Making the Last Iterate of SGD Information Theoretically Optimal

Prateek Jain PRAJAIN@MICROSOFT.COM
Microsoft Research, Bengaluru, Karnataka, India
Dheeraj Nagaraj DHEERAJ@MIT.EDU
Massachusetts Institute of Technology, Cambridge, Massachusetts, USA - 02139
Praneeth Netrapalli PRANEETH@MICROSOFT.COM
Microsoft Research, Bengaluru, Karnataka, India

Editors: Alina Beygelzimer and Daniel Hsu

Abstract

Stochastic gradient descent (SGD) is one of the most widely used algorithms for large scale optimization problems. While classical theoretical analysis of SGD for convex problems studies (suffix) *averages* of iterates and obtains information theoretically optimal bounds on suboptimality, the *last point* of SGD is, by far, the most preferred choice in practice. The best known results for last point of SGD (Shamir and Zhang, 2013) however, are suboptimal compared to information theoretic lower bounds by a $\log T$ factor, where T is the number of iterations. Harvey et al. (2018) shows that in fact, this additional $\log T$ factor is tight for standard step size sequences of $\Theta\left(\frac{1}{\sqrt{t}}\right)$ and $\Theta\left(\frac{1}{t}\right)$ for non-strongly convex and strongly convex settings, respectively. Similarly, even for subgradient descent (GD) when applied to non-smooth, convex functions, the best known step-size sequences still lead to $O(\log T)$-suboptimal convergence rates (on the final iterate). The main contribution of this work is to design new step size sequences that enjoy information theoretically optimal bounds on the suboptimality of *last point* of SGD as well as GD. We achieve this by designing a modification scheme, that converts one sequence of step sizes to another so that the last point of SGD/GD with modified sequence has the same suboptimality guarantees as the average of SGD/GD with original sequence. We also show that our result holds with high-probability. We validate our results through simulations which demonstrate that the new step size sequence indeed improves the final iterate significantly compared to the standard step size sequences.

Keywords: Stochastic Gradient Descent, Machine Learning, Convex Optimization.

1. Introduction

Stochastic Gradient Descent (SGD) is one of the most popular algorithms for solving large-scale empirical risk minimization (ERM) problems LeCun et al. (2015); Shalev-Shwartz et al. (2011); Akiba et al. (2017). The algorithm updates the iterates using stochastic gradients obtained by sampling data points uniformly at random. The algorithm has been studied for several decades Bubeck (2015) but there are still significant gaps between *practical implementations* and theoretical analyses. In particular, the standard analyses hold only for some kind of average of iterates, but

. Extended abstract. Full version appears as [arXiv:1904.12443, v2]

most practitioners just use the final iterate of SGD. So, Shamir (2012) asked the natural question of whether the final iterate of SGD, as opposed to average of iterates, is provably good. It was partly answered in Shamir and Zhang (2013) which gave sub-optimality bound for the last point of SGD but the obtained sub-optimality rates are $O(\log T)$ worse than the information theoretically optimal rates; T is the number of iterations.

Harvey et al. (2018) showed that the above result is tight for the standard step-size sequence used by most existing theoretical results. The extra logarithmic factor is not due to the stochastic nature of SGD. In fact, even for *subgradient descent* (GD) when applied to general non-smooth, convex functions, the last point's convergence rates are sub-optimal by $O(\log T)$ factor.

So, this work addresses the following two fundamental questions:

"Does there exist a step-size sequence for which the last point of SGD when applied to general convex functions as well as to strongly-convex functions has optimal error (sub-optimality) rate?", and,

"Does there exist a step-size sequence for which the last point of GD when applied to general non-smooth convex functions has optimal error (sub-optimality) rate?"

In this paper, we answer both the questions in the affirmative. That is, we provide novel step size sequences and show that the final iterate of SGD run with these step size sequences has the information theoretically optimal error (suboptimality) rate. In particular, for general non-smooth convex functions, our results ensure an error rate of $O(\frac{1}{\sqrt{T}})$ and for strongly-convex functions, the error rate is $O(\frac{1}{T})$. We also present high-probablity versions, i.e., we show that with probability at least $1-\delta$, the suboptimality is $O\left(\sqrt{\frac{\log\frac{1}{\delta}}{T}}\right)$ and $O\left(\frac{\log\frac{1}{\delta}}{T}\right)$ respectively. For GD, we show that a similarly modified step-size sequence leads to suboptimality of $O(\frac{1}{T})$ and $O(\frac{1}{\sqrt{T}})$ for non-smooth convex functions, with and with out strong convexity respectively, which is optimal.

In general, SGD takes the iterates near the optimum value but since the objective isn't smooth near the optimizer x^*, the gradients don't become small even when the points are close to x^*. Standard step sizes don't decay appreciably with time to ensure fast enough convergence to x^*. Therefore the iterates x_t, after going close to x^*, start oscillating around it without actually approaching it. Our new step sizes, ensure that the step sizes decay fast enough after a certain point, making the iterates go closer to the optimum x^*. The exact mode of this decay ensures that the last iterate approaches the optimum at the information theoretic rate.

