AFRL-RQ-WP-TR-2019-0182

FOUNDATIONS OF SCALABLE NONCONVEX OPTIMIZATION

Ali Jadbabaie, Suvrit Sra, Stefanie Jegelka, and Alexander Rakhlin

Massachusetts Institute of Technology

OCTOBER 2019
Final Report

AIR FORCE RESEARCH LABORATORY
AEROSPACE SYSTEMS DIRECTORATE
WRIGHT-PATTERSON AIR FORCE BASE, OH 45433-7542
AIR FORCE MATERIEL COMMAND
UNITED STATES AIR FORCE

NOTICE AND SIGNATURE PAGE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

This report is the result of contracted fundamental research deemed exempt from public affairs security and policy review in accordance with SAF/AQR memorandum dated 10 Dec 08 and AFRL/CA policy clarification memorandum dated 16 Jan 09. This report is available to the general public, including foreign nationals.

Copies may be obtained from the Defense Technical Information Center (DTIC) (https://discover.dtic.mil/).

AFRL-RQ-WP-TR-2019-0182 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

*//Signature on File//

DEAN E. BRYSON
Work Unit Manager
Design and Analysis Branch

//Signature on File//

CHARLES TYLER
Chief, Design and Analysis Branch
Aerospace Vehicles Division

//Signature on File//

PHILIP S. BERAN, PhD
Technical Advisor, Design and Analysis Branch
Aerospace Vehicles Division

This report is published in the interest of scientific and technical information exchange and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

*Disseminated copies will show "//Signature//" stamped or typed above the signature blocks.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
October 2019	Final	18 April 2018 – 11 October 2019

4. TITLE AND SUBTITLE
FOUNDATIONS OF SCALABLE NONCONVEX OPTIMIZATION

5a. CONTRACT NUMBER
FA8650-18-2-7838

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
61101E

6. AUTHOR(S)
Ali Jadbabaie, Suvrit Sra, Stefanie Jegelka, and Alexander Rakhlin

5d. PROJECT NUMBER
N/A

5e. TASK NUMBER

5f. WORK UNIT NUMBER
Q1YR

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Massachusetts Institute of Technology
77 Massachusetts Avenue
Cambridge, MA 02139

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Air Force Research Laboratory
Aerospace Systems Directorate
Wright-Patterson Air Force Base, OH 45433-7542
Air Force Materiel Command
United States Air Force

Defense Advanced Research Projects Agency/Tactical Technology Office (DARPA/TTO)
3701 N. Fairfax Drive
Arlington, VA 22203

10. SPONSORING/MONITORING AGENCY ACRONYM(S)
AFRL/RQVC

11. SPONSORING/MONITORING AGENCY REPORT NUMBER(S)
AFRL-RQ-WP-TR-2019-0182

12. DISTRIBUTION/AVAILABILITY STATEMENT
DISTRIBUTION STATEMENT A. Approved for public release. Distribution is unlimited.

13. SUPPLEMENTARY NOTES
This report is the result of contracted fundamental research, which is deemed exempt from Public Affairs Office security and policy review in accordance with Deputy Assistant Secretary of the Air Force (Science, Technology, Engineering) (SAF/AQR) memorandum dated 10 Dec 08 and Air Force Research Laboratory Executive Director (AFRL/CA) policy clarification memorandum dated 16 Jan 09.

This report contains .pdf attachments related to the research effort. Click the paperclip icon in the left vertical toolbar, and double click the referenced attachment. The file names correspond to the primary reference in each numbered section within the report.

14. ABSTRACT
This research program focused on creating a new paradigm for scalable nonconvex optimization. Four principal investigators with backgrounds in optimization, machine learning, and statistics were involved. The project led to development of new theories for understanding the acceleration phenomena in convex and nonconvex optimization in Euclidean and non-Euclidean spaces, with applications of training deep neural networks. The project also led to new theories about limitations and expressivity of neural networks, and to the first complete results for characterization of the optimization landscape of deep linear neural networks, leading to new results that supported the concept that local minima are global. Adding minimal nonlinearities changed the picture, with local optima whose performance are worse than linear classifiers. A first set of complete results on Bayesian optimization was developed, corresponding to settings in which even function evaluations are expensive, as well as gradients and higher order derivatives. Furthermore, the powers of graph neural networks, which have become a popular new framework for modeling large scale data with graph structures, was rigorously characterized. Finally, over fitting and generalization was analyzed, showing that the standard view in machine learning/statistics that interpolation leads to overfitting is not quite accurate in high dimensions, providing an explanation for unreasonable effectiveness of over-parameterized neural networks.

15. SUBJECT TERMS
nonconvex optimization, machine learning, scalable algorithms, foundational mathematics

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT:	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON (Monitor)
a. REPORT	b. ABSTRACT	c. THIS PAGE			Dean E. Bryson
Unclassified	Unclassified	Unclassified	SAR	23	19b. TELEPHONE NUMBER (Include Area Code) (312) 713-7137

TABLE OF CONTENTS

1. Executive Summary 1
2. Technical Progress 5
3. Comprehensive Project Report 6
3.1 Understanding acceleration in large scale, first-order optimization 6
3.2 Geometry of Acceleration in Non-Euclidean environments 7
3.3 Understanding Optimization landscape of Empirical Risk Minimization in deep neural networks 7
3.4 New Approaches to Scalable Bayesian Optimization 8
3.5 Efficient escape of saddle points and reaching stationary points in non-convex optimization 9
3.6 Understanding trade-offs between Over-fitting, interpolation, and generalization in optimization for statistical learning. 10
3.7 Finite sample expressive power of small-width ReLU networks 10
3.8 Efficient nonconvex empirical risk minimization via adaptive sample size methods 11
3.9 Achieving acceleration in distributed optimization via direct discretization of the Heavy-Ball ODE 11
3.10 On increasing self-confidence in non-Bayesian social learning over time-varying directed graphs 12
3.11 Interpolation as a learning mechanism. 12
3.12 Stable Optimization with Gaussian Processes 13
3.13 Small nonlinearities in activation functions create bad local minima in neural networks 14
3.14 Efficiently testing local optimality and escaping saddles for ReLU networks 14
3.15 R-SPIDER: A Fast Riemannian Stochastic Optimization Algorithm with Curvature Independent Rate 14
3.16 Consistency of Interpolation with Laplace Kernels is a High-Dimensional Phenomenon 15
3.17 What Can Neural Networks Reason About? 15
3.18 How Powerful are Graph Neural Networks? 15
3.19 Small ReLU networks are powerful memorizers: a tight analysis of memorization capacity. 16
3.20 Are deep ResNets provably better than linear predictors? 16
3.21 Competitive Contagion with Spare Seeding. 17
3.22 A Separation Principle for Joint Sensor and Actuator Scheduling with Guaranteed Performance Bound. 17
3.23 Non-Bayesian Social Learning with Uncertain Models over Time-Varying Directed Graphs. 18
3.24 Non-Bayesian Social Learning with Gaussian Uncertain Models 18
3.25 Non-Bayesian Social Learning with Uncertain Models 18

