Ioannis Mitliagkas — Research Statement — August 2018

Long-term vision: Making machine learning and modern artificial intelligence tools more accessible to small organizations and the public. This vision requires a wider education agenda, but also technical innovation on the optimization and tuning of machine learning components as well as big systems.

Modern data-driven applications rely on powerful computational infrastructure and skilled machine learning experts, to deliver analytical insights. Big organizations like web companies, retailers and government have the resources to attract experts and maintain powerful infrastructure. On the other hand, social groups, NGOs and academic research units lack funding and struggle to attract the people and maintain infrastructure. To make machine learning accessible, it is necessary that we bridge the gap between tools, human capital and infrastructure. Work towards this ambitious vision can be broken into three major thrusts:

1. **Optimize** popular tools to make them resource-efficient. This includes work on faster algorithms and systems that perform well with less computation and less data, lowering the cost of entry for small groups. A big part of my research effort goes into understanding the fundamental limits of popular techniques and coming up with resource-light alternative algorithms and systems.
2. **Automate** the tuning and assembly of machine learning pipelines. Modern machine learning is often compared to alchemy: practitioners report impressive results, achieved through painstaking trial-and-error selection of models and hyperparameters. The goal is for modern machine learning to become a science: we ought to develop the systematic methodology and theory that guide these practices with a guaranteed result. Tuning complex systems requires understanding their *computational and statistical behavior* and providing *data-dependent guarantees*—a focus of my research.
3. **Educate** a new generation of experts to develop these machine learning tools and a wider base of non-experts capable of using them. To that end, we use generous funding from public organizations to train more people in research. We are also developing a new professional masters program at Mila/UdeM, as well as one-year graduate diplomas in ML.

Research overview

I am interested in the theory, algorithms and systems in the intersection of statistics and computation. I aim to understand the theory and practical desiderata, and take a high-level look at entire ML pipelines, improving the various components to make a better overall system. My research touches upon themes of *optimization, sampling, analytical guarantees and tuning*, all necessary for bringing this vision about. It spans high-dimensional statistics, machine learning, optimization, statistical inference and large-scale distributed systems.

New theoretical insights on classic tools. Methods like principal component analysis [16], PageRank [14], stochastic gradient descent (SGD) [10, 9] and Gibbs sampling [11, 7] are ubiquitous, because they have proved their merit time and again. My work considers classic, massively deployed tools from a new perspective. In some cases it provides new, data-dependent guarantees [15, 11, 16], introduces new ways to prove the fundamental limits of problems via lower bounds [19, 18], or reveals previously unknown interactions between algorithms and hardware [10]. Other times it introduces modifications to extend use cases, improve performance and reduce the resource footprint [14, 16]. Working on classic tools means research with a high potential for impact.

Implementation and Systems. Theoretical and algorithmic breakthroughs have a bigger impact when translated into a real system. I find that working on a system prototype is useful as a proof of concept, as a source of new technical challenges and research questions, and as an experimentation platform. In Omnivore [12], we used my theoretical understanding of asynchrony [10] in our prototype deep learning system that was an order of magnitude faster than state-of-the-art systems. Following up on that work, we implemented and deployed a deep learning system on 10,000 machines [8]. In [14], I modified GraphLab to improve its PageRank performance. In [15], I implemented and deployed our graph algorithm on hundreds of AWS nodes using MapReduce.

A relationship with industry and research labs is a valuable source of technical challenges and collaborations. I maintain active collaborations with colleagues at Google Brain, Microsoft Research, and SAIL and ElementAI in Montréal. My asynchrony work [10] resulted into a joint project with Intel and NERSC and a publication at Supercomputing 2017 [8]. My work on momentum-based adaptive optimizers for deep learning has been picked up by industrial labs [6]. My past work on PageRank for large graphs [14] was motivated by interactions with Teradata and booking.com. I have given numerous invited talks at industrial labs, a process that I consider to be important for effective technology transfer.

