Open Problem: Do Good Algorithms Necessarily Query Bad Points?

Rong Ge RONGGE@CS.DUKE.EDU
Duke University, Durham NC, USA

Prateek Jain PRAJAIN@MICROSOFT.COM
Microsoft Research India

Sham M. Kakade SHAM@CS.WASHINGTON.EDU
University of Washington, Seattle WA, USA

Rahul Kidambi RKIDAMBI@UW.EDU
University of Washington, Seattle WA, USA

Dheeraj M. Nagaraj DHEERAJ@MIT.EDU
MIT, Boston MA, USA

Praneeth Netrapalli PRANEETH@MICROSOFT.COM
Microsoft Research, Bengaluru KA, India

Editors: Alina Beygelzimer and Daniel Hsu

Abstract

Folklore results in the theory of Stochastic Approximation indicates the (minimax) optimality of Stochastic Gradient Descent (SGD) (Robbins and Monro, 1951) with polynomially decaying step sizes and iterate averaging (Ruppert, 1988; Polyak and Juditsky, 1992) for classes of stochastic convex optimization. Basing of these folkore results and some recent developments, this manuscript considers a more subtle question: does *any* algorithm necessarily (information theoretically) have to query iterates that are sub-optimal infinitely often?

Keywords: Stochastic Approximation, Stochastic Gradient Descent, Iterate Averaging, Minimax Optimality

1. Introduction and Problem Setup

Stochastic Approximation studies the minimization of an objective function written as an expectation, and one that admits a sample based access, i.e.,:

$$w^* \in \arg\min_w F(w) = \mathbb{E}_{\xi\sim\mathcal{D}}[f(w;\xi)] \tag{1}$$

A variety of questions in modern machine learning and, more generally, ones in various scientific disciplines can be written as an instance of (1). In this context, we assume access to a stochastic first order oracle (Nemirovsky and Yudin, 1983), which, when queried at any iterate w_t returns a stochastic gradient $\nabla f(w_t;\xi_t)$ that is unbiased, i.e.,

$$\mathbb{E}_{\xi_t}[\nabla f(w_t;\xi_t)|w_t,\mathcal{F}_{t-1}] = \nabla F(w_t),$$

where, $\mathcal{F}_{t-1}$ is the filtration formed by $\{\xi_j\}_{j=1}^{t-1}$. Equipped with access to the stochastic first order oracle, stochastic gradient descent (SGD) (Robbins and Monro, 1951) is a popular technique used

to minimize the objective of the form (1). SGD works with the following (simple) update rule:

$$w_{t+1} \leftarrow w_t - \eta_t \cdot \nabla f(w_t; \xi_t), \text{ where,}$$

$\eta_t > 0$ is the learning rate/step size. Iterate averaged SGD (Ruppert, 1988; Polyak and Juditsky, 1992) involves running SGD and returning

$$\bar{w} = \frac{1}{n} \sum_{t=1}^{n} w_t.$$

In the literature, the stochastic first order oracle (Nemirovsky and Yudin, 1983) has typically assumed to satisfy the following two types of assumptions on its second moment. The first one assumes a uniform bound on the variance at every point in the domain (Robbins and Monro, 1951; Nemirovsky and Yudin, 1983), i.e.

$$\mathbb{E}_{\xi \sim \mathcal{D}}[||\nabla f(w; \xi) - \nabla F(w)||^2] \leq \sigma^2. \tag{2}$$

The other assumption is with regards to the covariance of the stochastic gradient at w^* (Polyak and Juditsky, 1992; Bach and Moulines, 2013; Jain et al., 2016, 2017), i.e.,

$$\mathbb{E}_{\xi \sim \mathcal{D}}[\nabla f(w^*; \xi) \nabla f(w^*; \xi)^\top] \preceq \sigma^2 \nabla^2 F(w^*). \tag{3}$$

Distinctions between assumptions (2), (3) are discussed in Jain et al. (2017). Iterate averaged SGD, coupled with polynomially decaying step sizes, i.e., $\eta_t = 1/t^\alpha$, $\alpha \in (0.5, 1)$ attains minimax rates (Nemirovsky and Yudin, 1983; Lehmann and Casella, 1998; Van der Vaart, 2000; Raginsky and Rakhlin, 2011; Agarwal et al., 2012) in an *anytime* sense (i.e. for all large values of n), for classes of stochastic convex optimization (Ruppert, 1988; Polyak and Juditsky, 1992).

2. Behavior of SGD's Final Iterate

The anytime optimal behavior of iterate averaged SGD with polynomially decaying stepsizes presents an interesting research program that involves understanding the behavior of the query points for classes of SGD style algorithms. Before considering the strictly harder problem of achieving anytime optimal behavior (i.e. when running an algorithm *without* the knowledge of the algorithm end time T), we discuss the situation when the end time T is known and fixed in advance.

2.1. Behavior of SGD's final iterate with a fixed end time T

In the context of non-smooth optimization (with or without strong convexity), under assumption (2), a recent work of Jain et al. (2019) presented a nuanced step size schedule that achieves minimax rates when the time horizon T is fixed in advance.

In the context of smooth stochastic convex optimization (with or without strong convexity), the situation is far from being resolved. We will detail two lines of work that attempt to make progress on various issues arising in the smooth stochastic approximation setting:

For the special case of streaming least squares regression, working with assumption (3), the work of Ge et al. (2019) presented the step-decay schedule (which is a geometrically decaying stepsize procedure) which achieves near minimax rates upto a $\log \kappa$ factor (where, κ is the condition

number) in strongly convex case and $\log T$ factor in the smooth case, when the horizon T is known in advance. Surprisingly, Ge et al. (2019) show that there exists *no* polynomially decaying step size procedure, i.e., where, $\eta_t = 1/t^\alpha,\ \alpha \in [0.5, 1]$ that can achieve a rate that is separated from the minimax rate by a factor of $\Omega(\kappa)$ (in the strongly convex case) and $\Omega(\sqrt{T})$ in the smooth case. These lower bounds shown by Ge et al. (2019) present telling evidence towards the (un-)suitability of standard polynomially decaying stepsizes towards obtaining optimal behavior of the final iterate for general classes of smooth (or smooth and strongly convex) stochastic convex optimization.

