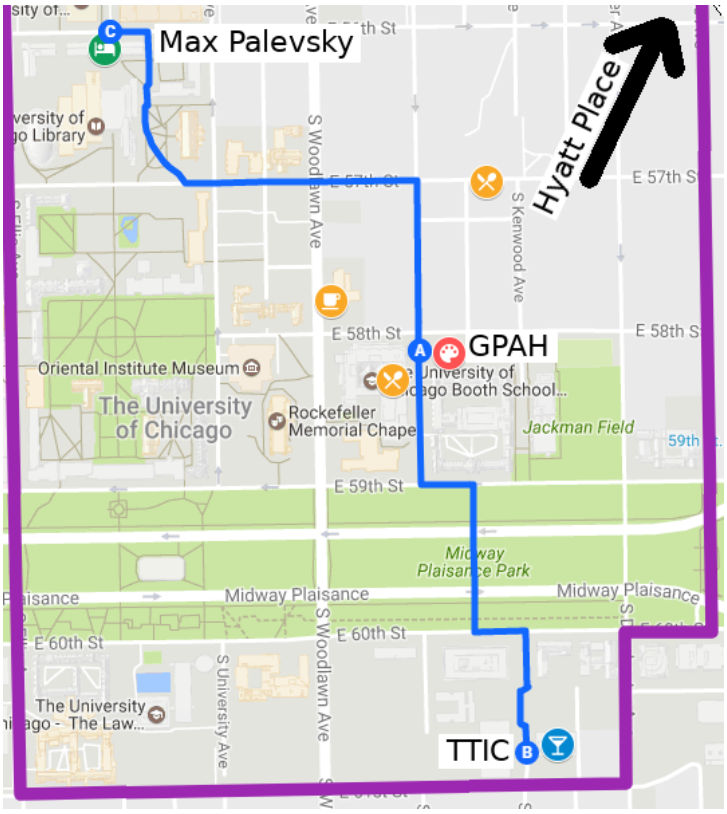
MMLS 2017 Booklet

Monday, June 19		
9:30 - 11:15	TTIC	Continental breakfast. (TTIC Colloquium: 10:00-11:00.)
11:15 - 11:30	GPAH	Opening remarks: Po-Ling Loh.
11:30 - 12:20	GPAH	**Plenary speaker:** Devavrat Shah . (Chair: Po-Ling Loh .) *Latent Variable Model Estimation via Collaborative Filtering.*
12:20 - 2:50	GPAH	Lunch, posters.
2:50 - 3:30	GPAH	**Invited talks** (chair: Mesrob Ohannessian). **2:50:** Rina Foygel *Projected Gradient Descent with Nonconvex Constraints.* **3:10:** Maxim Raginsky *Non-Convex Learning via Stochastic Gradient Langevin Dynamics.*
3:30 - 4:00	GPAH	Coffee Break.
4:00 - 4:50	GPAH	**Plenary speaker:** Rayid Ghani . (Chair: Nati Srebro .) *Machine Learning for Public Policy: Opportunities and Challenges*
4:50 - 5:30	GPAH	**Invited talks** (chair: Laura Balzano). **4:50:** Dimitris Papailiopoulos *Gradient Diversity Empowers Distributed Learning.* **5:10:** Alan Ritter *Large-Scale Learning for Information Extraction.*
5:45 - 7:00	TTIC	Reception, with remarks by Sadaoki Furui (TTIC President).
Tuesday, June 20		
8:30 - 9:50	GPAH	Continental breakfast.
9:00 - 9:50	GPAH	~~**Bonus speaker:** Larry Wasserman . (Chair: Mladen Kolar .)~~ ~~*Locally Optimal Testing.*~~ [**Cancelled.**]
9:50 - 10:50	GPAH	**Invited talks** (chair: Misha Belkin). **9:50:** Srinadh Bhojanapalli *Effectiveness of Local Search for Low Rank Recovery* **10:10:** Niao He *Learning From Conditional Distributions.* **10:30:** Clayton Scott *Nonparametric Preference Completion.*
10:50 - 11:20	GPAH	Coffee break.
11:20 - 12:20	GPAH	**Invited talks** (chair: Jason Lee). **11:20:** Lev Reyzin *On the Complexity of Learning from Label Proportions.* **11:40:** Ambuj Tewari *Random Perturbations in Online Learning.* **12:00:** Risi Kondor *Multiresolution Matrix Factorization.*
12:20 - 1:50	GPAH	Lunch, posters.
1:50 - 2:40	GPAH	**Plenary speaker:** Corinna Cortes . (Chair: Matus Telgarsky .) *Harnessing Neural Networks.*
2:40 - 2:50	GPAH	Coffee break.
2:50 - 3:50	GPAH	Panel discussion.

Further MMLS info:
Please google "Midwest ML Symposium";
equivalently, visit `http://mjt.web.engr.illinois.edu/mmls_17/` .
Questions or concerns?
Find a co-organizer, or email <mmls2017@ttic.edu>.
Check the website for an exit survey!

From the annotated map link on `http://mjt.web.engr.illinois.edu/mmls_17/` :



We graciously thank our sponsors:

1 Plenary talks

- **Speaker:** Corinna Cortes (Google Research NYC)

  Title: Harnessing Neural Networks.

  Abstract: Deep learning has demonstrated impressive performance gain in many machine learning applications. However, uncovering and realizing these performance gains is not always straightforward. Discovering the right network architecture is critical for accuracy and often requires a human in the loop. Some network architectures occasionally produce spurious outputs, and the outputs have to be restricted to meet the needs of an application. Finally, realizing the performance gain in a production system can be difficult because of extensive inference times.

  In this talk we discuss methods for making neural networks efficient in production systems. We also discuss an efficient method for automatically learning the network architecture, called AdaNet. We provide theoretical arguments for the algorithm and present experimental evidence for its effectiveness.

  Bio: Corinna Cortes is the Head of Google Research, NY, where she is working on a broad range of theoretical and applied large-scale machine learning problems. Prior to Google, Corinna spent more than ten years at AT&T Labs - Research, formerly AT&T Bell Labs, where she held a distinguished research position. Corinna's research work is well-known in particular for her contributions to the theoretical foundations of support vector machines (SVMs), for which she jointly with Vladimir Vapnik received the 2008 Paris Kanellakis Theory and Practice Award, and her work on data-mining in very large data sets for which she was awarded the AT&T Science and Technology Medal in the year 2000. Corinna received her MS degree in Physics from University of Copenhagen and joined AT&T Bell Labs as a researcher in 1989. She received her Ph.D. in computer science from the University of Rochester in 1993.

- **Speaker:** Rayid Ghani (University of Chicago)

  Title: Machine Learning for Public Policy: Opportunities and Challenges

  Abstract: Can machine learning help reduce police violence and misconduct? Can it help prevent children from getting lead poisoning? Can it help cities better target limited resources to improve lives of citizens? We're all aware of the impact machine learning is making in the corporate word right now but it's impact in public policy is still in the early stages. In this talk, I'll discuss opportunities and challenges learned from our work at University of Chicago while working on dozens of projects over the past few years with non-profits and governments on high-impact social challenges. I'll discuss these examples and describe what machine learning and social science research challenges need to be tackled, and what tools and techniques need to be developed in order to have a social and policy impact with machine learning.

  Bio: Rayid Ghani is the Director of the Center for Data Science & Public Policy and a Senior Fellow at the Harris School of Public Policy and the Computation Institute at the University of Chicago. Rayid is a reformed computer scientist and wanna-be social scientist, but mostly just wants to increase the use of data-driven approaches in solving large public policy and social challenges. Among other areas, Rayid works with governments and non-profits in policy areas such as health, criminal justice, education, public safety, economic development, and urban infrastructure. Rayid is also passionate about teaching practical data science and started the Eric & Wendy Schmidt Data Science for Social Good Fellowship at UChicago that trains computer scientists, statisticians, and social scientists from around the world to work on data science problems with social impact.

  Before joining the University of Chicago, Rayid was the Chief Scientist of the Obama 2012 Election Campaign where he focused on data, analytics, and technology to target and influence voters, donors, and volunteers. Previously, Rayid was a Research Scientist and led the Machine Learning group at Accenture Labs. Rayid did his graduate work in Machine Learning at Carnegie Mellon University and is actively involved in organizing Data Science related conferences and workshops. In his ample free time, Rayid works with non-profits to help them with their data, analytics and digital efforts and strategy.

