty quantification. In this work, after illustrating numerical difficulties for the Smoothed Concomitant Lasso formulation, we propose a modification we coined Smoothed Concomitant Lasso, aimed at increasing numerical stability. We propose an efficient and accurate solver leading to a computational cost no more expansive than the one for the Lasso. We leverage on standard ingredients behind the success of fast Lasso solvers: a coordinate descent algorithm, combined with safe screening rules to achieve speed efficiency, by eliminating early irrelevant features.

#80 Learning Treewidth-Bounded Bayesian Networks with Thousands of Variables

Mauro Scanagatta (Idsia)
Giorgio Corani (Idsia)
Cassio P de Campos (Queen's Univ. Belfast)
Marco Zaffalon (IDSIA)

We present a method for learning treewidth-bounded Bayesian networks from data sets containing thousands of variables. Bounding the treewidth of a Bayesian network greatly reduces the complexity of inferences. Yet, being a global property of the graph, it considerably increases the difficulty of the learning process. Our novel algorithm accomplishes this task, scaling both to large domains and to large treewidths. Our novel approach consistently outperforms the state of the art on experiments with up to thousands of variables.

#81 Ancestral Causal Inference

Sara Magliacane (VU Univ. Amsterdam)
Tom Claassen
Joris M Mooij (Radboud Univ. Nijmegen)

Constraint-based causal discovery from limited data is a notoriously difficult challenge due to the many borderline independence test decisions. Several approaches to improve the reliability of the predictions by exploiting redundancy in the independence information have been proposed recently. Though promising, existing approaches can still be greatly improved in terms of accuracy and scalability. We present a novel method that reduces the combinatorial explosion of the search space by using a more coarse-grained representation of causal information, drastically reducing computation time. Additionally, we propose a method to score causal predictions based on their confidence. Crucially, our implementation also allows one to easily combine observational and interventional data and to incorporate various types of available background knowledge. We prove soundness and asymptotic consistency of our method and demonstrate that it can outperform the state-of-the-art on synthetic data, achieving a speedup of several orders of magnitude. We illustrate its practical feasibility by applying it on a challenging protein data set.

#82 Visual Question Answering with Question Representation Update

Ruiyu Li (CUHK)
Jiaya Jia (CUHK)

We propose a framework for reasoning over natural language questions and visual images. Given a natural language question about an image, our model updates the question representation iteratively by selecting image regions relevant to the query and learns to give the correct answer. Our model contains several reasoning layers, exploiting the complex logical relation in visual question answering (VQA) task. The proposed network is end-to-end trainable through back-propagation, where its weights are initialized using pre trained convolutional neural network (CNN) and gated recurrent unit (GRU). We evaluate our method on challenging datasets of COCO-QA and VQA, demonstrating competitive performance.

#83 Identification and Overidentification of Linear Structural Equation Models

Bryant Chen (UCLA)

In this paper, we address the problems of identifying linear structural equation models and discovering the constraints they imply. We first extend the half-trek criterion to cover a broader class of models and apply our extension to finding testable constraints implied by the model. We then show that any semi-Markovian linear model can be recursively decomposed into simpler sub-models, resulting in improved identification and constraint discovery power. Finally, we show that, unlike the existing methods developed for linear models, the resulting method subsumes the identification and constraint discovery algorithms for non-parametric models.

#84 On Valid Optimal Assignment Kernels and Applications to Graph Classification

Nils M. Kriege (TU Dortmund)
Pierre-Louis Giscard (Univ. of York)
Richard Wilson (Univ. of York)

The success of kernel methods has initiated the design of novel positive semidefinite functions, in particular for structured data. A leading design paradigm for this is the convolution kernel, which decomposes structured objects into their parts and sums over all pairs of parts. Assignment kernels, in contrast, are obtained from an optimal bijection between parts, which can provide a more valid notion of similarity. In general however, optimal assignments yield indefinite functions, which complicates their use in kernel methods. We characterize a class of base kernels used to compare parts that guarantees positive semidefinite optimal assignment kernels. These base kernels give rise to hierarchies from which the optimal assignment kernels are computed in linear time by histogram intersection. We apply these results by developing the Weisfeiler-Lehman optimal assignment kernel for graphs. It provides high classification accuracy on widely-used benchmark data sets improving over the original Weisfeiler-Lehman kernel.

#85 Constraints Based Convex Belief Propagation
Yaniv Tenzer (The Hebrew Univ.)
Alex Schwing
Kevin Gimpel
Tamir Hazan

Inference in Markov random fields subject to consistency structure is a fundamental problem that arises in many real-life applications. In order to enforce consistency, classical approaches utilize consistency potentials or encode constraints over feasible instances. Unfortunately this comes at the price of a serious computational bottleneck. In this paper we suggest to tackle consistency by incorporating constraints on beliefs. This permits derivation of a closed-form message-passing algorithm which we refer to as the Constraints Based Convex Belief Propagation (CBCBP). Experiments show that CBCBP outperforms the standard approach while being at least an order of magnitude faster.

#86 Combinatorial Energy Learning for Image Segmentation
Jeremy B Maitin-Shepard (Google)
Viren Jain (Google)
Michal Januszewski (Google)
Peter Li
Pieter Abbeel (OpenAI / UC Berkeley / Gradescope)

We introduce a new machine learning approach for image segmentation that uses a neural network to model the conditional energy of a segmentation given an image. Our approach, combinatorial energy learning for image segmentation (CELIS) places a particular emphasis on modeling the inherent combinatorial nature of dense image segmentation problems. We propose efficient algorithms for learning deep neural networks to model the energy function, and for local optimization of this energy in the space of supervoxel agglomerations. We extensively evaluate our method on a publicly available 3-D microscopy dataset with 25 billion voxels of ground truth data. On an 11 billion voxel test set, we find that our method improves volumetric reconstruction accuracy by more than 20% as compared to two state-of-the-art baseline methods: graph-based segmentation of the output of a 3-D convolutional neural network trained to predict boundaries, as well as a random forest classifier trained to agglomerate supervoxels that were generated by a 3-D convolutional neural network.

#87 A scalable end-to-end Gaussian process adapter for irregularly sampled time series classification
Steve Li (UMass Amherst)
Benjamin M Marlin

We present a general framework for classification of sparse and irregularly-sampled time series. The properties of such time series can result in substantial uncertainty about the values of the underlying temporal processes, while making the data difficult to deal with using standard classification methods that assume fixed-dimensional feature spaces. To address these challenges, we propose an uncertainty-aware classification framework based on a special computational layer we refer to as the Gaussian process adapter that can connect irregularly sampled time series data to to any black-box classifier learnable using gradient descent. We show how to scale up the required computations based on combining the structured kernel interpolation framework and the Lanczos approximation method, and how to discriminatively train the Gaussian process adapter in combination with a number of classifiers end-to-end using backpropagation.

#88 Stochastic Variance Reduction Methods for Saddle-Point Problems
Balamurugan Palaniappan
Francis Bach

We consider convex-concave saddle-point problems where the objective functions may be split in many components, and extend recent stochastic variance reduction methods (such as SVRG or SAGA) to provide the first large-scale linearly convergent algorithms for this class of problems which is common in machine learning. While the algorithmic extension is straightforward, it comes with challenges and opportunities: (a) the convex minimization analysis does not apply and we use the notion of monotone operators to prove convergence, showing in particular that the same algorithm applies to a larger class of problems, such as variational inequalities, (b) there are two notions of splits, in terms of functions, or in terms of partial derivatives, (c) the split does need to be done with convex-concave terms, (d) non-uniform sampling is key to an efficient algorithm, both in theory and practice, and (e) these incremental algorithms can be easily accelerated using a simple extension of the "catalyst" framework, leading to an algorithm which is always superior to accelerated batch algorithms.

#89 Dimensionality Reduction of Massive Sparse Datasets Using Coresets
Dan Feldman
Mikhail Volkov (MIT)
Daniela Rus (MIT)

In this paper we present a practical solution with performance guarantees to the problem of dimensionality reduction for very large scale sparse matrices. We show applications of our approach to computing the Principle Component Analysis (PCA) of any n\times d matrix, using one pass over the stream of its rows. Our solution uses coresets: a scaled subset of the n rows that approximates their sum of squared distances to {every} k-dimensional {affine} subspace. An open theoretical problem has been to compute such a coreset that is independent of both n and d. An open practical problem has been to compute a non-trivial approximation to the PCA of very large but sparse databases such as the Wikipedia document-term matrix in a reasonable time. We answer both of these questions affirmatively. Our main technical result is a new framework for deterministic coreset constructions based on a reduction to the problem of counting items in a stream.

#90 Efficient state-space modularization for planning: theory, behavioral and neural signatures
Daniel McNamee (Univ. of Cambridge)
Daniel M Wolpert (Univ. of Cambridge)
Mate Lengyel (Univ. of Cambridge)

Even in state-spaces of modest size, planning is plagued by the ``curse of dimensionality". This problem is particularly acute in human and animal cognition given the limited capacity of working memory, and the time pressures under which planning often occurs in the natural environment. Hierarchically organized modular representations have long been suggested to underlie the capacity of biological systems to efficiently and flexibly plan in complex environments. However, the principles underlying efficient modularization remain obscure, making it difficult to identify its behavioral and neural signatures. Here, we develop a normative theory of efficient state-space representations which partitions an environment into distinct modules by minimizing the average (information theoretic) description length of planning within the environment,

thereby optimally trading off the complexity of planning across and within modules. We show that such optimal representations provide a unifying account for a diverse range of hitherto unrelated phenomena at multiple levels of behavior and neural representation. We show that recently identified characteristics of route compression and the strong correlation between hippocampal activity and state ``degree centrality'' in spatial cognition arise naturally from the optimal modularization predicted by our theory. In addition, by extending the theory with an efficient neural encoding scheme for trajectories, we show that the theory also predicts the appearance of ``task-bracketing'' in goal-directed control, and ``start'' and ``stop'' signals in operant conditioning paradigms.

#91 Adaptive Newton Method for Empirical Risk Minimization to Statistical Accuracy

Aryan Mokhtari (Univ. of Pennsylvania)
Hadi Daneshmand (ETH Zurich)
Aurelien Lucchi
Thomas Hofmann
Alejandro Ribeiro (Univ. of Pennsylvania)

We consider empirical risk minimization for large-scale datasets. We introduce Ada Newton as an adaptive algorithm that uses Newton's method with adaptive sample sizes. The main idea of Ada Newton is to increase the size of the training set by a factor larger than one in a way that the minimization variable for the current training set is in the local neighborhood of the optimal argument of the next training set. This allows to exploit the quadratic convergence property of Newton's method and reach the statistical accuracy of each training set with only one iteration of Newton's method. We show theoretically and empirically that Ada Newton can double the size of the training set in each iteration to achieve the statistical accuracy of the full training set with about two passes over the dataset.

#92 RETAIN: Interpretable Predictive Model in Healthcare using Reverse Time Attention Mechanism

Edward Choi (Georgia Institute of Technolog)
Mohammad Taha Bahadori (Gatech)
Jimeng Sun

Accuracy and interpretation are two goals of any successful predictive models. Most existing works have to suffer the tradeoff between the two by either picking complex black box models such as recurrent neural networks (RNN) or relying on less accurate traditional models with better interpretation such as logistic regression. To address this dilemma, we present REverse Time AttentIoN model (RETAIN) for analyzing EHR data that achieves high accuracy while remaining clinically interpretable. RETAIN is a two-level neural attention model that can find influential past visits and significant clinical variables within those visits (e.g,. key diagnoses). RETAIN mimics physician practice by attending the EHR data in a reverse time order so that more recent clinical visits will likely get higher attention. Experiments on a large real EHR dataset of 14 million visits from 263K patients over 8 years confirmed the comparable predictive accuracy and computational scalability to the state-of-the-art methods such as RNN. Finally, we demonstrate the clinical interpretation with concrete examples from RETAIN.

#93 Joint quantile regression in vector-valued RKHSs

Maxime Sangnier (LTCI)
Olivier Fercoq
Florence d'Alché-Buc

Addressing the will to give a more complete picture than an average relationship provided by standard regression, a novel framework for estimating and predicting simultaneously several conditional quantiles is introduced. The proposed methodology leverages kernel-based multi-task learning to curb the embarrassing phenomenon of quantile crossing, with a one-step estimation procedure and no post-processing. Moreover, this framework comes along with theoretical guarantees and an efficient coordinate descent learning algorithm. Numerical experiments on benchmark and real datasets highlight the enhancements of our approach regarding the prediction error, the crossing occurrences and the training time.

#94 Learnable Visual Markers

Oleg Grinchuk (Skolkovo Institute of Science and Technology)
Vadim Lebedev (Skolkovo Institute of Science and Technology)
Victor Lempitsky

We propose a new approach to designing visual markers (analogous to QR-codes, markers for augmented reality, and robotic fiducial tags) based on the advances in deep generative networks. In our approach, the markers are obtained as color images synthesized by a deep network from input bit strings, whereas another deep network is trained to recover the bit strings back from the photos of these markers. The two networks are trained simultaneously in a joint backpropagation process that takes characteristic photometric and geometric distortions associated with marker fabrication and capture into account. Additionally, a stylization loss based on statistics of activations in a pretrained classification network can be inserted into the learning in order to shift the marker appearance towards some texture prototype. In the experiments, we demonstrate that the markers obtained using our approach are capable of retaining bit strings that are long enough to be practical. The ability to automatically adapt markers according to the usage scenario and the desired capacity as well as the ability to combine information encoding with artistic stylization are the unique properties of our approach. As a byproduct, our approach provides an insight on the structure of patterns that are most suitable for recognition by ConvNets and on their ability to distinguish composite patterns.

#95 Exponential expressivity in deep neural networks through transient chaos

Ben Poole (Stanford Univ.)
Subhaneil Lahiri (Stanford Univ.)
Maithreyi Raghu (Cornell Univ.)
Jascha Sohl-Dickstein
Surya Ganguli (Stanford)

We combine Riemannian geometry with the mean field theory of high dimensional chaos to study the nature of signal propagation in deep neural networks with random weights. Our results reveal a phase transition in the expressivity of random deep networks, with networks in the chaotic phase computing nonlinear functions whose global curvature grows exponentially with depth, but not with width. We prove that this generic class of random functions cannot be efficiently computed by any shallow network, going beyond prior work that restricts their analysis to single functions. Moreover, we formally quantify and demonstrate the long conjectured idea that deep networks can disentangle exponentially curved manifolds in input space into flat manifolds in hidden space. Our theoretical framework for analyzing the expressive power of deep networks is broadly applicable and provides a basis for quantifying previously abstract notions about the geometry of deep functions.

#96 On Multiplicative Integration with Recurrent Neural Networks

Yuhuai Wu (Univ. of Toronto)
Saizheng Zhang (Univ. of Montreal)
Ying Zhang (Univ. of Montreal)
Yoshua Bengio (U. Montreal)
Russ Salakhutdinov (Univ. of Toronto)

We introduce a general simple structural design called "Multiplicative Integration" (MI) to improve recurrent neural networks (RNNs). MI changes the way of how the information flow gets integrated in the computational building block of an RNN, while introducing almost no extra parameters. The new structure can be easily embedded into many popular RNN models, including LSTMs and GRUs. We empirically analyze its learning behaviour and conduct evaluations on several tasks using different RNN models. Our experimental results demonstrate that Multiplicative Integration can provide a substantial performance boost over many of the existing RNN models.

#97 Interpretable Nonlinear Dynamic Modeling of Neural Trajectories

Yuan Zhao (Stony Brook Univ.)
Memming Park

A central challenge in neuroscience is understanding how neural system implements computation through its dynamics. We propose a nonlinear time series model aimed at characterizing interpretable dynamics from neural trajectories. Our model incorporates prior assumption about globally contractional dynamics to avoid overly enthusiastic extrapolation outside of the support of observed trajectories. We show that our model can recover qualitative features of the phase portrait such as attractors, slow points, and bifurcation, while also producing reliable long-term future predictions in a variety of dynamical models and in real neural data.

#98 Globally Optimal Training of Generalized Polynomial Neural Networks with Nonlinear Spectral Methods

Antoine Gautier (Saarland Univ.)
Quynh N Nguyen (Saarland Univ.)
Matthias Hein (Saarland Univ.)

The optimization problem behind neural networks is highly non-convex. Training with stochastic gradient descent and variants requires careful parameter tuning and provides no guarantee to achieve the global optimum. In contrast we show under quite weak assumptions on the data that a particular class of feedforward neural networks can be trained globally optimal with a linear convergence rate. Up to our knowledge this is the first practically feasible method which achieves such a guarantee. While the method can in principle be applied to deep networks, we restrict ourselves for simplicity in this paper to one- and two hidden layer networks. Our experiments confirms that these models are already rich enough to achieve good performance on a series of real-world datasets.

#99 Linear Feature Encoding for Reinforcement Learning

Zhao Song (Duke Univ.)
Ron E Parr
Xuejun Liao (Duke Univ.)
Lawrence Carin

Feature construction is of vital importance in reinforcement learning, as the quality of a value function or policy is largely determined by the corresponding features. The recent successes of deep reinforcement learning (RL) only increase the importance of understanding feature construction. Typical deep RL approaches use a linear output layer, which means that deep RL can be interpreted as a feature construction/encoding network followed by linear value function approximation. This paper develops and evaluates a theory of linear feature encoding. We extend theoretical results on feature quality for linear value function approximation from the uncontrolled case to the controlled case. We then develop a supervised linear feature encoding method that is motivated by insights from linear value function approximation theory, as well as empirical successes from deep RL. The resulting encoder is a surprisingly effective method for linear value function approximation using raw images as inputs.

#100 Graphical Time Warping for Joint Alignment of Multiple Curves

Yizhi Wang (Virginia Tech)
David J Miller (The Pennsylvania State Univ.)
Kira Poskanzer (Univ. of California)
Yue Wang (Virginia Tech)
Lin Tian (The Univ. of California)
Guoqiang Yu

Dynamic time warping (DTW) is a fundamental technique in time series analysis to compare one curve to another with a flexible time-warping function. However, it was designed for modeling a single pair of curves. In many applications, such as in metabolomics and imaging series analysis, we need to do alignment simultaneously for multiple pairs. Because the underlying warping functions are often related, independent applications of DTW to each of the pairs offer only a sub-optimal solution. Yet, it was largely unknown how to efficiently conduct a joint alignment with all warping functions simultaneously considered, since any given warping function is constrained by the rest and the dynamic programming strategy cannot be applied. In this paper, we report a discovery that the joint alignment problem can be transformed into a network flow problem and hence can be exactly and efficiently solved by max flow algorithm with a guarantee of global optimality. We name the proposed approach as graphical time warping (GTW), emphasizing the graphical nature of the solution and the fact that the dependency structure of warping functions can be flexibly encoded in a graph. Modifications of DTW like windowing and weighting are readily derivable from GTW. We further discussed the optimal tuning of hyperparameters in GTW. We illustrated the power of GTW using both synthetic data and a real case study of astrocyte calcium movie.

#101 Mixed Linear Regression with Multiple Components

Kai Zhong (UT AUSTIN)
Prateek Jain (Microsoft Research)
Inderjit S Dhillon

In this paper, we study the mixed linear regression (MLR) problem, where the goal is to recover multiple underlying linear models from their unlabeled linear measurements. We propose a non-convex objective function which we show is {\em locally strongly convex} in the neighborhood of the ground truth. We use a tensor method for initialization so that the initial models are in the local strong convexity region. After that we can employ general convex optimization algorithms to minimize the objective function. To the best of our knowledge, our approach provides first exact recovery guarantees for the MLR problem with K\geq 2 components. Moreover, our method has near-optimal computational complexity \Otilde(Nd) as well as near-optimal sample complexity \Otilde(d) for constant K. Furthermore, we show that our non-convex formulation can be extended to solving the {\em subspace clustering} problem as well. In particular, we show that when initialized within a small constant distance to the true subspaces, our method converges to the global optima (and recovers true subspaces) in time {\em linear} in the number of points. This represents a significant step towards solving the open problem of subspace clustering in linear

(in number of data points) time. Furthermore, our empirical results indicate that even with random initialization, our approach converges to the global optima in linear time and significantly improves upon the existing methods.

#102 Statistical Inference for Pairwise Graphical Models Using Score Matching

Ming Yu (The Univ. of Chicago)
Mladen Kolar
Varun Gupta (Univ. of Chicago)

Probabilistic graphical models have been widely used to model complex systems and aid scientific discoveries. As a result, there is a large body of literature focused on consistent model selection. However, scientists are often interested in understanding uncertainty associated with the estimated parameters, which current literature has not addressed thoroughly. In this paper, we propose a novel estimator for edge parameters for pairwise graphical models based on Hyv\"arinen scoring rule. Hyv\"arinen scoring rule is especially useful in cases where the normalizing constant cannot be obtained efficiently in a closed form. We prove that the estimator is √{n}-consistent and asymptotically Normal. This result allows us to construct confidence intervals for edge parameters, as well as, hypothesis tests. We establish our results under conditions that are typically assumed in the literature for consistent estimation. However, we do not require that the estimator consistently recovers the graph structure. In particular, we prove that the asymptotic distribution of the estimator is robust to model selection mistakes and uniformly valid for a large number of data-generating processes. We illustrate validity of our estimator through extensive simulation studies.

#103 Hardness of Online Sleeping Combinatorial Optimization Problems

Satyen Kale
Chansoo Lee
David Pal

We show that several online combinatorial optimization problems that admit efficient no-regret algorithms become computationally hard in the sleeping setting where a subset of actions becomes unavailable in each round. Specifically, we show that the sleeping versions of these problems are at least as hard as PAC learning DNF expressions, a long standing open problem. We show hardness for the sleeping versions of Online Shortest Paths, Online Minimum Spanning Tree, Online k-Subsets, Online k-Truncated Permutations, Online Minimum Cut, and Online Bipartite Matching. The hardness result for the sleeping version of the Online Shortest Paths problem resolves an open problem presented at COLT 2015 [Koolen et al., 2015].

#104 An algorithm for L1 nearest neighbor search via monotonic embedding

Xinan Wang (UCSD)
Sanjoy Dasgupta

Fast algorithms for nearest neighbor (NN) search have in large part focused on L2 distance. Here we develop an approach for L1 distance that begins with an explicit and exact embedding of the points into L2. We show how this embedding can efficiently be combined with random projection methods for L2 NN search, such as locality-sensitive hashing or random projection trees. We rigorously establish the correctness of the methodology and show by experimentation that it is competitive in practice with available alternatives.

#105 On Local Maxima in the Population Likelihood of Gaussian Mixture Models: Structural Results and Algorithmic Consequences

Chi Jin (UC Berkeley)
Yuchen Zhang
Sivaraman Balakrishnan (CMU)
Martin J Wainwright (UC Berkeley)
Michael I Jordan

The algorithmic task of estimating the parameters of a mixture of Gaussians, given access to samples, is a fundamental and challenging problem which has drawn attention from a variety of researchers, notably in statistics and theoretical computer science. From an empirical perspective, the Expectation-Maximization (EM) algorithm is widely used, and is often observed to recover reasonable parameter estimates. However, the EM algorithm is only guaranteed to converge to a local stationary point of the likelihood function, and provides no guarantees of converging to a set of parameters that are statistically meaningful. In this paper, we study the favorable situation in which we receive (infinitely many) samples drawn from a mixture of k Gaussians (k is known to the user, and the mixture model is correctly specified), the components of which are additionally assumed to be uniformly weighted, spherical with unit-variance, and well-separated. Our first main result shows that even in this setting the population (infinite-sample) likelihood function has local maxima that are arbitrarily worse (in terms of their log-likelihood value) than any global optimum, answering the open question of Srebro [2007]. Our second main result shows that the EM algorithm (and gradient EM algorithm) with certain random initialization schemes will converge to a bad critical points with at least 1-e^{-Ω(k)} probability. We further show gradient EM algorithm will not converge to strict saddle points almost surely, suggesting our second main result is most likely due to the presence of bad local maxima. Our results highlight the necessity of careful initialization when using the EM algorithm in practice.

#106 Learning User Perceived Clusters with Feature-Level Supervision

Ting-Yu Cheng
Guiguan Lin
xinyang gong (NTHU)
Kang-Jun Liu
Shan-Hung Wu (National Tsing Hua Univ.)

Semi-supervised clustering algorithms have been proposed to identify data clusters that align with user perceived ones via the aid of side information such as seeds or pairwise constrains. However, traditional side information is mostly at the instance level and subject to the sampling bias, where non-randomly sampled instances in the supervision can mislead the algorithms to wrong clusters. In this paper, we propose learning from the feature-level supervision. We show that this kind of supervision can be easily obtained in the form of perception vectors in many applications. Then we present novel algorithms, called Perception Embedded (PE) clustering, that exploit the perception vectors as well as traditional side information to find clusters perceived by the user. Extensive experiments are conducted on real datasets and the results demonstrate the effectiveness of PE empirically.

#107 InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets

Xi Chen (UC Berkeley and OpenAI)
Xi Chen (UC Berkeley and OpenAI)
Yan Duan (UC Berkeley)
Rein Houthooft (Ghent Univ., iMinds, UC Berkeley, OpenAI)
John Schulman (OpenAI)
Ilya Sutskever
Pieter Abbeel (OpenAI / UC Berkeley / Gradescope)

This paper describes InfoGAN, an information-theoretic extension to the Generative Adversarial Network that is able to learn disentangled representations in a completely unsupervised manner. InfoGAN is a generative adversarial network that also maximizes the mutual information between a small subset of the latent variables and the observation. We derive a lower bound to the mutual information objective that can be optimized efficiently, and show that our training procedure can be interpreted as a variation of the Wake-Sleep algorithm. Specifically, InfoGAN successfully disentangles writing styles from digit shapes on the MNIST dataset, pose from lighting of 3D rendered images, and background digits from the central digit on the SVHN dataset. It also discovers visual concepts that include hair styles, presence/absence of eyeglasses, and emotions on the CelebA face dataset. Experiments show that InfoGAN learns interpretable representations that are competitive with representations learned by existing fully supervised methods.

#108 Neural universal discrete denoiser

Taesup Moon (DGIST)
Seonwoo Min
Byunghan Lee
Sung R. Yoon

We present a new framework of applying deep neural networks (DNN) to devise a universal discrete denoiser. Unlike other approaches that utilize supervised learning for denoising, we do not require any additional training data. In such setting, while the ground-truth label, i.e., the clean data, is not available, we devise ``pseudo-labels'' and a novel objective function such that DNN can be trained in a same way as supervised learning to become a discrete denoiser. We experimentally show that our resulting algorithm, dubbed as Neural DUDE, significantly outperforms the previous state-of-the-art in several applications with a systematic rule of choosing the hyperparameter, which is an attractive feature in practice.

#109 A primal-dual method for constrained consensus optimization

Necdet Serhat Aybat (Penn State Univ.)
Erfan Yazdandoost Hamedani (Penn State Univ.)

We consider cooperative multi-agent consensus optimization problems over an undirected network of agents, where only those agents connected by an edge can directly communicate. The objective is to minimize the sum of agent-specific composite convex functions over agent-specific private conic constraint sets; hence, the optimal consensus decision should lie in the intersection of these private sets. We provide convergence rates both in sub-optimality, infeasibility and consensus violation; examine the effect of underlying network topology on the convergence rates of the proposed decentralized algorithms; and discuss how to extend these methods to time-varying topology.

#110 Simple and Efficient Weighted Minwise Hashing

Anshumali Shrivastava (Rice Univ.)

Weighted minwise hashing (WMH) is one of the fundamental subroutine, required by many celebrated approximation algorithms, commonly adopted in industrial practice for large -scale search and learning. The resource bottleneck with WMH is the computation of multiple (typically a few hundreds to thousands) independent hashes of the data. We propose a simple rejection type sampling scheme based on a carefully designed red-green map, where we show that the number of rejected sample has exactly the same distribution as weighted minwise sampling. The running time of our method, for many practical datasets, is an order of magnitude smaller than existing methods. Experimental evaluations, on real datasets, show that for computing 500 WMH, our proposal can be 60000x faster than the Ioffe's method without losing any accuracy. Our method is also around 100x faster than approximate heuristics capitalizing on the efficient ``densified'' one permutation hashing schemes~\cite{Proc:OneHashLSH_ICML14}. Given the simplicity of our approach and its significant advantages, we hope that it will replace existing implementations in practice.

#111 Eliciting Categorical Data for Optimal Aggregation

Chien-Ju Ho (Cornell Univ.)
Rafael Frongillo
Yiling Chen

Models for collecting and aggregating categorical data on crowdsourcing platforms typically fall into two broad categories: those assuming agents honest and consistent but with heterogeneous error rates, and those assuming agents strategic and seek to maximize their expected reward. The former often leads to tractable aggregation of elicited data, while the latter usually focuses on optimal elicitation and does not consider aggregation. In this paper, we develop a Bayesian model, wherein agents have differing quality of information, but also respond to incentives. Our model generalizes both categories and enables the joint exploration of optimal elicitation and aggregation. This model enables our exploration, both analytically and experimentally, of optimal aggregation of categorical data and optimal multiple-choice interface design.

#112 Depth from a Single Image by Harmonizing Overcomplete Local Network Predictions

Ayan Chakrabarti
Jingyu Shao (UCLA)
Greg Shakhnarovich

A single color image can contain many cues informative towards different aspects of local geometric structure. We approach the problem of monocular depth estimation by using a neural network to produce a mid-level representation that summarizes these cues. This network is trained to characterize local scene geometry by predicting, at every image location, depth derivatives of different orders, orientations and scales. However, instead of a single estimate for each derivative, the network outputs probability distributions that allow it to express confidence about some coefficients, and ambiguity about others. Scene depth is then estimated by harmonizing this overcomplete set of network predictions, using a globalization procedure that finds a single consistent depth map that best matches all the local derivative distributions. We demonstrate the efficacy of this approach through evaluation on the NYU v2 depth data set.

#113 SEBOOST - Boosting Stochastic Learning Using Subspace Optimization Techniques

Elad Richardson (Technion)
Rom Herskovitz
Boris Ginsburg
Michael Zibulevsky

We present SEBOOST, a technique for boosting the performance of existing stochastic optimization methods. SEBOOST applies a secondary optimization process in the subspace spanned by the last steps and descent directions. The method was inspired by the SESOP optimization method for large-scale problems, and has been adapted for the stochastic learning framework. It can be applied on top of any existing optimization method with no need to tweak the internal algorithm. We show that the method is able to boost the performance of different algorithms, and make them more robust to changes in their hyper-parameters. As the boosting steps of SEBOOST are applied between large sets of descent steps, the additional subspace optimization hardly increases the overall computational burden. We introduce two hyper-parameters that control the balance between the baseline method and the secondary optimization process. The method was evaluated on several deep learning tasks, demonstrating promising results.

#114 Reshaped Wirtinger Flow for Solving Quadratic Systems of Equations

Huishuai Zhang (Syracuse Univ.)
Yingbin Liang (Syracuse Univ.)

We study the problem of recovering a vector X\in R^n from its magnitude measurements y_i=|\langle {a}_i, X>|, i=1,..., m. Our work is along the line of the Wirtinger flow (WF) approach, which solves the problem by minimizing a nonconvex loss function via a gradient algorithm and can be shown to converge to a global optimal point under good initialization. In contrast to the smooth loss function used in WF, we adopt a nonsmooth but lower-order loss function, and design a gradient-like algorithm (referred to as reshaped-WF). We show that for random Gaussian measurements, reshaped-WF enjoys geometric convergence to a global optimal point as long as the number m of measurements is at the order of \mathcal{O}(n), where n is the dimension of the unknown X. This improves the sample complexity of WF, and achieves the same sample complexity as truncated-WF but without truncation at gradient step. Furthermore, reshaped-WF costs less computationally than WF, and runs faster numerically than both WF and truncated-WF. Bypassing higher-order variables in the loss function and truncations in the gradient loop, analysis of reshaped-WF is substantially simplified.

#115 Training and Evaluating Multimodal Word Embeddings with Large-scale Web Annotated Images

Junhua Mao (UCLA)
Jiajing Xu (Pinterest)
Kevin Jing
Alan L Yuille

In this paper, we focus on training and evaluating effective word embeddings with both text and visual information. More specifically, we introduce a large-scale dataset with 300 million sentences describing over 40 million images crawled and downloaded from publicly available Pins (i.e. an image with sentence descriptions uploaded by users) on Pinterest. This dataset is more than 200 times larger than MS COCO, the standard large-scale image dataset with sentence descriptions. In addition, we construct an evaluation dataset to directly assess the effectiveness of word embeddings in terms of finding semantically similar or related words and phrases. The word/phrase pairs in this evaluation dataset are collected from the click data with millions of users in a recommendation system, thus contain rich semantic relationships. Based on these datasets, we propose and compare several Recurrent Neural Networks (RNNs) based multimodal (text and image) models. Experiments show that our model benefits from incorporating the visual information into the word embeddings, and a weight sharing strategy is crutial for learning such multimodal embeddings. The datasets introduced in this paper will be released upon acceptance.

#116 Online ICA: Understanding Global Dynamics of Nonconvex Optimization via Diffusion Processes

Chris Junchi Li (Princeton Univ.)
Zhaoran Wang (Princeton Univ.)
Han Liu

Solving statistical learning problems often involves nonconvex optimization. Despite the empirical success of nonconvex statistical optimization methods, their global dynamics, especially convergence to the desirable local minima, remain less well understood in theory. In this paper, we propose a new analytic paradigm based on diffusion processes to characterize the global dynamics of nonconvex statistical optimization. As a concrete example, we study stochastic gradient descent (SGD) for the tensor decomposition formulation of independent component analysis. In particular, we cast different phases of SGD into diffusion processes, i.e., solutions to stochastic differential equations. Initialized from an unstable equilibrium, the global dynamics of SGD transit over three consecutive phases: (i) an unstable Ornstein-Uhlenbeck process slowly departing from the initialization, (ii) the solution to an ordinary differential equation, which quickly evolves towards the desirable local minimum, and (iii) a stable Ornstein-Uhlenbeck process oscillating around the desirable local minimum. Based on these three phases, we discuss the global rate of convergence of SGD based upon stopping time analysis. Our proof techniques are based upon Stroock and Varadhan's weak convergence of Markov chains to diffusion processes, which are of independent interest.

#117 Variational Information Maximizing Exploration

Rein Houthooft (Ghent Univ., iMinds, UC Berkeley, OpenAI)
Xi Chen (UC Berkeley and OpenAI)
Xi Chen (UC Berkeley and OpenAI)
Yan Duan (UC Berkeley)
John Schulman (OpenAI)
Filip De Turck (Ghent Univ. - iMinds)
Pieter Abbeel (OpenAI / UC Berkeley / Gradescope)

Scalable and effective exploration remains a key challenge in reinforcement learning (RL). While there are methods with optimality guarantees in the setting of discrete state and action spaces, these methods cannot be applied in high-dimensional deep RL scenarios. As such, most contemporary RL relies on simple heuristics such as epsilon-greedy exploration or adding Gaussian noise to the controls. This paper introduces Variational Information Maximizing Exploration (VIME), an exploration strategy based on maximization of information gain about the agent's belief of environment dynamics. We propose a practical implementation, using variational inference in Bayesian neural networks which efficiently handles continuous state and action spaces. VIME modifies the MDP reward function, and can be applied with several different underlying RL algorithms. We demonstrate that VIME achieves significantly better performance compared to heuristic exploration methods across a variety of continuous control tasks and algorithms, including tasks with very sparse rewards.

#118 Deconvolving Feedback Loops in Recommender Systems
Ayan Sinha (Purdue)
David Gleich
Karthik Ramani (Purdue Univ.)

Collaborative filtering is a popular technique to infer users' preferences on new content based on the collective information of all users preferences. Recommender systems then use this information to make personalized suggestions to users. When users accept these recommendations it creates a feedback loop in the recommender system, and these loops iteratively influence the collaborative filtering algorithm's predictions over time. We investigate whether it is possible to identify items affected by these feedback loops. We state sufficient assumptions to deconvolve the feedback loops while keeping the inverse solution tractable. We furthermore develop a metric to unravel the recommender system's influence on the entire user-item rating matrix. We use this metric on synthetic and real-world datasets to (1) identify the extent to which the recommender system affects the final rating matrix, (2) rank frequently recommended items, and (3) distinguish whether a user's rated item was recommended or an intrinsic preference. Our results indicate that it is possible to recover the ratings matrix of intrinsic user preferences using a single snapshot of the ratings matrix without any temporal information.

#119 A Non-parametric Learning Method for Confidently Estimating Patient's Clinical State and Dynamics
William Hoiles (Univ. of California)
Mihaela Van Der Schaar

Estimating patient's clinical state from multiple concurrent physiological streams plays an important role in determining if a therapeutic intervention is necessary and for triaging patients in the hospital. In this paper we construct a non-parametric learning algorithm to estimate the clinical state of a patient. The algorithm addresses several known challenges with clinical state estimation such as eliminating bias introduced by therapeutic intervention censoring, increasing the timeliness of state estimation while ensuring a sufficient accuracy, and the ability to detect anomalous clinical states. These benefits are obtained by combining the tools of non-parametric Bayesian inference, permutation testing, and generalizations of the empirical Bernstein inequality. The algorithm is validated using real-world data from a cancer ward in a large academic hospital.

#120 Semiparametric Differential Graph Models
Pan Xu (Univ. of Virginia)
Quanquan Gu (Univ. of Virginia)

In many cases of network analysis, it is more attractive to study how a network varies under different conditions than an individual static network. We propose a novel graphical model, namely Latent Differential Graph Model, where the networks under two different conditions are represented by two semiparametric elliptical distributions respectively, and the variation of these two networks (i.e., differential graph) is characterized by the difference between their latent precision matrices. We propose an estimator for the differential graph based on quasi likelihood maximization with nonconvex regularization. We show that our estimator attains a faster statistical rate in parameter estimation than the state-of-the-art methods, and enjoys oracle property under mild conditions. Thorough experiments on both synthetic and real world data support our theory.

#121 A Non-convex One-Pass Framework for Generalized Factorization Machines and Rank-One Matrix Sensing
Ming Lin
Jieping Ye

We develop an efficient alternating framework for learning Factorization Machine (FM) on steaming data with provable guarantees. When the feature is d-dimension and the target second order coefficient matrix in FM is of rank k, our algorithm converges linearly, achieves O(ϵ) recovery error after retrieving O(k^{3}dlog(1/ϵ)) training instances, consumes O(kd) memory in one-pass of dataset and only requires matrix-vector product operations in each iteration. The key ingredient of our framework is a construction of an estimation sequence endowed with a so-called Conditionally Independent RIP condition. As special cases of FM, our framework can be applied to symmetric or asymmetric rank-one matrix sensing problems, such as inductive matrix completion and phrase retrieval.

#122 Sublinear Time Orthogonal Tensor Decomposition
Zhao Song (UT-Austin)
David Woodruff
Huan Zhang (UC-Davis)

A recent work (Wang et. al., NIPS 2015) gives the fastest known algorithms for orthogonal tensor decomposition with provable guarantees. Their algorithm is based on the technique of linear sketching, which requires reading the input tensor to create a sketch. We show that one can achieve the same theoretical guarantees as in their work in sublinear time, i.e., without even reading most of the input tensor! Our method is based on importance sampling from the current iterate vector in known tensor decomposition methods, but does not suffer the shortcomings of previous sampling-based schemes, such as uniform sampling which misses spiky elements or previous biased sampling which requires reading the tensor to create a sampling distribution. Our algorithm is orders of magnitude faster than existing methods, which we show both in theory and demonstrate with empirical results.

#123 Achieving budget-optimality with adaptive schemes in crowdsourcing
Ashish Khetan (Univ. of Illinois Urbana-)
Sewoong Oh

Adaptive schemes, where tasks are assigned based on the data collected thus far, are widely used in practical crowdsourcing systems to efficiently allocate the budget. However, existing theoretical analyses of crowdsourcing systems suggest that the gain of adaptive task assignments is minimal. To bridge this gap, we investigate this question under a strictly more general probabilistic model, which has been recently introduced to model practical crowdsourcing datasets. Under this generalized Dawid-Skene model, we characterize the fundamental trade-off between budget and accuracy, and introduce a novel adaptive scheme that matches this fundamental limit. We further quantify the gain of adaptivity, by comparing the trade-off with the one for non-adaptive schemes, and confirm that the gain is significant and can be made arbitrarily large depending on the distribution of the difficulty level of the tasks at hand.

#124 Joint Line Segmentation and Transcription for End-to-End Handwritten Paragraph Recognition
Theodore Bluche (A2iA)

Offline handwriting recognition systems require cropped text line images for both training and recognition. On the one hand, the annotation of position and transcript at line level is costly to obtain. On the other hand, automatic line segmentation algorithms are prone to errors, compromising the subsequent recognition. In this paper, we propose a modification of the popular and efficient Multi-Dimensional Long Short-Term Memory Recurrent Neural Networks (MDLSTM-RNNs) to enable end-to-end processing of handwritten paragraphs. More particularly, we replace the collapse layer transforming the two-dimensional representation into a sequence of predictions by a recurrent version which can select one line at a time. In the proposed model, a neural network performs a kind of implicit line segmentation by computing attention weights on the image representation. The experiments on paragraphs of Rimes and IAM databases yield results that are competitive with those of networks trained at line level, and constitute a significant step towards end-to-end transcription of full documents.

#125 Human Decision-Making under Limited Time
Pedro A Ortega
Alan A Stocker

Abstract Subjective expected utility theory assumes that decision-makers possess unlimited computational resources to reason about their choices; however, virtually all decisions in everyday life are made under resource constraints---i.e. decision-makers are bounded in their rationality. Here we experimentally tested the predictions made by a formalization of bounded rationality based on ideas from statistical mechanics and information-theory. We systematically tested human subjects in their ability to solve combinatorial puzzles under different time limitations. We found that our bounded-rational model accounts well for the data. The decomposition of the fitted model parameter into the subjects' expected utility function and resource parameter provide interesting insight into the subjects' information capacity limits. Our results confirm that humans gradually fall back on their learned prior choice patterns when confronted with increasing resource limitations.

#126 Joint M-Best-Diverse Labelings as a Parametric Submodular Minimization
Alexander Kirillov (TU Dresden)
Sasha Shekhovtsov
Carsten Rother
Bogdan Savchynskyy

We consider the problem of jointly inferring the~M-best diverse labelings for a binary (high-order) submodular energy of a graphical model. Recently it was shown that this problem can be solved to a global optimum, for many practically interesting diversity measures. It was noted that the labelings are, so-called, nested. This nestedness property also holds for labelings of a class of parametric submodular minimization problems, where different values of the global parameter Γ give rise to different solutions. The popular example of the parametric submodular minimization is the monotonic parametric max-flow problem, which is also widely used for computing multiple labelings. As the main contribution of this work we establish a close relationship between diversity with submodular energies and the parametric submodular minimization. In particular, the joint M-best diverse labelings can be obtained by running a {\em non-parametric} submodular minimization (in the special case - max-flow) solver for~M different values of~Γ {\em in parallel}, for certain diversity measures. Importantly, the values for~Γ can be computed in a closed form in advance, prior to any optimization. These theoretical results suggest two simple yet efficient algorithms for the joint M-best diverse problem, which outperform competitors in terms of runtime and quality of results. In particular, as we show in the paper, the new methods compute the exact M-best diverse labelings faster than a popular method of Batra et al., which in some sense only obtains approximate solutions.

#127 Even Faster SVD Decomposition Yet Without Agonizing Pain
Zeyuan Allen-Zhu (Princeton Univ.)
Yuanzhi Li (Princeton Univ.)

We study k-SVD that is to obtain the first k singular vectors of a matrix A approximately. Recently, a few breakthroughs have been discovered on k-SVD: Musco and Musco [1] provided the first gap-free theorem for the block Krylov method, Shamir [2] discovered the first variance-reduction stochastic method, and Bhojanapalli et al. [3] provided the fastest O(nnz(A) + poly(1/eps))-type of algorithm using alternating minimization. In this paper, we improve the above breakthroughs by providing a new framework for solving k-SVD. In particular, we obtain faster gap-free convergence speed outperforming [1], we obtain the first accelerated AND stochastic method outperforming [3]. In the NNZ running-time regime, we outperform [3] without even using alternating minimization for certain parameter regimes.

#128 Fast and accurate spike sorting of high-channel count probes with KiloSort
Marius Pachitariu
Nicholas A Steinmetz (UCL)
Shabnam N Kadir
Matteo Carandini (UCL)
Daniel D Harris (UCL)

New silicon technology is enabling large-scale electrophysiological recordings in vivo from hundreds to thousands of channels. Interpreting these recordings requires scalable and accurate automated methods for spike sorting, which should minimize the time required for manual curation of the results. Here we introduce KiloSort, a new integrated spike sorting framework that uses template matching both during spike detection and during spike clustering. KiloSort models the electrical voltage as a sum of template waveforms triggered on the spike times, which allows overlapping spikes to be identified and resolved. Unlike previous algorithms that compress the data with PCA, KiloSort operates on the raw data which allows it to construct a more accurate model of the waveforms. Processing times are faster than in previous algorithms thanks to batch-based optimization on GPUs. We compare KiloSort to an established algorithm and show favorable performance, at much reduced processing times. A novel post-clustering merging step based on the continuity of the templates further reduced substantially the number of manual operations required on this data, for the neurons with near-zero error rates, paving the way for fully automated spike sorting of multichannel electrode recordings.

#129 BBO-DPPs: Batched Bayesian Optimization via Determinantal Point Processes

Tarun Kathuria (Microsoft Research)
Amit Deshpande
Pushmeet Kohli

Bayesian optimization has emerged as a powerful tool for optimizing noisy black box functions. One good example in machine learning is hyper-parameter optimization where each evaluation of the target function may require training a model which may involve days or even weeks of computation. Most methods for Bayesian Optimization only allow sequential exploration of the parameter space. However, it is often desirable to propose batches or sets of parameter values to explore simultaneously, especially when there are large parallel processing facilities at our disposal. Batch methods require modeling the interaction between the different evaluations in the batch, which can be expensive in complex scenarios. In this paper, we propose a new approach for parallelizing Bayesian optimization by modeling the diversity of a batch via Determinantal point processes (DPPs) whose kernels are learned automatically. This allows us to generalize a previous result as well as prove better regret bounds based on DPP sampling. Our experiments on a variety of synthetic and real-world robotics and hyper-parameter optimization tasks indicate that our DPP-based methods, especially those based on DPP sampling, outperform state-of-the-art methods.

#130 Stochastic Multiple Choice Learning for Training Diverse Deep Ensembles

Stefan Lee (Indiana Univ.)
Senthil Purushwalkam Shiva Prakash (Carnegie Mellon)
Michael Cogswell (Virginia Tech)
Viresh Ranjan (Virginia Tech)
David Crandall (Indiana Univ.)
Dhruv Batra

Many practical perception systems exist within larger processes which often include interactions with users or additional components that are capable of evaluating the quality of predicted solutions. In these contexts, it is beneficial to provide these oracle mechanisms with multiple highly likely hypotheses rather than a single prediction. In this work, we pose the task of producing multiple outputs as a learning problem over an ensemble of deep networks -- introducing a novel stochastic gradient descent based approach to minimize the loss with respect to an oracle. Our method is simple to implement, agnostic to both architecture and loss function, and parameter-free. Our approach achieves lower oracle error compared to existing methods on a wide range of tasks and deep architectures. We also show qualitatively that solutions produced from our approach often provide interpretable representations of task ambiguity.

#131 Optimal Sparse Linear Encoders and Sparse PCA

Malik Magdon-Ismail (Rensselaer)
Christos Boutsidis

Principal components analysis~(PCA) is the optimal linear encoder of data. Sparse linear encoders (e.g., sparse PCA) produce more interpretable features that can promote better generalization. (\rn{1}) Given a level of sparsity, what is the best approximation to PCA? (\rn{2}) Are there efficient algorithms which can achieve this optimal combinatorial tradeoff? We answer both questions by providing the first polynomial-time algorithms to construct {optimal} sparse linear auto-encoders; additionally, we demonstrate the performance of our algorithms on real data.

#132 Using Social Dynamics to Make Individual Predictions: Variational Inference with Stochastic Kinetic Model

Zhen Xu (SUNY at Buffalo)
Wen Dong
Sargur N Srihari

The availability of large-scale data in social networks and sensor networks offers an unprecedented opportunity to predict state changing events at the individual level. Examples of such events are disease infection, rumor propagation and opinion transition in elections, etc. Unlike previous research focusing on the collective effects of social systems, we want to make efficient inferences on the individual level. In order to cope with dynamic interactions within a large number of individuals, we introduce the stochastic kinetic model to capture adaptive transition kernel. In addition, we propose an efficient variational inference algorithm whose complexity grows linearly with the number of individuals. We performed epidemic dynamics experiments on wireless sensor network data collected from more than ten thousand people over three years. The proposed algorithm was used to track disease transmission, and predict the probability of infection for each individual. It is more efficient than sampling while achieving high accuracy.

#133 Learning Additive Exponential Family Graphical Models via \(ell_{2,1}\)-norm Regularized M-Estimation

Xiaotong Yuan (Nanjing Univ. of Informat)
Ping Li
Tong Zhang
Qingshan Liu
Guangcan Liu (NUIST)

We investigate a subclass of semi-parametric exponential family graphical models of which the sufficient statistics are defined by arbitrary additive forms. We propose two \(ℓ_{2,1}\)-norm regularized maximum likelihood estimators to learn the model parameters from i.i.d. samples. The first one is a joint MLE estimator which estimates all the parameters simultaneously. The second one is a node-wise conditional MLE estimator which estimates the parameters for each node individually. For both estimators, statistical analysis shows that under mild conditions the extra flexibility gained by the additive exponential family models comes at almost no cost of statistical efficiency. A Monte-Carlo approximation method is developed to efficiently optimize the proposed estimators. The advantages of our estimators over Gaussian graphical models and Nonparanormal estimators are demonstrated on several synthetic and real data sets.

#134 Residual Networks are Exponential Ensembles of Relatively Shallow Networks

Andreas Veit (Cornell Univ.)
Michael J Wilber
Serge Belongie (Cornell Univ.)

In this work, we introduce a novel interpretation of residual networks showing they are exponential ensembles. This observation is supported by a large-scale lesion study that demonstrates they behave just like ensembles at test time. Subsequently, we perform an analysis showing these ensembles mostly consist of networks that are each relatively shallow. For example, contrary to our expectations, most of the gradient in a residual network with 110 layers comes from an ensemble of very short networks, i.e., only 10-34 layers deep. This suggests that in addition to describing neural networks in terms of width and depth, there is a third dimension: multiplicity, the size of the implicit ensemble. Ultimately, residual networks do

not resolve the vanishing gradient problem by preserving gradient flow throughout the entire depth of the network - rather, they avoid the problem simply by ensembling many short networks together. This insight reveals that depth is still an open research question and invites the exploration of the related notion of multiplicity.

#135 Full-Capacity Unitary Recurrent Neural Networks

Scott Wisdom (Univ. of Washington)
Thomas Powers
John Hershey
Jonathan Le Roux
Les Atlas

Recurrent neural networks are generally plagued by vanishing and exploding gradient problems. Unitary recurrent networks, which use unitary weight matrices, have recently been proposed as a means to avoid these issues, but in previous experiments the weight matrices were restricted to a parameterized family of unitary matrices. However, an open question remains: how much of the set of unitary matrices can be represented using the proposed family of unitary matrices, and how does this limit what can be learned? To address this question, we introduce a new concept, called ``Givens capacity,'' which is intended to indicate the representation power of a unitary matrix. We define the Givens capacity of a unitary matrix as the minimum number of Givens operators--two-dimensional unitary matrices embedded in an identity matrix--that can be multiplied together to construct that unitary matrix. In particular, we demonstrate that a recently proposed unitary parameterization cannot represent the entire unitary group for hidden state dimension greater than 22, and thus has restricted capacity to represent all possible input-output sequence mappings. In contrast, we show how a complete, full-capacity unitary recurrence matrix can be optimized over the differential Stiefel manifold of unitary matrices. We then confirm the utility of our claim by empirically evaluating these full-capacity unitary recurrent networks on both synthetic and natural data, achieving superior performance compared to both LSTMs and the original restricted-capacity unitary recurrent networks.

#136 Quantum Perceptron Models

Ashish Kapoor
Nathan Wiebe (Microsoft Research)
Krysta Svore

We demonstrate how quantum computation can provide non-trivial improvements in the computational and statistical complexity of the perceptron model. We develop two quantum algorithms for perceptron learning. The first algorithm exploits quantum information processing to determine a separating hyperplane using a number of steps sublinear in the number of data points N, namely O(√{N}). The second algorithm illustrates how the classical mistake bound of O(1/Γ2) can be further improved to O(1/{√{Γ}}) through quantum means, where Γ denotes the margin. Such improvements are achieved through the application of quantum amplitude amplification to the version space interpretation of the perceptron model.

#137 Mapping Estimation for Discrete Optimal Transport

Michaël Perrot (Univ. of Saint-Etienne)
Nicolas Courty
Rémi Flamary
Amaury Habrard (Univ. of Saint-Etienne)

We are interested in the computation of the transport map of an Optimal Transport problem. Most of the computational approaches of Optimal Transport use the Kantorovich relaxation of the problem to learn a probabilistic coupling γ but do not address the problem of learning the transport map T linked to the original Monge problem. Consequently, it lowers the potential usage of such methods in contexts where out-of-samples computations are mandatory. In this paper we propose a new way to jointly learn the coupling and an approximation of the transport map. We use a jointly convex formulation which can be efficiently optimized. Additionally, jointly learning the coupling and the transport map allows to smooth the result of the Optimal Transport and generalize it on out-of-samples examples. Empirically, we show the interest and the relevance of our method in two tasks: domain adaptation and image editing.

#138 Stochastic Gradient Geodesic MCMC Methods

Chang Liu (Tsinghua Univ.)
Jun Zhu
Yang Song (Stanford Univ.)

We propose two stochastic gradient MCMC methods for sampling from Bayesian posterior distributions defined on Riemann manifolds with a known geodesic flow, e.g. hyperspheres. Our methods are the first scalable sampling methods on these manifolds, with the aid of stochastic gradients. Novel dynamics are conceived and second-order integrators are developed. By adopting embedding techniques and the geodesic integrator, the methods do not require a global coordinate system of the manifold and do not involve inner iterations. Synthetic experiments show the validity of the method, and its application to the challenging inference for spherical topic models indicate practical usability and efficiency.

#139 Variational Information Maximization for Feature Selection

Shuyang Gao
Greg Ver Steeg
Aram Galstyan

Feature selection is one of the most fundamental problems in machine learning. An extensive body of work on information-theoretic feature selection exists which is based on maximizing mutual information between subset of features and class labels. Practical methods are forced to rely on approximations due to the difficulty of estimating mutual information. We demonstrate that approximations made by existing methods are based on unrealistic assumptions. We formulate a more flexible and general class of assumptions based on variational distributions and use them to tractably generate lower bounds for mutual information. These bounds define a novel information-theoretic framework for feature selection, which we prove to be optimal under tree graphical models with proper choice of variational distributions. Our experiments demonstrate that the proposed method strongly outperforms existing information-theoretic feature selection approaches.

#140 A Minimax Approach to Supervised Learning

Farzan Farnia (Stanford Univ.)
David Tse (Stanford Univ.)

Given a task of predicting Y from X, a loss function L, and a set of probability distributions Gamma, what is the optimal decision rule minimizing the worst-case expected loss over Gamma? In this paper, we address this question by introducing a generalization of the principle of maximum entropy. Applying this principle to sets of distributions with a proposed structure, we develop a general minimax approach for supervised learning problems, that reduces to the maximum likelihood problem over generalized linear models. Through this framework, we develop a classification algorithm called the minimax SVM. This algorithm, which is a relaxed version of the standard SVM, minimizes the worst-case 0-1 loss over the structured set of distribution, and by our numerical experiments can outperform the SVM.

#141 Fast Distributed Submodular Cover: Public-Private Data Summarization

Baharan Mirzasoleiman (ETH Zurich)
Morteza Zadimoghaddam (Google Research)
Amin Karbasi

In this paper, we introduce the public-private framework of data summarization motivated by privacy concerns in personalized recommender systems and online social services. Such systems have usually access to massive data generated by a large pool of users. A major fraction of the data is public and is visible to (and can be used for) all users. However, each user can also contribute some private data that should not be shared with other users to ensure her privacy. The goal is to provide a succinct summary of massive dataset, ideally as small as possible, that is customized to each user: a summary can contain elements from the public data (for diversity) and user's private data (for personalization). To formalize the above challenge, we assume that the scoring function according to which a user evaluates the utility of her summary satisfies submodularity, a widely used notion in data summarization applications. Thus, we model the data summarization targeted to each user as an instance of a submodular cover problem. However, when the data is massive it is infeasible to use the centralized (and sequential in nature) greedy algorithm to find a customized summary even for a single user. Moreover, for a large pool of users, it is too time consuming to find such summaries separately. Instead, we develop a fast distributed algorithm for submodular cover, FastCover, that provides a succinct summary in one shot and for all users. We show that the solution provided by FastCover is competitive with that of the centralized algorithm with the number of rounds that is exponentially smaller than state of the art results. Moreover, we have implemented FastCover on Spark to demonstrate its practical performance on a number of concrete applications, including personalized location recommendation, personalized movie recommendation, and vertex cover on tens of millions of data points and varying number of users.

#142 Domain Separation Networks

Konstantinos Bousmalis (Google Brain)
George Trigeorgis (Google)
Nathan Silberman (Google)
Dilip Krishnan (Google)
Dumitru Erhan (Google)

The cost of large scale data collection and annotation often makes the application of machine learning algorithms to new tasks or datasets prohibitively expensive. One approach circumventing this cost is training models on synthetic data where annotations are provided automatically. Despite their appeal, such models often fail to generalize from synthetic to real images, necessitating domain adaptation algorithms to manipulate these models before they can be successfully applied. Existing approaches focus either on mapping representations from one domain to the other, or on learning to extract features that are invariant to the domain from which they were extracted. However, by focusing only on creating a mapping or shared representation between the two domains, they ignore the individual characteristics of each domain. We suggest that explicitly modeling what is unique to each domain can improve a model's ability to extract domain-invariant features. Inspired by work on private-shared component analysis, we explicitly learn to extract image representations that are partitioned into two subspaces: one component which is private to each domain and one which is shared across domains. Our model is trained not only to perform the task we care about in the source domain, but also to use the partitioned representation to reconstruct the images from both domains. Our novel architecture results in a model that outperforms the state-of-the-art on a range of unsupervised domain adaptation scenarios and additionally produces visualizations of the private and shared representations enabling interpretation of the domain adaptation process.

