AFRL-AFOSR-VA-TR-2021-0099

Learning Data Representations via Nonconvex Optimization

Soltanolkotabi, Mahdi
UNIVERSITY OF SOUTHERN CALIFORNIA
3720 S FLOWER ST FL 3
LOS ANGELES, CA, 90007
USA

08/17/2021
Final Technical Report

Air Force Research Laboratory
Air Force Office of Scientific Research
Arlington, Virginia 22203
Air Force Materiel Command

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From - To)*
17-08-2021	Final	01 Feb 2018 - 31 Jan 2021

4. TITLE AND SUBTITLE
Learning Data Representations via Nonconvex Optimization

5a. CONTRACT NUMBER

5b. GRANT NUMBER
FA9550-18-1-0078

5c. PROGRAM ELEMENT NUMBER
61102F

6. AUTHOR(S)
Mahdi Soltanolkotabi

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
UNIVERSITY OF SOUTHERN CALIFORNIA
3720 S FLOWER ST FL 3
LOS ANGELES, CA 90007
USA

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
AF Office of Scientific Research
875 N. Randolph St. Room 3112
Arlington, VA 22203

10. SPONSOR/MONITOR'S ACRONYM(S)
AFRL/AFOSR RTA2

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
AFRL-AFOSR-VA-TR-2021-0099

12. DISTRIBUTION/AVAILABILITY STATEMENT
A Distribution Unlimited: PB Public Release

13. SUPPLEMENTARY NOTES

14. ABSTRACT
In this project we developed a unified understanding of how to design and analyze efficient nonconvex optimization algorithms aimed at learning interpretable representations from data. These data representations can in turn enable automatic knowledge extraction from observed low-level sensory data enhancing a variety of applications. Our main results during the second year can be summarized in three categories: (1) understanding the optimization landscape of data representation tasks such as matrix factorization and shallow neural network training, (2) developing principled approaches to utilizing prior knowledge in data representation tasks and characterizing the reduction in the size of the training data that results from such usage of prior knowledge, and (3) developing algorithmic variations that can be implemented on often unreliable modern cloud infrastructure.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			FARIBA FAHROO
U	U	U	UU	7	**19b. TELEPHONE NUMBER** *(Include area code)* 426-8429

Final Performance Report for AFOSR award #FA9550-18-1-0078 Learning Data Representations via Nonconvex Optimization

PI: Mahdi Soltanolkotabi, University of Southern California
Program Manager: Dr. Fariba Fahroo & Dr. Warren Adams

Abstract

In this project we developed a unified understanding of how to design and analyze efficient nonconvex optimization algorithms aimed at learning interpretable representations from data. These data representations can in turn enable automatic knowledge extraction from observed low-level sensory data enhancing a variety of applications. Our main results during the second year can be summarized in three categories: (1) understanding the optimization landscape of data representation tasks such as matrix factorization and shallow neural network training, (2) developing principled approaches to utilizing prior knowledge in data representation tasks and characterizing the reduction in the size of the training data that results from such usage of prior knowledge, and (3) developing algorithmic variations that can be implemented on often unreliable modern cloud infrastructure.

The capability to automatically extract operationally relevant information from disparate data sources abroad aircrafts, ships, combat and weapon systems, etc. can tremendously enhance warfighting capabilities and greatly assist analysts in efficiently processing data. However, such data are often unlabeled, noisy and contains a rich mixture of information from many sources aggregated into a single recording or document. Therefore, *learning data representations* that help identify and disentangle the underlying explanatory factors hidden in such mixture data is fundamental to our ability to extract new and deep insights from data.

In this project we have developed a unified understanding of how to design and analyze efficient nonconvex optimization algorithms aimed at learning interpretable representations from data. These data representations can in turn enable automatic knowledge extraction from observed low-level sensory data. Our main results can be summarized in the following categories.

