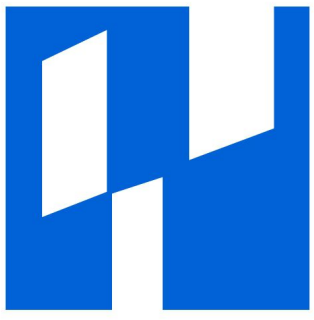
Program Booklet

Second TRIPODS PI meeting
University of California, Santa Cruz
Silicon Valley Campus
October 23-24, 2018

Baskin Engineering
UC SANTA CRUZ



Program At a Glance

Tuesday, 10/23

08:30 - 08:40	Welcome remarks - Lise Getoor (UCSC) and Abel Rodriguez (UCSC)
08:40 - 10:10	Research Session 1: Randomized & Stochastic Algorithms. Leaders: Piotr Indyk (MIT) & Katya Scheinberg (Lehigh)
10:10 - 10:30	Coffee Break
10:30 - 12:00	Research Session 2: Nonconvex Optimization & Deep Learning. Leaders: Xiaoming Huo (GATech) & Maryam Fazel (UWash)
12:00 - 01:00	Lunch
01:00 - 01:30	National Science Foundation.
01:30 - 03:00	Research Session 3: Topology, Geometry & Graphs. Leaders: Tamal Dey (OSU) & Sesh Comandur (UCSC)
03:00 - 03:30	Coffee Break
03:30 - 05:00	Industry Panel. Moderator: Michael Mahoney (UCB) Panelists: Deepak Agrawal (LinkedIn), Erin Ledell (H20), Edo Liberty (Amazon), Peter Norvig (Google), Ashok Srinivastava (Intuit).
05:00 - 05:10	Break
05:10 - 05:50	Poster Spotlights
05:50 - 08:30	Poster Session, with food & drinks. Evening Talk: Cathryn Carson (UCB) - A Historical Perspective on the Emerging Field of Data Science

Wednesday, 10/24

08:30 - 10:00	Data Science Education. Leader: Helen Zhang (UA). Speakers: Stephen Wright (UWisc), Michael Mahoney (UCB)
10:00 - 10:30	Coffee Break
10:30 - 12:00	Institution and Infrastructure Panel. Moderator: Abel Rodriguez (UCSC). Panelists: Hélène Barcelo (MSRI), Jeff Brock (ICERM), Bob Brown (Purdue NSF STC), David Ribes (UWash), Dimitri Shlyakhtenko (IPAM).
12:00 - 01:30	Lunch - PM & PI Closed Session
01:30 - 02:30	Research Session 4: Responsible Data Science. Leader: Lise Getoor (UCSC). Guest Speaker: Bill Howe (UWash)
02:30 - 03:00	Closing discussion - Next steps

Detailed Program

Tuesday, 10/23

08:30 - 08:40 Welcome remarks - Lise Getoor (UCSC) and Abel Rodriguez (UCSC)

08:40 - 10:10 Research Session 1: Randomized & Stochastic Algorithms. Leaders: Piotr Indyk (MIT) & Katya Scheinberg (Lehigh)

- 08:40 - 08:50 Overview of Stochastic Algorithms (Katya Scheinberg, Lehigh)
- 08:50 - 09:00 Overview of Randomized (Sub-linear) Algorithms (Piotr Indyk, MIT)
- 09:00 - 09:10 Gradient Sampling Methods for Nonconvex Nonsmooth Optimization (Frank E. Curtis, Lehigh)
- 09:10 - 09:20 Newton-Type Methods for Non-Convex Optimization Under Inexact Hessian Information, (Michael Mahoney, UCB)
- 09:20 - 09:30 Convergence Rates of Stochastic Algorithms in Nonsmooth Nonconvex Optimization (Dmitriy Drusvyatskiy, UWash)
- 09:30 - 09:50 Lightning presentations:
  - Differentially Private Identity and Equivalence Testing of Discrete Distributions (Maryam Aliakbarpour, MIT)
  - Rethinking learning rate schedules for stochastic optimization (Rahul Kidambi, UWash)
  - Approximate Nearest Neighbors in Limited Space (Tal Wagner, MIT)
  - Distributed Learning over a Network of Agents under Communication Constraints (Thinh T. Doan, GATech)
  - Learning-Based Frequency Estimation Algorithms (Ali Vakilian, MIT)
- 09:50 - 10:10 Discussion

10:10 - 10:30 Coffee Break

10:30 - 12:00 Research Session 2: Nonconvex Optimization & Deep Learning. Leaders: Xiaoming Huo (GATech) & Maryam Fazel (UWash)

- 10:30 - 10:45 Overview: lessons learned from recent Georgia Tech and Wisconsin-Washington workshops on nonconvex optimization & deep learning (X. Huo, S. Wright, M. Fazel)
- 10:45 - 11:45 Lightning presentations:
  - A Smoother Way to Train Structured Prediction Models (Z. Harchaoui, UWash)

- Implicit Self-Regularization in Deep Neural Networks: Evidence from Random Matrix Theory & Implications for Learning (M. Mahoney, UCB)
- The Gap Between Worst-Case Complexity and Empirical Performance in Nonconvex Optimization (F. Curtis, Lehigh)
- Sharp Convergence Rates for Langevin Dynamics in the Nonconvex Setting (M. Jordan, UCB)
- Convergence of Newton-Conjugate-Gradient Algorithms to Second-Order Points (S. Wright, UWisc)
- Convergence of Policy Gradient Methods for the Linear Quadratic Regulator (M. Fazel, UWash)
- Nonconvex Geometry and Algorithm for Sparse Blind Deconvolution (John Wright, Columbia)

11:45 - 12:00 Discussion

12:00 - 1:00 Lunch

01:00 - 01:30 National Science Foundation. Introductions: Tracy Kimbrel (Program Director, Division of Computing and Communication Foundations). Speakers: Deborah Lockhart (Deputy Assistant Director, Directorate for Mathematical & Physical Sciences), Walter (Rance) Cleaveland (Division Director, CCF), Juan C. Meza (Division Director, DMS)

01:30 - 02:00 Research Session 3: Topology, Geometry & Graphs. Leaders: Tamal Dey (OSU) & Sesh Comandur (UCSC)

01:30 - 01:35 Introduction to Topological Data Analysis (Tamal Dey, OSU)

01:35 - 02:10 Lightning presentations:

- Use of Topological Methods in Data Analysis (Yusu Wang, OSU)
- From TGDA for 3D Shapes, dynamic shapes, and networks (Facundo Memoli, OSU)
- Stochastic processes on and of networks (David Sivakoff, OSU)
- On Symplectic Optimization (Michael Jordan, UCB)
- Linear and sublinear-time algorithms for optimal transport (Jonathan Weed, MIT)
- Using the tangent map to explain nonlinear dimension reduction, (David Glickenstein, UA)
- Current and Future Role of Topological and Geometric Methods for Data Analysis (Sebastian Kurtek, OSU)

02:10 - 02:15 Discussion

02:15 - 02:25 Foundational data science challenges for graphs (C. Seshadhri, UCSC)

02:25 - 02:50 Spotlight presentations:

- Variational Perspective on Local Graph Clustering (Michael Mahoney, UCB)
- Multi-Level Steiner Trees (Joe Watkins, UA)
- Riemannian embedding models for relational data (Abel Rodriguez, UCSC)
- Improved estimators for network-based sampling of hard-to-reach populations (Sebastian Roch, UWisc)
- Estimating degree distributions, without see the whole graph (C. Seshadhri, UCSC)

02:50 - 03:00 Discussion

03:00 - 03:30 Coffee Break

03:30 - 05:00 Industry Panel. Moderator: Michael Mahoney (UCB)
Panelists: Deepak Agrawal (LinkedIn), Erin Ledell (H20), Edo Liberty (Amazon), Peter Norvig (Google), Ashok Srinivastava (Intuit).

05:00 - 05:10 Break

05:10 - 08:30 Poster session

05:10 - 05:50 Poster spotlights

1. Do we need 2nd order optimization methods in machine learning? (Albert Berahas, Lehigh)
2. A Unifying Framework of High-Dimensional Sparse Estimation with Difference-of-Convex (DC) Regularizations (Shanshan Cao, GATech)
3. ATOMO: Communication-efficient Learning via Atomic Sparsification (Zachary Charles, UWisc)
4. Random sampling and efficient algorithm for multiscale PDEs (Ke Chen, UWisc)
5. The Importance of Forgetting: Topological Insights from Neural Data (Samir Chowdhury, OSU)
6. Leveraged volume sampling for linear regression (Michal Derezinski, UCB)
7. Distributed Learning over a Network of Agents under Communication Constraints (Thinh Doan, GATech)
8. Large Scale Training of Neural Networks Using Robust Optimization and Hessian Information (Amir Gholami, UCB)
9. Bayesian Dynamic Feature Partition with Massive Data (Rene Gutierrez, UCSC)
10. Optimization in high-dimensional statistics with nonconvex constraints (Wooseok Ha, UCB)
11. Efficient Distributed Hessian Free Algorithm for Large-scale Empirical Risk Minimization via Accumulating Sample Strategy (Majid Jahani, Lehigh)
12. Beyond Single Examples: Interpreting Black-Box Models via Set Influence (Rajiv Khanna, UCB)