- **Speaker:** Devavrat Shah (MIT)

  Title: Latent Variable Model Estimation via Collaborative Filtering.

  Abstract: Much of modern data is generated by humans and drives decisions made in a variety of settings, such as recommendations for online portals, demand prediction in retail, matching buyers and sellers on social platforms, or denoising crowdsourced labels. Due to the complexities of human behavior, the precise data model is often unknown, creating a need for flexible models with minimal assumptions. A minimal property that is natural for many datasets is "exchangeability", i.e. invariant under relabeling of the dataset, which naturally leads to a nonparametric latent variable model a la Aldous and Hoover (early 1980s). The corresponding inference problem can be formulated as matrix or graphon estimation.

  In this talk, using inspiration from the classical Taylors expansion for differentiable functions, we shall propose an estimation algorithm that is syntactically identical to the basic version of the classical collaborative filtering for rec-

ommendation systems. We shall discuss statistical performance of the algorithm as well as its natural variation to overcome data sparsity.

The talk is based on joint works with (a) Christina Lee (MIT), Yihua Li (MS) and Dogyoon Song (MIT), and (b) Christina Borgs (MSR), Jennifer Chayes (MSR) and Christina Lee (MIT).

Bio: Devavrat Shah is a Professor with the department of Electrical Engineering and Computer Science at Massachusetts Institute of Technology. His current research interests are at the interface of Statistical Inference and Social Data Processing. His work has been recognized through prize paper awards in Machine Learning, Operations Research and Computer Science, as well as career prizes including 2010 Erlang prize from the INFORMS Applied Probability Society and 2008 ACM Sigmetrics Rising Star Award. He is a distinguished young alumni of his alma mater IIT Bombay.

- **Speaker:** Larry Wasserman (CMU)

  Title: Locally Optimal Testing.

  Abstract: We consider a fundamental problem in Statistics: testing if a sample was drawn from a given distribution. Despite the long history of this problem, much is still not known especially when we allow for high dimensions and low smoothness. In this setting, we find the local minimax testing rates and we give some tests that are optimal. We show that these tests have much higher power than standard tests. This is joint work with Sivaraman Balakrishnan.

  Bio: Larry Wasserman is Professor, Department of Statistics and Machine Learning Department, Carnegie Mellon University. In his spare time, he hunts big game.

2 Invited talks

- **Speaker:** Srinadh Bhojanapalli

  Title: Effectiveness of Local Search for Low Rank Recovery

  Abstract: Local search methods such as gradient descent have been extremely effective in finding good solutions for non-convex objectives such as in low rank recovery. In this talk we will present recent results characterizing this success of local search from both optimization and learning perspective.

  Bio: Srinadh Bhojanapalli obtained his Ph.D. in Electrical and Computer Engineering from The University of Texas at Austin in 2015. Prior to that he obtained Bachelors in Technology from Indian Institute of Technology Bombay in 2010. He has spent some summers as an intern at Microsoft research India and Ebay research labs. He is currently a research assistant professor at Toyota Technological Institute at Chicago, and as a second full-time job is Matus Telgarsky's best friend.

  His research is primarily focused on designing algorithms for large scale machine learning problems with rigorous statistical guarantees. He is interested in low rank recovery, neural networks, non-convex optimization and sublinear time algorithms.

- **Speaker:** Rina Foygel

  Title: Projected Gradient Descent with Nonconvex Constraints.

  Abstract: Nonconvex optimization arises in many applications of high-dimensional statistics and data analysis, where data models and regularization terms can both often exhibit nonconvexity. While convex programs for structured signal recovery have been widely studied, comparatively little is known about the theoretical properties of nonconvex optimization methods. In this talk I will discuss the problem of projected gradient descent over nonconvex constraints, where the local geometry of the constraint set is closely tied to its convergence behavior. By measuring the local concavity of the constraint set, we can give concrete guarantees for convergence of projected gradient descent. Furthermore, by relaxing these geometric conditions, we can allow for approximate calculation of the projection step to speed up the algorithm.

  Bio: Rina Foygel Barber is an Assistant Professor of Statistics at the University of Chicago. She received her PhD in Statistics from the University of Chicago in 2012, then was a NSF Postdoctoral Fellow at Stanford University in 2012-2013 before joining the faculty at Chicago in January 2014. Her research focuses on high-dimensional inference, sparse and low rank modeling, nonconvex optimization, and applications to medical imaging.

- **Speaker:** Niao He

  Title: Learning From Conditional Distributions.

  Abstract: While reinforcement learning received much attention recent years to address Markov decision problems, most existing reinforcement learning algorithms either are purely heuristic or lack sample efficiency. A key challenge boils down to the difficulty with learning from conditional distributions with limited samples and minimizing objectives

involving nested expectations. We propose a novel stochastic-approximation-based method that benefits from a fresh employ of Fenchel duality and kernel embedding techniques. To the best of our knowledge, this is the first algorithm that allows to take only one sample at a time from the conditional distribution and comes with provable theoretical guarantee. The proposed algorithm achieves the state-of-the-art empirical performances for the policy evaluation task on several benchmark datasets comparing to the existing algorithms such as gradient-TD2 and residual gradient.

Bio: Niao He is currently assistant professor with the Department of Industrial & Enterprise Systems Engineering and the Coordinated Science Laboratory at University of Illinois at Urbana-Champaign. She completed her M.S. in Computational Science & Engineering and Ph.D. in Operations Research from Georgia Institute of Technology in 2015. Her research interests span the areas of large-scale optimization and machine learning.

- **Speaker:** Risi Kondor

  Title: Multiresolution Matrix Factorization.

  Abstract: The sheer size of today's datasets dictates that learning algorithms compress or reduce their input data and/or make use of parallelism. Multiresolution Matrix Factorization (MMF) makes a connection between such computational strategies and some classical themes in Applied Mathematics, namely Multiresolution Analysis and Multigrid Methods. In particular, the similarity matrices appearing in data often have multiresolution structure, which can be exploited both for learning and to facilitate computation (joint work with Nedelina Teneva, Pramod Mudrakarta, Yi Ding, Jonathan Eskreis-Winkler and Vikas Garg).

  Bio: Risi Kondor is an assistant professor at the Computer Science and Statistics departments at The University of Chicago. Risi obtained his B.A. in Mathematics and Theoretical Physics from Cambridge, followed by an M.Sc. in Machine Learning from Carnegie Mellon and a PhD from Columbia, and postdoc positions at the Gatsby Unit (UCL) and Caltech.

- **Speaker:** Dimitris Papailiopoulos

  Title: Gradient Diversity Empowers Distributed Learning.

  Abstract: Distributed implementations of mini-batch SGD exhibit speedup saturation and decaying generalization beyond a particular batch-size. I will suggest that high similarity between concurrently processed gradients may be a cause of this performance degradation. I will introduce Gradient Diversity that measures the dissimilarity between concurrent gradient updates, and show its key role in the performance of mini-batch SGD. Through extensive experiments and analysis, we will see that problems with high gradient diversity are more amenable to parallelization and offer better generalization guarantees. I will then discuss how popular heuristics like dropout, quantization, and Langevin dynamics can improve it.

  Bio: Dimitris Papailiopoulos is an Assistant Professor of Electrical and Computer Engineering at the University of Wisconsin-Madison and a Faculty Fellow of the Grainger Institute for Engineering. Between 2014 and 2016, Papailiopoulos was a postdoctoral researcher in EECS at UC Berkeley and a member of the AMPLab. His research interests span machine learning, coding theory, and distributed algorithms, with a current focus on coordination-avoiding parallel machine learning and the use of erasure codes to speed up distributed computation. Dimitris earned his Ph.D. in ECE from UT Austin in 2014, under the supervision of Alex Dimakis. In 2015, he received the IEEE Signal Processing Society, Young Author Best Paper Award.

- **Speaker:** Maxim Raginsky

  Title: Non-Convex Learning via Stochastic Gradient Langevin Dynamics.

