Main research contributions

Ioannis Mitliagkas and group

August 2021

This document proves an executive summary of the work on machine learning optimization by me and my group in the past 4 years.

1 Smooth games and their numerical methods

The success of GANs [14] for generative modeling have recently spurred interest in machine learning for the optimization of smooth games, where interacting agents minimize different objectives (e.g. the generator and discriminator for GAN). Smooth games also appear in various setups like domain adaptation [3] and particular formulations of reinforcement learning [22]. On the other hand, this multi-objective optimization gives rise to much different behavior than in standard objective minimization and could benefit from tailored optimization algorithms different than just simple SGD [9].

Our work In this line of research, we pushed forward the analysis and development of algorithms for smooth game optimization in the deterministic and stochastic setting.

Our 2019 AISTATS paper was one of the first pieces of work to motivate and push the ML community for a deeper foundational analysis of adversarial problems and their methods [13]. Researchers in the field had been using the same methods as in single-objective optimization and also using the same hyper-parameter values. First we showed that positive momentum is problematic for that class of problems and proved that negative values of momentum are often optimal [13]. We then pointed out that the ML and mathematical optimization communities do not know what rates are optimal; the fundamental limits were missing. Our 2019 ICML work addresses this issue by providing the first linear lower bounds [17] and condition numbers for smooth games.

Then, equipped with an idea of optimality, we tackled the question of acceleration. First we used classic spectral tools used on classic linear systems work to establish optimal methods for quadratic games [8]. Then, we provided a deeper study of the popular extragradient method. That work provides the tightest and most general guarantees for this very important method in the field [7].
Finally, we tackled the question of *stochasticity* which was known to be more insidious in adversarial problems [11]. We adopted the recently proposed *Hamiltonian* family of methods and provided the first global non-asymptotic last-iterate convergence guarantees a class of stochastic games notably including some non-convex non-concave problems [19]. In a recent pre-print we dug deeper into the connection between stochasticity. We we introduce the expected co-coercivity condition, explain its benefits, and provide the first last-iterate convergence guarantees of SGDA and SCO under this condition for solving a class of stochastic variational inequality problems that are potentially non-monotone.

Motivated by the importance and increasing popularity of work in the area, we organized two consecutive NeurIPS workshops:

- **BRIDGING GAME THEORY AND DEEP LEARNING**, [https://sgo-workshop.github.io](https://sgo-workshop.github.io)

2 Modern optimization and deep learning

At the heart of the training algorithm for deep networks lies an optimization algorithm. Variants of stochastic gradient descent (SGD) have become the workhorse for modern large-scale optimization typical of machine learning [10], but many open questions remain.

**Our work** The goal of this research axis has been to develop and analyze new optimization algorithms in the context of modern deep learning.

In an oral SysML 2019 paper we gave an adaptive momentum method, with empirically better generalization properties that Adam and other popular optimizers [26]. In collaboration with colleagues at Google Brain, we then turned our attention to parameter-free stochastic versions of SGD, which provide the optimal variance reduction of online optimization and work for deterministic methods without different tuning. Our paper was selected for oral presentation at NeurIPS 2019 [4]. More recently, my students and I have gotten involved in a deep, fundamental study of different definitions of the condition number typically used in optimization (AISTATS 2021) [15]. We believe that this work is bound to have deep repercussions on our definitions for optimality and acceleration, but could also lead to better methods. We are currently following up on that foundational work.

On a slightly different thread we have been exploring alternatives to back-propagation; i.e. methods that do not exactly calculate the gradient. We first published to ICLR 2019 an empirical exploration of variants of backpropagation with better performance on LSTM models [6]. We also have a paper under submission on the analysis of feedback alignment, proposed in [18].
The generalization properties of deep learning

Optimization seeks to minimize an objective (e.g. the training error of a neural network) while the goal in supervised learning is to generalize well (i.e. small test error on unseen data). The empirical success of large overparameterized models have recently made the community revisit the interplay between optimization and generalization. In particular, modern neural networks have the capacity to overfit the training data and yet standard SGD algorithms often appear to yield networks with good generalization [25, 5]. A promising line of work to explain this behavior is to investigate the implicit regularization bias of optimization algorithms on modern architectures [21, 16, 23, 24].

Our work We further our theoretical understanding of deep learning by studying the basics of the bias-variance decomposition, providing robust methodology for large-scale empirical generalization studies and provide methodology and analysis the problems in the wide area of out-of-distribution generalization.

We provided the first modern, large-scale measurement of bias and variance in the predictions of neural networks [20]. There we showed that the classic understanding of the behavior of variance was incorrect; a phenomenon later dubbed as double descent. More recently, in a NeurIPS 2020 collaboration with colleagues from UoT, we borrowed robust tools from causal inference to provide robust methodology for the large-scale empirical study of generalization performance in neural networks [12].

On a second thread, we look at the problem of out-of-distribution generalization; a learning setting where the test examples are not drawn from the same distribution as the training examples. We start from the slightly more limited setting of domain generalization and provide a method based on distribution matching [2]. Motivated by this work, we look deeper into the fundamental limits of out-of-distribution generalization [1].

References