1. Executive Summary

A bold new research program focused on creating a new paradigm for scalable nonconvex optimization was proposed. From the invention of linear programming by George Dantzig in 1947, optimization has had a profound effect on all walks of life, but most importantly on military operations. Over the past 50 years there have been remarkable theoretical, numerical, and computational advances in all forms of optimization, yet most of the theoretical and complexity-theoretic advances have been in the field of convex optimization. Standard approaches for handling nonconvexity have been to use variants of gradient descent, or stochastic gradient algorithms (or various versions of Newton and quasi-Newton) methods to find stationary points (and not necessary minima) of nonconvex problems. In fact, verifying whether a point is a minimum itself is a computationally hard problem. As a result, nonconvex optimization has become mostly an art of choosing good initial guesses and/or heuristics based on annealing methods. Motivated by two application domains, we proposed a shift in paradigm beyond convexity, and to explore recent advances in various fields of pure and applied mathematics to exploit the geometric structure of a large class of discrete and continuous problems and go beyond heuristic approaches. Our research effort was motivated by application domains which form a cross-cutting thrust of the proposed effort on large scale statistical and machine learning. We proposed new approaches for continuous nonconvex optimization. First, we explored global optimality for nonconvex optimization problems, and developed an understanding of accelerated optimization algorithms. We also proposed how to exploit the geometric structure, by developing accelerated approaches for optimization over manifolds. Next, we investigated a variety of techniques for nonconvex optimization, and address how one can quantify fast escape from saddle points and propose to develop a rigorous complexity theory and convergence rates for nonconvex minimization. In the third task, we explored using derivative-free techniques in conjunction with the geometry of the problem to tackle nonconvexity and develop a deep understanding of the interplay of statistical learning and optimization.

We advanced the theoretical understanding of rectified linear unit (ReLU) networks from two different approaches. Initially, we provided new results about the finite sample expressive power of small-width ReLU networks. We improved state of the art literature by providing new bounds on the ability of ReLU networks to learn arbitrary data sets, with respect to the number and width of their hidden layers. Moreover, we provided a novel algorithm for testing local optimality of escaping saddle points for ReLU networks. In nonconvex functions, saddle points are a major limitation in traditional training methods, given that the satisfaction of first order conditions is not enough for the verification of local optimal. In turn saddle points might hinder the performance of such approaches. The correct identification of such saddle points and being able to move away from them are major contributions to the performance improvement of general deep learning methods.

One particularly relevant problem in non-convex optimization resulting from training deep learning models is the empirical risk minimization (ERM). We provided novel results for the minimization of such problem where the loss function is possibly non-convex. Our new adaptive sample size method can iteratively find a solution to ERM based on an iterative construction of solutions with a relatively small number of samples and avoid saddle points.

Also related to large-scale learning problems, we have studied novel interpretations and mechanism design for efficient learning using interpolation of data. We have shown that interpolation strategies can be used and still perform well in terms of out of sample prediction. With the hiring of the new postdoctoral scholar we introduced the distributed approach for the solution of optimization problems over networks. A first result related to distributed optimization over networks is the design of a new class of distributed algorithms based on discretization approaches of a differential equation. This new algorithm builds upon our results from Runge-Kutta integrators for accelerated centralized algorithms and extends these results for new algorithms that can be executed over networks. A second result is with respect of large-scale social learning algorithms, that models the learning process in social networks and provides new approaches for the design of distributed estimation and optimization algorithms. We provided the first necessary and sufficient condition that guarantees that a distributed optimization algorithm will achieve a network-wide solution to an optimization problem even if the links have decaying weights and the graph topology is directed and changes with time.

We have provided new insights into the impacts of nonlinearities in neural networks. Particularly, we have shown that the results on spurious local minima in linear neural networks are not useful for the study of the nonlinear ones. We demonstrated that even the slightest non-linearity introduces a plethora of spurious local minima, making the result about linear neural networks not robust.

Additionally, we developed a novel algorithm for testing optimality and escaping saddle points in neural networks with ReLU activation functions, which has been shown to be hard due to the presence of non-differentiable points. We exploit the geometry of the problem to provide a method that reduces the total computation to the solution of quadratic problem at each hidden node. In a best scenario this translates to a single equality constraint quadratic program, while in the worst case we show that the complexity is exponential only in the number of inequality constraints.

On the topic of escaping saddle points, we also contributed a new result for non-convex problems with constraints. We provide a framework that generates a sequence of iterates that reach an approximate second order stationary point and provide a corresponding upper bound for its iteration complexity. We characterize the overall complexity of reaching such approximate second order stationary point. Also, we provide a homologous result for the stochastic case.

Moreover, we have provided a fast optimization method for smooth stochastic problems over Riemanian manifolds. We extended the existing SPIDER algorithm and achieve a better convergence rate than other known methods. We also show a curvature-independent convergence rate for both convex and non-convex cases.

We also contributed in the study of adversarially robust optimization by developing a new algorithm for Gaussian process optimization, called StableOpt, that is robust to adversarial perturbations. We provide the sample complexity of the algorithm to reach an arbitrary optimal point and provide the corresponding lower bound.

Another front we explored is the efficient solution of non-convex ERM problems. We have provided an adaptive sample size method that iteratively finds approximate local minima to the ERM problem with a few samples, and iteratively increases the sample size in a fast converging region of the problem. This new method is computationally efficient, and we showed the required sample sizes to guarantee a desired accuracy for an approximation of a local minima, without getting attracted to a saddle point.

Finally, we have provided new theoretical insights into the relation between interpolation and generalization in learning. Initially, we have shown that interpolation, in the reproducing kernel Hilbert space, generalizes well for high dimensional datasets, but not for low-dimensional ones. This is a new result that presents a purely high dimensional phenomena in learning theory. This is particularly relevant for model machine learning applications, such as deep learning where usually the number of parameters is high, and yet some empirical evidence suggests memorization of the date yields to generalization. Moreover, we showed that for ridgeless regression can generalize, even if simple interpolation is used. We show that this is an implicit regularization phenomenon, due to the high dimension of the input data, the curvature of the kernel function and the geometric properties of the data. We provided an upper bound for out-of-sample error as well as empirical evidence of this phenomenon.