  Abstract: Stochastic Gradient Langevin Dynamics (SGLD) is a popular variant of Stochastic Gradient Descent, where properly scaled isotropic Gaussian noise is added to an unbiased estimate of the gradient at each iteration. This modest change allows SGLD to escape local minima and suffices to guarantee asymptotic convergence to global minimizers for sufficiently regular non-convex objectives. In this talk, I will present a nonasymptotic analysis in the context of non-convex learning problems, giving finite-time guarantees for SGLD to find approximate minimizers of both empirical and population risks. As in the asymptotic setting, the analysis relates the discrete-time SGLD Markov chain to a continuous-time diffusion process. A new tool that drives the results is the use of weighted transportation cost inequalities to quantify the rate of convergence of SGLD to a stationary distribution in the Euclidean 2-Wasserstein distance. This talk is based on joint work with Sasha Rakhlin and Matus Telgarsky.

  Bio: Maxim Raginsky received the B.S. and M.S. degrees in 2000 and the Ph.D. degree in 2002 from Northwestern University, all in electrical engineering. He has held research positions with Northwestern, University of Illinois at Urbana-Champaign (where he was a Beckman Foundation Fellow from 2004 to 2007), and Duke University. In 2012, he returned to UIUC, where he is currently an Assistant Professor and William L. Everitt Fellow in Electrical and Computer Engineering. He is also a faculty member of the Signals, Inference, and Networks (SINE) group and the Decision and Control group in the Coordinated Science Laboratory. Dr. Raginsky received a Faculty Early Career Development (CAREER) Award from the National Science Foundation in 2013. His research interests lie at the intersection of

information theory, machine learning, and control. He is a member of the editorial boards of Foundations and Trends in Communications and Information Theory and IEEE Transactions on Network Science and Engineering.

- **Speaker:** Lev Reyzin

  Title: On the Complexity of Learning from Label Proportions.

  Abstract: In the problem of learning with label proportions (also known as the problem of estimating class ratios), the training data is unlabeled, and only the proportions of examples receiving each label are given. The goal is to learn a hypothesis that predicts the proportions of labels on the distribution underlying the sample. This model of learning is useful in a wide variety of settings, including predicting the number votes for candidates in political elections from polls. In this talk, I will formalize the setting and discuss some of the results on the computational complexity of learning from label proportions. This is based on joint work with Benjamin Fish.

  Bio: Lev Reyzin is a tenure-track Assistant Professor in the MCS group at UIC's mathematics department. His research spans the theory and practice of machine learning. Previously, Lev was a Simons Postdoctoral Fellow at Georgia Tech, and before that, an NSF CI-Fellow atYahoo! Research, where he tackled problems in computational advertising. Lev received his Ph.D. on an NSF fellowship from Yale under Dana Angluin and his undergraduate degree from Princeton. His work has earned awards at ICML, COLT, and AISTATS and has been funded by the National Science Foundation and the Army Research Office.

- **Speaker:** Alan Ritter

  Title: Large-Scale Learning for Information Extraction.

  Abstract: Much of human knowledge is readily available on the internet, however natural language is notoriously difficult for computers to interpret. In this talk, I present some recent advances in extracting structured knowledge from text with an eye toward realtime information in massive data streams found on social media. I argue we can not exclusively rely on traditional methods that learn from small, hand-annotated datasets if we hope to extract a broad range of relations and events from diverse text genres at scale. As an alternative to human labeling, I will describe an approach that reasons about latent variables to learn robust information extraction models from large, opportunistically gathered datasets. As a concrete example, I will present a new approach to resolving time expressions without relying on any hand-coded rules or manually annotated data. By leveraging a database of known events as distant supervision, in conjunction with large quantities of in-domain data, our approach can outperform off-the-shelf resolvers on noisy user-generated data.

  Bio: Alan Ritter is an assistant professor in Computer Science at Ohio State University. His research interests include natural language processing, social media analysis and machine learning. Alan completed his PhD at the University of Washington and was a postdoctoral fellow in the Machine Learning Department at Carnegie Mellon University. He has received an NDSEG fellowship, a best student paper award at IUI, an NSF CRII and has served as an area chair for ACL, EMNLP and NAACL.

- **Speaker:** Clayton Scott

  Title: Nonparametric Preference Completion.

  Abstract: In the problem of collaborative preference completion, there is a pool of items, a pool of users, and a partially observed item-user rating matrix, and the goal is to recover the personalized ranking of items for each user. We investigate a nearest-neighbor type algorithm and establish its consistency under a flexible nonparametric model for the ratings. To our knowledge, this is the first consistency result for the collaborative preference completion problem in a nonparametric setting. We also report experiments on the Netflix and Movielens datasets suggesting that our algorithm has some advantages over existing neighborhood-based methods and that its performance is comparable to some state-of-the art matrix factorization methods. This is joint work with Julian Katz-Samuels.

  Bio: Clayton Scott is an associate professor in the Department of Electrical Engineering and Computer Science at the University of Michigan. He received his AB in Mathematics from Harvard University, and Masters and PhD in Electrical Engineering from Rice University. His research interests include statistical machine learning theory, methods, and applications.

- **Speaker:** Ambuj Tewari

  Title: Random Perturbations in Online Learning.

  Abstract: I will give a quick overview of some recent advances in our understanding of online learning methods that use random perturbations. For example, the classical algorithm known as Follow The Perturbed Leader (FTPL) can be viewed through the lens of stochastic smoothing, a tool that has proven popular within convex optimization. This perspective leads to a unified analysis of FTPL as well as clarifies connections with Follow The Regularized Leader (FTRL) algorithms. Furthermore, Radermacher complexity, a measure of the capacity of a function class to overfit, can be viewed through a perturbation perspective yielding an online algorithm called Follow the Sampled Leader (FTSL).

We recently established Hannan consistency of FTSL by making use of an anti-concentration result known as the Littlewood-Offord theorem. (Talk is based on joint work with Jacob Abernethy, Chansoo Lee, and Zifan Li.)

Bio: Ambuj Tewari is an assistant professor in the Department of Statistics and the Department of EECS (by courtesy) at the University of Michigan, Ann Arbor. His is also affiliated with the Michigan Institute for Data Science (MIDAS). He obtained his PhD under the supervision of Peter Bartlett at the University of California at Berkeley. His research interests lie in machine learning including statistical learning theory, online learning, reinforcement learning and control theory, network analysis, and optimization for machine learning. He collaborates with scientists to seek novel applications of machine learning in mobile health, learning analytics, and computational chemistry. His research has been recognized with paper awards at COLT 2005, COLT 2011, and AISTATS 2015. He received an NSF CAREER award in 2015 and a Sloan Research Fellowship in 2017.

3 Posters

1. **Author(s):** Albert Berahas (Northwestern University), Raghu Bollapragada, Jorge Nocedal

   Title: Are Newton-Sketch and subsampled Newton methods effective in practice?

   Abstract: The concepts of subsampling and sketching have recently received much attention by the optimization and statistics communities. In this talk, we focus on their numerical performance, with the goal of providing new insights into their theoretical and computational properties. We pay particular attention to Newton-Sketch and subsampled Newton methods, as well as techniques for solving the Newton equations approximately. Our tests are performed on a collection of optimization problems arising in machine learning.

2. **Author(s):** Rawad Bitar (Illinois Institute of Technology), Parimal Parag, Salim El Rouayheb

   Title: Secure distributed computing on untrusted workers

   Abstract: We consider the setting of a master server who possesses confidential data (genomic, medical data, etc.) and wants to run intensive computations on it, as part of a machine learning algorithm for example. The master wants to distribute these computations to untrusted workers who have volunteered or are incentivized to help with this task. However, the data must be kept private (in an information theoretic sense) and not revealed to the individual workers. The workers may be busy and will take a random time to finish the task assigned to them. We are interested in reducing the aggregate delay experienced by the master. We focus on linear computations as an essential operation in many iterative algorithms. A known solution is to use a linear secret sharing scheme to divide the data into secret shares on which the workers can compute. We propose to use instead new secure codes, called Staircase codes, introduced previously by two of the authors. We study the delay induced by Staircase codes which is always less than that of secret sharing. The reason is that secret sharing schemes need to wait for the responses of a fixed fraction of the workers, whereas Staircase codes offer more flexibility in this respect. For instance, for codes with rate $R = 1/2$ Staircase codes can lead to up to 40% reduction in delay compared to secret sharing.