13. Towards improved design of parallel and accelerated stochastic gradient methods (Rahul Kidambi, UWash)
14. An Efficient Pruning Algorithm for Robust Isotonic Regression (Cong Han Lim, GATech)
15. Thresholded Hierarchical Clustering of Gene-Level Association Statistics (Melissa McGuirl, Brown)
16. *Learning as shooting (Vincent Roulet, UWash)* [1]
17. A Geometric Variational Approach to Bayesian Inference (Abhijoy Saha, OSU)
18. Multi-Level Representation of Large-Scale Network Datasets (Faryad Sahneh, UA)
19. Optimal balancing of time-dependent confounders for marginal structural models (Michele Santacatterina, Cornell)
20. Constraint Programming Pushes the Limits of NMR Spectroscopy (Benjamin Sherman, UCSC)
21. Graph Reconstruction by Discrete Morse Theory (Ryan Slechta, OSU)
22. *Simplifying Filtrations for Persistent Homology via Edge Contraction (Ryan Slechta, OSU)* [2]
23. Trajectory Planning for Autonomous Vehicles for Optimal Environmental Monitoring (Sisi Song, UCSC)
24. Escaping saddle points efficiently in constrained optimization problems (Yue Sun, UWash)
25. Phase retrieval via randomized Kaczmarz: theoretical guarantees (Yan Shuo Tan, UCB)
26. *Provable and Practical Distance Estimation and Nearest Neighbor Search in Limited Space (Tal Wagner, MIT)* [3]
27. Faster convex optimization with higher-order smoothness via rescaled and accelerated gradient flows (Andre Wibisono, GATech)
28. Distance-Based Independence Screening for Canonical Analysis (Chuanping Yu, GATech)
29. Spherical Latent Factor Model (Xingchen Yu, UCSC)
30. An Alternative View: When Does SGD Escape Local Minima? (Yang Yuan, MIT)
31. Nonconvex Sparse Blind Deconvolution (Yuqian Zhang, Cornell)

05:50 - 07:00 Reception and poster discussions

07:00 - 07:30 A Historical Perspective on the Emerging Field of Data Science (Cathryn Carson, UCB)

07:30 - 08:30 Reception and poster discussions

[1] Poster canceled

[2] Poster will be presented, but no spotlight talk will be given

[3] Presenting a lighting talk in Research Session 1 instead of a poster spotlight

Wednesday, 10/24

08:30 - 10:00 Coffee Break

08:30 - 10:00 Data Science Education. Leader: Helen Zhang (UA). Speakers: Stephen Wright (UWisc), Michael Mahoney (UCB)

10:00 - 10:30 Coffee Break

10:30 - 12:00 Institution and Infrastructure Panel. Moderator: Abel Rodriguez (UCSC). Panelists: Hélène Barcelo (MSRI), Jeff Brock (ICERM), Bob Brown (Purdue NSF STC), David Ribes (UWash), Dimitri Shlyakhtenko (IPAM).

12:00 - 01:30 Lunch - PM & PI Closed Session

01:30 - 2:30 Research Session 4: Responsible Data Science. Leader: Lise Getoor (UCSC)

- 01:30 - 01:40 Introduction to Responsible Data Science (Lise Getoor, UCSC)
- 01:40 - 02:00 A Research Agenda on Foundations of Responsible Data Science (FoRDS) (Bill Howe, UWash)
- 02:00 - 02:10 Optimal balancing of time-dependent confounders for marginal structural models (Michele Santacatterina, Cornell)
- 02:10 - 02:30 Discussion

02:30 - 3:00 Closing discussion - Next steps

Research Session Abstracts

Randomized & Stochastic Algorithms Research Session

Overview of stochastic algorithms and their convergence analysis
Katya Scheinberg, Lehigh University

We will give a brief description of the rich variety of stochastic algorithms that have been proposed in the literature for machine learning problems and we will outline what sort of convergence guarantees have been derived for these algorithms.

Overview of randomized (sub-linear) algorithms
Piotr Indyk, MIT

This talk will complement Katya's introduction, providing the background for the randomized algorithms covered in this session. I will start from a brief overview of randomized algorithms, and then zoom into algorithms that use sub-linear time, space or communication.

Gradient Sampling Methods for Nonconvex Nonsmooth Optimization
Frank E. Curtis, Lehigh University

We discuss a relatively new class of algorithms for minimizing nonconvex and nonsmooth functions. The central idea is to approximate the minimum norm element of the subdifferential of a function at a point through random sampling of gradients. Over the past decade since the basic framework was proposed and analyzed, there have been various enhancements and extensions, e.g., toward the development of adaptive, second-order variants and approaches for solving constrained problems. Gradient sampling methods have some connections with randomized subgradient approaches, but have some practical advantages.

Newton-Type Methods for Non-Convex Optimization Under Inexact Hessian Information
Michael Mahoney, UC Berkeley

We consider variants of trust-region and cubic regularization methods for non-convex optimization, in which the Hessian matrix is approximated. Under mild conditions on the inexact Hessian, and using approximate solution of the corresponding sub-problems, we provide iteration complexity to achieve ε -approximate second-order optimality which have shown to be tight. Our Hessian approximation conditions constitute a major relaxation over the existing ones in the literature. Consequently, we are able to show that such mild conditions allow for the construction of the approximate Hessian through various random sampling methods. In this light, we consider the canonical problem of finite-sum minimization, provide appropriate uniform and non-uniform sub-sampling strategies to construct such Hessian approximations, and obtain optimal iteration complexity for the corresponding sub-sampled trust-region and cubic regularization methods.

Convergence Rates of Stochastic Algorithms in Nonsmooth Nonconvex Optimization
Dmitriy Drusvyatskiy, University of Washington

Stochastic iterative methods lie at the core of large-scale optimization and its modern applications to data science. Though such algorithms are routinely and successfully used in practice, few performance guarantees are available outside of smooth or convex settings. In this talk, I will describe a framework for designing and analyzing stochastic methods on a large class of nonsmooth and nonconvex problems, with provable efficiency guarantees. The problem class subsumes such important tasks as phase retrieval, robust PCA, and minimization of risk measures, while the methods include stochastic subgradient, Gauss-Newton, and proximal point iterations. The main thread of the proposed framework is appealingly intuitive: numerous stochastic methods can be interpreted as inexact gradient descent on an implicit smoothing of the problem.

Differentially Private Identity and Equivalence Testing of Discrete Distributions
Maryam Aliakbarpour, MIT

We study the fundamental problems of identity and equivalence testing over a discrete population from random samples. Our goal is to develop efficient testers while guaranteeing differential privacy to the individuals of the population. We provide sample-efficient differentially private testers for these problems. Our theoretical results significantly improve over the best known algorithms for identity testing, and are the first results for private equivalence testing. The conceptual message of our work is that there exist private hypothesis testers that are nearly as sample-efficient as their non-private counterparts. We perform an experimental evaluation of our algorithms on synthetic data. Our experiments illustrate that our private testers achieve small type I and type II errors with sample size *sublinear* in the domain size of the underlying distributions.

Rethinking learning rate schedules for stochastic optimization
Rahul Kidambi, University of Washington

There is a stark disparity between the learning rate schedules used in the practice of large scale machine learning and what are considered admissible learning rate schedules prescribed in the theory of stochastic approximation. Recent results, such as in the 'super-convergence' methods which use oscillating learning rates, serve to emphasize this point even more. One plausible explanation is that non-convex neural network training procedures are better suited to the use of fundamentally different learning rate schedules, such as the ``cut the learning rate every constant number of epochs'' method (which more closely resembles an exponentially decaying learning rate schedule); note that this widely used schedule is in stark contrast to the polynomial decay schemes prescribed in the stochastic approximation literature, which are indeed shown to be (worst case) optimal for classes of convex optimization problems.

The main contribution of this work shows that the picture is far more nuanced, where we do not even need to move to non-convex optimization to show other learning rate schemes can be far more effective. In fact, even for the simple case of stochastic linear regression with a fixed time horizon, the rate achieved by any polynomial decay scheme is sub-optimal compared to the statistical minimax rate (by a factor of condition number); in contrast the ''cut the learning rate every constant number of epochs'' provides an exponential improvement (depending only logarithmically on the condition number) compared to any polynomial decay scheme. Finally, it is

important to ask if our theoretical insights are somehow fundamentally tied to quadratic loss minimization (where we have circumvented minimax lower bounds for more general convex optimization problems)? Here, we conjecture that recent results which make the gradient norm small at a near optimal rate, for both convex and non-convex optimization, may also provide more insights into learning rate schedules used in practice.

Approximate Nearest Neighbors in Limited Space
Tal Wagner, MIT

We consider the (1 +eps)-approximate nearest neighbor search problem: given a set X of n points in a d-dimensional space, build a data structure that, given any query point y, finds a point $x \in X$ whose distance to y is at most $(1 + \text{eps}) \min_{x \in X} ||x - y||$ for an accuracy parameter $\text{eps} \in (0, 1)$. Our main result is a data structure that occupies only O(1/eps^2 n log(n) log(1/eps)) bits of space, assuming all points coordinates have O(log n) bits of precision. This improves over the best previously known space bound obtained via the celebrated randomized dimensionality reduction method due to Johnson and Lindenstrauss.

Distributed Learning over a Network of Agents under Communication Constraints
Thinh T. Doan, Georgia Tech

The rapid development of low-cost sensors, smart devices, communication networks, and learning algorithms has enabled data driven decision making in large-scale multi-agent systems; prominent examples include mobile robotic networks, smart grids, and autonomous systems. The key challenge in these systems is in handling the vast quantities of information shared between the agents in order to find an optimal or near-optimal policy that maximizes an objective function. This needs to be done under computation and communication constraints. Distributed information processing, which is not only amenable to low cost implementation but can also be implemented in real-time, has been recognized as an important approach to address this challenge.

