Making Gradient Descent Optimal for Strongly Convex Stochastic Optimization
(Extended Abstract)

Ohad Shamir
Microsoft Research New England
ohadsh@microsoft.com

Abstract

Stochastic gradient descent (SGD) is a simple and popular method to solve stochastic optimization problems which arise in machine learning. For strongly convex problems, its convergence rate was known to be $O((\log(T))/T)$, by running SGD for $T$ iterations and returning the average point. However, recent results showed that using a different algorithm, one can get an optimal $O(1/T)$ rate. This might lead one to believe that standard SGD is suboptimal, and maybe should even be replaced as a method of choice. In this paper, we investigate the optimality of SGD in a stochastic setting. We show that for smooth problems, the algorithm attains the optimal $O(1/T)$ rate. However, for non-smooth problems, the convergence rate with averaging might really be $\Omega((\log(T))/T)$, and this is not just an artifact of the analysis. On the flip side, we show that a simple modification of the averaging step suffices to recover the $O(1/T)$ rate, and no other change of the algorithm is necessary. We also present experimental results which support our findings, and point out open problems.

1 Introduction

Stochastic gradient descent (SGD) is one of the simplest and most popular first-order methods to solve convex learning problems. Given a convex loss function and a training set of $T$ examples, SGD can be used to obtain a sequence of $T$ predictors, whose average has a generalization error which converges (with $T$) to the optimal one in the class of predictors we consider. The common framework to analyze such first-order algorithms is via stochastic optimization, where our goal is to optimize an unknown convex function $F$, given only unbiased estimates of $F$’s subgradients (see Sec. 2 for a more precise definition).

An important special case is when $F$ is strongly convex (intuitively, can be lower bounded by a quadratic function). Such functions arise, for instance, in Support Vector Machines and other regularized learning algorithms. For such problems, there is a well-known $O((\log(T))/T)$ convergence guarantee for SGD with averaging. This rate is obtained using the analysis of the algorithm in the harder setting of online learning [2], combined with an online-to-batch conversion (see [3] for more details).

Surprisingly, a recent paper by Hazan and Kale [3] showed that in fact, an $O((\log(T))/T)$ is not the best that one can achieve for strongly convex stochastic problems. In particular, an optimal $O(1/T)$ rate can be obtained using a different algorithm, which is somewhat similar to SGD but is more complex (although with comparable computational complexity). A very similar algorithm was also presented recently by Juditsky and Nesterov [5].

These results left an important gap: Namely, whether the true convergence rate of SGD, possibly with some sort of averaging, might also be $O(1/T)$, and the known $O((\log(T))/T)$ result is just an
artifact of the analysis. Indeed, the whole motivation of [3] was that the standard online analysis is too loose to analyze the stochastic setting properly. Perhaps a similar looseness applies to the analysis of SGD as well? This question has immediate practical relevance: if the new algorithms enjoy a better rate than SGD, it might indicate they will work better in practice, and that practitioners should abandon SGD in favor of them.

In this work, we study the convergence rate of SGD for stochastic strongly convex problems, with the following contributions:

- We begin by recalling and extending known results to show that if $F$ is not only strongly convex, but also smooth (with respect to the optimum), then SGD with and without averaging achieves the optimal $\mathcal{O}(1/T)$ convergence rate.
- We then show that for non-smooth $F$, there are cases where the convergence rate of SGD with averaging is $\Omega(\log(T)/T)$. In other words, the $\mathcal{O}(\log(T)/T)$ bound for general strongly convex problems is real, and not just an artifact of the currently-known analysis.
- However, we show that one can recover the optimal $\mathcal{O}(1/T)$ convergence rate, by a simple modification of the averaging step: Instead of averaging of $T$ points, we only average the last $\alpha T$ points, where $\alpha \in (0, 1)$ is arbitrary. Thus, to obtain an optimal rate, one does not need to use an algorithm significantly different than SGD, such as those discussed earlier.
- We perform an empirical study on both artificial and real-world data, which supports our findings.

In this extended abstract, we only sketch the results, and omit many important details, which appear in the full version of our paper.

2 Preliminaries

We consider the standard setting of convex stochastic optimization, using first-order methods. Our goal is to minimize a convex function $F$ over some convex domain $\mathcal{W}$ (which is assumed to be a subset of some Hilbert space). However, we do not know $F$, and the only information available is through a stochastic gradient oracle, which given some $w \in \mathcal{W}$, produces a vector $\hat{g}$, whose expectation $\mathbb{E}[\hat{g}] = g$ is a subgradient of $F$ at $w$. Using a bounded number $T$ of calls to this oracle, we wish to find a point $w_T$ such that $F(w_t)$ is as small as possible. In particular, we will assume that $F$ attains a minimum at some $w^* \in \mathcal{W}$, and our analysis will provide bounds on $\mathbb{E}[F(w_T) - F(w^*)]$, where the expectation is over the randomness of the algorithm.