3. **Author(s):** Vishnu Boddeti (Michigan State University), Ryo Yonetani, Kris Kitani, Yoichi Sato

   Title: Privacy-preserving learning using doubly permuted homomorphic encryption

   Abstract: We propose a privacy-preserving framework for learning visual classifiers by leveraging distributed private image data. This framework is designed to aggregate multiple classifiers updated locally using private data and to ensure that no private information about the data is exposed during its learning procedure. We utilize a homomorphic cryptosystem that can aggregate the local classifiers while they are encrypted and thus kept secret. To overcome the high computational cost of homomorphic encryption of high-dimensional classifiers, we (1) impose sparsity constraints on local classifier updates and (2) propose a novel efficient encryption scheme named doubly-permuted homomorphic encryption (DPHE) which is tailored to sparse high-dimensional data. DPHE (i) decomposes sparse data into its constituent non-zero values and their corresponding support indices, (ii) applies homomorphic encryption only to the non-zero values, and (iii) employs double permutations on the support indices to make them secret. Our experimental evaluation on several public datasets demonstrates that the proposed approach significantly outperforms other privacy-preserving methods and achieves comparable performance against state-of-the-art visual recognition methods without privacy preservation.

4. **Author(s):** Jerry Chee (University of Chicago), Panos Toulis

   Title: Convergence diagnostics for stochastic gradient descent

   Abstract: Iterative procedures in stochastic optimization, such as stochastic gradient descent with constant learning rate, are generally characterized by a transient phase and a stationary phase. During the transient phase the procedure moves fast towards a region of interest, and during the stationary phase the procedure typically oscillates around a single stationary point. In this paper, we develop a statistical diagnostic to detect such phase transition, which relies on earlier results from stochastic approximation. We present theoretical and experimental results suggesting that beyond

our estimate of stationarity the iterates do not depend on the initial conditions. The diagnostic is illustrated in an application to speed up convergence of stochastic gradient descent, by halving the learning rate each time stationarity is detected. This leads to impressive speed gains that are empirically comparable to state-of-the-art.

5. **Author(s):** Sheng Chen (University of Minnesota), Arindam Banerjee

   Title: Robust structured estimation with single-index models

   Abstract: In this work, we investigate the general single-index models (SIMs) in high dimension. Specifically we propose two types of robust estimators for the recovery of structured parameter, which generalize several existing algorithms for one-bit compressed sensing (1-bit CS). With very minimal assumption on noise, the recovery guarantees can be established for the generalized estimators under suitable conditions, which allow general structures of underlying parameter. Moreover, our generalized estimator is novelly instantiated for SIMs with monotone transfer function, and the obtained estimator can better leverage the monotonicity. Experimental results are provided to support our theoretical analyses.

6. **Author(s):** Bosu Choi (Michigan State University), Mark Iwen

   Title: Fast algorithms for high dimensional sparse problems

   Abstract: Compressive sensing based sampling theory reduces the sample number to $O(s \log N)$ required to reconstruct s-sparse signals with length N while the traditional sampling theory requires $O(N)$. Despite the small sample number, many decoding algorithms have $O(N)$ runtime complexity. When it comes to high dimensional domains, N becomes exponential in domain dimension so that we face the curse of dimensionality. In our work, we assume sparsity in a high dimensional bounded orthonormal system and in order to overcome the curse, we approximate the active support of domain in a greedy way. In this way, our algorithm is exponentially faster than the traditional compressive sensing algorithms.

7. **Author(s):** Efrén Cruz Cortés (University of Michigan), Clayton Scott

   Title: Fixed bandwidth kernel density estimation

   Abstract: Consistency of the KDE requires that the kernel bandwidth tends to zero as the sample size grows. In this work, we investigate the question of whether consistency is still possible when the bandwidth is fixed, if we consider a more general class of weighted KDEs. To answer this question in the affirmative, we introduce the fixed-bandwidth KDE (fbKDE), obtained by solving a quadratic program, that consistently estimates any square-integrable density with compact support. Rates of convergence are also established for the positive definite radial kernels under appropriate smoothness conditions on the kernel and the unknown density. Furthermore, we perform experiments to demonstrate that the fbKDE compares favorably to the standard KDE and another previously proposed weighted KDE.

8. **Author(s):** Xiaowu Dai (UW-Madison)

   Title: Estimation for varying coefficient model with longitudinal data

   Abstract: Smoothing splines estimates for varying coefficient models was proposed by Hastie and Tibshirani (1993) to address repeated measurements. Although there exists efficient algorithms, e.g., the backfitting schemes, it remains unclear about the sampling properties of this estimator. We obtain sharp results on the minimax rates of convergences and show that smoothing spline estimators achieve the optimal rates of convergence for both prediction and estimation problems. Numerical results are obtained to demonstrate the theoretical developments.

9. **Author(s):** Zhongtian Dai (TTIC), Matthew Walter

   Title: Notepad-augmented environments in reinforcement learning

   Abstract: Observing the recent success of recurrent neural network-based models in reinforcement learning, we inquire about the importance of *memory* to achieving high performance in *abstract* partially observable Markov decision processes. We compare deterministic and stochastic memory architectures in terms of *training efficiency* and *interpretability*. To combine the advantages of both architectures in settings where continuous relaxation is feasible, we also explored training a stochastic memory architecture using *differentiable* Concrete distribution.

10. **Author(s):** Aniket Anand Deshmukh (University of Michigan), Urun Dogan, Clayton Scott

    Title: Multi-task learning for contextual bandits

    Abstract: Contextual bandits are a form of multi-armed bandit in which the agent has access to predictive side information (known as the context) for each arm at each time step, and have been used to model personalized news recommendation, ad placement, and other applications. In this work, we propose a multi-task learning framework for contextual bandit problems. Like multi-task learning in the batch setting, the goal is to leverage similarities in contexts for different arms so as to improve the agent's ability to predict rewards from contexts. We propose an upper confidence bound-based multi-task learning algorithm for contextual bandits, establish a corresponding regret bound,

and interpret this bound to quantify the advantages of learning in the presence of high task (arm) similarity. We also describe a scheme for estimating task similarity from data, and demonstrate our algorithm's performance on several data sets.

11. **Author(s):** Aditya Deshpande (UIUC), Jiajun Lu, Mao-Chuang Yeh, Min Jin Chong, David Forsyth

    Title: Learning diverse image colorization

    Abstract: Colorization is an ambiguous problem, with multiple viable colorizations. Our method produces multiple realistic colorizations with better diversity than CVAE and cGAN.

12. **Author(s):** Thinh Doan (UIUC), Carolyn Beck, R. Srikant

    Title: On the stability of distributed gradient-based consensus methods under communication delays

    Abstract: Motivated by applications in machine learning and statistics, we study distributed optimization problems that are defined over a network of processors. The objective function is composed of a sum of local functions where each function is known by only one processor. A standard example from statistical machine learning is the problem of minimizing an average loss function over large training data. In this problem, large data sets of the order of terabytes are distributed across a network of processors, where each processor computes the empirical loss over a local subset of data. The processors, therefore, must communicate to determine parameters that minimize the loss over the entire data set resulting in a need for distributed algorithms.

    The nature of distributed approaches necessarily requires communications among processors while these processors may differ in their computational capabilities. As a result, delays in inter-processor message exchange are inevitable. In this talk, we study the convergence rate of a popular distributed gradient-based consensus algorithm in the presence of network delays, which has been identified as a significant open problem. We consider the special case uniform, but possibly arbitrarily large, communication delays between the processors. Our main contribution is to show that convergence occurs at rate $\mathcal{O}\left(\frac{ne^{\tau}\ln(t)}{(1-\sigma_2)^2\sqrt{t}}\right)$, where n is the number of processors, σ_2 is a parameter representing the spectral properties of network connectivity of the processors, and τ is the delay between any pair of processors. We note that such an explicit formula for the convergence rate is not available for any consensus-based methods in solving distributed optimization problems with communication delays. We also provide a lower bound to show that the exponential dependence on delay is inevitable. Finally, we present simulation results to illustrate the tightness of our theoretical bounds.

