Making Gradient Descent Optimal for Strongly Convex Stochastic Optimization

Alexander Rakhlin RAKHLIN@WHARTON.UPENN.EDU
University of Pennsylvania

Ohad Shamir OHADSH@MICROSOFT.COM
Microsoft Research New England

Karthik Sridharan SKARTHIK@WHARTON.UPENN.EDU
University of Pennsylvania

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 $\mathcal{O}(\log(T)/T)$, 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 $\mathcal{O}(1/T)$ 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 $\mathcal{O}(1/T)$ 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 $\mathcal{O}(1/T)$ 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 $\mathcal{O}(\log(T)/T)$ convergence guarantee for SGD with averaging. This rate is obtained using the analysis of the algorithm in the harder setting of online learning (Hazan et al., 2007), combined with an online-to-batch conversion (see (Hazan & Kale, 2011) for more details).

Surprisingly, a recent paper by Hazan and Kale (Hazan & Kale, 2011) showed that in fact, an $\mathcal{O}(\log(T)/T)$ is not the best that one can achieve for strongly convex stochastic problems. In particular, an optimal $\mathcal{O}(1/T)$ rate can be obtained using a different algorithm, which is somewhat similar to SGD but is more complex (although with comparable computational complexity)[1]. A very similar algorithm was also presented recently by Juditsky and Nesterov (Juditsky & Nesterov, 2010).

Appearing in *Proceedings of the 29th International Conference on Machine Learning*, Edinburgh, Scotland, UK, 2012.

[1] Roughly speaking, the algorithm divides the T iterations into exponentially increasing epochs, and runs stochastic gradient descent with averaging on each one. The resulting point of each epoch is used as the starting point of the next epoch. The algorithm returns the resulting point of the last epoch.

These results left an important gap: Namely, whether the true convergence rate of SGD, possibly with some sort of averaging, might also be $\mathcal{O}(1/T)$, and the known $\mathcal{O}(\log(T)/T)$ result is just an artifact of the analysis. Indeed, the whole motivation of (Hazan & Kale, 2011) 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 paper, we study the convergence rate of SGD for stochastic strongly convex problems, with the following contributions:

- First, we extend 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 (in expectation and in high probability) by a simple modification of the averaging step: Instead of averaging of T points, we only average the last αT 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.

Moreover, our rate upper bounds are shown to hold in expectation, as well as in high probability (up to a $\log(\log(T))$ factor). While the focus here is on getting the optimal rate in terms of T, we note that our upper bounds are also optimal in terms of other standard problem parameters, such as the strong convexity parameter and the variance of the stochastic gradients.

Following the paradigm of (Hazan & Kale, 2011), we analyze the algorithm directly in the stochastic setting, and avoid an online analysis with an online-to-batch conversion. This also allows us to prove results which are more general. In particular, the standard online analysis of SGD requires the step size of the algorithm at round t to equal $1/\lambda t$, where λ is the strong convexity parameter of F. In contrast, our analysis copes with any step size $c/\lambda t$, as long as c is not too small.

In terms of related work, we note that the performance of SGD in a stochastic setting has been extensively researched in stochastic approximation theory (see for instance (Kushner & Yin, 2003)). However, these results are usually obtained under smoothness assumptions, and are often asymptotic, so we do not get an explicit bound in terms of T which applies to our setting. We also note that a finite-sample analysis of SGD in the stochastic setting was recently presented in (Bach & Moulines, 2011). However, the focus there was different than ours, and also obtained bounds which hold only in expectation rather than in high probability. More importantly, the analysis was carried out under stronger smoothness assumptions than our analysis, and to the best of our understanding, does not apply to general, possibly non-smooth, strongly convex stochastic optimization problems. For example, smoothness assumptions may not cover the application of SGD to support vector machines (as in (Shalev-Shwartz et al., 2011)), since it uses a non-smooth loss function, and thus the underlying function F we are trying to stochastically optimize may not be smooth.

2. Preliminaries

We use bold-face letters to denote vectors. Given some vector $\mathbf{w}$, we use w_i to denote its i-th coordinate. Similarly, given some indexed vector $\mathbf{w}_t$, we let $w_{t,i}$ denote its i-th coordinate. We let $\mathbf{1}_A$ denote the indicator function for some event A.

