2022-2023 Grand Challenge Award Final Report

Awardee: **Rachel Ward, Professor**
Mathematics

Research Award Title: **Robust and Accelerated Randomized Algorithms for Scientific Machine Learning**

Research Summary

The optimization procedure used to train neural networks is often stated concisely as "stochastic gradient descent, applied to non-convex neural network functions". While stochastic gradient descent (SGD) in its plainest form is well understood theoretically on smooth and convex functions, its strong behavior on non-convex neural network functions is still a mystery. Moreover, the variant of SGD used in practice and key to convergence behavior fast enough to scale is stochastic gradient descent with *heavy ball momentum (HBM)*. Heavy ball momentum maintains a "look back" auxiliary variable that tracks the previous stochastic gradient directions in such a way that the algorithm can converge faster empirically in general and provably when in the case of non-stochastic gradient descent.

During this Grand Challenge, collaborators and I made significant progress towards understanding

1. the fast convergence behavior of SGD on non-convex matrix factorization problems and linear neural networks, and
2. how SGD with heavy ball momentum can maintain the fast coverage speed of gradient descent with heavy ball momentum, provided minibatch size is above a critical threshold.

Understanding SGD with momentum

Non-stochastic gradient descent with heavy ball momentum is classically shown to attain the optimal linear convergence rate proportional to the square root of the condition number on quadratic objective functions. However, going from deterministic to stochastic gradient descent, the classical proof breaks down due to its delicate dependency on eigenvalues of the non-symmetric momentum update matrix. In fact, previous papers have provided counterexamples showing that heavy ball momentum does not lead to acceleration at all in SGD when very small minibatches are used to form the stochastic gradients.

During this award period, with Raghu Bollapragada (ORIE, UT Austin) and Tyler Chen (Courant Instructor, NYU) we derived the first provable guarantee that Stochastic Gradient Descent with minibatch heavy ball momentum provably attains gradient descent-level acceleration, provided the stochastic minibatch size is above a precise critical size [1]. This resolves the seemingly contradictory previous negative theoretical results and emprical success, as large minibatch sizes are used in practice. An illustration summarizing the context of our contribution can be found in Figure 1. *Our paper has been accepted in IMA Journal of Numerical Analysis and should appear shortly, as we have just submitted final proofs.

*I have been trying to prove (or disprove) such a result for over a decade, this result was made possible using recent concentration inequalities for products of random (not necessarily symmetric) matrices (co-developed by myself, Joel Tropp, Jon Niles-Weed, and De Huang).

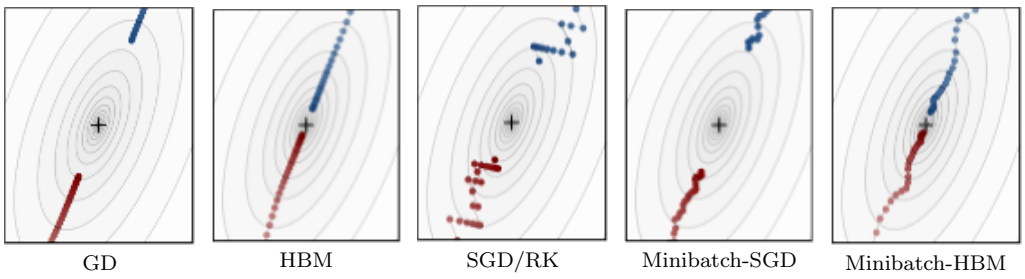

Figure 1: Sample convergence trajectories for various iterative methods applied to a quadratic objective with $n = 200$ and $d = 2$. **From left to right**: Gradient Descent with heavy ball momentum (HBM) allows for accelerated convergence over plain gradient descent (GD), and the proof is classical. Stochastic gradient descent (SDR/RK) is a lower per iteration cost variant of GD which is practical in large-scale machine learning settings, and the use of minibatching (Minibatch-SGD) reduces the variance of the iterates. In practice, minibatching and momentum are often used simultaneously for fast convergence (Minibatch-HBM); we provided the first theoretical guarantees on quadratic objectives, provided the minibatch size is larger than a critical threshold.

Understanding SGD convergence for Matrix Factorization

A second key research outcome from the Grand Challenge Award period is a joint paper with Tamara Kolda who was a distinguished visiting scholar at the Oden Institute in Spring 2023 during the award period. Our paper [2] provides a first sharp convergence analysis of (alternating) Gradient Descent on non-convex Matrix Factorization problems, which represent simple prototypes for non-convex neural network functions in the overparameterized regime. Our paper proves that by initializing the factor matrices in gradient descent according to a particular asymmetric random initialization, gradient descent converges linearly on these objective functions – as if the objective function were strongly convex – because the gradient iterations can be proven to concentrate on a strongly convex slice of the non-convex landscape. Our results hint at a general algorithmic advantage to asymmetric initialization in gradient descent, beyond matrix factorization: initializing the matrix weights in gradient descent in an asymmetric fashion, where the entries of one matrix are divided by the step-size and the entries in the other matrix are multiplied by the step-size, is crucial to inducing a faster convergence rate by "making sure" that the optimization trajectory stays in a strongly convex slice of the non-convex landscape. Our paper was published in Neurips 2023 [2]. A second key finding is understanding in a precise quantitative sense how overparameterizing the optimization problem by increasing the matrix factor sizes directly leads to a faster convergence rate.

Fast convergence of gradient descent with momentum on linear neural networks

During summer 2023, along with Molei Tao (Georgia Tech) and colleagues, we generalized and combined the previous two research works. We extended the analysis of Gradient Descent with momentum to a class of non-convex matric factorization problems which includes linear neural networks. This work which is currently under review at Neurips 2024 [3], represents *the first analysis of gradient descent with momentum showing an optimal convergence rate the optimal quantitative convergence rate for a general class of neural network functions.* Beyond the theoretical proof, the result corroborates the improvement offered by overparameterization, and suggestes that asymmetric balanced random initialization can benefit neural network training more broadly. This work is a major step towards understanding the mathematical foundations of large-scale optimization mechanisms such as momentum, and why these mechanisms unlock fast convergence on non-convex neural network objectives.

Talks and Honors

I was an invited plenary speaker at ICIAM in August 2023 in Tokyo, Japan, where I presented on this work. I was an invited speaker at the International Congress of Mathematicians (ICM) in 2022 which only occurs once every 4 years. It was a great honor to receive invitations to speak at both the premier international conference in pure math and the premier international conference in pure and applied/industrial mathematics, just one year apart.
During the 2022-2023 year, I received numerous invitations to give talks at universities and conferences. I declined most invitations to spend time with my kids during the fleeting few years they are young and still happy to be around me.

References

1. Raghu Bollapragade, Tyler Chen, and Rachel Ward. "On the fast convergence of minibatch heavy ball momentum". In: *To appear, IMA Journal of Numerical Analysis* (2024).

2. Rachel Ward and Tamara Kolda. "Convergenec of alternating gradient descent for matrix factorization". In: *Advances in Neural Information Processing Systems 36* (2023), pp 22369 - 22382.

3. Zhenghao Xu et al. "Provable Acceleration of Nesterov's Accelerated Gradient for Asymmetric Matrix Factorization and Linear Neural Networks". In: *In submission* (2024).