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 hyperparameter 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 con- vergence 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:

- SMOOTH GAMES OPTIMIZATION AND MACHINE LEARNING WORKSHOP, `https://sgo-workshop.github.io/index_2018.html`
- BRIDGING GAME THEORY AND DEEP LEARNING, `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 algernatives to backpropagation; 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].

3 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 UofT, we borrowed robust tools from causal inference to provide robust methodology for the large-scale empirical study of generalization performance in neaural 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

[1] Kartik Ahuja, Ethan Caballero, Dinghuai Zhang, Yoshua Bengio, Ioannis Mitliagkas, and Irina Rish. Invariance principle meets information bottleneck for out-of-distribution generalization. *arXiv preprint arXiv:2106.06607*, 2021.

[2] Isabela Albuquerque, João Monteiro, Mohammad Darvishi, Tiago H Falk, and Ioannis Mitliagkas. Generalizing to unseen domains via distribution matching. *arXiv preprint arXiv:1911.00804*, 2019.

[3] Isabela Albuquerque, João Monteiro, Tiago H Falk, and Ioannis Mitliagkas. Adversarial target-invariant representation learning for domain generalization. *arXiv preprint arXiv:1911.00804*, 2019.

[4] Sébastien MR Arnold, Pierre-Antoine Manzagol, Reza Babanezhad, Ioannis Mitliagkas, and Nicolas Le Roux. Reducing the variance in online optimization by transporting past gradients. *arXiv preprint arXiv:1906.03532*, 2019.

[5] Devansh Arpit, Stanislaw Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S. Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, and Simon Lacoste-Julien. A closer look at memorization in deep networks. In *International Conference on Machine Learning (ICML)*, 2017.

[6] Devansh Arpit, Bhargav Kanuparthi, Giancarlo Kerg, Nan Rosemary Ke, Ioannis Mitliagkas, and Yoshua Bengio. h-detach: Modifying the lstm gradient towards better optimization. *arXiv preprint arXiv:1810.03023*, 2018.

[7] Waïss Azizian, Ioannis Mitliagkas, Simon Lacoste-Julien, and Gauthier Gidel. A tight and unified analysis of gradient-based methods for a whole spectrum of differentiable games. In *International Conference on Artificial Intelligence and Statistics*, pages 2863–2873. PMLR, 2020.

[8] Waïss Azizian, Damien Scieur, Ioannis Mitliagkas, Simon Lacoste-Julien, and Gauthier Gidel. Accelerating smooth games by manipulating spectral shapes. In *International Conference on Artificial Intelligence and Statistics*, pages 1705–1715. PMLR, 2020.

[9] David Balduzzi, Sebastien Racaniere, James Martens, Jakob Foerster, Karl Tuyls, and Thore Graepel. The mechanics of n-player differentiable games. In *International Conference on Machine Learning (ICML)*, 2018.

[10] Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *Siam Review*, 60(2):223–311, 2018.

[11] Tatjana Chavdarova, Gauthier Gidel, François Fleuret, and Simon Lacoste-Julien. Reducing noise in GAN training with variance reduced extragradient. In *Advances in Neural Information Processing Systems*, 2019.

[12] Gintare Karolina Dziugaite, Alexandre Drouin, Brady Neal, Nitarshan Rajkumar, Ethan Caballero, Linbo Wang, Ioannis Mitliagkas, and Daniel M Roy. In search of robust measures of generalization. *arXiv preprint arXiv:2010.11924*, 2020.

[13] Gauthier Gidel, Reyhane Askari Hemmat, Mohammad Pezeshki, Rémi Le Priol, Gabriel Huang, Simon Lacoste-Julien, and Ioannis Mitliagkas. Negative momentum for improved game dynamics. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2019.

[14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative

adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014.

[15] Charles Guille-Escuret, Manuela Girotti, Baptiste Goujaud, and Ioannis Mitliagkas. A study of condition numbers for first-order optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 1261–1269. PMLR, 2021.

[16] Suriya Gunasekar, Jason D Lee, Daniel Soudry, and Nati Srebro. Implicit bias of gradient descent on linear convolutional networks. In *Advances in Neural Information Processing Systems*, 2018.

[17] Adam Ibrahim, Waıss Azizian, Gauthier Gidel, and Ioannis Mitliagkas. Linear lower bounds and conditioning of differentiable games. In *International Conference on Machine Learning*, pages 4583–4593. PMLR, 2020.

[18] Timothy P Lillicrap, Daniel Cownden, Douglas B Tweed, and Colin J Akerman. Random synaptic feedback weights support error backpropagation for deep learning. *Nature communications*, 7(1):1–10, 2016.

[19] Nicolas Loizou, Hugo Berard, Alexia Jolicoeur-Martineau, Pascal Vincent, Simon Lacoste-Julien, and Ioannis Mitliagkas. Stochastic hamiltonian gradient methods for smooth games. In *International Conference on Machine Learning*, pages 6370–6381. PMLR, 2020.

[20] Brady Neal, Sarthak Mittal, Aristide Baratin, Vinayak Tantia, Matthew Scicluna, Simon Lacoste-Julien, and Ioannis Mitliagkas. A modern take on the bias-variance tradeoff in neural networks. *arXiv preprint arXiv:1810.08591*, 2018.

[21] Behnam Neyshabur. *Implicit Regularization in Deep Learning*. PhD thesis, TTIC, 2017.

[22] David Pfau and Oriol Vinyals. Connecting generative adversarial networks and actor-critic methods. *arXiv preprint arXiv:1610.01945*, 2016.

[23] Qian Qian and Xiaoyuan Qian. The implicit bias of AdaGrad on separable data. In *Advances in Neural Information Processing Systems*, 2019.

[24] Navid Azizan Ruhi, Sahin Lale, and Babak Hassibi. Stochastic mirror descent on overparameterized nonlinear models: Convergence, implicit regularization, and generalization. *arXiv preprint arXiv:1906.03830*, 2019.

[25] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations (ICLR)*, 2017.

[26] Jian Zhang and Ioannis Mitliagkas. Yellowfin and the art of momentum tuning. *arXiv preprint arXiv:1706.03471*, 2017.