One of our main achievements are contributions toward the understanding of the generalization capabilities, memorization and performance of neural networks. Initially, we focused on a more complete characterization of the reasoning capabilities of neural networks. Particularly, we investigated on the empirical evidence that certain architectures in neural networks achieve better performance in reasoning tasks when such architecture resemble certain algorithmic structures. We build upon the recent evidence that graph neural networks (GNNs) achieve better performance than less structure networks. Our results suggest that GNNs can have potential applications to learn traditionally reasoning based algorithms like dynamic programming. Also, we shed some light into the possible design of architectures for complex reasoning. Along the same lines, we propose a theoretical formalization for the analysis of the expressive power of GNNs to capture graph structures. We show that certain networks such as graph convolutional networks cannot learn to distinguish between some carefully designed simple graphs. Moreover, we construct an architecture that is provable the most expressive in the class of GNNs. Both results are backed up by benchmark numerical analysis.

We also focused on the memorization power of ReLU networks. Particularly, we focus on showing tight bounds, i.e., a sufficient and necessary condition indicating the number of hidden nodes necessary to perfectly memorize a dataset with a relatively small ReLU network (3-layer). We also showed a generalization result for arbitrary L-layer networks. These bounds are supported by an analysis of stochastic gradient descent showing that under certain initialization conditions a memorization of the global minima is achieved, that is, a small empirical risk point is achieved. We also broadened our current understanding of the performance of residual networks (RNs) in comparison with linear predictors. The state-of-the-art understanding indicates that single residual block RNs outperforms any linear predictor. We extended this analysis to multiple residual blocks. We showed that deep RNs have critical points that are either as good as a linear predictor or have strictly negative Hessian. This, in turn indicates that the optimization landscape can improve with multiple skip-connections.

On a second research thrust, we studied a novel strategic model for information diffusion in social networks. Given that a company can seed strategic individuals in the market, such company can improve its market share. We built on a duopoly game representation between the firms. We studied the effect of the network structure on the optimal seeding strategies. We derived conditions under which Nash equilibrium leads to sparse seeding in large populations. Moreover, we explored the problem of sparse sensor and actuation in linear dynamical systems. We provided a novel understanding of the problem showing a separation principle where sensing actuating schedules can be found separately. However, these problems cannot be solved or approximated in polynomial time for time-invariant schedules. We proposed a time-varying solution that can be computed in polynomial time.

Finally, we developed a new model for distributed inference over networks, when the statistics about the hypotheses are not know precisely but need to be estimated from finite amounts of data. We showed that traditional non-Bayesian approaches can converge almost surely to the wrong hypothesis and thus, provide a novel model for which such uncertain decisions can be made, at the same time we provide a more general understanding of the effects of finite amounts of data in the constructions of the statistics of the hypotheses. We showed that such method can be implemented over arbitrary time-varying directed networks. We also extended this analysis to Gaussian observations.

2. Technical Progress

Here we provide a brief description of the technical accomplishments of the project. A detailed description of each bullet point is made below.

- Understanding acceleration in large scale, first-order optimization
- Geometry of Acceleration in Non-Euclidean environments
- Understanding Optimization landscape of Empirical Risk Minimization in deep neural networks
- New Approaches to Scalable Bayesian Optimization
- Efficient escape of saddle points and reaching stationary points in non-convex optimization
- Understanding trade-offs between Over-fitting, interpolation, and generalization in optimization for statistical learning.
- Finite sample expressive power of small-width ReLU networks
- Efficient nonconvex empirical risk minimization via adaptive sample size methods
- Achieving acceleration in distributed optimization via direct discretization of the Heavy-Ball ODE
- On increasing self-confidence in non-Bayesian social learning over time-varying directed graphs
- Interpolation as a learning mechanism.
- Stable Optimization with Gaussian Processes
- Small nonlinearities in activation functions create bad local minima in neural networks
- Efficiently testing local optimality and escaping saddles for ReLU networks
- R-SPIDER: A Fast Riemannian Stochastic Optimization Algorithm with Curvature Independent Rate
- Adversarially Robust Optimization with Gaussian Processes
- Consistency of Interpolation with Laplace Kernels is a High-Dimensional Phenomenon
- What Can Neural Networks Reason About?
- How Powerful are Graph Neural Networks?
- Small ReLU networks are powerful memorizers: a tight analysis of memorization capacity.
- Are deep Residual Networks provably better than linear predictors?
- Competitive Contagion with Sparse Seeding.
- A Separation Principle for Joint Sensor and Actuator Scheduling with Guaranteed Performance Bound.
- Non-Bayesian Social Learning with Uncertain Models over Time-Varying Directed Graphs.
- Non-Bayesian Social Learning with Gaussian Uncertain Models
- Non-Bayesian Social Learning with Uncertain Models

3. Comprehensive Project Report

A summary for our advances and new developments in each area is provided in the next sections.

3.1 Understanding acceleration in large scale, first-order optimization

Zhang, J., Mokhtari, A., Sra, S., & Jadbabaie, A. (2018). Direct Runge-Kutta discretization achieves acceleration. In Advances in Neural Information Processing Systems (pp. 3900-3909).

Our research focused on understanding the acceleration of gradient-based optimization methods which lead to provably-correct design of fast optimization algorithms. Particularly, we are able to achieve acceleration by directly discretizing a second order ordinary differential equation (ODE) related to the continuous limit of Nesterov's accelerated gradient method. While development of accelerated gradient-based optimization algorithms goes back to the work of Polyak (1967) and Nesterov (1983), many of the aspects of acceleration still remain a mystery. It is well-understood that gradient descent can be viewed as discretization of a first order ODE $\dot{x} = -\nabla f(x)$, and suffers from slow convergence rate. Momentum-based acceleration methods instead rely on a second order ODE of the form $\ddot{x} + b\dot{x} + \nabla f(x) = 0$. Recently, Boyd, Su, and Candes [2014] showed that when the stepsize in Nesterov's acceleration scheme goes to zero, one can recover the above ODE with a time-varying friction coefficient $b = 3/t$. However, up to now it remained a mystery how one can recover an accelerated gradient algorithm from the above second order ODE. In fact recent results by Wilson, Wibisono, and Jordan in a 2015 paper in Proceedings of the National Academy of Sciences suggested that the only way to achieve a stable accelerated gradient algorithm, one might have to use symplectic integrators that preserve the mechanical properties of the continuous-time dynamical system while discretizing.

In our recent work we (Jadbabaie, and Sra), together with our student Jingzhao Zhang and our postdoctoral scholar Aryan Mokhtari show that when the function is smooth enough, acceleration can be achieved by a stable discretization of this ODE using standard Runge-Kutta integrators. Specifically, we prove that under the standard assumptions of Lipschitz-gradient, convexity and order-(s+2) differentiability, the sequence of iterates generated by discretizing the proposed second-order ODE converges to the optimal solution at a rate of O(N^(−2s/s+1)), where s is the order of the Runge-Kutta numerical integrator. Furthermore, we introduce a new local flatness condition on the objective, under which rates even faster than O(N^(−2)) can be achieved with low-order integrators and only gradient information. Notably, this flatness condition is satisfied by several standard loss functions used in machine learning.