We will focus on an important special case of the problem, characterized by $F$ being a strongly convex function. Formally, we say that a function $F$ is $\lambda$-strongly convex, if for all $w, w' \in \mathcal{W}$ and any subgradient $g$ of $F$ at $w$, $F(w') \geq F(w) + \langle g, w' - w \rangle + \frac{\lambda}{2} \|w' - w\|^2$. Another possible property of $F$ we will consider is smoothness, at least with respect to the optimum $w^*$. Formally, a function $F$ is $\mu$-smooth with respect to $w^*$ if for all $w \in \mathcal{W}$ and any subgradient $g$ of $F$ at $w$, $F(w) - F(w^*) \leq \frac{\mu}{2} \|w - w^*\|^2$. Such functions arise, for instance, in logistic and least-squares regression, and in general for learning linear predictors where the loss function has a Lipschitz-continuous gradient.

The algorithm we focus on is stochastic gradient descent (SGD). The SGD algorithm is parameterized by step sizes $\eta_1, \ldots, \eta_T$, and is defined as follows:

1. Initialize $w_1 \in \mathcal{W}$ arbitrarily (or randomly)
2. For $t = 1, \ldots, T$:
   - Query the stochastic gradient oracle at $w_t$ to get a random $\hat{g}_t$ such that $\mathbb{E}[\hat{g}_t] = g_t$ is a subgradient of $F$ at $w_t$.
   - Let $w_{t+1} = \Pi_{\mathcal{W}}(w_t - \eta_t \hat{g}_t)$, where $\Pi_{\mathcal{W}}$ is the projection operator on $\mathcal{W}$.

This algorithm returns a sequence of points $w_1, \ldots, w_T$. To obtain a single point, one can use several strategies. Perhaps the simplest one is to return the last point, $w_{T+1}$. Another procedure, for which the standard online analysis of SGD applies [2], is to return the average point

$$w_T = \frac{1}{T}(w_1 + \ldots + w_T).$$
For stochastic optimization of $\lambda$-strongly functions, the standard analysis (through online learning) focuses on the step size $\eta_t$ being exactly $1/\lambda t$ [2]. Our analysis will consider more general step-sizes $c/\lambda t$, where $c$ is a constant. In general, we will assume that regardless of how $w_1$ is initialized, it holds that $E[\|\hat{g}_t\|^2] \leq G^2$ for some fixed constant $G$.

## 3 Smooth Functions

We begin by considering the case where the expected function $F(\cdot)$ is both strongly convex and smooth with respect to $w^*$. Our starting point is to show a $O(1/T)$ for the last point obtained by SGD. We note that this result is well-known in the stochastic approximation literature (see for instance [1] and references therein), and we include it here for completeness.

**Theorem 1.** Suppose $F$ is $\lambda$-strongly convex and $\mu$-smooth with respect to $w^*$ over a convex set $\mathcal{W}$, and that $E[\|\hat{g}_t\|^2] \leq G^2$. Then if we pick $\eta_t = c/\lambda t$ for some constant $c > 1/2$, it holds for any $T$ that

$$E[F(w_T) - F(w^*)] \leq \frac{1}{2} \max \left\{ \frac{c}{2 - 1/c}, \frac{\mu G^2}{\lambda^2 T} \right\}.$$ 

The theorem is an immediate corollary of the following lemma, and the definition of $\mu$-smoothness with respect to $w^*$.

**Lemma 1.** Suppose $F$ is $\lambda$-strongly convex over a convex set $\mathcal{W}$, and that $E[\|\hat{g}_t\|^2] \leq G^2$. Then if we pick $\eta_t = c/\lambda t$ for some constant $c > 1/2$, it holds for any $T$ that

$$E[\|w_T - w^*\|^2] \leq \max \left\{ \frac{c}{2 - 1/c}, \frac{G^2}{\lambda^2 T} \right\}.$$ 

With this result in hand, we now turn to discuss the behavior of the average point $\bar{w}_T = (w_1 + \ldots + w_T)/T$, and show that for smooth $F$, it also enjoys an optimal $O(1/T)$ convergence rate.

**Theorem 2.** Suppose $F$ is $\lambda$-strongly convex and $\mu$-smooth with respect to $w^*$ over a convex set $\mathcal{W}$, and that $E[\|\hat{g}_t\|^2] \leq G^2$. Then if we pick $\eta_t = c/\lambda t$ for some constant $c > 1/2$, it holds for any $T$ that

$$E[F(\bar{w}_T) - F(w^*)] \leq 2 \max \left\{ \frac{\mu G^2}{\lambda^2}, \frac{4\mu G}{\lambda}, \frac{\mu G}{\lambda} \sqrt{\frac{4c}{2 - 1/c}} \right\} \frac{1}{T}.$$ 

## 4 Non-Smooth Functions

We now turn to the discuss the more general case where the function $F$ may not be smooth (i.e. there is no constant $\mu$ for which $F$ is $\mu$-smooth uniformly for all $w \in \mathcal{W}$). In the context of learning, this may happen when we try to learn a predictor with respect to a non-smooth loss function, such as the hinge loss.