Recent work highlights

Asynchrony induces momentum [10, 12, 8, 6]. In the seminal theory paper for this line of work [10], I showed that running SGD asynchronously can be viewed as adding a momentum-like term to the SGD iteration. An important implication is that tuning the momentum parameter is necessary when deploying large asynchronous systems, and that some systems in use by big labs are tuned suboptimally. In cases of heavy asynchrony, using a *negative algorithmic momentum* value can improve performance. Our results, tested in the form of a prototype system [12], gained attention from industry and national labs. We used our novel theoretical understanding of the interaction between system and optimization dynamics from to provide an efficient hyperparameter optimizer and achieve performance 1.9× to 12× better than the fastest state-of-the-art systems. Expanding on this work, and in collaboration with Intel and NERSC at Lawrence Berkeley Labs, we implemented a very large scale deep learning system, consisting of almost 10,000 machines, and used it for science applications [8].

Optimal scan order in Gibbs sampling [11, 7]. Gibbs sampling iteratively draws variables from their conditional distributions. There are two common scan orders: random and systematic scan. It had been conjectured that the mixing times of random scan and systematic scan do not differ by more than a logarithmic factor. We showed by counterexample that this is not the case, and proved that the mixing times do not differ by more than a polynomial factor under mild conditions [11]. To prove these relative bounds, I introduced a method of augmenting the state space to study systematic scan using conductance. In a different paper [7] we took a coupling-based analysis method by Dobrushin, and repurposed it to get model-dependent guarantees and customized scan sequences for Gibbs samplers that target specific variables.

Streaming PCA [16, 13]. Known phase transition results suggest that in the *noisy setting* the number of samples required to recover principal components is $O(dimension)$. This means that batch algorithms require $O(dimension^2)$ memory and storage and motivates a *memory-limited, single-pass streaming* algorithm. My work was the first to provide an algorithm along with global convergence guarantees and tight characterization of the sample complexity for the streaming PCA problem [16]. It is easily parallelizable and can handle an overwhelming number of sample entry erasures [13]. It has been implemented by Julia and R developers in the StreamingPCA and OnlinePCA packages, which have been downloaded over a thousand times.

Current and proposed work

In the first months of my work as an assistant professor, I started four new projects, on the optimization dynamics of adversarial objectives, on the generalization properties of neural networks, an exploration of Stein's method for sampling from complex distributions, and an application of deep learning on 3D data. These directions, along with other work I contributed to, have yielded 5 manuscripts under review (cf. CV) and 2 accepted papers [4, 5]. On these projects I supervised 6 students and interns and mentored 4 more. The proposed agenda follows.

Optimization

There are often overlooked insights hiding in popular methods, waiting to be discovered. Using exact tools, like families of orthogonal polynomials, to analyze system and optimization dynamics is a promising direction with great expository value. As a first example, we published a paper on accelerated PCA [4]. This kind of analysis also led us to a momentum-based adaptive optimizer of deep learning [6].

Robustness properties of Polyak's momentum. My recent work explored tuning rules for the heavy ball method, also known as Polyak's momentum [6]. An interesting side-effect was the discovery of a previously unreported property: when tuned optimally, Polyak's momentum *equalizes the rates of convergence* along all directions (equivalently, all optimization variables). For example, on a quadratic problem the dynamics of multi-variable optimization decompose into separable scalar dynamics along the eigenvectors of the Hessian. When the momentum value is optimal (or higher), those independent scalar dynamics follow the exact same rate of convergence. That rate depends solely on the value of momentum, not on variations of curvature or step size. This property is not necessary for achieving acceleration and is unique to Polyak's momentum; Nesterov's accelerated gradient (NAG) does not possess this property. I plan to explain the implications of this property on convex and non-convex settings, focusing on the dynamics, ability to escape saddle points and generalization performance.