13. **Author(s):** Justin Eldridge (Ohio State University), Mikhail Belkin, Yusu Wang

    Title: Graphons, mergeons, and so on!

    Abstract: In this work we develop a theory of hierarchical clustering for graphs. Our modeling 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.

14. **Author(s):** Sinong Geng (UW-Madison), Zhaobin Kuang, David Page

    Title: A screening rule for ℓ_1-regularized Ising model estimation

    Abstract: We discover a screening rule for ℓ_1-regularized Ising model estimation. The simple closed-form screening rule is a necessary and sufficient condition for exactly recovering the blockwise structure of a solution under any given regularization parameters. With enough sparsity, the screening rule can be combined with exact and inexact optimization procedures to deliver solutions efficiently in practice. The screening rule is especially suitable for large-scale exploratory data analysis, where the number of variables in the dataset can be thousands while we are only interested in the relationship among a handful of variables within moderate-size clusters for interpretability. Experimental results on various datasets demonstrate the efficiency and insights gained from the introduction of the screening rule.

15. **Author(s):** Wooseok Ha (University of Chicago), Rina Foygel Barber

    Title: Alternating minimization and alternating descent for nonconvex optimization problems

    Abstract: Many optimization problems in high-dimensional statistics and signal processing involve two decision variables to be minimized, where the variables often reflect different structures of the signals being considered. Alternating minimization is a widely used method for solving such optimization problems, but the general properties of alternating minimization has not yet been understood well in some settings. In this work, we study and analyze the performance of alternating minimization under the setting where the variables are constrained to nonconvex sets under standard assumptions on the loss function such as restricted strong convexity. Since in practice performing the exact alternating

minimization might be intractable, we also approximate it with projected gradient descent steps and show that alternating descent approximates alternating minimization quickly, therefore obtaining fast convergence guarantees to the optimal. Our analysis depends strongly on the notion of local concavity coefficients, which have been recently proposed to measure and quantify the nonconvexity of a general nonconvex constraint set. We demonstrate our conditions on two important classes of the problems, low rank + sparse decomposition and multivariate regression problem, and provide some simulation results to see the empirical performance of the algorithms.

16. **Author(s):** Mark Harmon (Northwestern University), Diego Klabjan

    Title: Activation ensembles for deep neural networks

    Abstract: We propose a new methodology of designing activation functions within a neural network at each layer. We call this technique an "activation ensemble" because it allows the use of multiple activation functions at each layer. This is done by introducing additional variables, α, at each activation layer of a network to allow for multiple activation functions to be active at each neuron. By design, activations with larger α values at a neuron is equivalent to having the largest magnitude. Hence, those higher magnitude activations are "chosen" by the network. We implement the activation ensembles on a variety of datasets using an array of Feed Forward and Convolutional Neural Networks. By using the activation ensemble, we achieve superior results compared to traditional techniques. In addition, because of the flexibility of this methodology, we more deeply explore activation functions and the features that they capture.

17. **Author(s):** Bin Hu (UW-Madison), Laurent Lessard, Stephen Wright, Peter Seiler, Anders Rantzer

    Title: A unified analysis of stochastic optimization methods using dissipation inequalities and quadratic constraints

    Abstract: Empirical risk minimization (ERM) is a central topic for machine learning research, and is typically solved using stochastic optimization techniques. This poster will present a sequence of recent work which unifies the analysis of such stochastic optimization methods using dissipativity theory and quadratic constraints. Specifically, we will apply a unified dissipativity approach to derive sufficient conditions for linear rate certifications of SGD, SAG, SAGA, Finito, SDCA, SVRG, and SGD with momentum. The derived conditions are all in the forms of linear matrix inequalities (LMIs). We solve these resultant LMIs and obtain analytical proofs of new convergence rates for various stochastic methods (with or without individual convexity). Our proposed analysis can be automated for a large class of stochastic methods under various assumptions. In addition, the derived LMIs can always be numerically solved to provide clues for constructions of analytical proofs.

18. **Author(s):** Vahan Huroyan (University of Minnesota), Gilad Lerman

    Title: Distributed robust subspace recovery

    Abstract: We study Robust Subspace Recovery (RSR) in distributed settings. We consider a huge data set in an ad hoc network without a central processor, where each node has access only to one chunk of the data set. We assume that part of the whole data set lies around a low-dimensional subspace and the other part is composed of outliers that lie away from that subspace. The goal is to recover the underlying subspace for the whole data set, without transferring the data itself between the nodes.

19. **Author(s):** Vamsi Ithapu (UW-Madison), Vikas Singh, Risi Kondor

    Title: Decoding deep networks

    Abstract: The necessity of depth in efficient neural network learning has led to a family of designs referred to as *very deep* networks(e.g.,GoogLeNethas22layers).As depth further increases, the need for appropriate tools to explore the space of hidden layers becomes paramount. Beyond answering interesting theoretical questions as to what exactly the hidden representations represent in terms of information gain, "decoding" such inter and intra class relationships among hidden representations are critical for model selection (e.g., do I add another layer?). Using a recently proposed technique that models hierarchical (parsimonious) structure in matrices, in this work, we try toinfer semantic relationships among deep representations, even when they are not explicitly trained to do so. We show that this modeling technique is anexploratory tool providing human-interpretable feedback as one modulates the network architecture, thereby aiding in choosing the appropriate architecture for a given dataset.

20. **Author(s):** Gauri Jagatap (Iowa State University), Chinmay Hegde

    Title: Fast and sample-efficient algorithms for structured phase retrieval

    Abstract: We consider the problem of recovering a signal, from magnitude-only measurements. This is a stylized version of the classical phase retrieval problem, and is a fundamental challenge in nano- and bio-imaging systems, astronomical imaging, and speech processing. It is well known that the above problem is ill-posed, and therefore some additional assumptions on the signal and/or the measurements are necessary.

    In this paper, we first study the case where the underlying signal x is s-sparse. For this case, we develop a novel recovery algorithm that we call Compressive Phase Retrieval with Alternating Minimization, or CoPRAM. Our algorithm is

simple and can be obtained via a natural combination of the classical alternating minimization approach for phase retrieval with the CoSaMP algorithm for sparse recovery. Despite its simplicity, we prove that our algorithm achieves a sample complexity of $O(s^2 \log n)$ with Gaussian measurements, which matches the best known existing results; moreover, it also demonstrates linear convergence in theory and practice. An appealing feature of our algorithm is that it requires no extra tuning parameters other than the signal sparsity level s.

We then consider the case where the underlying signal arises from "structured" sparsity models. We specifically examine the case of block-sparse signals with uniform block size of b. For this problem, we design a recovery algorithm that we call Block CoPRAM that further reduces the sample complexity to $O((s^2/b) \log n)$. For sufficiently large block lengths of $b \sim s$, this bound equates to $O(s \log n)$. To our knowledge, this constitutes the first end-to-end algorithm for phase retrieval where the Gaussian sample complexity has a sub-quadratic dependence on the sparsity level of the signal.

21. **Author(s):** Michelle Jarboe (AMIGA Tech Initiative)

    Title: Way to go, Alpha Go: The game-changing AI program revolutionizing machine learning

    Abstract: Recent highlights of Google DeepMind's genius AI program, AlphaGo, and how it brings machine learning into play in the gaming world of competitive Go, one of the world's most complex strategy games. Citing research done by DeepMind, quotes, and recent articles. Topics: DeepMind; machine learning (ML); artificial intelligence (AI); deep neural network (DNN); neurons; value networks; policy networks; supervised learning (SL); reinforcement learning (RL); Monte Carlo tree search; convolutional neural network (CNN; ConvNet); multi-layered artificial neural network (ANN); algorithm; deep learning (DL); simulation; self-play games; Google; Demis Hassabis; Tribeca Film Festival movie; Lee Sedol; Google DeepMind Challenge Match; China; Chinese Go Association; Ke Jie; The Future of Go Summit; AlphaGo's retirement from Go; the future of AlphaGo; insights; infographics.