1 Understanding the optimization landscape of data representation tasks

Perhaps the most prominent method for learning data representations is training neural networks and low-rank models such as those arising in matrix factorization. However, the training algorithms involved for learning these representations are often nonconvex and it is completely unclear why the find good models. Furthermore, many modern learning tasks involve fitting nonlinear models to data which are trained in an overparameterized regime where the parameters of the model exceed the size of the training dataset. Due to this overparameterization, the training loss may have infinitely many global minima and it is critical to understand the properties of the solutions found by first-order optimization schemes such as (stochastic) gradient descent starting from different

initializations. In a series of papers we demonstrated that when the loss has certain properties over a minimally small neighborhood of the initial point, first order methods such as (stochastic) gradient descent have a few intriguing properties: (1) the iterates converge at a geometric rate to a global optima even when the loss is nonconvex, (2) among all global optima of the loss the iterates converge to one with a near minimal distance to the initial point, (3) the iterates take a near direct route from the initial point to this global optima. As evident from the papers below we have successfully applied this framework to a variety of problems including neural network training, low-rank factorization, training generalized linear model and learning ReLU nonlinearities. This line of inquiry has resulted in the following publications:

- Theoretical insights into the optimization landscape of over-parameterized shallow neural networks. M. Soltanolkotabi, A. Javanmard, and J. D. Lee. IEEE Trans. on Info. Theory.
- Overparameterized Nonlinear Learning: Gradient Descent Takes the Shortest Path? S. Oymak and M. Soltanolkotabi. International Conference on Machine Learning (ICML 2019).
- Fitting ReLUs via SGD and Quantized SGD. M. Mousavi Kalan, M. Soltanolkotabi, and A. Avestimehr. International Symposium on Information Theory (ISIT 2019).
- Towards moderate overparameterization: global convergence guarantees for training shallow neural networks. S. Oymak and M. Soltanolkotabi. Journal on Selected Areas of Information Theory, Deep Learning: Mathematical Foundations and Applications to Information Science, 2020.
- Gradient Descent with Early Stopping is Provably Robust to Label Noise for Overparameterized Neural Networks. M. Li, M. Soltanolkotabi, and S. Oymak. International Conference on Artificial Intelligence and Statistics (AISTATS 2020).
- Generalization Guarantees for Neural Networks via Harnessing the Low-rank Structure of the Jacobian. S. Oymak, Z. Fabian, M. Li, and M. Soltanolkotabi. Submitted.
- End-to-end Learning of a Convolutional Neural Network via Deep Tensor Decomposition. S. Oymak and M. Soltanolkotabi. Information and Inference, 2021.

2 Utilizing prior knowledge

We have developed a principled approach to utilize a priori knowledge in learning data representations by utilizing iterative shrinkage schemes. We have also precisely characterized the precise sample complexity (or size of training data) required for our approach to work. We have specialized this general theory in the context of the phase retrieval problem. The mathematical results in our work paves the way for a new generation of data-driven imaging systems at nano-scale that can utilize prior information to significantly reduce acquisition time and enhance image reconstruction, enabling nano-scale imaging at unprecedented speeds and resolutions. This work has resulted in the following papers.

- Structured signal recovery from quadratic measurements: Breaking sample complexity barriers via nonconvex optimization. M. Soltanolkotabi. IEEE Trans. on Info. Theory.

- Accelerated Wirtinger Flow for Multiplexed Fourier Ptychographic Microscopy. E. Bostan, M. Soltanolkotabi, D. Ren, and L. Waller. IEEE International Conference on Image Processing (ICIP 2018).

- 3D Phase Retrieval at Nano-Scale via Accelerated Wirtinger Flow. Z. Fabian, J. Haldar, R. Leahy, M. Soltanolkotabi. EUSIPCO 2020.

- Accelerated Wirtinger Flow: A Fast Algorithm for Ptychography. R. Xu, M. Soltanolkotabi, J. P. Haldar, W. Unglaub, J. Zusman, A. F. J. Levi, R. M. Leahy. Available on the PI's website.