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 $\mathbf{w} \in \mathcal{W}$, produces a vector $\hat{\mathbf{g}}$, whose expectation $\mathbb{E}[\hat{\mathbf{g}}] = \mathbf{g}$ is a subgradient of F at $\mathbf{w}$. Using a bounded number T of calls to this oracle, we wish to find a point $\mathbf{w}_T$ such that $F(\mathbf{w}_t)$ is as small as possible. In particular, we will assume that F attains a minimum at some $\mathbf{w}^* \in \mathcal{W}$, and our analysis provides bounds on $F(\mathbf{w}_t) - F(\mathbf{w}^*)$ either in expectation or in high probability (the high probability results are stronger, but require more effort and have slightly worse dependence on some problem parameters). The application of this framework to learning is straightforward (see for instance (Shalev-Shwartz et al., 2009)):

given a hypothesis class $\mathcal{W}$ and a set of T i.i.d. examples, we wish to find a predictor $\mathbf{w}$ whose expected loss $F(\mathbf{w})$ is close to optimal over $\mathcal{W}$. Since the examples are chosen i.i.d., the subgradient of the loss function with respect to any individual example can be shown to be an unbiased estimate of a subgradient of F.

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 *λ-strongly convex*, if for all $\mathbf{w}, \mathbf{w}' \in \mathcal{W}$ and any subgradient $\mathbf{g}$ of F at $\mathbf{w}$,

$$F(\mathbf{w}') \geq F(\mathbf{w}) + \langle \mathbf{g}, \mathbf{w}' - \mathbf{w} \rangle + \frac{\lambda}{2}\|\mathbf{w}' - \mathbf{w}\|^2. \quad (1)$$

Another possible property of F we will consider is smoothness, at least with respect to the optimum $\mathbf{w}^*$. Formally, a function F is *μ-smooth with respect to $\mathbf{w}^*$* if for all $\mathbf{w} \in \mathcal{W}$,

$$F(\mathbf{w}) - F(\mathbf{w}^*) \leq \frac{\mu}{2}\|\mathbf{w} - \mathbf{w}^*\|^2. \quad (2)$$

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 $\mathbf{w}_1 \in \mathcal{W}$ arbitrarily (or randomly)
2. For $t = 1, \ldots, T$:
   - Query the stochastic gradient oracle at $\mathbf{w}_t$ to get a random $\hat{\mathbf{g}}_t$ such that $\mathbb{E}[\hat{\mathbf{g}}_t] = \mathbf{g}_t$ is a subgradient of F at $\mathbf{w}_t$.
   - Let $\mathbf{w}_{t+1} = \Pi_{\mathcal{W}}(\mathbf{w}_t - \eta_t \hat{\mathbf{g}}_t)$, where $\Pi_{\mathcal{W}}$ is the projection operator on $\mathcal{W}$.

This algorithm returns a sequence of points $\mathbf{w}_1, \ldots, \mathbf{w}_T$. To obtain a single point, one can use several strategies. Perhaps the simplest one is to return the last point, $\mathbf{w}_{T+1}$. Another procedure, for which the standard online analysis of SGD applies (Hazan et al., 2007), is to return the average point

$$\bar{\mathbf{w}}_T = \frac{1}{T}(\mathbf{w}_1 + \ldots + \mathbf{w}_T).$$

For stochastic optimization of λ-strongly functions, the standard analysis (through online learning) focuses on the step size η_t being exactly $1/\lambda t$ (Hazan et al., 2007). Our analysis will consider more general step-sizes $c/\lambda t$, where c is a constant. We note that a step size of $\Theta(1/t)$ is necessary for the algorithm to obtain an optimal convergence rate (see Appendix A in the full version (Rakhlin et al., 2011)).

In general, we will assume that regardless of how $\mathbf{w}_1$ is initialized, it holds that $\mathbb{E}[\|\hat{\mathbf{g}}_t\|^2] \leq G^2$ for some fixed constant G. Note that this is a somewhat weaker assumption than (Hazan & Kale, 2011), which required that $\|\hat{\mathbf{g}}_t\|^2 \leq G^2$ with probability 1, since we focus only on bounds which hold in expectation. These types of assumptions are common in the literature, and are generally implied by taking $\mathcal{W}$ to be a bounded domain, or alternatively, assuming that $\mathbf{w}_1$ is initialized not too far from $\mathbf{w}^*$ and F satisfies certain technical conditions (see for instance the proof of Theorem 1 in (Shalev-Shwartz et al., 2011)).

Full proofs of our results are provided in Appendix B of the full version of this paper (Rakhlin et al., 2011).

3. Smooth Functions

We begin by considering the case where the expected function $F(\cdot)$ is both strongly convex and smooth with respect to $\mathbf{w}^*$. Our starting point is to show a $\mathcal{O}(1/T)$ for the *last* point obtained by SGD. This result is well known in the literature (see for instance (Nemirovski et al., 2009)) and we include a proof for completeness. Later on, we will show how to extend it to a high-probability bound.

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

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

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

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

$$\mathbb{E}\left[\|\mathbf{w}_T - \mathbf{w}^*\|^2\right] \leq \max\left\{4\ ,\ \frac{c}{2 - 1/c}\right\}\frac{G^2}{\lambda^2 T}.$$

We now turn to discuss the behavior of the average point $\bar{\mathbf{w}}_T = (\mathbf{w}_1 + \ldots + \mathbf{w}_T)/T$, and show that for smooth F, it also enjoys an optimal $\mathcal{O}(1/T)$ convergence rate (with even better dependence on c).