#143 Multimodal Residual Learning for Visual QA

Jin-Hwa Kim (Seoul National Univ.)
Sang-Woo Lee (Seoul National Univ.)
Donghyun Kwak (Seoul National Univ.)
Min-Oh Heo (Seoul National Univ.)
Jeonghee Kim (Naver Labs)
Jung-Woo Ha (Naver Labs)
Byoung-Tak Zhang (Seoul National Univ.)

Deep neural networks continue to advance the state-of-the-art of image recognition tasks with various methods. However, applications of these methods to multimodality remain limited. We present Multimodal Residual Networks (MRN) for the multimodal residual learning of visual question-answering, which extends the idea of the deep residual learning. Unlike the deep residual learning, MRN effectively learns the joint representation from visual and language information. The main idea is to use element-wise multiplication for the joint residual mappings exploiting the residual learning of the attentional models in recent studies. Various alternative models introduced by multimodality are explored based on our study. We achieve the state-of-the-art results on the Visual QA dataset for both Open-Ended and Multiple-Choice tasks. Moreover, we introduce a novel method to visualize the attention effect of the joint representations for each learning block using back-propagation algorithm, even though the visual features are collapsed without spatial information.

#144 Optimizing affinity-based binary hashing using auxiliary coordinates

Ramin Raziperchikolaei (UC Merced)
Miguel A. Carreira-Perpinan (UC Merced)

In supervised binary hashing, one wants to learn a function that maps a high-dimensional feature vector to a vector of binary codes, for application to fast image retrieval. This typically results in a difficult optimization problem, nonconvex and nonsmooth, because of the discrete variables involved. Much work has simply relaxed the problem during training, solving a continuous optimization, and truncating the codes a posteriori. This gives reasonable results but is quite suboptimal. Recent work has tried to optimize the objective directly over the binary codes and achieved better results, but the hash function was still learned a posteriori, which remains suboptimal. We propose a general framework for learning hash functions using affinity-based loss functions that uses auxiliary coordinates. This closes the loop and optimizes jointly over the hash functions and the binary codes so that they gradually match each other. The resulting algorithm can be seen as a corrected, iterated version of the procedure of optimizing first over the codes and then learning the hash function. Compared to this, our optimization is guaranteed to obtain better hash functions while being not much slower, as demonstrated experimentally in various supervised datasets. In addition, our framework facilitates the design of optimization algorithms for arbitrary types of loss and hash functions.

#145 Coresets for Scalable Bayesian Logistic Regression

Jonathan Huggins (MIT)
Trevor Campbell (MIT)
Tamara Broderick (MIT)

The use of Bayesian models in large-scale data settings is attractive because of the rich hierarchical models, uncertainty quantification, and prior specification they provide. Standard Bayesian inference algorithms are computationally expensive, however, making their direct application to large datasets difficult or infeasible. Recent work on scaling Bayesian inference has focused on modifying the underlying algorithms to, for example, use only a random data subsample at each iteration. We leverage the insight that data is often redundant to instead obtain a weighted subset of the data (called a coreset) that is much smaller than the original dataset. We can then use this small coreset in any number of existing posterior inference algorithms without modification. In this paper, we develop an efficient coreset construction algorithm for Bayesian logistic regression models. We provide theoretical guarantees on the size and approximation quality of the coreset -- both for fixed, known datasets, and in expectation for a wide class of data generative models. The proposed approach also permits efficient construction of the coreset in both streaming and parallel settings, with minimal additional effort. We demonstrate the efficacy of our approach on a number of synthetic and real-world datasets, and find that, in practice, the size of the coreset is independent of the original dataset size.

#146 The Parallel Knowledge Gradient Method for Batch Bayesian Optimization

Jian Wu (Cornell Univ.)
Peter Frazier

In many applications of black-box optimization, one can evaluate multiple points simultaneously, e.g. when evaluating the performances of several different neural network architectures in a parallel computing environment. In this paper, we develop a novel batch Bayesian optimization algorithm --- the parallel knowledge gradient method. By construction, this method provides the one-step Bayes optimal batch of points to sample. We provide an efficient strategy for computing this Bayes-optimal batch of points, and we demonstrate that the parallel knowledge gradient method finds global optima significantly faster than previous batch Bayesian optimization algorithms on both synthetic test functions and when tuning hyperparameters of practical machine learning algorithms, especially when function evaluations are noisy.

#147 Learning Multiagent Communication with Backpropagation

Sainaa Sukhbaatar (NYU)
arthur szlam
Rob Fergus (New York Univ.)

Many tasks in AI require the collaboration of multiple agents. Typically, the communication protocol between agents is manually specified and not altered during training. In this paper we explore a simple neural model, called CommNN, that uses continuous communication for fully cooperative tasks. The model consists of multiple agents and the communication between them is learned alongside their policy. We apply this model to a diverse set of tasks, demonstrating the ability of the agents to learn to communicate amongst themselves, yielding improved performance over non-communicative agents and baselines. In some cases, it is possible to interpret the language devised by the agents, revealing simple but effective strategies for solving the task at hand.

#148 Optimal Binary Classifier Aggregation for General Losses

Akshay Balsubramani (UC San Diego)
Yoav S Freund

We address the problem of aggregating an ensemble of binary classifiers in a semi-supervised setting. Recently, this problem was solved optimally using a game-theoretic approach, but that analysis was specific to the 0-1 loss. In this paper, we generalize the minimax optimal algorithm of the previous work to a very general, novel class of loss functions, including but not limited to all convex surrogates, while extending its performance and efficiency guarantees. The result is a family of parameter-free ensemble aggregation algorithms which use labeled and unlabeled data; these are as efficient as linear learning and prediction for convex risk minimization, but work without any relaxations on many non-convex loss functions. The prediction algorithms take a form familiar in decision theory, applying sigmoid functions to a generalized notion of ensemble margin, but without the assumptions typically made in margin-based learning.

#149 The Generalized Reparameterization Gradient

Francisco R Ruiz (Columbia Univ.)
Michalis Titsias RC AUEB
David Blei

The reparameterization gradient has become a widely used method for obtaining Monte Carlo gradients to optimize the variational objective. However, this technique only applies when fitting approximate Gaussian distributions. In this paper, we introduce the generalized reparameterization gradient, a method that extends the reparameterization gradient to a wider class of variational distributions. Generalized reparameterizations use invertible transformations of the latent variables which lead to transformed distributions that weakly depend on the variational parameters. This results in new Monte Carlo gradients that combine reparameterization gradients and score function gradients. We demonstrate our approach on variational inference for two complex probabilistic models, a deep exponential family and a nonconjugate factorization model. The generalized reparameterization is effective: even a single sample from the variational distribution is enough to obtain a low-variance gradient.

#150 Conditional Generative Moment-Matching Networks

Yong Ren (Tsinghua Univ.)
Jun Zhu
Jialian Li (Tsinghua Univ.)
Yucen Luo

Maximum mean discrepancy (MMD) has been successfully applied to learn deep generative models for characterizing a joint distribution of variables via kernel mean embedding. In this paper, we present conditional generative moment-matching networks (CGMMN), which learn a conditional distribution given some input variables based on a conditional maximum mean discrepancy (CMMD) criterion. The learning is performed by stochastic gradient descent with the gradient calculated by back-propagation. We evaluate CGMMN on a wide range of tasks, including predictive modeling, contextual generation, and Bayesian dark knowledge, which distills knowledge from a Bayesian model by learning a relatively small CGMMN student network. Our results demonstrate competitive performance in all the tasks.

#151 A Credit Assignment Compiler for Joint Prediction

Kai-Wei Chang
He He (Univ. of Maryland)
Stephane Ross (Google)
Hal Daume III
John Langford

Many machine learning applications involve jointly predicting multiple mutually1 dependent output variables. Learning to search is a family of methods where the complex decision problem is cast into a sequence of decisions via a search space. Although these methods have shown promise both in theory and in practice, implementing them has been burdensomely awkward In this paper, we show the search space can be defined by an arbitrary imperative program, turning learning to search into a credit assignment compiler. Altogether with the algorithmic improvements for the compiler, we radically reduce the complexity of programming and the running time. We demonstrate the feasibility of our approach on multiple joint prediction tasks. In all cases, we obtain accuracies as high as alternative approaches, at drastically reduced execution and programming time.

#152 Short-Dot: Computing Large Linear Transforms Distributedly Using Coded Short Dot Products

Sanghamitra Dutta (Carnegie Mellon Univ.)
Viveck Cadambe (Pennsylvania State Univ.)
Pulkit Grover (Carnegie Mellon Univ.)

Faced with saturation of Moore's law and increasing size and dimensionality of data, system designers have increasingly resorted to parallel distributed computing to reduce computation time of machine-learning algorithms. However, distributed computing is often bottlenecked by a small fraction of slow processors called ``stragglers'' that reduce the speed of computation because the sink node has to wait for all processors to complete their processing. To combat the effect of stragglers, recent literature proposes introducing redundant computations, e.g.~using repetition-based strategies or erasure codes. Introducing redundancy allows the sink node to complete the computation using outputs from a only subset of the processors. In this paper, we propose a novel coding technique, that we call ``Short-Dot,'' for introducing redundant computations in computing dense linear transforms of long dense vectors. Instead of computing dense dot products as required in the original computation, we construct a larger number of redundant and sparse dot products (hence shorter) that can be computed more efficiently at individual processors. Further, only a subset of these sparse dot products are required at the sink node to finish the computation of the linear transform successfully. We demonstrate through probabilistic analysis as well as experiments on computing clusters that the proposed strategy offers significant speed-up compared to existing techniques. We also derive fundamental limits on the sparsity that can be achieved by any such strategy and compare it to that achieved by our strategy.

#153 Spatio-Temporal Hilbert Maps for Continuous Occupancy Representation in Dynamic Environments

Ransalu Senanayake (The Univ. of Sydney)
Lionel Ott (The Univ. of Sydney)
Simon O'Callaghan (NICTA)
Fabio Ramos (The Univ. of Sydney)

We consider the problem of building continuous occupancy representations in dynamic environments for robotics applications. The problem has been hardly discussed previously due to the complexity of patterns in urban environments, which have both spatial and temporal dependencies. We address the problem as learning a kernel classifier on an efficient feature space. The key novelty of our approach is the incorporation of variations in the time domain into the spatial domain. We propose two methods, 1) propagating motion uncertainty into the kernel using a hierarchical model 2) capturing long-term occupancy using doubly-stochastic functional gradients. The main benefit of the first approach is that it can directly predict the occupancy state of the map in the future from past observations, being a valuable tool for robot trajectory planning under uncertainty. Both approaches preserve the main computational benefits of static Hilbert maps – using stochastic gradient descent for fast optimization of model parameters and incremental updates as new data are obtained. Experiments conducted in road intersections of an urban environment demonstrated that spatio-temporal Hilbert maps can accurately model changes in the map while outperforming other techniques on various aspects.

#154 Learning HMMs with Nonparametric Emissions via Spectral Decompositions of Continuous Matrices

Kirthevasan Kandasamy (CMU)
Maruan Al-Shedivat (CMU)
Eric P Xing (Carnegie Mellon Univ.)

Recently, there has been a surge of interest in using spectral methods for estimating latent variable models. However, it is usually assumed that the distribution of the observations conditioned on the latent variables is either discrete or belongs to a parametric family. In this paper, we study the estimation of an m-state hidden Markov model (HMM) with only smoothness assumptions, such as H\"olderian conditions, on the emission probabilities. By leveraging some recent advances in continuous linear algebra and numerical analysis, we develop a computationally efficient spectral algorithm for learning nonparametric HMMs. Our technique is based on computing an SVD on nonparametric estimates of density functions by viewing them as {continuous matrices}. We derive sample complexity bounds via concentration results for nonparametric density estimation and novel perturbation theory results for these continuous matrices. We implement our method using Chebyshev polynomial approximations. Our method is competitive with other baselines on synthetic and real problems and is also very computationally efficient.

#155 Integrator Nets

Hakan Bilen (Univ. of Oxford)
Andrea Vedaldi

Modern discriminative predictors have been shown to match natural intelligences in specific perceptual tasks in image classification, object and part detection, boundary extraction, etc. However, a major advantage that natural intelligences still have is that they work well for all perceptual problems together, solving them efficiently and coherently in an integrated manner. In order to capture some of these advantages in machine perception, we ask two questions: whether deep neural networks can learn universal image representations, useful not only for a single task but for all of them, and how the solutions to the different tasks can be integrated in this framework. We answer by proposing a new architecture, which we call multinet, in which not only deep image features are shared between tasks, but where tasks can interact in a recurrent manner by encoding the results of their analysis in a common shared representation of the data. In this manner, we show that the performance of individual tasks in standard benchmarks can be improved first by sharing features between them and then, more significantly, by integrating their solutions in the common representation.

#156 Blind Attacks on Machine Learners

Alex Beatson (Princeton Univ.)
Zhaoran Wang (Princeton Univ.)
Han Liu

We study statistical estimation when the training set is drawn from a mixture of the distribution of interest and a malicious distribution chosen by a "blind attacker" who does not observe the distribution of interest or the learner's training set. We analyze minimax rates of convergence in two variants of this setting: firstly, where an "informed learner" knows the malicious distribution, and secondly, where a "blind learner" knows only the proportion of malicious data and the family to which the malicious distribution belongs. These analyses exhibit limits on a learner's ability to learn and the potential for an attacker to harm learning when a data injection is performed by a blind attacker. We provide examples of simple yet effective attacks in both the informed learner and blind learner settings and derive resulting lower bounds on the learner's minimax risk for some simple statistical models.

#157 Optimistic Gittins Indices

Eli Gutin (Massachusetts Institute of Tec)
Vivek Farias

Starting with the Thomspon sampling algorithm, recent years have seen a resurgence of interest in Bayesian algorithms for the Multi-armed Bandit (MAB) problem. These algorithms seek to exploit prior information on arm biases and while several have been shown to be regret optimal, their design has not emerged from a principled approach. In contrast, if one cared about Bayesian regret discounted over an infinite horizon at a fixed, pre-specified rate, the celebrated Gittins index theorem offers an optimal algorithm. Unfortunately, the Gittins analysis does not appear to carry over to minimizing Bayesian regret over all sufficiently large horizons and computing a Gittins index is onerous relative to essentially any incumbent index scheme for the Bayesian MAB problem. The present paper proposes a sequence of 'optimistic' approximations to the Gittins index. We show that the use of these approximations in concert with the use of an increasing discount factor appears to offer a compelling alternative to a variety of index schemes proposed for the Bayesian MAB problem in recent years. In addition, we show that the simplest of these approximations yields regret that matches the Lai-Robbins lower bound, including achieving matching constants.

#158 Sub-sampled Newton Methods with Non-uniform Sampling

Peng Xu (Stanford Univ.)
Jiyan Yang (Stanford Univ.)
Farbod Roosta-Khorasani (Univ. of California Berkeley)
Chris Ré
Michael W Mahoney

We consider the problem of finding the minimizer of a convex function F: \mathbb R^d \rightarrow \mathbb R of the form F(w) \ defeq \sum_{i=1}^n f_i(w) + R(w) where a low-rank factorization of \ nabla^2 f_i(w) is readily available.We consider the regime where n \ gg d. We propose randomized Newton-type algorithms that exploit \textit{non-uniform} sub-sampling of \{\nabla^2 f_i(w)\}_{i=1}^{n}, as well as inexact updates, as means to reduce the computational complexity, and are applicable to a wide range of problems in machine learning. Two non-uniform sampling distributions based on {\it block norm squares} and {\it block partial leverage scores} are considered. Under certain assumptions, we show that our algorithms inherit a linear-quadratic convergence rate in w and achieve a lower computational complexity compared to similar existing methods. In addition, we show that our algorithms exhibit more robustness and better dependence on problem specific quantities, such as the condition number. We numerically demonstrate the advantages of our algorithms on several real datasets.

#159 Learned Region Sparsity and Diversity Also Predicts Visual Attention

Zijun Wei (Stony Brook)
Hossein Adeli
Minh Hoai
Greg Zelinsky
Dimitris Samaras

Learned region sparsity has achieved state-of-the-art performance in classification tasks by exploiting and integrating a sparse set of local information into global decisions. The underlying mechanisms resemble how people sample information from the image with their eye-movements when making similar decisions. In this paper we enhance the learned region sparsity model with the biologically plausible mechanism of Inhibition of Return, to impose diversity on the selected regions. We investigated whether these mechanisms of sparsity and diversity correspond to visual attention by testing our model on three different types of visual search tasks. We report state-of-the-art results in predicting the location of human visual attention, even though we only trained on image-level labels without object location annotation. Notably the enhanced model's classification performance remains the same as the original. This work sheds some light on the possible visual attention mechanisms in the brain and argues for inclusion of attention-based mechanisms for improving computer vision techniques.

#160 Adaptive Concentration Inequalities for Sequential Decision Problems

Shengjia Zhao (Tsinghua Univ.)
Enze Zhou (Tsinghua Univ.)
Ashish Sabharwal (Allen Institute for AI)
Stefano Ermon

A key challenge in sequential decision problems is to determine how many samples are needed for an agent to make reliable decisions with good probabilistic guarantees. We introduce Hoeffding-like concentration inequalities that hold for a random, adaptively chosen number of samples. Our inequalities are tight under natural assumptions and can greatly simplify the analysis of common sequential decision problems. In particular, we apply them to sequential hypothesis testing, best arm identification, and sorting. The resulting algorithms rival or exceed the state of the art both theoretically and empirically.

#161 Cooperative Graphical Models

Josip Djolonga (ETH Zurich)
Stefanie Jegelka (MIT)
Sebastian Tschiatschek (ETH Zurich)
Andreas Krause

We study a rich family of distributions that capture variable interactions significantly more expressive than those representable with low-treewidth or pairwise graphical models, or log-supermodular models. We call these cooperative graphical models. Yet, this family retains structure: by carefully exploiting this structure, we develop efficient inference techniques that combine the polyhedral structure of submodular functions in new ways with variational inference methods to obtain both lower and upper bounds on the partition function. While our fully convex upper bound is minimized as an SDP or via tree-reweighted belief propagation, our lower bound is tightened via belief propagation or mean-field algorithms. The resulting algorithms are easy to implement and, as our experiments show, effectively obtain good bounds and marginals for synthetic and real-world examples.

#162 Correlated-PCA: Principal Components' Analysis when Data and Noise are Correlated
Namrata Vaswani
Han Guo (Iowa State Univ.)

Given a matrix of observed data, Principal Components Analysis (PCA) computes a small number of orthogonal directions that contain most of the variability of the data. The PCA problem and provably accurate solutions for it have been in use for decades. However, to the best of our knowledge, all existing theoretical guarantees for it assume that the data and the noise are mutually independent, or at least uncorrelated. This is valid in practice often, but not always. In this paper, we study the PCA problem in the setting where the data and noise vectors can be correlated. We obtain a correctness result for the standard eigenvalue decomposition (EVD) based solution to the PCA problem under the correlated model. One limitation of this result is that its sample complexity grows as f^2 where f is the condition number of the true data covariance matrix. To address this limitation, we develop and analyze a simple generalization of EVD, that we call cluster-EVD. Under a clustering assumption on the data covariance eigenvalues, we argue that cluster-EVD has a significantly smaller sample complexity.

#163 Hierarchical Object Representation for Open-Ended Object Category Learning and Recognition
Hamidreza Kasaei (IEETA)

Most robots lack the ability to learn new objects from past experiences. To migrate a robot to a new environment one must often completely redesign and remodel the knowledge-base that it is running with. Since in open-ended domains the set of categories to be learned is not predefined, it is not feasible to assume that one can pre-program all object categories required by robots. Therefore, autonomous robots must have the ability to continuously execute learning and recognition in a concurrent and interleaved fashion. This paper proposes an open-ended 3D object recognition system which concurrently learns both the object categories and the statistical features for encoding objects. In particular, we propose an extension of Latent Dirichlet Allocation to learn structural semantic features (i.e. topics) from low-level feature co-occurrences for each category independently. Moreover, topics in each category are discovered in an unsupervised fashion and are updated incrementally using new object views. The approach contains similarities with the organization of visual cortex and builds a hierarchy of increasingly sophisticated representations. Results show the fulfilling performance of this approach on different types of objects. Moreover, this system demonstrates the capability of learning from few training examples and competes with state-of-the-art systems.

#164 Optimal Tagging with Markov Chain Optimization
Nir Rosenfeld (Hebrew Univ. of Jerusalem)
Amir Globerson (Tel Aviv Univ.)

Many information systems use tags and keywords to describe and annotate content. These allow for efficient organization and categorization of items, as well as facilitate relevant search queries. As such, the selected set of tags for an item can have a considerable effect on the volume of traffic that eventually reaches an item. In settings where tags are chosen by an item's creator, who in turn is interested in maximizing traffic, a principled approach for choosing tags can prove valuable. In this paper we introduce the problem of optimal tagging, where the task is to choose a subset of tags for a new item such that the probability of a browsing user reaching that item is maximized. We formulate the problem by modeling traffic using a Markov chain, and asking how transitions in this chain should be modified to maximize traffic into a certain state of interest. The resulting optimization problem involves maximizing a certain function over subsets, under a cardinality constraint. We show that the optimization problem is NP-hard, but has a (1-1/e)-approximation via a simple greedy algorithm due to monotonicity and submodularity. Furthermore, the structure of the problem allows for an efficient computation of the greedy step. To demonstrate the effectiveness of our method, we perform experiments on three tagging datasets, and show that the greedy algorithm outperforms other baselines.

#165 Bayesian optimization for automated model selection
Gustavo Malkomes (Washington Univ.)
Charles Schaff (Washington Univ. in St. Louis)
Roman Garnett

Despite the success of kernel-based nonparametric methods, kernel selection still requires considerable expertise, and is often described as a "black art." We present a sophisticated method for automatically searching for an appropriate kernel from an infinite space of potential choices. Previous efforts in this direction have focused on traversing a kernel grammar, only examining the data via computation of marginal likelihood. Our proposed search method is based on Bayesian optimization in model space, where we reason about model evidence as a function to be maximized. We explicitly reason about the data distribution and how it induces similarity between potential model choices in terms of the explanations they can offer for observed data. In this light, we construct a novel kernel between models to explain a given dataset. Our method is capable of finding a model that explains a given dataset well without any human assistance, often with fewer computations of model evidence than previous approaches, a claim we demonstrate empirically.

#166 Multi-view Anomaly Detection via Robust Probabilistic Latent Variable Models
Tomoharu Iwata
Makoto Yamada

We propose probabilistic latent variable models for multi-view anomaly detection, which is the task of finding instances that have inconsistent views given multi-view data. With the proposed model, all views of a non-anomalous instance are assumed to be generated from a single latent vector. On the other hand, an anomalous instance is assumed to have multiple latent vectors, and its different views are generated from different latent vectors. By inferring the number of latent vectors used for each instance with Dirichlet process priors, we obtain multi-view anomaly scores. The proposed model can be seen as a robust extension of probabilistic canonical correlation analysis for noisy multi-view data. We present Bayesian inference procedures for the proposed model based on a stochastic EM algorithm. The effectiveness of the proposed model is demonstrated in terms of performance when detecting multi-view anomalies and imputing missing values in multi-view data with anomalies.

#167 Inference by Reparameterization in Neural Population Codes

Rajkumar Vasudeva Raju (Rice Univ.)
Xaq Pitkow

Behavioral experiments on humans and animals suggest that the brain performs probabilistic inference to interpret its environment. Here we present a new general-purpose, biologically-plausible neural implementation of approximate inference. The neural network represents uncertainty using Probabilistic Population Codes (PPCs), which are distributed neural representations that naturally encode probability distributions, and support marginalization and evidence integration in a biologically-plausible manner. By connecting multiple PPCs together as a probabilistic graphical model, we represent multivariate probability distributions. Approximate inference in graphical models can be accomplished by message-passing algorithms that disseminate local information throughout the graph. An attractive and often accurate example of such an algorithm is Loopy Belief Propagation (LBP), which uses local marginalization and evidence integration operations to perform approximate inference efficiently even for complex models. Unfortunately, a subtle feature of LBP renders it neurally implausible. However, LBP can be elegantly reformulated as a sequence of Tree-based Reparameterizations (TRP) of the graphical model. We re-express the TRP updates as a nonlinear dynamical system with both fast and slow timescales, and show that this produces a neurally plausible solution. By combining all of these ideas, we show that a network of PPCs can represent multivariate probability distributions and implement the TRP updates to perform probabilistic inference. Simulations with Gaussian graphical models demonstrate that the neural network inference quality is comparable to the direct evaluation of LBP and robust to noise, and thus provides a promising mechanism for general probabilistic inference in the population codes of the brain.

#168 Efficient Neural Codes under Metabolic Constraints

Zhuo Wang (Univ. of Pennsylvania)
Xue-Xin Wei (Univ. of Pennsylvania)
Alan A Stocker
Daniel D Lee (Univ. of Pennsylvania)

Neural codes are inevitably shaped by various kinds of biological constraints, {e.g.} noise and metabolic cost. Here we formulate a coding framework which explicitly deals with noise and the metabolic costs associated with the neural representation of information, and analytically derive the optimal neural code for monotonic response functions and arbitrary stimulus distributions. Our framework can be applied to a neuronal pool of various sizes. For a single neuron, the theory predicts a family of optimal response functions depending on the metabolic budget and noise characteristics. Interestingly, the well-known histogram equalization solution can be viewed as a special case when metabolic resources are unlimited. For a pair of neurons, our theory suggests that under more substantial metabolic constraints, ON-OFF coding is an increasingly more efficient coding scheme compared to ON-ON or OFF-OFF. For a larger neural population, the theory predicts that the optimal code should divide the neurons into an ON-pool and an OFF-pool; neurons in the same pool have similar yet non-identical response functions. These analytical results may provide a theoretical basis for the predominant segregation into ON- and OFF-cells in early visual processing areas. Overall, the theory provides a unified framework for optimal neural codes with monotonic tuning curves in the brain, and makes predictions that can be directly tested with physiological experiments.

#169 Learning Deep Parsimonious Representations

Renjie Liao (UofT)
Alex Schwing
Richard Zemel
Raquel Urtasun

In this paper we aim at facilitating generalization for deep networks while supporting interpretability of the learned representations. Towards this goal, we propose a clustering based regularization that encourages parsimonious representations. Our k-means style objective is easy to optimize and flexible supporting various forms of clustering, including sample and spatial clustering as well as co-clustering. We demonstrate the effectiveness of our approach on the tasks of unsupervised learning, classification, fine grained categorization and zero-shot learning.

#170 An equivalence between high dimensional Bayes optimal inference and M-estimation

Madhu Advani (Stanford Univ.)
Surya Ganguli (Stanford)

Due to the computational difficulty of performing MMSE (minimum mean squared error) inference, maximum a posteriori (MAP) is often used as a surrogate. However, the accuracy of MAP is suboptimal for high dimensional inference, where the number of model parameters is of the same order as the number of samples. In this work we demonstrate how MMSE performance is asymptotically achievable via optimization with an appropriately selected convex penalty and regularization function which are a smoothed version of the widely applied MAP algorithm. Our findings provide a new derivation and interpretation for recent optimal M-estimators discovered by El Karoui, et. al. PNAS 2013 as well as extending to non-additive noise models. We demonstrate the performance of these optimal M-estimators with numerical simulations. Overall, at the heart of our work is the revelation of a remarkable equivalence between two seemingly very different computational problems: namely that of high dimensional Bayesian integration, and high dimensional convex optimization. In essence we show that the former computationally difficult integral may be computed by solving the latter, simpler optimization problem.

#171 Minimizing Quadratic Functions in Constant Time

Kohei Hayashi (AIST)
Yuichi Yoshida (NII)

A sampling-based optimization method for quadratic functions is proposed. Our method approximately solves the following n-dimensional quadratic minimization problem in constant time, which is independent of n: z^*=\min_{\bv \in \bbR^n}\bracket{\bv}{A \bv} + n\bracket{\bv}{\diag(\bd)\bv} + n\bracket{\bb}{\bv}, where A \in \bbR^{n \times n} is a matrix and \bd,\bb \in \bbR^n are vectors. Our theoretical analysis specifies the number of samples k(δ, ϵ) such that the approximated solution z satisfies |z - z^*| = O(ϵ n^2) with probability 1-δ. The empirical performance (accuracy and runtime) is positively confirmed by numerical experiments.

#172 Learning Structured Sparsity in Deep Neural Networks

Wei Wen (Univ. of Pittsburgh)
Chunpeng Wu (Univ. of Pittsburgh)
Yandan Wang (Univ. of Pittsburgh)
Yiran Chen (Univ. of Pittsburgh)
Hai Li (Univ. of Pittsburg)

High demand for computation resources severely hinders deployment of large-scale Deep Neural Networks (DNN) in resource constrained devices. In this work, we propose a Structured Sparsity Learning (SSL) method to regularize the structures (i.e., filters, channels, filter shapes, and layer depth) of DNNs. SSL can: (1) learn a compact structure from a bigger DNN to reduce computation cost; (2) obtain a hardware-friendly structured sparsity of DNN to efficiently accelerate the DNN's evaluation. Experimental results show that SSL achieves on average 5.1X and 3.1X speedups of convolutional layer computation of AlexNet against CPU and GPU, respectively, with off-the-shelf libraries. These speedups are about twice speedups of non-structured sparsity; (3) regularize the DNN structure to improve classification accuracy. The results show that for CIFAR-10, regularization on layer depth reduces a 20-layer Deep Residual Network (ResNet) to 18 layers while improves the accuracy from 91.25% to 92.60%, which is still higher than that of original ResNet with 32 layers. For AlexNet, SSL reduces the error by ~1%.

#173 Adversarial Multiclass Classification: A Risk Minimization Perspective

Rizal Fathony (U. of Illinois at Chicago)
Anqi Liu
Kaiser Asif
Brian Ziebart

Recently proposed adversarial classification methods have shown promising results for cost sensitive and multivariate losses. In contrast with empirical risk minimization (ERM) methods, which use convex surrogate losses to approximate the desired non-convex target loss function, adversarial methods minimize non-convex losses by treating the properties of the training data as being uncertain and worst case within a minimax game. Despite this difference in formulation, we recast adversarial classification under zero-one loss as an ERM method with a novel prescribed loss function. We demonstrate a number of theoretical and practical advantages over the very closely related hinge loss ERM methods. This establishes adversarial classification under the zero-one loss as a method that fills the long standing gap in multiclass hinge loss classification, simultaneously guaranteeing Fisher consistency and universal consistency, while also providing dual parameter sparsity and high accuracy predictions in practice.

#174 Unified Methods for Exploiting Piecewise Structure in Convex Optimization

Tyler B Johnson (Univ. of Washington)
Carlos Guestrin

We study the task of rapidly identifying important components of a convex optimization problem for the purpose of achieving fast convergence times. By considering a novel problem formulation—the minimization of a sum of piecewise functions—we describe a principled and general mechanism for exploiting piecewise structure in convex optimization. This result directly leads to a theoretically justified working set algorithm and a novel screening test, which generalize and improve upon many prior results on exploiting structure in optimization.

#175 Fast and Provably Good Seedings for k-Means

Olivier Bachem (ETH Zurich)
Mario Lucic (ETH Zurich)
Hamed Hassani (ETH Zurich)
Andreas Krause

Seeding - the task of finding initial cluster centers - is critical in obtaining high-quality clusterings for k-Means. However, k-means++ seeding, the state of the art algorithm, does not scale well to massive datasets as it is inherently sequential and requires k full passes through the data. It was recently shown that Markov chain Monte Carlo sampling can be used to efficiently approximate the seeding step of k-means++. However, this result requires assumptions on the data generating distribution. We propose a simple yet fast seeding algorithm that produces *provably* good clusterings even *without assumptions* on the data. Our analysis shows that the algorithm allows for a favourable trade-off between solution quality and computational cost, speeding up k-means++ seeding by up to several orders of magnitude. We validate our theoretical results in extensive experiments on a variety of real-world data sets.

#176 Testing for Differences in Gaussian Graphical Models: Applications to Brain Connectivity

Eugene Belilovsky (CentraleSupelec)
Gaël Varoquaux
Matthew B Blaschko (KU Leuven)

Functional brain networks are well described and estimated from data with Gaussian Graphical Models (GGMs), e.g.\ using sparse inverse covariance estimators. Comparing functional connectivity of subjects in two populations calls for comparing these estimated GGMs. Our goal is to identify differences in GGMs known to have similar structure. We characterize the uncertainty of differences with confidence intervals obtained using a parametric distribution on parameters of a sparse estimator. Sparse penalties enable statistical guarantees and interpretable models even in high-dimensional and low-sample settings. Characterizing the distributions of sparse models is inherently challenging as the penalties produce a biased estimator. Recent work invokes the sparsity assumptions to effectively remove the bias from a sparse estimator such as the lasso. These distributions can be used to give confidence intervals on edges in GGMs, and by extension their differences. However, in the case of comparing GGMs, these estimators do not make use of any assumed joint structure among the GGMs. Inspired by priors from brain functional connectivity we derive the distribution of parameter differences under a joint penalty when parameters are known to be sparse in the difference. This leads us to introduce the debiased multi-task fused lasso, whose distribution can be characterized in an efficient manner. We then show how the debiased lasso and multi-task fused lasso can be used to obtain confidence intervals on edge differences in GGMs. We validate the techniques proposed on a set of synthetic examples as well as neuro-imaging dataset created for the study of autism.

#177 Synthesis of MCMC and Belief Propagation

Sung-Soo Ahn (KAIST)
Michael Chertkov (Los Alamos National Laboratory)
Jinwoo Shin (KAIST)

Markov Chain Monte Carlo (MCMC) and Belief Propagation (BP) are the most popular algorithms for computational inference in Graphical Models (GM). In principle, MCMC is an exact probabilistic method which, however, often suffers from exponentially slow mixing. In contrast, BP is a deterministic method, which is typically fast, empirically very successful, however in general lacking control of accuracy over loopy graphs. In this paper, we introduce MCMC algorithms correcting the approximation error of BP, i.e., we provide a way to compensate for BP errors via a consecutive BP-aware MCMC. Our framework is based on the Loop Calculus (LC) approach which allows to express the BP error as a sum of weighted generalized loops. Although the full series is computationally intractable, it is known that a truncated series, summing up all 2-regular loops, is computable in polynomial-time for planar pair-wise binary GMs and it also provides a highly accurate approximation empirically. Motivated by this, we, first, propose a polynomial-time approximation MCMC scheme for the truncated series of general (non-planar) pair-wise binary models. Our main idea here is to use the Worm algorithm, known to provide fast mixing in other (related) problems, and then design an appropriate rejection scheme to sample 2-regular loops. Furthermore, we also design an efficient rejection-free MCMC scheme for approximating the full series. The main novelty underlying our design is in utilizing the concept of cycle basis, which provides an efficient decomposition of the generalized loops. In essence, the proposed MCMC schemes run on transformed GM built upon the non-trivial BP solution, and our experiments show that this synthesis of BP and MCMC outperforms both direct MCMC and bare BP schemes.

#178 Value Iteration Networks

Aviv Tamar
Sergey Levine
Pieter Abbeel (OpenAI / UC Berkeley / Gradescope)
YI WU (UC Berkeley)
Garrett Thomas (UC Berkeley)

We introduce the value iteration network (VIN): a fully differentiable neural network with a `planning module' embedded within. VINs can learn to plan, and are suitable for predicting outcomes that involve planning-based reasoning, such as policies for reinforcement learning. Key to our approach is a novel differentiable approximation of the value-iteration algorithm, which can be represented as a convolutional neural network, and trained end-to-end using standard backpropagation. We evaluate VIN based policies on discrete and continuous path-planning domains, and on a natural-language based search task. We show that by learning an explicit planning computation, VIN policies generalize better to new, unseen domains.

#179 Sequential Neural Models with Stochastic Layers

Marco Fraccaro (DTU)
Søren Kaae Sønderby (KU)
Ulrich Paquet (DeepMind)
Ole Winther (DTU)

How can we efficiently propagate uncertainty in a latent state representation with recurrent neural networks? This paper introduces stochastic recurrent neural networks which glue a deterministic recurrent neural network and a state space model together to form a stochastic and sequential neural generative model. The clear separation of deterministic and stochastic layers allows a structured variational inference network to track the factorization of the model's posterior distribution. By retaining both the nonlinear recursive structure of a recurrent neural network and averaging over the uncertainty in a latent path, like a state space model, we improve the state of the art results on the Blizzard and TIMIT speech modeling data sets by a large margin, while achieving comparable performances to competing methods on polyphonic music modeling.

#180 Graphons, mergeons, and so on!

Justin Eldridge (The Ohio State Univ.)
Mikhail Belkin
Yusu Wang (The Ohio State Univ.)

In this work we develop a theory of hierarchical clustering for graphs. Our modelling assumption is that graphs are sampled from a graphon, which is a powerful and general model for generating graphs and analyzing large networks. Graphons are a far richer class of graph models than stochastic blockmodels, the primary setting for recent progress in the statistical theory of graph clustering. We define what it means for an algorithm to produce the ``correct" clustering, give sufficient conditions in which a method is statistically consistent, and provide an explicit algorithm satisfying these properties.

#181 Hierarchical Clustering via Spreading Metrics

Aurko Roy (Georgia Tech)
Sebastian Pokutta (GeorgiaTech)

We study the cost function for hierarchical clusterings introduced by [Dasgupta, 2015] where hierarchies are treated as first-class objects rather than deriving their cost from projections into flat clusters. It was also shown in [Dasgupta, 2015] that a top-down algorithm returns a hierarchical clustering of cost at most \(O\left(α_n log n\ right)\) times the cost of the optimal hierarchical clustering, where \(α_n\) is the approximation ratio of the Sparsest Cut subroutine used. Thus using the best known approximation algorithm for Sparsest Cut due to Arora-Rao-Vazirani, the top down algorithm returns a hierarchical clustering of cost at most \(O\left(log^{3/2} n\ right)\) times the cost of the optimal solution. We improve this by giving an \(O(log{n})\)-approximation algorithm for this problem. Our main technical ingredients are a combinatorial characterization of ultrametrics induced by this cost function, deriving an Integer Linear Programming (ILP) formulation for this family of ultrametrics, and showing how to iteratively round an LP relaxation of this formulation by using the idea of {sphere growing} which has been extensively used in the context of graph partitioning. We also prove that our algorithm returns an \(O(log{n})\)-approximate hierarchical clustering for a generalization of this cost function also studied in [Dasgupta, 2015]. Experiments show that the hierarchies found by using the ILP formulation as well as our rounding algorithm often have better projections into flat clusters than the standard linkage based algorithms. We conclude with an inapproximability result for this problem, namely that no polynomial sized LP or SDP can be used to obtain a constant factor approximation for this problem.

#182 Deep Learning for Predicting Human Strategic Behavior

Jason S Hartford (Univ. of British Columbia)
James R Wright (Univ. of British Columbia)
Kevin Leyton-Brown

Predicting the behavior of human participants in strategic settings is an important problem in many domains. Most existing work either assumes that participants are perfectly rational, or attempts to directly model each participant's cognitive processes based on insights from cognitive psychology and experimental economics. In this work, we present an alternative, a deep learning approach that automatically performs cognitive modeling without relying on such expert knowledge. We introduce a novel architecture that allows a single network to generalize across different input and output dimensions by using matrix units rather than scalar units, and show that its performance significantly outperforms that of the previous state of the art, which relies on expert-constructed features.

#183 Global Analysis of Expectation Maximization for Mixtures of Two Gaussians

Ji Xu (Columbia Univ.)
Daniel Hsu
(Columbia Univ.)

Expectation Maximization (EM) is among the most popular algorithms for estimating parameters of statistical models. However, EM, which is an iterative algorithm based on the maximum likelihood principle, is generally only guaranteed to find stationary points of the likelihood objective, and these points may be far from any maximizer. This article addresses this disconnect between the statistical principles behind EM and its algorithmic properties. Specifically, it provides a global analysis of EM for specific models in which the observations comprise an i.i.d. sample from a mixture of two Gaussians. This is achieved by (i) studying the sequence of parameters from idealized execution of EM in the infinite sample limit, and fully characterizing the limit points of the sequence in terms of the initial parameters; and then (ii) based on this convergence analysis, establishing statistical consistency (or lack thereof) for the actual sequence of parameters produced by EM.

#184 Supervised learning through the lens of compression

Ofir David (Technion - Israel institute of technology)
Shay Moran (Technion - Israel institue of Technology)
Amir Yehudayoff (Technion - Israel institue of Technology)

This work continues the study of the relationship between sample compression schemes and statistical learning, which has been mostly investigated within the framework of binary classification. We first extend the investigation to multiclass categorization: we prove that in this case learnability is equivalent to compression of logarithmic sample size and that the uniform convergence property implies compression of constant size. We use the compressibility-learnability equivalence to show that (i) for multiclass categorization, PAC and agnostic PAC learnability are equivalent, and (ii) to derive a compactness theorem for learnability. We then consider supervised learning under general loss functions: we show that in this case, in order to maintain the compressibility-learnability equivalence, it is necessary to consider an approximate variant of compression. We use it to show that PAC and agnostic PAC are not equivalent, even when the loss function has only three values.

#185 Matrix Completion has No Spurious Local Minimum

Rong Ge
Jason Lee (UC Berkeley)
Tengyu Ma (Princeton Univ.)

Matrix completion is a basic machine learning problem that has wide applications, especially in collaborative filtering and recommender systems. Simple non-convex optimization algorithms are popular and effective in practice. Despite recent progress in proving various non-convex algorithms converge from a good initial point, it remains unclear why random or arbitrary initialization suffices in practice. We prove that the commonly used non-convex objective function for matrix completion has no spurious local minima \--- all local minima must also be global. Therefore, many popular optimization algorithms such as (stochastic) gradient descent can provably solve matrix completion with \textit{arbitrary} initialization in polynomial time.

#186 Clustering with Same-Cluster Queries

Hassan Ashtiani (Univ. of Waterloo)
Shrinu Kushagra (Univ. of Waterloo)
Shai Ben-David (U. Waterloo)

We propose a framework for Semi-Supervised Active Clustering framework (SSAC), where the learner is allowed to interact with a domain expert, asking whether two given instances belong to the same cluster or not. We study the query and computational complexity of clustering in this framework. We consider a setting where the expert conforms to a center-based clustering with a notion of margin. We show that there is a trade off between computational complexity and query complexity; We prove that for the case of k-means clustering (i.e., when the expert conforms to a solution of k-means), having access to relatively few such queries allows efficient solutions to otherwise NP hard problems. In particular, we provide a probabilistic polynomial-time (BPP) algorithm for clustering in this setting that asks O\big(k^2log k + klog n) same-cluster queries and runs with time complexity O\big(knlog n) (where k is the number of clusters and n is the number of instances). The success of the algorithm is guaranteed for data satisfying the margin condition under which, without queries, we show that the problem is NP hard. We also prove a lower bound on the number of queries needed to have a computationally efficient clustering algorithm in this setting.

#187 MetaGrad: Multiple Learning Rates in Online Learning

Tim van Erven
Wouter M Koolen

In online convex optimization it is well known that certain subclasses of objective functions are much easier than arbitrary convex functions. We are interested in designing adaptive methods that can automatically get fast rates in as many such subclasses as possible, without any manual tuning. Previous adaptive methods are able to interpolate between strongly convex and general convex functions. We present a new method, MetaGrad, that adapts to a much broader class of functions, including exp-concave and strongly convex functions, but also various types of stochastic and non-stochastic functions without any curvature. For instance, MetaGrad can achieve logarithmic regret on the unregularized hinge loss, even though it has no curvature, if the data come from a favourable probability distribution. MetaGrad's main feature is that it simultaneously considers multiple learning rates. Unlike all previous methods with provable regret guarantees, however, its learning rates are not monotonically decreasing over time and are not tuned based on a theoretically derived bound on the regret. Instead, they are weighted directly proportional to their empirical performance on the data using a tilted exponential weights master algorithm.

#188 Unsupervised Feature Extraction by Time-Contrastive Learning and Nonlinear ICA
Aapo Hyvarinen
Hiroshi Morioka (Univ. of Helsinki)

Nonlinear independent component analysis (ICA) provides an appealing framework for unsupervised feature learning, but the models proposed so far are not identifiable. Here, we first propose a new intuitive principle of unsupervised deep learning from time series which uses the nonstationary structure of the data. Our learning principle, time-contrastive learning (TCL), finds a representation which allows optimal discrimination of time segments (windows). Surprisingly, we show how TCL can be related to a nonlinear ICA model, when ICA is redefined to include temporal nonstationarities. In particular, we show that TCL combined with linear ICA estimates the nonlinear ICA model up to point-wise transformations of the sources, and this solution is unique --- thus providing the first identifiability result for nonlinear ICA which is rigorous, constructive, as well as very general.

#189 Phased LSTM: Accelerating Recurrent Network Training for Long or Event-based Sequences
Daniel Neil (Institute of Neuroinformatics)
Michael Pfeiffer (Institute of Neuroinformatics)
Shih-Chii Liu

Recurrent Neural Networks (RNNs) have become the state-of-the-art choice for extracting patterns from temporal sequences. Current RNN models are ill suited to process irregularly sampled data triggered by events generated in continuous time by sensors or other neurons. Such data can occur, for example, when the input comes from novel event-driven artificial sensors which generate sparse, asynchronous streams of events or from multiple conventional sensors with different update intervals. In this work, we introduce the Phased LSTM model, which extends the LSTM unit by adding a new time gate. This gate is controlled by a parametrized oscillation with a frequency range which require updates of the memory cell only during a small percentage of the cycle. Even with the sparse updates imposed by the oscillation, the Phased LSTM network achieves faster convergence than regular LSTMs on tasks which require learning of long sequences. The model naturally integrates inputs from sensors of arbitrary sampling rates, thereby opening new areas of investigation for processing asynchronous sensory events that carry timing information. It also greatly improves the performance of LSTMs in standard RNN applications, and does so with an order-of-magnitude fewer computes.

#190 Tractable Operations for Arithmetic Circuits of Probabilistic Models
Yujia Shen
Arthur Choi
Adnan Darwiche

We consider tractable representations of probability distributions and the polytime operations they support. In particular, we consider a recently proposed arithmetic circuit representation, the Probabilistic Sentential Decision Diagram (PSDD). We show that PSDD supports a polytime multiplication operator, while they do not support a polytime operator for summing-out variables. A polytime multiplication operator make PSDDs suitable for a broader class of applications compared to arithmetic circuits, which do not in general support multiplication. As one example, we show that PSDD multiplication leads to a very simple but effective compilation algorithm for probabilistic graphical models: represent each model factor as a PSDD, and then multiply them.

#191 Using Fast Weights to Attend to the Recent Past
Jimmy Ba (Univ. of Toronto)
Geoffrey E Hinton (Google)
Volodymyr Mnih
Joel Z Leibo (Google DeepMind)
Catalin Ionescu (Google)

Until recently, research on artificial neural networks was largely restricted to systems with only two types of variable: Neural activities that represent the current or recent input and weights that learn to capture regularities among inputs, outputs and payoffs. There is no good reason for this restriction. Synapses have dynamics at many different time-scales and this suggests that artificial neural networks might benefit from variables that change slower than activities but much faster than the standard weights. These ``fast weights'' can be used to store temporary memories of the recent past and they provide a neurally plausible way of implementing the type of attention to the past that has recently proven helpful in sequence-to-sequence models. By using fast weights we can avoid the need to store copies of neural activity patterns.

#192 Bayesian Intermittent Demand Forecasting for Large Inventories
Matthias W Seeger (Amazon)
David Salinas (Amazon)
Valentin Flunkert (Amazon)

We present a scalable and robust Bayesian method for demand forecasting in the context of a large e-commerce platform, paying special attention to intermittent and bursty target statistics. Inference is approximated by the Newton-Raphson algorithm, reduced to linear-time Kalman smoothing, which allows us to operate on several orders of magnitude larger problems than previous related work. In a study on large real-world sales datasets, our method outperforms competing approaches on fast and medium moving items.

#193 Blazing the trails before beating the path: Sample-efficient Monte-Carlo planning
Jean-Bastien Grill (Inria Lille - Nord Europe)
Michal Valko (Inria Lille - Nord Europe)
Remi Munos (Google DeepMind)

We study the sampling-based planning problem in Markov decision processes (MDPs) that we can access only through a generative model, usually referred to as Monte-Carlo planning. Our objective is to return a good estimate of the optimal value function at any state while minimizing the number of calls to the generative model, i.e. the sample complexity. We propose a new algorithm, TrailBlazer, able to handle MDPs with a finite or an infinite number of transitions from state-action to next states. TrailBlazer is an adaptive algorithm that exploits possible structures of the MDP by exploring only a subset of states reachable by following near-optimal policies. We provide bounds on its sample complexity that depend on a measure of the quantity of near-optimal states. The algorithm behavior can be considered as an extension of Monte-Carlo sampling (for estimating an expectation) to problems that alternate maximization (over actions) and expectation (over next states). Finally, another appealing feature of TrailBlazer is that it is simple to implement and computationally efficient.

#194 SDP Relaxation with Randomized Rounding for Energy Disaggregation

Kiarash Shaloudegi
András György
Csaba Szepesvari (U. Alberta)
Wilsun Xu (Univ. of Alberta)

We develop a scalable, computationally efficient method for the task of energy disaggregation for home appliance monitoring. In this problem the goal is to estimate the energy consumption of each appliance based on the total energy-consumption signal of a household. The current state of the art models the problem as inference in factorial HMMs, and finds an approximate solution to the resulting quadratic integer program via quadratic programming. Here we take a more principled approach, better suited to integer programming problems, and find an approximate optimum by combining convex semidefinite relaxations with randomized rounding, as well as with a scalable ADMM method that exploits the special structure of the resulting semidefinite program. Simulation results demonstrate the superiority of our methods both in synthetic and real-world datasets.

#195 Markov Chain Sampling in Discrete Probabilistic Models with Constraints

Chengtao Li (MIT)
Suvrit Sra (MIT)
Stefanie Jegelka (MIT)

We study probability measures induced by set functions with constraints. Such measures arise in a variety of real-world settings, where often limited resources, prior knowledge, or other pragmatic considerations can impose hard constraints (e.g., cardinality constraints). For a variety of such probabilistic models, we present theoretical results on mixing times of Markov chains, and show sufficient conditions under which the associated chains mix rapidly. In the unconstrained case, under a further assumption on the probability measure being strongly Rayleigh, we obtain sharper results. As an important corollary, this implies an unconditional proof of fast mixing for a Markov Chain sampler for determinantal point processes. We illustrate our claims by empirically verifying the dependence of mixing times on the key factors that govern our theoretical bounds.

#196 Unsupervised Learning of 3D Structure from Images

Danilo Jimenez Rezende
Ali Eslami (Google DeepMind)
Shakir Mohamed (Google DeepMind)
Peter Battaglia (Google DeepMind)
Max Jaderberg
Nicolas Heess

A key goal of computer vision is to recover the underlying 3D structure that gives rise to 2D observations of the world. If endowed with 3D understanding, agents can abstract away from the complexity of the rendering process to form stable, disentangled representations of scene elements. In this paper we learn strong deep generative models of 3D structures, and recover these structures from 2D images via probabilistic inference. We demonstrate high-quality samples and report log-likelihoods on several datasets, including ShapeNet \cite{chang2015shapenet}, and establish the first benchmarks in the literature. We also show how these models and their inference networks can be trained jointly, end-to-end, and directly from 2D images without any use of ground-truth 3D labels. This demonstrates for the first time the feasibility of learning to infer 3D representations of the world in a purely unsupervised manner.

#197 The Multiple Quantile Graphical Model

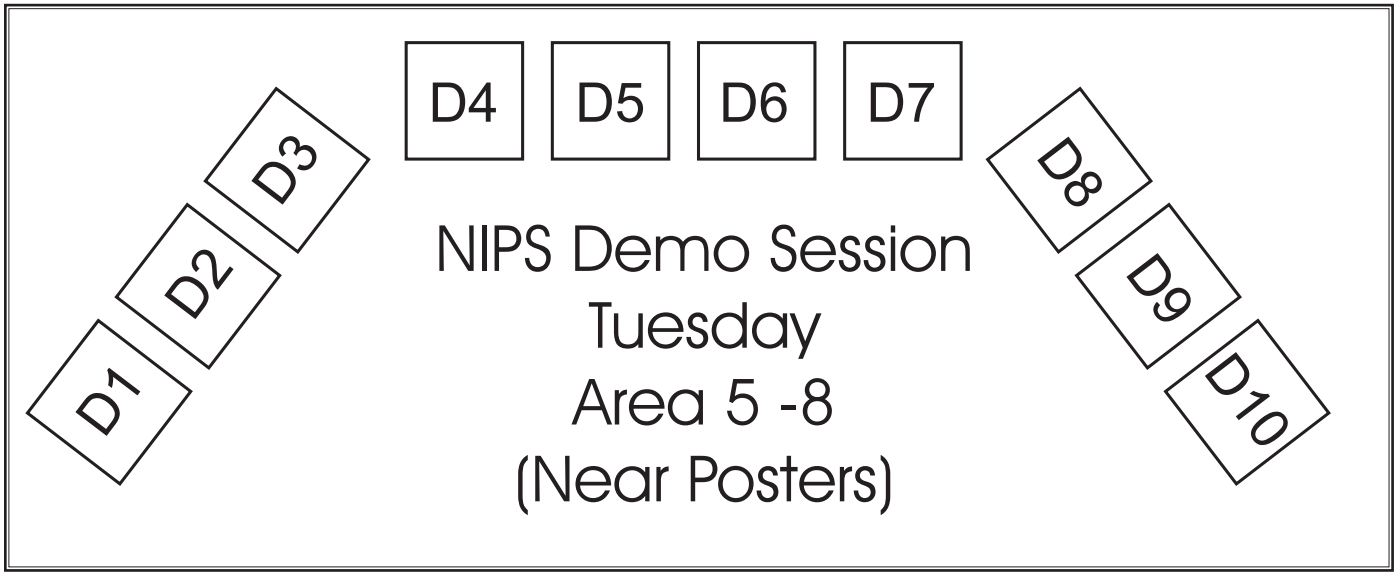
Alnur Ali (Carnegie Mellon Univ.)
J. Zico Kolter
Ryan J Tibshirani

We introduce the Multiple Quantile Graphical Model (MQGM), which extends the neighborhood selection approach of Meinshausen and Buhlmann for learning sparse graphical models. The latter is defined by the basic subproblem of modeling the conditional mean of one variable as a sparse function of all others. Our approach models a set of conditional quantiles of one variable as a sparse function of all others, and hence offers a much richer, more expressive class of conditional distribution estimates. We establish that, under suitable regularity conditions, the MQGM identifies the exact conditional independencies with probability tending to one as the problem size grows, even outside of the usual homoskedastic Gaussian data model. We develop an efficient algorithm for fitting the MQGM using the alternating direction method of multipliers. We also describe a strategy for sampling from the joint distribution that underlies the MQGM estimate. Lastly, we present detailed experiments that demonstrate the flexibility and effectiveness of the MQGM in modeling hetereoskedastic non-Gaussian data.

#198 Linear Contextual Bandits with Knapsacks

Shipra Agrawal
Nikhil Devanur (Microsoft Research)

We consider the linear contextual bandit problem with resource consumption, in addition to reward generation. In each round, the outcome of pulling an arm is a reward as well as a vector of resource consumptions. The expected values of these outcomes depend linearly on the context of that arm. The budget/capacity constraints require that the sum of these vectors doesn't exceed the budget in each dimension. The objective is once again to maximize the total reward. This problem turns out to be a common generalization of classic linear contextual bandits (linContextual), bandits with knapsacks (BwK), and the online stochastic packing problem (OSPP). We present algorithms with near-optimal regret bounds for this problem. Our bounds compare favorably to results on the unstructured version of the problem, where the relation between the contexts and the outcomes could be arbitrary, but the algorithm only competes against a fixed set of policies accessible through an optimization oracle. We combine techniques from the work on linContextual, BwK and OSPP in a nontrivial manner while also tackling new difficulties that are not present in any of these special cases.



D1 Deep Reinforcement Learning for Robotics in DIANNE

Steven Bohez, Elias De Coninck, Sam Leroux, Tim Verbelen

While deep RL has experienced major progress the last years, especially for robotics, integration of learning frameworks with physical and simulated systems is not trivial. This demo will show a practical application of deep RL in robotics using the DIANNE framework. A KUKA YouBot will be tasked to find and retrieve certain objects (e.g. soda cans) within a confined area, relying on a combination of (high-dimensional) sensor inputs. Sensors will be attached to both the robot itself as well as fixed in the environment. For efficiency (and safety), initial training and exploration is performed in a simulated environment using VREP, in which a virtual YouBot gathers experience in order to learn and improve a deep neural network policy. Once sufficiently trained, this policy is than transferred to a physical YouBot in order to finetune it to the real setup. To assist the physical YouBot in evaluating the deep policy, it is equipped with a Nvidia Jetson TX1 embedded GPU. Under the hood, this setup is automated using the DIANNE framework (http://dianne.intec.ugent.be, http://hdl.handle.net/1854/LU-8080319), which on the one hand facilitates designing and training deep learning models, and on the other hand easily integrates with e.g. ROS and VREP to set up environments for reinforcement learning. DIANNE can automatically collect experience from RL agents, use that experience to train RL policies and models and finally update the agent to the newest policy parameters.

D2 Detecting Unexpected Obstacles for Self-Driving Cars: Fusing Deep Learning and Geometric Modeling

Sebastian Ramos, Peter Pinggera, Stefan gehrig, Uwe Franke, Carsten Rother

Our demonstration shows a vision-based system that addresses a challenging and rarely addressed problem for self-driving cars: the detection of generic, small, and unexpected road hazards, such as lost cargo. To the best of our knowledge, our proposed approach to this unsolved problem is the first that leverages both, appearance and contextual cues via a deep convolutional neural network and geometric cues from a stereo-based approach, all combined in a Bayesian framework. Our visual detection framework achieves a very high detection performance with low false positive rates and proves to be robust to illumination changes, varying road appearance as well as 3D road profiles. Our system is able to reliably detect critical obstacles of very low heights (down to 5cm) even at large distances (up to 100m), operating at 22 Hz on our self-driving platform.

D3 Real-time interactive sequence generation with Recurrent Neural Network ensembles

Memo Akten

The demonstration allows users to gesturally 'conduct' the generation of text. We propose a method of real-time continuous control and 'steering' of sequence generation using an ensemble of RNNs, dynamically altering the mixture weights of the models. We demonstrate the method using character based LSTM networks and a gestural interface allowing users to 'conduct' the generation of text.