Another interesting line of work in the known (finite) horizon case involves minimizing the gradient norm (as opposed to the function value). Along this line of thought, with assumption (2), the work of Allen-Zhu (2018) presents improvements akin to ones offered by the work of Ge et al. (2019) for general classes of smooth (and smooth plus strongly convex) stochastic optimization.

2.2. Anytime behavior of the final iterate

For non-smooth optimization (with/without strong convexity), working with assumption (2), the work of Shamir and Zhang (2012) obtains a final iterate achieves minimax rates (in an anytime sense) upto a $\log(T)$ factor. In a sense, a recent work of Harvey et al. (2018) indicates that any choice of stepsizes that are close to the standard stepsize schedules for this problem will yield iterates that query sub-optimal function values (that are off by multiplicative $\log(T)$ factors) infinitely often.

Switching contexts to the smooth and strongly convex stochastic optimization, the work of Ge et al. (2019) indicates SGD's final iterate *regardless* of the choice of its stepsize sequence must query iterates that are off the minimax rate by a factor of nearly the condition number of the problem.

To state our open problem, we define the notion of a "non-adaptive" algorithm as one that queries an iterate w_t which can be *any* fixed (potentially non-stationary) linear combination of previously queried stochastic gradients defined prior to the start of the algorithm, i.e.:

$$w_t = \sum_{j<t} \alpha_j^{(t)} \nabla f(w_j; \xi_j), \text{ where, } \alpha_j^{(t)} \in \mathbb{R}\ \forall\ j.$$

With this definition in place, our question is phrased as follows:
Consider a non-adaptive algorithm with access to a stochastic first order oracle. The question we ask is: Does this algorithm have to query sub-optimal iterates (compared to the minimax rates) infinitely often (i.e. in a $\limsup$ sense)?

The notion of sub-optimality follows owing to Harvey et al. (2018) (where the final iterate is at least a $\log(T)$ factor worse compared to the minimax rate) for non-smooth stochastic convex optimization, and that of Ge et al. (2019) (where the final iterate is nearly a condition number factor away from the minimax rate) for optimizing strongly convex least squares. Whilst Harvey et al. (2018); Ge et al. (2019) begin to make progress towards answering this question, they are fairly limited in that their query points are the final iterate of SGD (as opposed to dealing with *any* non-adaptive procedure). Special cases of the question includes:

- Can we bridge the $\log T$ factor indicated as necessary by the work of Harvey et al. (2018) for the non-smooth stochastic convex optimization (Shamir and Zhang, 2012; Jain et al., 2019)?
- Can we bridge the gap of a condition number for SGD's final iterate for least squares (Ge et al., 2019) under conditions satisfied by a non-adaptive algorithm?
- With regards to gradient norm (as opposed to function value), can we make progress towards understanding query point sub-optimality in various cases of stochastic convex optimization?

References

Alekh Agarwal, Peter L. Bartlett, Pradeep Ravikumar, and Martin J. Wainwright. Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization. *IEEE Transactions on Information Theory*, 2012.

Zeyuan Allen-Zhu. How to make the gradients small stochastically. *CoRR*, abs/1801.02982, 2018.

Francis R. Bach and Eric Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate O(1/n). In *NIPS 26*, 2013.

Rong Ge, Sham M. Kakade, Rahul Kidambi, and Praneeth Netrapalli. The step decay schedule: A near optimal, geometrically decaying learning rate procedure. *CoRR*, 2019. URL https://arxiv.org/abs/1904.12838.

Nicholas J. A. Harvey, Christopher Liaw, Yaniv Plan, and Sikander Randhawa. Tight analyses for non-smooth stochastic gradient descent. *CoRR*, 2018. URL http://arxiv.org/abs/1812.05217.

Prateek Jain, Sham M Kakade, Rahul Kidambi, Praneeth Netrapalli, and Aaron Sidford. Parallelizing stochastic approximation through mini-batching and tail-averaging. *arXiv preprint arXiv:1610.03774*, 2016.

Prateek Jain, Sham M Kakade, Rahul Kidambi, Praneeth Netrapalli, and Aaron Sidford. Accelerating stochastic gradient descent. *arXiv preprint arXiv:1704.08227*, 2017.

Prateek Jain, Dheeraj Nagaraj, and Praneeth Netrapalli. Making the last iterate of sgd information theoretically optimal. *CoRR*, 2019. URL https://arxiv.org/abs/1904.12443.

Erich L. Lehmann and George Casella. *Theory of Point Estimation*. Springer Texts in Statistics. Springer, 1998. ISBN 9780387985022.

Arkadi S. Nemirovsky and David B. Yudin. *Problem Complexity and Method Efficiency in Optimization*. John Wiley, 1983.

Boris T. Polyak and Anatoli B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, volume 30, 1992.

Maxim Raginsky and Alexander Rakhlin. Information-based complexity, feedback and dynamics in convex programming. *IEEE Transactions on Information Theory*, 2011.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, vol. 22, 1951.

David Ruppert. Efficient estimations from a slowly convergent robbins-monro process. *Tech. Report, ORIE, Cornell University*, 1988.

Ohad Shamir and Tong Zhang. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. *CoRR*, abs/1212.1824, 2012.

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.