3 Large-scale and distributed learning over cloud infrastructure

In addition to the traditional challenges that exist for distributed computing (e.g., resource allocation, task scheduling, and data-distribution), we face several new fundamental challenges for distributed computing over modern cloud infrastructure such as Amazon EC2. The main challenge is that a variety of hardware failures lead to slowdown and poor communication in some machines (a.k.a. stragglers). To address this challenge we we leverage coded computing, a new research area that we have been recently developing, which brings to bear ideas from modern coding/information theory to alleviate key bottlenecks, such as bandwidth, latency, and robustness, in large-scale distribute data analysis. In our work we create carefully designed redundancy in the data (via random encoding techniques) to mitigate the effect of stragglers. We have also developed novel mathematical understanding for this framework demonstrating its effectiveness in a variety of settings. We also characterize fundamental trade-offs between convergence rate, size of data set, accuracy, computational load (or data redundancy), and straggler toleration in this framework.

We also developed Lagrange Coded Computing (LCC), a new framework to simultaneously provide (1) resiliency against stragglers that may prolong computations; (2) security against Byzantine (or malicious) workers that deliberately modify the computation for their benefit; and (3) (information-theoretic) privacy of the dataset amidst possible collusion of workers. LCC, which leverages the well-known Lagrange polynomial to create computation redundancy in a novel coded form across workers, can be applied to any computation scenario in which the function of interest is an arbitrary multivariate polynomial of the input dataset, hence covering many computations of interest in machine learning. LCC enables secure and private computing in distributed settings, improving the computation and communication efficiency of the state-of-the-art. Furthermore, we prove the optimality of LCC by showing that it achieves the optimal tradeoff between resiliency, security, and privacy, i.e., in terms of tolerating the maximum number of stragglers and adversaries, and providing data privacy against the maximum number of colluding workers. Finally, we show via experiments on Amazon EC2 that LCC speeds up the conventional uncoded implementation of distributed least-squares linear regression by up to 13.43×, and also achieves a 2.36×-12.65× speedup over the state-of-the-art straggler mitigation strategies.

These efforts have collectively lead to the following papers:

- Near-Optimal Straggler Mitigation for Distributed Gradient Methods. S. Li, M. Mousavi Kalan, S. Avestimehr, and M. Soltanolkotabi. The 7th International Workshop on Parallel and Distributed Computing for Large Scale Machine Learning and Big Data Analytics.

- Fundamental Resource Trade-offs for Encoded Distributed Optimization. A. Avestimehr, M. Mousavi Kalan, and M. Soltanolkotabi. To appear in Information and Inference 2021.
- Lagrange Coded Computing: Optimal Design for Resiliency, Security and Privacy. Q. Yu, S. Li, N. Raviv, M. Mousavi Kalan, M. Soltanolkotabi, and S. Avestimehr. International Conference on Artificial Intelligence and Statistics (AISTATS 2019).

4 Demystifying deep image priors

Convolutional Neural Networks (CNNs) have emerged as highly successful tools for image generation, recovery, and restoration. This success is often attributed to large amounts of training data. However, recent experimental findings challenge this view and instead suggest that a major contributing factor to this success is that convolutional networks impose strong prior assumptions about natural images. A surprising experiment that highlights this architectural bias towards natural images is that one can remove noise and corruptions from a natural image without using any training data, by simply fitting (via gradient descent) a randomly initialized, over-parameterized convolutional generator to the single corrupted image. While this over-parameterized network can fit the corrupted image perfectly, surprisingly after a few iterations of gradient descent one obtains the uncorrupted image. This intriguing phenomena enables state-of-the-art CNN-based denoising and regularization of linear inverse problems such as compressive sensing. During the last year we took a step towards demystifying this experimental phenomena by attributing this effect to particular architectural choices of convolutional networks, namely convolutions with fixed interpolating filters. We then formally characterized the dynamics of fitting a two layer convolutional generator to a noisy signal and proved that earlystopped gradient descent denoises/regularizes. We provide a pictorial depiction of such denoising in Figure 1 for reference. This effort has led to the following publications

- Compressive sensing with un-trained neural networks: Gradient descent finds the smoothest approximation. R. Heckel and M. Soltanolkotabi. International Conference on Machine Learning (ICML 2020).
- Denoising and Regularization via Exploiting the Structural Bias of Convolutional Generators. R. Heckel and M. Soltanolkotabi. International Conference on Learning Representations (ICLR 2020).
- Compressed Sensing with Deep Image Prior and Learned Regularization. D. V. Veen, A. Jalal, M. Soltanolkotabi, E. Price, S. Vishwanath, A. G. Dimakis. Submitted 2021.