D4 Project Malmo - Minecraft for AI Research

Katja Hofmann, Matthew A Johnson, Fernando Diaz, Alekh Agarwal, Tim Hutton, David Bignell, Evelyne Viegas

Project Malmo is an open source artificial intelligence (AI) experimentation platform, designed to support fundamental research. Rapid progress in many areas of AI research requires experimentation in interactive settings (agents interact with an environment) that are complex, diverse, dynamic and open, and that provide increasingly more difficult challenges as technology progresses. Project Malmo achieves such flexibility by building on top of Minecraft, a popular computer game with millions of players. The game Minecraft is particularly appealing due to its open ended nature, collaboration with other players, and creativity in game-play.

In this demo, we show the capabilities of the Project Malmo platform and the kind of research they enable. These range from 3D navigation tasks to interactive scenarios where agents converse, compete or collaborate with one another or humans to achieve a goal. The platform is designed to foster collaboration and openness. The result is a cross-platform (Windows, MacOS, Linux), cross-language (e.g., C/C++, Java, C#, Python, Lua) experimentation environment that uses standard data formats to easily exchange tasks and recorded data. Recently, the platform was publicly released as open source software.

D5 Autonomous exploration, active learning and human guidance with open-source Poppy humanoid robot platform and Explauto library

Sébastien Forestier Yoan Mollard
Pierre-Yves Oudeyer

Our demonstration presents an open-source hardware and software platform which allows non-roboticists researchers to conduct machine learning experiments to benchmark algorithms for autonomous exploration and active learning. In particular, in addition to showing the general properties of the platform such as its modularity and usability, we will demonstrate the online functioning of a particular algorithm which allows efficient learning of multiple forward and inverse models and can leverage information from human guidance. A first aspect of our demonstration is to illustrate the ease of use of the 3D printed low-cost Poppy humanoid robotic platform, that allows non-roboticists to quickly set up and program robotic experiments. A second aspect is to show how the Explauto library allows systematic comparison and evaluation of active learning and exploration algorithms in sensorimotor spaces, through a Python API to select already implemented exploration algorithms. The third idea is to showcase Active Model Babbling, an efficient exploration algorithm dynamically choosing which task/goal space to explore and particular goals to reach, and integrating social guidance from humans in real time to drive exploration towards particular objects or actions.

D6 Movidius Fathom: Deep Learning in a USB stick

Cormac Brick Sofiane Yous
Marko Vitez Ian F Hunter
Jack Dashwood

Movidius will present a demonstration of the Fathom Neural Compute Stick, a modular deep learning accelerator in the form of a standard USB stick. Featuring a full-fledged Myriad 2 Vision Processing Unit (VPU), the Fathom Neural Compute Stick allows you to easily integrate your custom trained neural networks in quickly deployable applications. Because of the efficiency and ultra-low power operation of Myriad 2 VPU, the Fathom Neural Compute Stick does not require an external power supply, and can run neural networks on-device in real-time, while only consuming a single Watt of power. Fathom makes it easy to profile, tune and optimize your standard Torch7, TensorFlow or Caffe neural network. Once you find your optimal operational point, Fathom allows your network to run with accelerated performance in embedded environments such as smart cameras, drones, virtual reality headsets and robots

D7 Brain-machine interface spelling device based on reinforcement learning

Inaki Iturrate Ricardo Chavarriaga

The current demonstration will show a novel EEG-based brain-machine interface (BMI) spelling device. It combines real-time decoding of brain activity signals with a reinforcement learning approach to rapidly infer the characters the user wants to write We have developed a communication interface based on multimodal signals that allows users to communicate using different input devices depending on their condition. The proposed solution is an enhanced version of classical matrix-based systems in which machine learning techniques are used to speed up communication and reduce the user's workload. Importantly, the implementation of these techniques also takes into account the speed at which the user can deliver the input and ensures that user's input mistakes have small impact on the communication performance. The interface is composed of a character matrix, in which a moving cursor automatically scans the characters. At the same time, the user gives feedback to the device about the correctness of the movements. Contrasting to conventional systems, the cursor does not move in a pre-defined manner; instead it moves towards the most probable character to be written. This probability is estimated based on a language model and the feedback provided by the user.

D8 Content-based Related Video Recommendations

Joonseok Lee

This is a demo of related video recommendations, seeded from random YouTube videos, and based purely on video content signals. Traditional recommendation systems using collaborative filtering (CF) approaches suggest related videos for a given seed based on how many users have watched a particular candidate video right after watching the seed video. This does not take the video content into account but relies on aggregate user behavior. In this demo, we focus on the cold-start problem, where either the seed and/or the candidate video are freshly uploaded (or undiscovered). We model this as a video content-based similarity learning problem, and learn deep video embeddings trained to predict ground-truth video relationships (identified by a CF co-watch-based system) but using only visual content. It embeds any new video into a 1024-dimensional representation based on its content and pairwise video similarity is computed simply as a dot product in the embedding space. We show that the learned video embeddings generalize beyond simple visual similarity and are able to capture complex semantic relationships.

D9 Logically Complex Symbol Grounding for Interactive Robots by Seq2seq Learning with an LSTM-RNN

Tatsuro Yamada Shingo Murata
Hiroaki Arie Tetsuya Ogata

This study applied the sequence to sequence (seq2seq) learning method for recurrent neural networks (RNN) to learning for interactive robots, which respond to a human's linguistic instructions by generating appropriate behavior. This study extended the method by constructing target data not as unimodal language sequences, but as multimodal sequences of words, vision, and the robot's joint angles. By using them to train the RNN, the robot can acquire the ability to deal with interactive tasks online. In this scheme, not only the relationships between instructions and corresponding behaviors but also the task progression pattern, that is, the repetition of instruction, behavior, and waiting for subsequent instructions, can be autonomously learned from the data, so the execution of the task is achieved by continuous forward propagation alone. This proposal has the following novelty: (1) We implemented a long short-term memory (LSTM)-RNN model trained by the seq2seq method, which is mainly used in the field of natural language processing, for interactive robots in the aforementioned extended way. (2) We dealt with the logical operators "true," "false," "and," and "or," which have not been dealt with in previous studies on integrative learning of language and robot behavior in the research field called symbol emergence in robotics.

WEDNESDAY SESSIONS

9:00 - 9:50 am - INVITED TALK:
Machine Learning and Likelihood-Free Inference in Particle Physics - Kyle Cranmer Area 1 + 2

9:50 - 10:10 am - AWARD TALK:
Matrix Completion has No Spurious Local Minimum Area 1 + 2
Arong Ge · Jason Lee · Tengyu Ma

10:10 - 10:40 am - Coffee Break - P1 & P2

10:40 - 12:20 pm - Track 1: Algorithms Area 3

- **Achieving the KS threshold in the general stochastic block model with linearized acyclic belief propagation**
  Emmanuel Abbe, Colin Sandon
- **Orthogonal Random Features**
  Felix X Yu, Ananda Theertha Suresh, Krzysztof M Choromanski, Daniel N Holtmann-Rice, Sanjiv Kumar
- **Poisson-Gamma dynamical systems**
  Aaron Schein, Hanna Wallach, Mingyuan Zhou
- **The Multiscale Laplacian Graph Kernel**
  Risi Kondor, Horace Pan
- **Stochastic Online AUC Maximization**
  Yiming Ying, Longyin Wen, Siwei Lyu

10:40 - 12:20 pm - Track 2: Applications Area 1 + 2

- **Large-Scale Price Optimization via Network Flow**
  Shinji Ito, Ryohei Fujimaki
- **Probabilistic Modeling of Future Frames from a Single Image**
  Tianfan Xue, Jiajun Wu, Katie Bouman, Bill Freeman
- **Supervised Word Mover's Distance**
  Gao Huang, Chuan Guo, Matt J Kusner, Yu Sun, Fei Sha, Kilian Q Weinberger
- **Beyond Exchangeability: The Chinese Voting Process**
  Moontae Lee, Jin Jin, David Mimno
- **Protein contact prediction from amino acid co-evolution using convolutional networks for graph-valued images**
  Vladimir Golkov, Marcin J Skwark, Antonij Golkov, Alexey Dosovitskiy, Thomas Brox, Jens Meiler, Daniel Cremers

12:20 - 3:00 pm - LUNCH ON YOUR OWN

3:00 - 3:50 pm - INVITED TALK:
Dynamic Legged Robots
Marc Raibert Area 1 + 2

3:50 - 4:20 pm - Coffee Break - P1 & P2

4:20 - 5:40 pm - Track 1: Deep Learning Area 1 + 2

- **Deep Learning without Poor Local Minima**
  Kenji Kawaguchi
- **Universal Correspondence Network**
  Christopher B Choy, Manmohan Chandraker, JunYoung Gwak, Silvio Savarese
- **Learning to Poke by Poking: Experiential Learning of Intuitive Physics**
  Pulkit Agrawal, Ashvin V Nair, Pieter Abbeel, Jitendra Malik, Sergey Levine
- **Learning What and Where to Draw**
  Scott E Reed, Zeynep Akata, Santosh Mohan, Samuel Tenka, Bernt Schiele, Honglak Leer
- **Weight Normalization: A Simple Reparameterization to Accelerate Training of Deep Neural Networks**
  Tim Salimans, Diederik P Kingma

4:20 - 5:40 pm - Track 2: Optimization Area 3

- **Without-Replacement Sampling for Stochastic Gradient Methods**
  Ohad Shamir
- **Regularized Nonlinear Acceleration**
  Damien Scieur, Alexandre d'Aspremont, Francis Bach
- **Linear-Memory and Decomposition-Invariant Linearly Convergent Conditional Gradient Algorithm for Structured Polytopes**
  Dan Garber, Dan Garber, Ofer Meshi
- **Generalization of ERM in Stochastic Convex Optimization: The Dimension Strikes Back**
  Vitaly Feldman
- **Bayesian Optimization with Robust Bayesian Neural Networks**
  Jost Tobias Springenberg, Aaron Klein, Stefan Falkner, Frank Hutter

Wednesday, Dec 7th, 9:00 - 9:50 am

Machine Learning and Likelihood-Free Inference in Particle Physics

Area 1 & 2

Kyle Cranmer (New York Univ.)

Particle physics aims to answer profound questions about the fundamental building blocks of the Universe through enormous data sets collected at experiments like the Large Hadron Collider at CERN. Inference in this context involves two extremes. On one hand the theories of fundamental particle interactions are described by quantum field theory, which is elegant, highly constrained, and highly predictive. On the other hand, the observations come from interactions with complex sensor arrays with uncertain response, which lead to intractable likelihoods. Machine learning techniques with high-capacity models offer a promising set of tools for coping with the complexity of the data; however, we ultimately want to perform inference in the language of quantum field theory. I will discuss likelihood-free inference, generative models, adversarial training, and other recent progress in machine learning from this point of view.

Kyle Cranmer is an Associate Professor of Physics at New York University and affiliated with NYU's Center for Data Science. He is an experimental particle physicists working, primarily, on the Large Hadron Collider, based in Geneva, Switzerland.

He was awarded the Presidential Early Career Award for Science and Engineering in 2007 and the National Science Foundation's Career Award in 2009. Professor Cranmer developed a framework that enables collaborative statistical modeling, which was used extensively for the discovery of the Higgs boson in July, 2012. His current interests are at the intersection of physics and machine learning and include inference in the context of intractable likelihoods, development of machine learning models imbued with physics knowledge, adversarial training for robustness to systematic uncertainty, the use of generative models in the physical sciences, and integration of reproducible workflows in the inference pipeline.

Wednesday, Dec 7th, 9:50 - 10:10 am

Award Talk: Matrix Completion has No Spurious Local Minimum

Area 1 & 2

Rong Ge (Princeton)
Jason Lee (UC Berkely)
Tengyu Ma (Princeton)

Matrix completion is a basic machine learning problem that has wide applications, especially in collaborative filtering and recommender systems. Simple non-convex optimization algorithms are popular and effective in practice. Despite recent progress in proving various non-convex algorithms converge from a good initial point, it remains unclear why random or arbitrary initialization suffices in practice. We prove that the commonly used non-convex objective function for matrix completion has no spurious local minima --- all local minima must also be global. Therefore, many popular optimization algorithms such as (stochastic) gradient descent can provably solve matrix completion with \textit{arbitrary} initialization in polynomial time.

Wednesday, Dec 7th, 3:00 - 3:50 pm

Dynamic Legged Robots

Area 1 & 2

Marc Raibert
(Boston Dynamics)

A new generation of high-performance robots is leaving the laboratory and entering the world. They operate in offices, homes and the field, where ordinary vehicles can not go. They use sensors to see the world around them in order to navigate, interact and understand. Their agility, dexterity, autonomy and intelligence are evolving in ways that promise to free us from the tasks that no human should have to perform. The presentation will give a status report on the work Boston Dynamics is doing to help develop advanced mobile manipulation robots.

Marc Raibert founded Boston Dynamics in 1992 as a spin-off from MIT. Boston Dynamics develops some of the world's most advanced dynamic robots, such as BigDog, Atlas, Cheetah and Spot. These robots are inspired by the remarkable ability of animals to move with agility, mobility, dexterity and speed. A key ingredient of these robots is their dynamic behavior, which contributes to their effectiveness in real-world tasks and their life-like qualities. Before starting Boston Dynamics, Raibert was Professor of Electrical Engineering and Computer Science at MIT from 1986 to 1995. Before that he was Associate Professor of Computer Science and a member of the Robotics Institute at Carnegie Mellon from 1980 to 1986. While at CMU and MIT Raibert founded the Leg Laboratory, a lab that helped establish the scientific basis for highly dynamic robots. Raibert has been a member of the National Academy of Engineering since 2008.

Sessions 10:40 am - 12:20 pm
Algorithms @ Area 3

Achieving the KS threshold in the general stochastic block model with linearized acyclic belief propagation

Emmanuel Abbe
Colin Sandon

The stochastic block model (SBM) has long been studied in machine learning and network science as a canonical model for clustering and community detection. In the recent years, new developments have demonstrated the presence of threshold phenomena for this model, which have set new challenges for algorithms. For the {\it detection} problem in symmetric SBMs, Decelle et al.\ conjectured that the so-called Kesten-Stigum (KS) threshold can be achieved efficiently. This was proved for two communities, but remained open from three communities. We prove this conjecture here, obtaining a more general result that applies to arbitrary SBMs with linear size communities. The developed algorithm is a linearized acyclic belief propagation (ABP) algorithm, which mitigates the effects of cycles while provably achieving the KS threshold in O(n \ln n) time. This extends prior methods by achieving universally the KS threshold while reducing or preserving the computational complexity. ABP is also connected to a power iteration method on a generalized nonbacktracking operator, formalizing the spectral-message passing interplay described in Krzakala et al., and extending results from Bordenave et al.

Orthogonal Random Features

Felix X Yu
Ananda Theertha Suresh
Krzysztof M Choromanski
Daniel N Holtmann-Rice
Sanjiv Kumar (Google)

We present an intriguing discovery related to Random Fourier Features: replacing multiplication by a random Gaussian matrix with multiplication by a properly scaled random orthogonal matrix significantly decreases kernel approximation error. We call this technique Orthogonal Random Features (ORF), and provide theoretical and empirical justification for its effectiveness. Motivated by the discovery, we further propose Structured Orthogonal Random Features (SORF), which uses a class of structured discrete orthogonal matrices to speed up the computation. The method reduces the time cost from \mathcal{O}(d^2) to \mathcal{O}(d log d), where d is the data dimensionality, with almost no compromise in kernel approximation quality compared to ORF. Experiments on several datasets verify the effectiveness of ORF and SORF over the existing methods. We also provide discussions on using the same type of discrete orthogonal structure for a broader range of kernels and applications.

Poisson-Gamma dynamical systems

Aaron Schein (UMass Amherst)
Hanna Wallach (Microsoft Research)
Mingyuan Zhou

This paper presents a dynamical system based on the Poisson-Gamma construction for sequentially observed multivariate count data. Inherent to the model is a novel Bayesian nonparametric prior that ties and shrinks parameters in a powerful way. We develop theory about the model's infinite limit and its steady-state. The model's inductive bias is demonstrated on a variety of real-world datasets where it is shown to learn interpretable structure and have superior predictive performance.

The Multiscale Laplacian Graph Kernel

Risi Kondor
Horace Pan (UChicago)

Many real world graphs, such as the graphs of molecules, exhibit structure at multiple different scales, but most existing kernels between graphs are either purely local or purely global in character. In contrast, by building a hierarchy of nested subgraphs, the Multiscale Laplacian Graph kernels (MLG kernels) that we define in this paper can account for structure at a range of different scales. At the heart of the MLG construction is another new graph kernel, called the Feature Space Laplacian Graph kernel (FLG kernel), which has the property that it can lift a base kernel defined on the vertices of two graphs to a kernel between the graphs. The MLG kernel applies such FLG kernels to subgraphs recursively. To make the MLG kernel computationally feasible, we also introduce a randomized projection procedure, similar to the Nystro¨ m method, but for RKHS operators.

Stochastic Online AUC Maximization

Yiming Ying
Longyin Wen (State University of New York at Albany)
Siwei Lyu (State University of New York at Albany)

Area under ROC (AUC) is a metric which is widely used for measuring the classification performance for imbalanced data. It is of theoretical and practical interest to develop online learning algorithms that maximizes AUC for large-scale data. A specific challenge in developing online AUC maximization algorithm is that the learning objective function is usually defined over a pair of training examples of opposite classes, and existing methods achieves on-line processing with higher space and time complexity. In this work, we propose a new stochastic online algorithm for AUC maximization. In particular, we show that AUC optimization can be equivalently formulated as a convex-concave saddle point problem. From this saddle representation, a stochastic online algorithm (SOLAM) is proposed which has time and space complexity of one datum. We establish theoretical convergence of SOLAM with high probability and demonstrate its effectiveness and efficiency on standard benchmark datasets.

Sessions 10:40 - 12:20 pm
Applications @ Area 1 + 2

Large-Scale Price Optimization via Network Flow

Shinji Ito (NEC Coorporation)
Ryohei Fujimaki

This paper deals with price optimization, which is to find the best pricing strategy that maximizes revenue or profit, on the basis of demand forecasting models. Though recent advances in regression technologies have made it possible to reveal price-demand relationship of a number of multiple products, most existing price optimization methods, such as mixed integer programming formulation, cannot handle tens or hundreds of products because of their high computational costs. To cope with this problem, this paper proposes a novel approach based on network flow algorithms. We reveal a connection between supermodularity of the revenue and cross elasticity of demand. On the basis of this connection, we propose an efficient algorithm that employs network flow algorithms. The proposed algorithm can handle hundreds or thousands of products, and returns an exact optimal solution under an assumption regarding cross elasticity of demand. Even in case in which the assumption does not hold, the proposed algorithm can efficiently find approximate solutions as good as can other state-of-the-art methods, as empirical results show.

Probabilistic Modeling of Future Frames from a Single Image

Tianfan Xue
Jiajun Wu (MIT)
Katie Bouman (MIT)
Bill Freeman

We study the problem of synthesizing a number of likely future frames from a single input image. In contrast to traditional methods, which have tackled this problem in a deterministic or non-parametric way, we propose a novel approach which models future frames in a probabilistic manner. Our proposed method is therefore able to synthesize multiple possible next frames using the same model. Solving this challenging problem involves low- and high-level image and motion understanding for successful image synthesis. Here, we propose a novel network structure, namely a Cross Convolutional Network, that encodes images as feature maps and motion information as convolutional kernels to aid in synthesizing future frames. In experiments, our model performs well on both synthetic data, such as 2D shapes and animated game sprites, as well as on real-wold video data. We show that our model can also be applied to tasks such as visual analogy-making, and present analysis of the learned network representations.

Supervised Word Mover's Distance

Gao Huang
Chuan Guo (Cornell University)
Matt J Kusner
Yu Sun
Fei Sha (University of Southern California)
Kilian Q Weinberger

Accurately measuring the similarity between text documents lies at the core of many real world applications of machine learning. These include web-search ranking, document recommendation, multi-lingual document matching, and article categorization. Recently, a new document metric, the word mover's distance (WMD), has been proposed with unprecedented results on kNN-based document classification. The WMD elevates high quality word embeddings to document metrics by formulating the distance between two documents as an optimal transport problem between the embedded words. However, the document distances are entirely unsupervised and lack a mechanism to incorporate supervision when available. In this paper we propose an efficient technique to learn a supervised metric, which we call the Supervised WMD (S-WMD) metric. Our algorithm learns document distances that measure the underlying semantic differences between documents by leveraging semantic differences between individual words discovered during supervised training. This is achieved with an linear transformation of the underlying word embedding space and tailored word-specific weights, learned to minimize the stochastic leave-one-out nearest neighbor classification error on a per-document level. We evaluate our metric on eight real-world text classification tasks on which S-WMD consistently outperforms almost all of our 26 competitive baselines.

Beyond Exchangeability: The Chinese Voting Process

Moontae Lee (Cornell University)
Jin Jin (Cornell University)
David Mimno (Cornell University)

Many online communities present user-contributed responses, such as reviews of products and answers to questions. User-provided helpfulness votes can highlight the most useful responses, but voting is a social process that can gain momentum based on the popularity of responses and the polarity of existing votes. We propose the Chinese Voting Process (CVP) which models the evolution of helpfulness votes as a self-reinforcing process dependent on position and presentation biases. We evaluate this model on Amazon product reviews and more than 80 StackExchange forums, measuring the intrinsic quality of individual responses and behavioral coefficients of different communities.

Protein contact prediction from amino acid co-evolution using convolutional networks for graph-valued images

Vladimir Golkov (Technical University of Munich)
Marcin J Skwark (Vanderbilt University)
Antonij Golkov (University of Augsburg)
Alexey Dosovitskiy (University of Freiburg)
Thomas Brox (University of Freiburg)
Jens Meiler (Vanderbilt University)
Daniel Cremers (Technical University of Munich)

Proteins are the "building blocks of life", the most abundant organic molecules, and the central focus of most areas of biomedicine. Protein structure is strongly related to protein function, thus structure prediction is a crucial task on the way to solve many biological questions. A contact map is a compact representation of the three-dimensional structure of a protein via the pairwise contacts between the amino acid constituting the protein. We use a convolutional network to calculate protein contact maps from inferred statistical coupling between positions in the protein sequence. The input to the network has an image-like structure amenable to convolutions, but every "pixel" instead of color channels contains a bipartite undirected edge-weighted graph. We propose several methods for treating such "graph-valued images" in a convolutional network. The proposed method outperforms state-of-the-art methods by a large margin. It also allows for a great flexibility with regard to the input data, which makes it useful for studying a wide range of problems.

Sessions 4:20 - 5:40 pm
Deep Learning 2 @ Area 1 + 2

Deep Learning without Poor Local Minima

Kenji Kawaguchi (MIT)

In this paper, we prove a conjecture published in 1989 and also partially address an open problem announced at the Conference on Learning Theory (COLT) 2015. For an expected loss function of a deep nonlinear neural network, we prove the following statements under the independence assumption adopted from recent work: 1) the function is non-convex and non-concave, 2) every local minimum is a global minimum, 3) every critical point that is not a global minimum is a saddle point, and 4) the property of saddle points differs for shallow networks (with three layers) and deeper networks (with more than three layers). Moreover, we prove that the same four statements hold for deep linear neural networks with any depth, any widths and no unrealistic assumptions. As a result, we present an instance, for which we can answer to the following question: how difficult to directly train a deep model in theory? It is more difficult than the classical machine learning models (because of the non-convexity), but not too difficult (because of the nonexistence of poor local minima and the property of the saddle points). We note that even though we have advanced the theoretical foundations of deep learning, there is still a gap between theory and practice.

Universal Correspondence Network

Christopher B Choy (Stanford University)
Manmohan Chandraker (NEC Labs America)
JunYoung Gwak (Stanford University)
Silvio Savarese (Stanford University)

We present a deep learning framework for accurate visual correspondences and demonstrate its effectiveness for both geometric and semantic matching, spanning across rigid motions to intra-class shape or appearance variations. In contrast to previous CNN-based approaches that optimize a surrogate patch similarity objective, we use deep metric learning to directly learn a feature space that preserves either geometric or semantic similarity. Our fully convolutional architecture, along with a novel correspondence contrastive loss allows faster training by effective reuse of computations, accurate gradient computation through the use of thousands of examples per image pair and faster testing with O(n) feedforward passes for n keypoints, instead of O(n^2) for typical patch similarity methods. We propose a convolutional spatial transformer to mimic patch normalization in traditional features like SIFT, which is shown to dramatically boost accuracy for semantic correspondences across intra-class shape variations. Extensive experiments on KITTI, PASCAL and CUB-2011 datasets demonstrate the significant advantages of our features over prior works that use either hand-constructed or learned features.

Learning to Poke by Poking: Experiential Learning of Intuitive Physics

Pulkit Agrawal (UC Berkeley)
Ashvin V Nair (UC Berkeley)
Pieter Abbeel (OpenAI / UC Berkeley / Gradescope)
Jitendra Malik
Sergey Levine (University of Washington)

We investigate an experiential learning paradigm for acquiring an internal model of intuitive physics. Our model is evaluated on a real-world robotic manipulation task that requires displacing objects to target locations by poking. The robot gathered over 400 hours of experience by executing more than 50K pokes on different objects. We propose a novel approach based on deep neural networks for modeling the dynamics of robot's interactions directly from images, by jointly estimating forward and inverse models of dynamics. The inverse model objective provides supervision to construct informative visual features, which the forward model can then predict and in turn regularize the feature space for the inverse model. The interplay between these two objectives creates useful, accurate models that can then be used for multi-step decision making. This formulation has the additional benefit that it is possible to learn forward models in an abstract feature space and thus alleviate the need of predicting pixels. Our experiments show that this joint modeling approach outperforms alternative methods. We also demonstrate that active data collection using the learned model further improves performance.

Learning What and Where to Draw

Scott E Reed (University of Michigan)
Zeynep Akata (Max Planck Institute for Informatics)
Santosh Mohan (University of MIchigan)
Samuel Tenka (University of MIchigan)
Bernt Schiele
Honglak Lee (University of Michigan)

Generative Adversarial Networks (GANs) have recently demonstrated the capability to synthesize compelling real-world images, such as room interiors, album covers, manga, faces, birds, and flowers. While existing models can synthesize images based on global constraints such as a class label or caption, they do not provide control over pose or object location. We propose a new model, the Generative Adversarial What-Where Network (GAWWN), that synthesizes images given instructions describing what content to draw in which location. We show high-quality 128 $\times$ 128 image synthesis on the Caltech-UCSD Birds dataset, conditioned on both informal text descriptions and also object location. Our system exposes control over both the bounding box around the bird and its constituent parts. By modeling the conditional distributions over part locations, our system also enables conditioning on arbitrary subsets of parts (e.g. only the beak and tail), yielding an efficient interface for picking part locations.

Weight Normalization: A Simple Reparameterization to Accelerate Training of Deep Neural Networks

Tim Salimans
Diederik P Kingma

We present weight normalization: a reparameterization of the weight vectors in a neural network that decouples the length of those weight vectors from their direction. By reparameterizing the weights in this way we improve the conditioning of the optimization problem and we speed up convergence of stochastic gradient descent. Our reparameterization is inspired by batch normalization but does not introduce any dependencies between the examples in a minibatch. This means that our method can also be applied successfully to recurrent models such as LSTMs and to noise-sensitive applications such as deep reinforcement learning or generative models, for which batch normalization is less well suited. Although our method is much simpler, it still provides much of the speed-up of full batch normalization. In addition, the computational overhead of our method is lower, permitting more optimization steps to be taken in the same amount of time. We demonstrate the usefulness of our method on applications in supervised image recognition, generative modelling, and deep reinforcement learning.

Sessions 4:20 - 5:40 pm
Optimization @ Area 3

Without-Replacement Sampling for Stochastic Gradient Methods

Ohad Shamir (Weizmann Institute of Science)

Stochastic gradient methods for machine learning and optimization problems are usually analyzed assuming data points are sampled *with* replacement. In contrast, sampling *without* replacement is far less understood, yet in practice it is very common, often easier to implement, and usually performs better. In this paper, we provide competitive convergence guarantees for without-replacement sampling under several scenarios, focusing on the natural regime of few passes over the data. Moreover, we describe a useful application of these results in the context of distributed optimization with randomly-partitioned data, yielding a nearly-optimal algorithm for regularized least squares (in terms of both communication complexity and runtime complexity) under broad parameter regimes. Our proof techniques combine ideas from stochastic optimization, adversarial online learning and transductive learning theory, and can potentially be applied to other stochastic optimization and learning problems.

Regularized Nonlinear Acceleration

Damien Scieur (INRIA - ENS)
Alexandre d'Aspremont
Francis Bach

We describe a convergence acceleration technique for generic optimization problems. Our scheme computes estimates of the optimum from a nonlinear average of the iterates produced by any optimization method. The weights in this average are computed via a simple and small linear system, whose solution can be updated online. This acceleration scheme runs in parallel to the base algorithm, providing improved estimates of the solution on the fly, while the original optimization method is running. Numerical experiments are detailed on classical classification problems.

Linear-Memory and Decomposition-Invariant Linearly Convergent Conditional Gradient Algorithm for Structured Polytopes

Dan Garber
Dan Garber
Ofer Meshi

Recently, several works have shown that natural modifications of the classical conditional gradient method (aka Frank-Wolfe algorithm) for constrained convex optimization, provably converge with a linear rate when the feasible set is a polytope, and the objective is smooth and strongly-convex. However, all of these results suffer from two significant shortcomings: i) large memory requirement due to the need to store an explicit convex decomposition of the current iterate, and as a consequence, large running-time overhead per iteration ii) the worst case convergence rate depends unfavorably on the dimension In this work we present a new conditional gradient variant and a corresponding analysis that improves on both of the above shortcomings. In particular, both memory and computation overheads are only linear in the dimension, and in addition, in case the optimal solution is sparse, the new convergence rate replaces a factor which is at least linear in the dimension in previous works, with a linear dependence on the number of non-zeros in the optimal solution At the heart of our method, and corresponding analysis, is a novel way to compute decomposition-invariant away-steps. While our theoretical guarantees do not apply to any polytope, they apply to several important structured polytopes that capture central concepts such as paths in graphs, perfect matchings in bipartite graphs, marginal distributions that arise in structured prediction tasks, and more. Our theoretical findings are complemented by empirical evidence that shows that our method delivers state-of-the-art performance.

Generalization of ERM in Stochastic Convex Optimization: The Dimension Strikes Back

Vitaly Feldman

In stochastic convex optimization the goal is to minimize a convex function F(x) \doteq \E_{f\sim D}[f(x)] over a convex set \K \subset \R^d where D is some unknown distribution and each f(\cdot) in the support of D is convex over \K. The optimization is based on i.i.d.~samples f^1,f^2,\ldots,f^n from D. A common approach to such problems is empirical risk minimization (ERM) that optimizes F_S(x) \doteq \frac{1}{n}\sum_{i\leq n} f^i(x). Here we consider the question of how many samples are necessary for ERM to succeed and the closely related question of uniform convergence of F_S to F over \K. We demonstrate that in the standard ℓ_p/ℓ_q setting of Lipschitz-bounded functions over a \K of bounded radius, ERM requires sample size that scales linearly with the dimension d. This nearly matches standard upper bounds and improves on Ω(log d) dependence proved for ℓ_2/ℓ_2 setting in (Shalev-Shwartz et al. 2009). In stark contrast, these problems can be solved using dimension-independent number of samples for ℓ_2/ℓ_2 setting and log d dependence for ℓ_1/ℓ_\infty setting using other approaches. We also demonstrate that for a more general class of range-bounded (but not Lipschitz-bounded) stochastic convex programs an even stronger gap appears already in dimension 2.

Bayesian Optimization with Robust Bayesian Neural Networks

Jost Tobias Springenberg (University of Freiburg)
Aaron Klein (University of Freiburg)
Stefan Falkner (University of Freiburg)
Frank Hutter (University of Freiburg)

Bayesian optimization is a prominent method for optimizing expensive to evaluate black-box functions that is prominently applied to tuning the hyperparameters of machine learning algorithms. Despite its successes, the prototypical Bayesian optimization approach - using Gaussian process models - does not scale well to either many hyperparameters or many function evaluations. Attacking this lack of scalability and flexibility is thus one of the key challenges of the field. We present a general approach for using flexible parametric models (neural networks) for Bayesian optimization, staying as close to a truly Bayesian treatment as possible. We obtain scalability through stochastic gradient Hamiltonian Monte Carlo, whose robustness we improve via a scale adaptation. Experiments including multi-task Bayesian optimization with 21 tasks, parallel optimization of deep neural networks and deep reinforcement learning show the power and flexibility of this approach.

#1 **Unsupervised Learning from Noisy Networks with Applications to Hi-C Data**
Bo Wang, Junjie Zhu, Armin Pourshafeie

#2 **Towards Unifying Hamiltonian Monte Carlo and Slice Sampling**
Yizhe Zhang, Xiangyu Wang, Changyou Chen, Ricardo Henao, Kai Fan, Lawrence Carin

#3 **Differential Privacy without Sensitivity**
Kentaro Minami, HItomi Arai, Issei Sato, Hiroshi Nakagawa

#4 **Generalized Correspondence-LDA Models (GC-LDA) for Identifying Functional Regions in the Brain**
Tim Rubin, Sanmi Koyejo, Michael Jones, Tal Yarkoni

#5 **Kronecker Determinantal Point Processes**
Zelda E. Mariet, Suvrit Sra

#6 **Variance Reduction in Stochastic Gradient Langevin Dynamics**
Avinava Dubey, Sashank J. Reddi, Sinead A Williamson, Barnabas Poczos, Alex J Smola, Eric P Xing

#7 **Online Pricing with Strategic and Patient Buyers**
Michal Feldman, Tomer Koren, Roi Livni, Yishay Mansour, Aviv Zohar

#8 **Exploiting the Structure: Stochastic Gradient Methods Using Raw Clusters**
Zeyuan Allen-Zhu, Yang Yuan, Karthik Sridharan

#9 **Clustering Signed Networks with the Geometric Mean of Laplacians**
Pedro Mercado, Francesco Tudisco, Matthias Hein

#10 **Robust Spectral Detection of Global Structures in the Data by Learning a Regularization**
Pan Zhang

#11 **Learning Volumetric 3D Object Reconstruction from Single-View with Projective Transformations**
Xinchen Yan, Jimei Yang, Ersin Yumer, Yijie Guo, Honglak Lee

#12 **Launch and Iterate: Reducing Prediction Churn**
Mahdi Milani Fard, Quentin Cormier, Kevin Canini, Maya Gupta

#14 **Data Poisoning Attacks on Factorization-Based Collaborative Filtering**
Bo Li, Yining Wang, Aarti Singh, Yevgeniy Vorobeychik

#15 **Scaling Memory-Augmented Neural Networks with Sparse Reads and Writes**
Jack Rae, Jonathan J Hunt, Ivo Danihelka, Tim Harley, Andrew W Senior, Greg Wayne, Alex Graves, Timothy Lillicrap

#16 **Optimal Architectures in a Solvable Model of Deep Networks**
Jonathan Kadmon, Haim Sompolinsky

#17 **Scalable Adaptive Stochastic Optimization Using Random Projections**
Gabriel Krummenacher, Brian McWilliams, Yannic Kilcher, Joachim M Buhmann, Nicolai Meinshausen

#18 **Spectral Learning of Dynamic Systems from Nonequilibrium Data**
Hao Wu, Frank Noe

#19 **Local Minimax Complexity of Stochastic Convex Optimization**
sabyasachi chatterjee, John C Duchi, John Lafferty, Yuancheng Zhu

#20 **A Theoretically Grounded Application of Dropout in Recurrent Neural Networks**
Yarin Gal, Zoubin Ghahramani

#21 **Brains on Beats**
Umut Güçlü, Jordy Thielen, Michael Hanke, Marcel van Gerven

#22 **A Communication-Efficient Parallel Algorithm for Decision Tree**
dreamqi Meng, Guolin Ke, Taifeng Wang, Wei Chen, Qiwei Ye, Zhi-Ming Ma, Tieyan Liu

#23 **Leveraging Sparsity for Efficient Submodular Data Summarization**
Erik Lindgren, Shanshan Wu, Alex Dimakis

#24 **Avoiding Imposters and Delinquents: Adversarial Crowdsourcing and Peer Prediction**
Jacob Steinhardt, Gregory Valiant, Moses Charikar

#25 **Designing smoothing functions for improved worst-case competitive ratio in online optimization**
Reza Eghbali, Maryam Fazel

#26 **The Forget-me-not Process**
Kieran Milan, Joel Veness, James Kirkpatrick, Michael Bowling, Anna Koop, Demis Hassabis

#27 **Generating Videos with Scene Dynamics**
Carl Vondrick, Hamed Pirsiavash, Antonio Torralba

#28 **The Robustness of Estimator Composition**
Pingfan Tang, Jeff M Phillips

#29 **Improved Deep Metric Learning with Multi-class N-pair Loss Objective**
Kihyuk Sohn

#30 **Preference Completion from Partial Rankings**
Suriya Gunasekar, Sanmi Koyejo, Joydeep Ghosh

#31 **Bayesian optimization under mixed constraints with a slack-variable augmented Lagrangian**
Victor Picheny, Robert B Gramacy, Stefan Wild, Sebastien Le Digabel

#32 **Privacy Odometers and Filters: Pay-as-you-Go Composition**
Ryan M Rogers, Salil Vadhan, Aaron Roth, Jonathan Ullman

#33 **Large Margin Discriminant Dimensionality Reduction in Prediction Space**
Ehsan Saberian, Jose Costa Pereira, Nuno Nvasconcelos

#34 **Tight Complexity Bounds for Optimizing Composite Objectives**
Blake E Woodworth, Nati Srebro

#35 **Automatic Neuron Detection in Calcium Imaging Data Using Convolutional Networks**
Noah Apthorpe, Alexander Riordan, Robert Aguilar, Jan Homann, Yi Gu, David Tank, H. Sebastian Seung

#36 **Hierarchical Deep Reinforcement Learning: Integrating Temporal Abstraction and Intrinsic Motivation**
Tejas D Kulkarni, Karthik Narasimhan, Ardavan Saeedi, Josh Tenenbaum

#37 **Conditional Image Generation with Pixel CNN Decoders**
Aaron van den Oord, Nal Kalchbrenner, Lasse Espeholt, koray kavukcuoglu, Oriol Vinyals, Alex Graves

#38 **Natural-Parameter Networks: A Class of Probabilistic Neural Networks**
Hao Wang, Xingjian SHI, Dit-Yan Yeung

#39 **Long-term Causal Effects via Behavioral Game Theory**
Panos Toulis, David C Parkes

#40 **PerforatedCNNs: Acceleration through Elimination of Redundant Convolutions**
Michael Figurnov, Aizhan Ibraimova, Dmitry P Vetrov, Pushmeet Kohli

#41 **A Probabilistic Programming Approach To Probabilistic Data Analysis**
Feras Saad, Vikash K Mansinghka

#42 **Learning Bayesian networks with ancestral constraints**
Eunice Yuh-Jie Chen, Yujia Shen, Arthur Choi, Adnan Darwiche

#43 **Solving Random Systems of Quadratic Equations via Truncated Generalized Gradient Flow**
Gang Wang, Georgios Giannakis

#44 **Balancing Suspense and Surprise: Timely Decision Making with Endogenous Information Acquisition**
Ahmed M. Alaa Ibrahim, Mihaela Van Der Schaar

#45 **Blind Optimal Recovery of Signals**
Dmitry Ostrovsky, Zaid Harchaoui, Anatoli Juditsky, Arkadi S Nemirovski

#46 **Spatiotemporal Residual Networks for Video Action Recognition**
Christoph Feichtenhofer, Axel Pinz, Richard Wildes

#47 **CMA-ES with Optimal Covariance Update and Storage Complexity**
Oswin Krause, Dídac Rodríguez Arbonès, Christian Igel

#48 **An End-to-End Approach for Natural Language to IFTTT Program Translation**
Chang Liu, Xinyun Chen, Richard Shin, Mingcheng Chen, Dawn Song

#49 **The Sound of APALM Clapping: Faster Nonsmooth Nonconvex Optimization with Stochastic Asynchronous PALM**
damekdavis Davis, Brent Edmunds, Madeleine Udell

#50 **Efficient Algorithm for Streaming Submodular Cover**
Ashkan Norouzi-Fard, Abbas Bazzi, Ilija Bogunovic, Marwa El Halabi, Ya-Ping Hsieh, Volkan Cevher

#51 **Attend, Infer, Repeat: Fast Scene Understanding with Generative Models**
Ali Eslami, Nicolas Heess, Theophane Weber, Yuval Tassa, David Szepesvari, koray kavukcuoglu, Geoffrey E Hinton

#52 **An ensemble diversity approach to supervised binary hashing**
Miguel A. Carreira-Perpinan, Ramin Raziperchikolaei

#53 **End-to-End Goal-Driven Web Navigation**
Rodrigo Nogueira, Kyunghyun Cho

#54 **The Power of Adaptivity in Identifying Statistical Alternatives**
Kevin Jamieson, Daniel Haas, Benjamin Recht

#55 **A Probabilistic Framework for Deep Learning**
Ankit B Patel, Minh Tan Nguyen, Richard Baraniuk

#56 **Minimax Estimation of Maximum Mean Discrepancy with Radial Kernels**
Ilya Tolstikhin, Bharath K. Sriperumbudur, Prof. Bernhard Schölkopf

#57 **Adaptive Neural Compilation**
Rudy R Bunel, Alban Desmaison, Pawan K Mudigonda, Pushmeet Kohli, Philip Torr

#58 **Tagger: Deep Unsupervised Perceptual Grouping**
Klaus Greff, Antti Rasmus, Mathias Berglund, Tele Hotloo Hao, Harri Valpola

#59 **A scaled Bregman theorem with applications**
Richard Nock, Aditya Menon, Cheng Soon Ong

#60 **Learning feed-forward one-shot learners**
Luca Bertinetto, João F. Henriques, Jack Valmadre, Philip Torr, Andrea Vedaldi

#61 **Error Analysis of Generalized Nyström Kernel Regression**
Hong Chen, Haifeng Xia, Heng Huang

#62 **Breaking the Bandwidth Barrier: Geometrical Adaptive Entropy Estimation**
Weihao Gao, Sewoong Oh, Pramod Viswanath

#63 **Asynchronous Parallel Greedy Coordinate Descent**
Yang You, Xiangru Lian, Ji Liu, Hsiang-Fu (Rofu) Yu, Inderjit S Dhillon, James Demmel, Cho-Jui Hsieh

#64 **Structured Prediction Theory Based on Factor Graph Complexity**
Corinna Cortes, Vitaly Kuznetsov, Mehryar Mohri, Scott Yang

#65 **Parameter Learning for Log-supermodular Distributions**
Tatiana Shpakova, Francis Bach

#66 **Exact Recovery of Hard Thresholding Pursuit**
Xiaotong Yuan, Ping Li, Tong Zhang

#67 **A New Liftable Class for First-Order Probabilistic Inference**
Seyed Mehran Kazemi, Angelika Kimmig, Guy Van den Broeck, David Poole

#68 **Variational Inference in Mixed Probabilistic Submodular Models**
Josip Djolonga, Sebastian Tschiatschek, Andreas Krause

#69 **Unifying Count-Based Exploration and Intrinsic Motivation**
Marc Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, Remi Munos

#70 **Approximate maximum entropy principles via Goemans-Williamson with applications to provable variational methods**
Andrej Risteski, Yuanzhi Li

#71 **A Multi-step Inertial Forward-Backward Splitting Method for Non-convex Optimization**
Jingwei Liang, Jalal Fadili, Gabriel Peyré

#72 **Fast and Flexible Monotonic Functions with Ensembles of Lattices**
Mahdi Milani Fard, Kevin Canini, Andy Cotter, Jan Pfeifer, Maya Gupta

#73 **Architectural Complexity Measures of Recurrent Neural Networks**
Saizheng Zhang, Yuhuai Wu, Tong Che, Zhouhan Lin, Roland Memisevic, Russ Salakhutdinov, Yoshua Bengio

#74 **Online Convex Optimization with Unconstrained Domains and Losses**
Ashok Cutkosky, Kwabena A Boahen

#75 **Split LBI: An Iterative Regularization Path with Structural Sparsity**
Chendi Huang, Xinwei Sun, Jiechao Xiong, Yuan Yao

#76 **Variational Autoencoder for Deep Learning of Images, Labels and Captions**
Yunchen Pu, Zhe Gan, Ricardo Henao, Xin Yuan, Chunyuan Li, Andrew Stevens, Lawrence Carin

#77 **Recovery Guarantee of Non-negative Matrix Factorization via Alternating Updates**
Yuanzhi Li, Yingyu Liang, Andrej Risteski

#78 **Proximal Deep Structured Models**
Shenlong Wang, Sanja Fidler, Raquel Urtasun

#79 **Safe Policy Improvement by Minimizing Robust Baseline Regret**
Mohammad Ghavamzadeh, Marek Petrik, Yinlam Chow

#80 **A Pseudo-Bayesian Algorithm for Robust PCA**
Tae-Hyun Oh, Yasuyuki Matsushita, In Kweon, David Wipf

#81 **Learning values across many orders of magnitude**
Hado van Hasselt, Baguez Aguez, Matteo Hessel, Volodymyr Mnih, David Silver

#82 **Single Pass PCA of Matrix Products**
Shanshan Wu, Srinadh Bhojanapalli, Sujay Sanghavi, Alex Dimakis

#83 **Convolutional Neural Fabrics**
Shreyas Saxena, Jakob Verbeek

#84 **Generative Shape Models: Joint Text Recognition and Segmentation with Very Little Training Data**
Xinghua Lou, Ken Kansky, Wolfgang Lehrach, CC Laan, Bhaskara Marthi, D. Phoenix, Dileep George

#85 **Mixed vine copulas as joint models of spike counts and local field potentials**
Arno Onken, Stefano Panzeri

#86 **Optimal Black-Box Reductions Between Optimization Objectives**
Zeyuan Allen-Zhu, Elad Hazan

#87 **Dialog-based Language Learning**
Jason E Weston

#88 **Online Bayesian Moment Matching for Topic Modeling with Unknown Number of Topics**
Wei-Shou Hsu, Pascal Poupart

#89 **A Sparse Interactive Model for Matrix Completion with Side Information**
Jin Lu, Guannan Liang, Jiangwen Sun, Jinbo Bi

#90 **Truncated Variance Reduction: A Unified Approach to Bayesian Optimization and Level-Set Estimation**
Ilija Bogunovic, Jonathan Scarlett, Andreas Krause, Volkan Cevher

#91 **On Mixtures of Markov Chains**
Rishi Gupta, Ravi Kumar, Sergei Vassilvitskii

#92 **High Dimensional Structured Superposition Models**
Qilong Gu, Arindam Banerjee

#93 **Finite Sample Prediction and Recovery Bounds for Ordinal Embedding**
Lalit Jain, Kevin Jamieson, Rob Nowak

#94 **What Makes Objects Similar: A Unified Multi-Metric Learning Approach**
Han-Jia Ye, De-Chuan Zhan, Xue-Min Si, Yuan Jiang, Zhi-Hua Zhou

#95 **Unsupervised Learning of Spoken Language with Visual Context**
David Harwath, Antonio Torralba, James Glass

#96 **Cyclades: Conflict-free Asynchronous Machine Learning**
Xinghao Pan, Maximilian Lam, Stephen Tu, Dimitrios Papailiopoulos, Ce Zhang, Michael I Jordan, Kannan Ramchandran, Chris Ré, Benjamin Recht

#97 **Disease Trajectory Maps**
Peter Schulam, Raman Arora

#98 **Fast ε-free Inference of Simulation Models with Bayesian Conditional Density Estimation**
George Papamakarios, Iain Murray

#99 **Stochastic Structured Prediction under Bandit Feedback**
Artem Sokolov, Julia Kreutzer, Stefan Riezler

#100 **Learning under uncertainty: a comparison between R-W and Bayesian approach**
Crane Huang, Martin Paulus

#101 **Minimax Optimal Alternating Minimization for Kernel Nonparametric Tensor Learning**
Taiji Suzuki, Heishiro Kanagawa, Hayato Kobayashi, Nobuyuki Shimizu, Yukihiro Tagami

#102 **On the Recursive Teaching Dimension of VC Classes**
Xi Chen, Xi Chen, Yu Cheng, Bo Tang

#103 **Dimension-Free Iteration Complexity of Finite Sum Optimization Problems**
Yossi Arjevani, Ohad Shamir

#104 **f-GAN: Training Generative Neural Samplers using Variational Divergence Minimization**
Sebastian Nowozin, Botond Cseke, Ryota Tomioka

#105 **Low-Rank Regression with Tensor Responses**
Guillaume Rabusseau, Hachem Kadri

#106 **Double Thompson Sampling for Dueling Bandits**
Huasen Wu, Xin Liu

#107 **Linear dynamical neural population models through nonlinear embeddings**
Yuanjun Gao, Evan W Archer, Liam Paninski, John Cunningham

#108 **Regret Bounds for Non-decomposable Metrics with Missing Labels**
Nagarajan Natarajan, Prateek Jain

#109 **Dynamic matrix recovery from incomplete observations under an exact low-rank constraint**
Liangbei Xu, Mark Davenport

#110 **Rényi Divergence Variational Inference**
Yingzhen Li, Richard E Turner

#111 **Confusions over Time: An Interpretable Bayesian Model to Characterize Trends in Decision Making**
Himabindu Lakkaraju, Jure Leskovec

#112 **Adaptive Averaging in Accelerated Descent Dynamics**
Walid Krichene, Alexandre Bayen, Peter L Bartlett

#113 **Bayesian Optimization for Probabilistic Programs**
Tom Rainforth, Tuan-Anh Le, Jan-Willem van de Meent, Michael A Osborne, Frank Wood

#114 **Efficient Globally Convergent Stochastic Optimization for Canonical Correlation Analysis**
Weiran Wang, Jialei Wang, Dan Garber, Dan Garber, Nati Srebro

#115 **A Unified Approach for Learning the Parameters of Sum-Product Networks**
Han Zhao, Pascal Poupart, Geoffrey J Gordon

#116 **Feature-distributed sparse regression: a screen-and-clean approach**
Jiyan Yang, Michael W Mahoney, Michael Saunders, Yuekai Sun

#117 **Backprop KF: Learning Discriminative Deterministic State Estimators**
Tuomas Haarnoja, Anurag Ajay, Sergey Levine, Pieter Abbeel

#118 **Swapout: Learning an ensemble of deep architectures**
Saurabh Singh, Derek Hoiem, David Forsyth

#119 **Assortment Optimization Under the Mallows model**
Antoine Desir, Vineet Goyal, Srikanth Jagabathula, Danny Segev

#120 **Operator Variational Inference**
Rajesh Ranganath, Dustin Tran, Jaan Altosaar, David Blei

#121 **Select-and-Sample for Spike-and-Slab Sparse Coding**
Abdul-Saboor Sheikh, Jörg Lücke

#122 **Fast recovery from a union of subspaces**
Chinmay Hegde, Piotr Indyk, Ludwig Schmidt

#123 **Ladder Variational Autoencoders**
Casper Kaae Sønderby, Tapani Raiko, Lars Maaløe, Søren Kaae Sønderby, Ole Winther

#124 **SPALS: Fast Alternating Least Squares via Implicit Leverage Scores Sampling**
Dehua Cheng, Richard Peng, Yan Liu, Kimis Perros

#125 **CRF-CNN: Modeling Structured Information in Human Pose Estimation**
Xiao Chu, Wanli Ouyang, hongsheng Li, Xiaogang Wang

#126 **A Consistent Regularization Approach for Structured Prediction**
Carlo Ciliberto, Lorenzo Rosasco, Alessandro Rudi

#127 **Refined Lower Bounds for Adversarial Bandits**
Sébastien Gerchinovitz, Tor Lattimore

#128 **Learning Deep Embeddings with Histogram Loss**
Evgeniya Ustinova, Victor Lempitsky

#129 **Solving Marginal MAP Problems with NP Oracles and Parity Constraints**
Yexiang Xue, zhiyuan li, Stefano Ermon, Carla P Gomes, Bart Selman

#130 **Kernel Bayesian Inference with Posterior Regularization**
Yang Song, Jun Zhu, Yong Ren

#131 **Learning Influence Functions from Incomplete Observations**
Xinran He, Ke Xu, David Kempe, Yan Liu

#132 **General Tensor Spectral Co-clustering for Higher-Order Data**
Tao Wu, Austin R Benson, David Gleich

#133 **Bayesian latent structure discovery from multi-neuron recordings**
Scott Linderman, Ryan P Adams, Jonathan W Pillow

#134 **Estimating the Size of a Large Network and its Communities from a Random Sample**
Lin Chen, Amin Karbasi, Forrest W. Crawford

#135 **Wasserstein Training of Restricted Boltzmann Machines**
Grégoire Montavon, Klaus-Robert Müller, Marco Cuturi

#136 **Deep ADMM-Net for Compressive Sensing MRI**
yan yang, Jian Sun, Huibin Li, Zongben Xu

#137 **Maximization of Approximately Submodular Functions**
Thibaut Horel, Yaron Singer

#138 **Combining Low-Density Separators with CNNs**
Yu-Xiong Wang, Martial Hebert

#139 **Learning Sensor Multiplexing Design through Back-propagation**
Ayan Chakrabarti

#140 **High resolution neural connectivity from incomplete tracing data using nonnegative spline regression**
Kameron D Harris, Stefan Mihalas, Eric Shea-Brown

#141 **Learning a Probabilistic Latent Space of Object Shapes via 3D Generative-Adversarial Modeling**
Chengkai Zhang, Jiajun Wu, Tianfan Xue, Bill Freeman, Josh Tenenbaum

#142 **Learning Sparse Gaussian Graphical Models with Overlapping Blocks**
Mohammad Javad Hosseini, Su-In Lee

#143 **Multi-step learning and underlying structure in statistical models**
Maia Fraser

#144 **Dynamic Network Surgery for Efficient DNNs**
Yiwen Guo, Anbang Yao, Yurong Chen

#145 **Active Nearest-Neighbor Learning in Metric Spaces**
Aryeh Kontorovich, Sivan Sabato, Ruth Urner

#146 **Discriminative Gaifman Models**
Mathias Niepert

#147 **Professor Forcing: A New Algorithm for Training Recurrent Networks**
Alex M Lamb, Anirudh Goyal, Ying Zhang, Saizheng Zhang, Aaron C Courville, Yoshua Bengio

#148 **Pruning Random Forests for Prediction on a Budget**
Feng Nan, Joseph Wang, Venkatesh Saligrama

#149 **Multistage Campaigning in Social Networks**
Mehrdad Farajtabar, Xiaojing Ye, Sahar Harati, Le Song, Hongyuan Zha

#150 **Coevolutionary Latent Feature Processes for Continuous-Time User-Item Interactions**
Yichen Wang, Nan Du, Rakshit Trivedi, Le Song

#151 **Coordinate-wise Power Method**
Qi Lei, Kai Zhong, Inderjit S Dhillon

#152 **Barzilai-Borwein Step Size for Stochastic Gradient Descent**
Conghui Tan, Shiqian Ma, Yu-Hong Dai, Yuqiu Qian

#153 **Fast learning rates with heavy-tailed losses**
Vu C Dinh, Lam S Ho, Binh Nguyen, Duy Nguyen

#154 **CliqueCNN: Deep Unsupervised Exemplar Learning**
Miguel A Bautista, Artsiom Sanakoyeu, Ekaterina Tikhoncheva, Bjorn Ommer

#155 **Guided Policy Search as Approximate Mirror Descent**
William H Montgomery, Sergey Levine

#156 **Structured Sparse Regression via Greedy Hard Thresholding**
Prateek Jain, Nikhil Rao, Inderjit S Dhillon

#157 **Learning in Games: Robustness of Fast Convergence**
Dylan J Foster, zhiyuan li, Thodoris Lykouris, Karthik Sridharan, Eva Tardos

#158 **Measuring the reliability of MCMC inference with Bidirectional Monte Carlo**
Roger B Grosse, Siddharth Ancha, Dan Roy

#159 **Average-case hardness of RIP certification**
Tengyao Wang, Quentin Berthet, Yaniv Plan

#160 **Provable Efficient Online Matrix Completion via Non-convex Stochastic Gradient Descent**
Chi Jin, Sham Kakade, Praneeth Netrapalli

#161 **Infinite Hidden Semi-Markov Modulated Interaction Point Process**
matt zhang, Peng Lin, Ting Guo, Yang Wang, Yang Wang, Fang Chen

#162 **Selective inference for group-sparse linear models**
Fan Yang, Rina Foygel Barber, Prateek Jain, John Lafferty

#163 **Deep Neural Networks with Inexact Matching for Person Re-Identification**
Arulkumar Subramaniam, Moitreya Chatterjee, Anurag Mittal

#164 **Accelerating Stochastic Composition Optimization**
Mengdi Wang, Ji Liu

#165 **Learning Bound for Parameter Transfer Learning**
Wataru Kumagai

#166 **Can Active Memory Replace Attention?**
Łukasz Kaiser, Samy Bengio

#167 **Understanding the Effective Receptive Field in Deep Convolutional Neural Networks**
Wenjie Luo, Yujia Li, Raquel Urtasun, Richard Zemel

#168 **Local Similarity-Aware Deep Feature Embedding**
Chen Huang, Chen Change Loy, Xiaoou Tang

#169 **End-to-End Kernel Learning with Supervised Convolutional Kernel Networks**
Julien Mairal

#170 **Single-Image Depth Perception in the Wild**
Weifeng Chen, Zhao Fu, Dawei Yang, Jia Deng

#171 **Toward Deeper Understanding of Neural Networks: The Power of Initialization and a Dual View on Expressivity**
Amit Daniely, Roy Frostig, Yoram Singer

#172 **R-FCN: Object Detection via Region-based Fully Convolutional Networks**
jifeng dai, Yi Li, Kaiming He, Jian Sun

#173 **Consistent Estimation of Functions of Data Missing Non-Monotonically and Not at Random**
Ilya Shpitser

#174 **Without-Replacement Sampling for Stochastic Gradient Methods**
Ohad Shamir

#175 **Probabilistic Modeling of Future Frames from a Single Image**
Tianfan Xue, Jiajun Wu, Katie Bouman, Bill Freeman

#176 **Learning What and Where to Draw**
Scott E Reed, Zeynep Akata, Santosh Mohan, Samuel Tenka, Bernt Schiele, Honglak Lee

#177 **Stochastic Online AUC Maximization**
Yiming Ying, Longyin Wen, Siwei Lyu

#178 **Deep Learning without Poor Local Minima**
Kenji Kawaguchi

#179 **Regularized Nonlinear Acceleration**
Damien Scieur, Alexandre d'Aspremont, Francis Bach

#180 **Learning to Poke by Poking: Experiential Learning of Intuitive Physics**
Pulkit Agrawal, Ashvin V Nair, Pieter Abbeel, Jitendra Malik, Sergey Levine

#181 **Weight Normalization: A Simple Reparameterization to Accelerate Training of Deep Neural Networks**
Tim Salimans, Diederik P Kingma

#182 **Linear-Memory and Decomposition-Invariant Linearly Convergent Conditional Gradient Algorithm for Structured Polytopes**
Dan Garber, Dan Garber, Ofer Meshi

#183 **Achieving the KS threshold in the general stochastic block model with linearized acyclic belief propagation**
Emmanuel Abbe, Colin Sandon

#184 **Orthogonal Random Features**
Felix X Yu, Ananda Theertha Suresh, Krzysztof M Choromanski, Daniel N Holtmann-Rice, Sanjiv Kumar

#185 **Universal Correspondence Network**
Christopher B Choy, Manmohan Chandraker, JunYoung Gwak, Silvio Savarese

#186 **The Multiscale Laplacian Graph Kernel**
Risi Kondor, Horace Pan

#187 **Generalization of ERM in Stochastic Convex Optimization: The Dimension Strikes Back**
Vitaly Feldman

#188 **Large-Scale Price Optimization via Network Flow**
Shinji Ito, Ryohei Fujimaki

#189 **Bayesian Optimization with Robust Bayesian Neural Networks**
Jost Tobias Springenberg, Aaron Klein, Stefan Falkner, Frank Hutter

#190 **Protein contact prediction from amino acid co-evolution using convolutional networks for graph-valued images**
Vladimir Golkov, Marcin J Skwark, Antonij Golkov, Alexey Dosovitskiy, Thomas Brox, Jens Meiler, Daniel Cremers

#191 **Supervised Word Mover's Distance**
Gao Huang, Chuan Guo, Matt J Kusner, Yu Sun, Fei Sha, Kilian Q Weinberger

#192 **Beyond Exchangeability: The Chinese Voting Process**
Moontae Lee, Jin Jin, David Mimno

#193 **Poisson-Gamma dynamical systems**
Aaron Schein, Hanna Wallach, Mingyuan Zhou

#194 **Interpretable Distribution Features with Maximum Testing Power**
Wittawat Jitkrittum, Zoltán Szabó, Kacper P Chwialkowski, Arthur Gretton

#195 **Dense Associative Memory for Pattern Recognition**
Dmitry Krotov, John J. Hopfield

#196 **Relevant sparse codes with variational information bottleneck**
Matthew Chalk, Olivier Marre, Gasper Tkacik

#197 **Examples are not enough, learn to criticize! Criticism for Interpretability**
Been Kim, Sanmi Koyejo, Rajiv Khanna

#198 **Showing versus doing: Teaching by demonstration**
Mark K Ho, Michael Littman, James MacGlashan, Fiery Cushman, Joe Austerweil, Joe L Austerweil

#1 Unsupervised Learning from Noisy Networks with Applications to Hi-C Data

Bo Wang (Stanford Univ.)
Junjie Zhu (Stanford Univ.)
Armin Pourshafeie (Stanford Univ.)