Accelerated stochastic optimization. NAG and the heavy ball provide accelerated convergence for convex problems in the full batch setting. Tail-averaging yields an optimal asymptotic rate in the online stochastic

setting. Until recently, there was no method that provably achieves both: an accelerated rate during early phases of stochastic optimization and an optimal asymptotic rate in the later phases. Work by Jain et al. proposes such an accelerated stochastic method for least squares regression. That method requires the careful tuning of 3 hyperparameters, and uses an intricate sequence of 3 convex combinations of different iterates to achieve this result. The method betrays no insight about its inner workings or the problem at hand. It is for these reasons desirable to design an accelerated stochastic optimization method that is based on simpler, already understood, components and would depend of fewer hyperparameters. The proposed work relies on a gradient estimator that is unbiased (like in SGD), but can achieve the optimal asymptotic variance decay for specific classes of problems (for example, quadratics) by optimally adjusting step sizes over time. This kind of estimator, in combination with an outer loop of acceleration, based on NAG or the heavy ball method, will be the main object of study.

Dynamics of adversarial optimization. Following up on the theme of studying optimization dynamics in special settings [10], I plan to study the dynamics of differentiable games. One prime example in ML are Generative Adversarial Networks (GANs). The ML community has been using tools meant for optimization, like gradient descent. However, adversarial objectives demonstrate very special dynamics that slow down convergence; these dynamics are rotational, and are reminiscent of momentum dynamics. In a recent NIPS submission we explore this idea and prove that using negative momentum can help with certain GANs [3]. I plan to explore this direction deeper, especially through the study of operators carefully designed to minimize rotational dynamics.

Inference

Targeted, optimized inference. The main hyperparameter in Gibbs sampling, is the *scan*: the order in which each variable is sampled. In [11] we studied the relative efficacy of *systematic*, and *uniform random scans*. Non-uniform scans, however, can lead to more accurate inferences both in theory and in practice. This effect is particularly pronounced when certain variables are of greater inferential interest. A first step in optimizing the scan order is our recent paper that uses couplings to optimize the scan sequence, applicable on fast-mixing distributions [7]. I plan to explore other approaches for optimized, targeted scan sequences for general distributions.

Stein's method and efficient sampling. A recent line of work in the statistics and machine learning literature has been using Stein's method to facilitate sampling from complex distributions; e.g. when the normalization constant is not known, but we still have access to the score function. Stein variational gradient descent uses a set of particles to approximate the target distribution, and poses the problem as optimization. Other work applies the Stein operator in reproducing kernel Hilbert spaces allowing for even simpler sampling. Most of this literature is still missing some important algebraic and geometric insights about these operators and spaces. Along with a mathematically inclined student I work with, I plan to explore some of these insights and potential methods that come out of a geometric interpretation of the Stein operator, starting with a projection-based method for fitting the set of particles to the target distribution, and the idea of using negatively weighted particles.

Deep learning

Neural networks and generalization. In a recently started project with one of my students we are studying the generalization properties of wide and deep neural networks under the lens of a bias-variance decomposition. The textbook bias-variance decomposition picture suggests a trade-off: complex models can achieve low bias, but they suffer because of high variance in the prediction. The latter implies bad generalization. Preliminary results, however, suggest that deep networks of increasing width also exhibit decreasing variance. This is a new approach to explaining the good generalization of over-parametrized neural networks. In the same thread, I plan to start a research project to study data- an distribution-dependent generalization bounds.

Robustness to adversarial attacks. In the past year I have co-authored two manuscripts on defenses for adversarial attacks ([2, 1] under review). The question of adversarial robustness is interesting and critical for the use of deep learning on mission-critical applications. I plan to pursue theoretical work in the area.

The University of Montréal and Mila have offered a productive and hospitable home for my research. I am always looking for strong students to push the boundary on all of these exciting research directions.

References

[1] V. Verma, A. Lamb, C. Beckham, A. Courville, I. Mitliagkas, and Y. Bengio. Manifold mixup: Encouraging meaningful on-manifold interpolation as a regularizer. *arXiv preprint arXiv:1806.05236*, 2018.

[2] A. Lamb, J. Binas, A. Goyal, D. Serdyuk, S. Subramanian, I. Mitliagkas, and Y. Bengio. Fortified networks: Improving the robustness of deep networks by modeling the manifold of hidden representations. *arXiv preprint arXiv:1804.02485*, 2018.