5 Precise Tradeoffs in Adversarial Training

Despite breakthrough performance, modern learning models are known to be highly vulnerable to small adversarial perturbations in their inputs. While a wide variety of recent adversarial training methods have been effective at improving robustness to perturbed inputs (robust accuracy), often this benefit is accompanied by a decrease in accuracy on benign inputs (standard accuracy), leading to a tradeoff between often competing objectives. Complicating matters further, recent empirical evidence suggest that a variety of other factors (size and quality of training data, model size,

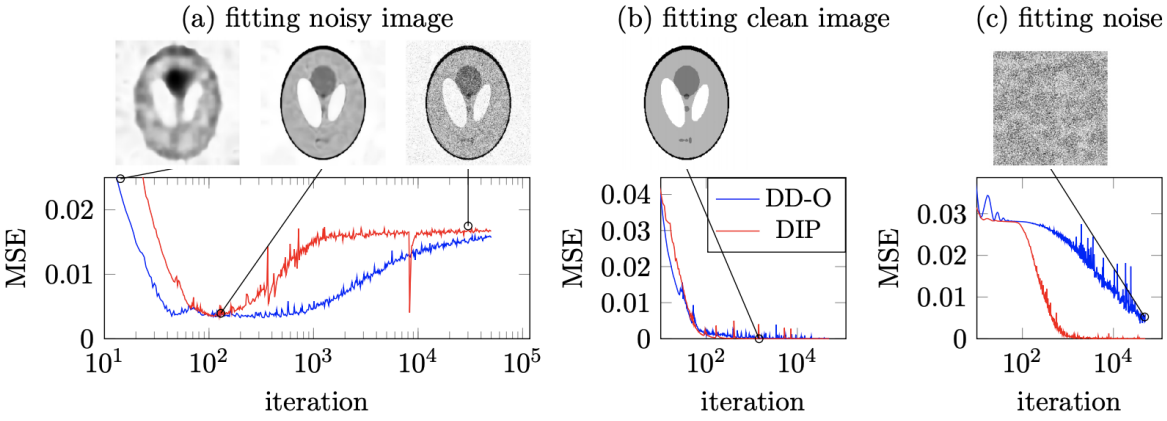


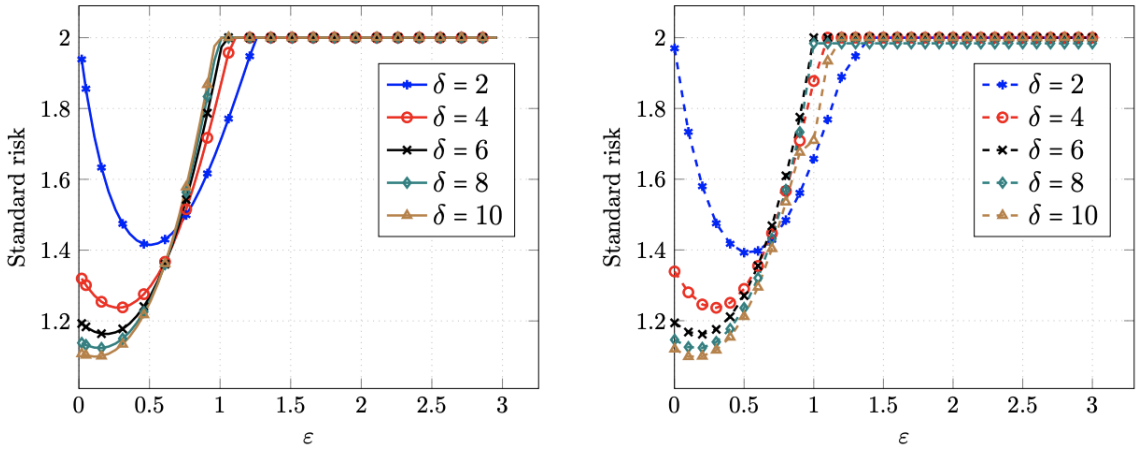
Figure 1: Fitting an over-parameterized Deep Decoder (DD-O) and the deep image prior (DIP) to a (a) noisy image, (b) clean image, and (c) pure noise. Here, MSE denotes Mean Square Error of the network output with respect to the clean image in (a) and fitted images in (b) and (c). While the network can fit the noise due to over-parameterization, it fits natural images in significantly fewer iterations than noise. Hence, when fitting a noisy image, the image component is fitted faster than the noise component which enables denoising via early stopping.