Complex networks play an important role in a plethora of disciplines in natural science. It poses an essential challenge in network science to clean up the noisy observed networks. Existing methods utilize labeled data to alleviate the noise effect in the network. However, labeled data is usually expensive to collect while unlabeled data can be gathered cheaply. Therefore, we propose an optimization framework to mine useful structures from noisy networks in an unsupervised manner. Local structures are exploited together with global patterns in the network through the optimization framework. We extend our method to incorporate multi-resolution networks in order to add further resistance to large noise. Our framework is generalizable to utilize partial labels to enhance the performance. We apply our method to multi-resolution Hi-C data to recover clusters in genomic regions. Additionally, we use Capture-C-generated partial labels to further denoise the Hi-C network. We empirically demonstrate the effectiveness of our framework in denoising the network and improving community detection results.

#2 Towards Unifying Hamiltonian Monte Carlo and Slice Sampling

Yizhe Zhang (Duke Univ.)
Xiangyu Wang (Duke Univ.)
Changyou Chen
Ricardo Henao
Kai Fan (Duke Univ.)
Lawrence Carin

We unify slice sampling and Hamiltonian Monte Carlo (HMC) sampling, demonstrating their connection via the Hamiltonian-Jacobi equation from Hamiltonian mechanics. This insight enables extension of HMC and slice sampling to a broader family of samplers, called Monomial Gamma Samplers (MGS). We provide a theoretical analysis of the mixing performance of such samplers, proving that in the limit of a single parameter, the MGS draws decorrelated samples from the desired target distribution. We further show that as this parameter tends toward this limit, performance gains are achieved at a cost of increasing numerical difficulty and some practical convergence issues. Our theoretical results are validated with synthetic data and real-world applications.

#3 Differential Privacy without Sensitivity

Kentaro Minami (The Univ. of Tokyo)
HItomi Arai (The Univ. of Tokyo)
Issei Sato (The Univ. of Tokyo)
Hiroshi Nakagawa

The exponential mechanism is a general method to construct a randomized estimator that satisfies (\varepsilon, 0)-differential privacy. Recently, Wang et al. showed that the Gibbs posterior, which is a data-dependent probability distribution that contains the Bayesian posterior, is essentially equivalent to the exponential mechanism under certain boundedness conditions on the loss function. While the exponential mechanism provides a way to build an (\varepsilon, 0)-differential private algorithm, it requires boundedness of the loss function, which is quite stringent for some learning problems. In this paper, we focus on (\varepsilon, δ)-differential privacy of Gibbs posteriors with convex and Lipschitz loss functions. Our result extends the classical exponential mechanism, allowing the loss functions to have an unbounded sensitivity.

#4 Generalized Correspondence-LDA Models (GC-LDA) for Identifying Functional Regions in the Brain

Tim Rubin (Indiana Univ.)
Sanmi Koyejo (UIUC)
Michael Jones (Indiana Univ.)
Tal Yarkoni (Univ. of Texas at Austin)

This paper presents Generalized Correspondence-LDA (GC-LDA), a generalization of the Correspondence-LDA model that allows for variable spatial representations to be associated with topics, and increased flexibility in terms of the strength of the correspondence between data types induced by the model. We present three variants of GC-LDA, each of which associates topics with a different spatial representation, and apply them to a corpus of neuroimaging data. In the context of this dataset, each topic corresponds to a functional brain region, where the region's spatial extent is captured by a probability distribution over neural activity, and the region's cognitive function is captured by a probability distribution over linguistic terms. We illustrate the qualitative improvements offered by GC-LDA in terms of the types of topics extracted with alternative spatial representations, as well as the model's ability to incorporate a-priori knowledge from the neuroimaging literature. We furthermore demonstrate that the novel features of GC-LDA improve predictions for missing data.

#5 Kronecker Determinantal Point Processes

Zelda E. Mariet (MIT)
Suvrit Sra (MIT)

Determinantal Point Processes (DPPs) are probabilistic models over all subsets a ground set of N items. They have recently gained prominence in several applications that rely on diverse subsets. However, their applicability to large problems is still limited due to O(N^3) complexity of core tasks such as sampling and learning. We enable efficient sampling and learning for DPPs by introducing KronDPP, a DPP model whose kernel matrix decomposes as a tensor product of multiple smaller kernel matrices. This decomposition immediately enables fast exact sampling. But contrary to what one may expect, leveraging the Kronecker product structure for speeding up DPP learning turns out to be more difficult. We overcome this challenge, and derive batch and stochastic optimization algorithms for efficiently learning the parameters of a KronDPP.

#6 Variance Reduction in Stochastic Gradient Langevin Dynamics

Avinava Dubey (Carnegie Mellon Univ.)
Sashank J. Reddi (Carnegie Mellon Univ.)
Sinead A Williamson
Barnabas Poczos
Alex J Smola
Eric P Xing (Carnegie Mellon Univ.)

Stochastic gradient-based MCMC methods such as Langevin dynamics are useful tools for posterior inference on large scale datasets in many machine learning applications. These methods scale to large datasets by using noisy gradients calculated using a mini-batch or subset of the dataset. However, the high variance inherent in these noisy gradients degrades performance and leads to slower mixing. In this paper, we present techniques for reducing variance in stochastic Langevin dynamics, yielding novel stochastic MCMC methods that improve performance by reducing the variance in the stochastic gradient. We show that our proposed method has better theoretical guarantees on convergence rate than stochastic Langevin dynamics. This is complemented by impressive empirical results obtained on a variety of real world datasets, and on four different machine learning tasks (regression, classification, independent component analysis and mixture model). These theoretical and empirical contributions combine to make a compelling case for using variance reduction in stochastic MCMC methods.

#7 Online Pricing with Strategic and Patient Buyers

Michal Feldman (TAU)
Tomer Koren (Technion---Israel Inst. of Technology)
Roi Livni (Huji)
Yishay Mansour (Microsoft)
Aviv Zohar (huji)

We consider a seller with an unlimited supply of a single good, who is faced with a stream of T buyers. lowest price in that window of time, or not buy at all. Each buyer has a window of time in which she would like to purchase, and would buy at the lowest price in that window, provided that this price is lower than her private value (and otherwise, does not buy at all). In this setting, we give an algorithm that attains O(T^{2/3}) regret over any sequence of T buyers, and prove that no algorithm can perform better in the worst case. and derive a tight regret bound of \Theta(T^{2/3}) over T transactions. constant

#8 Exploiting the Structure: Stochastic Gradient Methods Using Raw Clusters

Zeyuan Allen-Zhu (Princeton Univ.)
Yang Yuan (Cornell Univ.)
Karthik Sridharan (Univ. of Pennsylvania)

The amount of data available in the world is growing faster than our ability to deal with it. However, if we take advantage of the internal structure, data may become much smaller for machine learning purposes. In this paper we focus on one of the fundamental machine learning tasks, empirical risk minimization (ERM), and provide faster algorithms with the help from the clustering structure of the data. We introduce a simple notion of raw clustering that can be efficiently computed from the data, and propose two algorithms based on clustering information. Our accelerated algorithm ClusterACDM is built on a novel Haar transformation applied to the dual space of the ERM problem, and our variance-reduction based algorithm ClusterSVRG introduces a new gradient estimator using clustering. Our algorithms outperform their classical counterparts ACDM and SVRG respectively.

#9 Clustering Signed Networks with the Geometric Mean of Laplacians

Pedro Mercado (Saarland Univ.)
Francesco Tudisco (Saarland Univ.)
Matthias Hein (Saarland Univ.)

Signed networks allow to model positive and negative relationships. We analyze existing extensions of spectral clustering to signed networks. It turns out that existing approaches do not recover the ground truth clustering in several situations where either the positive or the negative network structures contain no noise. Our analysis shows that these problems arise as existing approaches take some form of arithmetic mean of the Laplacians of the positive and negative part. As a solution we propose to use the geometric mean of the Laplacians of positive and negative part and show that it outperforms the existing approaches. While the geometric mean of matrices is computationally expensive, we show that eigenvectors of the geometric mean can be computed efficiently, leading to a numerical scheme for sparse matrices which is of independent interest.

#10 Robust Spectral Detection of Global Structures in the Data by Learning a Regularization

Pan Zhang (Institute of Theoretical Physics)

Spectral methods are popular in detecting global structures in the given data that can be represented as a matrix. However when the data matrix is sparse or noisy, classic spectral methods usually fail to work, due to localization of eigenvectors (or singular vectors) induced by the sparsity or noise. In this work, we propose a general method to solve the localization problem by learning a regularization matrix from the localized eigenvectors. Using matrix perturbation analysis, we demonstrate that the learned regularizations suppress down the eigenvalues associated with localized eigenvectors and enable us to recover the informative eigenvectors representing the global structure. We show applications of our method in several inference problems: community detection in networks, clustering from pairwise similarities, rank estimation and matrix completion problems. Using extensive experiments, we illustrate that our method solves the localization problem and works down to the theoretical detectability limits in different kinds of synthetic data. This is in contrast with existing spectral algorithms based on data matrix, non-backtracking matrix, Laplacians and those with rank-one regularizations, which perform poorly in the sparse case with noise.

#11 Learning Volumetric 3D Object Reconstruction from Single-View with Projective Transformations

Xinchen Yan (Univ. of Michigan)
Jimei Yang
Ersin Yumer (Adobe Research)
Yijie Guo (Univ. of Michigan)
Honglak Lee (Univ. of Michigan)

Understanding the 3D world is a fundamental problem in machine learning, computer vision and robotics. However, learning a good representation of 3D objects is still an open problem due to the high dimensionality of the data and many factors involved. In this work, we investigate the task of single-view 3D object reconstruction from a learning agent's perspective. By formulating the learning procedure as an interaction between 3D and 2D representations, we propose to learn such interactions using an encoder-decoder network with a novel loss defined by the projective transformation. We demonstrate the capacity of the encode-decoder network in generating 3D volume from a single 2D image with three sets of experiments: (1) learning from single-class objects; (2) learning from multi-class objects and (3) testing on novel object classes. Results show that superior performance and better generalization ability are obtained for 3D object reconstruction when the projection loss involved. More importantly, with the projection loss, the encoder-decoder network can be properly trained only from 2D observations without accessing the ground-truth 3D volumes of target objects.

#12 Launch and Iterate: Reducing Prediction Churn

Mahdi Milani Fard
Quentin Cormier (Google)
Kevin Canini
Maya Gupta

Practical applications of machine learning often involve successive training iterations with changes to features and training examples. Ideally, changes in the output of any new model should only be improvements (wins) over the previous iteration, but in practice the predictions may change neutrally for many examples, resulting in extra net-zero wins and losses, that we refer to as churn. These changes in the predictions are problematic for usability for some applications, and make it harder and more expensive to measure if a change is statistically significant positive. In this paper, we formulate the problem and present a stabilization operator to regularize a classifier towards a previous classifier. We use a Markov chain Monte Carlo stabilization operator to produce a model with more consistent predictions but without degrading accuracy. We investigate the properties of the proposal with theoretical analysis. Experiments on benchmark datasets for different classification algorithms demonstrate the method and the substantial of churn reduction it can provide.

#14 Data Poisoning Attacks on Factorization-Based Collaborative Filtering

Bo Li (Vanderbilt Univ.)
Yining Wang (Carnegie Mellon Univ.)
Aarti Singh (Carnegie Mellon Univ.)
Yevgeniy Vorobeychik (Vanderbilt Univ.)

Recommendation and collaborative filtering systems are important in modern information and e-commerce applications. As these systems are becoming increasingly popular in industry, their outputs could affect business decision making, introducing incentives for an adversarial party to compromise the availability or integrity of such systems. We introduce a data poisoning attack on collaborative filtering systems. We demonstrate how a powerful attacker with full knowledge of the learner can generate malicious data so as to maximize his/her malicious objectives, while at the same time mimicking normal user behaviors to avoid being detected. While the complete knowledge assumption seems extreme, it enables a robust assessment of the vulnerability of collaborative filtering schemes to highly motivated attacks. We present efficient solutions for two popular factorization-based collaborative filtering algorithms: the alternative minimization formulation and the nuclear norm minimization method. Finally, we test the effectiveness of our proposed algorithms on real-world data and discuss potential defensive strategies.

#15 Scaling Memory-Augmented Neural Networks with Sparse Reads and Writes

Jack Rae (DeepMind)
Jonathan J Hunt
Ivo Danihelka
Tim Harley (DeepMind)
Andrew W Senior
Greg Wayne
Alex Graves
Timothy Lillicrap (DeepMind)

Recently introduced architectures in which neural networks are augmented with memory, such as Neural Turing Machines and Memory Networks, have the ability to learn algorithmic solutions to tasks. These models appear promising for applications such as language modeling and machine translation. However, they scale poorly in both space and time as the amount of memory grows --- limiting their applicability to real-world domains. Here, we present an end-to-end differentiable memory access scheme, which we call Sparse Access Memory (SAM), that retains the representational power of the original approach whilst training efficiently with very large memories. We show that SAM achieves asymptotic lower bounds in space and time complexity, and find that an implementation runs 1,000\times faster and 3,000\times more memory efficient than non-sparse models. SAM can learn with comparable data efficiency to existing models on a set of problems, including synthetic and natural data, and scale to tasks with 100,000s of time steps and memories.

#16 Optimal Architectures in a Solvable Model of Deep Networks

Jonathan Kadmon (Hebrew Univ.)
Haim Sompolinsky

Deep neural networks have received a considerable attention due to the success of their training for real world machine learning applications. They are also of great interest to the understanding of sensory processing in cortical sensory hierarchies. The purpose of this work is to advance our theoretical understanding of the computational benefits of these architectures. Using a simple model of clustered noisy inputs and a simple learning rule, we provide analytically derived recursion relations describing the propagation of the signals along the deep network. Analysis of these equations, and defining performance measures, we show that this model network has an optimal depth and explore the dependence of the optimal architecture on the system parameters.

#17 Scalable Adaptive Stochastic Optimization Using Random Projections

Gabriel Krummenacher (ETH Zurich)
Brian McWilliams (Disney Research)
Yannic Kilcher (ETH Zurich)
Joachim M Buhmann (ETH Zurich)
Nicolai Meinshausen

ADAGRAD has gained popularity as an optimization technique for stochastic em- pirical risk minimization and in particular for training deep neural networks. The most commonly used and studied variant maintains a diagonal matrix approxima- tion to second order information by accumulating past gradients which are used to tune the step size adaptively. In certain situations the full-matrix variant of ADAGRAD is expected to attain better performance, however in high dimensions it is computationally impractical. We present ADA-LR and RADAGRAD two compu- tationally efficient approximations to full-matrix ADAGRAD based on randomized dimensionality reduction. They are able to capture correlations in the gradients and achieve similar performance to full-matrix ADAGRAD but at a computational cost comparable to the diagonal variant. We show that the regret of ADA-LR is close to the regret of full-matrix ADAGRAD which can have an up-to exponentially smaller dependence on the dimension than the diagonal variant. Empirically, we show that ADA-LR and RADAGRAD perform similarly to full-matrix ADAGRAD. On the task of training neural networks, our proposed methods achieve faster convergence than diagonal ADAGRAD.

#18 Spectral Learning of Dynamic Systems from Nonequilibrium Data

Hao Wu (Free Univ. of Berlin)
Frank Noe

Abstract Observable operator models (OOMs) and related models are one of the most important and powerful tools for modeling and analyzing stochastic systems. They can exactly describe dynamics of finite-rank systems, and be efficiently learned from data by moment based algorithms. Almost all OOM learning algorithms are developed based on the assumption of equilibrium data which is very difficult to guarantee in real life, especially for complex processes with large time scales. In this paper, we derive a nonequilibrium learning algorithm for OOMs, which dismisses this assumption and can effectively extract the equilibrium dynamics of a system from nonequilibrium observation data. In addition, we propose binless OOMs for the application of nonequilibrium learning to continuous-valued systems. In comparison with the other OOMs with continuous observations, binless OOMs can achieve consistent estimation from nonequilibrium data with only linear computational complexity.

#19 Local Minimax Complexity of Stochastic Convex Optimization

sabyasachi chatterjee (Univ. of Chicago)
John C Duchi
John Lafferty
Yuancheng Zhu (Univ. of Chicago)

We extend the traditional worst-case, minimax analysis of stochastic convex optimization by introducing a localized form of minimax complexity for individual functions. Our main result gives function-specific lower and upper bounds on the number of stochastic subgradient evaluations needed to optimize either the function or its ``hardest local alternative" to a given numerical precision. The bounds are expressed in terms of a localized and computational analogue of the modulus of continuity that is central to statistical minimax analysis. We show how the computational modulus of continuity can be explicitly calculated in concrete cases, and relates to the curvature of the function at the optimum. We also prove a superefficiency result that demonstrates it is a meaningful benchmark, acting as a computational analogue of the Fisher information in statistical estimation. The nature and practical implications of the results are demonstrated in simulations.

#20 A Theoretically Grounded Application of Dropout in Recurrent Neural Networks

Yarin Gal (Univ. of Cambridge)
Zoubin Ghahramani

Recurrent neural networks (RNNs) stand at the forefront of many recent developments in deep learning. Yet a major difficulty with these models is their tendency to overfit, with dropout shown to fail when applied to recurrent layers. Recent results at the intersection of Bayesian modelling and deep learning offer a Bayesian interpretation of common deep learning techniques such as dropout. This grounding of dropout in approximate Bayesian inference suggests an extension of the theoretical results, offering insights into the use of dropout with RNN models. We apply this new variational inference based dropout technique in LSTM and GRU models, assessing it on language modelling and sentiment analysis tasks. The new approach outperforms existing techniques, and to the best of our knowledge improves on the single model state-of-the-art in language modelling with the Penn Treebank (73.4 test perplexity). This extends our arsenal of variational tools in deep learning.

#21 Brains on Beats

Umut Güçlü (Radboud Univ.)
Jordy Thielen (Radboud Univ.)
Michael Hanke (Otto-von-Guericke Univ. Magdeburg)
Marcel van Gerven (Radboud Univ.)

We developed task-optimized deep neural networks (DNNs) that achieved state-of-the-art performance in different evaluation scenarios for automatic music tagging. These DNNs were subsequently used to probe the neural representations of music. Representational similarity analysis revealed the existence of a representational gradient across the superior temporal gyrus (STG). Anterior STG was shown to be more sensitive to low-level stimulus features encoded in shallow DNN layers whereas posterior STG was shown to be more sensitive to high-level stimulus features encoded in deep DNN layers.

#22 A Communication-Efficient Parallel Algorithm for Decision Tree

dreamqi Meng (Peking Univ.)
Guolin Ke (Microsoft Research)
Taifeng Wang (Microsoft Research)
Wei Chen (Microsoft Research)
Qiwei Ye (Microsoft Research)
Zhi-Ming Ma (Academy of Mathematics and Systems Science)
Tieyan Liu (Microsoft Research)

Decision tree (and its extensions such as Gradient Boosting Decision Trees and Random Forest) is a widely used machine learning algorithm, due to its practical effectiveness and model interpretability. With the emergence of big data, there is an increasing need to parallelize the training process of decision tree. However, most existing attempts along this line suffer from high communication costs. In this paper, we propose a new algorithm, called {Parallel Voting Decision Tree (PV-Tree)}, to tackle this challenge. After partitioning the training data onto a number of (e.g., M) machines, this algorithm performs both local voting and global voting in each iteration. For local voting, the top-k attributes are selected from each machine according to its local data. Then, the indices of these top attributes are aggregated by a server, and the globally top-2k attributes are determined by a majority voting among these local candidates. Finally, the full-grained histograms of the globally top-2k attributes are collected from local machines in order to identify the best (most informative) attribute and its split point. PV-Tree can achieve a very low communication cost (independent of the total number of attributes) and thus can scale out very well. Furthermore, theoretical analysis shows that this algorithm can learn a near optimal decision tree, since it can find the best attribute with a large probability. Our experiments on real-world datasets show that PV-Tree significantly outperforms the existing parallel decision tree algorithms in the tradeoff between accuracy and efficiency.

#23 Leveraging Sparsity for Efficient Submodular Data Summarization

Erik Lindgren (Univ. of Texas at Austin)
Shanshan Wu (UT Austin)
Alex Dimakis

The facility location problem is widely used for summarizing large datasets and has found applications such as sensor placement, image retrieval, and clustering. A significant problem is that optimizing such functions typically requires the calculation of pairwise benefits for all items in the dataset, which is computationally infeasible for large problems. For this reason, recent work proposed to only calculate nearest-neighbor benefits and showed how this idea can dramatically accelerate this submodular optimization problem. One limitation of the existing work was that several strong assumptions were required to obtain provable approximation guarantees. In this paper we show that solving the sparsified problem will be close to optimal under minimal assumptions. We then analyze a different method of sparsification that is a better model for nearest neighbor methods such as Locality Sensitive Hashing (LSH) to accelerate the nearest-neighbor computations and extend the use of the problem to a broader family of similarities. We experimentally validate our approach using multiple datasets and similarity measures and demonstrate that it rapidly gives interpretable summaries for large datasets.

#24 Avoiding Imposters and Delinquents: Adversarial Crowdsourcing and Peer Prediction

Jacob Steinhardt (Stanford Univ.)
Gregory Valiant
Moses Charikar (Stanford Univ.)

We consider a crowdsourcing model in which n workers are asked to rate the quality of n items previously generated by other workers. An unknown set of α n workers generate reliable ratings, while the remaining workers may behave arbitrarily and possibly adversarially. The manager of the experiment can also manually evaluate the quality of a small number of items, and wishes to curate together almost all of the high-quality items with at most an fraction of low-quality items. Perhaps surprisingly, we show that this is possible with an amount of work required of the manager, and each worker, that does not scale with n: the dataset can be curated with $O(1/\beta\alpha\epsilon^4)$ ratings per worker, and $O(1/\beta\epsilon^2)$ ratings by the manager, where β is the fraction of high-quality items. Our results extend to the more general setting of peer prediction, including peer grading in online classrooms.

#25 Designing smoothing functions for improved worst-case competitive ratio in online optimization

Reza Eghbali (Univ. of washington)
Maryam Fazel (Univ. of Washington)

Online optimization covers problems such as online resource allocation, online bipartite matching, adwords (a central problem in e-commerce and advertising), and adwords with separable concave returns. We analyze the worst case competitive ratio of two primal-dual algorithms for a class of online convex (conic) optimization problems that contains the previous examples as special cases defined on the positive orthant. We derive a sufficient condition on the objective function that guarantees a constant worst case competitive ratio (greater than or equal to \frac{1}{2}) for monotone objective functions. Using the same framework, we also derive the competitive ratio for problems with objective functions that are not monotone. We provide new examples of online problems on the positive orthant and the positive semidefinite cone that satisfy the sufficient condition. We show how smoothing can improve the competitive ratio of these algorithms, and in particular for separable functions, we show that the optimal smoothing can be derived by solving a convex optimization problem. This result allows us to directly optimize the comptitive ratio bound over a class of smoothing functions, and hence {design} effective smoothing customized for a given cost function.

#26 The Forget-me-not Process

Kieran Milan (DeepMind)
Joel Veness
James Kirkpatrick (DeepMind)
Michael Bowling
Anna Koop (Univ. of Alberta)
Demis Hassabis

We introduce the Forget-me-not Process, an efficient, non-parametric meta-algorithm for online probabilistic sequence prediction for piecewise stationary, repeating sources. Our method works by taking a Bayesian approach to partition a stream of data into postulated task-specific segments, while simultaneously building a model for each task. We provide regret guarantees with respect to piecewise stationary data sources under the logarithmic loss, and validate the method empirically across a range of sequence prediction and task identification problems.

#27 Generating Videos with Scene Dynamics

Carl Vondrick (MIT)
Hamed Pirsiavash
Antonio Torralba

Understanding object motions and scene dynamics is a core problem in computer vision. For both video recognition tasks and video generation tasks, a model of how objects move and scenes transform is needed. However, creating a model of dynamics is challenging because there is a vast number of ways that objects and scenes can change. In this paper, we introduce an approach that learns some of this temporal knowledge from large amounts of unlabeled video. We present a generative adversarial network that learns to generate tiny videos with fairly realistic motions for some scene categories. Our experiments suggest that carefully designing the generator network to explicitly model an active foreground and a stationary background can yield slightly more realistic videos. Moreover, our experiments suggest that a representation emerges that is useful for classifying human actions. We believe that generative video models can have a large impact in many applications, such as simulations, forecasting, and video recognition tasks.

#28 The Robustness of Estimator Composition

Pingfan Tang (Univ. of Utah)
Jeff M Phillips (Univ. of Utah)

We formalize notions of robustness for composite estimators via the notion of a breakdown point. A composite estimator successively applies two (or more) estimators: on data decomposed into disjoint parts, it applies the first estimator on each part, then the second estimator on the outputs of the first estimator. And so on, if the composition is of more than two estimators. Informally, the breakdown point is the minimum fraction of data points which if significantly modified will also significantly modify the output of the estimator, so it is typically desirable to have a large breakdown point. Our main result shows that, under mild conditions on the individual estimators, the breakdown point of the composite estimator is the product of the breakdown points of the individual estimators. We also demonstrate several scenarios, ranging from regression to statistical testing, where this analysis is easy to apply, useful in understanding worst case robustness, and sheds powerful insights onto the associated data analysis.

#29 Improved Deep Metric Learning with Multi-class N-pair Loss Objective

Kihyuk Sohn

Deep metric learning has gained much popularity in recent years, following the success of deep learning. However, existing frameworks of deep metric learning based on contrastive loss and triplet loss often suffer from slow convergence, partially because they employ only one negative example while not interacting with the other negative classes in each update. As a result, one needs additional expensive data sampling process or combination with classification loss in order to yield competitive performance. In this paper, we propose to address this problem with a new metric learning objective called multi-class N-pair loss. The proposed objective firstly generalizes triplet loss by allowing joint comparison among more than one negative examples–more specifically, N-1 negative examples–and secondly reduces the computational burden of evaluating deep embedding vectors via an efficient batch construction strategy using only N pairs of examples, instead of (N+1)×N. We demonstrate the superiority of our proposed multi-class N-pair loss to the triplet loss as well as other competing loss functions for a variety of tasks on several visual recognition benchmark, including fine-grained object recognition and verification, image clustering and retrieval, and face verification and identification.

#30 Preference Completion from Partial Rankings

Suriya Gunasekar (UT Austin)
Sanmi Koyejo (UIUC)
Joydeep Ghosh (UT Austin)

We propose a novel and efficient algorithm for the preference completion problem, which involves jointly estimating individualized rankings for a set of entities over a shared set of items, based on a limited number of observed affinity values. Our approach exploits the observation that while preferences are often recorded as numerical scores, the predictive quantity of interest is the underlying rankings. Thus, attempts to closely match the recorded scores may lead to overfitting and impair generalization performance. Instead, we propose an estimator that directly fits the underlying rank order, combined with nuclear norm constraints to encourage low rank parameters. Besides (approximate) correctness of the ranking order, the proposed estimator makes no generative assumption on the numerical scores of the observations. One consequence is that the proposed estimator can fit any consistent entity--specific partial ranking over a subset of the items represented as a directed acyclic graph (DAG), generalizing standard techniques that can only fit preference scores. Despite this generality, for supervision representing total or blockwise total orders, the computational complexity of our algorithm is within a log factor of the standard algorithms for nuclear norm regularization based estimates for matrix completion. We further show promising empirical results for a novel and challenging application of collaboratively ranking of the associations between brain--regions and cognitive neuroscience terms.

#31 Bayesian optimization under mixed constraints with a slack-variable augmented Lagrangian

Victor Picheny (Institut National de la Recherche Agronomique)
Robert B Gramacy (Virginia Tech)
Stefan Wild (Argonne National Lab)
Sebastien Le Digabel (École Polytechnique de Montréal)

An augmented Lagrangian (AL) can convert a constrained optimization problem into a sequence of simpler (e.g., unconstrained) problems which are then usually solved with local solvers. Recently, surrogate-based Bayesian optimization (BO) sub-solvers have been successfully deployed in the AL framework for a more global search in the presence of inequality constraints; however a drawback was that expected improvement (EI) evaluations relied on Monte Carlo. Here we introduce an alternative slack variable AL, and show that in this formulation the EI may be evaluated with library routines. The slack variables furthermore facilitate equality as well as inequality constraints, and mixtures thereof. We show our new slack "ALBO" compares favorably to the original. Its superiority over conventional alternatives is reinforced on several new mixed constraint examples.

#32 Privacy Odometers and Filters: Pay-as-you-Go Composition

Ryan M Rogers (Univ. of Pennsylvania)
Salil Vadhan (Harvard Univ.)
Aaron Roth
Jonathan Ullman

In this paper we initiate the study of adaptive composition in differential privacy when the length of the composition, and the privacy parameters themselves can be chosen adaptively, as a function of the outcome of previously run analyses. This case is much more delicate than the setting covered by existing composition theorems, in which the algorithms themselves can be chosen adaptively, but the privacy parameters must be fixed up front. Indeed, it isn't even clear how to define differential privacy in the adaptive parameter setting. We proceed by defining two objects which cover the two main use cases of composition theorems. A privacy filter is a stopping time rule that allows an analyst to halt a computation before his pre-specified privacy budget is exceeded. A privacy odometer allows the analyst to track realized privacy loss as he goes, without needing to pre-specify a privacy budget. We show that unlike the case in which privacy parameters are fixed, in the adaptive parameter setting, these two use cases are distinct. We show that there exist privacy filters with bounds comparable (up to constants) with existing privacy composition theorems. We also give a privacy odometer that nearly matches non-adaptive private composition theorems, but is sometimes worse by a small asymptotic factor. Moreover, we show that this is inherent, and that any valid privacy odometer in the adaptive parameter setting must lose this factor, which shows a formal separation between the filter and odometer use-cases.

#33 Large Margin Discriminant Dimensionality Reduction in Prediction Space

Ehsan Saberian (Netflix)
Jose Costa Pereira (UC San Diego)
Nuno Nvasconcelos (UC San Diego)

In this paper we establish a duality between boosting and SVM, and use this to derive a novel discriminant dimensionality reduction algorithm. In particular, using the multiclass formulation of boosting and SVM we note that both use a combination of mapping and linear classification to maximize the multiclass margin. In SVM this is implemented using a pre-defined mapping (induced by the kernel) and optimizing the linear classifiers. In boosting the linear classifiers are pre-defined and the mapping (predictor) is learned through combination of weak learners. We argue that the intermediate mapping, e.g. boosting predictor, is preserving the discriminant aspects of the data and by controlling the dimension of this mapping it is possible to achieve discriminant low dimensional representations for the data. We use the aforementioned duality and propose a new method (LADDER) that jointly learns the mapping and the linear classifiers in an efficient manner. This leads to a data-driven mapping which can embed data into any number of dimensions. Experimental results show that this embedding can significantly improve performance on tasks such as hashing and image/scene classification.

#34 Tight Complexity Bounds for Optimizing Composite Objectives

Blake E Woodworth (Toyota Technological Institute)
Nati Srebro

We provide tight upper and lower bounds on the complexity of minimizing the average of m convex functions using gradient and prox oracles of the component functions. We show a significant gap between the complexity of deterministic vs randomized optimization. For smooth functions, we show that accelerated gradient descent (AGD) and accelerated SVRG (A-SVRG) are optimal in the deterministic and randomized settings respectively, and that a gradient oracle is sufficient for the optimal rate. For non-smooth functions, having access to prox oracles reduces the complexity and we present optimal methods based on smoothing AGD that improve over methods using just gradient accesses.

#35 Automatic Neuron Detection in Calcium Imaging Data Using Convolutional Networks

Noah Apthorpe (Princeton Univ.)
Alexander Riordan (Princeton Univ.)
Robert Aguilar (Princeton Univ.)
Jan Homann (Princeton Univ.)
Yi Gu (Princeton Univ.)
David Tank (Princeton Univ.)
H. Sebastian Seung (Princeton Univ.)

Calcium imaging is an important technique for monitoring the activity of thousands of neurons simultaneously. As calcium imaging datasets grow in size, automated detection of individual neurons is becoming important. Here we apply a supervised learning approach to this problem and show that convolutional networks can achieve near-human accuracy and superhuman speed. Accuracy is superior to the popular PCA/ICA method, based on precision and recall relative to ground truth annotation by a human expert. These results suggest that convolutional networks are an efficient and flexible tool for the analysis of large-scale calcium imaging data.

#36 Hierarchical Deep Reinforcement Learning: Integrating Temporal Abstraction and Intrinsic Motivation

Tejas D Kulkarni (MIT)
Karthik Narasimhan (MIT)
Ardavan Saeedi (MIT)
Josh Tenenbaum

Learning goal-directed behavior in environments with sparse feedback is a major challenge for reinforcement learning algorithms. The primary difficulty arises due to insufficient exploration, resulting in an agent being unable to learn robust value functions. Intrinsically motivated agents can explore new behavior for its own sake rather than to directly solve problems. Such intrinsic behaviors could eventually help the agent solve tasks posed by the environment. We present hierarchical-DQN (h-DQN), a framework to integrate hierarchical value functions, operating at different temporal scales, with intrinsically motivated deep reinforcement learning. A top-level value function learns a policy over intrinsic goals, and a lower-level function learns a policy over atomic actions to satisfy the given goals. h-DQN allows for flexible goal specifications, such as functions over entities and relations. This provides an efficient space for exploration in complicated environments. We demonstrate the strength of our approach on two problems with very sparse, delayed feedback: (1) a complex discrete stochastic decision process with stochastic transitions, and (2) the classic ATARI game `Montezuma's Revenge'.

#37 Conditional Image Generation with Pixel CNN Decoders

Aaron van den Oord (DeepMind)
Nal Kalchbrenner
Lasse Espeholt
koray kavukcuoglu (DeepMind)
Oriol Vinyals
Alex Graves

Recent progress in generative models has shown that state of the art log-likelihood models like PixelRNNs can produce visually high quality samples. An important unanswered problem is to be able to model the distribution of images conditioned on a given arbitrary context. In this work, we explore conditional image generation with a new image density model based on the recent PixelCNN algorithm. The model can be conditioned on a class label or more generally on any latent vector representation that can be produced from an encoder or inference process. We demonstrate that when conditioned on class labels, the model can generate a large diversity of realistic scenes that contain animals or objects. When conditioned on a single image of unseen face, the model is able to generate a variety of new portraits of the same person with different facial expressions, poses and lighting conditions. Additionally the gated convolutional layers in the proposed model improve the baseline performance of PixelCNN to match PixelRNN models whilst significantly reducing the computational requirements and achieving state of the art log-likelihood performance on the ImageNet dataset.

#38 Natural-Parameter Networks: A Class of Probabilistic Neural Networks

Hao Wang (HKUST)
Xingjian SHI
Dit-Yan Yeung

Neural networks (NN) have achieved state-of-the-art performance in various applications. Unfortunately in applications where training data is insufficient, they are often prone to overfitting. One effective way to alleviate this problem is to exploit the Bayesian approach by using Bayesian neural networks (BNN). Another shortcoming of NN is the lack of flexibility to customize different distributions for the weights and neurons according to the data, as is often done in probabilistic graphical models. To address these problems, we propose a class of probabilistic neural networks, dubbed natural-parameter networks (NPN), as a novel Bayesian treatment of NN. NPN allows the usage of arbitrary exponential-family distributions to model the weights and neurons. Different from traditional NN and BNN, NPN takes distributions as input and goes through layers of transformation before producing distributions to match the target output distributions. As a Bayesian treatment, efficient backpropagation (BP) is performed to learn the natural parameters for the distributions over both the weights and neurons. The output distributions of each layer, as byproducts, may be used as second-order representations for the associated tasks such as link prediction. Experiments on real-world datasets show that NPN can achieve state-of-the-art performance.

#39 Long-term Causal Effects via Behavioral Game Theory

Panos Toulis (Univ. of Chicago)
David C Parkes (Harvard Univ.)

Recently, A/B experiments have become extremely important for reliably comparing a new treatment B (e.g., pricing policy) against the baseline treatment A, say, in terms of revenue. In such comparisons within a multiagent economy (e.g., the Uber ecosystem) the effect of treatment B relative to A is said to be {\em causal} if the revenue when {\em all} agents operate under B is different than the revenue when {\em all} agents operate under A. Since in an experiment only a fraction of agents are observed to operate under one particular treatment, classical statistical methods have been developed to {\em estimate} the aforementioned causal effect from the observed experimental data. But one crucial shortcoming of the classical methodology is that it doesn't take into account the dynamical nature of the response of multiagent systems to experimental treatments. As agents adapt to the new pricing policy B the causal effect on revenue defined after such adaptation period, the {\em long-term causal effect}, is not captured by classical methodology even though it is clearly more representative of the comparative value of the new policy. Here, we formalize the problem and propose a methodology to estimate long-term causal effects. Central to our approach is a model of agent behaviors that (i) predicts how agents would behave under different assignments, and, (ii) predicts how agents would behave in the long term by leveraging game theory. We formally show that these two prediction tasks enable the estimation of long-term causal effects under suitable assumptions, which we state explicitly.

#40 PerforatedCNNs: Acceleration through Elimination of Redundant Convolutions

Michael Figurnov (Skolkovo Inst. of Sc and Tech)
Aizhan Ibraimova (Skolkovo Institute of Science and Tech.)
Dmitry P Vetrov
Pushmeet Kohli

We propose a novel approach to reduce the computational cost of evaluation of convolutional neural networks, a factor that has hindered their deployment in low-power devices such as mobile phones. Inspired by the loop perforation technique from source code optimization, we speed up the bottleneck convolutional layers by skipping their evaluation in some of the spatial positions. We propose and analyze several strategies of choosing these positions. We demonstrate that perforation can accelerate modern convolutional networks such as AlexNet and VGG-16 by a factor of 2x - 4x. Additionally, we show that perforation is complementary to the recently proposed acceleration method of Zhang et al.

#41 A Probabilistic Programming Approach To Probabilistic Data Analysis

Feras Saad (MIT)
Vikash K Mansinghka (MIT)

Probabilistic techniques are central to data analysis, but different approaches can be challenging to apply, combine, and compare. This paper introduces conditional generative population models (CGPMs), a computational abstraction that extends directed graphical models and can be used to describe and compose a broad class of probabilistic data analysis techniques. Examples include discriminative machine learning, hierarchical Bayesian models, multivariate kernel methods, clustering algorithms, and arbitrary probabilistic programs. The paper also demonstrates the integration of CGPMs into BayesDB, a probabilistic programming platform that can express data analysis tasks using a structured query language. The practical value is illustrated in two ways. First, the paper describes an analysis that identifies satellite data records which probably violate Kepler's Third Law by composing causal probabilistic programs with nonparametric Bayes in under 20 lines of probabilistic code. Second, for several representative data analysis tasks, it reports the lines of code and accuracy of various CGPMs, plus comparisons with standard baseline solutions from Python and MATLAB libraries.

#42 Learning Bayesian networks with ancestral constraints

Eunice Yuh-Jie Chen (UCLA)
Yujia Shen
Arthur Choi
Adnan Darwiche

We consider the problem of learning Bayesian networks optimally, when subject to background knowledge in the form of ancestral constraints. Our approach is based on a recently proposed framework for optimal structure learning based on non-decomposable scores, which is general enough to accommodate ancestral constraints. The proposed framework exploits oracles for learning structures using decomposable scores, which cannot accommodate ancestral constraints since they are non-decomposable. We show how to empower these oracles by passing them decomposable constraints that they can handle, which are inferred from ancestral constraints that they cannot handle. Empirically, we demonstrate that our approach can be orders-of-magnitude more efficient than alternative frameworks, such as those based on integer linear programming.

#43 Solving Random Systems of Quadratic Equations via Truncated Generalized Gradient Flow

Gang Wang (Univ. of Minnesota)
Georgios Giannakis (Univ. of Minnesota)

This paper puts forth a novel algorithm, termed {truncated generalized gradient flow} (TGGF), to solve for \bm{x}\in$R$^n/\mathbb{C}^n a system of m quadratic equations y_i=|\langle\bm{a}_i,\bm{x}>|^2, i=1,2,\ldots,m, which even for \left\{\bm{a}_i\in$R$^n/\mathbb{C}^n\right\}_{i=1}^m random is known to be {NP-hard} in general. We prove that as soon as the number of equations m is on the order of the number of unknowns n, TGGF recovers the solution exactly (up to a global unimodular constant) with high probability and complexity growing linearly with the time required to read the data \left\{\left(\bm{a}_i;\,y_i\right)\right\}_{i=1}^m. Specifically, TGGF proceeds in two stages: s1) A novel {orthogonality-promoting} initialization that is obtained with simple power iterations; and, s2) a refinement of the initial estimate by successive updates of scalable {truncated generalized gradient iterations}. The former is in sharp contrast to the existing spectral initializations, while the latter handles the rather challenging nonconvex and nonsmooth {amplitude-based} cost function. Numerical tests demonstrate that: i) The novel orthogonality-promoting initialization method returns more accurate and robust estimates relative to its spectral counterparts; and ii) even with the same initialization, our refinement/truncation outperforms Wirtinger-based alternatives, all corroborating the superior performance of TGGF over state-of-the-art algorithms.

#44 Balancing Suspense and Surprise: Timely Decision Making with Endogenous Information Acquisition

Ahmed M. Alaa Ibrahim (UCLA)
Mihaela Van Der Schaar

We develop a Bayesian model for decision-making under time pressure with endogenous information acquisition. In our model, the decision-maker decides when to observe (costly) information by sampling an underlying continuous-time stochastic process (time series) that conveys information about the potential occurrence/non-occurrence of an adverse event which will terminate the decision-making process. In her attempt to predict the occurrence of the adverse event, the decision-maker follows a policy that determines when to acquire information from the time series (continuation), and when to stop acquiring information and make a final prediction (stopping). We show that the optimal policy has a "rendezvous" structure, i.e. a structure in which whenever a new information sample is gathered from the time series, the optimal "date" for acquiring the next sample becomes computable. The optimal interval between two information samples balances a trade-off between the decision maker's "surprise", i.e. the drift in her posterior belief after observing new information, and "suspense", i.e. the probability that the adverse event occurs in the time interval between two information samples. Moreover, we characterize the continuation and stopping regions in the decision-maker's state-space, and show that they depend not only on the decision-maker's beliefs, but also on the "context", i.e. the current realization of the time series.

#45 Blind Optimal Recovery of Signals

Dmitry Ostrovsky (Univ. Grenoble Alpes)
Zaid Harchaoui (NYU)
Anatoli Juditsky
Arkadi S Nemirovski (Gerogia Institute of Technology)

We consider the problem of recovering a signal observed in Gaussian noise. If the set of signals is convex and compact, and can be specified

beforehand, one can use classical linear estimators that achieve a risk within a constant factor of the minimax risk. However, when the set is unspecified, designing an estimator that is blind to the hidden structure of the signal remains a challenging problem. We propose a new family of estimators to recover signals observed in Gaussian noise. Instead of specifying the set where the signal lives, we assume the existence of a well-performing linear estimator. Proposed estimators enjoy exact oracle inequalities and can be efficiently computed through convex optimization. We present several numerical illustrations that show the potential of the approach.

#46 Spatiotemporal Residual Networks for Video Action Recognition

Christoph Feichtenhofer (Graz Univ. of Technology)
Axel Pinz (Graz Univ. of Technology)
Richard Wildes (York Univ. Toronto)

Two-stream Convolutional Networks (ConvNets) have shown strong performance for human action recognition in videos. Recently, Residual Networks (ResNets) have arisen as a new technique to train extremely deep architectures. In this paper, we introduce spatiotemporal ResNets as a combination of these two approaches. Our novel architecture generalizes ResNets for the spatiotemporal domain by introducing residual connections in two ways. First, we inject residual connections between the appearance and motion pathways of a two-stream architecture to allow spatiotemporal interaction between the two streams. Second, we transform pretrained image ConvNets into spatiotemporal networks by equipping these with learnable convolutional filters that are initialized as temporal residual connections and operate on adjacent feature maps in time. This approach slowly increases the spatiotemporal receptive field as the depth of the model increases and naturally integrates image ConvNet design principles. The whole model is trained end-to-end to allow hierarchical learning of complex spatiotemporal features. We evaluate our novel spatiotemporal ResNet using two widely used action recognition benchmarks where it exceeds the previous state-of-the-art.

#47 CMA-ES with Optimal Covariance Update and Storage Complexity

Oswin Krause
Dídac Rodríguez Arbonès (Univ. of Copenhagen)
Christian Igel

The covariance matrix adaptation evolution strategy (CMA-ES) is arguably one of the most powerful real-valued derivative-free optimization algorithms, finding many applications in machine learning. The CMA-ES is a Monte Carlo method, sampling from a sequence of multi-variate Gaussian distributions. Given the function values at the sampled points, updating and storing the covariance matrix dominates the time and space complexity in each iteration of the algorithm. We propose a numerically stable quadratic-time covariance matrix update scheme with minimal memory requirements based on maintaining triangular Cholesky factors. This requires a modification of the cumulative step-size adaption (CSA) mechanism in the CMA-ES, in which we replace the inverse of the square root of the covariance matrix by the inverse of the triangular Cholesky factor. Because the triangular Cholesky factor changes smoothly with the matrix square root, this modification does not change the behavior of the CMA-ES in terms of required objective function evaluations as verified empirically. Thus, the described algorithm can and should replace the standard CMA-ES if updating and storing the covariance matrix matters.

#48 An End-to-End Approach for Natural Language to IFTTT Program Translation

Chang Liu (Univ. of Maryland)
Xinyun Chen (Shanghai Jiaotong Univ.)
Richard Shin
Mingcheng Chen (Univ. of Illinois)
Dawn Song (UC Berkeley)

Recently, there is an increasing interest in automatically translating natural language describing "if-then" rule into executable programs. Existing works tackle this problem using sophisticated semantic parsing techniques. In this work, we seek answers to the question whether deep learning techniques can reduce the level of human efforts while achieving the same goal. In this work, we show that a bi-directional LSTM model, which can be trained end-to-end, can outperform handcrafted state-of-the-art [14] 15 points on accuracy, and reduce the error rate by a half. We also design a new feedforward network model, called Hierarchical Attention Model, which can also outperform [14] by 10 points, be trained end-to-end, and is easier to be trained than recurrent networks. We demonstrate that Hierarchical Attention Model can yield a much better performance than RNN on a one shot learning setting of our problem, which has its own application merits.

#49 The Sound of APALM Clapping: Faster Nonsmooth Nonconvex Optimization with Stochastic Asynchronous PALM

damekdavis Davis (Cornell Univ.)
Brent Edmunds (Univ. of California)
Madeleine Udell

We introduce the Stochastic Asynchronous Proximal Alternating Linearized Minimization (SAPALM) method, a block coordinate stochastic proximal-gradient method for solving nonconvex, nonsmooth optimization problems. SAPALM is the first asynchronous parallel optimization method that provably converges on a large class of nonconvex, nonsmooth problems. We prove that SAPALM matches the best known rates of convergence --- among synchronous or asynchronous methods --- on this problem class. We provide upper bounds on the number of workers for which we can expect to see a linear speedup, which match the best bounds known for less complex problems, and show that in practice SAPALM achieves this linear speedup. We demonstrate state-of-the-art performance on several matrix factorization problems.

#50 Efficient Algorithm for Streaming Submodular Cover

Ashkan Norouzi-Fard (EPFL)
Abbas Bazzi (EPFL)
Ilija Bogunovic (EPFL Lausanne)
Marwa El Halabi (I)
Ya-Ping Hsieh
Volkan Cevher

We initiate the study of the classical Submodular Cover (SC) problem in the data streaming model, that we refer to as the Streaming Submodular Cover (SSC) problem. We show that any single pass streaming algorithm that uses sublinear memory in the size of the stream, is bound to fail in providing any non-trivial approximation guarantees for SSC. To cope with that, we consider the relaxed version of SSC, where we only seek to find a partial cover. We design the first Efficient bicriteria Submodular Cover Streaming (ESC-streaming) algorithm for this problem, and provide theoretical guarantees for its performance that match our experimental results. Our algorithm finds solutions that are competitive with the near-optimal offline greedy algorithm despite requiring only a single pass over the data stream. In our experiments, we evaluate the performance of ESC-streaming on active set selection and large-scale graph cover problems.

#51 Attend, Infer, Repeat: Fast Scene Understanding with Generative Models

Ali Eslami (DeepMind)
Nicolas Heess
Theophane Weber
Yuval Tassa (DeepMind)
David Szepesvari (DeepMind)
koray kavukcuoglu (DeepMind)
Geoffrey E Hinton (Google)

We present a framework for efficient inference in structured image models that explicitly reason about objects. We achieve this by performing probabilistic inference using a recurrent neural network that attends to scene elements and processes them one at a time. Crucially, the model itself learns to choose the appropriate number of inference steps. We use this scheme to learn to perform inference in partially specified 2D models (variable-sized variational auto-encoders) and fully specified 3D models (probabilistic renderers). We show that such models learn to identify multiple objects - counting, locating and classifying the elements of a scene - without any supervision, e.g., decomposing 3D images with various numbers of objects in a single forward pass of a neural network at unprecedented speed. We further show that the networks produce accurate inferences when compared to supervised counterparts, and that their structure leads to improved generalization.

#52 An ensemble diversity approach to supervised binary hashing

Miguel A. Carreira-Perpinan (UC Merced)
Ramin Raziperchikolaei (UC Merced)

Binary hashing is a well-known approach for fast approximate nearest-neighbor search in information retrieval. Much work has focused on affinity-based objective functions involving the hash functions or binary codes. These objective functions encode neighborhood information between data points and are often inspired by manifold learning algorithms. They ensure that the hash functions differ from each other through constraints or penalty terms that encourage codes to be orthogonal or dissimilar across bits, but this couples the binary variables and complicates the already difficult optimization. We propose a much simpler approach: we train each hash function (or bit) independently from each other, but introduce diversity among them using techniques from classifier ensembles. Surprisingly, we find that not only is this faster and trivially parallelizable, but it also improves over the more complex, coupled objective function, and achieves state-of-the-art precision and recall in experiments with image retrieval.

#53 End-to-End Goal-Driven Web Navigation

Rodrigo Nogueira (New York Univ.)
Kyunghyun Cho (Univ. of Montreal)

We propose a goal-driven web navigation as a benchmark task for evaluating an agent with abilities to understand natural language and plan on partially observed environments. In this challenging task, an agent navigates through a website, which is represented as a graph consisting of web pages as nodes and hyperlinks as directed edges, to find a web page in which a query appears. The agent is required to have sophisticated high-level reasoning based on natural languages and efficient sequential decision-making capability to succeed. We release a software tool, called WebNav, that automatically transforms a website into this goal-driven web navigation task, and as an example, we make WikiNav, a dataset constructed from the English Wikipedia. We extensively evaluate different variants of neural net based artificial agents on WikiNav and observe that the proposed goal-driven web navigation well reflects the advances in models, making it a suitable benchmark for evaluating future progress. Furthermore, we extend the WikiNav with question-answer pairs from Jeopardy! and test the proposed agent based on recurrent neural networks against strong inverted index based search engines. The artificial agents trained on WikiNav outperforms the engined based approaches, demonstrating the capability of the proposed goal-driven navigation as a good proxy for measuring the progress in real-world tasks such as focused crawling and question-answering.

#54 The Power of Adaptivity in Identifying Statistical Alternatives

Kevin Jamieson (UC Berkeley)
Daniel Haas
Benjamin Recht

This paper studies the trade-off between two different kinds of pure exploration: breadth versus depth. We focus on the most biased coin problem, asking how many total coin flips are required to identify a ``heavy" coin from an infinite bag containing both ``heavy" coins with mean \theta_1 \in (0,1), and ``light" coins with mean \theta_0 \in (0,\theta_1), where heavy coins are drawn from the bag with proportion α \in (0,1/2). When α,\theta_0,\theta_1 are unknown, the key difficulty of this problem lies in distinguishing whether the two kinds of coins have very similar means, or whether heavy coins are just extremely rare. While existing solutions to this problem require some prior knowledge of the parameters \theta_0,\theta_1,α, we propose an adaptive algorithm that requires no such knowledge yet still obtains near-optimal sample complexity guarantees. In contrast, we provide a lower bound showing that non-adaptive strategies require at least quadratically more samples. In characterizing this gap between adaptive and nonadaptive strategies, we make connections to anomaly detection and prove lower bounds on the sample complexity of differentiating between a single parametric distribution and a mixture of two such distributions.

#55 A Probabilistic Framework for Deep Learning

Ankit B Patel (Baylor College of Medicine and Rice Univ.)
Minh Tan Nguyen (Rice Univ.)
Richard Baraniuk

We develop a probabilistic framework for deep learning based on the Deep Rendering Model (DRM), a generative probabilistic model that explicitly captures variations in the data due to latent nuisance variables. We demonstrate that max-sum message passing in the DRM corresponds directly to the operations in deep convolutional neural networks (DCNs). Our framework provides new insights into the success and shortcomings of DCNs as well as a principled route to their improvement. DRM training via the Expectation-Maximization (EM) algorithm is a powerful alternative to DCN back-propagation, and initial training results are promising. DRM-based classification outperforms DCNs in supervised digit classification, training 2-3 times faster and achieving better accuracy(1.21% vs 1.30%), and they show comparable results to prior art in semi-supervised and unsupervised learning tasks (with no hyper-parameter tuning nor regularization). In sum, our theoretical and training results demystify the structure of DCNs and support a unified approach to supervised, unsupervised, and semi-supervised learning.

#56 Minimax Estimation of Maximum Mean Discrepancy with Radial Kernels

Ilya Tolstikhin
Bharath K. Sriperumbudur
Prof. Bernhard Schölkopf

Maximum Mean Discrepancy (MMD) is a distance on the space of probability measures which has found numerous applications in machine learning and nonparametric testing. This distance is based on the notion of embedding probabilities in a reproducing kernel Hilbert space. In this paper, we present the first known lower bounds for the estimation of MMD based on finite samples. Our lower bounds hold for any radial universal kernel on \R^d and match the existing upper bounds up to constants that depend only on the properties of the kernel. Using these lower bounds, we establish the minimax rate optimality of the empirical estimator and its U-statistic variant, which are usually employed in applications.

#57 Adaptive Neural Compilation

Rudy R Bunel (Oxford Univ.)
Alban Desmaison (Oxford)
Pawan K Mudigonda (Univ. of Oxford)
Pushmeet Kohli
Philip Torr

This paper proposes an adaptive neural-compilation framework to address the problem of efficient program learning. Traditional code optimisation strategies used in compilers are based on applying pre-specified set of transformations that make the code faster to execute without changing its semantics. In contrast, our work involves adapting programs to make them more efficient while considering correctness only on a target input distribution. Our work is inspired from differentiable representations of programs. We show that it is possible to compile programs written in a low-level language to a differentiable representation. We also show how programs in this representation can be optimised to make them efficient on a target distribution of inputs. Experimental results demonstrate that our approach enables learning specifically-tuned algorithms for given data distributions with a high success rate.

#58 Tagger: Deep Unsupervised Perceptual Grouping

Klaus Greff (IDSIA)
Antti Rasmus (The Curious AI Company)
Mathias Berglund (The Curious AI Company)
Tele Hotloo Hao (The Curious AI Company)
Harri Valpola (The Curious AI Company)

We present a framework for efficient perceptual inference that explicitly reasons about segmentation of its inputs and features. By enriching the representations of a neural network we enable it to group the representations of different objects in an iterative manner. Crucially we are not training the network for a specific segmentation, but it learns the grouping process in a unsupervised way or alongside any supervised task. By letting the system amortize the iterative inference of the groupings, we achieve very fast convergence. In contrast to other recently proposed methods to deal with multi-object scenes our system does not assume the inputs to be images and can therefore directly deal with other modalities. We use this system for multi-digit classification in very cluttered images that require texture segmentation. We find an improvement for classification performance over ConvNets even though our network is fully connected. Furthermore we observe that our system performs much better than the baseline Ladder model on semi- supervised learning on our dataset, indicating that segmentation can also improve the sample-efficiency.