The main contribution of this work is to propose a new variant of distributed consensus-based gradient methods, which gives us more flexibility in handling communication constraints over the network. In particular, we first show that such method is applicable to the case when the agents are only allowed to exchange their quantized values due to their finite communication bandwidth. We provide an explicit formula for its rates of convergence as a function on the underlying network topology and the communication capacity shared between agents. Second, the method can also be used to study the impact of communication delays over the network. Finally, one can view the proposed method as a distributed variant of the popular stochastic approximation, which is applicable to broad applications in many areas.

Learning-Based Frequency Estimation Algorithms
Ali Vakilian, MIT

Estimating the frequencies of elements in a data stream is a fundamental task in data analysis and machine learning. The problem is typically addressed using streaming algorithms which can process very large data using limited storage. Today's streaming algorithms, however, cannot exploit patterns in their input to improve performance. We propose a new class of algorithms that automatically learn relevant patterns in the input data and use them to improve its frequency estimates. The proposed algorithms combine the benefits of machine learning with the formal

guarantees available through algorithm theory. We prove that our learning-based algorithms have lower estimation errors than their non-learning counterparts. We also evaluate our algorithms on two real-world datasets and demonstrate empirically their performance gains.

Nonconvex Optimization & Deep Learning Research Session

Overview: Lessons Learned from Recent Georgia Tech and Wisconsin-Washington Workshops on Nonconvex Optimization & Deep Learning
Xiaoming Huo, Georgia Tech, Stephen Wright, University of Wisconsin-Madison, and Maryam Fazel, University of Washington

A Smoother Way to Train Structured Prediction Models
Zaid Harchaoui. University of Washington

We present a framework allowing one to perform smoothing on the inference used by structured prediction methods. Smoothing breaks the non-smoothness inherent to structured prediction objectives, without the need to resort to convex duality, and paves the way to the use of fast primal gradient-based optimization algorithms. We illustrate the proposed framework by developing an novel primal incremental gradient-based optimization algorithm for the structural support vector machine. The algorithm blends an extrapolation scheme for acceleration and an adaptive smoothing scheme for gradient-based optimization. We establish its worst-case complexity bounds. We present experiment results on two real-world problems, namely named entity recognition and visual object localization. Experimental results show that the proposed framework allows one to develop competitive primal optimization algorithms for structured prediction efficiently leveraging inference routines.

Implicit Self-Regularization in Deep Neural Networks: Evidence from Random Matrix Theory and Implications for Learning
Michael Mahoney, UC Berkeley

Random Matrix Theory (RMT) is applied to analyze the weight matrices of Deep Neural Networks (DNNs), including both production quality, pre-trained models such as AlexNet and Inception, and smaller models trained from scratch, such as LeNet5 and a miniature-AlexNet. Empirical and theoretical results clearly indicate that the DNN training process itself implicitly implements a form of *Self-Regularization*, implicitly sculpting a more regularized energy or penalty landscape. In particular, the empirical spectral density (ESD) of DNN layer matrices displays signatures of traditionally-regularized statistical models, even in the absence of exogenously specifying traditional forms of explicit regularization, such as Dropout or Weight Norm constraints. Building on relatively recent results in RMT, most notably its extension to Universality classes of Heavy-Tailed matrices, and applying them to these empirical results, we develop a theory to identify *5+1 Phases of Training*, corresponding to increasing amounts of *Implicit Self-Regularization*. These phases can be observed during the training process as well as in the final learned DNNs. For smaller and/or older DNNs, this implicit Self-Regularization is like traditional Tikhonov regularization, in that there is a ``size scale'' separating signal from noise. For state-of-the-art DNNs, however, we identify a novel form of *Heavy-Tailed Self-Regularization*, similar to the self-organization seen in the statistical physics of disordered systems. This results from correlations arising at all size scales, which for DNNs arises implicitly due to the training process itself. This Implicit Self-Regularization can depend strongly on the many knobs of the

training process. In particular, by exploiting the generalization gap phenomena, we demonstrate that we can cause a small model to exhibit all 5+1 phases of training simply by changing the batch size. This demonstrates that---all else being equal---DNN optimization with larger batch sizes leads to less-well implicitly-regularized models, and it provides an explanation for the generalization gap phenomena. Our results suggest that large, well-trained DNN architectures should exhibit Heavy-Tailed Self-Regularization, and we discuss the theoretical and practical implications of this.

The Gap Between Worst-Case Complexity and Empirical Performance in Nonconvex Optimization
Frank E. Curtis, Lehigh University

It has long been observed that better worst-case complexity guarantees in the context of convex optimization often (though not always) translates into improved empirical performance. Now that worst-case complexity has become a central focus of research on algorithms for solving nonconvex problems, the question becomes: Does improved complexity lead to improved performance? Unfortunately, little evidence has been provided that clearly justifies this translation. In this work, we suggest that part of the issue is the manner in which worst-case complexity is measured in nonconvex optimization. We propose a new strategy for characterizing performance that may better capture the behavior of algorithms in practice.
Paper: https://arxiv.org/abs/1802.01062

Sharp Convergence Rates for Langevin Dynamics in the Nonconvex Setting
Michael Jordan, UC Berkeley

We study the problem of sampling from a distribution for which the negative logarithm of the target density is L-smooth everywhere and m-strongly convex outside a ball of radius R, but potentially nonconvex inside this ball. We study both overdamped and underdamped Langevin MCMC and establish upper bounds on the time required to obtain a sample from a distribution that is within ε of the target distribution in 1-Wasserstein distance. For the first-order method (overdamped Langevin MCMC), the time complexity is $\tilde{\mathcal{O}}\lrp{e^{cLR^2}\frac{d}{\epsilon^2}}$, where d is the dimension of the underlying space. For the second-order method (underdamped Langevin MCMC), the time complexity is $\tilde{\mathcal{O}}\lrp{e^{cLR^2}\frac{\sqrt{d}}{\epsilon}}$ for an explicit positive constant c. Surprisingly, the convergence rate is only polynomial in the dimension d and the target accuracy ε.
Links: https://arxiv.org/abs/1805.01648

Convergence of Newton-Conjugate-Gradient Algorithms to Second-Order Points
Steve Wright, University of Wisconsin-Madison

The Newton-CG method is a classical approach to minimization of smooth functions of many variables. One attractive feature is that it does not require explicit evaluation of the Hessian, only the ability to evaluate a Hessian-vector product with an arbitrary vector, an operation whose complexity is comparable to an evaluation of the gradient. We describe a line-search variant of this approach that also exploits negative-curvature directions, in a framework that guarantees convergence to approximate second-order necessary points with complexity guarantees. Computational results shed light on the practical performance of this method and several competitors.
Link: https://arxiv.org/abs/1803.02924

Convergence of Policy Gradient Methods for the Linear Quadratic Regulator
Maryam Fazel, University of Washington

Policy gradient methods for reinforcement learning and continuous control are popular in practice, but lack theoretical guarantees even for the simplest case of linear dynamics and a quadratic cost, i.e., the Linear Quadratic Regulator (LQR) problem. A difficulty is that unlike the classical approaches, these methods must solve a nonconvex optimization problem to find the optimal control policy. We show that despite the nonconvexity, gradient descent starting from a stabilizing policy converges to the globally optimal policy, and we discuss how this can help understand policy gradient type methods.
Link: http://proceedings.mlr.press/v80/fazel18a/fazel18a.pdf

Nonconvex Geometry and Algorithm for Sparse Blind Deconvolution
John Wright, Columbia

We consider the problem of modeling a given dataset as superpositions of basic motifs. This simple model arises from several important applications, including microscopy image analysis, neural spike sorting, and image deblurring. This motif-finding problem can be phrased as "short-and-sparse" blind deconvolution, in which the goal is to recover a short motif (convolution kernel) from its convolution with a random spike train. We assume the kernel to have unit Frobenius norm, and formulate it as a nonconvex optimization problem over the sphere. By analyzing the optimization landscape, we argue that when the target spike train is sufficiently sparse, then on a region of the sphere, every local minimum is equivalent to the ground truth. This geometric characterization implies that efficient methods obtain the ground truth under the same conditions.

Topological and Geometric Data Analysis Research Session

Introduction to Topological Data Analysis
Tamal Dey, Ohio State University

In recent years, Data analysis based on topological methods has proved to be effective because of its footing on solid mathematical and algorithmic theories, robustness against noise and scale invariance. We will introduce the basic methodology with some examples.

The use of topological methods in data analysis
Yusu Wang, Ohio State University

Topological objects and ideas can capture certain essential structures behind not just domains where data are sampled from, but also functions / maps defined on them. In this talk, we aim to provide some examples to demonstrate where and how topological ideas may help in data analysis. In particular, we will show how topological methods help to provide (1) a generic yet flexible framework for feature vectorization, (2) tools / algorithms to identify and provide hidden structures; and (3) visual analytic platforms for high dimensional data exploration.

From TGDA for 3D Shapes, dynamic shapes, and networks
Facundo Memoli, Ohio State University

We'll describe different applications of persistent homology features to characterization of 3D shapes, dynamic shapes, and classification of network valued data.

Stochastic processes on and of networks
David Sivakoff, Ohio State University

Classical results on discrete, spatial stochastic processes are limited to lattices and trees. Driven by applications (e.g., to epidemiology and social science), there has been a lot of interest in stochastic models on complex networks, and most recently, on dynamic networks. We will give an overview of the theoretical state of the art and future challenges.