As discussed earlier, SGD with averaging is known to have a rate of at most $O(\log(T)/T)$. In the previous section, we saw that for smooth $F$, the rate is actually $O(1/T)$. Moreover, [3] showed that for using a different algorithm than SGD, one can obtain a rate of $O(1/T)$ even in the non-smooth case. This might lead us to believe that an $O(1/T)$ rate for SGD is possible in the non-smooth case, and that the $O(\log(T)/T)$ analysis is simply not tight.

However, this intuition turns out to be wrong. Below, we show that there are strongly convex stochastic optimization problems in Euclidean space, in which the convergence rate of SGD with averaging is lower bounded by $\Omega(\log(T)/T)$. Thus, the logarithm in the bound is not merely a shortcoming in the standard online analysis of SGD, but is really a property of the algorithm.

We begin with a relatively simple example (Thm. 3), which shows the essence of the idea. We let $F$ be the 1-strongly convex function $F(w) = \frac{1}{2}||w||^2 + w_1$, which has a global minimum at $0$, and let $\mathcal{W} = [0,1]^d$. Suppose the stochastic gradient oracle, given a point $w_t$, returns the gradient estimate $\hat{g}_t = w_t + (Z_1, 0, \ldots, 0)$, where $Z_1$ is uniformly distributed over $[-1, 3]$. The intuition of the construction is that the global optimum lies at a corner of $\mathcal{W}$, so averaging the points returned by SGD does not help us in getting closer to it.
Theorem 3. For the stochastic optimization problem described above, if SGD is initialized at any point in \( W \), and ran with \( \eta_t = c/t \), then for any \( T \geq T_0 + 1 \), where \( T_0 = \max\{2, c/2\} \), we have

\[
\mathbb{E}[F(\bar{w}_T) - F(w^*)] \geq \frac{c}{16T} \sum_{t=0}^{T-1} \frac{1}{t}.
\]

When \( c \) is considered a constant, this lower bound is \( \Omega(\log(T)/T) \).

While being relatively straightforward, this example is not fully satisfying, since it crucially relies on the fact that \( w^* \) is on the border of \( W \). In strongly convex problems, \( w^* \) usually lies in the interior of \( W \), so perhaps the \( \Omega(\log(T)/T) \) lower bound does not hold in such cases. Our main result, presented below, shows that this is not the case, and that even if \( w^* \) is well inside the interior of \( W \), an \( \Omega(\log(T)/T) \) rate for SGD with averaging can be unavoidable. The intuition is that we construct a non-smooth \( F \), which forces \( w_t \) to approach the optimum from just one direction, creating the same effect as in the previous example.

Theorem 4. There exists a stochastic 1-strongly convex problem over \( W = [-1, 1]^d \), with a global minimum \( w^* \) at 0 and \( \mathbb{E}[\|\hat{g}_t\|^2] \leq d + 63 \), with the following property: If SGD is initialized at any point \( w_0 \) with \( w_{1,1} \geq 0 \), and ran with \( \eta_t = c/t \), then for any \( T \geq T_0 + 2 \), where \( T_0 = \max\{2, 6c + 1\} \), we have

\[
\mathbb{E}[F(\bar{w}_T) - F(w^*)] \geq \frac{3c}{16T} \sum_{t=0}^{T-1} \left( \frac{1}{t} \right) - \frac{T_0}{T}.
\]

When \( c \) is considered a constant, this lower bound is \( \Omega(\log(T)/T) \).

We note that the requirement of \( w_{1,1} \geq 0 \) can be relaxed at the cost of some additional second-order factors.

5 Recovering an \( O(1/T) \) Rate for SGD with \( \alpha \)-Suffix Averaging

In the previous section, we showed that SGD with averaging may have a rate of \( \Omega(\log(T)/T) \) for non-smooth \( F \). To get the optimal \( O(1/T) \) rate for any \( F \), we might turn to the algorithms of [3] and [5]. However, these algorithms constitute a significant departure from standard SGD. In this section, we show that it is actually possible to get an \( O(1/T) \) rate using a much simpler modification of the algorithm: given the sequence of points \( w_1, \ldots, w_T \) provided by SGD, instead of returning the average \( \bar{w}_T = (w_1 + \ldots + w_T)/T \), we average and return just a suffix, namely

\[
\bar{w}_T^\alpha = \frac{w_{(1-\alpha)T+1} + \ldots + w_T}{\alpha T},
\]

for some constant \( \alpha \in (0, 1) \) (assuming \( \alpha T \) and \( (1-\alpha)T \) are integers). We call this procedure \( \alpha \)-suffix averaging.