#59 A scaled Bregman theorem with applications

Richard Nock (Data61 and ANU)
Aditya Menon
Cheng Soon Ong (Data61)

Bregman divergences play a central role in the design and analysis of a range of machine learning algorithms. This paper explores the use of Bregman divergences to establish reductions between such algorithms and their analyses. We present a new scaled isodistortion theorem involving Bregman divergences (scaled Bregman theorem for short) which shows that certain "Bregman distortions" (employing a potentially non-convex generator) may be exactly re-written as a scaled Bregman divergence computed over transformed data. Admissible distortions include geodesic distances on curved manifolds and projections or gauge-normalisation, while admissible data include scalars, vectors and matrices. Our theorem allows one to leverage to the wealth and convenience of Bregman divergences when analysing algorithms relying on the aforementioned Bregman distortions. We illustrate this with three novel applications of our theorem: a reduction from multi-class density ratio to class-probability estimation, a new adaptive projection free yet norm-enforcing dual norm mirror descent algorithm, and a reduction from clustering on flat manifolds to clustering on curved manifolds. Experiments on each of these domains validate the analyses and suggest that the scaled Bregman theorem might be a worthy addition to the popular handful of Bregman divergence properties that have been pervasive in machine learning.

#60 Learning feed-forward one-shot learners

Luca Bertinetto (Univ. of Oxford)
João F. Henriques (Univ. of Oxford)
Jack Valmadre (Univ. of Oxford)
Philip Torr
Andrea Vedaldi

One-shot learning is usually tackled by using generative models or discriminative embeddings. Discriminative methods based on deep learning, which are very effective in other learning scenarios, are ill-suited for one-shot learning as they need large amounts of training data. In this paper, we propose a method to learn the parameters of a deep model in one shot. We construct the learner as a second deep network, called a learnet, which predicts the parameters of a pupil network from a single exemplar. In this manner we obtain an efficient feed-forward one-shot learner, trained end-to-end by minimizing a one-shot classification objective in a learning to learn formulation. In order to make the construction feasible, we propose a number of factorizations of the parameters of the pupil network. We demonstrate encouraging results by learning characters from single exemplars in Omniglot, and by tracking visual objects from a single initial exemplar in the Visual Object Tracking benchmark.

#61 Error Analysis of Generalized Nyström Kernel Regression

Hong Chen (Univ. of Texas)
Haifeng Xia (Huazhong Agricultural Univ.)
Heng Huang (Univ. of Texas Arlington)

Nystr\"{o}m method has been used successfully to improve the computational efficiency of kernel ridge regression (KRR). Recently, theoretical analysis of Nystr\"{o}m KRR, including generalization bound and convergence rate, has been established based on reproducing kernel Hilbert space (RKHS) associated with the symmetric positive semi-definite kernel. However, in real world applications, RKHS is not always optimal and kernel function is not necessary to be symmetric or positive semi-definite. In this paper, we consider the generalized Nystr\"{o}m kernel regression (GNKR) with ℓ_2 coefficient regularization, where the kernel just requires the continuity and boundedness. Error analysis is provided to characterize its generalization performance and the column norm sampling is introduced to construct the refined hypothesis space. In particular, the fast learning rate with polynomial decay is reached for the GNKR. Experimental analysis demonstrates the satisfactory performance of GNKR with the column norm sampling.

#62 Breaking the Bandwidth Barrier: Geometrical Adaptive Entropy Estimation

Weihao Gao (UIUC)
Sewoong Oh
Pramod Viswanath (UIUC)

Estimators of information theoretic measures such as entropy and mutual information from samples are a basic workhorse for many downstream applications in modern data science. State of the art approaches have been either geometric (nearest neighbor (NN) based) or kernel based (with bandwidth chosen to be data independent and vanishing sub linearly in the sample size). In this paper we combine both these approaches to design new estimators of entropy and mutual information that strongly outperform all state of the art methods. Our estimator uses bandwidth choice of fixed k-NN distances; such a choice is both data dependent and linearly vanishing in the sample size and necessitates a bias cancellation term that is universal and independent of the underlying distribution. As a byproduct, we obtain a unified way of obtaining both kernel and NN estimators. The corresponding theoretical contribution relating the geometry of NN distances to asymptotic order statistics is of independent mathematical interest.

#63 Asynchronous Parallel Greedy Coordinate Descent

Yang You (UC Berkeley)
Xiangru Lian (Univ. of Rochester)
Ji Liu
Hsiang-Fu (Rofu) Yu (Univ. of Texas at Austin)
Inderjit S Dhillon
James Demmel (UC Berkeley)
Cho-Jui Hsieh

In this paper, we propose and study an Asynchronous parallel Greedy Coordinat Descent (Asy-GCD) algorithm for minimizing a smooth function with bounded constraints. At each iteration, workers asynchronously conduct greedy coordinate descent updates on a block of variables. In the first part of the paper, we analyze the theoretical behavior of Asy-GCD and prove a linear convergence rate. In the second part, we develop an efficient kernel SVM solver based on Asy-GCD in the shared memory multi-core setting. Since our algorithm is fully asynchronous—each core does not need to idle and wait for the other cores—the resulting algorithm enjoys good speedup and outperforms existing multi-core kernel SVM solvers including asynchronous stochastic coordinate descent and multi-core LIBSVM.

#64 Structured Prediction Theory Based on Factor Graph Complexity

Corinna Cortes
Vitaly Kuznetsov (Courant Institute)
Mehryar Mohri
Scott Yang (New York Univ.)

We present a general theoretical analysis of structured prediction. By introducing a new complexity measure that explicitly factors in the structure of the output space and the loss function, we are able to derive new data-dependent learning guarantees for a broad family of losses and for hypothesis sets with an arbitrary factor graph decomposition. We extend this theory by leveraging the principle of Voted Risk Minimization (VRM) and showing that learning is possible with complex factor graphs. We both present new learning bounds in this advanced setting as well as derive two new families of algorithms, Voted Conditional Random Fields and Voted Structured Boosting, which can make use of very complex features and factor graphs without overfitting. Finally, we also validate our theory through experiments on several datasets.

#65 Parameter Learning for Log-supermodular Distributions

Tatiana Shpakova (Inria - ENS Paris)
Francis Bach

We consider log-supermodular models on binary variables, which are probabilistic models with negative log-densities which are submodular. These models provide probabilistic interpretations of common combinatorial optimization tasks such as image segmentation. In this paper, we focus primarily on parameter estimation in the models from known upper-bounds on the intractable log-partition function. We show that the bound based on separable optimization on the base polytope of the submodular function is always inferior to a bound based on ``perturb-and-MAP'' ideas. Then, to learn parameters, given that our approximation of the log-partition function is an expectation (over our own randomization), we use a stochastic subgradient technique to maximize a lower-bound on the log-likelihood. This can also be extended to conditional maximum likelihood. We illustrate our new results in a set of experiments in binary image denoising, where we highlight the flexibility of a probabilistic model to learn with missing data.

#66 Exact Recovery of Hard Thresholding Pursuit

Xiaotong Yuan (Nanjing Univ. of Informat)
Ping Li
Tong Zhang

The Hard Thresholding Pursuit (HTP) is a class of truncated greedy descent methods for finding sparse solutions of ℓ_0-constrained loss minimization problems. The HTP-style methods have been shown to have strong approximation guarantee and impressive numerical performance in high dimensional statistical learning applications. However, the current theoretical treatment of these methods has traditionally been restricted to the analysis of parameter estimation consistency. It remains an open problem to analyze the support recovery performance (a.k.a.,sparsistency) of this type of methods for recovering the global minimizer of the original NP-hard problem. In this paper, we bridge this gap by showing, for the first time, that exact recovery of the global sparse minimizer is possible for HTP-style methods under proper restricted strong condition number conditions. We further show that HTP-style methods are able to recover the support of certain relaxed sparse solutions with arbitrary restricted strong condition number. Numerical results on simulated data confirms our theoretical predictions.

#67 A New Liftable Class for First-Order Probabilistic Inference

Seyed Mehran Kazemi (UBC)
Angelika Kimmig (KU Leuven)
Guy Van den Broeck
David Poole (UBC)

Statistical relational models provide compact encodings of probabilistic dependencies in relational domains, but result in highly intractable graphical models. The goal of lifted inference is to carry out probabilistic inference without needing to reason about each individual separately, by instead treating exchangeable, undistinguished objects as a whole. In this paper, we study the domain recursion inference rule, which, despite its central rule in early theoretical results on domain-lifted inference, has later been believed redundant. We show that this rule is more powerful than expected, and in fact significantly extends the range of models for which lifted inference runs in time polynomial in the number of individuals in the domain. This includes an open problem called S4, the symmetric transitivity model, and a first-order logic encoding of the birthday paradox. We further identify the new class FO2+ of domain-liftable theories, which subsumes both FO2 and recursively unary theories, the only classes of domain-liftable theories known so far, and show that using domain recursion can achieve exponential speedup even in theories that cannot fully be lifted with the extended set of inference rules.

#68 Variational Inference in Mixed Probabilistic Submodular Models

Josip Djolonga (ETH Zurich)
Sebastian Tschiatschek (ETH Zurich)
Andreas Krause

We consider the problem of variational inference in probabilistic models with both log-submodular and log-supermodular higher-order potentials. These models can represent arbitrary distributions over binary variables and thus generalize the commonly used pairwise Markov random fields and models with log-supermodular potentials only. While approximate inference methods for models with purely log-submodular or log-supermodular potentials exist, we provide the first efficient inference schemes that can handle models with both types of potentials by exploiting recent advances in the field of discrete optimization. We demonstrate the effectiveness of our approach in a large set of experiments, where our model allows reasoning about preferences for sets of items with complements and substitutes.

#69 Unifying Count-Based Exploration and Intrinsic Motivation

Marc Bellemare (DeepMind)
Sriram Srinivasan
Georg Ostrovski (DeepMind)
Tom Schaul
David Saxton (DeepMind)
Remi Munos (DeepMind)

We consider an agent's uncertainty about its environment and the problem of generalizing this uncertainty across observations. Specifically, we focus on the problem of exploration in non-tabular reinforcement learning. Drawing inspiration from the intrinsic motivation literature, we use sequential density models to measure uncertainty, and propose a novel algorithm for deriving a pseudo-count from an arbitrary sequential density model. This technique enables us to generalize count-based exploration algorithms to the non-tabular case. We apply our techniques to Atari 2600 games, providing sensible pseudo-counts from raw pixels. We derive exploration bonuses from these pseudo-counts and obtain significantly improved exploration in a number of hard games, including the infamously difficult Montezuma's Revenge.

#70 Approximate maximum entropy principles via Goemans-Williamson with applications to provable variational methods

Andrej Risteski (Princeton Univ.)
Yuanzhi Li (Princeton Univ.)

The well known maximum-entropy principle due to Jaynes, which states that given mean parameters, the maximum entropy distribution matching them is in an exponential family has been very popular in machine learning due to its "Occam's razor" interpretation. Unfortunately, calculating the potentials in the maximum entropy distribution is intractable [BGS14]. We provide computationally efficient versions of this principle when the mean parameters are pairwise moments: we design distributions that approximately match given pairwise moments, while having entropy which is comparable to the maximum entropy distribution matching those moments. We additionally provide surprising applications of the approximate maximum entropy principle to designing provable variational methods for partition function calculations for Ising models without any assumptions on the potentials of the model. More precisely, we show that we can get approximation guarantees for the log-partition function comparable to those in the low-temperature limit, which is the setting of optimization of quadratic forms over the hypercube. ([AN06])

#71 A Multi-step Inertial Forward-Backward Splitting Method for Non-convex Optimization

Jingwei Liang (GREYC)
Jalal Fadili
Gabriel Peyré

In this paper, we propose a multi-step inertial Forward--Backward splitting algorithm for minimizing the sum of two non-necessarily convex functions, one of which is proper lower semi-continuous while the other is differentiable with a Lipschitz continuous gradient. We first prove global convergence of the scheme with the help of the Kurdyka–Łojasiewicz property. Then, when the non-smooth part is also partly smooth relative to a smooth submanifold, we establish finite identification of the latter and provide sharp local linear convergence analysis. The proposed method is illustrated on a few problems arising from statistics and machine learning.

#72 Fast and Flexible Monotonic Functions with Ensembles of Lattices

Mahdi Milani Fard
Kevin Canini
Andy Cotter
Jan Pfeifer (Google)
Maya Gupta

For many machine learning problems, there are some inputs that are known to be positively (or negatively) related to the output, and in such cases training the model to respect that monotonic relationship can provide regularization, and makes the model more interpretable. However, flexible monotonic functions are computationally challenging to learn beyond a few features. We break through this barrier by learning ensembles of monotonic calibrated interpolated look-up tables (lattices). A key contribution is an automated algorithm for selecting feature subsets for the ensemble base models. We demonstrate that compared to random forests, these ensembles produce similar or better accuracy, while providing guaranteed monotonicity consistent with prior knowledge, smaller model size and faster evaluation.

#73 Architectural Complexity Measures of Recurrent Neural Networks

Saizheng Zhang (Univ. of Montreal)
Yuhuai Wu (Univ. of Toronto)
Tong Che (IHES)
Zhouhan Lin (Univ. of Montreal)
Roland Memisevic (Univ. of Montreal)
Russ Salakhutdinov (Univ. of Toronto)
Yoshua Bengio (U. Montreal)

In this paper, we systematically analyse the connecting architectures of recurrent neural networks (RNNs). Our main contribution is twofold: first, we present a rigorous graph-theoretic framework describing the connecting architectures of RNNs in general. Second, we propose three architecture complexity measures of RNNs: (a) the recurrent depth, which captures the RNN's over-time nonlinear complexity, (b) the feedforward depth, which captures the local input-output nonlinearity (similar to the "depth" in feedforward neural networks (FNNs)), and (c) the recurrent skip coefficient which captures how rapidly the information propagates over time. Our experimental results show that RNNs might benefit from larger recurrent depth and feedforward depth. We further demonstrate that increas- ing recurrent skip coefficient offers performance boosts on long term dependency problems.

#74 Online Convex Optimization with Unconstrained Domains and Losses

Ashok Cutkosky (Stanford Univ.)
Kwabena A Boahen (Stanford Univ.)

We propose an online convex optimization algorithm (RescaledExp) that achieves optimal regret in the unconstrained setting without prior knowledge of any bounds on the loss functions. We prove a lower bound showing a strong separation between the regret of existing algorithms that require a known bound on the loss functions and any algorithm that does not require such knowledge. RescaledExp matches this lower bound asymptotically in the number of iterations. RescaledExp is naturally hyperparameter-free and we demonstrate empirically that it matches prior optimization algorithms with hyperparameter optimization.

#75 Split LBI: An Iterative Regularization Path with Structural Sparsity

Chendi Huang (Peking Univ.)
Xinwei Sun
Jiechao Xiong (Peking Univ.)
Yuan Yao

An iterative regularization path with structural sparsity is proposed in this paper based on variable splitting and the Linearized Bregman Iteration, hence called {Split LBI}. Despite its simplicity, Split LBI outperforms the popular generalized Lasso in both theory and experiments. A theory of path consistency is presented that equipped with a proper early stopping, Split LBI may achieve model selection consistency under a family of Irrepresentable Conditions which can be weaker than the necessary and sufficient condition for generalized Lasso. Furthermore, some ℓ_2 error bounds are also given at the minimax optimal rates. The utility and benefit of the algorithm are illustrated by applications on both traditional image denoising and a novel example on partial order ranking.

#76 Variational Autoencoder for Deep Learning of Images, Labels and Captions

Yunchen Pu (Duke Univ.)
Zhe Gan (Duke)
Ricardo Henao
Xin Yuan (Bell Labs)
Chunyuan Li (Duke)
Andrew Stevens (Duke Univ.)
Lawrence Carin

A novel variational autoencoder is developed to model images, as well as associated labels or captions. The Deep Generative Deconvolutional Network (DGDN) is used as a decoder of the latent image features, and a deep Convolutional Neural Network (CNN) is used as an image encoder; the CNN is used to approximate a distribution for the latent DGDN features/code. The latent code is also linked to generative models for labels (Bayesian support vector machine) or captions (recurrent neural network). When predicting a label/caption for a new image at test, averaging is performed across the distribution of latent codes; this is computationally efficient as a consequence of the learned CNN-based encoder. Since the framework is capable of modeling the image in the presence/absence of associated labels/captions, a new semi-supervised setting is manifested for CNN learning with images; the framework even allows unsupervised CNN learning, based on images alone.

#77 Recovery Guarantee of Non-negative Matrix Factorization via Alternating Updates

Yuanzhi Li (Princeton Univ.)
Yingyu Liang
Andrej Risteski (Princeton Univ.)

Non-negative matrix factorization is a popular tool for decomposing data into feature and weight matrices under non-negativity constraints. It enjoys practical success but is poorly understood theoretically. This paper proposes an algorithm that alternates between updating the features and weights, and shows that assuming a generative model of the data, it provably recovers the ground-truth under fairly mild conditions. In particular, its only essential requirement on features is linear independence. Furthermore, the algorithm can tolerate adversarial noise that can potentially be as large as the signal. The analysis of the algorithm relies on a carefully designed coupling between two Lyapunov functions, which we believe is of independent interest.

#78 Proximal Deep Structured Models

Shenlong Wang (Univ. of Toronto)
Sanja Fidler
Raquel Urtasun

Many problems in real-world applications involve predicting continuous-valued random variables that are statistically related. In this paper, we propose a powerful deep structured model that is able to learn complex non-linear functions which encode the dependencies between continuous output variables. We show that inference in our model using proximal methods can be efficiently solved as a feed-foward pass of a special type of deep recurrent neural network. We demonstrate the effectiveness of our approach in the tasks of image denoising, depth refinement and optical flow estimation.

#79 Safe Policy Improvement by Minimizing Robust Baseline Regret

Mohammad Ghavamzadeh
Marek Petrik
Yinlam Chow (Stanford Univ.)

An important problem in sequential decision-making under uncertainty is to use limited data to compute a safe policy, i.e., a policy that is guaranteed to perform at least as well as a given baseline strategy. In this paper, we develop and analyze a new model-based approach to compute a safe policy when we have access to an inaccurate dynamics model of the system with known accuracy guarantees. Our proposed robust method uses this (inaccurate) model to directly minimize the (negative) regret w.r.t. the baseline policy. Contrary to the existing approaches, minimizing the regret allows one to improve the baseline policy in states with accurate dynamics and seamlessly fall back to the baseline policy, otherwise. We show that our formulation is NP-hard and propose an approximate algorithm. Our empirical results on several domains show that even this relatively simple approximate algorithm can significantly outperform standard approaches.

#80 A Pseudo-Bayesian Algorithm for Robust PCA

Tae-Hyun Oh (KAIST)
Yasuyuki Matsushita (Osaka Univ.)
In Kweon (KAIST)
David Wipf

Commonly used in many applications, robust PCA represents an algorithmic attempt to reduce the sensitivity of classical PCA to outliers. The basic idea is to learn a decomposition of some data matrix of interest into low rank and sparse components, the latter representing unwanted outliers. Although the resulting problem is typically NP-hard, convex relaxations provide a computationally-expedient alternative with theoretical support. However, in practical regimes performance guarantees break down and a variety of non-convex alternatives, including Bayesian-inspired models, have been proposed to boost estimation quality. Unfortunately though, without additional a priori knowledge none of these methods can significantly expand the critical operational range such that exact principal subspace recovery is possible. Into this mix we propose a novel pseudo-Bayesian algorithm that explicitly compensates for design weaknesses in many existing non-convex approaches leading to state-of-the-art performance with a sound analytical foundation.

#81 Learning values across many orders of magnitude

Hado van Hasselt
Baguez Aguez
Matteo Hessel (DeepMind)
Volodymyr Mnih
David Silver

Most learning algorithms are not invariant to the scale of the function that is being approximated. We propose to adaptively normalize the targets used in learning. This is useful in value-based reinforcement learning, where the magnitude of appropriate value approximations can change over time when we update the policy of behavior. Our main motivation is prior work on learning to play Atari games, where the rewards were all clipped to a predetermined range. This clipping facilitates learning across many different games with a single learning algorithm, but a clipped reward function can result in qualitatively different behavior. Using the adaptive normalization we can remove this domain-specific heuristic without diminishing overall performance.

#82 Single Pass PCA of Matrix Products

Shanshan Wu (UT Austin)
Srinadh Bhojanapalli (TTI Chicago)
Sujay Sanghavi
Alex Dimakis

In this paper we present a new algorithm for computing a low rank approximation of the product A^TB by taking only a single pass of the two matrices A and B. The straightforward way to do this is to (a) first sketch A and B individually, and then (b) find the top components using PCA on the sketch. Our algorithm in contrast retains additional summary information about A,B (e.g. row and column norms etc.) and subsequently uses these in a subtle way to get an improved approximation from the sketches. Our main analytical result establishes a comparable spectral norm guarantee to existing two-pass methods; in addition we also provide results from an Apache Spark implementation that shows better computational and statistical performance on real-world and synthetic evaluation datasets.

#83 Convolutional Neural Fabrics

Shreyas Saxena (INRIA)
Jakob Verbeek

Despite the success of convolutional neural networks, selecting the optimal architecture for a given task remains an open problem. Instead of aiming to select a single optimal architecture, we propose to use a ``fabric'' model that embeds an exponentially large number of CNN architectures. The fabric consists of a 3D trellis that connects response maps at different layers, scales, and channels with a sparse homogeneous connectivity pattern. The only hyper-parameters of the model (nr.\ of channels and layers) are not critical for performance. While individual CNN architectures can be recovered as paths in the trellis, the trellis can in addition ensemble all embedded architectures together, sharing their weights where their paths overlap. The trellis parameters can be learned using standard methods based on back-propagation, at a cost that scales linearly in the fabric size. We present benchmark results competitive with the state of the art for image classification on MNIST and CIFAR10, and for semantic segmentation on the Part Labels dataset.

#84 Generative Shape Models: Joint Text Recognition and Segmentation with Very Little Training Data

Xinghua Lou (Vicarious FPC Inc)
Ken Kansky
Wolfgang Lehrach
CC Laan
Bhaskara Marthi
D. Phoenix
Dileep George

We demonstrate that a generative model for object shapes can achieve state of the art results on challenging scene text recognition tasks, and with orders of magnitude fewer training images than required for competing discriminative methods. In addition to transcribing text from challenging images, our method performs fine-grained instance segmentation of characters. We show that our model is more robust to both affine transformations and non-affine deformations compared to previous approaches.

#85 Mixed vine copulas as joint models of spike counts and local field potentials

Arno Onken (IIT)
Stefano Panzeri (IIT)

Concurrent measurements of neural activity at multiple scales become increasingly important for studying brain function. However, statistical methods for their concurrent analysis are currently lacking. Here we introduce such techniques in a framework based on vine copulas with mixed margins to construct multivariate stochastic models. These models can describe detailed mixed interactions between discrete variables such as neural spike counts, and continuous variables such as local field potentials. We propose efficient methods for likelihood calculation, inference, sampling and mutual information estimation within this framework. We test our methods on artificial data and demonstrate applicability on mixed data generated by a biologically realistic neural network. Our methods hold promise to considerably improve statistical analysis of neural data recorded simultaneously at different scales.

#86 Optimal Black-Box Reductions Between Optimization Objectives

Zeyuan Allen-Zhu (Princeton Univ.)
Elad Hazan

The diverse world of machine learning applications has given rise to a plethora of algorithms and optimization methods, finely tuned to the specific regression or classification task at hand. We reduce the complexity of algorithm design for machine learning by reductions: we develop reductions that take a method developed for one setting and apply it to the entire spectrum of smoothness and strong-convexity in applications. Furthermore, unlike existing results, our new reductions are {optimal} and more {practical}. We show how these new reductions give rise to new and faster running times on training linear classifiers for various families of loss functions, and conclude with experiments showing their successes also in practice.

#87 Dialog-based Language Learning

Jason E Weston

A long-term goal of machine learning research is to build an intelligent dialog agent. Most research in natural language understanding has focused on learning from fixed training sets of labeled data, with supervision either at the word level (tagging, parsing tasks) or sentence level (question answering, machine translation). This kind of supervision is not realistic of how humans learn, where language is both learned by, and used for, communication. In this work, we study dialog-based language learning, where supervision is given naturally and implicitly in the response of the dialog partner during the conversation. We study this setup in two domains: the bAbI dataset of (Weston et al., 2015) and large-scale question answering from (Dodge et al., 2015). We evaluate a set of baseline learning strategies on these tasks, and show that a novel model incorporating predictive lookahead is a promising approach for learning from a teacher's response. In particular, a surprising result is that it can learn to answer questions correctly without any reward-based supervision at all.

#88 Online Bayesian Moment Matching for Topic Modeling with Unknown Number of Topics

Wei-Shou Hsu (Univ. of Waterloo)
Pascal Poupart

Latent Dirichlet Allocation (LDA) is a very popular model for topic modeling as well as many other problems with latent groups. It is both simple and effective. When the number of topics (or latent groups) is unknown, the Hierarchical Dirichlet Process (HDP) provides an elegant non-parametric extension; however, it is a complex model and it is difficult to incorporate prior knowledge since the distribution over topics is implicit. We propose two new models that extend LDA in a simple and intuitive fashion by directly expressing a distribution over the number of topics. We also propose a new online Bayesian moment matching technique to learn the parameters and the number of topics of those models based on streaming data. The approach achieves higher log-likelihood than batch and online HDP on several corpora.

#89 A Sparse Interactive Model for Matrix Completion with Side Information

Jin Lu (Univ. of Connecticut)
Guannan Liang (Univ. of Connecticut)
Jiangwen Sun (Univ. of Connecticut)
Jinbo Bi (Univ. of Connecticut)

Matrix completion methods can benefit from side information besides the partially observed matrix. The use of side features describing the row and column entities of a matrix has been shown to reduce the sample complexity for completing the matrix. We propose a novel sparse formulation that explicitly models the interaction between the row and column side features to approximate the matrix entries. Unlike early methods, this model does not require the low-rank condition on the model parameter matrix. We prove that when the side features can span the latent feature space of the matrix to be recovered, the number of observed entries needed for an exact recovery is O(log N) where N is the size of the matrix. When the side features are corrupted latent features of the matrix with a small perturbation, our method can achieve an ϵ-recovery with O(log N) sample complexity, and maintains a Ω(N^{3/2}) rate similar to classfic methods with no side information. An efficient linearized Lagrangian algorithm is developed with a strong guarantee of convergence. Empirical results show that our approach outperforms three state-of-the-art methods both in simulations and on real world datasets.

#90 Truncated Variance Reduction: A Unified Approach to Bayesian Optimization and Level-Set Estimation

Ilija Bogunovic (EPFL Lausanne)
Jonathan Scarlett
Andreas Krause
Volkan Cevher

We present a new algorithm, truncated variance reduction (TruVaR), that addresses Bayesian optimization (BO) and level-set estimation (LSE) with Gaussian processes in a unified fashion. The algorithm greedily shrinks the total variance, up to a truncation threshold, within a set of potential maximizers (BO) or unclassified points (LSE), which is updated based on confidence bounds. TruVaR is effective in several important settings that are typically non-trivial to incorporate into existing algorithms, including cases with pointwise costs, heteroscedastic noise, and multi-task settings. We provide a general theoretical guarantee for TruVaR covering the former two of these, and use it to recover and strengthen existing results. Moreover, we provide a new result for a setting where one can select from a number of noise levels having associated costs. We demonstrate the effectiveness of the algorithm on both synthetic and real-world data sets.

#91 On Mixtures of Markov Chains

Rishi Gupta (Stanford)
Ravi Kumar
Sergei Vassilvitskii (Google)

We study the problem of reconstructing a mixture of Markov chains from the trajectories generated by random walks through the state space. Under mild non-degeneracy conditions, we show that we can uniquely reconstruct the underlying chains by only considering trajectories of length three, which represent triples of states. Our algorithm is spectral in nature, and is easy to implement. Experimental results show it outperforming the natural EM method for this problem.

#92 High Dimensional Structured Superposition Models

Qilong Gu (Univ. of Minnesota)
Arindam Banerjee

High dimensional superposition models characterize observations using parameters which can be written as a sum of multiple component parameters, each with its own structure, e.g., sum of low rank and sparse matrices. In this paper, we consider general superposition models which allow sum of any number of component parameters, and each component structure can be characterized by any norm. We present a simple estimator for such models, give a geometric condition under which the components can be accurately estimated, characterize sample complexity of the estimator, and give non-asymptotic bounds on the componentwise estimation error. We use tools from empirical processes and generic chaining for the statistical analysis, and our results, which substantially generalize prior work on superposition models, are in terms of Gaussian widths of suitable spherical caps.

#93 Finite Sample Prediction and Recovery Bounds for Ordinal Embedding

Lalit Jain (Univ. of Michigan)
Kevin Jamieson (UC Berkeley)
Rob Nowak (Univ. of Wisconsin Madison)

The goal of ordinal embedding is to represent items as points in a low-dimensional Euclidean space given a set of constraints in form of distance comparisons like ``item i is closer to item j than item k''. Ordinal constraints like this often come from human judgments. To account for errors and variation in judgments, we consider the noisy situation in which the given constraints are independently corrupted by reversing the correct constraint with some probability. This paper makes several new contributions to this problem. First, we derive prediction error bounds for ordinal embedding with noise by exploiting the fact that the rank of a distance matrix of points in \ R^d is at most d+2. These bounds characterize how well a learned embedding predicts new comparative judgments. Second, we investigate the special case of a known noise model and study the Maximum Likelihood estimator. Third, knowledge of the noise model enables us to relate prediction errors to embedding accuracy. This relationship is highly non-trivial since we show that the linear map corresponding to distance comparisons is non-invertible, but there exists a nonlinear map that is invertible. Fourth, two new algorithms for ordinal embedding are proposed and evaluated in experiments.

#94 What Makes Objects Similar: A Unified Multi-Metric Learning Approach

Han-Jia Ye
De-Chuan Zhan
Xue-Min Si (Nanjing Univ.)
Yuan Jiang (Nanjing Univ.)
Zhi-Hua Zhou

Linkages are essential representations of similarity measuring and can be considered from spatial or semantic perspectives. Spatial linkages are explicitly generated from heterogenous data in localities. While semantic linkages can be with different reasons, e.g., multiple physical meanings hidden behind social relationships. As a consequence, similarities lying on the linkages are overdetermined in real applications. Existing metric learning methods mainly focus on spatial linkages, leaving the rich semantic factors unconsidered. In this paper, we propose a Unified Multi-Metric Learning (UM2L) framework to exploit types of multiple metrics from either localities or semantics. In UM2L, a group of operators are introduced for distance characterizations, and we find this mechanism can introduce flexible configurations for different similarity measurements both spatially and semantically. Besides, we propose a uniform solver for UM2L and it is guaranteed to converge. Extensive experiments on diverse applications reveal the superior classification performance and interpretability of UM2L. Visualization effects also validate the ability of our framework on discovering physical meanings as well.

#95 Unsupervised Learning of Spoken Language with Visual Context

David Harwath (MIT CSAIL)
Antonio Torralba (MIT CSAIL)
James Glass (MIT CSAIL)

Humans learn to speak before they can read or write, so why can't computers do the same? In this paper, we present a deep neural network model capable of rudimentary spoken language acquisition using untranscribed audio training data, whose only supervision comes in the form of contextually relevant visual images. We describe the collection of our data comprised of over 120,000 spoken audio captions for the Places image dataset and evaluate our model on an image search and annotation task. We also provide some visualizations which suggest that our model is learning to recognize meaningful words within the caption spectrograms.

#96 Cyclades: Conflict-free Asynchronous Machine Learning

Xinghao Pan (UC Berkeley)
Maximilian Lam (UC Berkeley)
Stephen Tu (UC Berkeley)
Dimitrios Papailiopoulos
Ce Zhang (Stanford)
Michael I Jordan
Kannan Ramchandran
Chris Ré
Benjamin Recht

We present Cyclades, a general framework for parallelizing stochastic optimization algorithms in a shared memory setting. Cyclades is asynchronous during model updates, and requires no memory locking mechanisms, similar to Hogwild!-type algorithms. Unlike Hogwild!, Cyclades introduces no conflicts during parallel execution, and offers a black-box analysis for provable speedups across a large family of algorithms. Due to its inherent cache locality and conflict-free nature, our multi-core implementation of Cyclades consistently outperforms Hogwild!-type algorithms on sufficiently sparse datasets, leading to up to 40% speedup gains compared to Hogwild!, and up to 5\times gains over asynchronous implementations of variance reduction algorithms.

#97 Disease Trajectory Maps

Peter Schulam (Johns Hopkins Univ.)
Raman Arora

Medical researchers are coming to appreciate that many diseases are in fact complex, heterogeneous syndromes composed of subpopulations that express different variants of a related complication. Time series data extracted from individual electronic health records (EHR) offer an exciting new way to study subtle differences in the way these diseases progress over time. In this paper, we focus on answering two questions that can be asked using these databases of time series. First, we want to understand whether there are individuals with similar disease trajectories and whether there are a small number of degrees of freedom that account for differences in trajectories across the population. Second, we want to understand how important clinical outcomes are associated with disease trajectories. To answer these questions, we propose the Disease Trajectory Map (DTM), a novel probabilistic model that learns low-dimensional representations of sparse and irregularly sampled time series. We propose a stochastic variational inference algorithm for learning the DTM that allows the model to scale to large modern medical datasets. To demonstrate the DTM, we analyze data collected on patients with the complex autoimmune disease, scleroderma. We find that DTM learns meaningful representations of disease trajectories that the representations are significantly associated with important clinical outcomes.

#98 Fast ε-free Inference of Simulation Models with Bayesian Conditional Density Estimation

George Papamakarios (Univ. of Edinburgh)
Iain Murray (Univ. of Edinburgh)

Many statistical models can be simulated forwards but have intractable likelihoods. Approximate Bayesian Computation (ABC) methods are used to infer properties of these models from data. Traditionally these methods approximate the posterior over parameters by conditioning on data being inside an ε-ball around the observed data, which is only correct in the limit $\varepsilon \to 0$. Monte Carlo methods can then draw samples from the approximate posterior to approximate predictions or error bars on parameters. These algorithms critically slow down as $\varepsilon \to 0$, and in practice draw samples from a broader distribution than the posterior. We propose a new approach to likelihood-free inference based on Bayesian conditional density estimation. Preliminary inferences based on limited simulation data are used to guide later simulations. In some cases, learning an accurate parametric representation of the entire true posterior distribution requires fewer model simulations than Monte Carlo ABC methods need to produce a single sample from an approximate posterior.

#99 Stochastic Structured Prediction under Bandit Feedback

Artem Sokolov (Heidelberg Univ.)
Julia Kreutzer (Heidelberg Univ.)
Stefan Riezler (Heidelberg Univ.)

Stochastic structured prediction under bandit feedback follows a learning protocol where on each of a sequence of iterations, the learner receives an input, predicts an output structure, and receives partial feedback in form of a task loss evaluation of the predicted structure. We introduce stochastic approximation algorithms that apply this learning scenario to probabilistic structured prediction, with a focus on asymptotic convergence and ease of elicitability of feedback. We present simulation experiments for complex natural language processing tasks, showing fastest empirical convergence and smallest empirical variance for stochastic optimization of a non-convex pairwise preference learning objective compared to stochastic optimization of related non-convex and convex objectives.

#100 Learning under uncertainty: a comparison between R-W and Bayesian approach

Crane Huang (LIBR)
Martin Paulus (LIBR)

Accurately differentiating between what is truly unpredictably random and systematic changes that occur at random can have profound effect on affect and cognition. To examine the underlying computational principles that guide different learning behavior in an uncertain environment, we compared an R-W model with a fixed learning rate and a Bayesian approach with a belief of environmental stationarity in a visual search task with different volatility levels. Both R-W model and the Bayesian approach reflected an individual's estimation of the environmental volatility, and there is a strong correlation between the learning rate in R-W model and the belief of stationarity in the Bayesian approach in different volatility conditions. In a low volatility condition, R-W model indicates that learning rate positively correlates with lose-shift rate, but not choice optimality (inverted U shape). The Bayesian approach indicates that the belief of environmental stationarity positively correlates with choice optimality, but not lose-shift rate (inverted U shape). In addition, we showed that comparing to Expert learners, individuals with high lose-shift rate (sub-optimal learners) had significantly higher learning rate estimated from R-W model and lower belief of stationarity from the Bayesian model.

#101 Minimax Optimal Alternating Minimization for Kernel Nonparametric Tensor Learning

Taiji Suzuki
Heishiro Kanagawa
Hayato Kobayashi
Nobuyuki Shimizu
Yukihiro Tagami

We investigate the statistical performance and computational efficiency of the alternating minimization procedure for nonparametric tensor learning. Tensor modeling has been widely used for capturing the higher order relations between multimodal data sources. In addition to a linear model, a nonlinear tensor model has been received much attention recently because of its high flexibility. We consider an alternating minimization procedure for a general nonlinear model where the true function consists of components in a reproducing kernel Hilbert space (RKHS). In this paper, we show that the alternating minimization method achieves linear convergence as an optimization algorithm and that the generalization error of the resultant estimator yields the minimax optimality. We apply our algorithm to some multitask learning problems and show that the method actually shows favorable performances.

#102 On the Recursive Teaching Dimension of VC Classes

Xi Chen (Columbia Univ.)
Xi Chen (Columbia Univ.)
Yu Cheng (U of Southern California)
Bo Tang (Univ. of Oxford)

The recursive teaching dimension (RTD) of a concept class C \subseteq \{0, 1\}^n, introduced by Zilles et al. [ZLHZ11], is a complexity parameter measured by the worst-case number of labeled examples needed to learn any target concept of C in the recursive teaching model. In this paper, we study the quantitative relation between RTD and the well-known learning complexity measure VC dimension (VCD), and improve the best known upper and (worst-case) lower bounds on the recursive teaching dimension with respect to the VC dimension. Given a concept class C \subseteq \{0, 1\}^n with VCD(C) = d, we first show that RTD(C) is at most d 2^{d+1}. This is the first upper bound for RTD(C) that depends only on VCD(C), independent

of the size of the concept class |C| and its~domain size n. Before our work, the best known upper bound for RTD(C) is O(d 2^d log log |C|), obtained by Moran et al. [MSWY15]. We remove the log log |C| factor. We also improve the lower bound on the worst-case ratio of RTD(C) to VCD(C). We present a family of classes \{ C_k \}_{k \ge 1} with VCD(C_k) = 3k and RTD(C_k)=5k, which implies that the ratio of RTD(C) to VCD(C) in the worst case can be as large as 5/3. Before our work, the largest ratio known was 3/2 as obtained by Kuhlmann [Kuh99]. Since then, no finite concept class C has been known to satisfy RTD(C) > (3/2) VCD(C).

#103 Dimension-Free Iteration Complexity of Finite Sum Optimization Problems

Yossi Arjevani (Weizmann Institute of Science)
Ohad Shamir (Weizmann Institute of Science)

Many canonical machine learning problems boil down to a convex optimization problem with a finite sum structure. However, whereas much progress has been made in developing faster algorithms for this setting, the inherent limitations of these problems are not satisfactorily addressed by existing lower bounds. Indeed, current bounds focus on first-order optimization algorithms, and only apply in the often unrealistic regime where the number of iterations is less than \cO(d/n) (where d is the dimension and n is the number of samples). In this work, we extend the framework of Arjevani et al. \cite{arjevani2015lower,arjevani2016iteration} to provide new lower bounds, which are dimension-free, and go beyond the assumptions of current bounds, thereby covering standard finite sum optimization methods, e.g., SAG, SAGA, SVRG, SDCA without duality, as well as stochastic coordinate-descent methods, such as SDCA and accelerated proximal SDCA.

#104 f-GAN: Training Generative Neural Samplers using Variational Divergence Minimization

Sebastian Nowozin (Microsoft Research)
Botond Cseke (Microsoft Research)
Ryota Tomioka (MSRC)

Generative neural networks are probabilistic models that implement sampling using feedforward neural networks: they take a random input vector and produce a sample from a probability distribution defined by the network weights. These models are expressive and allow efficient computation of samples and derivatives, but cannot be used for computing likelihoods or for marginalization. The generative-adversarial training method allows to train such models through the use of an auxiliary discriminative neural network. We show that the generative-adversarial approach is a special case of an existing more general variational divergence estimation approach. We show that any f-divergence can be used for training generative neural networks. We discuss the benefits of various choices of divergence functions on training complexity and the quality of the obtained generative models.

#105 Low-Rank Regression with Tensor Responses

Guillaume Rabusseau (Aix-Marseille Univ.)
Hachem Kadri

This paper proposes an efficient algorithm (HOLRR) to handle regression tasks where the outputs have a tensor structure. We formulate the regression problem as the minimization of a least square criterion under a multilinear rank constraint, a difficult non convex problem. HOLRR computes efficiently an approximate solution of this problem, with solid theoretical guarantees. A kernel extension is also presented. Experiments on synthetic and real data show that HOLRR outperforms multivariate and multilinear regression methods and is considerably faster than existing tensor methods.

#106 Double Thompson Sampling for Dueling Bandits

Huasen Wu (Univ. of California at Davis)
Xin Liu (Univ. of California)

In this paper, we propose a Double Thompson Sampling (D-TS) algorithm for dueling bandit problems. As its name suggests, D-TS selects both the first and the second candidates according to Thompson Sampling. Specifically, D-TS maintains a posterior distribution for the preference matrix, and chooses the pair of arms for comparison according to two sets of samples independently drawn from the posterior distribution. This simple algorithm applies to general Copeland dueling bandits, including Condorcet dueling bandits as its special case. For general Copeland dueling bandits, we show that D-TS achieves O(K^2 log T) regret. Moreover, using a back substitution argument, we refine the regret to O(K log T + K^2 log log T) in Condorcet dueling bandits and many practical Copeland dueling bandits. In addition, we propose an enhancement of D-TS, referred to as D-TS+, that reduces the regret by carefully breaking ties. Experiments based on both synthetic and real-world data demonstrate that D-TS and D-TS^+ significantly improve the overall performance, in terms of regret and robustness.

#107 Linear dynamical neural population models through nonlinear embeddings

Yuanjun Gao (Columbia Univ.)
Evan W Archer
Liam Paninski
John Cunningham

A body of recent work in modeling neural activity focuses on recovering low-dimensional latent features that capture the statistical structure of large-scale neural populations. Most such approaches have focused on linear generative models, where inference is computationally tractable. Here, we propose fLDS, a general class of nonlinear generative models that permits the firing rate of each neuron to vary as an arbitrary smooth function of a latent, linear dynamical state. This extra flexibility allows the model to capture a richer set of neural variability than a purely linear model, but retains an easily visualizable low-dimensional latent space. To fit this class of non-conjugate models we propose a variational inference scheme, along with a novel approximate posterior capable of capturing rich temporal correlations across time. We show that our techniques permit inference in a wide class of generative models. We also show in application to two neural datasets that, compared to state-of-the-art neural population models, fLDS captures a much larger proportion of neural variability with a small number of latent dimensions, providing superior predictive performance and interpretability.

#108 Regret Bounds for Non-decomposable Metrics with Missing Labels

Nagarajan Natarajan (Microsoft Research)
Prateek Jain (Microsoft Research)

We consider the problem of recommending relevant labels (items) for a given data point (user). In particular, we are interested in the practically important setting where the evaluation is with respect to non-decomposable (over labels) performance metrics like the F_1 measure, {and} the training data has missing labels. To this end, we propose a generic framework that given a performance metric \Psi, can devise a regularized objective function and a threshold such that all the values in the predicted score vector above and only above the threshold are selected to be positive. We show that the regret or generalization error in the given metric \Psi is bounded ultimately by estimation error of certain underlying parameters. In particular, we derive regret bounds under three popular settings: a) collaborative filtering, b) multilabel classification, and c) PU (positive-unlabeled) learning. For each of the above problems, we can obtain precise non-asymptotic regret bound which is small even when a large fraction of labels is missing. Our empirical results on synthetic and benchmark datasets demonstrate that by explicitly modeling for missing labels and optimizing the desired performance metric, our algorithm indeed achieves significantly better performance (like F_1 score) when compared to methods that do not model missing label information carefully.

#109 Dynamic matrix recovery from incomplete observations under an exact low-rank constraint

Liangbei Xu (Gatech)
Mark Davenport

Low-rank matrix factorizations arise in a wide variety of applications -- including recommendation systems, topic models, and source separation, to name just a few. In these and many other applications, it has been widely noted that by incorporating temporal information and allowing for the possibility of time-varying models, significant improvements are possible in practice. However, despite the reported superior empirical performance of these dynamic models over their static counterparts, there is limited theoretical justification for introducing these more complex models. In this paper we aim to address this gap by studying the problem of recovering a dynamically evolving low-rank matrix from incomplete observations. First, we propose the locally weighted matrix smoothing (LOWEMS) framework as one possible approach to dynamic matrix recovery. We then establish error bounds for LOWEMS in both the {\em matrix sensing} and {\em matrix completion} observation models. Our results quantify the potential benefits of exploiting dynamic constraints both in terms of recovery accuracy and sample complexity. To illustrate these benefits we provide both synthetic and real-world experimental results.

#110 Rényi Divergence Variational Inference

Yingzhen Li (Univ. of Cambridge)
Richard E Turner

This paper introduces the variational Rényi bound (VR) that extends traditional variational inference to Rényi's alpha-divergences. This new family of variational lower-bounds unifies a number of existing variational methods, and enables a smooth interpolation from the evidence lower-bound to the log (marginal) likelihood that is controlled by the value of alpha that parametrises the divergence. The reparameterization trick, Monte Carlo approximation and stochastic optimisation methods are deployed to obtain a unified implementation for the VR bound optimisation. We further consider negative alpha values and propose a novel variational inference method as a new special case in the proposed framework. Experiments on Bayesian neural networks and variational auto-encoders demonstrate the wide applicability of the VR bound.

#111 Confusions over Time: An Interpretable Bayesian Model to Characterize Trends in Decision Making

Himabindu Lakkaraju (Stanford Univ.)
Jure Leskovec

We propose Confusions over Time (CoT), a novel generative framework which facilitates a multi-granular analysis of the decision making process. The CoT not only models confusions or error properties of individual decision makers and their evolution over time, but also allows us to obtain diagnostic insights into the collective decision making process in an interpretable manner. To this end, the CoT models the confusions of the decision makers and their evolution over time via time-dependent confusion matrices. Interpretable insights are obtained by grouping similar decision makers (and items being judged) into clusters and representing each such cluster with an appropriate prototype and identifying the most important features characterizing the cluster via a subspace feature indicator vector. Experimentation with real world data on bail decisions, asthma treatments, and insurance policy approval decisions demonstrates that CoT can accurately model and explain the confusions of decision makers and their evolution over time.

#112 Adaptive Averaging in Accelerated Descent Dynamics

Walid Krichene (UC Berkeley)
Alexandre Bayen (UC Berkeley)
Peter L Bartlett

We study accelerated descent dynamics for constrained convex optimization. This dynamics can be described naturally as a coupling of a dual variable accumulating gradients at a given rate \eta(t), and a primal variable obtained as the weighted average of the mirrored dual trajectory, with weights w(t). Using a Lyapunov argument, we give sufficient conditions on \eta and w to achieve a desired convergence rate. As an example, we show that the replicator dynamics (an example of mirror descent on the simplex) can be accelerated using a simple averaging scheme. We then propose an adaptive averaging heuristic which adaptively computes the weights to speed up the decrease of the Lyapunov function. We provide guarantees on adaptive averaging in continuous-time, and give numerical experiments in discrete-time to compare it with existing heuristics, such as adaptive restarting. The experiments indicate that adaptive averaging performs at least as well as adaptive restarting, with significant improvements in some cases.

#113 Bayesian Optimization for Probabilistic Programs

Tom Rainforth (Univ. of Oxford)
Tuan-Anh Le (Univ. of Oxford)
Jan-Willem van de Meent (Univ. of Oxford)
Michael A Osborne
Frank Wood

We present the first general purpose framework for marginal maximum a posteriori estimation of probabilistic program variables. By using a series of code transformation, the marginal likelihood of any probabilistic program, and therefore of any graphical model, can be optimized with respect to an arbitrary subset of its sampled variables. To carry out this optimization, we develop the first Bayesian optimization package to directly exploit the source code of its target, leading to innovations in problem-independent hyperpriors, unbounded optimization and implicit constraint satisfaction; delivering significant performance improvements over prominent existing packages. We present applications of our method to a number of tasks including engineering design and parameter optimization.

#114 Efficient Globally Convergent Stochastic Optimization for Canonical Correlation Analysis

Weiran Wang
Jialei Wang (Univ. of Chicago)
Dan Garber
Dan Garber
Nati Srebro

We study the stochastic optimization of canonical correlation analysis (CCA), whose objective is nonconvex and does not decouple over training samples. Although several stochastic gradient based optimization algorithms have been recently proposed to solve this problem, no global convergence guarantee was provided by any of them. Inspired by the alternating least squares/power iterations formulation of CCA, and the shift-and-invert preconditioning method for PCA, we propose two globally convergent meta-algorithms for CCA, both of which transform the original problem into sequences of least squares problems that need only be solved approximately. We instantiate the meta-algorithms with state-of-the-art SGD methods and obtain time complexities that significantly improve upon that of previous work. Experimental results demonstrate their superior performance.

#115 A Unified Approach for Learning the Parameters of Sum-Product Networks

Han Zhao (Carnegie Mellon Univ.)
Pascal Poupart
Geoffrey J Gordon

We present a unified approach for learning the parameters of Sum-Product networks (SPNs). We prove that any complete and decomposable SPN is equivalent to a mixture of trees where each tree corresponds to a product of univariate distributions. Based on the mixture model perspective, we characterize the objective function when learning SPNs based on the maximum likelihood estimation (MLE) principle and show that the optimization problem can be formulated as a signomial program. We construct two parameter learning algorithms for SPNs by using sequential monomial approximations (SMA) and the concave-convex procedure (CCCP), respectively. The two proposed methods naturally admit multiplicative updates, hence effectively avoiding the projection operation. With the help of the unified framework, we also show that, in the case of SPNs, CCCP leads to the same algorithm as Expectation Maximization (EM) despite the fact that they are different in general.

#116 Feature-distributed sparse regression: a screen-and-clean approach

Jiyan Yang (Stanford Univ.)
Michael W Mahoney
Michael Saunders (Stanford Univ.)
Yuekai Sun (Univ. of Michigan)

Most existing approaches to distributed sparse regression assume the data is partitioned by samples. However, for high-dimensional data, it is more natural to partition the data by features. We propose an approach to distributed sparse regression when the data is partitioned by features rather than samples. One benefit of our approach is it allows practitioners to tailor the approach to various distributed computing platforms by trading-off the total amount of data (in bits) sent over the communication network and the number of rounds of communication.

#117 Backprop KF: Learning Discriminative Deterministic State Estimators

Tuomas Haarnoja (UC Berkeley)
Anurag Ajay (UC Berkeley)
Sergey Levine (Univ. of Washington)
Pieter Abbeel (OpenAI / UC Berkeley / Gradescope)

Generative state estimators based on probabilistic filters and smoothers are one of the most popular classes of state estimators for robots and autonomous vehicles. However, generative models have limited capacity to handle rich sensory observations, such as camera images, since they must model the entire distribution over sensor readings. Discriminative models do not suffer from this limitation, but are typically more complex to train as latent variable models for state estimation. We present an alternative approach where the parameters of the latent state distribution are directly optimized as a deterministic computation graph, resulting in a simple and effective gradient descent algorithm for training discriminative state estimators. We show that this procedure can be used to train state estimators that use complex input, such as raw camera images, which must be processed using expressive nonlinear function approximators such as convolutional neural networks. Our model can be viewed as a type of recurrent neural network, and the connection to probabilistic filtering allows us to design a network architecture that is particularly well suited for state estimation. We evaluate our approach on tracking task with raw image inputs. The results show significant improvement over both standard generative approaches and regular recurrent neural networks.

#118 Swapout: Learning an ensemble of deep architectures

Saurabh Singh (UIUC)
Derek Hoiem (UIUC)
David Forsyth (UIUC)

We describe Swapout, a new stochastic training method, that outperforms ResNets of identical network structure yielding impressive results on CIFAR-10 and CIFAR-100. Swapout samples from a rich set of architectures including dropout, stochastic depth and residual architectures as special cases. When viewed as a regularization method swapout not only inhibits co-adaptation of units in a layer, similar to dropout, but also across network layers. We conjecture that swapout achieves strong regularization by implicitly tying the parameters across layers. When viewed as an ensemble training method, it samples a much richer set of architectures than existing methods such as dropout or stochastic depth. We propose a parameterization that reveals connections to exiting architectures and suggests a much richer set of architectures to be explored. We show that our formulation suggests an efficient training method and validate our conclusions on CIFAR-10 and CIFAR-100 matching state of the art accuracy. Remarkably, our 32 layer wider model performs similar to a 1001 layer ResNet model.

#119 Assortment Optimization Under the Mallows model

Antoine Desir (Columbia Univ.)
Vineet Goyal
Srikanth Jagabathula
Danny Segev

We consider the assortment optimization problem when customer preferences follow the Mallows distribution. The assortment optimization problem focuses on determining the revenue/profit maximizing subset of products from a large universe of products; it is an important decision that is commonly faced by retailers in determining what to offer their customers. We use the popular Mallows distribution to model customer demand. There are two key challenges: (a) Mallows distribution lacks a closed-form expression (and requires summing an exponential number of terms) to compute the choice probability and hence, the expected revenue/profit per customer; and (b) finding the best subset may require an exhaustive search. Our key contributions are a closed-form expression for computing the choice probability under the Mallows model and a compact mixed integer linear program (MIP) formulation for the assortment problem.

#120 Operator Variational Inference

Rajesh Ranganath (Princeton Univ.)
Dustin Tran (Columbia Univ.)
Jaan Altosaar (Princeton Univ.)
David Blei

Variational inference is an umbrella term for algorithms which cast Bayesian inference as optimization. Classically, variational inference uses the Kullback-Leibler divergence to define the optimization. While this divergence offers many convenient computational properties, it is designed for convenience, sacrificing many desirable statistical and computational properties of the approximation. In this paper, we revisit variational inference from its core principle, designing variational objectives that handle an explicit tradeoff of the computational and statistical properties one seeks from inference. We use operators to formalize a broad class of variational objectives. This class contains currently used objectives such as the Kullback-Leibler divergence as well as many new ones. Operators can be used to characterize different properties of the objective, such as those that scale to massive datasets or those that permit richer approximating families. We develop a single black box algorithm that optimizes any objective in the class with respect to variational parameters, while simultaneously adapting the objective to avoid the worst traits of the approximation. We study different operators on an instructive example and also on deep generative models of images.

#121 Select-and-Sample for Spike-and-Slab Sparse Coding

Abdul-Saboor Sheikh (SAP Labs Berlin)
Jörg Lücke

Probabilistic inference serves as a popular model for neural processing. It is still unclear, however, how approximate probabilistic inference can be accurate and scalable to very high-dimensional continuous latent spaces. Especially as typical posteriors for sensory data can be expected to exhibit complex latent dependencies including multiple modes. Here, we study an approach that can efficiently be scaled while maintaining a richly structured posterior approximation under these conditions. As example model we use spike-and-slab sparse coding for V1 processing, and combine latent subspace selection with Gibbs sampling (select-and-sample). Unlike factored variational approaches, the method can maintain large numbers of posterior modes and complex latent dependencies. Unlike pure sampling, the method is scalable to very high-dimensional latent spaces. Among all sparse coding approaches with non-trivial posterior approximations (MAP or ICA-like models), we report the largest-scale results. In applications we firstly verify the approach by showing competitiveness in standard denoising benchmarks. Secondly, we use its scalability to, for the first time, study highly-overcomplete settings for V1 encoding using sophisticated posterior representations. More generally, our study shows that very accurate probabilistic inference for multi-modal posteriors with complex dependencies is tractable, functionally desirable and consistent with models for neural inference.

#122 Fast recovery from a union of subspaces

Chinmay Hegde
Piotr Indyk (MIT)
Ludwig Schmidt (MIT)

We address the problem of recovering a high-dimensional but structured vector from linear observations in a general setting where the vector can come from an arbitrary union of subspaces. This setup includes well-studied problems such as compressive sensing and low-rank matrix recovery. We show how to design more efficient algorithms for the union-of subspace recovery problem by using *approximate* projections. Instantiating our general framework for the low-rank matrix recovery problem gives the fastest provable running time for an algorithm with optimal sample complexity. Moreover, we give fast approximate projections for 2D histograms, another well-studied low-dimensional model of data. We complement our theoretical results with experiments demonstrating that our framework also leads to improved time and sample complexity empirically.

#123 Ladder Variational Autoencoders

Casper Kaae Sønderby (Univ. of Copenhagen)
Tapani Raiko
Lars Maaløe (Technical Univ. of Denmark)
Søren Kaae Sønderby (KU)
Ole Winther (Technical Univ. of Denmark)

Variational autoencoders are powerful models for unsupervised learning. However deep models with several layers of dependent stochastic variables are difficult to train which limits the improvements obtained using these highly expressive models. We propose a new inference model, the Ladder Variational Autoencoder, that recursively corrects the generative distribution by a data dependent approximate likelihood in a process resembling the recently proposed Ladder Network. We show that this model provides state of the art predictive log-likelihood and tighter log-likelihood lower bound compared to the purely bottom-up inference in layered Variational Autoencoders and other generative models. We provide a detailed analysis of the learned hierarchical latent representation and show that our new inference model is qualitatively different and utilizes a deeper more distributed hierarchy of latent variables. Finally, we observe that batch normalization and deterministic warm-up (gradually turning on the KL-term) are crucial for training variational models with many stochastic layers.

#124 SPALS: Fast Alternating Least Squares via Implicit Leverage Scores Sampling

Dehua Cheng (Univ. of Southern California)
Richard Peng
Yan Liu
Kimis Perros (Georgia Institute of Technology)

Tensor CANDECOMP/PARAFAC (CP) decomposition is a powerful but computationally challenging tool in modern data analytics. In this paper, we show ways of sampling intermediate steps of alternating minimization algorithms for computing low rank tensor CP decompositions, leading to the sparse alternating least squares (SPALS) method. Specifically, we sample the the Khatri-Rao product, which arises as an intermediate object during the iterations of alternating least squares. This product captures the interactions between different tensor modes, and form the main computational bottleneck for solving many tensor related tasks. By exploiting the spectral structures of the matrix Khatri-Rao product, we provide efficient access to its statistical leverage scores. When applied to the tensor CP decomposition, our method leads to the first algorithm that runs in sublinear time per-iteration and approximates the output of deterministic alternating least squares algorithms. Empirical evaluations of this approach show significantly speedups over existing randomized and deterministic routines for performing CP decomposition. On a tensor of the size 2.4m by 6.6m by 92k with over 2 billion nonzeros formed by Amazon product reviews, our routine converges in two minutes to the same error as deterministic ALS.