Our step-size sequence requires that the number of iterations or horizon T is known apriori. In contrast, standard step-size sequences do not require T apriori, and hence guarantee any-time results. Information about T apriori helps us in ensuring that we do not drop step-size too early; only after we are close to the optimum, does the step size drop rapidly. In fact, we conjecture that in absence of apriori information about T, *no step-size* sequence can ensure the information theoretically optimal error rates for final iterate of SGD.

Our results utilize a general step size modification scheme which ensures that the upper bounds for the average function value with the original step sizes gets transferred to the last iterate when the modified step sizes are used. A key technical contribution of the paper is the construction of a sequence of averaging schemes which are 'good' with high probability such that the last averaging scheme consists only of the last iterate and hence lets us conclude that the last iterate is 'good' with high probability.

Related Work: Averaging was used first in the stochastic approximation setting by Polyak and Juditsky (1992) to show optimal rates of convergence. Gradient Descent type methods have been

shown to achieve information theoretically optimal error rates in the convex and strongly convex settings when averaging of iterates is used (Nemirovsky and Yudin (1983),Zinkevich (2003),Cesa-Bianchi et al. (2004), Kakade and Tewari (2009), Epoch GD in Hazan and Kale (2014) , SGD Rakhlin et al. (2012) and Lacoste-Julien et al. (2012)). The question of the last iterate was first considered in Shamir and Zhang (2013) and it gives a bound of $O(\frac{\log T}{\sqrt{T}})$ and $O(\frac{\log T}{T})$ in expectation for the general case and strongly convex case respectively. Harvey et al. (2018) show matching high probability bounds and show that for the standard step sizes ($O\left(\frac{1}{\sqrt{t}}\right)$ in the general case and $O\left(\frac{1}{t}\right)$ in the strongly convex case), the logarithmic-suboptimal bounds are tight.

2. Conclusions and Discussion

We studied the fundamental question of sub-optimality of the last point of SGD/GD for general non-smooth convex functions as well as for strongly-convex functions. We proposed a novel step-size sequence that leads to information theoretically optimal rates in both the above mentioned settings. Our result proves a more general result for any "modified step-size" of a decaying standard step-size, and uses a novel technique of tracking best iterate in each time-interval and ensuring that the later iterates do not significantly deviate from the best iterate in the previous time interval. We also provide a high-probability bound using a super-martingale technique from Harvey et al. (2018). Simulations show that our step-size indeed leads to better last point than the standard step-size sequences.

Our approach fundamentally exploits an assumption that we apriori know the total number of iterations T. Hence, our result does not provide an any-time algorithm. In contrast, existing any-time results have an extra $\log T$ multiplicative factor in the sub-optimality. We conjecture that this gap is fundamental and every *any-time* algorithm would suffer from the extra $\log T$ factor.

Acknowledgments

This research was partially supported by ONR N00014-17-1-2147 and MIT-IBM Watson AI Lab.

References

Takuya Akiba, Shuji Suzuki, and Keisuke Fukuda. Extremely large minibatch sgd: Training resnet-50 on imagenet in 15 minutes. *arXiv preprint arXiv:1711.04325*, 2017.

Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

Nicolo Cesa-Bianchi, Alex Conconi, and Claudio Gentile. On the generalization ability of on-line learning algorithms. *IEEE Transactions on Information Theory*, 50(9):2050–2057, 2004.

Nicholas JA Harvey, Christopher Liaw, Yaniv Plan, and Sikander Randhawa. Tight analyses for non-smooth stochastic gradient descent. *arXiv preprint arXiv:1812.05217*, 2018.

Elad Hazan and Satyen Kale. Beyond the regret minimization barrier: optimal algorithms for stochastic strongly-convex optimization. *The Journal of Machine Learning Research*, 15(1): 2489–2512, 2014.

Sham M Kakade and Ambuj Tewari. On the generalization ability of online strongly convex programming algorithms. In *Advances in Neural Information Processing Systems*, pages 801–808, 2009.

Simon Lacoste-Julien, Mark Schmidt, and Francis Bach. A simpler approach to obtaining an o (1/t) convergence rate for the projected stochastic subgradient method. *arXiv preprint arXiv:1212.2002*, 2012.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436, 2015.

Arkadii Semenovich Nemirovsky and David Borisovich Yudin. Problem complexity and method efficiency in optimization. 1983.

Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.

Alexander Rakhlin, Ohad Shamir, Karthik Sridharan, et al. Making gradient descent optimal for strongly convex stochastic optimization. In *ICML*, volume 12, pages 1571–1578. Citeseer, 2012.

Shai Shalev-Shwartz, Yoram Singer, Nathan Srebro, and Andrew Cotter. Pegasos: Primal estimated sub-gradient solver for svm. *Mathematical programming*, 127(1):3–30, 2011.

Ohad Shamir. Open problem: Is averaging needed for strongly convex stochastic gradient descent? In *Conference on Learning Theory*, pages 47–1, 2012.

Ohad Shamir and Tong Zhang. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In *International Conference on Machine Learning*, pages 71–79, 2013.

Tong Zhang. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *Proceedings of the twenty-first international conference on Machine learning*, page 116. ACM, 2004.

Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pages 928–936, 2003.