On Symplectic Optimization
Michael Jordan, UC Berkeley

Accelerated gradient methods have had significant impact in machine learning -- in particular the theoretical side of machine learning -- due to their ability to achieve oracle lower bounds. But their heuristic construction has hindered their full integration into the practical machine-learning algorithmic toolbox, and has limited their scope. In this paper we build on recent work which casts acceleration as a phenomenon best explained in continuous time, and we augment that picture by providing a systematic methodology for converting continuous-time dynamics into discrete-time algorithms while retaining oracle rates. Our framework is based on ideas from Hamiltonian dynamical systems and symplectic integration. These ideas have had major impact in many areas in applied mathematics, but have not yet been seen to have a relationship with optimization.

Linear and sublinear-time algorithms for optimal transport
Jonathan Weed, MIT

Optimal transport is a concept from analysis that has attracted a lot of recent attention in the data science community for its ability to handle data with geometric structure. I will describe a few recent ideas, inspired by statistical applications, which have led to very simple algorithms for solving optimal transport problems fast.

Using the tangent map to explain nonlinear dimension reduction
David Glickenstein, University of Arizona

Nonlinear dimension reduction techniques such as t-SNE are extremely important for tasks such as data visualization. However, the nonlinearity of these techniques often obscures what precisely the dimension reduction is doing. In fact, the nonlinearity can mean that the dimension reduction is capturing different information on each particular data set being visualized, which indicates that it could be beneficial to consider the differential map of tangents. We will explore using the differential of the dimension reduction map to probe what the dimension reduction is doing in a neighborhood of a particular data set.

Current and Future Role of Topological and Geometric Methods for Data Analysis
Sebastian Kurtek, Ohio State University

This talk will briefly discuss the role of topological and geometric methods in data analysis, and the opportunity for these methods to tackle important problems in various application areas including materials science and neuroscience.

Graphs Research Session

Foundational data science challenges for graphs
C. Seshadhri, UC Santa Cruz

This talk will cover some of the foundational challenges of data science research, in the context of graphs. This is a huge topic, so the talk merely provides my(opic) view. I will attempt to connect these challenges with the other session themes, as well as the research of other TRIPODS groups.

Variational Perspective on Local Graph Clustering
Michael Mahoney, UC Berkeley

Modern graph clustering applications require the analysis of large graphs and this can be computationally expensive. In this regard, local spectral graph clustering methods aim to identify well-connected clusters around a given "seed set" of reference nodes without accessing the entire graph. The celebrated Approximate Personalized PageRank (APPR) algorithm in the seminal paper by Andersen et al. is one such method. APPR was introduced and motivated purely from an algorithmic perspective. In other words, there is no a priori notion of objective function/optimality conditions that characterizes the steps taken by APPR. Here, we derive a novel variational formulation which makes explicit the actual optimization problem solved by APPR. In doing so, we draw connections between the local spectral algorithm of and an iterative shrinkage-thresholding algorithm (ISTA). In particular, we show that, appropriately initialized ISTA applied to our variational formulation can recover the sought-after local cluster in a time that only depends on the number of non-zeros of the optimal solution instead of the entire graph. In the process, we show that an optimization algorithm which apparently requires accessing the entire graph, can be made to behave in a completely local manner by accessing only a small number of nodes. This viewpoint builds a bridge across two seemingly disjoint fields of graph processing and numerical optimization, and it allows one to leverage well-studied, numerically robust, and efficient optimization algorithms for processing today's large graphs.
Link: https://arxiv.org/abs/1602.01886

Multi-Level Steiner Trees
Joe Watkins, University of Arizona

Given an undirected, connected graph $G=(V,E)$ with non-negative edge costs and a set of terminals $T \subseteq V$, the objective of the classical Steiner tree problem is to find a minimum-cost edge set $E' \subseteq E$ that spans the terminals. The problem is APX-hard with the best known approximation algorithm having a ratio of $\rho = \ln(4) + \varepsilon < 1.39$. We study a natural generalization, the multi-level Steiner tree (MLST) problem: given a nested sequence of terminals $T_1 \subset \cdots \subset T_k \subseteq V$, compute nested edge sets $E_1 \subseteq \cdots \subseteq E_k \subseteq E$ that span the corresponding terminal sets with minimum total cost. The MLST problem and variants thereof have been studied under names such as Quality-of-Service Multicast

tree, Grade-of-Service Steiner tree, and Multi-Tier tree with several known approximation results. We present two natural heuristics with approximation factor~$O(k)$ that lead to a composite algorithm that requires $2k$ Steiner tree computations. Its approximation ratio is solving using linear programming. We compare five algorithms experimentally on several classes of graphs using Erdos-Renyi, random geometric, Watts-Strogatz, and Barabasi-Albert network generation models for varying $|V|$, k, and terminal selection methods. We also implemented an integer linear program for MLST to provide ground truth. Our combined algorithm outperforms the others both in theory and in practice when the number of levels is up to $k \leq 22$.
Link: https://arxiv.org/abs/1804.02627

Riemannian embedding models for relational data
Abel Rodriguez, UC Santa Cruz

We describe a novel class of factor models for binary data. Rather than embedding multivariate discrete response onto a low dimensional Euclidean space, these models embed them into a more general (prespecified) low-dimensional manifold. This approach endows the model with greater expressive power without sacrificing interpretability. We will particularly focus on models for spherical embeddings, which are readily motivated by applications in political science.

Improved estimators for network-based sampling of hard-to-reach populations
Sebastien Roch, University of Wisconsin-Madison

To sample marginalized and/or hard-to-reach populations, respondent-driven sampling (RDS) and similar techniques reach their participants via peer referral. Under a Markov model for RDS, previous research has shown that if the typical participant refers too many contacts, then the variance of common estimators does not decay like O(n^{−1}), where n is the sample size. This implies that confidence intervals will be far wider than under a typical sampling design. Here we show that generalized least squares (GLS) can effectively reduce the variance of RDS estimates. In particular, a theoretical analysis indicates that the variance of the GLS estimator is O(n^{−1}). We also derive two classes of feasible GLS estimators. These results point the way to entirely different classes of estimators that account for the network structure beyond node degree. (Joint with Karl Rohe.)
Link: http://www.pnas.org/content/early/2018/09/24/1706699115

Estimating degree distributions, without see the whole graph
C. Seshadhri, UC Santa Cruz

The degree distribution is one of the most fundamental properties used in the analysis of massive graphs. There is a large literature on graph sampling, where the goal is to estimate properties (especially the degree distribution) of a large graph through a small, random sample. The degree distribution estimation poses a significant challenge, due to its heavy-tailed nature and the large variance in degrees. In contrast with previous statistical approaches, we apply recent mathematical techniques from the field of sublinear algorithms. This approach leads to an algorithm that has a complex sampling strategy, but we get provable bounds on its accuracy. A corollary of our main result is a provably sublinear algorithm for any degree distribution bounded below by a power law. The algorithm also behaves well on large real-world instances, typically observing less than 2% of the graph for accurate estimates at all scales of the degree distribution.
Link: https://arxiv.org/abs/1710.08607 (WWW 2018)

Evening Talk

A Historical Perspective on the Emerging Field of Data Science
Cathryn Carson, UC Berkeley

Abstract: Data science is a shimmering concept. We may not agree on exactly what it is –- but it gets at changes underway that are serious and real. Both in academic disciplines and out there in the world, something is emerging in the space where massive or pervasive data, its computational handling, and its analytical manipulation come together to underwrite inferential conclusions and actions in a "datafied" world. This talk tracks the emergence of this new field as an example of integrative, interdisciplinary work against a larger historical backdrop and in real contemporary experience.

Bio: Cathryn Carson is a professor of the History of Science at the University of California, Berkeley. She is currently serving as Faculty Lead of Berkeley's Data Science Education Program in the new Division of Data Sciences. She co-chaired the university's Data Science Education Rapid Action Team (2014-15) that articulated the broad-based, integrative vision for Berkeley's data science curriculum, and she chaired the Faculty Advisory Board of the campus-wide Data Science Planning Initiative (2015-16), which developed the blueprint for Berkeley's current organizational realignment around data science. Her research has dealt with the intellectual, cultural, and political history of the twentieth-century sciences, especially physics; the integration of social scientific and humanistic perspectives into engineering education; the organization and management of contemporary research universities; and the history and ethnography of data science.

Data Science Education Session

Session Chair: Helen Zhang, University of Arizona
Speakers: Steve Wright, University of Wisconsin-Madison and Michael Mahoney, UC Berkeley

Both speakers will talk about DS leadership summit report section on education and share their education experiences and activities of training DS undergraduates, graduates, and postdocs through theirs TRIPODS institutes. In particular, Stephen Wright talk about their Madison summer school, workshop, joint advisers, seminars, and designing a cross-disciplinary undergrad DS major. Michael will talk about the new data science major in Berkeley, the introductory mathematics of data class he is teaching, and hs experiences with PCMI and MMDS.

Responsible Data Science Research Session

Introduction to Responsible Data Science
Lise Getoor, UC Santa Cruz

I'll give a brief introduction to responsible data science, and highlight a few areas of special interest to the TRIPODS community.

A Research Agenda on Foundations of Responsible Data Science (FoRDS)
Bill Howe, University of Washington

As the deployment of automated decision tools in society continues to accelerate, their interactions with fundamental questions in law, in the social sciences, and in public policy have become impossible to ignore. Although this technology holds the promise of reducing costs, reducing errors, and improving objectivity, there is enormous potential for harm: As we train algorithms on biased data, we are amplifying, operationalizing, and, most insidiously, legitimizing the historical discrimination and opacity that the technology was in part intended to address.