#125 CRF-CNN: Modeling Structured Information in Human Pose Estimation

Xiao Chu (Cuhk)
Wanli Ouyang
hongsheng Li (cuhk)
Xiaogang Wang (Chinese Univ. of Hong Kong)

Deep convolutional neural networks (CNN) have achieved great success. On the other hand, modeling structural information has been proved critical in many vision problems. It is of great interest to integrate them effectively. In a classical neural network, there is no message passing between neurons in the same layer. In this paper, we propose a CRF-CNN framework which can simultaneously model structural information in both output and hidden feature layers in a probabilistic way, and it is applied to human pose estimation. A message passing scheme is proposed, so that in various layers each body joint receives messages from all the others in an efficient way. Such message passing can be implemented with convolution between features maps in the same layer, and it is also integrated with feedforward propagation in neural networks. Finally, a neural network implementation of end-to-end learning CRF-CNN is provided. Its effectiveness is demonstrated through experiments on two benchmark datasets.

#126 A Consistent Regularization Approach for Structured Prediction

Carlo Ciliberto (MIT)
Lorenzo Rosasco
Alessandro Rudi

We propose and analyze a regularization framework for structured prediction problems. We characterize a large class of loss functions that allows to naturally embed structured outputs in a linear space. We exploit this fact to design learning algorithms using a surrogate loss approach and regularization techniques. Consistency and finite sample bounds are proved characterizing the generalization properties of the proposed methods. Experimental results are provided to demonstrate its practical usefulness.

#127 Refined Lower Bounds for Adversarial Bandits

Sébastien Gerchinovitz
Tor Lattimore

We provide new lower bounds on the regret that must be suffered by adversarial bandit algorithms. The new results show that recent upper bounds that either (a) hold with high-probability or (b) depend on the total loss of the best arm or (c) depend on the quadratic variation of the losses, are close to tight. Besides this we prove two impossibility results. First, the existence of a single arm that is optimal in every round cannot improve the regret in the worst case. Second, the regret cannot scale with the effective range of the losses. In contrast, both results are possible in the full-information setting.

#128 Learning Deep Embeddings with Histogram Loss

Evgeniya Ustinova (Skoltech)
Victor Lempitsky

We suggest a new loss for learning deep embeddings. The key characteristics of the new loss is the absence of tunable parameters and very good results obtained across a range of datasets and problems. The loss is computed by estimating two distribution of similarities for positive (matching) and negative (non-matching) point pairs, and then computing the probability of a positive pair to have a lower similarity score than a negative pair based on these probability estimates. We show that these operations can be performed in a simple and piecewise-differentiable manner using 1D histograms with soft assignment operations. This makes the proposed loss suitable for learning deep embeddings using stochastic optimization. The experiments reveal favourable results compared to recently proposed loss functions.

#129 Solving Marginal MAP Problems with NP Oracles and Parity Constraints

Yexiang Xue (Cornell Univ.)
zhiyuan li (Tsinghua Univ.)
Stefano Ermon
Carla P Gomes (Cornell Univ.)
Bart Selman

Arising from many applications at the intersection of decision making and machine learning, Marginal Maximum A Posteriori (Marginal MAP) Problems unify the two main classes of inference, namely maximization (optimization) and marginal inference (counting), and are believed to have higher complexity than both of them. We propose XOR_MMAP, a novel approach to solve the Marginal MAP Problem, which represents the intractable counting subproblem with queries to NP oracles, subject to additional parity constraints. XOR_MMAP provides a constant factor approximation to the Marginal MAP Problem, by encoding it as a single optimization in polynomial size of the original problem. We evaluate our approach in several machine learning and decision making applications, and show that our approach outperforms several state-of-the-art Marginal MAP solvers.

#130 Kernel Bayesian Inference with Posterior Regularization

Yang Song (Stanford Univ.)
Jun Zhu
Yong Ren (Tsinghua Univ.)

We propose a vector-valued regression problem whose solution is equivalent to the reproducing kernel Hilbert space (RKHS) embedding of the Bayesian posterior distribution. This equivalence provides a new understanding of kernel Bayesian inference. Moreover, the optimization problem induces a new regularization for the posterior embedding estimator, which is faster and has comparable performance to the squared regularization in kernel Bayes' rule. This regularization coincides with a former thresholding approach used in kernel POMDPs whose consistency remains to be established. Our theoretical work solves this open problem and provides consistency analysis in regression settings. Based on our optimizational formulation, we propose a flexible Bayesian posterior regularization framework which for the first time enables us to put regularization at the distribution level. We apply this method to nonparametric state-space filtering tasks with extremely nonlinear dynamics and show performance gains over all other baselines.

#131 Learning Influence Functions from Incomplete Observations

Xinran He (USC)
Ke Xu (USC)
David Kempe (USC)
Yan Liu

We study the problem of learning influence functions under incomplete observations of node activations. Incomplete observations are a major concern as most (online and real-world) social networks are not fully observable. We establish both proper and improper PAC learnability of influence functions under uniformly randomly missing observations. Proper PAC learnability under the Discrete-Time Linear Threshold (DLT) and Discrete-Time Independent Cascade (DIC) models is established by essentially reducing incomplete observations to complete observations in a modified graph. Our improper PAC learnability result applies for the DLT and DIC models as well as the Continuous-Time Independent Cascade (CIC) model. It is based on a parametrization in terms of reachability features, and also gives rise to an efficient and practical heuristic. Experiments on synthetic and real-world datasets demonstrate the ability of our method to compensate even for fairly large loss rates.

#132 General Tensor Spectral Co-clustering for Higher-Order Data

Tao Wu (Purdue Univ.)
Austin R Benson (Stanford Univ.)
David Gleich

Spectral clustering and co-clustering are well-known techniques in data analysis, and recent work has extended spectral clustering to square, symmetric tensors and hypermatrices derived from a network. We develop a new tensor spectral co-clustering method that simultaneously clusters the rows, columns, and slices of a nonnegative three-mode tensor and generalizes to tensors with any number of modes. The algorithm is based on a new random walk model which we call the super-spacey random surfer. We show that our method out-performs state-of-the-art co-clustering methods on several synthetic datasets with ground truth clusters and then use the algorithm to analyze several real-world datasets.

#133 Bayesian latent structure discovery from multi-neuron recordings

Scott Linderman
Ryan P Adams
Jonathan W Pillow

Neural circuits contain heterogeneous groups of neurons that differ in type, location, connectivity, and basic response properties. However, traditional methods for dimensionality reduction and clustering are ill-suited to recovering the structure underlying the organization of neural circuits. In particular, they do not take advantage of the rich temporal dependencies in multi-neuron recordings and fail to account for the noise in neural spike trains. Here we describe new tools for inferring latent structure from simultaneously recorded spike train data using a hierarchical extension of a multi-neuron point process model commonly known as the generalized linear model (GLM). Our approach combines the GLM with flexible graph-theoretic priors governing the relationship between latent features and neural connectivity patterns. Fully Bayesian inference via Pólya-gamma augmentation of the resulting model allows us to classify neurons and infer latent dimensions of circuit organization from correlated spike trains. We demonstrate the effectiveness of our method with applications to synthetic data and multi-neuron recordings in primate retina, revealing latent patterns of neural types and locations from spike trains alone.

#134 Estimating the Size of a Large Network and its Communities from a Random Sample

Lin Chen (Yale Univ.)
Amin Karbasi
Forrest W. Crawford (Yale Univ.)

Most real-world networks are too large to be measured or studied directly and there is substantial interest in estimating global network properties from smaller sub-samples. One of the most important global properties is the number of vertices/nodes in the network. Estimating the number of vertices in a large network is a major challenge in computer science, epidemiology, demography, and intelligence analysis. In this paper we consider a population random graph G = (V;E) from the stochastic block model (SBM) with K communities/blocks. A sample is obtained by randomly choosing a subset W and letting G(W) be the induced subgraph in G of the vertices in W. In addition to G(W), we observe the total degree of each sampled vertex and its block membership. Given this partial information, we propose an efficient PopULation Size Estimation algorithm, called PULSE, that correctly estimates the size of the whole population as well as the size of each community. To support our theoretical analysis, we perform an exhaustive set of experiments to study the effects of sample size, K, and SBM model parameters on the accuracy of the estimates. The experimental results also demonstrate that PULSE significantly outperforms a widely-used method called the network scale-up estimator in a wide variety of scenarios. We conclude with extensions and directions for future work.

#135 Wasserstein Training of Restricted Boltzmann Machines

Grégoire Montavon
Klaus-Robert Müller
Marco Cuturi

Restricted Boltzmann machines are able to learn highly complex, multimodal, structured and multiscale real-world data distributions. Parameters of the model are usually learned by minimizing the Kullback-Leibler (KL) divergence from training samples to the learned model. We

propose in this work a novel approach for Boltzmann machine training which assumes that a meaningful metric between observations is given. This metric can be represented by the Wasserstein distance between distributions, for which we derive a gradient with respect to the model parameters. Minimization of this new objective leads to generative models with different statistical properties. We demonstrate their practical potential on data completion and denoising, for which the metric between observations plays a crucial role.

#136 Deep ADMM-Net for Compressive Sensing MRI

yan yang (Xi'an Jiaotong Univ.)
Jian Sun (Xi'an Jiaotong Univ.)
Huibin Li
Zongben Xu

Compressive Sensing (CS) is an effective approach for fast Magnetic Resonance Imaging (MRI). It aims at reconstructing MR image from a small number of sampled data in k-space, and accelerating the data acquisition in MRI. To improve the current MRI system in reconstruction accuracy and computational speed, in this paper, we propose a novel deep architecture, dubbed ADMM-Net. The ADMM-Net is defined over a data flow graph, which is derived from the iterative procedures in Alternating Direction Method of Multipliers (ADMM) algorithm optimizing a general CS-based MRI model. In the training phase, all parameters of the net, e.g., image transforms, shrinkage functions, etc., are discriminatively learned end-to-end using L-BFGS algorithm. In the testing phase, it has computational overhead similar to ADMM but uses optimized parameters learned from the training data for CS-based reconstruction task. Experiments on MRI image reconstruction under different sampling ratios in k-space demonstrate that it significantly improves the baseline ADMM algorithm and achieves state-of-the-art reconstruction accuracies with fast computational speed.

#137 Maximization of Approximately Submodular Functions

Thibaut Horel (Harvard Univ.)
Yaron Singer

We study the problem of maximizing a function that is approximately submodular under a cardinality constraint. Approximate submodularity implicitly appears in a wide range of applications as in many cases errors in evaluation of a submodular function break submodularity. Say that F is \eps-approximately submodular if there exists a submodular function f such that (1-\eps)f(S) \leq F(S)\leq (1+\eps)f(S) for all subsets S. We are interested in characterizing the query-complexity of maximizing F subject to a cardinality constraint k as a function of the error level \eps > 0. We provide both lower and upper bounds: for \eps < n^{-1/2} we show an exponential query-complexity lower bound. In contrast, when \eps > {1}/{k} or under a stronger bounded curvature assumption, we give constant approximation algorithms.

#138 Combining Low-Density Separators with CNNs

Yu-Xiong Wang (Carnegie Mellon Univ.)
Martial Hebert (Carnegie Mellon Univ.)

This work explores using CNNs for recognition of novel categories from few examples. Inspired by the transferability analysis of CNNs, we introduce an additional unsupervised pre-training stage that exposes multiple higher layer units to large amounts of unlabeled real-world images. By encouraging these units to learn diverse sets of low-density separators across the unlabeled data, we capture a more generic, richer description of the visual world, which decouples these units from ties to a specific set of categories. We propose an unsupervised margin maximization that jointly estimates compact high-density regions and infers low-density separators. The low-density separator (LDS) module can be plugged into any or all of the top layers of a standard CNN architecture. The resulting CNNs, with enhanced generality, significantly improve the performance in scene classification, fine-grained recognition, and action recognition with small training samples.

#139 Learning Sensor Multiplexing Design through Back-propagation

Ayan Chakrabarti

Recent progress on many imaging and vision tasks has been driven by the use of deep feed-forward neural networks, which are trained by propagating gradients of a loss defined on the final output, back through the network up to the first layer that operates directly on the image. We propose back-propagating one step further---to learn camera sensor designs jointly with networks that carry out inference on the images they capture. In this paper, we specifically consider the design and inference problems in a typical color camera---where the sensor is able to measure only one color channel at each pixel location, and computational inference is required to reconstruct a full color image. We learn the camera sensor's color multiplexing pattern by encoding it as layer whose learnable weights determine which color channel, from among a fixed set, will be measured at each location. These weights are jointly trained with those of a reconstruction network that operates on the corresponding sensor measurements to produce a full color image. Our network achieves significant improvements in accuracy over the traditional Bayer pattern used in most color cameras. It automatically learns to employ a sparse color measurement approach similar to that of a recent design, and moreover, improves upon that design by learning an optimal layout for these measurements.

#140 High resolution neural connectivity from incomplete tracing data using nonnegative spline regression

Kameron D Harris (Univ. of Washington)
Stefan Mihalas (Allen Institute for Brain Science)
Eric Shea-Brown (Univ. of Washington)

Whole-brain neural connectivity data are now available from viral tracing experiments, which reveal the connections between a source injection site and elsewhere in the brain. These hold the promise of revealing spatial patterns of connectivity throughout the mammalian brain. To achieve this goal, we seek to fit a weighted, nonnegative adjacency matrix among 100 μm brain "voxels" using viral tracer data. Despite a multi-year experimental effort, the problem remains severely underdetermined: Injection sites provide incomplete coverage, and the number of voxels is orders of magnitude larger than the number of injections. Furthermore, projection data are missing within the injection site because local connections there are not separable from the injection signal. We use a novel machine-learning algorithm to meet these challenges and develop a spatially explicit, voxel-scale connectivity map of the mouse visual system. Our method combines three features: a matrix completion loss for missing data, a smoothing spline penalty term to regularize the problem, and (optionally) a low rank factorization. We demonstrate the consistency of our estimator using synthetic data and then apply it to newly available Allen Mouse Brain Connectivity Atlas data for the visual system. Our algorithm is significantly more predictive than current state of the art approaches which assume regions to be homogeneous. We demonstrate the efficacy of a low rank version on visual cortex data and discuss the possibility of extending this to a whole-brain connectivity matrix at the voxel scale.

#141 Learning a Probabilistic Latent Space of Object Shapes via 3D Generative-Adversarial Modeling

Chengkai Zhang
Jiajun Wu (MIT)
Tianfan Xue
Bill Freeman
Josh Tenenbaum

We study the problem of 3D object generation. We propose a novel framework which generates 3D objects from a probabilistic space by leveraging recent advances in volumetric convolutional networks and generative adversarial nets. The benefits of our model are three-fold: first, the use of a generative adversarial criterion, instead of traditional heuristic criteria, enables the generator to capture object structure implicitly and to synthesize realistic objects; second, the generator establishes a compact mapping from a low-dimensional probabilistic space to the space of 3D objects, so that we can sample objects without a reference image or CAD model, and explore the 3D object manifold; third, the adversarial discriminator provides a powerful 3D shape descriptor which, learned without supervision, has wide applications in 3D object recognition. Experiments demonstrate that our method generates high-quality 3D objects, and achieves very impressive performance on 3D object recognition, comparable with supervised learning algorithms.

#142 Learning Sparse Gaussian Graphical Models with Overlapping Blocks

Seyed Mohammad Javad Hosseini (Univ. of Washington)
Su-In Lee

We present a novel framework, called GRAB (GRaphical models with overlApping Blocks), to capture densely connected components in a network estimate. GRAB takes as input a data matrix of p variables and n samples, and jointly learns both a network among p variables and densely connected groups of variables (called `blocks'). GRAB has four major novelties as compared to existing network estimation methods: 1) It does not require the blocks to be given a priori. 2) Blocks can overlap. 3) It can jointly learn a network structure and overlapping blocks. 4) It solves a joint optimization problem with the block coordinate descent method that is convex in each step. We show that GRAB reveals the underlying network structure substantially better than four state-of-the-art competitors on synthetic data. When applied to cancer gene expression data, GRAB outperforms its competitors in revealing known functional gene sets and potentially novel genes that drive cancer.

#143 Multi-step learning and underlying structure in statistical models

Maia Fraser (Univ. of Ottawa)

In multi-step learning, where a final learning task is accomplished via a sequence of intermediate learning tasks, the intuition is that successive steps or levels transform the initial data into representations more and more ``suited" to the final learning task. A related principle arises in transfer-learning where Baxter (2000) proposed a theoretical framework to study how learning multiple tasks transforms the inductive bias of a learner. The most widespread multi-step learning approach is semi-supervised learning with two steps: unsupervised, then supervised. Several authors (Castelli-Cover, 1996; Balcan-Blum, 2005;, Niyogi, 2008; Ben-David et al, 2008) have analyzed SSL, with Balcan-Blum (2005) proposing a version of the PAC learning framework augmented by a ``compatibility function" to link concept class and unlabeled data distribution. We propose to analyze SSL and other multi-step learning approaches, much in the spirit of Baxter's framework, by defining a learning problem generatively as a full statistical model on X \times Y. We show this implies a natural compatibility function which agrees with that of Balcan-Blum in the identifiable case and in general corresponds to choice of a factorization of the model. This viewpoint sheds geometric insight on generative vs. distribution-free frameworks: the former correspond to learning in a model that is a fiber sub-bundle of the large product model of the latter. As tool, we define a notion of Γ-uniform shattering for statistical models. We use this to give conditions on the marginal and conditional models which imply an advantage for multi-step learning approaches. In particular, we recover a more general version of a result of Poggio et al (2012): under mild hypotheses a multi-step approach which learns features invariant under successive factors of a finite group of invariances has sample complexity requirements that are additive rather than multiplicative in the size of the subgroups.

#144 Dynamic Network Surgery for Efficient DNNs

Yiwen Guo (Intel Labs China)
Anbang Yao
Yurong Chen

Deep learning has become a ubiquitous technology to improve machine intelligence. However, most of the existing deep models are structurally very complex, making them difficult to be deployed on the mobile platforms with limited computational power. In this paper, we propose a novel network compression method called dynamic network surgery, which can remarkably reduce the network complexity by making on-the-fly connection pruning. Unlike the previous methods which accomplish this task in a greedy way, we properly incorporate connection splicing into the whole process to avoid incorrect pruning and make it as a continual network maintenance. The effectiveness of our method is proved with experiments. Without any accuracy loss, our method can efficiently compress the number of parameters in LeNet-5 and AlexNet by a factor of \bm{108}\times and \bm{17.7}\times respectively, proving that it outperforms the recent pruning method by considerable margins.

#145 Active Nearest-Neighbor Learning in Metric Spaces

Aryeh Kontorovich
Sivan Sabato (Ben-Gurion Univ. of the Negev)
Ruth Urner (MPI Tuebingen)

We propose a pool-based non-parametric active learning algorithm for general metric spaces, called MArgin Regularized Metric Active Nearest Neighbor (MARMANN), which outputs a nearest-neighbor classifier. We give prediction error guarantees that depend on the noisy-margin properties of the input sample, and are competitive with those obtained by previously proposed passive learners. We prove that the label complexity of MARMANN is significantly lower than that of any passive learner with similar error guarantees. Our algorithm is based on a generalized sample compression scheme and a new label-efficient active model-selection procedure.

#146 Discriminative Gaifman Models

Mathias Niepert

We present Gaifman models, a novel family of relational machine learning models. Gaifman models learn knowledge base representations bottom up from representations of locally connected and bounded-size regions of the input. Considering local and bounded-size neighborhoods renders logical inference and learning tractable, mitigates the problem of overfitting, and facilitates weight sharing across neighborhoods. We present the core ideas of Gaifman models and apply them to large-scale relational learning problems. We also discuss the ways in which Gaifman models relate to some existing relational machine learning approaches.

#147 Professor Forcing: A New Algorithm for Training Recurrent Networks

Alex M Lamb (Montreal)
Anirudh Goyal (Univ. of Montreal)
Ying Zhang (Univ. of Montreal)
Saizheng Zhang (Univ. of Montreal)
Aaron C Courville (Univ. of Montreal)
Yoshua Bengio (U. Montreal)

The Teacher Forcing algorithm trains recurrent networks by supplying observed sequence values as inputs during training and using the network's own one-step-ahead predictions to do multi-step sampling. We introduce the Professor Forcing algorithm, which uses adversarial domain adaptation to encourage the dynamics of the recurrent network to be the same when training the network and when sampling from the network over multiple time steps. We apply Professor Forcing to language modeling, vocal synthesis on raw waveforms, handwriting generation, and image generation. Empirically we find that Professor Forcing acts as a regularizer, improving test likelihood on character level Penn Treebank and sequential MNIST. We also find that the model qualitatively improves samples, especially when sampling for a large number of time steps. This is supported by human evaluation of sample quality. Trade-offs between Professor Forcing and Scheduled Sampling are discussed. We produce T-SNEs showing that Professor Forcing successfully makes the dynamics of the network during training and sampling more similar.

#148 Pruning Random Forests for Prediction on a Budget

Feng Nan (Boston Univ.)
Joseph Wang (Boston Univ.)
Venkatesh Saligrama

We propose to prune a random forest (RF) for resource-constrained prediction. We first construct a RF and then prune it to optimize expected feature cost & accuracy. We pose pruning RFs as a novel 0-1 integer program with linear constraints that encourages feature re-use. We establish total unimodularity of the constraint set to prove that the corresponding LP relaxation solves the original integer program. We then exploit connections to combinatorial optimization and develop an efficient primal-dual algorithm, scalable to large datasets. In contrast to our bottom-up approach, which benefits from good RF initialization, conventional methods are top-down acquiring features based on their utility value and is generally intractable, requiring heuristics. Empirically, our pruning algorithm outperforms existing state-of-the-art resource-constrained algorithms.

#149 Multistage Campaigning in Social Networks

Mehrdad Farajtabar (Georgia Tech)
Xiaojing Ye (Georgia State Univ.)
Sahar Harati (Emory Univ.)
Le Song
Hongyuan Zha (Georgia Institute of Technology)

We consider control problems for multi-stage campaigning over social networks. The dynamic programming framework is employed to balance the high present reward and large penalty on low future outcome in the presence of extensive uncertainties. In particular, we establish theoretical foundations of optimal campaigning over social networks where the user activities are modeled as a multivariate Hawkes process, and we derive a time dependent linear relation between the intensity of exogenous events and several commonly used objective functions of campaigning. We further develop a convex dynamic programming framework for determining the optimal intervention policy that prescribes the required level of external drive at each stage for the desired campaigning result. Experiments on both synthetic data and the real-world MemeTracker dataset show that our algorithm can steer the user activities for optimal campaigning much more accurately than baselines.

#150 Coevolutionary Latent Feature Processes for Continuous-Time User-Item Interactions

Yichen Wang (Georgia Tech)
Nan Du
Rakshit Trivedi (Georgia Institute of Technolo)
Le Song

Discovering users' current interests, and matching them to the right service items such as groups and products is a fundamental task in recommender systems. As users interact with different services in continuous time, users' interests and services' features may drift, evolve and even coevolve over time. Traditional models based on static latent features or discretizing time into epochs can become ineffective for capturing the fine-grained temporal dynamics in the user-item interactions. In this paper, we propose a coevolutionary latent feature process for user-item interactions, which is designed to capture the coevolving nature of user and item features over continuous time. Despite the sophistication of the model, we show that the model parameters can be estimated efficiently from interaction traces via a convex optimization algorithm. We have evaluated our method over a variety of datasets, including online TV streaming data and discussion forum data, verifying that our method can lead to significant improvements in user behavior prediction compared to previous state-of-the-arts.

#151 Coordinate-wise Power Method

Qi Lei (UT AUSTIN)
Kai Zhong (UT AUSTIN)
Inderjit S Dhillon

In this paper, we propose a coordinate-wise version of the power method from an optimization viewpoint. The vanilla power method simultaneously updates all the coordinates of the iterate, which is essential for its convergence analysis. However, different coordinates converge to the optimal value at different speeds. Our proposed algorithm, which we call coordinate-wise power method, is able to select and update the most important k coordinates in O(kn) time at each iteration, where n is the dimension of the matrix and k <= n is the size of active set. Inspired by the "greedy" nature of our method, we further propose a greedy coordinate descent algorithm applied on a non-convex objective function specialized for symmetric matrices. We provide convergence analyses for both methods. Experimental results on both synthetic data and real data show that our methods achieve up to 14 times speedup over the vanilla power method. Meanwhile, due to their the coordinate-wise nature, our methods are very suitable to deal with the important case when data cannot fit into memory. Finally, we introduce how the coordinate-wise mechanism could be applied to other iterative methods that are used in machine learning.

#152 Barzilai-Borwein Step Size for Stochastic Gradient Descent

Conghui Tan (The Chinese Univ. of HK)
Shiqian Ma
Yu-Hong Dai
Yuqiu Qian (The Univ. of Hong Kong)

One of the major issues in stochastic gradient descent (SGD) methods is how to choose an appropriate step size while running the algorithm. Since the traditional line search technique does not apply for stochastic optimization methods, the common practice in SGD is either to use a diminishing step size, or to tune a step size by hand, which can be time consuming in practice. In this paper, we propose to use the Barzilai-Borwein (BB) method to automatically compute step sizes for SGD and its variant: stochastic variance reduced gradient (SVRG) method, which leads to two algorithms: SGD-BB and SVRG-BB. We prove that SVRG-BB converges linearly for strongly convex objective functions. As a by-product, we prove the linear convergence result of SVRG with Option I proposed in [10], whose convergence result has been missing in the literature. Numerical experiments on standard data sets show that the performance of SGD-BB and SVRG-BB is comparable to and sometimes even better than SGD and SVRG with best-tuned step sizes, and is superior to some advanced SGD variants.

#153 Fast learning rates with heavy-tailed losses

Vu C Dinh (Fred Hutchinson Cancer Center)
Lam S Ho (UCLA)
Binh Nguyen (Univ. of Science)
Duy Nguyen (Univ. of Wisconsin-Madison)

Recent progresses about fast learning rates with structured data have refined our understanding about settings under which such rates are possible, leading to the development of robust algorithms that can automatically adapt to data with hidden structures and achieve faster rates whenever possible. The literature, however, has mainly focused on bounded losses and little has been known about rates of learning in the unbounded cases. The purpose of this research is to study fast learning rates when the losses are not necessarily bounded and may have a distribution with heavy tails. To enable such analyses, we introduce two new conditions: (i) the envelope function \sup_{f \in \mathcal{F}}|ℓ \circ f|, where ℓ is the loss function and \mathcal{F} is the hypothesis class, exists and is L^r-integrable, and (ii) ℓ satisfies the multi-scale Bernstein's condition on \mathcal{F}. Under these assumptions, we prove that learning rate faster than O(n^{-1/2}) can be obtained and, depending on r and the multi-scale Bernstein's powers, can be arbitrarily close to O(n^{-1}). We then verify these assumptions and derive fast learning rates for the problem of vector quantization by k-means clustering with heavy-tailed distributions. The analyses enable us to obtain novel learning rates that extend and complement existing results in the literature from both theoretical and practical viewpoints.

#154 CliqueCNN: Deep Unsupervised Exemplar Learning

Miguel A Bautista (Heidelberg Univ.)
Artsiom Sanakoyeu (Heidelberg Univ.)
Ekaterina Tikhoncheva (Heidelberg Univ.)
Bjorn Ommer

Exemplar learning is a powerful paradigm for discovering visual similarities in an unsupervised manner. In this context, however, the recent breakthrough in deep learning could not yet unfold its full potential. With only a single positive sample, a great imbalance between one positive and many negatives, and unreliable relationships between most samples, training of convolutional neural networks is impaired. Given weak estimates of local distance we propose a single optimization problem to extract batches of samples with mutually consistent relations. Conflicting relations are distributed over different batches and similar samples are grouped into compact cliques. Learning exemplar similarities is framed as a sequence of clique categorization tasks. The CNN then consolidates transitivity relations within and between cliques and learns a single representation for all samples without the need for labels. The proposed unsupervised approach has shown competitive performance on detailed posture analysis and object classification.

#155 Guided Policy Search as Approximate Mirror Descent

William H Montgomery (Univ. of Washington)
Sergey Levine (Univ. of Washington)

Guided policy search algorithms can be used to optimize complex nonlinear policies, such as deep neural networks, without directly computing policy gradients in the high-dimensional parameter space. Instead, these methods use supervised learning to train the policy to mimic a "teacher" algorithm, such as a trajectory optimizer or a trajectory-centric reinforcement learning method. Guided policy search methods provide asymptotic local convergence guarantees by construction, but it is not clear how much the policy improves within a small, finite number of iterations. We show that guided policy search algorithms can be interpreted as an approximate variant of mirror descent, where the projection onto the constraint manifold is not exact. We derive a new guided policy search algorithm that is simpler and provides appealing improvement and convergence guarantees in simplified convex and linear settings, and show that in the more general nonlinear setting, the error in the projection step can be bounded. We provide empirical results on several simulated robotic manipulation tasks that show that our method is stable and achieves similar or better performance when compared to prior guided policy search methods, with a simpler formulation and fewer hyperparameters.

#156 Structured Sparse Regression via Greedy Hard Thresholding

Prateek Jain (Microsoft Research)
Nikhil Rao
Inderjit S Dhillon

Several learning applications require solving high-dimensional regression problems where the relevant features belong to a small number of (overlapping) groups. For very large datasets and under standard sparsity constraints, hard thresholding methods have proven to be extremely efficient, but such methods require NP hard projections when dealing with overlapping groups. In this paper, we show that such NP-hard projections can not only be avoided by appealing to submodular optimization, but such methods come with strong theoretical guarantees even in the presence of poorly conditioned data (i.e. say when two features have correlation \geq 0.99), which existing analyses cannot handle. These methods exhibit an interesting computation-accuracy trade-off and can be extended to significantly harder problems such as sparse overlapping groups. Experiments on both real and synthetic data validate our claims and demonstrate that the proposed methods are orders of magnitude faster than other greedy and convex relaxation techniques for learning with group-structured sparsity.

#157 Learning in Games: Robustness of Fast Convergence

Dylan J Foster (Cornell Univ.)
zhiyuan li (Tsinghua Univ.)
Thodoris Lykouris (Cornell Univ.)
Karthik Sridharan (Univ. of Pennsylvania)
Eva Tardos (Cornell Univ.)

We consider repeated games where players use regret minimizing algorithms for choosing their actions. We introduce a low approximate regret property, allowing a small a multiplicative approximation factor of regret. We show that this property is ubiquitous among learning algorithms with surprisingly small additive error term, which in turn implies fast convergence to approximate optimality in a large class of games. Our results improve recent work of Syrgkanis et al in a number of ways: we improve the speed of convergence by a factor of n, the number of players; we broaden the class of learning algorithms considered, and require only realized feedback, or even just bandit feedback in some cases, not the expectation over actions of other players; and show that convergence occurs with high probability. For learners with only bandit feedback, we present a novel algorithm which should be of independent interest as it provides a "small loss"-type bound with improved dependence on the number of actions. We also extend the results to dynamic population games, showing that many algorithms have the low approximate regret property even with shifting experts, and increasing the allowed churn in the results of Lykouris et al.

#158 Measuring the reliability of MCMC inference with Bidirectional Monte Carlo

Roger B Grosse
Siddharth Ancha (Univ. of Toronto)
Dan Roy

Markov chain Monte Carlo (MCMC) is one of the main workhorses of probabilistic inference, but it is notoriously hard to measure the accuracy of approximate posterior samples. In this work, we apply the recently introduced bidirectional Monte Carlo technique to evaluate MCMC-based posterior inference algorithms. More specifically, by running annealed importance sampling (AIS) chains both from prior to posterior and vice versa on simulated data, we obtain bounds on both directions of KL divergence between the true posterior distribution and the distribution of approximate samples. We integrate this algorithm into two probabilistic programming languages: WebPPL and Stan. We discuss methods to validate the relevance of experiments on simulated data to real-world datasets of interest. We have applied our tool to analyze the appropriate choice of model representation in both frameworks and to uncover a previously unknown bug in WebPPL.

#159 Average-case hardness of RIP certification

Tengyao Wang (Univ. of Cambridge)
Quentin Berthet
Yaniv Plan (Univ. of British Columbia)

The restricted isometry property (RIP) for design matrices gives guarantees for optimal recovery in sparse linear models. It is of high interest in compressed sensing and statistical learning. This property is particularly important for computationally efficient recovery methods. As a consequence, even though it is in general NP-hard to check that RIP holds, there have been substantial efforts to find tractable proxies for it. These would allow the construction of RIP matrices and the polynomial-time verification of RIP given an arbitrary matrix. We consider the framework of average-case certifiers, that never wrongly declare that a matrix is RIP, while being often correct for random instances. While there are such functions which are tractable in a suboptimal parameter regime, we show that this is a computationally hard task in any better regime. Our results are based on a new, weaker assumption on the problem of detecting dense subgraphs.

#160 Provable Efficient Online Matrix Completion via Non-convex Stochastic Gradient Descent

Chi Jin (UC Berkeley)
Sham Kakade
Praneeth Netrapalli (Microsoft Research)

Matrix completion, where we wish to recover a low rank matrix by observing a few entries from it, is a widely studied problem in both theory and practice with wide applications. Most of the provable algorithms so far on this problem have been restricted to the offline setting where they provide an estimate of the unknown matrix using all observations simultaneously. However, in many applications, the online version, where we observe one entry at a time and dynamically update our estimate, is more appealing. While existing algorithms are efficient for the offline setting, they could be highly inefficient for the online setting. In this paper, we propose the first provable, efficient online algorithm for matrix completion. Our algorithm starts from an initial estimate of the matrix and then performs non-convex stochastic gradient descent (SGD). After every observation, it performs a fast update involving only one row of two tall matrices, giving near linear total runtime. Our algorithm can be naturally used in the offline setting as well, where it gives competitive sample complexity and runtime to state of the art algorithms. Our proofs introduce a general framework to show that SGD updates tend to stay away from saddle surfaces and could be of broader interests to other non-convex problems.

#161 Infinite Hidden Semi-Markov Modulated Interaction Point Process

matt zhang (Nicta)
Peng Lin (Data61)
Ting Guo (Data61)
Yang Wang (Data61)
Yang Wang (Data61)
Fang Chen (Data61)

The correlation between events is ubiquitous and important for temporal events modelling. In many cases, the correlation exists between not only events' emitted observations, but also their arrival times. State space models (e.g., hidden Markov model) and stochastic interaction point process models (e.g., Hawkes process) have been studied extensively yet separately for the two types of correlations in the past. In this paper, we propose a Bayesian nonparametric approach that considers both types of correlations via unifying and generalizing hidden semi-Markov model and interaction point process model. The proposed approach can simultaneously model both the observations and arrival times of temporal events, and determine the number of latent states from data. A Metropolis-within-particle-Gibbs sampler with ancestor resampling is developed for efficient posterior inference. The approach is tested on both synthetic and real-world data with promising outcomes.

#162 Selective inference for group-sparse linear models

Fan Yang (Univ. of Chicago)
Rina Foygel Barber
Prateek Jain (Microsoft Research)
John Lafferty

We develop tools for selective inference in the setting of group sparsity, including the construction of confidence intervals and p-values for testing selected groups of variables. Our main technical result gives the precise distribution of the magnitude of the projection of the data onto a given subspace, and enables us to develop inference procedures for a broad class of group-sparse selection methods, including the group lasso, iterative hard thresholding, and forward stepwise regression. We give numerical results to illustrate these tools on simulated data and on health record data.

#163 Deep Neural Networks with Inexact Matching for Person Re-Identification

Arulkumar Subramaniam (IIT Madras)
Moitreya Chatterjee (IIT Madras)
Anurag Mittal (IIT Madras)

Person Re-Identification is the task of matching images of a person across multiple camera views. Almost all prior approaches attempt to learn the transformation that relates the different views of a person, from a training corpora. They then utilize this transformation pattern for matching a query image to those in a gallery image bank at test time. This necessitates learning good feature representations of the images and having a robust feature matching technique. Deep learning approaches, such as Convolutional Neural Networks (CNN), simultaneously learn both and have shown great promise recently. In this work, we propose two CNN-based architectures for person re-identification. In the first, given a pair of images, we extract feature maps from these images, via multiple stages of convolution and pooling. A novel inexact matching technique, then matches pixels in the first representation with those of the second. Furthermore, we search across a wider region in the second representation for matching. Our novel matching technique allows us to tackle the challenges posed by large viewpoint variation, illumination change or partial occlusion. Our approach shows a promising performance and requires only about half the parameters as a current state-of-the-art technique. Nonetheless it also suffers from false matches at times. In order to mitigate this issue, we propose a fused architecture that combines our inexact matching pipeline with a state-of-the-art exact matching technique. We observe substantial gains with the fused model over the current state-of-the-art on multiple challenging datasets of varying sizes, with jumps of upto about 21%.

#164 Accelerating Stochastic Composition Optimization

Mengdi Wang
Ji Liu

Consider the stochastic composition optimization problem where the objective is a composition of two expected-value functions. We propose a new stochastic first-order method, namely the accelerated stochastic compositional proximal gradient (ASC-PG) method, which updates based on queries to the sampling oracle using two different timescales. The ASC-PG is the first proximal gradient method for the stochastic composition problem that can deal with nonsmooth regularization penalty. We show that the ASC-PG exhibits faster convergence than the best known algorithms, and that it achieves the optimal sample-error complexity in several important special cases. We further demonstrate the application of ASC-PG to reinforcement learning and conduct numerical experiments.

#165 Learning Bound for Parameter Transfer Learning

Wataru Kumagai (Kanagawa Univ.)

We consider a transfer learning problem by parameter transfer approach, where a suitable parameter of feature mapping is learnd through one task and applied to another objective task. Then, we newly introduce the notion of transfer learnability, and thereby derive a first learning bound for parameter transfer algorithms under the local stability of parametric feature mappings. As an application of parameter transfer learning, we discuss the performance of sparse coding in self-taught learning. Although self-taught learning algorithms with plentiful unlabeled data show good empirical performance, its theoretical analysis has not been studied. In this paper, we give a first theoretical learning bound for self-taught learning under sparse model.

#166 Can Active Memory Replace Attention?

Łukasz Kaiser
Samy Bengio

Several mechanisms to focus attention of a neural network on selected parts of its input or memory have been used successfully in deep learning models in recent years. Attention has improved image classification, image captioning, speech recognition, generative models, and learning algorithmic tasks, but it had probably the largest impact on neural machine translation. Recently, similar improvements have been obtained using alternative mechanisms that do not focus on a single part of a memory but operate on all of it in parallel, in a uniform way. Such mechanism, which we call active memory, improved over attention in algorithmic tasks, image processing, and in generative modelling. So far, however, active memory has not improved over attention for most natural language processing tasks, in particular for machine translation. We analyze this shortcoming in this paper and propose an extended model of active memory that matches existing attention models on neural machine translation and generalizes better to longer sentences. We investigate this model and explain why previous active memory models did not succeed. Finally, we discuss when active memory brings most benefits and where attention can be a better choice.

#167 Understanding the Effective Receptive Field in Deep Convolutional Neural Networks

Wenjie Luo (Univ. of Toronto)
Yujia Li (Univ. of Toronto)
Raquel Urtasun
Richard Zemel

We study characteristics of receptive fields of units in deep convolutional networks. The receptive field size is a crucial issue in many visual tasks, as the output must respond to large enough areas in the image to capture information about large objects. We introduce the notion of an effective receptive field size, and show that it both has a Gaussian distribution and only occupies a fraction of the full theoretical receptive field size. We analyze the effective receptive field in several architecture designs, and the effect of sub-sampling, skip connections, dropout and nonlinear activations on it. This leads to suggestions for ways to address its tendency to be too small.

#168 Local Similarity-Aware Deep Feature Embedding

Chen Huang (Chinese Univ. of HongKong)
Chen Change Loy (The Chinese Univ. of HK)
Xiaoou Tang (The Chinese Univ. of Hong Kong)

Existing deep embedding methods in vision tasks are capable of learning a compact Euclidean space from images, where Euclidean distances correspond to a similarity metric. To make learning more effective and efficient, hard sample mining is usually employed, with samples identified through computing the Euclidean feature distance. However, the global Euclidean distance cannot faithfully characterize the true feature similarity in a complex visual feature space, where the intraclass distance in a high-density region may be larger than the interclass distance in low-density regions. In this paper, we introduce a Position-Dependent Deep Metric (PDDM) unit, which is capable of learning a similarity metric adaptive to local feature structure. The metric can be used to select genuinely hard samples in a local neighborhood to guide the deep embedding learning in an online and robust manner. The new layer is appealing in that it is pluggable to any convolutional networks and is trained end-to-end. Our local similarity-aware feature embedding not only demonstrates faster convergence and boosted performance on two complex image retrieval datasets, its large margin nature also leads to superior generalization results under the large and open set scenarios of transfer learning and zero-shot learning on ImageNet 2010 and ImageNet-10K datasets.

#169 End-to-End Kernel Learning with Supervised Convolutional Kernel Networks

Julien Mairal (Inria)

In this paper, we propose a new image representation based on a multilayer kernel machine that performs end-to-end learning. Unlike traditional kernel methods, where the kernel is handcrafted or adapted to data in an unsupervised manner, we learn how to shape the kernel for a supervised prediction problem. We proceed by generalizing convolutional kernel networks, which originally provide unsupervised image representations, and we derive backpropagation rules to optimize model parameters. As a result, we obtain a new type of convolutional neural network with the following properties: (i) at each layer, learning filters is equivalent to optimizing a linear subspace in a reproducing kernel Hilbert space (RKHS), where we project data; (ii) the networks may be learned with supervision or without; (iii) the model comes with a natural regularization function (the norm in the RKHS). We show that our method achieves reasonably competitive performance on some standard ``deep learning'' image classification datasets such as CIFAR-10 and SVHN, and also state-of-the-art results for image super-resolution, demonstrating the applicability of our approach to a large variety of image-related tasks.

#170 Single-Image Depth Perception in the Wild

Weifeng Chen (Univ. of Michigan)
Zhao Fu (Univ. of Michigan)
Dawei Yang (Univ. of Michigan)
Jia Deng

This paper studies single-image depth perception in the wild, i.e., recovering depth from a single image taken in unconstrained settings. We introduce a new dataset "Depth in the Wild" consisting of images in the wild annotated with relative depth between pairs of random points. We also propose a new algorithm that learns to estimate metric depth using annotations of relative depth. Compared to the state of the art, our algorithm is simpler and performs better. Experiments show that our algorithm, combined with existing RGB-D data and our new relative depth annotations, significantly improves single-image depth perception in the wild.

#171 Toward Deeper Understanding of Neural Networks: The Power of Initialization and a Dual View on Expressivity

Amit Daniely (Google Brain)
Roy Frostig (Stanford Univ.)
Yoram Singer (Google)

We develop a general duality between neural networks and compositional kernel Hilbert spaces. We introduce the notion of a computation skeleton, an acyclic graph that succinctly describes both a family of neural networks and a kernel space. Random neural networks are generated from a skeleton through node replication followed by sampling from a normal distribution to assign weights. The kernel space consists of functions that arise by compositions, averaging, and non-linear transformations governed by the skeleton's graph topology and activation functions. We prove that random networks induce representations which approximate the kernel space. In particular, it follows that random weight initialization often yields a favorable starting point for optimization despite the worst-case intractability of training neural networks.

#172 R-FCN: Object Detection via Region-based Fully Convolutional Networks

jifeng dai (Microsoft)
Yi Li (Tsinghua Univ.)
Kaiming He (Microsoft)
Jian Sun (Microsoft)

We present region-based, fully convolutional networks for accurate and efficient object detection. In contrast to previous region-based detectors such as Fast/Faster R-CNN that apply a costly per-region subnetwork hundreds of times, our region-based detector is fully convolutional with almost all computation shared on the entire image. To achieve this goal, we propose position-sensitive score maps to address a dilemma between translation-invariance in image classification and translation-variance in object detection. Our method can thus naturally adopt fully convolutional image classifier backbones, such as the latest Residual Networks (ResNets), for object detection. We show competitive results on the PASCAL VOC datasets (e.g., 83.6% mAP on the 2007 set) with the 101-layer ResNet. Meanwhile, our result is achieved at a test-time speed of 170ms per image, 2.5-20x faster than the Faster R-CNN counterpart. Code will be made publicly available.

#173 Consistent Estimation of Functions of Data Missing Non-Monotonically and Not at Random

Ilya Shpitser

Missing records are a perennial problem in analysis of complex data of all types, when the target of inference is some function of the full data law. In simple cases, where data is missing at random or completely at random (Rubin, 1976), well-known adjustments exist that result in consistent estimators of target quantities. Assumptions underlying these estimators are generally not realistic in practical missing data problems. Unfortunately, consistent estimators in more complex cases where data is missing not at random, and where no ordering on variables induces monotonicity of missingness status are generally not known. In this paper, we propose a general class of consistent estimators for cases where data is missing not at random, and missingness status is non-monotonic. Our estimators, which are generalized inverse probability weighting estimators, make no assumptions on the underlying full data law, and only assume a joint model of missingness status conditional on the data. The missingness status model we use can be viewed as a version of a conditional Markov random field (MRF) corresponding to a chain graph. The independence assumptions embedded in our model permit identification from the observed data law, and admit a natural fitting procedure based on the pseudo likelihood approach of (Besag, 1975). We illustrate our approach with a simple simulation study.

#174 Without-Replacement Sampling for Stochastic Gradient Methods

Ohad Shamir (Weizmann Institute of Science)

Stochastic gradient methods for machine learning and optimization problems are usually analyzed assuming data points are sampled *with* replacement. In contrast, sampling *without* replacement is far less understood, yet in practice it is very common, often easier to implement, and usually performs better. In this paper, we provide competitive convergence guarantees for without-replacement sampling under several scenarios, focusing on the natural regime of few passes over the data. Moreover, we describe a useful application of these results in the context of distributed optimization with randomly-partitioned data, yielding a nearly-optimal algorithm for regularized least squares (in terms of both communication complexity and runtime complexity) under broad parameter regimes. Our proof techniques combine ideas from stochastic optimization, adversarial online learning and transductive learning theory, and can potentially be applied to other stochastic optimization and learning problems.

#175 Probabilistic Modeling of Future Frames from a Single Image

Tianfan Xue
Jiajun Wu (MIT)
Katie Bouman (MIT)
Bill Freeman

We study the problem of synthesizing a number of likely future frames from a single input image. In contrast to traditional methods, which have tackled this problem in a deterministic or non-parametric way, we propose a novel approach which models future frames in a probabilistic manner. Our proposed method is therefore able to synthesize multiple possible next frames using the same model. Solving this challenging problem involves low- and high-level image and motion understanding for successful image synthesis. Here, we propose a novel network structure, namely a Cross Convolutional Network, that encodes images as feature maps and motion information as convolutional kernels to aid in synthesizing future frames. In experiments, our model performs well on both synthetic data, such as 2D shapes and animated game sprites, as well as on real-wold video data. We show that our model can also be applied to tasks such as visual analogy-making, and present analysis of the learned network representations.

#176 Learning What and Where to Draw

Scott E Reed (Univ. of Michigan)
Zeynep Akata (Max Planck Institute for Informatics)
Santosh Mohan (Univ. of MIchigan)
Samuel Tenka (Univ. of MIchigan)
Bernt Schiele
Honglak Lee (Univ. of Michigan)

Generative Adversarial Networks (GANs) have recently demonstrated the capability to synthesize compelling real-world images, such as room interiors, album covers, manga, faces, birds, and flowers. While existing models can synthesize images based on global constraints such as a class label or caption, they do not provide control over pose or object location. We propose a new model, the Generative Adversarial What-Where Network (GAWWN), that synthesizes images given instructions describing what content to draw in which location. We show high-quality 128 $\times$ 128 image synthesis on the Caltech-UCSD Birds dataset, conditioned on both informal text descriptions and also object location. Our system exposes control over both the bounding box around the bird and its constituent parts. By modeling the conditional distributions over part locations, our system also enables conditioning on arbitrary subsets of parts (e.g. only the beak and tail), yielding an efficient interface for picking part locations.

#177 Stochastic Online AUC Maximization

Yiming Ying
Longyin Wen (State Univ. of New York at Albany)
Siwei Lyu (State Univ. of New York at Albany)

Area under ROC (AUC) is a metric which is widely used for measuring the classification performance for imbalanced data. It is of theoretical and practical interest to develop online learning algorithms that maximizes AUC for large-scale data. A specific challenge in developing online AUC maximization algorithm is that the learning objective function is usually defined over a pair of training examples of opposite classes, and existing methods achieves on-line processing with higher space and time complexity. In this work, we propose a new stochastic online algorithm for AUC maximization. In particular, we show that AUC optimization can be equivalently formulated as a convex-concave saddle point problem. From this saddle representation, a stochastic online algorithm (SOLAM) is proposed which has time and space complexity of one datum. We establish theoretical convergence of SOLAM with high probability and demonstrate its effectiveness and efficiency on standard benchmark datasets.

#178 Deep Learning without Poor Local Minima

Kenji Kawaguchi (MIT)

In this paper, we prove a conjecture published in 1989 and also partially address an open problem announced at the Conference on Learning Theory (COLT) 2015. For an expected loss function of a deep nonlinear neural network, we prove the following statements under the independence assumption adopted from recent work: 1) the function is non-convex and non-concave, 2) every local minimum is a global minimum, 3) every critical point that is not a global minimum is a saddle point, and 4) the property of saddle points differs for shallow networks (with three layers) and deeper networks (with more than three layers). Moreover, we prove that the same four statements hold for deep linear neural networks with any depth, any widths and no unrealistic assumptions. As a result, we present an instance, for which we can answer to the following question: how difficult to directly train a deep model in theory? It is more difficult than the classical machine learning models (because of the non-convexity), but not too difficult (because of the nonexistence of poor local minima and the property of the saddle points). We note that even though we have advanced the theoretical foundations of deep learning, there is still a gap between theory and practice.

#179 Regularized Nonlinear Acceleration

Damien Scieur (INRIA - ENS)
Alexandre d'Aspremont
Francis Bach

We describe a convergence acceleration technique for generic optimization problems. Our scheme computes estimates of the optimum from a nonlinear average of the iterates produced by any optimization method. The weights in this average are computed via a simple and small linear system, whose solution can be updated online. This acceleration scheme runs in parallel to the base algorithm, providing improved estimates of the solution on the fly, while the original optimization method is running. Numerical experiments are detailed on classical classification problems.

#180 Learning to Poke by Poking: Experiential Learning of Intuitive Physics

Pulkit Agrawal (UC Berkeley)
Ashvin V Nair (UC Berkeley)
Pieter Abbeel (OpenAI / UC Berkeley / Gradescope)
Jitendra Malik
Sergey Levine (Univ. of Washington)

We investigate an experiential learning paradigm for acquiring an internal model of intuitive physics. Our model is evaluated on a real-world robotic manipulation task that requires displacing objects to target locations by poking. The robot gathered over 400 hours of experience by executing more than 50K pokes on different objects. We propose a novel approach based on deep neural networks for modeling the dynamics of robot's interactions directly from images, by jointly estimating forward and inverse models of dynamics. The inverse model objective provides supervision to construct informative visual features, which the forward model can then predict and in turn regularize the feature space for the inverse model. The interplay between these two objectives creates useful, accurate models that can then be used for multi-step decision making. This formulation has the additional benefit that it is possible to learn forward models in an abstract feature space and thus alleviate the need of predicting pixels. Our experiments show that this joint modeling approach outperforms alternative methods. We also demonstrate that active data collection using the learned model further improves performance.

#181 Weight Normalization: A Simple Reparameterization to Accelerate Training of Deep Neural Networks

Tim Salimans
Diederik P Kingma

We present weight normalization: a reparameterization of the weight vectors in a neural network that decouples the length of those weight vectors from their direction. By reparameterizing the weights in this way we improve the conditioning of the optimization problem and we speed up convergence of stochastic gradient descent. Our reparameterization is inspired by batch normalization but does not introduce any dependencies between the examples in a minibatch. This means that our method can also be applied successfully to recurrent models such as LSTMs and to noise-sensitive applications such as deep reinforcement learning or generative models, for which batch normalization is less well suited. Although our method is much simpler, it still provides much of the speed-up of full batch normalization. In addition, the computational overhead of our method is lower, permitting more optimization steps to be taken in the same amount of time. We demonstrate the usefulness of our method on applications in supervised image recognition, generative modelling, and deep reinforcement learning.

#182 Linear-Memory and Decomposition-Invariant Linearly Convergent Conditional Gradient Algorithm for Structured Polytopes

Dan Garber
Dan Garber
Ofer Meshi

Recently, several works have shown that natural modifications of the classical conditional gradient method (aka Frank-Wolfe algorithm) for constrained convex optimization, provably converge with a linear rate when the feasible set is a polytope, and the objective is smooth and strongly-convex. However, all of these results suffer from two significant shortcomings: i) large memory requirement due to the need to store an explicit convex decomposition of the current iterate, and as a consequence, large running-time overhead per iteration ii) the worst case convergence rate depends unfavorably on the dimension In this work we present a new conditional gradient variant and a corresponding analysis that improves on both of the above shortcomings. In particular, both memory and computation overheads are only linear in the dimension, and in addition, in case the optimal solution is sparse, the new convergence rate replaces a factor which is at least linear in the dimension in previous works, with a linear dependence on the number of non-zeros in the optimal solution At the heart of our method, and corresponding analysis, is a novel way to compute decomposition-invariant away-steps. While our theoretical guarantees do not apply to any polytope, they apply to several important structured polytopes that capture central concepts such as paths in graphs, perfect matchings in bipartite graphs, marginal distributions that arise in structured prediction tasks, and more. Our theoretical findings are complemented by empirical evidence that shows that our method delivers state-of-the-art performance.

#183 Achieving the KS threshold in the general stochastic block model with linearized acyclic belief propagation

Emmanuel Abbe
Colin Sandon

The stochastic block model (SBM) has long been studied in machine learning and network science as a canonical model for clustering and community detection. In the recent years, new developments have demonstrated the presence of threshold phenomena for this model, which have set new challenges for algorithms. For the {\it detection} problem in symmetric SBMs, Decelle et al.\ conjectured that the so-called Kesten-Stigum (KS) threshold can be achieved efficiently. This was proved for two communities, but remained open from three communities. We prove this conjecture here, obtaining a more general result that applies to arbitrary SBMs with linear size communities. The developed algorithm is a linearized acyclic belief propagation (ABP) algorithm, which mitigates the effects of cycles while provably achieving the KS threshold in O(n \ln n) time. This extends prior methods by achieving universally the KS threshold while reducing or preserving the computational complexity. ABP is also connected to a power iteration method on a generalized nonbacktracking operator, formalizing the spectral-message passing interplay described in Krzakala et al., and extending results from Bordenave et al.

#184 Orthogonal Random Features

Felix X Yu
Ananda Theertha Suresh
Krzysztof M Choromanski
Daniel N Holtmann-Rice
Sanjiv Kumar (Google)

We present an intriguing discovery related to Random Fourier Features: replacing multiplication by a random Gaussian matrix with multiplication by a properly scaled random orthogonal matrix significantly decreases kernel approximation error. We call this technique Orthogonal Random Features (ORF), and provide theoretical and empirical justification for its effectiveness. Motivated by the discovery, we further propose Structured Orthogonal Random Features (SORF), which uses a class of structured discrete orthogonal matrices to speed up the computation. The method reduces the time cost from \mathcal{O}(d^2) to \mathcal{O}(d log d), where d is the data dimensionality, with almost no compromise in kernel approximation quality compared to ORF. Experiments on several datasets verify the effectiveness of ORF and SORF over the existing methods. We also provide discussions on using the same type of discrete orthogonal structure for a broader range of kernels and applications.

#185 Universal Correspondence Network

Christopher B Choy (Stanford Univ.)
Manmohan Chandraker (NEC Labs America)
JunYoung Gwak (Stanford Univ.)
Silvio Savarese (Stanford Univ.)

We present a deep learning framework for accurate visual correspondences and demonstrate its effectiveness for both geometric and semantic matching, spanning across rigid motions to intra-class shape or appearance variations. In contrast to previous CNN-based approaches that optimize a surrogate patch similarity objective, we use deep metric learning to directly learn a feature space that preserves either geometric or semantic similarity. Our fully convolutional architecture, along with a novel correspondence contrastive loss allows faster training by effective reuse of computations, accurate gradient computation through the use of thousands of examples per image pair and faster testing with O(n) feedforward passes for n keypoints, instead of O(n^2) for typical patch similarity methods. We propose a convolutional spatial transformer to mimic patch normalization in traditional features like SIFT, which is shown to dramatically boost accuracy for semantic correspondences across intra-class shape variations. Extensive experiments on KITTI, PASCAL and CUB-2011 datasets demonstrate the significant advantages of our features over prior works that use either hand-constructed or learned features.

#186 The Multiscale Laplacian Graph Kernel

Risi Kondor
Horace Pan (UChicago)

Many real world graphs, such as the graphs of molecules, exhibit structure at multiple different scales, but most existing kernels between graphs are either purely local or purely global in character. In contrast, by building a hierarchy of nested subgraphs, the Multiscale Laplacian Graph kernels (MLG kernels) that we define in this paper can account for structure at a range of different scales. At the heart of the MLG construction is another new graph kernel, called the Feature Space Laplacian Graph kernel (FLG kernel), which has the property that it can lift a base kernel defined on the vertices of two graphs to a kernel between the graphs. The MLG kernel applies such FLG kernels to subgraphs recursively. To make the MLG kernel computationally feasible, we also introduce a randomized projection procedure, similar to the Nystro¨ m method, but for RKHS operators.

#187 Generalization of ERM in Stochastic Convex Optimization: The Dimension Strikes Back

Vitaly Feldman

In stochastic convex optimization the goal is to minimize a convex function F(x) \doteq \E_{f\sim D}[f(x)] over a convex set \K \subset \ R^d where D is some unknown distribution and each f(\cdot) in the support of D is convex over \K. The optimization is based on i.i.d.~samples f^1,f^2,\ldots,f^n from D. A common approach to such problems is empirical risk minimization (ERM) that optimizes F_S(x) \doteq \frac{1}{n}\sum_{i\leq n} f^i(x). Here we consider the question of how many samples are necessary for ERM to succeed and the closely related question of uniform convergence of F_S to F over \K. We demonstrate that in the standard ℓ_p/ℓ_q setting of Lipschitz-bounded functions over a \K of bounded radius, ERM requires sample size that scales linearly with the dimension d. This nearly matches standard upper bounds and improves on Ω(log d) dependence proved for ℓ_2/ℓ_2 setting in (Shalev-Shwartz et al. 2009). In stark contrast, these problems can be solved using dimension-independent number of samples for ℓ_2/ℓ_2 setting and log d dependence for ℓ_1/ℓ_\infty setting using other approaches. We also demonstrate that for a more general class of range-bounded (but not Lipschitz-bounded) stochastic convex programs an even stronger gap appears already in dimension 2.