22. **Author(s):** Kwang-Sung Jun (UW-Madison), Francesco Orabona, Rebecca Willett, Stephen Wright

    Title: Improved strongly adaptive online learning using coin betting

    Abstract: This paper describes a new parameter-free online learning algorithm for changing environments. In comparing against algorithms with the same time complexity as ours, we obtain a strongly adaptive regret bound that is a factor of at least $\sqrt{\log(T)}$ better, where T is the time horizon. Empirical results show that our algorithm outperforms state-of-the-art methods in learning with expert advice and metric learning scenarios.

23. **Author(s):** Justin Khim (University of Pennsylvania), Po-Ling Loh

    Title: Permutation tests for infection graphs

    Abstract: We formulate and analyze a hypothesis testing problem for inferring the edge structure of an infection graph. Our model is as follows: A disease spreads over a network via contagion and random infection, where uninfected nodes contract the disease at a time corresponding to an independent exponential random variable and infected nodes transmit the disease to uninfected neighbors according to independent exponential random variables with an unknown rate parameter. A subset of nodes is also censored, meaning the infection statuses of the nodes are unobserved. Given the statuses of all nodes in the network, the goal is to determine the underlying graph. Our procedure consists of a permutation test, and we derive a condition in terms of automorphism groups of the graphs corresponding to the null and alternative hypotheses that ensures the validity of our test. Notably, the permutation test does not involve estimating unknown parameters governing the infection process; instead, it leverages differences in the topologies of the null and alternative graphs. We derive risk bounds for our testing procedure in settings of interest; provide extensions to situations involving relaxed versions of the algebraic condition; and discuss multiple observations of infection spreads. We conclude with experiments validating our results.

24. **Author(s):** Mladen Kolar (University of Chicago), Sanmi Koyejo, Suriya Gunasekar

    Title: Testing for group differences in high-dimensional graphical models with latent variables

    Abstract: We propose a novel procedure for learning the difference between two graphical models with latent variables. Linear convergence rate is established for an alternating gradient descent procedure with correct initialization. A debiasing procedure if proposed to further establish asymptotically normal estimator of the difference. Simulation studies illustrate performance of the procedure. We also illustrate the procedure on an application in neuroscience.

25. **Author(s):** Jaehoon Koo (Northwestern University), Diego Klabjan

    Title: Improved classification based on Deep Belief Networks

    Abstract: For better classification generative models are used to initialize the model and model features before training a classifier. Typically it is needed to solve separate unsupervised and supervised learning problems. Generative restricted Boltzmann machines and deep belief networks are widely used for unsupervised learning. We developed several supervised models based on DBN in order to improve this two phase strategy. Modifying the loss function to account for expectation with respect to the underlying generative model, introducing weight bounds, and multi-level

programming are applied in model development. The proposed models capture both unsupervised and supervised objectives effectively. The computational study verifies that our models perform better than the two-phase training approach.

26. **Author(s):** Drew Lazar (Ball State University), Lizhen Lin

    Title: Scale and curvature effects in principal geodesic analysis

    Abstract: There is growing interest in using the close connection between differential geometry and statistics to model smooth manifold-valued data. In particular, much work has been done recently to generalize principal component analysis (PCA), the method of dimension reduction in linear spaces, to Riemannian manifolds. One such generalization is known as principal geodesic analysis (PGA). In a novel fashion we will obtain Taylor expansions in scaling parameters introduced in the domain of objective functions in PGA. It is shown this technique not only leads to better closed-form approximations of PGA but also reveals the effects that scale, curvature and the distribution of data have on solutions to PGA and on their differences to first-order tangent space approximations. This approach should be able to be applied not only to PGA but also to other generalizations of PCA and more generally to other intrinsic statistics on Riemannian manifolds.

27. **Author(s):** Zhi Li (Michigan State University), Wei Shi, Ming Yan

    Title: A decentralized proximal-gradient method with network independent step-sizes and separated convergence rates

    Abstract: This poster considers the decentralized optimization problem with the objective function containing smooth terms and non-smooth terms. To find the optimal solution, a proximal-gradient scheme is studied. The proposed method has some advantages: First, agents can use uncoordinated step-sizes. The upper bounds on step-sizes can be as large as the one in the gradient descent, since the bounds are independent from the network topology, and only determined by local objective functions. Second, without non-smooth terms, linear convergence can be achieved under the strong convexity assumption. The dependence of the convergence rate on the objective functions and the network are separated, and the convergence rate of our new scheme is as good as one of the two convergence rates that match the typical rates for the general gradient descent and the consensus averaging. We also provide some numerical experiments to demonstrate the efficacy of the introduced algorithms and validate our theoretical results.

28. **Author(s):** Cong Han Lim (UW-Madison), Stephen Wright

    Title: k-group sparsity and group ordered-weighted sparsity for overlapping groups

    Abstract: The k-support and OWL norms generalize the ℓ_1 norm, providing better prediction accuracy and handling of correlated variables. We study the norms obtained from extending the k-support norm or OWL norm to the setting where we have overlapping groups. The resulting norms are in general NP-hard to compute, but they are tractable for certain collections of groups. To show this, we first develop a dynamic program for the problem of projecting onto the set of vectors supported by a fixed number of group. Our dynamic program utilizes tree decompositions and the complexity scales exponentially with the treewidth. This program can be converted to a convex formulation that, for the associated group structure, models the k-group support norms and an overlapping group variant of the ordered weighted ℓ_1 norm. Numerical results for these new norms are presented.

29. **Author(s):** Chaoyue Liu (Ohio State University), Mikhail Belkin

    Title: Clustering with Bregman divergences: An asymptotic analysis

    Abstract: 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.

30. **Author(s):** Siyuan Ma (Ohio State University), Mikhail Belkin

    Title: Diving into the shallows: A computational perspective on large-scale shallow learning

    Abstract: Remarkable recent success of deep neural networks has not been easy to analyze theoretically. It has been particularly hard to disentangle relative significance of architecture and optimization in achieving accurate classification on large datasets. On the flip side, shallow methods (such as kernel methods) have encountered obstacles in scaling to large data, despite excellent performance on smaller datasets, and extensive theoretical analysis. Practical methods, such as variants of gradient descent used so successfully in deep learning, seem to perform below par when applied to kernel methods. This difficulty has sometimes been attributed to the limitations of shallow architecture.

    In this paper we identify a basic limitation in gradient descent-based optimization methods when used in conjunctions with smooth kernels. An analysis demonstrates that only a vanishingly small fraction of the function space is reachable after a fixed number of gradient descent iterations drastically limiting its power and resulting in severe over-regularization. The issue is purely algorithmic, persisting even in the limit of infinite data.

To address this shortcoming in practice, we introduce EigenPro iteration, based on a simple preconditioning scheme using a small number of approximately computed eigenvectors. It turns out that even this small (and computationally inexpensive) amount of approximate second-order information results in significant improvement of performance for large-scale kernel methods. Using EigenPro in conjunction with stochastic gradient descent we demonstrate scalable state-of-the-art results for kernel methods on a modest computational budget of a few GPU-hours (compared to typically much larger computational expenditures to obtain best results in the literature).

Finally, we feel that these results show a need for a broader computational perspective on modern large-scale learning to complement more traditional statistical and convergence analyses. In particular, systematic analysis concentrating on the approximation power of algorithms with a fixed budget of computation will lead to progress both in theory and practice.

31. **Author(s):** Yintai Ma (Northwestern University), Diego Klabjan

    Title: Diminishing batch normalization network

    Abstract: Batch Normalization (BN) is very effective in accelerating the convergence of the neural network in training phase that it has become a common practice. We propose a new flavor of Diminishing Batch Normalization (DBN) algorithm and show for the first time the convergence analysis with nonconvex and convex cost function, respectively. For nonconvex case, DBN converges to a stationary point which is optimal with respect to trainable parameters. For convex case, DBN converges to the optimal value sublinearly. In the numerical experiment, we observe that the DBN performs as well as the original BN algorithm in terms of accelerating.