3.2 Geometry of Acceleration in Non-Euclidean environments

Zhang, H., & Sra, S. (2018, July). An estimate sequence for geodesically convex optimization. In Conference On Learning Theory (pp. 1703-1723).

Sra, S., Vishnoi, N. K., & Yildiz, O. (2018). On Geodesically Convex Formulations for the Brascamp-Lieb Constant. In Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2018). Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

We propose a Riemannian version of Nesterov's accelerated gradient algorithm (RAGD) and show that for geodesically smooth and strongly convex problems, within a neighborhood of the minimizer whose radius depends on the condition number as well as the sectional curvature of the manifold, RAGD converges to the minimizer with acceleration. Unlike past algorithms that require the exact solution to a nonlinear equation which in turn may be intractable, our algorithm is constructive and computationally tractable. Our proof exploits a new estimate sequence and a novel bound on the nonlinear metric distortion, both ideas may be of independent interest.

We have further studied two non-convex formulations for computing the optimal constant in the Brascamp-Lieb inequality, which is an inequality that can be thought of as extension of the Poincaré inequality which only concerns Gaussian probability distributions. The Brascamp–Lieb inequality is also related to the Cramér–Rao bound. While Brascamp–Lieb is an upper-bound, the Cramér–Rao bound lower-bounds the variance. We have shown that the above inequality corresponding to a given datum are geodesically log-concave on the manifold of positive definite matrices endowed with the Riemannian metric corresponding to the Hessian of the logdeterminant function. Recent work of authors like Garg and collaborators in the literature also implies a geodesically log-concave formulation of the Brascamp-Lieb constant through a reduction to the operator scaling problem. However, the dimension of the arising optimization problem in their reduction depends exponentially on the number of bits needed to describe the Brascamp-Lieb datum. The formulations presented here have dimensions that are polynomial in the bit complexity of the input datum.

3.3 Understanding Optimization landscape of Empirical Risk Minimization in deep neural networks

Yun, C., Sra, S., & Jadbabaie, A. (2018). Efficiently testing local optimality and escaping saddles for ReLU networks. arXiv preprint arXiv:1809.10858. Accepted to International Conference on Learning Representations (ICLR) 2019

In the past year, we have investigated the loss surface of deep linear and nonlinear neural networks. We show that for deep linear networks with differentiable losses, critical points after the multilinear parameterization inherit the structure of critical points of the underlying loss with linear parameterization. As corollaries we obtain results that local minima are global which subsume most previous results, while showing how to distinguish global minima from saddle points. For nonlinear neural networks, we prove two theorems showing that even for networks with one hidden layer, there can be spurious local minima. Indeed, for piecewise linear

nonnegative homogeneous activations (e.g., ReLU), we prove that for almost all practical datasets there exist infinitely many local minima that are not global. We have constructed a counterexample involving other activation functions (e.g., sigmoid, tanh, arctan, etc.), for which there exists a local minimum strictly inferior to the global minimum. This paper has been submitted for publication.

In our recent work we (Jadbabaie and Sra), together with our student Chulhee Yun, provide a theoretical algorithm for checking local optimality and escaping saddles at nondifferentiable points of empirical risks of two-layer ReLU networks. Our algorithm receives any parameter value and returns: local minimum, second-order stationary point, or a strict descent direction. The presence of M data points on the nondifferentiability of the ReLU divides the parameter space into at most 2^M regions, which makes analysis difficult. By exploiting polyhedral geometry, we reduce the total computation down to one convex quadratic program (QP) for each hidden node, O(M) (in)equality tests, and one (or a few) nonconvex QP. For the last QP, we show that our specific problem can be solved efficiently, in spite of nonconvexity. In the benign case, we solve one equality constrained QP, and we prove that projected gradient descent solves it exponentially fast. In the bad case, we have to solve a few more inequality constrained QPs, but we prove that the time complexity is exponential only in the number of inequality constraints. Our experiments show that either benign case or bad case with very few inequality constraints occurs, implying that our algorithm is efficient in most cases.

3.4 New Approaches to Scalable Bayesian Optimization

Wang, Z., Gehring, C., Kohli, P., & Jegelka, S. (2017). Batched large-scale Bayesian optimization in high-dimensional spaces. arXiv preprint arXiv:1706.01445.

Bayesian optimization (BO) has become an effective approach for black-box function optimization problems when function evaluations are expensive, and the optimum can be achieved within a relatively small number of queries. However, many cases, such as the ones with high-dimensional inputs, may require a much larger number of observations for optimization. Despite an abundance of observations thanks to parallel experiments, current BO techniques have been limited to merely a few thousand observations. In this paper, we propose ensemble Bayesian optimization (EBO) to address three current challenges in BO simultaneously: large-scale observations, high dimensional input spaces, and selections of batch queries that balance quality and diversity. The key idea of EBO is to operate on an ensemble of additive Gaussian process (GP) models, each of which possesses a randomized strategy to divide and conquer. We show unprecedented, previously impossible results of scaling up BO to tens of thousands of observations within minutes of computation.

3.5 Efficient escape of saddle points and reaching stationary points in non-convex optimization

Mokhtari, A., Ozdaglar, A., & Jadbabaie, A. (2018). Escaping saddle points in constrained optimization. In Advances in Neural Information Processing Systems (pp. 3629-3639).

There has been a recent revival of interest in large-scale, scalable, non-convex optimization, due to obvious applications in machine learning. In convex problems, finding a first-order stationary point is often sufficient since it leads to finding an approximate local (and hence global) minimum. However, in the nonconvex setting, even when the problem is unconstrained, convergence to a first-order stationary point is not enough as the critical point to which convergence is established might be a saddle point. It is therefore natural to look at higher order derivatives and search for a second-order stationary point. Indeed, under the assumption that all the saddle points are strict, in both unconstrained and constrained settings, convergence to a second order stationary point implies convergence to a local minimum. While convergence to a second order stationary point has been thoroughly investigated in the recent literature for the unconstrained setting; the iteration complexity of the convex-constrained setting has not been studied yet.

Our research focuses on escaping from saddle points in smooth nonconvex optimization problems subject to a convex set and achieving a second-order stationary point which is a good approximation of the global minimum in many nonconvex problems that appear in machine learning, including matrix completion, dictionary learning, phase retrieval, and certain classes of deep neural networks. In particular, we propose a generic framework that yields convergence to a second-order stationary point, if the convex constraint set is simple for a quadratic objective function. To be more precise, our results hold if one can find a constant factor approximate solution of a quadratic program subject to the constraint in polynomial time. Under this condition, we show that the sequence of iterates generated by the proposed framework reaches a second-order stationary point in a polynomial number of iterations. We further characterize the overall arithmetic operations to reach a second-order stationary point when the constraint set can be written as a set of quadratic constraints. Finally, we extend our results to the stochastic setting and characterize the number of stochastic gradient and Hessian evaluations to reach a second-order stationary point.