#188 Large-Scale Price Optimization via Network Flow

Shinji Ito (NEC Coorporation)
Ryohei Fujimaki

This paper deals with price optimization, which is to find the best pricing strategy that maximizes revenue or profit, on the basis of demand forecasting models. Though recent advances in regression technologies have made it possible to reveal price-demand relationship of a number of multiple products, most existing price optimization methods, such as mixed integer programming formulation, cannot handle tens or hundreds of products because of their high computational costs. To cope with this problem, this paper proposes a novel approach based on network flow algorithms. We reveal a connection between supermodularity of the revenue and cross elasticity of demand. On the basis of this connection, we propose an efficient algorithm that employs network flow algorithms. The proposed algorithm can handle hundreds or thousands of products, and returns an exact optimal solution under an assumption regarding cross elasticity of demand. Even in case in which the assumption does not hold, the proposed algorithm can efficiently find approximate solutions as good as can other state-of-the-art methods, as empirical results show.

#189 Bayesian Optimization with Robust Bayesian Neural Networks

Jost Tobias Springenberg (Univ. of Freiburg)
Aaron Klein (Univ. of Freiburg)
Stefan Falkner (Univ. of Freiburg)
Frank Hutter (Univ. of Freiburg)

Bayesian optimization is a prominent method for optimizing expensive to evaluate black-box functions that is prominently applied to tuning the hyperparameters of machine learning algorithms. Despite its successes, the prototypical Bayesian optimization approach - using Gaussian process models - does not scale well to either many hyperparameters or many function evaluations. Attacking this lack of scalability and flexibility is thus one of the key challenges of the field. We present a general approach for using flexible parametric models (neural networks) for Bayesian optimization, staying as close to a truly Bayesian treatment as possible. We obtain scalability through stochastic gradient Hamiltonian Monte Carlo, whose robustness we improve via a scale adaptation. Experiments including multi-task Bayesian optimization with 21 tasks, parallel optimization of deep neural networks and deep reinforcement learning show the power and flexibility of this approach.

#190 Protein contact prediction from amino acid co-evolution using convolutional networks for graph-valued images

Vladimir Golkov (Technical Univ. of Munich)
Marcin J Skwark (Vanderbilt Univ.)
Antonij Golkov (Univ. of Augsburg)
Alexey Dosovitskiy (Univ. of Freiburg)
Thomas Brox (Univ. of Freiburg)
Jens Meiler (Vanderbilt Univ.)
Daniel Cremers (Technical Univ. of Munich)

Proteins are the “building blocks of life”, the most abundant organic molecules, and the central focus of most areas of biomedicine. Protein structure is strongly related to protein function, thus structure

prediction is a crucial task on the way to solve many biological questions. A contact map is a compact representation of the three-dimensional structure of a protein via the pairwise contacts between the amino acid constituting the protein. We use a convolutional network to calculate protein contact maps from inferred statistical coupling between positions in the protein sequence. The input to the network has an image-like structure amenable to convolutions, but every "pixel" instead of color channels contains a bipartite undirected edge-weighted graph. We propose several methods for treating such "graph-valued images" in a convolutional network. The proposed method outperforms state-of-the-art methods by a large margin. It also allows for a great flexibility with regard to the input data, which makes it useful for studying a wide range of problems.

#191 Supervised Word Mover's Distance

Gao Huang
Chuan Guo (Cornell Univ.)
Matt J Kusner
Yu Sun
Fei Sha (Univ. of Southern California)
Kilian Q Weinberger

Accurately measuring the similarity between text documents lies at the core of many real world applications of machine learning. These include web-search ranking, document recommendation, multi-lingual document matching, and article categorization. Recently, a new document metric, the word mover's distance (WMD), has been proposed with unprecedented results on kNN-based document classification. The WMD elevates high quality word embeddings to document metrics by formulating the distance between two documents as an optimal transport problem between the embedded words. However, the document distances are entirely unsupervised and lack a mechanism to incorporate supervision when available. In this paper we propose an efficient technique to learn a supervised metric, which we call the Supervised WMD (S-WMD) metric. Our algorithm learns document distances that measure the underlying semantic differences between documents by leveraging semantic differences between individual words discovered during supervised training. This is achieved with an linear transformation of the underlying word embedding space and tailored word-specific weights, learned to minimize the stochastic leave-one-out nearest neighbor classification error on a per-document level. We evaluate our metric on eight real-world text classification tasks on which S-WMD consistently outperforms almost all of our 26 competitive baselines.

#192 Beyond Exchangeability: The Chinese Voting Process

Moontae Lee (Cornell Univ.)
Jin Jin (Cornell Univ.)
David Mimno (Cornell Univ.)

Many online communities present user-contributed responses, such as reviews of products and answers to questions. User-provided helpfulness votes can highlight the most useful responses, but voting is a social process that can gain momentum based on the popularity of responses and the polarity of existing votes. We propose the Chinese Voting Process (CVP) which models the evolution of helpfulness votes as a self-reinforcing process dependent on position and presentation biases. We evaluate this model on Amazon product reviews and more than 80 StackExchange forums, measuring the intrinsic quality of individual responses and behavioral coefficients of different communities.

#193 Poisson-Gamma dynamical systems

Aaron Schein (UMass Amherst)
Hanna Wallach (Microsoft Research)
Mingyuan Zhou

This paper presents a dynamical system based on the Poisson-Gamma construction for sequentially observed multivariate count data. Inherent to the model is a novel Bayesian nonparametric prior that ties and shrinks parameters in a powerful way. We develop theory about the model's infinite limit and its steady-state. The model's inductive bias is demonstrated on a variety of real-world datasets where it is shown to learn interpretable structure and have superior predictive performance.

#194 Interpretable Distribution Features with Maximum Testing Power

Wittawat Jitkrittum (Gatsby Unit)
Zoltán Szabó
Kacper P Chwialkowski (Gatsby Unit)
Arthur Gretton

Two semimetrics on probability distributions are proposed, given as the sum of differences of expectations of analytic functions evaluated at spatial or frequency locations (i.e, features). The features are chosen so as to maximize the distinguishability of the distributions, by optimizing a lower bound on test power for a statistical test using these features. The result is a parsimonious and interpretable indication of how and where two distributions differ locally. An empirical estimate of the test power criterion converges with increasing sample size, ensuring the quality of the returned features. In real-world benchmarks on high-dimensional text and image data, linear-time tests using the proposed semimetrics achieve comparable performance to the state-of-the-art quadratic-time maximum mean discrepancy test, while returning human-interpretable features that explain the test results.

#195 Dense Associative Memory for Pattern Recognition

Dmitry Krotov (Institute for Advanced Study)
John J. Hopfield (Princeton Neuroscience Institute)

We propose a model of associative memory having an unusual mathematical structure. Contrary to the standard case, which works well only in the limit when the number of stored memories is much smaller than the number of neurons, our model stores and reliably retrieves many more patterns than the number of neurons in the network. We propose a simple duality between this dense associative memory and neural networks commonly used in models of deep learning. On the associative memory side of this duality, a family of models that smoothly interpolates between two limiting cases can be constructed. One limit is referred to as the feature-matching mode of pattern recognition, and the other one as the prototype regime. On the deep learning side of the duality, this family corresponds to neural networks with one hidden layer and various activation functions, which transmit the activities of the visible neurons to the hidden layer. This family of activation functions includes logistics, rectified linear units, and rectified polynomials of higher degrees. The proposed duality makes it possible to apply energy-based intuition from associative memory to analyze computational properties of neural networks with unusual activation functions - the higher rectified polynomials which until now have not been used for training neural networks. The utility of the dense memories is illustrated for two test cases: the logical gate XOR and the recognition of handwritten digits from the MNIST data set.

#196 Relevant sparse codes with variational information bottleneck

Matthew Chalk (IST Austria)
Olivier Marre (Institut de la vision)
Gasper Tkacik (Institute of Science and Technology Austria)

In many applications, it is desirable to extract only the relevant aspects of data. A principled way to do this is the information bottleneck (IB) method, where one seeks a code that maximises information about a relevance variable, Y, while constraining the information encoded about the original data, X. Unfortunately however, the IB method is computationally demanding when data are high-dimensional and/or non-gaussian. Here we propose an approximate variational scheme for maximising a lower bound on the IB objective, analogous to variational EM. Using this method, we derive an IB algorithm to recover features that are both relevant and sparse. Finally, we demonstrate how kernelised versions of the algorithm can be used to address a broad range of problems with non-linear relation between X and Y.

#197 Examples are not enough, learn to criticize! Criticism for Interpretability

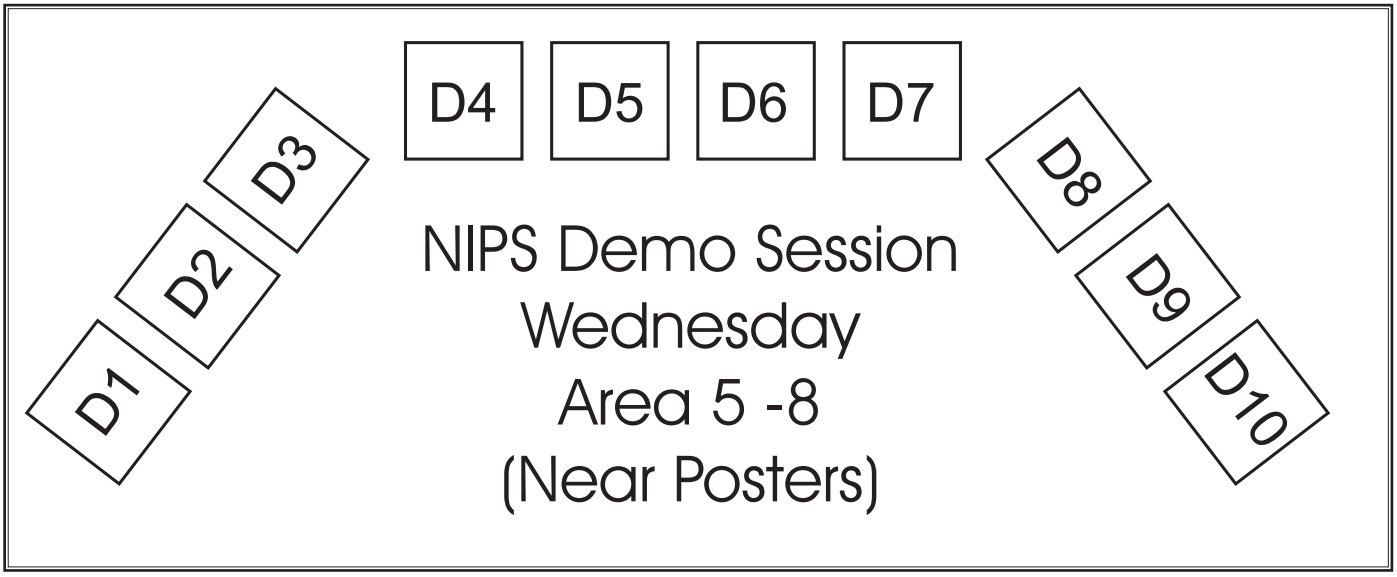
Been Kim
Sanmi Koyejo (UIUC)
Rajiv Khanna (UT Austin)

Example-based explanations are widely used in the effort to improve the interpretability of highly complex distributions. However, prototypes alone are rarely sufficient to represent the gist of the complexity. In order for users to construct better mental models and understand complex data distributions, we also need {\em criticism} to explain what are \textit{not} captured by prototypes. Motivated by the Bayesian model criticism framework, we develop \texttt{MMD-critic} which efficiently learns prototypes and criticism, designed to aid human interpretability. A human subject pilot study shows that the \texttt{MMD-critic} selects prototypes and criticism that are useful to facilitate human understanding and reasoning. We also evaluate the prototypes selected by \texttt{MMD-critic} via a nearest prototype classifier, showing competitive performance compared to baselines.

#198 Showing versus doing: Teaching by demonstration

Mark K Ho (Brown Univ.)
Michael Littman
James MacGlashan (Brown Univ.)
Fiery Cushman (Harvard Univ.)
Joe Austerweil (Univ. of Wisconsin-Madison)
Joe L Austerweil (Univ. of Wisconsin-Madison)

People often learn from others' demonstrations, and classic inverse reinforcement learning (IRL) algorithms have brought us closer to realizing this capacity in machines. In contrast, teaching by demonstration has been less well studied computationally. Here, we develop a novel Bayesian model for teaching by demonstration. Stark differences arise when demonstrators are intentionally teaching a task versus simply performing a task. In two experiments, we show that human participants systematically modify their teaching behavior consistent with the predictions of our model. Further, we show that even standard IRL algorithms benefit when learning from behaviors that are intentionally pedagogical. We conclude by discussing IRL algorithms that can take advantage of intentional pedagogy.



Wed Dec 7th 12:30 -- 3 pm @ Leonardo Da Vinci Square
Live Demo: Detecting Unexpected Obstacles for Self-Driving Cars

Sebastian Ramos
Stefan gehrig
Carsten Rother
Peter Pinggera
Uwe Franke

Our demonstration shows a vision-based system that addresses a challenging and rarely addressed problem for self-driving cars: the detection of generic, small, and unexpected road hazards, such as lost cargo. To the best of our knowledge, our proposed approach to this unsolved problem is the first that leverages both, appearance and contextual cues via a deep convolutional neural network and geometric cues from a stereo-based approach, all combined in a Bayesian framework. Our visual detection framework achieves a very high detection performance with low false positive rates and proves to be robust to illumination changes, varying road appearance as well as 3D road profiles. Our system is able to reliably detect critical obstacles of very low heights (down to 5cm) even at large distances (up to 100m), operating at 22 Hz on our self-driving platform.

D1 Neural Puppet
Tom White

Our proposed work shows NIPS attendees executing various high level manipulations of their own face images using generative neural networks. Portrait photos are encoded into the latent space of a variational autoencoder where attribute vectors can be applied. These include opening and closing the mouth, or adding or removing a smile. Images are then decoded from the latent space and videos are created showing these effects. Additionally, participants can define their own attribute vector by having two photos taken and using the difference between them. This new attribute vector can then be applied to provided reference images as a one-shot generalization.

D2 DeepGTTM-I: local grouping boundary analyzer
Masatoshi Hamanaka

We present a musical analyzer for a generative theory of tonal music (GTTM) that enables us to output the results obtained from analysis that are the similar to those obtained by musicologists on the basis of deep learning by learning the analysis results obtained by musicologists. Directly learning the relationship between an input score and output analysis result is impossible. Therefore, we first constructed a deep belief network (DBN) with latent musical knowledge that could output whether each GTTM rule was applicable or not on each note transition by learning the relationship between the scores and positions of applied grouping preference rules with deep learning. After learning all the grouping preference rules, the network underwent supervised fine-tuning by back propagation using the labeled datasets of local grouping boundaries.

D3 Adapting Microsoft's CNTK and ResNet-18 to Enable Strong-Scaling on Cray Systems
Mark Staveley

Cray XC30 Supercomputer Cray Programming Environment and Performance Tools Cray MPI Libraries Cray Aries Network Microsoft's CNTK modified with Cray MPI Primitives

D4 Automated simulation and replication of fMRI experiments
Leila Wehbe
Alexander G Huth
Fatma Deniz
Marie-Luise Kieseler
Jack Gallant

We present a free online tool that predicts brain activity from text, images and videos that users supply to the system. The tool is based on models that were trained using data from experiments in which subjects process complex, uncontrolled stimuli: they listen to hours of stories and watch real movies. The models can then predict brain activity for any new stimulus, effectively simulating the results of an fMRI experiment that was not performed. Our system allows users to run a variety of common analysis tools used in imaging such as computing contrasts between conditions, running statistical tests, and interactively visualizing the obtained brain maps in multiple views. This can be useful while planning a new fMRI experiment, and could be crucial for encouraging replicability of fMRI results: published results and activation maps can be easily compared against our predictions for the same stimuli.

D5 Adventures with Deep Generator Networks

Jason Yosinski, Anh Nguyen, Jeff Clune, Douglas K Bemis

This demonstration shows a Deep Generator Network (DGN) running live and with its output images used as input for (and optimized for) a few different types of networks: AlexNet / Inception-v4, a Caption convnet, and a Visual Relationship model. Users can investigate properties of these networks by interactively generating their own images using combinations of a webcam to provide input images and keyboard to type captions and relationships, as described below.

D6 Biometric applications of CNNs: get a job at Wupla!

Sergio Escalera, Isabelle Guyon, Baiyu Chen, Marc P Quintana, Umut Güçlü, Yağmur Güçlütürk, Xavier Baró, Rob van Lier, Carlos Andujar, Marcel A. J. van Gerven, Bernhard E Boser, Luke Wang

We show applications of convolutional neural networks to analyze human "imprints", ranging from personality traits to fingerprints.

You have passed all the technical pre-requisites to join "Impending Technologies", a very successful startup, this is your dream job. But wait, now you are asked to be fingerprinted and must go through personality tests to demonstrate "good character". You donate your fingerprint, then you are asked by a friendly avatar to present yourself for 15 seconds in front of a camera. You wait nervously for the outcome of your background check and the analysis of your personality traits. The friendly avatar lets you know that YES, you made it, you have been hired! Now you want to consult your personality trait analysis. We protect your privacy: to access it you must identify yourself with your fingerprint!

D7 Nullhop: Flexibly efficient FPGA CNN accelerator driven by DAVIS neuromorphic vision sensor

Hesham Mostafa, Alessandro Aimar, Enrico Calabrese, Antonio Rios-Navarro, Ricardo Tapiador, Iulia-Alexandra Lungu, Angel F. Jimenez-Fernandez, Federico Corradi, Alejandro Linares-Barranco, Shih-Chii Liu, Tobi Delbruck

CNNs like VGG19, Resnets-50 and GooLeNet show sparsity between 30-90% even after max pooling. During the inference stage, conventional computing architectures such as CPUs and GPUs typically fail to make efficient use of the sparse activations to accelerate the computation. In this demo, we show a novel convolutional neural network accelerator implemented on an FPGA that stores, and operates on, compressed sparse representations of the feature maps. Our implementation improves on state of the art for CNN accelerators that operate with high efficiency across varied kernel sizes and that take advantage of sparsity. In contrast to [1], it never decompresses the features map, so it uses zero clock cycles for zeros in the feature maps. In contrast to [2], its flexibility is maintained across a wide range of numbers of input and output feature maps in layers.

D8 Realistic Virtual Worlds and Human Actions for Video Understanding

César R. De Souza, Adrien Gaidon, Antonio M López

We present a live demo to explore interactively in 3D and in Virtual Reality (VR) our new Virtu-al Human Actions Dataset (VHAD). Virtual Worlds are rapidly gaining momentum as a reliable technique for visual training data generation. This is particularly the case for video, where manual labelling is extremely difficult or even impossible. This scarcity of adequate labeled training data is widely accepted as a major bottleneck of deep learning algorithms for important video under-standing tasks like action recognition. VHAD is a tentative solution to this issue, and consists in using modern game technology (esp. realistic rendering and physics engines) to generate large scale, densely labeled, high-quality syn-thetic video data without any manual intervention. In contrast to approaches using existing video games to record limited data from human game sessions (e.g., [7]), we build upon the more power-ful approach of "virtual world generation" [1,2], which can be seen as making a kind of serious game (dynamic virtual environment) to be played only by (game) AIs in order to generate training data for other (perceptual) AI algorithms. The objective of our demo is to introduce attendees to the benefits of using these realistic virtual worlds, and to allow them to identify new challenges and opportunities, both in terms of research and applications, in particular for action recognition, scene understanding, autonomous driving, deep learning, domain adaptation, multi-task learning, data generation, and related fundamental scientific problems. Our demo lets users navigate through the dynamic virtual worlds used in VHAD using state-of-the-art VR headsets.

D9 Interactive musical improvisation with Magenta

Adam Roberts, Sageev Oore, Curtis Hawthorne, Douglas Eck

Traditional search engines have three main processing phases: crawling, inverted index construction, and candidate documents retrieval. Our retrieval method, instead, has only one: Navigation. We built a neural net based agent that navigates through a website, such as Wikipedia, to find a web page that contains the answer to the question. During training, it learns to assign high probabilities to the hyperlinks that will lead to the correct page and it stops when it is confident that the current page contains the answer to the question.

D10 End-to-End Web Navigation

Rodrigo Frassetto Nogueira

Traditional search engines have three main processing phases: crawling, inverted -index construction, and candidate documents retrieval. Our retrieval method, instead, has only one: Navigation. We built a neural net based agent that navigates through a website, such as Wikipedia, to find a web page that contains the answer to the question. During training, it learns to assign high probabilities to the hyperlinks that will lead to the correct page and it stops when it is confident that the current page contains the answer to the question

THURSDAY SESSIONS

9:00 - 9:50 am - INVITED TALK:
Learning About the Brain: Neuroimaging and Beyond
Irina Rish — Area 1 + 2

9:50 - 10:40 am - INVITED TALK (Brieman Lecture)
Reproducible Research: the Case of the Human Microbiome
Susan Holmes — Area 1 + 2

10:40 - 11:10 am - Coffee Break - P1 & P2

11:10 - 12:10 pm - Interpretable Models - Areas 1 + 2

- **Interpretable Distribution Features with Maximum Testing Power**
  Wittawat Jitkrittum, Zoltán Szabó, Kacper P Chwialkowski, Arthur Gretton
- **Examples Are Not Enough, Learn To Criticize! Criticism for Interpretability**
  Been Kim, Sanmi Koyejo, Rajiv Khanna

11:10 - 12:10 pm - Neuroscience/Cognitive - Area 3

- **Showing versus doing: Teaching by demonstration**
  Mark K Ho, Michael Littman, James MacGlashan, Fiery Cushman, Joe L Austerweil
- **Relevant sparse codes with variational information bottleneck**
  Matthew Chalk, Olivier Marre, Gasper Tkacik
- **Dense Associative Memory for Pattern Recognition**
  Dmitry Krotov, John J. Hopfield

12:10 - 2:00 pm - LUNCH ON YOUR OWN

2:00 - 9:30 pm - SYMPOSIUM

- **Machine Learning and the Law** — Area 3
  Adrian Weller · Thomas D. Grant · Conrad McDonnell · Jatinder Singhi
- **Deep Learning Symposium** — Area 1 + 2
  Yoshua Bengio · Yann LeCun · Navdeep Jaitly · Roger B Grosse
- **Recurrent Neural Networks and Other Machines that Learn Algorithms** — Room 111 + 112
  Juergen Schmidhuber · Sepp Hochreiter · Alex Graves · Rupesh K Srivastava

4:00 - 4:30 pm - Coffee Break - P1 & P2

6:30 - 7:30 pm - Light Dinner

Thursday, Dec 8th, 9 - 9:50 am

Learning About the Brain: Neuroimaging and Beyond

Irina Rish (IBM TJ Watson Research Center)
Area 1 & 2

Quantifying mental states and identifying "statistical biomarkers" of mental disorders from neuroimaging data is an exciting and rapidly growing research area at the intersection of neuroscience and machine learning. Given the focus on gaining better insights about the brain functioning, rather than just learning accurate "black-box" predictors, interpretability and reproducibility of learned models become particularly important in this field. We will discuss promises and limitations of machine learning in neuroimaging, and lessons learned from applying various approaches, from sparse models to deep neural nets, to a wide range of neuroimaging studies involving pain perception, schizophrenia, cocaine addiction and other mental disorders. Moreover, we will also go "beyond the scanner" and discuss some recent work on inferring mental states from relatively cheap and easily collected data, such as speech and wearable sensors, with applications ranging from clinical settings ("computational psychiatry") to everyday life ("augmented human")

Irina Rish is a researcher at the Healthcare & Life Sciences department of IBM T.J. Watson Research Center. She received MS in Applied Mathematics from Moscow Gubkin Institute, Russia, and PhD in Computer Science from the University of California, Irvine. Her areas of expertise include artificial intelligence and machine learning, with a particular focus on probabilistic graphical models, sparsity and compressed sensing, active learning, and their applications to various domains, ranging from diagnosis and performance management of distributed computer systems ("autonomic computing") to predictive modeling and statistical biomarker discovery in neuroimaging and other biological data. Irina has published over 60 research papers, several book chapters, two edited books, and a monograph on Sparse Modeling, taught several tutorials and organized multiple workshops at machine-learning conferences, including NIPS, ICML and ECML. She holds 24 patents and several IBM awards. As an adjunct professor at the EE Department of Columbia University, she taught several advanced graduate courses on statistical learning and sparse signal modeling.

Thursday, Dec 8th, 9:50 - 10:40 am

Reproducible Research: the Case of the Human Microbiome

Susan Holmes
Area 1 & 2

Modern data sets usually present multiple levels of heterogeneity, some apparent such as the necessity of combining trees, graphs, contingency tables and continuous covariates, others concern latent factors and gradients. The biggest challenge in the analyses of these data comes from the necessity to maintain and percolate uncertainty throughout the analyses. I will present a completely reproducible workflow that combines the typical kernel multidimensional scaling approaches with Bayesian nonparametrics to arrive at visualizations that present honest projection regions.

This talk will include joint work with Kris Sankaran, Julia Fukuyama, Lan Huong Nguyen, Ben Callahan, Boyu Ren, Sergio Bacallado, Stefano Favaro, Lorenzo Trippa and the members of Dr Relman's research group at Stanford.

Brought up in the French School of Data Analysis (Analyse des Données) in the 1980's, Professor Holmes specializes in exploring and visualizing complex biological data.

She is interested in integrating the information provided by phylogenetic trees, community interaction graphs and metabolic networks with sequencing data and clinical covariates. She uses computational statistics, and Bayesian methods to draw inferences about many complex biological phenomena such as the human microbiome or the interactions between the immune system and cancer.

She teaches using R and BioConductor and tries to make everything she does freely available.

Sessions 10:40 AM - 12:20 PM

Interpretable Models @ Area 1 + 2

Interpretable Distribution Features with Maximum Testing Power

Wittawat Jitkrittum (Gatsby Unit) Zoltán Szabó
Kacper P Chwialkowski (Gatsby Unit) Arthur Gretton

Two semimetrics on probability distributions are proposed, given as the sum of differences of expectations of analytic functions evaluated at spatial or frequency locations (i.e, features). The features are chosen so as to maximize the distinguishability of the distributions, by optimizing a lower bound on test power for a statistical test using these features. The result is a parsimonious and interpretable indication of how and where two distributions differ locally. An empirical estimate of the test power criterion converges with increasing sample size, ensuring the quality of the returned features. In real-world benchmarks on high-dimensional text and image data, linear-time tests using the proposed semimetrics achieve comparable performance to the state-of-the-art quadratic-time maximum mean discrepancy test, while returning human-interpretable features that explain the test results.

Examples are not enough, learn to criticize! Criticism for Interpretability

Been Kim Sanmi Koyejo (UIUC)
Rajiv Khanna (UT Austin)

Example-based explanations are widely used in the effort to improve the interpretability of highly complex distributions. However, prototypes alone are rarely sufficient to represent the gist of the complexity. In order for users to construct better mental models and understand complex data distributions, we also need {\em criticism} to explain what are \textit{not} captured by prototypes. Motivated by the Bayesian model criticism framework, we develop \texttt{MMD-critic} which efficiently learns prototypes and criticism, designed to aid human interpretability. A human subject pilot study shows that the \texttt{MMD-critic} selects prototypes and criticism that are useful to facilitate human understanding and reasoning. We also evaluate the prototypes selected by \texttt{MMD-critic} via a nearest prototype classifier, showing competitive performance compared to baselines.

Neuroscience/Cognitive @ Area 3

Showing versus doing: Teaching by demonstration

Mark K Ho (Brown Univ.) Michael Littman
James MacGlashan (Brown Univ.) Fiery Cushman (Harvard)
Joe L Austerweil (Univ. of Wisconsin-Madison)

People often learn from others' demonstrations, and classic inverse reinforcement learning (IRL) algorithms have brought us closer to realizing this capacity in machines. In contrast, teaching by demonstration has been less well studied computationally. Here, we develop a novel Bayesian model for teaching by demonstration. Stark differences arise when demonstrators are intentionally teaching a task versus simply performing a task. In two experiments, we show that human participants systematically modify their teaching behavior consistent with the predictions of our model. Further, we show that even standard IRL algorithms benefit when learning from behaviors that are intentionally pedagogical. We conclude by discussing IRL algorithms that can take advantage of intentional pedagogy.

Relevant sparse codes with variational information bottleneck

Matthew Chalk (IST Austria) Olivier Marre (Institut de la vision)
Gasper Tkacik (Institute of Science and Technology Austria)

In many applications, it is desirable to extract only the relevant aspects of data. A principled way to do this is the information bottleneck (IB) method, where one seeks a code that maximises information about a relevance variable, Y, while constraining the information encoded about the original data, X. Unfortunately however, the IB method is computationally demanding when data are high-dimensional and/or non-gaussian. Here we propose an approximate variational scheme for maximising a lower bound on the IB objective, analogous to variational EM. Using this method, we derive an IB algorithm to recover features that are both relevant and sparse. Finally, we demonstrate how kernelised versions of the algorithm can be used to address a broad range of problems with non-linear relation between X and Y.

Dense Associative Memory for Pattern Recognition

Dmitry Krotov (Institute for Advanced Study)
John J. Hopfield (Princeton Neuroscience Institute)

We propose a model of associative memory having an unusual mathematical structure. Contrary to the standard case, which works well only in the limit when the number of stored memories is much smaller than the number of neurons, our model stores and reliably retrieves many more patterns than the number of neurons in the network. We propose a simple duality between this dense associative memory and neural networks commonly used in models of deep learning. On the associative memory side of this duality, a family of models that smoothly interpolates between two limiting cases can be constructed. One limit is referred to as the feature-matching mode of pattern recognition, and the other one as the prototype regime. On the deep learning side of the duality, this family corresponds to neural networks with one hidden layer and various activation functions, which transmit the activities of the visible neurons to the hidden layer. This family of activation functions includes logistics, rectified linear units, and rectified polynomials of higher degrees. The proposed duality makes it possible to apply energy-based intuition from associative memory to analyze computational properties of neural networks with unusual activation functions - the higher rectified polynomials which until now have not been used for training neural networks. The utility of the dense memories is illustrated for two test cases: the logical gate XOR and the recognition of handwritten digits from the MNIST data set.

SYMPOSIUM

THURSDAY SESSIONS: 2 - 9:30 PM

Machine Learning and the Law

Adrian Weller (Univ. of Cambridge)
Thomas D. Grant (Univ. of Cambridge)
Conrad McDonnell (Gray's Inn Tax Chambers)
Jatinder Singh (Univ. of Cambridge)

Advances in machine learning and artificial intelligence mean that predictions and decisions of algorithms are already in use in many important situations under legal or regulatory control, and this is likely to increase dramatically in the near future. Examples include deciding whether to approve a bank loan, driving an autonomous car, or even predicting whether a prison inmate is likely to offend again if released. This symposium will explore the key themes of privacy, liability, transparency and fairness specifically as they relate to the legal treatment and regulation of algorithms and data. Our primary goals are (i) to inform our community about important current and ongoing legislation (e.g. the EU's GDPR https://en.wikipedia.org/wiki/GeneralDataProtection_Regulation which introduces a "right to explanation"); and (ii) to bring together the legal and technical communities to help form better policy in the future.

Deep Learning Symposium

Yoshua Bengio (Univ. of Montreal)
Yann LeCun (New York Univ.)
Navdeep Jaitly (Google Brain)
Roger B Grosse (Univ. of Toronto)

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.

Recurrent Neural Networks and Other Machines that Learn Algorithms

Juergen Schmidhuber (IDSIA)
Sepp Hochreiter (Johannes Kepler Univ.)
Alex Graves (DeepMind)
Rupesh K Srivastava (IDSIA)

Soon after the birth of modern computer science in the 1930s, two fundamental questions arose: 1. How can computers learn useful programs from experience, as opposed to being programmed by human programmers? 2. How to program parallel multiprocessor machines, as opposed to traditional serial architectures? Both questions found natural answers in the field of Recurrent Neural Networks (RNNs), which are brain-inspired general purpose computers that can learn parallel-sequential programs or algorithms encoded as weight matrices.

Our first RNNaissance NIPS workshop dates back to 2003: http://people.idsia.ch/~juergen/rnnaissance.html . Since then, a lot has happened. Some of the most successful applications in machine learning (including deep learning) are now driven by RNNs such as Long Short-Term Memory, e.g., speech recognition, video recognition, natural language processing, image captioning, time series prediction, etc. Through the world's most valuable public companies, billions of people have now access to this technology through their smartphones and other devices, e.g., in the form of Google Voice or on Apple's iOS. Reinforcement-learning and evolutionary RNNs are solving complex control tasks from raw video input. Many RNN-based methods learn sequential attention strategies.

Here we will review the latest developments in all of these fields, and focus not only on RNNs, but also on learning machines in which RNNs interact with external memory such as neural Turing machines, memory networks, and related memory architectures such as fast weight networks and neural stack machines. In this context we will also will discuss asymptotically optimal program search methods and their practical relevance.

Our target audience has heard a bit about recurrent neural networks but will happy to hear again a summary of the basics, and then delve into the latest advanced stuff, to see and understand what has recently become possible. We are hoping for thousands of attendees.

All talks (mostly by famous experts in the field who have already agreed to speak) will be followed by open discussions. We will also have a call for posters. Selected posters will adorn the environment of the lecture hall. We will also have a panel discussion on the bright future of RNNs, and their pros and cons.

WORKSHOPS

FRIDAY WORKSHOPS - 8 am - 6:30 pm

Practical Bayesian Nonparametrics AC Barcelona Room 22
Nick Foti, Tamara Broderick, Trevor Campbell, Michael C. Hughes, Jeff Miller, Aaron Schein, Sinead A Williamson, Yanxun Xu

Interpretable Machine Learning for Complex Systems AC Barcelona, Sagrada Familia
Andrew G Wilson, Been Kim, William Herlands

Deep Reinforcement Learning Area 1
David Silver, Satinder Singh, Pieter Abbeel

Learning in High Dimensions with Structure Area 2
Nikhil Rao, Prateek Jain, Hsiang-Fu Yu, Ming Yuan, Francis Bach

Adversarial Training Area 3
David Lopez-Paz, Leon Bottou, Alec Radford

Nonconvex Optimization for Machine Learning: Theory and Practice Area 5 + 6
Hossein Mobahi, Anima Anandkumar, Percy S Liang, Stefanie Jegelka, Anna E Choromanska

Efficient Methods for Deep Neural Networks
Mohammad Rastegari, Matthieu Courbariaux Area 7 + 8

Intuitive Physics Hilton Diag. Mar, Blrm. C
Adam Lerer, Jiajun Wu, Josh Tenenbaum, Emmanuel Dupoux, Rob Fergus

The Future of Interactive Machine Learning Hilton Diag. Mar, Blrm. A
Kory Mathewson, Kaushik Subramanian, Mark K Ho, Robert Loftin, Joe L Austerweil, Anna Harutyunyan, Doina Precup, Layla El Asri, Matthew Gombolay, Jerry Zhu, Sonia Chernova, Charles L Isbell, Patrick M Pilarski, Weng-Keen Wong, Manuela Veloso, Julie A Shah, Matthew Taylor, Brenna Argall, Michael Littman

Cognitive Computation: Integrating Neural and Symbolic Approaches Hilton Diag. Mar, Blrm. B
Tarek R. Besold, Antoine Bordes, Greg Wayne, Artur Garcez

Extreme Classification: Multi-class and Multi-label Learning in Extremely Large Label Spaces Room 111
Moustapha Cisse, Manik Varma, Samy Bengio

Advances in Approximate Bayesian Inference Room 112
Tamara Broderick, Stephan Mandt, James McInerney, Dustin Tran, David Blei, Kevin P Murphy, Andrew Gelman, Michael I Jordan

Reliable Machine Learning in the Wild Room 113
Dylan Hadfield-Menell, Adrian Weller, David Duvenaud, Jacob Steinhardt, Percy S Liang

Representation Learning in Artificial and Biological Neural Networks Room 114
Leila Wehbe, Marcel Van Gerven, Moritz Grosse-Wentrup, Irina Rish, Brian Murphy, Georg Langs, Guillermo Cecchi, Anwar O Nunez-Elizalde

3D Deep Learning Room 115
Fu Yu, Joseph J Lim, Matt D Fisher, Qixing Huang, Jianxiong Xiao

Machine Learning for Health Room 116
Uri Shalit, Marzyeh Ghassemi, Jason Fries, Rajesh Ranganath, Theofanis Karaletsos, David Kale, Peter Schulam, Madalina Fiterau

Time Series Workshop Room 117
Oren Anava, Marco Cuturi, Azadeh Khaleghi, Vitaly Kuznetsov, Sasha Rakhlin

Crowdsourcing and Machine Learning Room 120 + 121
Adish Singla, Rafael Frongillo, Matteo Venanzi

Adaptive Data Analysis Room 122 + 123
Vitaly Feldman, Aaditya Ramdas, Aaron Roth, Adam Smith

Machine Learning for Intelligent Transportation Systems Room 124 + 125
Li Erran Li, Prof. Darrell

Imperfect Decision Makers: Admitting Real-World Rationality Room 127 + 128
Miroslav Karny, David H Wolpert, David Rios Insua, Tatiana V. Guy

Challenges in Machine Learning: Gaming and Education Room 129 + 130
Isabelle Guyon, Evelyne Viegas, Balázs Kégl, Ben Hamner, Sergio Escalera

Private Multi-Party Machine Learning Room 131 + 132
Borja Balle, Aurélien Bellet, David Evans, Adrià Gascón

Learning, Inference and Control of Multi-Agent Systems Room 133 + 134
Thore Graepel, Marc Lanctot, Joel Z Leibo, Guy Lever, Janusz Marecki, Frans A Oliehoek, Karl Tuyls, Vicky Holgate

Brains and Bits: Neuroscience meets Machine Learning Room 211
Allie Fletcher, Eva L Dyer, Jascha Sohl-Dickstein, Joshua T Vogelstein, Konrad Koerding, Jakob H Macke

Machine Intelligence @ NIPS Room 212
Tomas Mikolov, Baroni Marco, Armand Joulin, Germán Kruszewski, Angeliki Lazaridou, Klemen Simonic

People and machines: Public views on machine learning, and what this means for machine learning researchers VIP Room
Susannah Odell, Peter Donnelly, Jessica Montgomery

Neurorobotics: A Chance for New Ideas, Algorithms and Approaches VIP Room
Elmar Rueckert, Martin Riedmiller

WORKSHOPS

SATURDAY WORKSHOPS - 8 am - 6:30 pm

Bayesian Deep Learning — Area 1
Yarin Gal, Christos Louizos, Zoubin Ghahramani, Kevin P Murphy, Max Welling

Optimizing the Optimizers — Area 2
Maren Mahsereci, Alex J Davies, Philipp Hennig

Deep Learning for Action and Interaction — Area 3
Chelsea Finn, Raia Hadsell, David Held, Sergey Levine, Percy S Liang

Learning with Tensors: Why Now and How? — Area 5 + 6
Anima Anandkumar, Rong Ge, Yan Liu, Maximilian Nickel, Rose Yu

Continual Learning and Deep Networks — Area 7 + 8
Razvan Pascanu, Mark Ring, Tom Schaul

Machine Learning for Spatiotemporal Forecasting — Hilton Diag. Mar, Blrm. B
Florin Popescu, Sergio Escalera, Xavier Baró, Stephane Ayache, Isabelle Guyon

End-to-end Learning for Speech and Audio Processing — Hilton Diag. Mar, Blrm. A
John Hershey, Philemon Brakel

Let's Discuss: Learning Methods for Dialogue — Hilton Diag. Mar, Blrm. C
Hal Daume III, Paul Mineiro, Amanda Stent, Jason E Weston

Large Scale Computer Vision Systems — Room 111
Manohar Paluri, Lorenzo Torresani, Gal Chechik, Dario Garcia, Du Tran

OPT 2016: Optimization for Machine Learning — Room 112
Suvrit Sra, Francis Bach, Sashank J. Reddi, Niao He

Neural Abstract Machines & Program Induction — Room 113
Matko Bošnjak, Nando de Freitas, Tejas D Kulkarni, Arvind Neelakantan, Scott E Reed, Sebastian Riedel, Tim Rocktäschel

Towards an Artificial Intelligence for Data Science — Room 114
Charles Sutton, James Geddes, Zoubin Ghahramani, Padhraic Smyth, Chris Williams

The Future of Gradient-Based Machine Learning Software — Room 115
Alex Wiltschko, Zachary DeVito, Frederic Bastien, Pascal Lamblin

Machine Learning Systems — Room 116
Aparna Lakshmiratan, Li Erran Li, Siddhartha Sen, Sarah Bird, Hussein Mehanna

Bayesian Optimization: Black-box Optimization and Beyond — Room 117
Roberto Calandra, Bobak Shahriari, Javier Gonzalez, Frank Hutter, Ryan P Adams

Adaptive and Scalable Nonparametric Methods in Machine Learning — Room 120 + 121
Aaditya Ramdas, Bharath K. Sriperumbudur, Arthur Gretton, Han Liu, John Lafferty, Samory Kpotufe, Zoltán Szabó

Computing with Spikes — Room 122 + 123
Sander M Bohte, Thomas Nowotny, Cristina Savin, Davide Zambrano

Constructive Machine Learning — Room 127 + 128
Fabrizio Costa, Thomas Gärtner, Andrea Passerini, Francois Pachet

Machine Learning for Education — Room 129 + 130
Richard Baraniuk, Jiquan Ngiam, Christoph Studer, Phillip Grimaldi, Andrew Lan

Connectomics II: Opportunities and Challenges for Machine Learning — Room 131 + 132
Viren Jain, Srini C Turaga

"What If?" Inference and Learning of Hypothetical and Counterfactual Interventions in Complex Systems — Room 133 + 134
Ricardo Silva, John Shawe-Taylor, Adith Swaminathan, Thorsten Joachims

Brains and Bits: Neuroscience meets Machine Learning (2nd day) — Room 211

Machine Learning in Computational Biology — Room 212
Gerald Quon, Sara Mostafavi, James Y Zou, Barbara Engelhardt, Oliver Stegle, Nicolo Fusi

Neurorobotics: A Chance for New Ideas, Algorithms and Approaches (2nd day) — VIP Room

REVIEWERS

Ehsan Abbasi
Yasin Abbasi Yadkori
Sepehr Abbasi Zadeh
Jacob Abernethy
Vineet Abhishek
Sami Abu-El-Haija
Kamal Abuhassan
Massil Achab
Jayadev Acharya
Sreangsu Acharyya
Tudor Achim
Hanno Ackermann
Roy Adams
Ryan Adams
Tameem Adel
Ehsan Adeli
Shivani Agarwal
Amirali Aghazadeh
Zeljko Agic
John Agosta
Pulkit Agrawal
Shipra Agrawal
Sheeraz Ahmad
Amr Ahmed
Farahnaz Ahmed Wick
Sungjin Ahn
James Aimone
Zeynep Akata
Youhei Akimoto
Adedotun Akintayo
Ahmed Alaa
Ibrahim Alabdulmohsin
Xavier Alameda-Pineda
Andrea Albarelli
Hande Alemdar
Alexander Alemi
Alireza Alemi
Alnur Ali
Alexandre Allauzen
Felicity Allen
Zeyuan Allen-Zhu
Robin Allesiardo
Pedro Almagro Blanco
Jesús Alonso
André Altmann
Jaan Altosaar
Jose Alvarez
Mauricio Alvarez
Andrés Álvarez
Mohamed Aly
Carlos Alzate
Mitsuru Ambai
Christophe Ambroise
Mohamed Amer
Ehsan Amid
Kareem Amin
Massih-Reza Amini
Jesse Anderton
Alexandr Andoni
Micha Andriluka
Anelia Angelova
Bhargava Aniruddha
Asha Anoosheh
Luiza Antonie
Piotr Antonik
Mikio Aoi
Ron Appel
Rathinakumar Appuswamy
Forough Arabshahi
Julyan Arbel
David Arbour
Esteban Arcaute
Cedric Archambeau
Evan Archer
Ery Arias-Castro
Raman Arora
Devansh Arpit
Antonio Artes
John Arthur
Timothy Arthur Mann
Thierry Artières
Georgios Arvanitidis
Afsaneh Asaei
Khalid Ashraf
Yannis Assael
Ramon Astudillo
Onur Atan
Awais Athar
Ben Athiwaratkun
Jamal Atif
Mathieu Aubry
Julien Audiffren
Anne Auger
Moritz August
Michael Auli
Joe Austerweil
Konstantin Avrachenkov
Pranjal Awasthi
Yusuf Aytar
Mahdi Azarafrooz
George Azzopardi
Keith B Hall
Mohammad Babaeizadeh
Rohit Babbar
Artem Babenko
Stephen Bach
Olivier Bachem
Philip Bachman
François Bachoc
Pierre-Luc Bacon
Ashwinkumar Badanidiyuru
Mohammad Taha Bahadori
Lu Bai
Song Bai
Wenruo Bai
Xiang Bai
Christian Bailer
Raphaël Bailly
Narayanaswamy Balakrishnan
Sivaraman Balakrishnan
P. Balamurugan
Krishnakumar Balasubramaniam
David Balduzzi
Randall Balestriero
Eric Balkanski
Nicolas Ballas
Borja Balle
Samuel Balmand
Akshay Balsubramani
Tadas Baltrusaitis
Laura Balzano
Afonso Bandeira
Arindam Banerjee
Koyel Banerjee
Monami Banerjee
Prithish Banerjee
Taposh Banerjee
Arjun Bansal
Mohit Bansal
Chenglong Bao
Aharon Bar-Hillel
Omri Barak
Yoseph Barash
Pierre Barbillon
Remi Bardenet
Elias Bareinboim
Oren Barkan
Annalisa Barla
Matt Barnes
Andre Barreto
Simon Bartels
Daniel Bartz
Mehmet Basbug
Giacomo Bassetto
Raef Bassily
Nematollah (Kayhan) Batmanghelı
Dhruv Batra
Peter Battaglia
Josef Bauer
Konstantin Bauman
Stephen Becker
Behrouz Behmardi
Eugene Belilovsky
Vaishak Belle
Pierre Bellec
Marc Bellemare
Aurélien Bellet
Francois Belletti
Boris Belousov
Soukeina Ben Chikha
Nir Ben Zrihem
Alessio Benavoli
Rodrigo Benenson
Samy Bengio
Yoshua Bengio
Yuval Benjamini
Jonathan Berant
Maxime Berar
Philipp Berens
Felix Berkenkamp
Maxim Berman
Gedas Bertasius
Quentin Berthet
Omar Besbes
Michel Besserve
Alex Beutel
Jalaj Bhandari
Aditya Bhaskara
Chiranjib Bhattacharya
Indrajit Bhattacharya
Sharmodeep Bhattacharya
Kshipra Bhawalkar
Srinadh Bhojanapalli
Jinbo Bi
Nan Bi
Yatao Bian
Filippo Maria Bianchi
Michael Biehl
Felix Biessmann
Misha Bilenko
Mustafa Bilgic
Ilias Bilionis
Sarah Bird
George Biros
William Bishop
Anupam Biswas
Mathieu Blondel
Theodore Bluche
Charles Blundell
Joe Bockhorst
Vishnu Boddeti
Joschka Boedecker
Federica Bogo
Sander Bohte
Xavier Boix
Piotr Bojanowski
Martin Boldt
Tolga Bolukbasi
Edwin Bonilla
Antoine Bordes
Karsten Borgwardt
Jorg Bornschein
Diana Borsa
Davide Boscaini
Matko Bosnjak
Giulio Bottegal
Leon Bottou
Haitham Bou Ammar
Diane Bouchacourt
Djamel Bouchaffra
Guillaume Bouchard
Alexandre Bouchard-Côté
Sabri Boughorbel
Abdeslam Boularias
Nicolas Boumal
Y-Lan Boureau
Konstantinos Bousmalis
Christos Boutsidis
Charles Bouveyron
Claire Boyer
Kasia Bozek
Philemon Brakel
Sébastien Bratières
Johanni Brea
Ulf Brefeld
Wieland Brendel
Guy Bresler
Xavier Bresson
Mason Bretan
Mike Brodie
Russell Brooke
James Brooks
Gavin Brown
Frank Broz
Neil Bruce
Michael Brueckner
Joan Bruna
Emma Brunskill
Steven Brunton
Marcus Buckmann
Jakob Buhmann
Thang Bui
Alejandro Bujan
Michael Bukatin
Brian Bullins
Rudy Bunel
Wray Buntine
Yura Burda
Johannes Burge
Michael Burkhart
Evgeny Burnaev
Apostolos Burnetas
Robert Busa-Fekete
Lucian Busoniu
Simon Bussy
Arunkumar Byravan
Danilo Bzdok
Rebecca C. Steorts
Gabriel Cadamuro
Tiberio Caetano
Deng Cai
Jinzheng Cai
Mingbo Cai
Nian Cai
Ruichu Cai
Juan Caicedo
Burak Cakmak
Jeff Calder
Ben Calderhead
Jan-Peter Calliess
Ken Caluwaerts
Ewan Cameron
Raffaello Camoriano
Trevor Campbell
Gustau Camps-Valls
Stephane Canu
Tongyi Cao
Yu Cao
Zhu Cao
Cecile Caponi
Barbara Caputo
Constantine Caramanis
Jaime Carbonell
Cj Carey
David Carlson
Yair Carmon
François Caron
Joao Carreira
De Castro Yohann
Anthony Caterini
Romain Cazé
Miguel Cazorla
Elisa Celis
Nicolò Cesa-Bianchi
Volkan Cevher
Brahim Chaib-Draa
Ayan Chakrabarti
Soumen Chakrabarti
Rudrasis Chakraborty
Chee Seng Chan
Jeffrey Chan
Laiwan Chan
Parag Shridhar Chandakkar
Sarath Chandar
Muthukumaran Chandrasekaran
Kai-Wei Chang
Shawn Chang
Xiaojun Chang
Soravit Changpinyo
Nicolas Chapados
Laetitia Chapel
Adam Charles
Laurent Charlin
Shaunak Chatterjee
Sotirios Chatzis
Arindam Chaudhuri
Sougata Chaudhuri
Abhishek Chaurasia
Zhengping Che
Noshaba Cheema
Oussama Chelly
Bangrui Chen
Beidi Chen
Changyou Chen
Chao Chen
Chen Chen
Dawn Chen
Gang Chen
George Chen
Guoguo Chen
Hao Chen
Hong Chen
Jianshu Chen
Jie Chen
Kevin Chen
Liang Chen
Liang Chen
Liang-Chieh Chen
Lijie Chen
Lin Chen
Ling Chen
Linlin Chen
Lu-Hung Chen
Minhua Chen
Minmin Chen
Po-Hsuan Chen
Qian Chen
Terrence Chen
Wei Chen
Wei Chen
Wei Chen
Weifeng Chen
Xi Chen
Yao Chen
Yen-Chi Chen
Yichen Chen
Yining Chen
Yuansi Chen
Yudong Chen
Yun-Nung Chen
Yunpeng Chen
Yutian Chen
Yuxin Chen
Yuxin Chen
Zhen Chen
Zhitang Chen
Zihao Chen
Ching-An Cheng
De Cheng
Dehua Cheng
Li Cheng
Lu Cheng
Minhao Cheng
Weiwei Cheng
Veronika Cheplygina
Anoop Cherian
Misha Chertkov
Brian Cheung
Wang Chi Cheung
Sylvain Chevallier
Jackie Chi Kit Cheung
Kai-Yang Chiang
Silvia Chiappa
Davide Chicco
Max Chickering
Soumith Chintala
Julien Chiquet
Hamidreza Chitsaz
Hyeokjun Choe
Arthur Choi
Dave Choi
Edward Choi
Jaesik Choi
Jonghyun Choi
Seungjin Choi
Francois Chollet
Nicolas Chopin
Krzysztof Choromanski
Jan Chorowski
Biswarup Choudhury
Yinlam Chow
Girish Chowdhary
Samir Chowdhury
Christopher Choy
Dominique Chu
Hang Chu
Xiao Chu
Jeroen Chua
Yansong Chua
Fu-Lai Chung
Sueyeon Chung
Valentin Churavy
Greg Ciccarelli
Jesús Cid-Sueiro
Carlo Ciliberto
Moustapha Cisse
Philippe Ciuciu
Tom Claassen
Marc Claesen
Stephan Clémencon
Djork-Arné Clevert
Oliver Cliff
Christopher Clingerman
Alexander Cloninger
Edith Cohen
Joseph Paul Cohen
Nadav Cohen
Vincent Cohen-Addad
Olivier Collier
Michael Collins
Nicolo Colombo
Richard Combes
Danilo Comminiello
Alexis Conneau
Tim Cooijmans
Fabrizio Costa
Rui Costa
Omar Costilla-Reyes
Francois Cote
Andy Cotter
Gary Cottrell
Nicolas Courty
Simon Cousins
Benjamin Cowley
Benjamin Cowley
Fabio Cozman
Marco Cristani
Tom Croonenborghs
Balázs Csáji
Dominik Csiba
Adrián Csiszárik
Lixin Cui
Peng Cui
Xiaodong Cui
Zhen Cui
Mark Culp
Eugenio Culurciello
Nguyen Cuong
James Cussens
Marco Cusumano-Towner
Ashok Cutkosky
Marco Cuturi
Robert D. Kleinberg
Jason D. Lee
Dragos D. Margineantu
Hamid Dadkhahi
Andrew Dai
Bo Dai
Hanjun Dai
Lingzheng Dai
Qi Dai
Zhenwen Dai
Onkar Dalal
Oscar Dalmau
Andreas Damianou
Amit Daniely
Ivo Danihelka
Christoph Dann
Abir Das
Mrinal Das
Subhro Das
Gautam Dasarathy
Sanjoy Dasgupta
Sayantan Dasgupta
Jyotikrishna Dass
Jyotishka Datta
Mark Davenport
Ofir David
Jesse Davis
Charles Day
Abir De
Soham De
Bert De Brabandere
Cassio De Campos
Christopher De Sa
Ernesto De Vito
Aurélien Decelle
Moritz Deger
Kévin Degraux
Tom Dela Haije
Erick Delage
Olivier Delalleau
Charles-Alban Deledalle
George Deligiannidis
Nicolas Della Penna
Krzysztof Dembczynski
Jeremiah Deng
Jia Deng
Li Deng
Xiaoming Deng
Zhi-Hong Deng
Misha Denil
Aaron Dennis
Ludovic Denoyer
Stefan Depeweg
Julie Dequaire
Kosta Derpanis
Mattia Desana
Thomas Deselaers
Amit Deshpande
Antoine Desir
Guillaume Desjardins
Alban Desmaison
Christophe Devleeschouwer
Robin Devooght
Biswadip Dey
Debadeepta Dey
Arturo Deza
Sauptik Dhar
Paramveer Dhillon
Dotan Di Castro
Ilias Diakonikolas
Fernando Diaz Ledezma
Hamdi Dibeklioglu
Travis Dick
Luke Dickens
Lee Dicker
Sander Dieleman
Adji Bousso Dieng
Tom Diethe
Thomas Dietterich
Laura Dietz
Christos Dimitrakakis
Jie Ding
Nan Ding
Pak Lun Kevin Ding
Weicong Ding
Xinghao Ding
Yao-Xiang Ding
Yukun Ding
Yuzhen Ding
Zhengming Ding
Laurent Dinh
Vu Dinh
Nemanja Djuric
Nicolas Dobigeon
Elvis Dohmatob
Carlotta Domeniconi
Justin Domke
Frank Dondelinger
Bin Dong
Weisheng Dong
Yinpeng Dong
Alexey Dosovitskiy
Arnaud Doucet
Matthijs Douze
Doug Downey
John Drakopoulos
Anton Dries
Petros Drineas
Alexandre Drouin
Michal Drozdzal
Lan Du
Nan Du
Nan Du
Simon Du
Yang Du
Yong Du
Zhijuan Du
John Duchi
Ambedkar Dukkipati
Gabriel Dulac-Arnold
Lea Duncker
Robert Durrant
Ishan Durugkar
Haimonti Dutta
Sanghamitra Dutta
David Duvenaud
Pavel Dvurechensky
Jennifer Dy
Eva Dyer
Richard E. Turner
Eric Eaton
Elad Eban
Frederick Eberhardt
Imme Ebert-Uphoff
Douglas Eck
Alexander Ecker
Abbas Edalat
Ashley Edwards
Harrison Edwards
Katharina Eggensperger
Reza Eghbali
Hamid Eghbalzadeh
Michael Eickenberg
Jacob Eisenstein
Jason Eisner
Carl Ek
Shady El Damaty
Ethan Elenberg
Ehsan Elhamifar
Matthew Elkherj
Kevin Ellis
Dominik Endres
Alina Ene
Luis Enrique Sucar
Murat Erdogdu
Deniz Erdogmus
Dumitru Erhan
N. Benjamin Erichson
Jeffrey Erlich
Stefano Ermon
Jason Ernst
Hossein Esfandiari
S. M. Ali Eslami
Umberto Esposito
Slim Essid
Seyed Rasoul Etesami
Georgios Evangelopoulos
Robin Evans
Eyrun Eyjolfsdottir
Remi Eyraud
Eduardo F. Morales
Jalal Fadili
Aldo Faisal
Moein Falahatgar
Stefan Falkner
Kai Fan
Xinjie Fan
Xuhui Fan
Yanbo Fan
Ethan Fang
Wen-Chieh Fang
Yili Fang
Amir Massoud Farahmand
Mehrdad Farajtabar
Mahdi Fard
Farzan Farnia
Matteo Fasiolo
Rizal Fathony
Alhussein Fawzi
Valentina Fedorova
Christoph Feichtenhofer
Dan Feldman
Vitaly Feldman
Ian Fellows
Jiashi Feng
Jiashi Feng
Long Feng
Yunlong Feng
Raphaël Féraud
Olivier Fercoq
Kelwin Fernandes
Carlos Fernandez-Granda
Jacobo Fernandez-Vargas
Basura Fernando
Sira Ferradans
Cristian Ferrer
Hamid Feyzmahdavian
Sanja Fidler
Ila Fiete
Mario Figueiredo
Roman Filipovych
Sarah Filippi
Maurizio Filippone
Chelsea Finn
Marcelo Fiori
Amy Fire
Asja Fischer
John Fisher
Madalina Fiterau
Will Fithian
Boris Flach
Peter Flach
Arthur Flajolet
Rémi Flamary
Seth Flaxman
Tom Fletcher
Jakob Foerster
Dylan Foster
Nicholas Foti
James Foulds
Kimon Fountoulakis
Roy Fox
Marco Fraccaro
Edward Frady
Katerina Fragkiadaki
Vojtech Franc
Christopher Frantz
Maia Fraser
Jes Frellsen
Lex Fridman
Johannes Friedrich
Mario Fritz
Elisa Fromont
Magalie Fromont-Renoir
Rafael Frongillo
Pascal Frossard
Huazhu Fu
Qinbing Fu
Tianfan Fu
Xiao Fu
Yanwei Fu
Yun Fu
Kaito Fujii
Yasuhisa Fujii
Ryohei Fujimaki
Kazuto Fukuchi
Kenji Fukumizu
Glenn Fung
Tommaso Furlanello
Nicolo Fusi
Alexandros G. Dimakis
Pablo G. Moreno
Victor Gabillon
Sébastien Gadat
Akshay Gadde
Christian Gagné
Yarin Gal
Jürgen Gall
Matthias Galle
Patrick Gallinari
Aram Galstyan
Lingrui Gan
Zhe Gan
Ashwinkumar Ganesan
Yaroslav Ganin
Elad Ganmor
Ravi Ganti
Sam Ganzfried
Bin Gao
Hongchang Gao
Ji Gao
Jianfeng Gao
Katelyn Gao
Shenghua Gao
Weihao Gao
Yuanjun Gao
Dan Garber
Dario Garcia-Garcia
Guillermo Garcia-Hernando
Ravi Garg
Roman Garnett
Mike Gartrell
Gilles Gasso
Jan Gasthaus
Wolfgang Gatterbauer
Leon Gatys
Romaric Gaudel
Eric Gaussier
Bertrand Gauthier
Matan Gavish
Efstratios Gavves
Hong Ge
Jian Ge
Rong Ge
Wendong Ge
Andreas Geiger
Philipp Geiger
Matthieu Geist
Ian Gemp
Xin Geng
Zhi Geng
Robin Genuer
Petko Georgiev
Samuel Gerber
Sébastien Gerchinovitz
Gaston Gerchkovich
Pascal Germain
Sam Gershman
Pierre Geurts
Mehrdad Ghadiri
Bernard Ghanem
Mohamad Ghavamzadeh
Keyan Ghazi-Zahedi
Shalini Ghosh
Shaona Ghosh
Soumya Ghosh
Souvik Ghosh
Asish Ghoshal
Dimitrios Giannakis
Fabian Gieseke