Theorem 5. Consider SGD with \( \alpha \)-suffix averaging as described above, and with step sizes \( \eta_t = c/\lambda t \) where \( c > 1/2 \) is a constant. Suppose \( F = \lambda \)-strongly convex, and that \( \mathbb{E}[\|\hat{g}_t\|^2] \leq G \) for all \( t \). Then for any \( T \), it holds that

\[
\mathbb{E}[F(\bar{w}_T^\alpha) - F(w^*)] \leq \left( c + \left( \frac{c}{2} + c' \right) \log \left( \frac{1}{\alpha} \right) \right) \frac{G^2}{\lambda T},
\]

where \( c' = \max\left\{ \frac{c}{2}, \frac{1}{4 - 2/c} \right\} \).

Note that for any constant \( \alpha \in (0, 1) \), the bound above is \( O(G^2/\lambda T) \). This applies to any relevant step size \( c/\lambda t \), and matches the optimal guarantees in [3] up to constant factors. However, this is shown for standard SGD, as opposed to the more specialized algorithm of [3].

6 Experiments

In the full version of our paper, we provide an empirical study of how the algorithms behave, and compare it to our theoretical findings. In this extended abstract, we only textually summarize our main findings.
We studied the behavior of the following four algorithms:

1. **SGD-A**: Performing SGD and then returning the average point over all $T$ rounds.
2. **SGD-$\alpha$**: Performing SGD with $\alpha$-suffix averaging. We chose $\alpha = 1/2$ - namely, we return the average point over the last $T/2$ rounds.
3. **SGD-L**: Performing SGD and returning the point obtained in the last round.

First, as a simple sanity check, we measured the performance of these algorithms on a simple, strongly-convex stochastic optimization problem, which is also smooth. All 4 algorithms clearly achieved a $\Theta(1/T)$ rate, matching our theoretical analysis (Thm. 1, Thm. 2 and Thm. 5). The results also seem to indicate that SGD-A has a somewhat worse performance in terms of leading constants.

Second, as another simple experiment, we measured the performance of the algorithms on the non-smooth, strongly convex problem used in the proof of Thm. 4. As our theory indicates, SGD-A seems to have an $\Theta(\log(T)/T)$ convergence rate, whereas the other 3 algorithms all seem to have the optimal $\Theta(1/T)$ convergence rate. Among these algorithms, the SGD variants SGD-L and SGD-$\alpha$ seem to perform somewhat better than EPOCH-GD. Also, while the average performance of SGD-L and SGD-$\alpha$ are similar, SGD-$\alpha$ has less variance. This is reasonable, considering the fact that SGD-$\alpha$ returns an average of many points, whereas SGD-L return only the very last point.

Finally, we performed a set of experiments on real-world data. We used the same 3 binary classification datasets (CCAT, COV1 and ASTRO-PH) used by [6] and [4], to test the performance of optimization algorithms for Support Vector Machines using linear kernels. Each of these datasets is composed of a training set and a test set. Given a training set of instance-label pairs, $\{(x_i, y_i)\}_{i=1}^m$, we defined $F$ to be the standard (non-smooth) objective function of Support Vector Machines, namely

$$F(w) = \frac{\lambda}{2} \|w\|^2 + \frac{1}{m} \sum_{i=1}^m \max\{0, 1 - y_i \langle x_i, w \rangle\}.$$  \hspace{1cm} (1)

In all experiments, SGD-A performed the worst. The other 3 algorithms performed rather similarly, with SGD-$\alpha$ being slightly better on the COV1 dataset, and SGD-L being slightly better on the other 2 datasets.

### 7 Discussion and Open Questions

In this paper, we analyzed the behavior of SGD for strongly convex stochastic optimization problems. We demonstrated that this simple and well-known algorithm performs optimally whenever the underlying function is also smooth, but the standard averaging step can make it suboptimal for non-smooth problems. However, a simple modification of the averaging step suffices to recover the optimal rate, and a more sophisticated algorithm is not necessary. The experiments we presented seem to support this conclusion.

There are still several open issues remaining. For example, all our results are for bounds which hold in expectation, and it remains to be seen whether high-probability bounds with a similar rate can be obtained. Moreover, the $O(1/T)$ rate in the non-smooth case still requires some sort of averaging. However, in our experiments and other studies (e.g. [6]), returning the last point seem to perform very well. Our current theory does not cover this - at best, one can use Lemma 1 and Jensen’s inequality to argue that the last predictor has a convergence rate of $O(1/\sqrt{T})$, but the behavior in practice is clearly much better. Is it possible to show that SGD, without averaging, obtains an $O(1/T)$ for non-smooth strongly convex problems?

### References