etc.) affect this tradeoff in somewhat surprising ways. In the last year we provided a precise and comprehensive understanding of the role of adversarial training in the context of linear regression with Gaussian features. In particular, we characterized the fundamental tradeoff between the accuracies achievable by any algorithm regardless of computational power or size of the training data. Furthermore, we precisely characterized the standard/robust accuracy and the corresponding tradeoff achieved by a contemporary mini-max adversarial training approach in a high-dimensional regime where the number of data points and the parameters of the model grow in proportion to each other. Our theory for adversarial training algorithms also enabled the rigorous study of how a variety of factors (size and quality of training data, model overparametrization etc.) affect the tradeoff between these two competing accuracies. This effort has resulted in the following publications:

- Precise Statistical Analysis of Classification Accuracies for Adversarial Training. A. Javanmard and M. Soltanolkotabi. Submitted 2020.
- Precise Tradeoffs in Adversarial Training for Linear Regression. A. Javanmard, M. Soltanolkotabi, and H. Hassani. Conference on Learning Theory (COLT 2020).

6 Robustness

Most modern learning algorithms require high-quality training data to work effectively. Understanding the existing limitations of learning algorithms and how to overcome these short-comings



(a) Theoretical curves

(b) Empirical curves

Figure 2: Standard risk versus the adversary's power (ε) for several values of δ (ratio of number of samples to model size). Left panel corresponds to the theoretical prediction curves and the right panel empirical performance. We see that the theoretical predictions are a near perfect match with the empirical results. The empirical results are averaged over 100 different realizations of noise and features. As δ grows, we observe a slower decay in the standard risk at small ε due to adversarial training. For $\delta = 10$, the standard risk has a small initial slope with respect to ε and then starts to increase rapidly. Put differently, with larger δ, the negative effect of adversarial training on the standard risk starts at smaller ε.

is crucial for their deployment in safety critical systems spanning autonomous driving to healthcare technology. In our group we made progress towards such robustification in the following two fronts.

6.1 High-dimensional Robust Mean Estimation via gradient descent

We studied the problem of high-dimensional robust mean estimation in the presence of a constant fraction of adversarial outliers. A recent line of work has provided sophisticated polynomial-time algorithms for this problem with dimension-independent error guarantees for a range of natural distribution families. In our work, we showed that a natural non-convex formulation of the problem can be solved directly by iterative shrinkage schemes. Our approach leverages a novel structural lemma, roughly showing that any approximate stationary point of our non-convex objective gives a near-optimal solution to the underlying robust estimation task. Our work establishes an intriguing connection between algorithmic high-dimensional robust statistics and non-convex optimization, which may have broader applications to other robust estimation tasks. This work has resulted in the following publication:

- High-dimensional Robust Mean Estimation via Gradient Descent. Y. Cheng, I. Diakonikolas, R. Ge, and M. Soltanolkotabi. International Conference on Machine Learning (ICML 2020).

6.2 Agnostically learning ReLU nonlinearties

We considered the fundamental problem of ReLU regression, where the goal is to output the best fitting ReLU with respect to square loss given access to draws from some unknown distribution. We gave the first efficient, constant-factor approximation algorithm for this problem assuming the underlying distribution satisfies some weak concentration and anti-concentration conditions (and includes, for example, all log-concave distributions). This resolved an open problem of Goel et al., who proved hardness results for any exact algorithm for ReLU regression (up to an additive factor). Using more sophisticated techniques, we can also improve our results and obtain a polynomial-time approximation scheme for any subgaussian distribution. Given the aforementioned hardness results, these guarantees can not be substantially improved. This work has resulted in the following paper:

- Approximation Schemes for ReLU Regression. I. Diakonikolas, S. Goel, S. Karmalkar, A. Klivans, and M. Soltanolkotabi. Conference on Learning Theory (COLT 2020).