REVIEWERS

Sébastien Giguère
Ran Gilad Bachrach
Dar Gilboa
Jennifer Gillenwater
Remi Gilleron
Aditya Gilra
David Ginsbourger
Aristides Gionis
Christophe Giraud-Carrier
Ritwik Giri
Raja Giryes
Andrej Gisbrecht
Alex Gittens
Mario Valerio Giuffrida
Ilaria Giulini
Inmar Givoni
Katerina Gkirtzou
Tobias Glasmachers
Jesse Glass
David Gleich
Ben Glocker
Xavier Glorot
Hervé Glotin
Luke Godfrey
Marc Goessling
Georges Goetz
Chong Yang Goh
Nicolas Goix
Jacob Goldberger
Anna Goldenberg
Tom Goldstein
Matthew Golub
Ryan Gomes
Vicenc Gomez
Pedro Goncalves
Alon Gonen
Boqing Gong
Maoguo Gong
Mingming Gong
Pinghua Gong
Yunchao Gong
Fabio A. González
Javier González
Ian Goodfellow
Aditya Gopalan
Yannig Goude
Stephen Gould
Cyril Goutte
Stephan Gouws
Anirudh Goyal
Agnieszka Grabska-Barwinska
Håkan Grahn
Edith Grall-Maes
Robert Gramacy
Alexandre Gramfort
Edouard Grave
Mihajlo Grbovic
David Greenberg
Kristjan Greenewald
Ed Grefenstette
Klaus Greff
Karol Gregor
Russ Greiner
Remi Gribonval
Joshua Griffin
Tom Griffiths
Yuri Grinberg
Perry Groot
Roger Grosse
Aditya Grover
Steffen Grunewalder
Peter Grunwald
Audrunas Gruslys
Chongyang Gu
Quanquan Gu
Sergio Guadarrama
Ziyu Guan
Benjamin Guedj
Florimond Guéniat
Yann Guermeur
Arthur Guez
Sudipto Guha
Abhradeep Guha Thakurta
Vincent Guigue
Agathe Guilloux
Caglar Gulcehre
San Gultekin
Suriya Gunasekar
Aditya Guntuboyina
Chuan Guo
Han Guo
Hongyu Guo
Shangqi Guo
Xiaojie Guo
Xin Guo
Yanqing Guo
Yiwen Guo
Yuhong Guo
Maya Gupta
Saurabh Gupta
Sunil Gupta
Swati Gupta
Vishal Gupta
Mert Gürbüzbalaban
Todd Gureckis
Gleb Gusev
Eli Gutin
Michael Gutmann
Tatiana Guy
Cristóbal Guzmán
Junyoung Gwak
András György
Minh Ha Quang
Tuomas Haarnoja
Amaury Habrard
Dylan Hadfield-Menell
Raia Hadsell
Benjamin Haeffele
Saeid Haghighatshoar
Morteza Haghir Chehreghani
Paul Hahn
Davood Hajinezhad
Hannaneh Hajishirzi
Dilek Hakkani-Tür
Assaf Hallak
Jonas Hallgren
Mark Hamilton
Jihun Hamm
Barbara Hammer
Fred Hamprecht
Onur Hamsici
Bo Han
Bohyung Han
Jun Han
Lei Han
Shizhong Han
Song Han
Ankur Handa
Josiah Hanna
Lauren Hannah
Steve Hanneke
Sofie Hansen
Satoshi Hara
Tatsuya Harada
Mehrtash Harandi
Bharath Hariharan
Alena Harley
Kameron Harris
Jason Hartford
Anna Harutyunyan
David Harwath
Sadid Hasan
Leonard Hasenclever
Hamed Hassani
Negar Hassanpour
Kohei Hatano
Søren Hauberg
Jarvis Haupt
Anne-Claire Haury
Mike Hawrylycz
Kohei Hayashi
Zeeshan Hayder
Tamir Hazan
Bryan He
He He
Hi He
Hu He
Jingrui He
Kaiming He
Niao He
Ru He
Tong He
Xinran He
Zhenyu He
Jennifer Healey
Mohamed Hebiri
Yotam Hechtlinger
Reinhard Heckel
Nicolas Heess
Chinmay Hegde
Eric Heim
Markus Heinonen
Ruth Heller
Mikael Henaff
Ricardo Henao
Joao Henriques
James Hensman
Romain Herault
Mark Herbster
Tue Herlau
Daniel Hernandez Lobato
Ivan Herreros-Alonso
John Hershey
Matteo Hessel
Todd Hester
Irina Higgins
Jun-Ichiro Hirayama
Michael Hirsch
Satoru Hiwa
Rex Devon Hjelm
Chien-Ju Ho
Chin Pang Ho
Mark Ho
Minh Hoai
Trong Nghia Hoang
Toby Hocking
Judy Hoffman
Matt Hoffman
Katja Hofmann
Michael Hofmann
Thomas Hofmann
David Hofmeyr
Steven Hoi
William Hoiles
Dan Holtmann-Rice
Junya Honda
Paul Honeine
Chun Yu Hong
Jungpyo Hong
Mingyi Hong
Yi Hong
Antti Honkela
Jean Honorio
Thibaut Horel
Inbal Horev
Timothy Hospedales
Mohammad Javad Hosseini
Reshad Hosseini
Chenping Hou
Ming Hou
Michael Houle
Rein Houthooft
Dirk Hovy
Cho-Jui Hsieh
Ya-Ping Hsieh
Daniel Hsu
Bin Hu
Changwei Hu
Jinli Hu
Shell Hu
Wei Hu
Xiaolin Hu
Yang Hu
Zhe Hu
Zhiting Hu
Bert Huang
Biwei Huang
Chen Huang
Dingjiang Huang
Furong Huang
Gao Huang
He Huang
Heng Huang
Jia-Bin Huang
Jiaji Huang
Jiapei Huang
Jim Huang
Jonathan Huang
Junzhou Huang
Kejun Huang
Liang Huang
Longbo Huang
Po-Sen Huang
Ruitong Huang
Sheng-Jun Huang
Shuai Huang
Tzu-Kuo Huang
Weiran Huang
Wen Huang
Xuhui Huang
Yanlong Huang
Zhiwu Huang
Itay Hubara
Michael Hughes
Andreas Hula
Jan Humplik
Eric Hunsberger
Jonathan Hunt
Zhouyuan Huo
Michael Hurley
Vahan Huroyan
Ferenc Huszar
Michael Hütel
Marcus Hutter
Jan-Christian Hütter
Vân Huynh-Thu
Young Hwan Chang
Kyuyeon Hwang
Sung Jin Hwang
Youngha Hwang
Alexandre Hyafil
Stephanie Hyland
Antti Hyttinen
Aapo Hyvärinen
Mesrob I. Ohannessian
Forrest Iandola
Yasutoshi Ida
Eugene Ie
Christian Igel
Laura Igual
Alexander Ihler
Hal Iii
Daiki Ikami
Ilija Ilievski
Cindy Im
Jiwoong Im
Masaaki Imaizumi
Kwang In Kim
John Ingraham
Stratis Ioannidis
Catalin Ionescu
Geoffrey Irving
Atil Iscen
Masato Ishii
Phillip Isola
Olga Isupova
Vamsi Ithapu
Shinji Ito
Franck Iutzeler
Masakazu Iwamura
Tomoharu Iwata
Matthew J. W. Howard
Maryam Jaberi
Peter Jacko
Laurent Jacob
Max Jaderberg
Herbert Jaeger
Martin Jaggi
Patrick Jähnichen
Arjun Jain
Brijnesh Jain
Lalit Jain
Mihir Jain
Shanatnu Jain
Swayambhoo Jain
Ragesh Jaiswal
Amin Jalali
Amin Jalali
Kevin Jamieson
Vijay Janakiraman
Jeremy Jancsary
Hyeryung Jang
Benjamin Jantzen
Nathalie Japkowicz
Fethi Jarray
Adel Javanmard
Mehran Javanmardi
Kashif Javed
Pratik Jawanpuria
Neal Jean
Bruno Jedynak
Vladisav Jelisavcic
Rodolphe Jenatton
Bjørn Jensen
Ghassen Jerfel
Abhay Jha
Hui Ji
Qiang Ji
Shihao Ji
Shuiwang Ji
Kui Jia
Xu Jia
Yangqing Jia
Yantao Jia
Bai Jiang
Bingbing Jiang
Binyan Jiang
Biye Jiang
Bo Jiang
Duo Jiang
Heinrich Jiang
Hui Jiang
Ke Jiang
Meng Jiang
Wenhao Jiang
Zhewei Jiang
Zequn Jie
Danilo Jimenez Rezende
Chi Jin
Junqi Jin
Ming Jin
Peter Jin
Rong Jin
Liping Jing
Yu Jing
Wittawat Jitkrittum
Fredrik Johansson
Ramesh Johari
Matthew Johnson
Nicholas Johnson
Tyler Johnson
Matthew Joseph
Armand Joulin
Jérémie Jozefowiez
Anatoli Juditsky
Felix Juefei-Xu
Taeho Jung
Giuseppe Jurman
Preethi Jyothi
Michalis K. Titsias
Ata Kaban
Jad Kabbara
Maya Kabkab
Nirag Kadakia
Asim Kadav
Jonathan Kadmon
Hachem Kadri
Kushal Kafle
Gregory Kahn
Lars Kai Hansen
Łukasz Kaiser
Nal Kalchbrenner
Nathan Kallus
Alexandros Kalousis
Jayashree Kalpathy-Cramer
Shivaram Kalyanakrishnan
Tin Kam Ho
Parameswaran Kamalaruban
Gautam Kamath
Nitin Kamra
Varun Kanade
Motonobu Kanagawa
Pallika Kanani
Kirthevasan Kandasamy
Melih Kandemir
Vasilis Kandylas
Kai Kang
Keegan Kang
Tae Seung Kang
Yang Kang
Ingmar Kanitscheider
Sreeram Kannan
Purushottam Kar
Theofanis Karaletsos
Nikos Karampatziakis
Masayuki Karasuyama
Amin Karbasi
Maximilian Karl
Zohar Karnin
Miroslav Karny
Andrej Karpathy
Senanayak Sesh Kumar Karri
Shiva Kasiviswanathan
Michael Katehakis
Tarun Kathuria
Dimitrios Katselis
Gaurav Kaul
Yoshinobu Kawahara
Noriaki Kawamae
Ken-Ichi Kawarabayashi
Tomas Kazmar
Sander Keemink
Michael Kemmler
Kristian Kersting
Azadeh Khaleghi
Koosha Khalvati
Ahmed Shehab Khan
Nishith Khandwala
Rajiv Khanna
Daniel Khashabi
Ashish Khetan
Justin Khim
Khashayar Khosravi
Tushar Khot
Rahul Kidambi
Been Kim
Donghyun Kim
Dongwoo Kim
Gyuwan Kim
Hanjoo Kim
Hyunwoo Kim
Jisu Kim
Junmo Kim
Kee-Eung Kim
Myunghwan Kim
Saehoon Kim
Seyoung Kim
Tae-Kyun Kim
Taehwan Kim
Taesup Kim
Yong-Deok Kim
Yongdai Kim
Stefan Kinauer
Pieter-Jan Kindermans
Diederik Kingma
Brian Kingsbury
Katherine Kinnaird
Franz Kiraly
Alexander Kirillov
Jyrki Kivinen
Negar Kiyavash
Diego Klabjan
Arto Klami
Aaron Klein
Samantha Kleinberg
Maria Klodt
Marius Kloft
Jan Klopp
Olga Klopp
Christian Knoll
David Knowles
Boris Knyazev
Dmitry Kobak
Murat Kocaoglu
Mykel Kochenderfer
Sokol Koco
Toshiaki Koike-Akino
Mikko Koivisto
Iasonas Kokkinos
Mladen Kolar
Morten Kolbæk
Ritesh Kolte
Zico Kolter
Vladlen Koltun
Junpei Komiyama
Kishore Konda
Jakub Konecny
Deguang Kong
Linglong Kong
Xiangnan Kong
Aryeh Kontorovich
Ender Konukoglu
Jaehoon Koo
Julian Kooij
Irena Koprinska
Anoop Korattikara
Frédéric Koriche
Ramya Korlakai Vinayak
Urs Koster
Wojciech Kotlowski
Lars Kotthoff
Margarita Kotti
Satwik Kottur
Jan Koutnik
Adriana Kovashka
Sanmi Koyejo
Mark Kozdoba
Peter Krafft
Philipp Krähenbühl
Felix Krahmer
Jonathan Krause
Oswin Krause
Stefan Kremer
Georg Krempl
Walid Krichene
Nils Kriege
Akshay Krishnamurthy
Dilip Krishnan
Vivek Krishnan
Smita Krishnaswamy
Oliver Kroemer
Dmitry Krotov
Gabriel Krummenacher
Florent Krzakala
Kun Kuang
Yotaro Kubo
Alp Kucukelbir
Volodymyr Kuleshov
Alex Kulesza
Brian Kulis
Tejas Kulkarni
Meelis Kull
Wataru Kumagai
Anshul Kundaje
Alex Kurakin
Manabu Kuroki
Evgeny Kusmenko
Danil Kuzin
Vitaly Kuznetsov
Branislav Kveton
Nojun Kwak
Roland Kwitt
Heesung Kwon
Anastasios Kyrillidis
Jakub Łącki
Alexandre Lacoste
Jean Lafond
Paul Lagrée
Subhaneil Lahiri
Wei-Sheng Lai
Himabindu Lakkaraju
Balaji Lakshminarayanan
Chandrashekar Lakshminarayanan
Antony Lam
Remi Lam
Luc Lamontagne
Sylvain Lamprier
Andrew Lan
Guanghui Lan
Marc Lanctot
Adrian Lancucki
Nicholas Lane
Agata Lapedriza
Romain Laroche
Gustav Larsson
Kenneth Latimer
Pierre Latouche
Tor Lattimore
Chandrashekhar Lavania
Niklas Lavesson
Marc Law
Miguel Lázaro Gredilla
Khanh Hien Le
Ngan Le
Trung Le
Géraud Le Falher
Marine Le Morvan
Nicolas Le Roux
Anthony Lee
Christina Lee
Edward Lee
Jaekoo Lee
John Lee
Jongmin Lee
Kanghoon Lee
Kuang-Chih Lee
Minsik Lee
Moontae Lee
Namhoon Lee
Seunghak Lee
Su-In Lee
Yin-Tat Lee
Young Lee
Leonidas Lefakis
Robert Legenstein
Andreas Lehrmann
Eric Lei
Lihua Lei
Qi Lei
Yunwen Lei
Guy Leibovitz
Benedict Leimkuhler
Scott Leishman
Jose Leiva
Marc Lelarge
Victor Lempitsky
Vincent Lepetit
Nathan Lepora
Mathieu Lerasle
Adam Lerer
Gilad Lerman
Jonathan Leroux
Thibault Lesieur
David Leslie
Alex Leung
Sergey Levine
Ludmila Levkova
Clément Levrard
Kfir Levy
Andrew Li
Bin Li
Bo Li
Changsheng Li
Chenglong Li
Chengtao Li
Chun-Liang Li
Chunyuan Li
Dangna Li
Dong Li
Dongsheng Li
Fuxin Li
Hai Li
Hao Li
Hongwei Li
Huan Li
Huibin Li
Jerry Li
Jian Li
Jianguo Li
Jun Li
Junzhi Li
Ke Li
Lei Li
Meng Li
Ming Li
Ming Li
Ping Li
Ping Li
Qi Li
Rui Li
Ruiyu Li
Shao-Yuan Li
Shuai Li
Shuai Li
Songze Li
Steven Cheng-Xian Li
Tianyang Li
Wei Li
Weite Li
Weixin Li
Wenye Li
Wu-Jun Li
Xiangyang Li
Xiao Li
Xiaocheng Li
Ximing Li
Xingguo Li
Xudong Li
Yangyan Li
Yanjun Li
Yifeng Li
Yijun Li
Yingzhen Li
Yitan Li
Yixuan Li
Yu-Feng Li
Yuanlong Li
Yujia Li
Yujun Li
Zhe Li
Zhen Li
Zhenguo Li
Zhenhua Li
Zhiyuan Li
Zhizhong Li
Zhuoru Li
Xiangru Lian
Dawen Liang
Jian Liang
Jingwei Liang
Kun Liang
Xiaodan Liang
Yingyu Liang
Renjie Liao
Rui Liao
Xuejun Liao
Edo Liberty
Jan Malte Lichtenberg
Thomas Liebig
Katrina Ligett
Toby Lightheart
Timothy Lillicrap
Joseph Lim
Kar Wai Lim
Néhémy Lim
Woosang Lim
Zhan Wei Lim
Chih-Jen Lin
Hsuan-Tien Lin
Junhong Lin
Liang Lin
Min Lin
Ming Lin
Qihang Lin
Shou-De Lin
Tiger Lin
Tsung-Han Lin
Zhihui Lin
Zhouchen Lin
Zhouhan Lin
Scott Linderman
Erik Lindgren
Fredrik Lindsten
Julia Ling
Erik Linstead
Anna Little
Baoyuan Liu
Bo Liu
Chang Liu
Chang Liu
Chao Liu
Chenxi Liu
Dong Liu
Fei Liu
Guangcan Liu
Han Liu
Hantang Liu
Hanxiao Liu
Hongfu Liu
Ji Liu
Ji Liu
Jie Liu
Jinglan Liu
Jun Liu
Miao Liu
Ming-Yu Liu
Ping Liu
Shicong Liu
Shih-Chii Liu
Shu Liu
Sifei Liu
Song Liu
Sulin Liu
Tie-Yan Liu
Wei Liu
Weiwei Liu
Weiyang Liu
Xianming Liu
Xinwang Liu
Yan Liu
Yang Liu
Yashu Liu
Yi-Kai Liu
Yuanchao Liu
Yuhang Liu
Yunlong Liu
Yunshu Liu
Zhe Liu
Zhuanghua Liu
Ziqi Liu
Jesse Livezey
Lorenzo Livi
Roi Livni
Dan Lizotte
Andrey Lokhov
Hervé Lombaert
Maria Lomeli
Mingsheng Long
Marco Loog
Manuel Lopes
Adam Lopez
Antonio Lopez
David Lopez-Paz
Kin Lore
Marco Lorenzi
Tania Lorido-Botran
Ilya Loshchilov
Xinghua Lou
Gilles Louppe
Sébastien Loustau
Bryan Kian Hsiang Low

REVIEWERS

Daniel Lowd
Ryan Lowe
Chen Change Loy
Aurelie Lozano
José Lozano
Cewu Lu
Chaochao Lu
Chi-Jen Lu
Haiping Lu
Hongtao Lu
Jiasen Lu
Jin Lu
Jiwen Lu
Ming Lu
Peng Lu
Songtao@lastate.Edu Lu
Wei Lu
Xin Lu
Yao Lu
Yu Lu
Zhongqi Lu
Aurelien Lucchi
Mario Lucic
Elliot A. Ludvig
Jörg Lücke
Marti Lukac
Gediminas Luksys
Haipeng Luo
Haipeng Luo
Luo Luo
Minnan Luo
Ping Luo
Ping Luo
Wenjie Luo
Xiangyu Luo
Yan Luo
Khoa Luu
Jinglei Lv
Shaogao Lv
Christopher Lynn
Wouter M. Koolen
Cong Ma
Jing Ma
Keng Teck Ma
Ning Ma
Qianli Ma
Shiqian Ma
Shuai Ma
Xuezhe Ma
Yifei Ma
Yingju Ma
Lars Maaløe
Andrew Maas
Omid Madani
Takanori Maehara
Sara Magliacane
Alessandro Magnani
Sridhar Mahadevan
Dhruv Mahajan
Mehrdad Mahdavi
Hessam Mahdavifar
Mohammad Mahdian
Niru Maheswaranathan
A. Rupam Mahmood
Michael Mahoney
Ngoc Mai Tran
Odalric Maillard
Michael Maire
Jeremy Maitin-Shepard
Rajkumar Maity
Subhransu Maji
Konstantin Makarychev
Rahul Makhijani
Alan Malek
Mateusz Malinowski
Tomasz Malisiewicz
Rakesh Malladi
Jesus Malo
Brandon Malone
Antoine Mandel
Travis Mandel
Stephan Mandt
Daniel Mankowitz
Hassan Mansour
Yishay Mansour
Junhua Mao
Qi Mao
Xueyu Mao
Onaiza Maqbool
William March
Mario Marchand
Etienne Marcheret
Jakub Marecek
Radu Marinescu
Natasha Markuzon
Benjamin Marlin
Andre Marquand
Gautier Marti
James Martin
Trevor Martin
José Martín-Guerrero
Aleix Martinez
Manuel Martinez
Ruben Martinez-Cantin
Gonzalo Martínez-Muñoz
Luca Martino
Andre Martins
Georg Martius
Pekka Marttinen
Jérémie Mary
Tristan Mary-Huard
Jonathan Masci
Andrew Massimino
Lionel Mathelin
Kory Mathewson
Michael Mathieu
Catherine Matias
Yoshitatsu Matsuda
Tetsu Matsukawa
Loic Matthey
Tyler Maunu
Andreas Maurer
Jonathan May
N. Michael Mayer
Lucas Maystre
Arya Mazumdar
David Mcallester
Rowan Mcallister
Julian Mcauley
Calvin Mccarter
Michael Mccourt
Erik Mcdermott
Brian Mcfee
Ian Mcgraw
Sean Mcgregor
Kevin Mcguinness
James Mcinerney
Lane Mcintosh
Brendan Mcmahan
Daniel Mcnamee
Scott Mcquade
James Mcqueen
Brian Mcwilliams
Wannes Meert
Nishant Mehta
Song Mei
Eli Meirom
Yaron Meirovitch
Raghu Meka
Tim Melano
Talya Meltzer
Facundo Memoli
Gonzalo Mena
De Meng
Zibo Meng
Aditya Menon
Bjoern Menze
Farnood Merrikh Bayat
Ronaldo Messina
Florian Metze
Elad Mezuman
Antonio Valerio Miceli Barone
Pierre Michaud
Bertrand Michel
Martin Miguel
José Miguel Hernández-Lob
Lyudmila Mihaylova
Anton Milan
Brian Milch
Andrew Miller
David Mimno
Martin Renqiang Min
Seonwoo Min
Kentaro Minami
Paul Mineiro
Kean Ming Tan
Tom Minka
Nina Miolane
Benjamie Mirabelli
Bamdev Mishra
Rajat Mishra
Vinith Misra
Ioannis Mitliagkas
Ritwik Mitra
Hideyuki Miyahara
Martin Mladenov
Wiktor Mlynarski
Andriy Mnih
Volodymyr Mnih
Hossein Mobahi
Decebal Constantin Mocanu
Soheil Mohajer
Karthik Mohan
Karthika Mohan
Mahesh Mohan
Aryan Mokhtari
Marco Mondelli
Rajat Monga
Gregoire Montavon
Lucie Montuelle
Guido Montufar
Hyungil Moon
Kevin Moon
Taesup Moon
Cristopher Moore
Michael Morehead
Alexander Moreno
Francesc Moreno-Noguer
Jamie Morgenstern
Clayton Morrison
Emilie Morvant
Amit Moscovich
Benjamin Moseley
Thiago Mosqueiro
Hesham Mostafa
Sara Mostafavi
Roozbeh Mottaghi
Lili Mou
Xenia Mountrouidou
Janaina Mourao-Miranda
Mahta Mousavi
Youssef Mroueh
Nicole Mücke
Mayur Mudigonda
Christian Mueller
Jonas Mueller
Eran Mukamel
Nabanita Mukherjee
Soumendu Sundar Mukherjee
Klaus-Robert Muller
Lorenz Muller
Lyle Muller
Andreas Müller
Katharina Mülling
Andres Munoz
Enrique Munoz Ce Cote
Calvin Murdock
Vittorio Murino
Keerthiram Murugesan
Cameron Musco
Christopher Musco
Boaz Nadler
Shinichi Nakajima
Atsuyoshi Nakamura
Masahiro Nakano
Ndapandula Nakashole
Yuji Nakatsukasa
Eric Nalisnick
Vinay Namboodiri
Hongseok Namkoong
Feng Nan
Hemachandra Nandyala
Karthik Narasimhan
Akshay Narayan
Arun Narayanan
Houssam Nassif
Nagarajan Natarajan
Saketha Nath
Humberto Naves
Fatemeh Navidi
Seyedehsara Nayer
Eugene Ndiaye
Deanna Needell
Daniel Neil
Willie Neiswanger
Jelani Nelson
Aida Nematzadeh
Gergely Neu
Graham Neubig
Gerhard Neumann
Matey Neykov
Behnam Neyshabur
Bernard Ng
Joe Yue-Hei Ng
Yin Cheng Ng
Vien Ngo
Anh Nguyen
Hao Nguyen
Huy Nguyen
Quang Nguyen
Quynh Nguyen
Tan Nguyen
Vu Nguyen
Xuan Vinh Nguyen
Maximilian Nickel
Hannes Nickisch
Maria-Irina Nicolae
Mihalis Nicolaou
Alexandru Niculescu-Mizil
Feiping Nie
Siqi Nie
Juan Carlos Niebles
Mathias Niepert
Hirotaka Niitsuma
Ioannis Nikolentzos
Daniel Nikovski
Guanghan Ning
Yang Ning
Robert Nishihara
Atsushi Nitanda
Tohru Nitta
Gang Niu
Lingfeng Niu
Mu Niu
Bill Noble
Richard Nock
Hyeonwoo Noh
Yung-Kyun Noh
Mohammad Norouzi
Ehimwenma Nosakhare
Robert Nowak
Mariusz Nowostawski
Sebastian Nowozin
Rebecca Nugent
Masayuki Numao
Timothy O'Shea
Chris Oates
Thomas Oberlin
Oliver Obst
Jean-Marc Odobez
Dino Oglic
Barlas Oğuz
Sechan Oh
Sewoong Oh
Tae-Hyun Oh
Naoto Ohsaka
Wayne Oldford
Michael Oliver
Bruno Olshausen
Mohamed Omar
Arno Onken
Sageev Oore
José Oramas M.
Vicente Ordonez
Lorenzo Orecchia
Alexander Ororbia Ii
Pedro Ortega
Ian Osband
Christian Osendorfer
Simon Osindero
Dmitry Ostrovsky
Joachim Ott
Long Ouyang
Yunbo Ouyang
Megan Owen
Samet Oymak
Mete Ozay
Bahadir Ozdemir
Alexey Ozerov
Dmitry P. Vetrov
Marius Pachitariu
Leslie Pack Kaelbling
Krishnan Padmanabahn
Cosmin Paduraru
John Paisley
Joni Pajarinen
Ari Pakman
Christopher Pal
David Pal
Dipan Pal
Soumitra Pal
Sebastian Palacio
Konstantina Palla
Manohar Paluri
Binbin Pan
Sinno Pan
Wei Pan
Weiwei Pan
Xinlei Pan
Yunpeng Pan
Wei Pang
Rina Panigrahy
Liam Paninski
Maxim Panov
Angeliki Pantazi
Dimitris Papailiopoulos
George Papamakarios
George Papandreou
Laetitia Papaxanthos
Ulrich Paquet
Roberto Paredes
Dohyung Park
Jongsoo Park
Juhyun Park
Jung-Guk Park
Mijung Park
Seunghyun Park
Youngsuk Park
Ronald Parr
Emilio Parrado
Pekka Parviainen
Razvan Pascanu
Stephen Pasteris
Giorgio Patrini
Edouard Pauwels
Vladimir Pavlovic
Klaus Pawelzik
Jason Pazis
Mykola Pechenizkiy
Dmitry Pechyony
Marco Pedersoli
Fabian Pedregosa
Tomi Peltola
Jaakko Peltonen
Hanyang Peng
Richard Peng
Jeffrey Pennington
Anastasia Pentina
Fernando Pereira
Franz Pernkopf
Yura Perov
Laurent Perrinet
Valerio Perrone
Michael Perrot
Vrancx Peter
Zuzana Petrickova
Marek Petrik
Nathalie Peyrard
David Pfau
Jeff Phillips
Dinh Phung
Massimo Piccardi
Olivier Pietquin
Georgios Piliouras
Venkata Pillutla
Wei Ping
Bilal Piot
Matteo Pirotta
Hamed Pirsiavash
Xaq Pitkow
Emily Pitler
Barbara Plank
John Platt
Sergey Plis
Mark Plumbley
Eftychios Pnevmatikakis
Adam Pocock
Anastasia Podosinnikova
Matthias Poloczek
Adriano Polpo
Jean Ponce
Ting Kei Pong
Ben Poole
Alin Popa
Florin Popescu
Jean Pouget-Abadie
Stefanos Poulis
Pascal Poupart
Daniel Povey
Sandhya Prabhakaran
Rohit Prabhavalkar
L.A. Prashanth
Vittal Premachandran
Philippe Preux
Bob Price
Eric Price
Alexandre Proutiere
Yunchen Pu
Christian Puhrsch
Jay Pujara
Oriol Pujol Vila
Golan Pundak
Sanjay Purushotham
Nelly Pustelnik
Fei Qi
Yanjun Qi
Zhengling Qi
Zhongang Qi
Chao Qian
Guangwu Qian
Junyang Qian
Xiaoning Qian
Linbo Qiao
Yu Qiao
Biao Qin
Tao Qin
Chao Qu
Simeng Qu
Zheng Qu
Novi Quadrianto
Gerald Quon
Kush R. Varshney
Andrew Rabinovich
Neil Rabinowitz
Miki Racz
Vladan Radosavljevic
Jack Rae
Stephen Ragain
Aswin Raghavan
Maithra Raghu
Maxim Raginsky
Mostafa Rahmani
Holakou Rahmanian
Piyush Rai
Raviv Raich
Tom Rainforth
Nazneen Fatema Rajani
Parikshit Ram
Prajit Ramachandran
Kandan Ramakrishnan
Karthikeyan Ramamurthy
Ramprasaath Ramasamy Selvaraju
Arunselvan Ramaswamy
Harish Ramaswamy
Visvanathan Ramesh
Jan Ramon
Raul Ramos-Pollán
Rajesh Ranganath
Syama Rangapuram
Amal Rannen Triki
Marc'Aurelio Ranzato
Anup Rao
Nikhil Rao
Vinayak Rao
Janardhan Rao Doppa
Malte Rasch
Garvesh Raskutti
Mohammad Rastegari
Kashif Rasul
Alexander Ratner
Joel Ratsaby
Florian Raudies
Federico Raue
Siamak Ravanbakhsh
Sathya Ravi
Sujith Ravi
Ankit Rawat
Ramin Raziperchikolaei
Leonid Razoumov
Tabea Rebafka
Sashank Reddi
Ievgen Redko
Scott Reed
Henry Reeve
Lucas Rêgo Drumond
Chaouki Regoui
Felix Reinhart
Thodoris Rekatsinas
Konstantinos Rematas
Haiqing Ren
Mengye Ren
Steve Renals
Alexandre Rene
Vijay Rengarajan
Steven Rennie
Marcello Restelli
Achim Rettinger
Jan Reubold
Lev Reyzin
Hamid Rezatofighi
Alejandro Ribeiro
Elisa Ricci
Emile Richard
Jonas Richiardi
Oran Richman
Peter Richtarik
James Ridgway
Sebastian Riedel
Carles Riera Molina
Stefan Riezler
Ryan Rifkin
Guillem Rigaill
Sebastian Risi
Andrej Risteski
Daniel Ritchie
Mariano Rivera Meraz
Syed Ali Asad Rizvi
Brian Roark
Sylvain Robbiano
Jonathan Robert Ullman
Stéphane Robin
Pablo Robles Granda
Alexis Roche
Tim Rocktäschel
Erik Rodner
Pau Rodríguez López
Irene Rodriguez-Lujan
Jordan Rodu
Rebecca Roelofs
Ryan Rogers
Gregory Rogez
Mohammad Rohban
Cosma Rohilla Shalizi
Marcus Rohrbach
Gemma Roig
Lina Rojas-Barahona
Mateo Rojas-Carulla
Bernardino Romera-Paredes
Adriana Romero
Javier Romero
Nan Rong
Teemu Roos
Massimo Rosa
Derek Rose
Clemens Rosenbaum
David Rosenberg
Johnathan Rosenblatt
Nir Rosenfeld
Benjamin Rosman
Fabrice Rossi
Luca Rossi
Afshin Rostamizadeh
Volker Roth
Wolfgang Roth
Constantin Rothkopf
Sylvain Rousseau
Juho Rousu
Ryan Rowekamp
Mark Rowland
Aurko Roy
Sandipan Roy
Christopher Rozell
Leonel Rozo
Alessandro Rozza
Paul Rubenstein
Ran Rubin
Benjamin Rubinstein
Natali Ruchansky
Alessandro Rudi
Maja Rudolph
Ognjen Rudovic
Francisco Ruiz
Nicholas Ruozzi
Bryan Russell
Daniel Russo
Daniil Ryabko
Michael Ryoo
Feras Saad
Yunus Saatci
Régis Sabbadin
Anne Sabourin
Sushant Sachdeva
João Sacramento
Kayvan Sadeghi
Veeranjaneyulu Sadhanala
Ardavan Saeedi
Amir Saffari
Gonzalo Safont
Hesam Sagha
Levent Sagun
Shagan Sah
Aadirupa Saha
Barna Saha
Mojtaba Sahraee-Ardakan
Puja Sahu
Hiroto Saigo
Mehdi Sajjadi
Tomoya Sakai
Charbel Sakr
Jun Sakuma
Amir Salavati
Ardavan Salehi
Nobandegani
Venkatesh Saligrama
Hugh Salimbeni
Joseph Salmon
Mathieu Salzmann
Dimitris Samaras
Aven Samareh
Rajhans Samdani
Wojciech Samek
Ruben Sanchez-Romero
Fernando Sancho Caparrini
Mark Sandler
Maxime Sangnier
Amir Sani
Aswin Sankaranarayanan
Sriram Sankararaman
Scott Sanner
Raul Santos-Rodriguez
Soumik Sarkar
Raghunatha Sarma
Badrul Sarwar
Anand Sarwate
Hiroaki Sasaki
Yutaka Sasaki
Aziza Satkhozhina
Issei Sato
James Saunderson
Shervi Savafi
Pierre-André Savalle
Bogdan Savchynskyy
Cristina Savin
Nikolay Savinov
Andrew Saxe
Shreyas Saxena
Ali Sayed
Mauro Scanagatta
Simone Scardapane
Benjamin Scellier
Robert Schapire
Tom Schaul
Aaron Schein
Katya Scheinberg
Reinhold Scherer
Bernt Schiele
Leander Schietgat
Jürgen Schmidhuber
Ludwig Schmidt
Mark Schmidt
Mikkel Schmidt
Tanner Schmidt
Uwe Schmidt
Francois Schnitzler
Angela Schoellig
Peter Schulam
Bjoern Schuller
Thomas Schultz
Eric Schulz
Ingmar Schuster
Mike Schuster
Haim Schweitzer
Holger Schwenk
Alexander Schwing
Adam Scibior
Damien Scieur
Christoph Scnoerr
Clayton Scott
D. Sculley
Michèle Sebag
Hanie Sedghi
Matthias Seeger
Frank Seide
Daniel Seita
Dino Sejdinovic
Andrew Semelhago
Stanislau Semeniuta
Prithiraj Sen
Rajat Sen
Ransalu Senanayake
Ozan Sener
Andrew Senior
Minjoon Seo
Iulian Serban
Thomas Serre
Sohan Seth
H. Sebastian Seung
Laura Sevilla
Ahmad Shabbar Kazmi
Izhak Shafran
Fahad Shah
Nihar Shah
Dafna Shahaf
Shahin Shahrampour
Arash Shahriari
Greg Shakhnarovich
Uri Shalit
Naresh Shanbhag
Fanhua Shang
Karthikeyan Shanmugam
Weijia Shao
Amit Sharma
Gaurav Sharma
Tatyana Sharpee
James Sharpnack
Or Sheffet
Alexander Shekhovtsov
Daniel Sheldon
Christian Shelton
Chunhua Shen
Haichen Shen
Jie Shen
Li Shen
Weiwei Shen
Xiaobo Shen
Xin Shen
Yanyao Shen
Yuanyuan Shen
Zhiqiang Shen
Baoguang Shi
Baoxu Shi
Lei Shi
Qinfeng Shi
Xiaoshuang Shi
Xingjian Shi
Yinghuan Shi
Yuanming Shi
Zhiyuan Shi
Chuenkai Shie
Alistair Shilton
Hideaki Shimazaki
Shohei Shimizu
Nahum Shimkin
Hidetoshi Shimodaira
Yishai Shimoni
Jinwoo Shin
Sungho Shin
Lavi Shpigelman
Ilya Shpitser
Anshumali Shrivastava
Julian Shun
Hussein Sibai
Aaron Sidford
Leonid Sigal
Olivier Sigaud
Nathan Silberman
Ricardo Silva
David Silver
Edgar Simo-Serra
Karen Simonyan
Ozgur Simsek
Umut Simsekli

REVIEWERS

Vikas Sindhwani
Yaron Singer
Yoram Singer
Yoram Singer
Maneesh Singh
Sameer Singh
Shashank Singh
Vivek Singh
Tomas Singliar
Aman Sinha
Ayan Sinha
Kaushik Sinha
Mathieu Sinn
Benjamin Sirb
Justin Sirignano
Vidyashankar Sivakumar
Josef Sivic
Marcin Skwark
Martin Slawski
Paris Smaragdis
Marek Smieja
Cristian Sminchisescu
Kevin Smith
Christopher Smithers
Cees Snoek
Jasper Snoek
Marta Soare
Richard Socher
Maximilian Soelch
Mohammad Reza Soheili
Jascha Sohl-Dickstein
Kihyuk Sohn
Jure Sokolic
Nataliya Sokolovska
Justin Solomon
Mahdi Soltanolkotabi
Tasuku Soma
Friedrich Sommer
Stefan Sommer
Jeong-Woo Son
Casper Kaae Sønderby
Søren Sønderby
Dogyoon Song
Hyun Oh Song
Le Song
Shuang Song
Yang Song
Yangqiu Song
Zhao Song
Aureli Soria-Frisch
Humberto Sossa
Jose Sotelo
Alvaro Soto
Daniel Soudry
Pablo Sprechmann
Ryan Spring
Jost Tobias Springenberg
Sreenivas Sremath Tirumala
Vivek Srikumar
Suraj Srinivas
Srinivasan Sriram
Cyril Stark
Michael Stark
Nikolaas Steenbergen
Noah Stein
Jacob Steinhardt
Thomas Steinke
Christian Steinrücken
Ingo Steinwart
Ian Stevenson
Sebastian Stober
Quentin Stout
Bing Su
Hang Su
Hao Su
Hongyu Su
Pei-Hao Su
Shih-Chieh Su
Weijie Su
Alberto Suarez
Arun Sai Suggala
Mahito Sugiyama
Changho Suh
Suwon Suh
Heung-Il Suk
Sainbayar Sukhbaatar
Uygar Sümbül
Baochen Sun
Hailong Sun
Haoze Sun
He Sun
Jian Sun
Jiangwen Sun
Ke Sun
Min Sun
Qing Sun
Qingyun Sun
Quan Sun
Ruoyu Sun
Shizhao Sun
Tingni Sun
Wei Sun
Wen Sun
Yifan Sun
Zhun Sun
Wee Sun Lee
Ganesh Sundaramoorthi
Wonyong Sung
Charles Sutton
Johan Suykens
Hideyuki Suzuki
Taiji Suzuki
Michele Svanera
Lennart Svensson
Ola Svensson
Adith Swaminathan
Kevin Swersky
Paul Swoboda
Umar Syed
Umar Syed
Zeeshan Syed
Gabriel Synnaeve
Vasilis Syrgkanis
Zoltan Szabo
Botond Szabó
Sandor Szedmak
Enikő Székely
David Szepesvari
Arthur Szlam
Balazs Szorenyi
Yasuo Tabei
Amirhossein Taghvaei
Martin Takac
Akiko Takeda
Naoya Takeishi
Ichiro Takeuchi
Eiji Takimoto
Partha Talukdar
Erik Talvitie
Ameet Talwalkar
Kunal Talwar
Aviv Tamar
Conghui Tan
Tanmingkui Tan
Xiaoyang Tan
Toshiyuki Tanaka
Bo Tang
Charlie Tang
Cheng Tang
Jianbin Tang
Peng Tang
Shaojie Tang
Song Tang
Jun Tani
Daniel Tanneberg
Wesley Tansey
Chenyang Tao
Dacheng Tao
Wenbing Tao
Sasha Targ
Danny Tarlow
Davide Tateo
Nikolaj Tatti
Adam Tauman Kalai
Romain Tavenard
David Tax
Gavin Taylor
Yee-Whye Teh
Matus Telgarsky
Arthur Tenenhaus
Yoshikazu Terada
Evimaria Terzi
Heidi Tessmer
Justin Thaler
Vu Thang
Ananda Theertha Suresh
Fabian Theis
Georgios Theocharous
Sergios Theodoridis
Evangelos Theodorou
John Thickstun
Bertrand Thirion
Philipp Thomann
Albert Thomas
Nicolas Thome
Tian Tian
Yuandong Tian
Ryan Tibshirani
Radu Timofte
Daniel Ting
Ivan Titov
George Toderici
Iliya Tolstikhin
Ryota Tomioka
Marc Tommasi
Tatiana Tommasi
Mansi Tommaso
Mariya Toneva
Yan Tong
Antonio Torralba
José Torrecilla
Lorenzo Torresani
Andrea Torsello
Christopher Tosh
Alexander Toshev
Alessandra Tosi
Aristide Tossou
Panos Toulis
Brendan Tracey
Du Tran
Dustin Tran
Quoc Tran-Dinh
Volker Tresp
George Trigeorgis
Sebastian Trimpe
Samarth Tripathi
Rocco Tripodi
Jean-Baptiste Tristan
Ludovic Trottier
Chuan-Yung Tsai
Ivor Tsang
Sebastian Tschiatschek
Maksim Tsikhanovich
Kewei Tu
Kun Tu
Stephen Tu
Zhuowen Tu
George Tucker
Shubham Tulsiani
Borahan Tumer
Srinivas Turaga
Gyorgy Turan
Javier Turek
Rasul Tutunov
Niall Twomey
Himanshu Tyagi
Shashanka Ubaru
Eiji Uchibe
Madeleine Udell
Jonas Umlauft
Thomas Unterthiner
Utkarsh Upadhyay
Raquel Urtasun
Tanguy Urvoy
Nicolas Usunier
Daniel Vainsencher
Samuel Vaiter
Sergio Valadez-Godínez
Isabel Valera
Hamed Valizadegan
Michal Valko
Jack Valmadre
Jan-Willem Van De Meent
Guy Van Den Broeck
Laurens Van Der Maaten
Mark Van Der Wilk
Tim Van Erven
Jurgen Van Gael
Marcel Van Gerven
Hado Van Hasselt
Johan Van Horebeek
Twan Van Laarhoven
Jimmy Vandel
Bart Vandereycken
Pierre Vandergheynst
David Vandyke
Jarno Vanhatalo
Vincent Vanhoucke
Joaquin Vanschoren
Balakrishnan Varadarajan
Kiran Varanasi
Yogatheesan Varatharajah
Ehsan Variani
Gaël Varoquaux
Ara Vartanian
Sebastiano Vascon
Eleni Vasilaki
Flavian Vasile
Ashish Vaswani
Namrata Vaswani
Vsr Veeravasarapu
Andreas Veit
Alfredo Vellido
Dmytro Velychko
Suresh Venkatasubramanian
Dan Ventura
Samuel Ventura
Deepak Venugopal
Greg Ver Steeg
Jakob Verbeek
Nakul Verma
Claire Vernade
Paul Vernaza
Jean-Charles Vialatte
Brunel Victor-Emmanuel
Rene Vidal
Carmen Vidaurre
Marina Vidovic
Matthieu Vignes
Ashwin Vijayakumar
Sudheendra Vijayanarasimhan
Aravindan Vijayaraghavan
Silvia Villa
Ruben Villegas
Pascal Vincent
Giuseppe Vinci
Vibhav Vineet
Marina Vinyes
Seppo Virtanen
Nisheeth Vishnoi
Fabio Vitale
Constantine Vitt
Andreas Vlachos
Adrian Vladu
Joshua Vogelstein
Julia Vogt
Karsten Vogt
Carl Vondrick
Bang Vu
Slobodan Vucetic
Sara Wade
Aseem Wadhwa
Bo Waggoner
Hoi-To Wai
Yoav Wald
Christian Walder
Irene Waldspurger
Byron Wallace
Matthew Walter
Yali Wan
Baiyang Wang
Beilun Wang
Bo Wang
Changhu Wang
Chaohui Wang
Chong Wang
Chunye Wang
De Wang
Deqing Wang
Dilin Wang
Donghui Wang
Fangpo Wang
Fei Wang
Gang Wang
Hao Wang
Huahua Wang
Huazheng Wang
Jialei Wang
Jiang Wang
Jie Wang
Jingdong Wang
Jinzhuo Wang
Jun-Kun Wang
Lei Wang
Li Wang
Li-Lun Wang
Lingxiao Wang
Linnan Wang
Liwei Wang
Lu Wang
Mengdi Wang
Peng Wang
Pengyu Wang
Po-Wei Wang
Qianqian Wang
Qiurui Wang
Shaojun Wang
Shaoqing Wang
Shenlong Wang
Shuhui Wang
Siwei Wang
Suhang Wang
Taifeng Wang
Wei Wang
Weiqiang Wang
Weiran Wang
Xiangfeng Wang
Xiangyu Wang
Xiaolong Wang
Xiaoqian Wang
Xinan Wang
Xinggang Wang
Xueqin Wang
Yali Wang
Yan Wang
Yan Wang
Yang Wang
Yi Wang
Yichen Wang
Yilin Wang
Yining Wang
Yixin Wang
Yiyang Wang
Yizhen Wang
Yizhi Wang
Yu Wang
Yulong Wang
Yunhe Wang
Yuyang Wang
Yuyi Wang
Zhangyang Wang
Zhaoran Wang
Zheng Wang
Zilei Wang
Ziyu Wang
David Warde-Farley
Shinji Watanabe
Fabian Wauthier
Greg Wayne
Theophane Weber
Jan Wegner
Leila Wehbe
Louis Wehenkel
Dennis Wei
Ermin Wei
Kai Wei
Tao Wei
Xiu-Shen Wei
Yingying Wei
Zijun Wei
Sebastian Weichwald
Pan Weike
Asaf Weinstein
David Weiss
Jeremy Weiss
Roi Weiss
Adrian Weller
Jinming Wen
Tsung-Hsien Wen
Wei Wen
Yandong Wen
Zheng Wen
Bingni Wen Brunton
Paul Weng
Dong Wenyong
Adam White
Colin White
Martha White
Michael Wick
Nathan Wiebe
Aleksander Wieczorek
Jenna Wiens
Michael Wilber
Bryan Wilder
Rebecca Willett
Chris Williams
Joseph Williams
Andrew Wilson
James Wilson
David Wingate
Kevin Winner
Olivier Wintenberger
David Wipf
Anthony Wirth
Scott Wisdom
Peter Wittek
Paul Wohlhart
Christian Wolf
Benjamin Wolfe
Alex Wong
Eric Wong
David Woodruff
Blake Woodworth
John Wright
Anqi Wu
Chunpeng Wu
Guangbin Wu
Hao Wu
Hao Wu
Huasen Wu
Jiajun Wu
Jian Wu
Jianxin Wu
Qiang Wu
Qingyao Wu
Qingyun Wu
Shan-Hung Wu
Shanshan Wu
Si Wu
Tao Wu
Ying Wu
Ying Nian Wu
Yue Wu
Zifeng Wu
Bin Xiang
Yu Xiang
Yu Xiang
Houping Xiao
Jingjing Xiao
Bo Xie
Chen-Wei Xie
Jianwen Xie
Ning Xie
Pengtao Xie
Saining Xie
Yao Xie
Yusheng Xie
Bo Xin
Fuyong Xing
Felix Xinnan Yu
Caiming Xiong
Jiechao Xiong
Kai Xiong
Yunyang Xiong
Chang Xu
Chen Xu
Cheng Xu
Chenliang Xu
Guodong Xu
Huan Xu
Jia Xu
Jinglin Xu
Kevin Xu
Liangbei Xu
Linli Xu
Miao Xu
Min Xu
Peng Xu
Qianqian Xu
Shen Chen Xu
Weidi Xu
Yan Xu
Yi Xu
Ying Xu
Yuesheng Xu
Zhen Xu
Zheng Xu
Zheng Xu
Zhiqiang Xu
Tianfan Xue
Makoto Yamada
Kota Yamaguchi
Bowei Yan
Junchi Yan
Mengyuan Yan
Ming Yan
Shuicheng Yan
Songbai Yan
Xiaoran Yan
Xinchen Yan
Yan Yan
Yan Yan
Dawei Yang
Eunho Yang
Greg Yang
Haiqin Yang
Hao Yang
Jianwei Yang
Jie Yang
Jimei Yang
Meng Yang
Michael Yang
Ming-Hsuan Yang
Pei Yang
Scott Yang
Shan Yang
Shuo Yang
Tao Yang
Tianbao Yang
Wenzhuo Yang
Yi Yang
Ying Yang
Yingxiang Yang
Yingzhen Yang
Yun Yang
Zhi Yang
Zhihui Yang
Zhilin Yang
Zhirong Yang
Zhuoran Yang
Chen Yanover
Anbang Yao
Angela Yao
Quanming Yao
Yuan Yao
Zhijun Yao
Julian Yarkony
Fatos Yarman Vural
Grigory Yaroslavtsev
Yutaka Yasui
Dimitri Yatsenko
Erfan Yazdandoost Hamedani
Haishan Ye
Han-Jia Ye
Jianbo Ye
Jong Chul Ye
Mao Ye
Nan Ye
Xiaojing Ye
Yuting Ye
Ian En-Hsu Yen
Halid Ziya Yerebakan
Dit-Yan Yeung
Florian Yger
Xinyang Yi
Scott Yih
Dong Yin
Junming Yin
Aphinyanaphongs Yindalon
Bicheng Ying
Arjun Yogeswaran
Chang Yoo
Jaeyoon Yoo
Sungjoo Yoo
Yuya Yoshikawa
Chong You
Quanzeng You
Seungil You
Shan You
Mohammed Yousefhussien
Adams Wei Yu
Angela Yu
Byron Yu
Chun-Nam Yu
Dong Yu
Fisher Yu
Guoqiang Yu
Hao Yu
Hong Yu
Hsiang-Fu Yu
Jun Yu
Rose Yu
Shan Yu
Shipeng Yu
Stella Yu
Xiyu Yu
Yang Yu
Yaoliang Yu
Zhang Yu
Zhaofei Yu
Zhiding Yu
Changhe Yuan
Chun Yuan
Chunfeng Yuan
Ganzhao Yuan
Jiangye Yuan
Ke Yuan
Lufeng Yuan
Xiaotong Yuan
Xin Yuan
Yang Yuan
Ye Yuan
Yisong Yue
Ersin Yumer
Hyokun Yun
Se-Young Yun
Mikhail Yurochkin
Amir Zadeh
Morteza Zadimoghaddam
Bianca Zadrozny
Stefanos Zafeiriou
Tom Zahavy
Manzil Zaheer
Davide Zambrano
Giacomo Zanella
Fabio Zanzotto
Giovanni Zappella
Ali Zarezade
Maxim Zaslavsky
Alexey Zaytsev
Lenka Zdeborova
Pablo Zegers
Kuo Hao Zeng
Fabio Massimo Zennaro
Shuangfei Zhai
De-Chuan Zhan
Yifeng Zhan
Yusen Zhan
Amy Zhang
Anru Zhang
Aonan Zhang
Bibo Zhang
Byoung-Tak Zhang
Ce Zhang
Changqing Zhang
Changshui Zhang
Chicheng Zhang
Danke Zhang
Daoqiang Zhang
Debing Zhang
Dingwen Zhang
Hejia Zhang
Hengmin Zhang
Hong Zhang
Hongyang Zhang
Hongyi Zhang
Huan Zhang
Huishuai Zhang
Jiji Zhang
Junping Zhang
Kai Zhang
Lei Zhang
Liang Zhang
Libo Zhang
Lijun Zhang
Martin Zhang
Matt Zhang
Mengmi Zhang
Miaomiao Zhang
Min-Ling Zhang
Nevin Zhang
Pan Zhang
Peng Zhang
Qin Zhang
Quan Zhang
Quanshi Zhang
Richard Zhang
Richong Zhang
Rui Zhang
Ruiwen Zhang
Sai Zhang
Saizheng Zhang
Shaoting Zhang
Teng Zhang
Weizhong Zhang
Wen-Hao Zhang
Xiang Zhang
Xiaopeng Zhang
Xiaoqin Zhang
Xiaowei Zhang
Xin Zhang
Xinhua Zhang
Yinda Zhang
Ying Zhang
Yizhe Zhang
Yizhe Zhang
Yu Zhang
Yuan Zhang
Yuchen Zhang
Yunong Zhang
Yuting Zhang
Zhen Zhang
Zhenjie Zhang
Zhihua Zhang
Bo Zhao
Fang Zhao
Han Zhao
Handong Zhao
Hao Zhao
Jiaping Zhao
Liang Zhao
Peilin Zhao
Qi Zhao
Shengjia Zhao
Shenjian Zhao
Shijie Zhao
Tuo Zhao
Yi Zhao
Yue Zhao
Zhizhen Zhao
Li Zhaoping
Shandian Zhe
Elena Zheleva
Alice Zheng
Charles Zheng
Kai Zheng
Qinqing Zheng
Shuai Zheng
Shun Zheng
Stephan Zheng
Wenjie Zheng
Xiaoqing Zheng
Yin Zheng
Andrey Zhmoginov
Kai Zhong
Mingjun Zhong
Xiaowei Zhong
Yu Zhong
Ding-Xuan Zhou
Hong-Yu Zhou
Luowei Zhou
Mingyuan Zhou
Pengcheng Zhou
Qiang Zhou
Shuigeng Zhou
Tianyi Zhou
Wenxin Zhou
Xu Zhou
Yaqin Zhou
Yuxun Zhou
Zhi-Hua Zhou
En Zhu
Hongtu Zhu
Jianke Zhu
Jun Zhu
Jun Zhu
Junjie Zhu
Kai Zhu
Linhong Zhu
Rong Zhu
Rui Zhu
Xia Zhu
Xiaodan Zhu
Yuancheng Zhu
Yunzhang Zhu
Michael Zibulevsky
Brian Ziebart
Sandra Zilles
Daniel Zink
Andrew Zisserman
Masrour Zoghi
Daniel Zoran
James Zou
Quan Zou
Anastasios Zouzias
Or Zuk
Wangmeng Zuo
Alon Zweig =

AUTHOR INDEX

P = Poster

AUTHOR INDEX

AUTHOR INDEX

AUTHOR INDEX

AUTHOR INDEX

AUTHOR INDEX

Platinum Sponsors

AIG · Apple · Audi · Citadel · DC VC Data Collective · DeepMind · DiDi · facebook · Google · intel · KLA-Tencor · Microsoft · Tencent 腾讯 · WINTON

GOLD SPONSORS

- Adobe Research
- ALIBABA
- Amazon
- AtlaSense
- Baidu
- Criteo
- IBM Research
- NVIDIA
- The Voleon Group

SILVER SPONSORS

- 360 Technology Co.
- Automotive Safety Technologies
- Bloomberg
- Bosch
- Cubist Systematic Strategies
- D. E. Shaw Group
- eBay
- FeatureX
- G-Research
- Hutchin Hill Capital, LP
- Jump Trading
- Man AHL
- Next AI
- Optiver
- Qualcomm
- Renaissance Technologies
- Rosetta Analytics
- SAP SE
- Sentient Technologies
- Two Sigma Investments
- United Technologies Research

BRONZE SPONSORS

- Benevolent AI
- Beijing Institute of Big Data Research
- Cheetah Mobile
- Datatang Technology
- Disney Research
- Invenia Labs
- Maluuba Inc.
- Nokia
- Oracle
- Palantir Technologies
- Panasonic
- PDT Partners
- QuantumBlack
- RBC
- Recursion Pharmaceuticals
- Schibsted Media Group
- SigOpt, Inc.
- Telefonica
- Yandex