We refer to this area of research as foundations of responsible data science. By "responsible," we refer to systems that take into account the ethical and epistemic issues associated with the larger social contexts in which they are deployed. These issues relate to the sources of data on which a system might be trained, the decision-making contexts in which it is deployed, and the potentially competing goals among stakeholders that use it. The goal is to build a community around a research agenda in sociotechnical systems that reflect our shared societal values.

In this short talk, I will highlight some of the open research questions in the area, including causality-based algorithms in machine learning to combat discrimination, interactive "nutritional labels" for datasets and models, the role of semi-synthetic datasets to enable collaboration analysis of sensitive data, and, more broadly, the role that technology should or should not play in addressing social questions.

In addition, I will describe a planned NSF workshop in February 2019 on this topic bringing together experts across computer science, statistics, social sciences, law, and policy to articulate a new foundational research agenda in responsible data science.

Optimal balancing of time-dependent confounders for marginal structural models
Michele Santacatterina, Cornell University

Marginal structural models (MSMs) estimate the causal effect of a time-varying treatment in the presence of time-dependent confounding via weighted regression. The standard approach of using inverse probability of treatment weighting (IPTW) can lead to high-variance estimates due to extreme weights and be sensitive to model misspecification. Various methods have been proposed to partially address this, including truncation and stabilized-IPTW to temper extreme weights and covariate balancing propensity score (CBPS) to address treatment model misspecification. In this paper, we present Kernel Optimal Weighting (KOW), a convex-optimization-based approach that finds weights for fitting the MSM that optimally balance time-dependent confounders while simultaneously controlling for precision, directly addressing the above limitations. KOW directly minimizes the error in estimation due to time-dependent confounding via a new decomposition as a functional. We further extend KOW to control for informative censoring. We evaluate the performance of KOW in a simulation study, comparing it with IPTW, stabilized-IPTW, and CBPS. We demonstrate the use of KOW in studying the effect of treatment initiation on time-to-death among people living with HIV and the effect of negative advertising on elections in the United States.
Link: https://arxiv.org/pdf/1806.01083.pdf

Poster abstracts

1. **Do we need 2nd order optimization methods in machine learning?**
   Albert Berahas, Lehigh University

   In this poster, we attempt to address two questions: (1) if and when 2nd order optimization methods are needed for training deep neural networks (DNNs), and (2) when are stochastic gradient (SG) methods sufficient. We discuss several challenges that arise when using stochastic and batch quasi-Newton (QN) methods for training DNNs. Finally, we present preliminary numerical experiments.

2. **A Unifying Framework of High-Dimensional Sparse Estimation with Difference-of-Convex (DC) Regularizations**
   Shanshan Cao, Georgia Institute of Technology

   Under the linear regression framework, we study the variable selection problem when the underlying model is assumed to have a small number of nonzero coefficients (i.e., the underlying linear model is sparse). The celebrated Least Absolute Shrinkage and Selection Operator (LASSO) approach introduces the L1 penalty, which has demonstrated practical effectiveness. To overcome the bias in an LASSO estimator, some non-convex penalties have been proposed to substitute the L1 penalty that has been used in LASSO. The representative ones include the smoothly clipped absolute deviation (SCAD), the minimax concave penalty (MCP), capped-L1, and many more. Recent work has pointed out that nearly all existing non-convex penalties can be represented as a difference-of-convex (DC) functions, which can be expressed as the difference of two convex functions, while itself may not be convex. There is a large existing literature on the optimization problems when their objectives and/or constraints involve DC functions. Efficient numerical solutions have been proposed. Under the DC framework, directional-stationary (d-stationary) solutions are considered, and they are usually not unique. In this paper, we show that under some mild conditions, a certain subset of d-stationary solutions in an optimization problem (with a DC objective) has some ideal statistical properties: namely, asymptotic estimation consistency, asymptotic model selection consistency, asymptotic efficiency. The aforementioned properties are the ones that have been proven by many researchers for a range of proposed non-convex penalties in the sparse estimation. Our assumptions are either weaker than or comparable with those conditions that have been adopted in other existing work. This work shows that DC is a nice framework to offer a unified approach to this existing work where non-convex penalty is involved, which bridges the communities of optimization and statistics.

3. **ATOMO: Communication-efficient Learning via Atomic Sparsification**
   Zachary Charles, University of Wisconsin-Madison

   Distributed model training suffers from communication overheads due to frequent gradient updates transmitted between compute nodes. To mitigate these overheads, several studies propose the use of sparsified stochastic gradients. We argue that these are facets of a general sparsification method that can operate on any possible atomic decomposition. Notable examples include element-wise, singular value, and Fourier decompositions. We present ATOMO, a general framework for atomic sparsification of stochastic gradients. Given a gradient, an atomic decomposition, and a sparsity budget, ATOMO gives a random unbiased sparsification of the atoms minimizing variance. We show that methods such as QSGD and TernGrad are special cases

of ATOMO and show that sparsifiying gradients in their singular value decomposition (SVD), rather than the coordinate-wise one, can lead to significantly faster distributed training.

4. **Random sampling and efficient algorithm for multiscale PDEs**
   Ke Chen, University of Wisconsin-Madison

   We describe an efficient framework for multiscale PDE problems that uses random sampling to capture low-rank local solution spaces arising in a domain decomposition framework. In contrast to existing techniques, our method does not rely on detailed analytical understanding of specific multiscale PDEs, in particular, their asymptotic limits. Our framework is applied to two specific problems - a linear kinetic equation and an elliptic equation with oscillatory media - for which recover the asymptotic preserving scheme and numerical homogenization, respectively. Numerical results confirm the efficacy of our approach.

5. **The Importance of Forgetting: Topological Insights from Neural Data**
   Samir Chowdhury, Ohio State University

   We develop of a line of work initiated by Curto and Itskov towards understanding the amount of information contained in the spike trains of hippocampal place cells via topology considerations. Previously, it was established that simply knowing which groups of place cells fire together in an animal's hippocampus is sufficient to extract the global topology of the animal's physical environment. We model a system where collections of place cells group and ungroup according to short-term plasticity rules. In particular, we obtain the surprising result that in experiments with spurious firing, the accuracy of the extracted topological information decreases with the persistence (beyond a certain regime) of the cell groups. This suggests that synaptic transience, or forgetting, is a mechanism by which the brain counteracts the effects of spurious place cell activity.

6. **Leveraged volume sampling for linear regression**
   Michal Derezinski, UC Berkeley

   Suppose an n x d design matrix in a linear regression problem is given, but the response for each point is hidden unless explicitly requested. The goal is to sample only a small number k << n of the responses, and then produce a weight vector whose sum of squares loss over all points is at most $1+\varepsilon$ times the minimum. When k is very small (e.g., k=d), jointly sampling diverse subsets of points is crucial. One such method called volume sampling has a unique and desirable property that the weight vector it produces is an unbiased estimate of the optimum. It is therefore natural to ask if this method offers the optimal unbiased estimate in terms of the number of responses k needed to achieve a $1+\varepsilon$ loss approximation. Surprisingly we show that volume sampling can have poor behavior when we require a very accurate approximation -- indeed worse than some i.i.d. sampling techniques whose estimates are biased, such as leverage score sampling. We then develop a new rescaled variant of volume sampling that produces an unbiased estimate which avoids this bad behavior and has at least as good a tail bound as leverage score sampling: sample size $k=O(d \log d + d/\varepsilon)$ suffices to guarantee total loss at most $1+\varepsilon$ times the minimum with high probability. Thus, we improve on the best previously known sample size for an unbiased estimator, $k=O(d^2/\varepsilon)$. Our rescaling procedure leads to a new efficient algorithm for volume sampling which is based on a *determinantal rejection sampling* technique with potentially broader applications to determinantal point processes. Other contributions include introducing the combinatorics needed for rescaled volume sampling and developing tail bounds for sums of dependent random matrices which arise in the process.

7. **Distributed Learning over a Network of Agents under Communication Constraints**
   Thinh Doan, Georgia Institute of Technology

   The rapid development of low-cost sensors, communication networks, and learning algorithms has enabled data driven decision making in large-scale multi-agent systems; prominent examples include robotic networks and smart grids. The key challenge in these systems is in handling the vast quantities of information shared between the agents in order to find an optimal policy that maximizes an objective function. This needs to be done under computation and communication constraints. Distributed information processing, which is not only amenable to low cost implementation but can also be implemented in real-time, has been recognized as an important approach to address this challenge. In our recent work, we study the performance of distributed stochastic variants of popular gradient methods when the agents are only allowed to exchange their quantized values due to the finite communication bandwidth. In particular, we provide an explicit formula for their rates of convergence, which shows the dependence on the underlying network topology and the communication capacity shared between agents.

8. **Large Scale Training of Neural Networks Using Robust Optimization and Hessian Information**
   Amir Gholami, UC Berkeley

   Stochastic Gradient Descent (SGD) methods using randomly selected batches are widely-used to train neural network (NN) models. Performing design exploration to find the best NN for a particular task often requires extensive training with different models on a large dataset, which is very computationally expensive. The most straightforward method to accelerate this computation is to distribute the batch of SGD over multiple processors. To keep the distributed processors fully utilized requires commensurately growing the batch size; however, large batch training often times leads to degradation in accuracy, poor generalization, and even poor robustness to adversarial attacks. Existing solutions for large batch training either significantly degrade accuracy or require massive hyper-parameter tuning. To address this issue, we propose a novel large batch training method which combines recent results in adversarial training (to regularize against "sharp minima") and second order optimization (to use curvature information to change batch size adaptively during training). We extensively evaluate our method on Cifar-10/100, SVHN, TinyImageNet, and ImageNet datasets, using multiple NNs, including residual networks as well as smaller networks for mobile applications such as SqueezeNext. Our new approach exceeds the performance of the existing solutions in terms of both accuracy and the number of SGD iterations (up to 1% and 5x, respectively). We emphasize that this is achieved without any additional hyper-parameter tuning to tailor our proposed method in any of these experiments.