In this paper, we study the problem of escaping from saddle points in smooth nonconvex optimization problems subject to a convex set C. We propose a generic framework that yields convergence to a second-order stationary point of the problem, if the convex set C is simple for a quadratic objective function. Specifically, our results hold if one can find a ρ-approximate solution of a quadratic program subject to C in polynomial time, where $\rho < 1$ is a positive constant that depends on the structure of the set C. Under this condition, we show that the sequence of iterates generated by the proposed framework reaches an (ρ, γ)-second order stationary point (SOSP) in at most $O(\max\{ \rho^{-2} , \rho^{-3} \gamma^{-3} \})$ iterations. We further characterize the overall complexity of reaching an SOSP when the convex set C can be written as a set of quadratic constraints and the objective function Hessian has a specific structure over the convex set C. Finally, we extend our results to the stochastic setting and characterize the number of stochastic gradient and Hessian evaluations to reach an (ρ, γ)-SOSP.

3.6 Understanding trade-offs between Over-fitting, interpolation, and generalization in optimization for statistical learning.

Belkin, M., Rakhlin, A., & Tsybakov, A. B. (2018). Does data interpolation contradict statistical optimality?. arXiv preprint arXiv:1806.09471. (See attachment Reference 3.6.pdf.)

Liang, T., & Rakhlin, A. (2018). Just interpolate: Kernel "ridgeless" regression can generalize. arXiv preprint arXiv:1808.00387. (See attachment Reference Liang 3.6 and 3.11.pdf.)

Our research over the 18 months on this topic was focused on the following fundamental question: can a learning method be successful out-of-sample if it interpolates data? It is usually taught in both Machine Learning and Statistics courses that data memorization is a bad idea from generalization point of view. In joint work "Does data interpolation contradict statistical optimality?" (with M. Belkin and A. Tsybakov, arXiv:1806.09471v1) we challenged this point of view, showing that a classical nonparametric estimator (the Nadaraya-Watson estimator) with an appropriately chosen kernel fits the data exactly while being optimal in terms of out-of-sample performance. We continued this line of work in "Just Interpolate: Kernel 'Ridgeless' Regression Can Generalize" (with T. Liang, arXiv:1501.06598), showing that kernel ridge regression (a classical method in machine learning and statistics) can generalize even if the regularization it turned off (in which case the method achieves exact fit to data). The analysis uncovers a new implicit regularization mechanism that is due to high dimensionality of the data, curvature of the kernel function, and favorable geometric properties of the data. These two papers challenge the common belief that a statistical or learning procedure necessarily overfits if it interpolates the data. The motivation for looking at this question lies, in part, in the recent success of deep learning methods, which have the flexibility to fit data exactly. Our findings have implication for both theory and practice of machine learning, and suggest further avenues of investigation that includes the optimization side of the problem of interpolation.

3.7 Finite sample expressive power of small-width ReLU networks

Yun, C., Sra, S., & Jadbabaie, A. (2019). Small ReLU networks are powerful memorizers: a tight analysis of memorization capacity. Accepted to NeurIPS 2019. (See attachment Reference 3.7 and 3.19 (Yun).pdf.)

In our recent work we (Jadbabaie and Sra), together with our student Chulhee Yun, study universal finite sample expressivity of neural networks, defined as the capability to perfectly memorize arbitrary datasets. For scalar outputs, existing results require a hidden layer as wide as N to memorize N data points. In contrast, we prove that a 3-layer (2-hidden-layer) ReLU network with $4\sqrt{N}$ hidden nodes can perfectly fit any arbitrary dataset. For K-class classification, we prove that a 4-layer ReLU network with $4\sqrt{N}+4K$ hidden neurons can memorize arbitrary datasets. For example, a 4-layer ReLU network with only 8,000 hidden nodes can memorize datasets with N = 1,000,000 and K = 1,000 (e.g., ImageNet). Our results show that even small networks already have tremendous overfitting capability, admitting zero empirical risk for any dataset. We also extend our results to deeper and narrower networks, and prove converse results showing necessity of $\Omega(N)$ parameters for shallow networks.

3.8 Efficient nonconvex empirical risk minimization via adaptive sample size methods

Mokhtari, A., Ozdaglar, A., & Jadbabaie, A. (2019, April). Efficient Nonconvex Empirical Risk Minimization via Adaptive Sample Size Methods. In The 22nd International Conference on Artificial Intelligence and Statistics (pp. 2485-2494).

We are interested in finding a local minimizer of an ERM problem where the loss associated with each sample is possibly a nonconvex function. Unlike traditional deterministic and stochastic algorithms that attempt to solve the ERM problem for the full training set, we propose an adaptive sample size scheme to reduce the overall computational complexity of finding a local minimum. To be more precise, we first find an approximate local minimum of the ERM problem corresponding to a small number of samples and use the uniform convergence theory to show that if the population risk is a Morse function, by properly increasing the size of training set the iterates generated by the proposed procedure always stay close to a local minimum of the corresponding ERM problem. Therefore, eventually the proposed procedure finds a local minimum of the ERM corresponding to the full training set which happens to also be a local minimum of the ERM problem with high probability. We formally state the conditions on the size of the initial sample set and characterize the required accuracy for obtaining an approximate local minimum to ensure that the iterates always stay in a neighborhood of a local minimum and do not get attracted to saddle points.

In this paper, we are interested in finding a local minimizer of an ERM problem where the loss associated with each sample is possibly a nonconvex function. Unlike traditional deterministic and stochastic algorithms that attempt to solve the ERM problem for the full training set, we propose an adaptive sample size scheme to reduce the overall computational complexity of finding a local minimum. To be more precise, we first find an approximate local minimum of the ERM problem corresponding to a small number of samples and use the uniform convergence theory to show that if the population risk is a Morse function, by properly increasing the size of training set the iterates generated by the proposed procedure always stay close to a local minimum of the corresponding ERM problem. Therefore, eventually the proposed procedure finds a local minimum of the ERM corresponding to the full training set which happens to also be a local minimum of the ERM problem with high probability. We formally state the conditions on the size of the initial sample set and characterize the required accuracy for obtaining an approximate local minimum to ensure that the iterates always stay in a neighborhood of a local minimum and do not get attracted to saddle points.