32. **Author(s):** Blake Mason (UW-Madison), Lalit Jain, Robert Nowak

    Title: Learning low-dimensional metrics

    Abstract: This poster investigates the theoretical foundations of metric learning, focused on three key questions that are not fully addressed in prior work: 1) we consider learning general low-dimensional (low-rank) metrics as well as sparse metrics; 2) we develop upper and lower (minimax) bounds on the generalization error; 3) we quantify the sample complexity of metric learning in terms of the dimension of the feature space and the dimension/rank of the underlying metric; 4) we also bound the accuracy of the learned metric relative to the underlying true generative metric. All the results involve novel mathematical approaches to the metric learning problem, and also shed new light on the special case of ordinal embedding (aka non-metric multidimensional scaling).

33. **Author(s):** Sami Merhi (Michigan State University), Aditya Viswanathan, Mark Iwen

    Title: Recovery of compactly supported functions from spectrogram measurements via lifting

    Abstract: The problem of signal recovery (up to a global phase) from phaseless STFT measurements appears in many audio, engineering and imaging applications. Our principal motivation here, however, is ptychographic imaging in the 1-D setting. We present a novel phase retrieval method, that is partially inspired by the well known PhaseLift algorithm. The method is based on a lifted formulation of the infinite dimensional problem which is then later truncated for the sake of computation. We demonstrate the promise of the proposed approach through numerical experiments.

34. **Author(s):** Thanh Nguyen (Iowa State University), Chinmay Hegde, Raymond Wong

    Title: A neurally plausible algorithm for doubly-sparse dictionary learning

    Abstract: Dictionary learning (a.k.a. sparse coding) is a crucial subroutine in several signal processing and machine learning algorithms. Here the goal is to learn an overcomplete dictionary (or code) that can sparsely represent a given input dataset. However, storage, transmission, and processing of the learned code can be untenably high. To resolve this, we consider a doubly-sparse dictionary model where the dictionary is the product of a fixed sparse orthonormal basis and a data-adaptive sparse component. First, we introduce a simple, neurally plausible algorithm for doubly-sparse dictionary learning. Next, we asymptotically analyze its performance and demonstrate sample complexity benefits over existing, rigorous dictionary learning approaches.

35. **Author(s):** Greg Ongie (University of Michigan), Rebecca Willett, Robert Nowak, Laura Balzano

    Title: Algebraic variety models for high-rank matrix completion

    Abstract: We consider a generalization of low-rank matrix completion to the case where the data belongs to an algebraic variety, i.e., each data point is a solution to a system of polynomial equations. In this case the original matrix is possibly high-rank, but it becomes low-rank after mapping each column to a higher dimensional space of monomial features. Many well-studied extensions of linear models, including affine subspaces and their union, can be described by a variety model. In addition, varieties can be used to model a richer class of nonlinear quadratic and higher degree curves and surfaces. We study the sampling requirements for matrix completion under a variety model with a focus on a union of affine subspaces. We also propose an efficient matrix completion algorithm that minimizes a convex or non-convex surrogate of the rank of the matrix of monomial features. Our algorithm uses the well- known kernel trick

to avoid working directly with the high-dimensional monomial matrix. We show the proposed algorithm is able to recover synthetically generated data up to the predicted sampling complexity bounds. The proposed algorithm also outperforms standard low rank matrix completion and subspace clustering techniques in experiments with real data.

36. **Author(s):** Shane Settle (TTIC), Keith Levin, Herman Kamper, Karen Livescu

    Title: Query-by-example search with discriminative neural acoustic word embeddings

    Abstract: Query-by-example search often uses dynamic time warping (DTW) for comparing queries and proposed matching segments. Recent work has shown that comparing speech segments by representing them as fixed-dimensional vectors—acoustic word embeddings—and measuring their vector distance (e.g., cosine distance) can discriminate between words more accurately than DTW-based approaches. We consider an approach to query-by-example search that embeds both the query and database segments according to a neural model, followed by nearest-neighbor search to find the matching segments. Earlier work on embedding-based query-by-example, using template-based acoustic word embeddings, achieved competitive performance. We find that our embeddings, based on recurrent neural networks trained to optimize word discrimination, achieve substantial improvements in performance and run-time efficiency over the previous approaches.

37. **Author(s):** Mohammadreza Soltani (Iowa State University), Chinmay Hegde

    Title: Improved algorithms for matrix recovery from rank-one projections

    Abstract: We consider the problem of estimation of a low-rank matrix from a limited number of noisy rank-one projections. In particular, we propose two fast, non-convex proper algorithms for matrix recovery and support them with rigorous theoretical analysis. We show that the proposed algorithms enjoy linear convergence and that their sample complexity is independent of the condition number of the unknown true low-rank matrix. By leveraging recent advances in low-rank matrix approximation techniques, we show that our algorithms achieve computational speed-ups over existing methods. Finally, we complement our theory with some numerical experiments.

38. **Author(s):** Alex Stec (Northwestern University), Diego Klabjan, Jean Utke

    Title: Nest multi-instance classification

    Abstract: There are classification tasks that take as inputs groups of images rather than single images. In order to address such situations, we introduce a nested multi-instance deep network. The approach is generic in that it is applicable to general data instances, not just images. The network has several convolutional neural networks grouped together at different stages. This primarily differs from other previous works in that we organize instances into relevant groups that are treated differently. We also introduce methods to replace instances that are missing and a manual dropout when a whole group of instances is missing. With specific pretraining, we find that the model works to great effect on our data. For our two cases with four and seven classes, we obtain 75.9% and 68.7% accuracy, respectively.

39. **Author(s):** Amirhossein Taghvaei (UIUC), Jin Kim, Prashant Mehta

    Title: Critical points of linear networks

    Abstract: The subject of this poster pertains to the problem of representing and learning a linear transformation using a linear neural network. In recent years, there has been a growing interest in the study of such networks in part due to the success of deep learning. The main question in this research and also in our poster concerns existence and optimality properties of the critical points for the mean-squared loss function. An optimal control model (a certain regularized form of the loss function) is introduced and a learning algorithm (a variant of backprop) derived for the same using the Hamiltonian formulation of optimal control. The formulation is used to provide a complete characterization of the critical points in terms of the solutions of a nonlinear matrix-valued equation, referred to as the characteristic equation. Analytical and numerical tools from bifurcation theory are used to compute the solutions of the characteristic equation. The resulting critical points reveal some surprising conclusions on representations as well as the optimality properties of the same. Relationship to the backprop algorithm is described by considering a limit where the regularization parameter goes to zero.

40. **Author(s):** Andrea Trevino-Gavito (Northwestern University), Diego Klabjan

    Title: Deep learning for video data

    Abstract: We present an unsupervised deep learning architecture for clustering of videos. The focus is on videos where a small patch differentiates one instance from the other. The architecture combines transfer learning, crops from frames, and recurrent convolutional neural networks to extract embeddings by means of a sequence to sequence auto encoder and uses clustering over the learned representations. We apply this framework to ultrasound videos.

41. **Author(s):** Minzhe Wang (University of Chicago), Zheng Tracy Ke

    Title: A new SVD approach to optimal topic estimation

Abstract: We proposed a new SVD-based algorithm for topic recovery in the topic model, and developed it's high probability l1-error bound under probabilistic Latent Semantic Indexing model, which achieves the minimax lower bound. Extensive simulation comparisons show the superiority of our method over the other popular ones. Finally encouraging results have been obtained when we implemented our method on two real datasets. All these findings validate our discovery.

42. **Author(s):** Song Wang (UW-Madison), Yini Zhang, Chris Wells, Karl Rohe

    Title: Trump followers on Twitter before the 2016 election

    Abstract: With Twitter playing an indispensable role in Donald Trump's ascent to power, this study seeks to understand his Twitter followers and their engagement with him. In the study, a random sample of 330k Trump followers was selected with probability proportional to their follower counts and each followers friend list and 3200 most recent tweets were also harvested. Based on following relationships (who-follows-whom), we constructed a bipartite network and found 10 groups of Trump's follower clusters, including Trump supporters, conservatives, liberals, politically disengaged, etc. These clusters were validated through text analysis of tweets from Trump followers' timelines and their engagement with Trump. Then we topic-modeled of Trump's 8000 tweets during the campaign retweeted/replied by our sampled followers and found 20 themes such as criticizing Hilary Clinton, Obama, GOP opponents, media and announcing campaign information. We saw 1) different groups of followers interacted with Trump on Twitter at different time points; 2) different groups of followers reacted to Trumps tweets differently the politically engaged Trump followers, liberal or conservative, commonly engaged with Trump's tweets about Clinton, Obama, and media, whereas the politically disengaged tended to respond more to Trump's relatively standardized campaign tweets. We further explored the sentiments of the retweets/replies. Our approach can be replicated to explore other politicians' followership on Twitter and their communication strategies through it.