9. **Bayesian Dynamic Feature Partition with Massive Data**
   Rene Gutierrez, UC Santa Cruz

   Bayesian computation of high dimensional regression models using the Markov Chain Monte Carlo (MCMC) or its variants is too slow or completely prohibitive since these methods perform costly computations at each iteration of the sampling chain. Furthermore, this computational cost cannot usually be efficiently divided across a parallel architecture. These problems are aggravated if the data size is massive or data arrives sequentially over time (streaming or online settings). This article proposes a novel dynamic feature partitioned (DFP) approach for efficient online inference for high dimensional regression with streaming data. DFP constructs a pseudo posterior density of the parameters at every time, followed by quickly updating the pseudo-posterior when a new

block of data arrives. The pseudo posterior at every time suitably partitions the parameter space to exploit parallelization for efficient posterior computation even with a big parameter space. The proposed approach is applied to high dimensional regression with Gaussian scale mixture priors and spike and slab priors on the large parameter space and is found to yield state-of-the-art inferential performance.

10. **Optimization in high-dimensional statistics with nonconvex constraints**
    Wooseok Ha, UC Berkeley

    Many problems in modern statistics can be formulated as an optimization problem with structured constraints, where the constraints often exhibit nonconvexity such as sparsity or low rank. However, working with nonconvex constraints presents challenges from both a theoretical and practical point of view. In this talk, we discuss a convergence behavior on two widely used algorithms, projected gradient descent and alternating minimization method, in the presence of nonconvex constraints. A major tool allowing to handle the nonconvex constraints is the local concavity coefficient, which aims to measure the concavity of a general nonconvex set. In the setting of alternating minimization, our result further reveals important distinction between alternating and non-alternating methods. We demonstrate our framework on a range of specific examples with rank-constrained variables, including factor model and multitask regression.

11. **Efficient Distributed Hessian Free Algorithm for Large-scale Empirical Risk Minimization via Accumulating Sample Strategy**
    Majid Jahani, Lehigh University

    In this paper, we propose a Distributed Accumulated Newton Conjugate gradiEnt (DANCE) method in which sample size is gradually increasing to quickly obtain a solution whose empirical loss is under satisfactory statistical accuracy. Our proposed method is multistage in which the solution of a stage serves as a warm start for the next stage which contains more samples (including the samples in the previous stage). The proposed multistage algorithm reduces the number of passes over data to achieve the statistical accuracy of the full training set. Moreover, our algorithm in nature is easy to be distributed and shares the strong scaling property indicating that acceleration is always expected by using more computing nodes. Various iteration complexity results regarding descent direction computation, communication efficiency and stopping criteria are analyzed under convex setting. Our numerical results illustrate that the proposed algorithm can outperform other comparable methods for training machine learning tasks including neural networks.

12. **Beyond Single Examples: Interpreting Black-Box Models via Set Influence**
    Rajiv Khanna, UC Berkeley

    Research in both machine learning and psychology suggests that salient examples can help humans to interpret learning models. To this end, we take a novel look at black box interpretation of test predictions in terms of training examples. Our goal is to ask ``which training examples are most responsible for a given set of predictions''? To answer this question, we make use of Fisher kernels as the defining feature embedding of each data point, combined with Sequential Bayesian Quadrature (SBQ) for efficient selection of examples. In contrast to prior work, our method is able to seamlessly handle any sized subset of test predictions in a principled way. We theoretically analyze our approach, providing novel approximation bounds for SBQ by relating it to recently developed theory of weak submodularity. We also present application of the proposed approach to three use cases: cleaning training data, fixing mislabeled examples and data summarization.

13. Towards improved design of parallel and accelerated stochastic gradient methods

Rahul Kidambi, University of Washington

Stochastic Gradient Descent (SGD) has turned out to be the workhorse method for large scale training of Machine Learning models, particularly for the current state of the art Deep Learning methods. Several impressive theory and practical efforts have taken strides towards speeding up SGD. In this vein, this poster presents recent results on the precise behavior of parallelization of SGD, by considering two variants, one based on mini-batching and the other based on model averaging. Our result indicates linear parallelization speedups offered by mini-batch SGD until some (efficiently computable) problem dependent batch size threshold. For model averaging, our result indicates (problem dependent) regimes on optimization when model averaging offers linear parallelization speedups on the excess risk. With regards to acceleration, this paper considers momentum methods, which are commonly used in practice to make SGD converge faster. This talk presents a counterpoint to the widely held belief that the momentum technique allows SGD to converge faster, especially when used with small batch sizes. We then present Accelerated SGD, which is an algorithm that is used to make SGD converge provably faster across small or large batch sizes.

14. An Efficient Pruning Algorithm for Robust Isotonic Regression

Cong Han Lim, Georgia Institute of Technology

We study a generalization of the classic isotonic regression problem where we allow separable nonconvex objective functions, focusing on the case of estimators used in robust regression. A simple dynamic programming approach allows us to solve this problem to within ε -accuracy (of the global minimum) in time linear in $1/\varepsilon$ and the dimension. We can combine techniques from the convex case with branch-and-bound ideas to form a new algorithm for this problem that naturally exploits the shape of the objective function. Our algorithm achieves the best bounds for both the convex case (linear in $\log(1/\varepsilon)$) and the general nonconvex case, while performing much faster in practice than a straightforward dynamic programming approach, especially as the desired accuracy increases.

15. Thresholded Hierarchical Clustering of Gene-Level Association Statistics

Melissa McGuirl, Brown University

Until recently, genome wide association (GWA) studies generally focused on a single phenotype, or observable characteristic of interest. Existing and emerging GWA datasets, merged with medical record and/or survey data, enable testing for associations for dozens of phenotypes, yet methods for characterizing the shared genetic architecture of multiple traits are still not well established. In this work, we present a new method, thresholded hierarchical clustering of gene-level association tests results, for characterizing shared and divergent aspects of genetic architecture among multiple phenotypes. Our goal is to identify clusters of phenotypes that share a core set of significant genes which can be detected even in the presence of noise. Simulations show that our clustering method is sensitive to shared genes that are significantly mutated in cases across phenotypes and can detect clusters that vary in size from 3 to 5 phenotypes. We apply our method to characterize the genetic architecture of 33 traits in 349,468 European-ancestry individuals from the UK Biobank, and identify two clusters of phenotypes that differentiate metabolic and immunological phenotypes. Our results identify multiple genes that influence the development of of these traits and diseases, and thus are potential targets for novel therapies.

16. Learning as shooting

Vincent Roulet, University of Washington

Canceled

17. A Geometric Variational Approach to Bayesian Inference

Abhijoy Saha, Ohio State University

We propose a novel Riemannian geometric framework for variational inference in Bayesian models based on the nonparametric Fisher-Rao metric on the manifold of probability density functions. Under the square-root density representation, the manifold can be identified with the positive orthant of the unit hypersphere and the Fisher-Rao metric reduces to the standard L^2 metric. Exploiting such a Riemannian structure, we formulate the task of approximating the posterior distribution as a variational problem on the hypersphere based on the alpha-divergence. This provides a tighter lower bound on the marginal distribution when compared to, and a corresponding upper bound unavailable with, approaches based on the Kullback-Leibler divergence. We propose a novel gradient-based algorithm for the variational problem based on Frechet derivative operators and examine its properties. Through simulations and real-data applications, we demonstrate the utility of the proposed geometric framework and algorithm on several Bayesian models.

18. Multi-Level Representation of Large-Scale Network Datasets

Faryad Sahneh, University of Arizona

Many algorithms, tools, and online services exist to analyze and visualize network data. However, few can be applied to large-scale networks in a robust way that is intuitive and informative. We propose a multi-level tool-set for analyzing and visualizing large-scale networks. The rationale is that in most real-world networks, nodes associated with values that can be interpreted as importance and can be used to determine membership in different level in an abstract, application-independent notation. The notion of levels allows scraping out layers of most significant and informative structures in a large-scale network, which otherwise would stay hidden.

19. Optimal balancing of time-dependent confounders for marginal structural models

Michele Santacatterina, Cornell University

Marginal structural models (MSMs) estimate the causal effect of a time-varying treatment in the presence of time-dependent confounding via weighted regression. The standard approach of using inverse probability of treatment weighting (IPTW) can lead to high-variance estimates due to extreme weights and be sensitive to model misspecification. Various methods have been proposed to partially address this, including truncation and stabilized-IPTW to temper extreme weights and covariate balancing propensity score (CBPS) to address treatment model misspecification. In this project, we present Kernel Optimal Weighting (KOW), a convex-optimization-based approach that finds weights for fitting the MSM that optimally balance time-dependent confounders while simultaneously controlling for precision, directly addressing the above limitations. KOW directly minimizes the error in estimation due to time-dependent confounding via a new decomposition as a functional. We further extend KOW to control for informative censoring. We evaluate the performance of KOW in a simulation study, comparing it with IPTW, stabilized-IPTW, and CBPS. We demonstrate the use of KOW in studying

the effect of treatment initiation on time-to-death among people living with HIV and the effect of negative advertising on elections in the United States.