3.9 Achieving acceleration in distributed optimization via direct discretization of the Heavy-Ball ODE

Zhang, J., Uribe, C. A., Mokhtari, A., & Jadbabaie, A. (2019, July). Achieving acceleration in distributed optimization via direct discretization of the heavy-ball ODE. In 2019 American Control Conference (ACC) (pp. 3408-3413). IEEE.

We follow up our previous result showing that gradient-based optimization methods achieve acceleration by directly discretizing a second order ODE related to the continuous limit of Nesterov's accelerated gradient method. In our recent work we (Jadbabaie), together with student

Jingzhao Zhang and postdoctoral scholars Aryan Mokhtari and Cesar A. Uribe extend the dynamical system point of view to the design of accelerated algorithms for distributed large-scale optimization problems over networks. Particularly, we develop a distributed algorithm for solving problem of minimizing large but finite sum of convex functions over networks. The proposed algorithm is derived from directly discretizing the second-order heavy-ball differential equation and achieves acceleration: a convergence rate faster than distributed gradient descent-based methods for strongly convex objectives that may not be smooth. Notably, we achieve acceleration without resorting to well-known Nesterov's momentum approach. We provide numerical experiments and contrast the proposed method with recently proposed optimal distributed optimization algorithms.

3.10 On increasing self-confidence in non-Bayesian social learning over time-varying directed graphs

Uribe, C. A., & Jadbabaie, A. (2019, July). On Increasing Self-Confidence in Non-Bayesian Social Learning over Time-Varying Directed Graphs. In 2019 American Control Conference (ACC) (pp. 3532-3537). IEEE.

One important aspect of large-scale optimization problems is the availability of large quantities of data, which in turn can be distributed locally among several data centers. Given that the data is not available at a central location, distributed approaches play an important role to handle this limited information scenario. Several aspects of consensus-based algorithms have been studied in the literature. However, most of the existing results assume the existence of some persistent communication between agents or nodes where the data is located and processed.

In our recent work we (Jadbabaie), together with postdoctoral scholar Cesar A. Uribe, studied the convergence of the log-linear non-Bayesian social learning update rule, for a group of agents that collectively seek to identify a parameter that best describes a joint sequence of observations. Contrary to recent literature, we focus on the case where agents assign decaying weights to its neighbors, and the network is not connected at every time instant but over some finite time intervals. We provide a necessary and sufficient condition for the rate at which agents decrease the weights and still guarantees social learning.

3.11 Interpolation as a learning mechanism.

Liang, T., & Rakhlin, A. (2018). Just interpolate: Kernel "ridgeless" regression can generalize. arXiv preprint arXiv:1808.00387. (See attachment Reference Liang 3.6 and 3.11.pdf.)

We (Rakhlin), and student Xiyu Zhai, have been investigating the question of interpolation as a learning mechanism. The prior work by PI Rakhlin and T. Liang showed that one can interpolate the data using infinite-dimensional kernel Hilbert space functions, yet still perform well in terms of out of sample prediction. Our approach in that work relied on a high-dimensional phenomenon for random kernel matrices. A natural follow-up question was whether high dimensionality of the data is necessary to show good properties of interpolation. In the present work with Xiyu Zhai, we proved that interpolation cannot succeed in low dimensions, under

certain general assumptions. Our work sheds further light on the nature of interpolation as a learning mechanism.

In the absence of explicit regularization, kernel ridgeless regression with nonlinear kernels has the potential to fit the training data perfectly. It has been observed empirically, however, that such interpolated solutions can still generalize well on test data. We isolate a phenomenon of implicit regularization for minimum-norm interpolated solutions which is due to a combination of high dimensionality of the input data, curvature of the kernel function, and favorable geometric properties of the data such as an eigenvalue decay of the empirical covariance and kernel matrices. In addition to deriving a data-dependent upper bound on the out-of-sample error, we present experimental evidence suggesting that the phenomenon occurs in the MNIST dataset.

3.12 Stable Optimization with Gaussian Processes

Bogunovic, I., Scarlett, J., Jegelka, S., & Cevher, V. (2018). Adversarially robust optimization with Gaussian processes. In Advances in Neural Information Processing Systems (pp. 5760-5770).

GPs provide powerful means for sequentially optimizing a black-box function f that is costly to evaluate and for which noisy point evaluations are available. Since its introduction, this approach has successfully been applied to numerous applications, including robotics, algorithm parameter tuning, recommender systems, environmental monitoring, and many more. In many such applications, one is faced with various forms of uncertainty that are not accounted for by standard algorithms. In robotics, the optimization is often performed via simulations, creating a mismatch between the assumed function and the true one; in parameter tuning, the function is typically similarly mismatched due to limited training data; in recommendation systems and several other applications, the underlying function is inherently time-varying, so the returned solution may become increasingly stale over time; the list goes on.

We (Jegelka) address these considerations by studying the GP optimization problem with an additional requirement of stability or robustness: the returned point is perturbed by an adversary, and we seek to ensure that this perturbation degrades the function value as little as possible. This problem is of interest not only for attaining improved robustness to uncertainty, but also for settings where one seeks a region of good solutions rather than a single point, and for other related max-min optimization settings. We show that standard GP optimization algorithms do not exhibit the desired robustness properties and give a novel confidence-bound based algorithm StableOpt for this purpose. We rigorously establish the required number of samples for StableOpt to find a near-optimal point, and we complement this guarantee with an algorithm-independent lower bound. We experimentally demonstrate a variety of potential applications of interest on real-world data sets, and we show that StableOpt consistently succeeds in finding a stable maximizer where several baseline methods fail.

3.13 Small nonlinearities in activation functions create bad local minima in neural networks

Yun, C., Sra, S., & Jadbabaie, A. (2018). Small nonlinearities in activation functions create bad local minima in neural networks. arXiv preprint arXiv:1802.03487. Accepted to ICLR | 2019, Seventh International Conference on Learning.

We investigate the loss surface of neural networks. We prove that even for one-hidden-layer networks with slightest nonlinearity, the empirical risks have spurious local minima in most cases. Our results thus indicate that in general having no spurious local minima is a property limited to deep linear networks, and insights obtained from linear networks are not robust. Specifically, for ReLU(-like) networks we constructively prove that for almost all (in contrast to previous results) practical datasets there exist infinitely many local minima. We also present a counterexample for more general activations (sigmoid, tanh, arctan, ReLU, etc.), for which there exists a bad local minimum. Our results make the least restrictive assumptions relative to existing results on local optimality in neural networks. We complete our discussion by presenting a comprehensive characterization of global optimality for deep linear networks, which unifies other results on this topic.

3.14 Efficiently testing local optimality and escaping saddles for ReLU networks

Yun, C., Sra, S., & Jadbabaie, A. (2018). Efficiently testing local optimality and escaping saddles for ReLU networks. arXiv preprint arXiv:1809.10858. Accepted to ICLR | 2019, Seventh International Conference on Learning.