[3] G. Gidel, R. A. Hemmat, M. Pezeshki, G. Huang, R. Lepriol, S. Lacoste-Julien, and I. Mitliagkas. Negative momentum for improved game dynamics. *arXiv preprint arXiv:1807.04740*, 2018.

[4] C. De Sa, B. He, I. Mitliagkas, C. Ré, and P. Xu. Accelerated stochastic power iteration. *AISTATS*, 2018.

[5] P. Achlioptas, O. Diamanti, I. Mitliagkas, and L. Guibas. Learning representations and generative models for 3D point clouds. *ICML*, 2018.

[6] J. Zhang, I. Mitliagkas, and C. Ré. Yellowfin and the art of momentum tuning. *arXiv preprint arXiv:1706.03471*, 2017.

[7] I. Mitliagkas and L. Mackey. Improving Gibbs sampler scan quality with DoGS. In *International Conference on Machine Learning*, pages 2469–2477, 2017.

[8] T. Kurth, J. Zhang, N. Satish, E. Racah, I. Mitliagkas, M. M. A. Patwary, T. Malas, N. Sundaram, W. Bhimji, M. Smorkalov, et al. Deep learning at 15PF: supervised and semi-supervised classification for scientific data. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, page 7. ACM, 2017.

[9] J. Zhang, C. De Sa, I. Mitliagkas, and C. Ré. Parallel SGD: When does averaging help? *OptML, Workshop at NIPS 2016*, 2016.

[10] I. Mitliagkas, C. Zhang, S. Hadjis, and C. Ré. Asynchrony begets momentum, with an application to deep learning. *Proc. of Allerton Conf. on Communication, Control and Computing, Monticello, USA*, 2016.

[11] B. He, C. De Sa, I. Mitliagkas, and C. Ré. Scan order in gibbs sampling: Models in which it matters and bounds on how much. *NIPS*, 2016.

[12] S. Hadjis, C. Zhang, I. Mitliagkas, and C. Ré. Omnivore: An optimizer for multi-device deep learning on cpus and gpus. *Under Review*, 2016.

[13] I. Mitliagkas, C. Caramanis, and P. Jain. Streaming PCA with Many Missing Entries. *Preprint*, 2015.

[14] I. Mitliagkas, M. Borokhovich, A. Dimakis, and C. Caramanis. FrogWild! fast pagerank approximations on graph engines. *VLDB*, 2015.

[15] D. Papailiopoulos, I. Mitliagkas, A. Dimakis, and C. Caramanis. Finding dense subgraphs via low-rank bilinear optimization. In *ICML 2014*, pages 1890–1898, 2014.

[16] I. Mitliagkas, C. Caramanis, and P. Jain. Memory limited, streaming PCA. In *NIPS*, pages 2886–2894, 2013.

[17] I. Mitliagkas, N. D. Sidiropoulos, and A. Swami. Joint power and admission control for ad-hoc and cognitive underlay networks: Convex approximation and distributed implementation. *IEEE Transactions on Wireless Communications*, 2011.

[18] I. Mitliagkas, A. Gopalan, C. Caramanis, and S. Vishwanath. User rankings from comparisons: Learning permutations in high dimensions. In *Proc. of Allerton Conf. on Communication, Control and Computing, Monticello, USA*, 2011.

[19] I. Mitliagkas and S. Vishwanath. Strong information-theoretic limits for source/model recovery. In *Proc. of Allerton Conf. on Communication, Control and Computing, Monticello, USA*, 2010.

[20] I. Mitliagkas, N. D. Sidiropoulos, and A. Swami. Distributed joint power and admission control for ad-hoc and cognitive underlay networks. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 3014–3017. IEEE, 2010.

[21] I. Mitliagkas, N. D. Sidiropoulos, and A. Swami. Convex approximation-based joint power and admission control for cognitive underlay networks. In *Wireless Communications and Mobile Computing Conference, 2008. IWCMC'08. International*, pages 28–32. IEEE, 2008.