20. Constraint Programming Pushes the Limits of NMR Spectroscopy

Benjamin Sherman, UC Santa Cruz

The set of all possible methyl-group Nuclear Overhauser Effect (NOE) interactions in a protein can be modeled as a graph G with a few hundred vertices. Nuclear Magnetic Resonance (NMR) experiments measuring these interactions return a degraded graph H, that contains about 90% of the vertices and 50% of the edges of G. Conventional wisdom has been that such experiments are no good for proteins above a certain size because the resulting inverse problems are severely underconstrained, so that most vertices of H have multiple valid preimages in G. We show that this conventional wisdom is wrong: in spite of the fact that local considerations do indeed make virtually all vertices of H appear severely underconstrained, we prove that more than 70% of vertices have a unique preimage. Discovering the existence of this rigid core requires global considerations and computation. By showing how this computation can be performed efficiently, we significantly extend the applicability of NMR spectroscopy in structural biology.

21. Graph Reconstruction by Discrete Morse Theory

Ryan Slechta, Ohio State University

Recovering hidden graph-like structures from potentially noisy data is a fundamental task in modern data analysis. Recently, a persistence-guided discrete Morse-based framework to extract a geometric graph from low-dimensional data has become popular. However, to date, there is very limited theoretical understanding of this framework in terms of graph reconstruction. This paper makes a first step towards closing this gap. Specifically, first, leveraging existing theoretical understanding of persistence-guided discrete Morse cancellation, we provide a simplified version of the existing discrete Morse-based graph reconstruction algorithm. We then introduce a simple and natural noise model and show that the aforementioned framework can correctly reconstruct a graph under this noise model, in the sense that it has the same loop structure as the hidden ground-truth graph, and is also geometrically close. We also provide some experimental results for our simplified graph-reconstruction algorithm.

22. Simplifying Filtrations for Persistent Homology via Edge Contraction

Ryan Slechta, Ohio State University

Persistent homology is a popular data analysis technique that is used to capture the changing topology of a filtration associated with some simplicial complex K. These topological changes are summarized in a p-dimensional persistence diagram. We propose two contraction operators which when applied to K and its associated filtration, bound the perturbation in this persistence diagram. The first assumes that the underlying space of K is a 2-manifold and ensures that simplices are paired with the same simplices in the contracted complex as they are in the original. The second is for arbitrary n-complexes, and bounds the bottleneck distance between the initial and final p-dimensional persistence diagrams. In addition, we show how the second operator can efficiently compose with itself across multiple contractions. The paper concludes with experiments demonstrating the operator's utility on terrains and a brief discussion of future directions for research.

23. Trajectory Planning for Autonomous Vehicles for Optimal Environmental Monitoring
Sisi Song, UC Santa Cruz

Our work considers real-time optimal trajectory planning for autonomous vehicles used in investigating environmental phenomena. We develop and implement algorithms that generate optimal trajectories to either (1) reconstruct the environmental field with minimal error, or (2) find the global maximum of the field. Our algorithms use Gaussian process priors to model the unknown field and Bayesian sequential experimental design methods, which involve developing utility functions that directly address the two operational goals.

24. Escaping saddle points efficiently in constrained optimization problems
Yue Sun, University of Washington

We consider finding an approximate second order stationary point of a smooth function in two kinds of constrained setting. Firstly, when there is only equality constraints $c_i(x)=0$, we show that a perturbed gradient projection algorithm converges in a number of iterations that depend polynomially on appropriate smoothness and curvature parameters for the cost and the constraints, and only polylogarithmically on the dimension. We also model it as smooth manifold and show that perturbed gradient retraction algorithm converges in a number of iterations that depend polynomially on smoothness of function, curvature parameters of manifold and error between retraction and exponential map. Secondly, for inequality constraints, based on copositivity test problem, we show that the convergence rate of perturbed gradient projection algorithm only exponentially depends on the number of degenerated constraints, which matches NP-hardness of the problem. The general constrained optimization setting can be solved by a divide conquer strategy based on manifold equality constrained problem.

25. Phase retrieval via randomized Kaczmarz: theoretical guarantees
Yan Shuo Tan, UC Berkeley

We consider the problem of phase retrieval, i.e. that of solving systems of quadratic equations. A simple variant of the randomized Kaczmarz method was recently proposed for phase retrieval, and it was shown numerically to have a computational edge over state-of-the-art Wirtinger flow methods. We provide the first theoretical guarantee for the convergence of the randomized Kaczmarz method for phase retrieval. We show that it is sufficient to have as many Gaussian measurements as the dimension, up to a constant factor. Along the way, we introduce a sufficient condition on measurement sets for which the randomized Kaczmarz method is guaranteed to work. We show that Gaussian sampling vectors satisfy this property with high probability; this is proved using a chaining argument coupled with bounds on Vapnik–Chervonenkis (VC) dimension and metric entropy.

26. Provable and Practical Distance Estimation and Nearest Neighbor Search in Limited Space
Tal Wagner, Massachusetts Institute of Technology

We study the problem of compressing a high-dimensional pointset, while retaining the ability to estimate all distances up to a distortion of $1+\varepsilon$, and to report approximate nearest neighbors for new query points. We present a data structure that occupies only $O(\varepsilon^{-2} n \log(n) \log(1/\varepsilon))$ bits of space, where n is the number of data points, and all point coordinates have magnitude $n^{O(1)}$. This improves over the previously best known space bound of $O(\varepsilon^{-2} n \log(n)^2)$, obtained by the randomized dimensionality reduction method of Johnson and Lindenstrauss (1984). Furthermore, we show that our bounds are nearly tight. We also present a practical version of our method,

which has comparable theoretical guarantees, but is much simpler and amenable to implementation. Our experiments show that with appropriately tuned parameters, our method produces compressed representations whose size is comparable to those produced by Product Quantization (Jegou et al., 2011), a state of the art metric compression method. Based on joint works with Piotr Indyk and Ilya Razenshteyn.

27. Faster convex optimization with higher-order smoothness via rescaled and accelerated gradient flows

Andre Wibisono, Georgia Institute of Technology

Accelerated gradient methods play a central role in optimization, achieving the optimal convergence rates in many settings. While many extensions of Nesterov's original acceleration method have been proposed, it is not yet clear what is the natural scope of the acceleration concept. In this work, we study accelerated methods from a continuous-time perspective. We show there is a Bregman Lagrangian functional that generates a large class of accelerated methods in continuous time, including (but not limited to) accelerated gradient descent, its non-Euclidean extension, and accelerated higher-order gradient methods. We show that in continuous time, these accelerated methods correspond to traveling the same curve in spacetime at different speeds. This is in contrast to the family of rescaled gradient flows, which correspond to changing the distance in space. We show how to implement both the rescaled and accelerated gradient methods as algorithms in discrete time with matching convergence rates. These algorithms achieve faster convergence rates for convex optimization under higher-order smoothness assumptions. We will also discuss lower bounds and some open questions. Joint work with Ashia Wilson (MSR) and Michael Jordan (UC Berkeley).

28. Distance-Based Independence Screening for Canonical Analysis

Chuanping Yu, Georgia Institute of Technology

A new method named Distance-based Independence Screening for Canonical Analysis (DISCA) is introduced to reduce dimensions of two random vectors with arbitrary dimensions. DISCA is based on the distance-based independence measure, also known as the distance covariance, proposed by Szekely and Rizzo in 2007. Unlike the existing canonical analysis methods, DISCA does not need the assumption that the dimension of the reduced subspaces of the two random vectors are equal. Besides, it can be applied to any types of distributions, continuous or discrete, light- or heavy-tailed. Our method can be solved efficiently by adopting the difference-of-convex (DC) optimization algorithm and the alternating direction method of multipliers (ADMM). It also avoids the potentially numerically-intensive bootstrap method to determine the dimension of the reduced subspaces.

29. Spherical Latent Factor Model

Xingchen Yu, UC Santa Cruz

Classical latent factor models used for scaling of binary and polychotomous responses in social sciences assume that latent factors live in low-dimensional Euclidean space. This helps make the resulting scales interpretable, but is rather restrictive. For example, in the political science literature, these models used to scale preferences show very conservative or liberal legislators as centrists. Therefore, we introduce a generalization of traditional models in which latent traits live in spheres/hyperspheres. We demonstrate that our model overcomes the limitation of the Euclidean space in such applications while maintaining the compelling properties of the Euclidean space.

30. An Alternative View: When Does SGD Escape Local Minima?

Yang Yuan, MIT

Stochastic gradient descent (SGD) is widely used in machine learning. Although being commonly viewed as a fast but not accurate version of gradient descent (GD), it always finds better solutions than GD for modern neural networks. In order to understand this phenomenon, we take an alternative view that SGD is working on the convolved (thus smoothed) version of the loss function. We show that, even if the function *f* has many bad local minima or saddle points, as long as for every point *x*, the weighted average of the gradients of its neighborhoods is one point convex with respect to the desired solution *x*∗, SGD will get close to, and then stay around *x*∗with constant probability. More specifically, SGD will not get stuck at "sharp" local minima with small diameters, as long as the neighborhoods of these regions contain enough gradient information. The neighborhood size is controlled by step size and gradient noise. Our result identifies a set of functions that SGD provably works, which is much larger than the set of convex functions. Empirically, we observe that the loss surface of neural networks enjoys nice one point convexity properties locally, therefore our theorem helps explain why SGD works so well for neural networks.