We provide a theoretical algorithm for checking local optimality and escaping saddles at nondifferentiable points of empirical risks of two-layer ReLU networks. Our algorithm receives any parameter value and returns: local minimum, second-order stationary point, or a strict descent direction. The presence of M data points on the nondifferentiability of the ReLU divides the parameter space into at most 2^M regions, which makes analysis difficult. By exploiting polyhedral geometry, we reduce the total computation down to one convex quadratic program (QP) for each hidden node, O(M) (in)equality tests, and one (or a few) nonconvex QP. For the last QP, we show that our specific problem can be solved efficiently, in spite of nonconvexity. In the benign case, we solve one equality constrained QP, and we prove that projected gradient descent solves it exponentially fast. In the bad case, we have to solve a few more inequality constrained QPs, but we prove that the time complexity is exponential only in the number of inequality constraints. Our experiments show that either benign case or bad case with very few inequality constraints occurs, implying that our algorithm is efficient in most cases.

3.15 R-SPIDER: A Fast Riemannian Stochastic Optimization Algorithm with Curvature Independent Rate

Zhang, J., Zhang, H., & Sra, S. (2018). R-spider: A fast Riemannian stochastic optimization algorithm with curvature independent rate. arXiv preprint arXiv:1811.04194.

We study smooth stochastic optimization problems on Riemannian manifolds. Via adapting the recently proposed SPIDER algorithm proposed by Fang and collaborators (cf.

arXiv:1807.01695v2) to Riemannian manifolds, we can achieve faster rate than known algorithms in both the finite sum and stochastic settings. Unlike previous works, by not resorting to bounding iterate distances, our analysis yields curvature independent convergence rates for both the nonconvex and strongly convex cases.

3.16 Consistency of Interpolation with Laplace Kernels is a High-Dimensional Phenomenon

Rakhlin, A., & Zhai, X. (2018). Consistency of Interpolation with Laplace Kernels is a High-Dimensional Phenomenon. arXiv preprint arXiv:1812.11167.

We show that minimum-norm interpolation in the reproducing kernel Hilbert space (RKHS) corresponding to the Laplace kernel is not consistent if input dimension is constant. The lower bound holds for any choice of kernel bandwidth, even if selected based on data. The result supports the empirical observation that minimum-norm interpolation (that is, exact fit to training data) in RKHS generalizes well for some high-dimensional datasets, but not for low-dimensional ones.

3.17 What Can Neural Networks Reason About?

Xu, K., Li, J., Zhang, M., Du, S. S., Kawarabayashi, K. I., & Jegelka, S. (2019). What Can Neural Networks Reason About?. arXiv preprint arXiv:1905.13211.

Neural networks have successfully been applied to solving reasoning tasks, ranging from learning simple concepts like "close to", to intricate questions whose reasoning procedures resemble algorithms. Empirically, not all network structures work equally well for reasoning. For example, GNNs have achieved impressive empirical results, while less structured neural networks may fail to learn to reason. Theoretically, there is currently limited understanding of the interplay between reasoning tasks and network learning. In this paper, we develop a framework to characterize which tasks a neural network can learn well, by studying how well its structure aligns with the algorithmic structure of the relevant reasoning procedure. This suggests that GNNs can learn dynamic programming, a powerful algorithmic strategy that solves a broad class of reasoning problems, such as relational question answering, sorting, intuitive physics, and shortest paths. Our perspective also implies strategies to design neural architectures for complex reasoning. On several abstract reasoning tasks, we see empirically that our theory aligns well with practice.

3.18 How Powerful are Graph Neural Networks?

Xu, K., Hu, W., Leskovec, J., & Jegelka, S. (2018). How powerful are graph neural networks?. arXiv preprint arXiv:1810.00826.

GNNs are an effective framework for representation learning of graphs. GNNs follow a neighborhood aggregation scheme, where the representation vector of a node is computed by recursively aggregating and transforming representation vectors of its neighboring nodes. Many GNN variants have been proposed and have achieved state-of-the-art results on both node and graph classification tasks. However, despite GNNs revolutionizing graph representation learning,

there is limited understanding of their representational properties and limitations. Here, we present a theoretical framework for analyzing the expressive power of GNNs to capture different graph structures. Our results characterize the discriminative power of popular GNN variants, such as graph convolutional networks and GraphSAGE, and show that they cannot learn to distinguish certain simple graph structures. We then develop a simple architecture that is provably the most expressive among the class of GNNs and is as powerful as the Weisfeiler-Lehman graph isomorphism test. We empirically validate our theoretical findings on a number of graph classification benchmarks, and demonstrate that our model achieves state-of-the-art performance.

3.19 Small ReLU networks are powerful memorizers: a tight analysis of memorization capacity.

Yun, C., Sra, S., & Jadbabaie, A. (2019). Small ReLU networks are powerful memorizers: a tight analysis of memorization capacity. Accepted to Neural Information Processing Systems (NeurIPS) 2019 (Spotlight). (See attachment Reference 3.7 and 3.19 (Yun).pdf.)

We study finite sample expressivity, i.e., memorization power of ReLU networks. We show that 3-layer ReLU networks with $\Omega(\sqrt{N})$ hidden nodes can perfectly memorize most datasets with N points. We also prove that width $\Omega(\sqrt{N})$ is necessary and sufficient for memorizing N data points, proving tight bounds on memorization capacity. For deeper networks, we show that an L-layer network with W parameters in the hidden layers can memorize N data points if $W = \Omega(N)$. Combined with a recent upper bound $O(WlogW)$ on Vapnik Chervonenkis dimension, our construction is almost tight for any fixed L, i.e., the result cannot be further improved upon. Subsequently, we analyze memorization capacity of residual networks under a general position assumption; we prove results that substantially reduce the known requirement of N hidden nodes. Finally, we study dynamics of stochastic gradient descent (SGD), and show that when initialized near a memorizing global minimum of the empirical risk, SGD quickly finds a nearby point with small empirical risk

3.20 Are deep ResNets provably better than linear predictors?

Yun, C., Sra, S., & Jadbabaie, A. (2019). Are deep ResNets provably better than linear predictors?. arXiv preprint arXiv:1907.03922. Accepted to Neural Information Processing Systems (NeurIPS) 2

Recent results in the literature indicate that a residual network (ResNet) composed of a single residual block outperforms linear predictors, in the sense that all local minima in its optimization landscape are at least as good as the best linear predictor. However, these results are limited to a single residual block, instead of the deep ResNets composed of multiple residual blocks. We take a step towards extending this result to deep ResNets. We start by two motivating examples. First, we show that there exist datasets for which all local minima of a fully-connected ReLU network are no better than the best linear predictor, whereas a ResNet can have strictly better local minima. Second, we show that even at its global minimum, the representation obtained from the residual blocks of a 2-block ResNet do not necessarily improve monotonically over subsequent blocks, which highlights a fundamental difficulty in analyzing deep ResNets. Our main theorem

on deep ResNets shows under simple geometric conditions that, any critical point in the optimization landscape is either (i) at least as good as the best linear predictor; or (ii) the Hessian at this critical point has a strictly negative eigenvalue. Notably, our results show that even without using direct skip-connections from input layer to the last hidden layer, multiple skip-connections can improve the optimization landscape. Finally, we complement our results by showing benign properties of the near-identity regions of deep ResNets, showing size-independent upper bounds for the risk attained at critical points as well as the Rademacher complexity.