43. **Author(s):** Yutong Wang (University of Michigan), Laura Balzano, Clayton Scott

    Title: Joint analysis of bulk and single-cell RNA sequencing data via matrix factorization

    Abstract: RNA sequencing (RNA-Seq) technologies are developed by biologists to measure gene expression profiles of cells. Gene expression profiles are represented as nonnegative integer-valued vectors in G-dimensional space, where G is the number of genes being studied. The entries of the vectors represent the abundance of the genes present in a sample, which could be an entire bulk tissue or a single cell.

    Two scientific problems of interest are to (1) cluster single-cell measurements from a biological tissue into distinct types of cells and (2) obtain the true gene expression profile for each type of cells. Bulk RNA-Seq measures a mixture of the expression profiles for all cell types in the tissue with low noise. Mathematically, this corresponds to observing a convex combination of expression profiles of different types of cells. On the other hand, single-cell RNA-Seq measures the expression profile of a single cell with high noise. Mathematically, this corresponds to observing a noise corrupted expression profile of one particular cell type.

    Problem (2) can be approached by computing a nonnegative matrix factorization (NMF) of the bulk RNA-seq data matrix. However, NMF is only known to perform provably well under the so called separability condition, which is unlikely to be true for the gene expression data. We address the problem by using the single-cell data to guide the NMF of the bulk data matrix to choose the correct factorization when separability fails. To achieve this, we present an alternating algorithm for minimizing a regularized NMF objective. As a side effect, our methods also solve problem (1). We will show some empirical results on synthetic data which suggests the potential applicability of our methods.

44. **Author(s):** Papis Wongchaisuwat (Northwestern University), Diego Klabjan

    Title: Validating the truth of biomedical sentences

    Abstract: Abundant information obtained from diverse sources is not always reliable, e.g., inaccurate health information can lead to critical consequences. Existing approaches to validate the truthfulness of sentences mostly depend on additional information about sources which might not be readily available. We develop an algorithm to identify the trustworthiness of sentences and providing supporting evidence based on biomedical literature. Specifically, our proposed algorithm relies on the predications extracted from MEDLINE citations to indicate whether the sentence is semantically true or false. Supporting evidence is extracted based on a new algorithm utilizing knowledge from various biomedical ontologies.

45. **Author(s):** Min Xu (University of Pennsylvania), Varun Jog, Po-Ling Loh

    Title: Optimal rates for community estimation in the weighted stochastic block model

    Abstract: Community identification in a network is an important problem in fields such as social science, neuroscience, and genetics. Over the past decade, stochastic block models (SBMs) have emerged as a popular statistical framework for this problem. However, SBMs have an important limitation in that they are suited only for networks with unweighted edges; in various scientific applications, disregarding the edge weights may result in a loss of valuable information. We

study a weighted generalization of the SBM, in which observations are collected in the form of a weighted adjacency matrix and the weight of each edge is generated independently from an unknown probability density determined by the community membership of its endpoints. We characterize the optimal rate of misclustering error of the weighted SBM in terms of the Renyi divergence of order 1/2 between the weight distributions of within-community and between-community edges, substantially generalizing existing results for unweighted SBMs. Furthermore, we present a principled, computationally tractable algorithm based on discretization that achieves the optimal error rate without assuming knowledge of the weight densities.

46. **Author(s):** Ming Yu (University of Chicago), Varun Gupta, Mladen Kolar

    Title: An influence-receptivity model for topic based information cascades

    Abstract: We consider the problem of estimating the latent structure of a social network based on observational data on information diffusion processes, or cascades. Here for a given cascade, we only observe the time a node/agent is infected but not the source of infection. Existing literature has focused on estimating network diffusion matrix without any underlying assumptions on the structure of the network. We propose a novel model for inferring network diffusion matrix based on the intuition that an information datum is more likely to propagate among two nodes if they are interested in similar topics, which are common with the information. In particular, our model endows each node with an influence vector (how authoritative they are on each topic) and a receptivity vector (how susceptible they are on each topic). We show how this node-topic structure can be estimated from observed cascades. The estimated model can be used to build recommendation system based on the receptivity vectors, as well as for marketing based on the influence vectors.

47. **Author(s):** Yilin Zhang (UW-Madison), Marie Poux-Berth, Chris Wells, Karolina Koc-Michalska, Karl Rohe

    Title: Discovering political topics in Facebook discussion threads with spectral contextualization

    Abstract: To study the political engagement on Facebook during the 2012 French presidential election, we examine the Facebook posts of the leading eight candidates and the comments beneath these posts. We find evidence of both (i) candidate-centered structure, where citizens primarily comment on the wall of one candidate and (ii) issue-centered structure, where citizens' attention and expression is primarily directed towards a specific set of issues (e.g. economics, immigration, etc). To discover issue-centered structure, we develop Spectral Contextualization, a novel approach of simultaneously analyzing graph data (i.e. for each citizen, the list of posts that they comment on) with text data (i.e. what the posts and comments say). This technique scales to high-dimensional text (thousands of unique words), hundreds of thousands of citizens, and thousands of unique words. Using only the network, without any contextualizing information, spectral clustering finds a mixture of candidate and issue clusters. The contextualizing information helps to separate these two structures. We conclude by showing that the novel methodology is consistent under a statistical model.

48. **Author(s):** Yuting Zhang (University of Michigan), Luyao Yuan, Yijie Guo, Zhiyuan He, I-An Huang, Honglak Lee

    Title: Discriminative bimodal networks for visual localization and detection with natural language queries

    Abstract: Associating image regions with text queries has been recently explored as a new way to bridge visual and linguistic representations. A few pioneering approaches have been proposed based on recurrent neural language models trained generatively (e.g., generating captions), but achieving somewhat limited localization accuracy. To better address natural-language-based visual entity localization, we propose a discriminative approach. We formulate a discriminative bimodal neural network (DBNet), which can be trained by a classifier with extensive use of negative samples. Our training objective encourages better localization on single images, incorporates text phrases in a broad range, and properly pairs image regions with text phrases into positive and negative examples. Experiments on the Visual Genome dataset demonstrate the proposed DBNet significantly outperforms previous state-of-the-art methods both for localization on single images and for detection on multiple images. We we also establish an evaluation protocol for natural-language visual detection.

49. **Author(s):** Hao Zhou (UW-Madison), Sathya N. Ravi, Vamsi K. Ithapu, Sterling C. Johnson, Grace Wahba, Vikas Singh

    Title: When and how can multi-site datasets be pooled? Consistency, hypothesis tests and Alzheimer's disease

    Abstract: Many studies in biomedical and health sciences involve small sample sizes due to logistic or financial constraints. Often, identifying weak (but scientifically interesting) associations between a set of features and a response necessitates pooling datasets from multiple diverse labs or groups. However, the pooling meets trouble since the observed features might be distorted, causing heterogeneity across multiple datasets. We address this problem using an algorithm based on maximum mean discrepancy and show that the estimators are consistent. We further provide a hypothesis test to check whether the distortion is the single reason for heterogeneity, and another hypothesis test to check whether pooling improves regression performance after correction. With a focus on Alzheimer?s disease studies, we show empirical results in regimes suggested by our analysis, where pooling a locally acquired dataset with data from an international study improves power.

50. **Author(s):** Xiaofeng Zhu (Northwestern University), Diego Klabjan, Patrick Bless

Title: Semantic document distance measures and unsupervised document revision detection

Abstract: We model the document revision detection problem as a minimum cost branching problem that relies on computing document distances. Furthermore, we propose two new document distance measures, word vector-based Dynamic Time Warping (wDTW) and word vector-based Tree Edit Distance (wTED). Our revision detection system is designed for a large scale corpus and implemented in Apache Spark. We demonstrate that our system can more precisely detect revisions than state-of-the-art methods by utilizing the Wikipedia revision dumps and simulated data sets.