Link: https://arxiv.org/abs/1802.06175

31. Nonconvex Sparse Blind Deconvolution

Yuqian Zhang, Cornell University

Blind deconvolution is an ill-posed problem aiming to recover a convolution kernel and an activation signal from their convolution. We focus on the short and sparse variant, where the convolution kernel is short and the activation signal is sparsely and randomly supported. This variant models convolutional signals in several important application scenarios. The observation is invariant up to some mutual scaling and shift of the convolutional pairs. This scaled-shift symmetry is intrinsic to the convolution operator and imposes challenges for reliable algorithm design. We normalize the convolution kernel to have unit Frobenius norm and then cast the blind deconvolution problem as a nonconvex optimization problem over the sphere. We demonstrate that (i) under conditions, every local optimum is close to some shift truncation of the ground truth, and (ii) for a generic filter of length k, when the sparsity of activation signal satisfies $\theta < k^{2/3}$ and number of measurements m > poly(k), provable recovery of some shift truncation of the ground truth kernel can be obtained.

Panelist Biographies

Industry Panelists

Deepak Agarwal is a vice president of engineering at LinkedIn where he is responsible for all AI efforts across the company. He is well known for his work on recommender systems and has published a book on the topic. He has published extensively in top-tier computer science conferences and has coauthored several patents. He is a Fellow of the American Statistical Association and has served on the Executive Committee of Knowledge Discovery and Data Mining (KDD). Deepak regularly serves on program committees of various conferences in the field of AI and computer science. He is also an associate editor of two flagship statistics journals.

Erin Ledell is the Chief Machine Learning Scientist at H2O.ai in Mountain View, California. Erin has a Ph.D. in Biostatistics with a Designated Emphasis in Computational Science and Engineering from University of California, Berkeley. Her research focuses on automatic machine learning, ensemble machine learning and statistical computing. She also holds a B.S. and M.A. in Mathematics. Before joining H2O.ai, she was the Principal Data Scientist at Wise.io (acquired by GE Digital in 2016) and Marvin Mobile Security (acquired by Veracode in 2012), and the founder of DataScientific, Inc.

Edo Liberty is a Director of Research at AWS and the manager Amazon AI Labs. The Lab is a mix of scientists and engineers who build cutting edge machine learning systems and services for AWS customers. They support SageMaker, Kinesis, QuickSight and other yet-to-be-released services from AWS. Until mid 2016 he was the Senior Research Director at Yahoo. He was the head of Yahoo's Independent Research in New York where he focused on scalable machine learning and data mining for Yahoo critical applications. He received his B.Sc in Physics and Computer Science from Tel Aviv university and his Ph.D in Computer Science from Yale University, under the supervision of Steven Zucker. After that, he was a Post-Doctoral fellow at Yale in Program in Applied Mathematics. His personal research interests include fast dimensionality reduction, clustering, streaming and online algorithms, machine learning, and large scale numerical linear algebra. He is especially fond of randomized algorithms and high dimensional geometry.

Peter Norvig is a Director of Research at Google Inc. Previously he was head of Google's core search algorithms group, and of NASA Ames's Computational Sciences Division, making him NASA's senior computer scientist. He received the NASA Exceptional Achievement Award in 2001. He has taught at the University of Southern California and the University of California at Berkeley, from which he received a Ph.D. in 1986 and the distinguished alumni award in 2006. He was co-teacher of an Artifical Intelligence class that signed up 160,000 students, helping to kick off the current round of massive open online classes. His publications include the books *Artificial Intelligence: A Modern Approach* (the leading textbook in the field), *Paradigms of AI Programming: Case Studies in Common Lisp*, *Verbmobil: A Translation System for Face-to-Face Dialog*, and *Intelligent Help Systems for UNIX*. He is also the author of the Gettysburg Powerpoint Presentation and the world's longest palindromic sentence. He is a fellow of the AAAI, ACM, California Academy of Science and American Academy of Arts & Sciences.

Ashok N. Srivastava, Ph.D. is the Senior Vice President and Chief Data Officer at Intuit. He is responsible for setting the vision and direction for large-scale machine learning and AI across the enterprise to help power prosperity across the world. Previously, he was the VP of Big Data and Artificial Intelligence Systems and the Chief Data Scientist at Verizon. He is an Adjunct Professor at Stanford in the Electrical Engineering Department and was the Editor-in-Chief of the AIAA Journal of Aerospace Information Systems. Ashok is a Fellow of the IEEE, the American Association for the Advancement of Science (AAAS), and the American Institute of Aeronautics and Astronautics (AIAA). Ashok is the author of over 100 research articles in data mining, machine learning, and text mining, and has edited a book on Text Mining: Classification, Clustering, and Applications. He has won numerous awards including the IEEE Computer Society Technical Achievement Award for "pioneering contributions to intelligent information systems," the NASA Exceptional Achievement Medal for contributions to state-of-the-art data mining and analysis, the NASA Honor Award for Outstanding Leadership, the Distinguished Engineering Alumni Award from UC Boulder, the IBM Golden Circle Award, and the Department of Education Merit Fellowship.

Institution and Infrastructure Panelists

Hélène Barcelo is the Deputy Director of MSRI, a position she has held since July 1, 2008. As Deputy Director, she is in charge of overseeing all scientific activities at the Institute. A native of Québec, Canada, Hélène Barcelo received her PhD in mathematics in 1988 at the University of California, San Diego. After a three-year postdoctoral position at the University of Michigan, Ann Arbor, she moved to Arizona State University -Tempe campus (ASU). She is a Professor *Emerita* of Mathematics at ASU and a visiting scholar at the University of California, Berkeley. She received the Wexler Award (ASU) for distinguished teaching, and 4 doctoral and 8 master students completed their degree under her direction. She has held visiting positions at numerous universities and research institutes around the world. Professor Barcelo's research interests lie in algebraic combinatorics; more specifically, combinatorial representation theory and homotopy theories in relation to subspace arrangements. Prior to becoming Deputy Director, she spent the spring 2008 semester at MSRI as a Research Professor in the Combinatorial Representation Theory program; she was in residence at MSRI for the year-long

program in Combinatorics in 1996-97; and, she visited again in Fall 2004 during the Hyperplane Arrangements and Applications program. For many years (2001-09), Professor Barcelo was the Editor-in-Chief of the Journal of Combinatorial Theory, Series A, and she is currently a member of its advisory board. She has served on the executive committee of the International Conference in Formal Power Series and Algebraic Combinatorics (FPSAC) for several years (2001-08). She is also serving a four-year term on the Executive Committee of the American Mathematical Society and serves on its Committee on Publications.

Jeffrey Brock is Professor of Mathematics at Yale University, and Yale's inaugural Dean of Science (as of January, 2019). His research focuses on low dimensional geometry and topology, particularly hyperbolic geometry. His work on William Thurston's program to understand hyperbolic 3-manifolds led to their geometric classification and has developed into a study of geometric flows on their deformation spaces. More recently, he has developed interest in geometric and topological methods in analysis of large, complex data sets. He was an undergraduate at Yale, and obtained his Ph.D. at U.C. Berkeley, after which he held positions at Stanford and U. Chicago before moving to the Brown University Math Department, which he Chaired from 2013 to 2017. In 2016 he served as founding Director of Brown's Data Science Initiative, and as the lead PI on Brown's NSF TRIPODS institute grant. In his new role at Yale he will also serve as a co-PI on Yale's TRIPODS+X grant Investigations at the Interface of Data Science and Neuroscience. He was a Guggenheim Fellow in 2008 and is an elected Fellow of the American Mathematical Society.

Bob Brown is the Managing Director of the Center for Science of Information, a National Science Foundation Science and Technology Center. As managing director, Bob oversees all administrative aspects of the Center including budget and finance, programs, reporting, partner relations and personnel. The Center, with headquarters at Purdue University, advances the next generation of information theory through collaborative research and teaching. Supported by a National Science Foundation grant entitled "Emerging Frontiers of Science of Information," the Center for Science of Information is the first NSF-funded Science and Technology Center in Indiana. By assimilating elements of space, time, structure, semantics, and context, the Center deepens our understanding of information and applies these results to critical problems in society. Center partners include Bryn Mawr College, Howard University, Massachusetts Institute of Technology, Princeton University, Purdue University, Stanford University, Texas A&M University, University of California – Berkeley, University of California – San Diego, University of Hawaii, and University of Illinois. Bob has over twenty years of higher education administrative experience in the areas of grant management, finance, development and advancement. He received a B.S. and M.B.A. from Virginia Tech's Pamplin School of Business.

David Ribes is associate professor in the Department of Human Centered Design and Engineering (HCDE) and director of the Data Ecologies Lab (deLAB) at the University of Washington. He is a sociologist of science and technology who focuses on the development and sustainability of research infrastructures (i.e., networked information technologies for the support of interdisciplinary science); their relation to long-term changes in the conduct of science; and, transformations in objects of research. His current research investigates the emerging institutions of *data science* at multiple scales, such as changing scientific practices, budding regional or national organizations and novel public-private partnerships. David is regular contributor to the fields of Science and Technology Studies (STS) and Information Studies. His methods are ethnographic, archival-historical and comparative. See davidribes.com or dataecologi.es for more.

Dimitri (Dima) Shlyakhtenko was appointed Director of IPAM on July 1, 2017. Shlyakhtenko has been a member of the faculty at UCLA department of mathematics since 1998, and served as the department chair from 2012 to 2015. His research is on Operator Algebra and includes free probability theory, random matrix theory, as well as von Neumann algebras and L^2-invariants. He received his PhD from University of California, Berkeley in 1997 at the age of 22. He was the recipient of a Sloan Foundation Fellowship in 2001 and a Special Project Award from the Clay Mathematics Institute in 2002. The UCLA department of mathematics presented him with the R. Sorgenfrey Distinguished Teaching Award in 2004. He gave an invited talk at the International Congress of Mathematicians in Hyderabad, India in 2010.