3.21 Competitive Contagion with Spare Seeding.

M. Siami, A. Ajorlou, and A. Jadbabaie, "Competitive Contagion with Spare Seeding," IFAC Workshop on Distributed Estimation and Control in Networked Systems (NecSys19).

This paper studies a strategic model of marketing and product diffusion in social networks. We consider two firms offering substitutable products which can improve their market share by seeding the key individuals in the market. Consumers update their consumption level for each of the two products as the best response to the consumption of their neighbors in the previous period. This results in linear update dynamics for the product consumption. Each consumer receives externality from the consumption of each neighbor where the strength of the externality is higher for consumption of the products of the same firm. We represent the above setting as a duopoly game between the firms and introduce a novel framework that allows for sparse seeding to emerge as an equilibrium strategy. We then study the effect of the network structure on the optimal seeding strategies and the extent to which the strategies can be sparsified. In particular, we derive conditions under which near Nash equilibrium strategies can asymptotically lead to sparse seeding in large populations. The results are illustrated using a core-periphery network.

3.22 A Separation Principle for Joint Sensor and Actuator Scheduling with Guaranteed Performance Bound.

M. Siami, and A. Jadbabaie, "A Separation Principle for Joint Sensor and Actuator Scheduling with Guaranteed Performance Bounds," The 58th IEEE Conference on Decision and Control, Nice, France, 2019.

We study the problem of jointly designing a sparse sensor and actuator schedule for linear dynamical systems while guaranteeing a control/estimation performance that approximates the fully sensed/actuated setting. We further prove a separation principle, showing that the problem can be decomposed into finding sensor and actuator schedules separately. However, it is shown that this problem cannot be efficiently solved or approximated in polynomial, or even quasi-polynomial time for time-invariant sensor/actuator schedules; instead, we develop deterministic polynomial-time algorithms for a time-varying sensor/actuator schedule with guaranteed approximation bounds. Our main result is to provide a polynomial-time joint actuator and sensor schedule that on average selects only a constant number of sensors and actuators at each time step, irrespective of the dimension of the system. The key idea is to sparsify the controllability and observability Gramians while providing approximation guarantees for Hankel singular

values. This idea is inspired by recent results in theoretical computer science literature on sparsification

3.23 Non-Bayesian Social Learning with Uncertain Models over Time-Varying Directed Graphs.

Uribe, C. A., Hare, J. Z., Kaplan, L., & Jadbabaie, A. (2019). Non-Bayesian Social Learning with Uncertain Models over Time-Varying Directed Graphs. arXiv preprint arXiv:1909.04255. Accepted to The 58th IEEE Conference on Decision and Control, Nice, France, 2019.

We study the problem of non-Bayesian social learning with uncertain models, in which a network of agents seeks to cooperatively identify the state of the world based on a sequence of observed signals. In contrast with the existing literature, we focus our attention on the scenario where the statistical models held by the agents about possible states of the world are built from finite observations. We show that existing non-Bayesian social learning approaches may select a wrong hypothesis with non-zero probability under these conditions. Therefore, we propose a new algorithm to iteratively construct a set of beliefs that indicate whether a certain hypothesis is supported by the empirical evidence. This new algorithm can be implemented over time-varying directed graphs, with non-doubly stochastic weights.

3.24 Non-Bayesian Social Learning with Gaussian Uncertain Models

Hare, J. Z., Uribe, C. A., Kaplan, L., & Jadbabaie, A. (2019). 4.21 Non-Bayesian Social Learning with Gaussian Uncertain Models, Submitted to American Control Conference

Non-Bayesian social learning theory provides a framework for distributed inference of a group of agents interacting over a social network by sequentially communicating and updating beliefs about the unknown state of the world through likelihood updates from their observations. Typically, likelihood models are assumed known precisely. However, in many situations the models are generated from sparse training data due to lack of data availability, high cost of collection/calibration, limits within the communications network, and/or the high dynamics of the operational environment. Recently, social learning theory was extended to handle those model uncertainties for categorical models. In this paper, we introduce the theory of Gaussian uncertain models and study the properties of the beliefs generated by the network of agents. We show that even with finite amounts of training data, non-Bayesian social learning can be achieved and all agents in the network will converge to a consensus belief that provably identifies the best estimate for the state of the world given the set of prior information.

3.25 Non-Bayesian Social Learning with Uncertain Models

Hare, J. Z., Uribe, C. A., Kaplan, L., & Jadbabaie, A. (2019). Non-Bayesian Social Learning with Uncertain Models. arXiv preprint arXiv:1909.09228. Submitted to IEEE Transactions on Signal Processing.

Non-Bayesian social learning theory provides a framework that models distributed inference for a group of agents interacting over a social network. In this framework, each agent iteratively

forms and communicates beliefs about an unknown state of the world with their neighbors using a learning rule. Existing approaches assume agents have access to precise statistical models (in the form of likelihoods) for the state of the world. However, in many situations, such models must be learned from finite data. We propose a social learning rule that takes into account uncertainty in the statistical models using second-order probabilities. Therefore, beliefs derived from uncertain models are sensitive to the amount of past evidence collected for each hypothesis. We characterize how well the hypotheses can be tested on a social network, as consistent or not with the state of the world. We explicitly show the dependency of the generated beliefs with respect to the amount of prior evidence. Moreover, as the amount of prior evidence goes to infinity, learning occurs and is consistent with traditional social learning theory.

Major Announcements and Placements:

- PI Ali Jadbabaie was the lead organizer of a new DoD-supported conference at the interface of Learning, dynamical systems, and control theories. The conference is called L4DCand attracted nearly 400 researchers. The second conference will be held at Berkeley campus on June 10-11
- Lagrange postdoc Aryan Mokhtari joined the faculty of University of Texas at Austin as an Assistant Professor in the Electrical Engineering department
- Lagrange postdoc Cesar Uribe (who received partial support on this project), together with PI Ali Jadbabaie, collaborated with Dr. Lance Kaplan of Army Research Lab and his postdoc James Hare who is funded by a LUCI project from OSD.