Theorem 2. *Suppose F is λ-strongly convex and μ-smooth with respect to $\mathbf{w}^*$ over a convex set $\mathcal{W}$, and* *that $\mathbb{E}[\|\hat{\mathbf{g}}_t\|^2] \leq G^2$. Then if we pick $\eta_t = c/\lambda t$ for some constant $c > 1/2$, $\mathbb{E}[F(\bar{\mathbf{w}}_T) - F(\mathbf{w}^*)]$ is at most*

$$2\max\left\{\frac{\mu G^2}{\lambda^2}\ ,\ \frac{4\mu G}{\lambda}\ ,\ \frac{\mu G}{\lambda}\sqrt{\frac{4c}{2-1/c}}\right\}\frac{1}{T}.$$

A rough proof intuition is the following: Lemma 1 implies that the Euclidean distance of $\mathbf{w}_t$ from $\mathbf{w}^*$ is on the order of $1/\sqrt{t}$, so the squared distance of $\bar{\mathbf{w}}_T$ from $\mathbf{w}^*$ is on the order of $((1/T)\sum_{t=1}^{T} 1/\sqrt{t})^2 \approx 1/T$, and the rest follows from smoothness.

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 μ which satisfies Eq. (2) uniformly for all $\mathbf{w} \in \mathcal{W}$). 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 $\mathcal{O}(\log(T)/T)$. In the previous section, we saw that for smooth F, the rate is actually $\mathcal{O}(1/T)$. Moreover, (Hazan & Kale, 2011) showed that for using a different algorithm than SGD, one can obtain a rate of $\mathcal{O}(1/T)$ even in the non-smooth case. This might lead us to believe that an $\mathcal{O}(1/T)$ rate for SGD is possible in the non-smooth case, and that the $\mathcal{O}(\log(T)/T)$ 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 the following relatively simple example, which shows the essence of the idea. Let F be the 1-strongly convex function

$$F(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|^2 + w_1,$$

over the domain $\mathcal{W} = [0,1]^d$, which has a global minimum at $\mathbf{0}$. Suppose the stochastic gradient oracle, given a point $\mathbf{w}_t$, returns the gradient estimate $\hat{\mathbf{g}}_t = \mathbf{w}_t + (Z_t, 0, \ldots, 0)$, where Z_t is uniformly distributed over $[-1,3]$. It is easily verified that $\mathbb{E}[\hat{\mathbf{g}}_t]$ is a subgradient of $F(\mathbf{w}_t)$, and that $\mathbb{E}[\|\hat{\mathbf{g}}_t\|^2] \leq d+5$ which is a bounded quantity for fixed d.

The following theorem implies in this case, the convergence rate of SGD with averaging has a $\Omega(\log(T)/T)$ lower bound. The intuition for this is that the global optimum lies at a corner of $\mathcal{W}$, so SGD "approaches" it only from one direction. As a result, averaging the points returned by SGD actually hurts us.

Theorem 3. *Consider the strongly convex stochastic optimization problem presented above. If SGD is initialized at any point in $\mathcal{W}$, and ran with $\eta_t = c/t$, then for any $T \geq T_0 + 1$, where $T_0 = \max\{2, c/2\}$, we have*

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

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

While the lower bound scales with c, we remind the reader that one must pick $\eta_t = c/t$ with constant c for an optimal convergence rate in general (see discussion in Sec. 2).

This example is relatively straightforward but not fully satisfying, since it crucially relies on the fact that $\mathbf{w}^*$ is on the border of $\mathcal{W}$. In strongly convex problems, $\mathbf{w}^*$ usually lies in the interior of $\mathcal{W}$, 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 $\mathbf{w}^*$ is well inside the interior of $\mathcal{W}$, 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 $\mathbf{w}_t$ to approach the optimum from just one direction, creating the same effect as in the previous example.

In particular, let F be the 1-strongly convex function

$$F(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|^2 + \begin{cases} w_1 & w_1 \geq 0 \\ -7w_1 & w_1 < 0 \end{cases},$$

over the domain $\mathcal{W} = [-1,1]^d$, which has a global minimum at $\mathbf{0}$. Suppose the stochastic gradient oracle, given a point $\mathbf{w}_t$, returns the gradient estimate

$$\hat{\mathbf{g}}_t = \mathbf{w}_t + \begin{cases} (Z_t, 0, \ldots, 0) & w_1 \geq 0 \\ (-7, 0, \ldots, 0) & w_1 < 0 \end{cases},$$

where Z_t is a random variable uniformly distributed over $[-1,3]$. It is easily verified that $\mathbb{E}[\hat{\mathbf{g}}_t]$ is a subgradient of $F(\mathbf{w}_t)$, and that $\mathbb{E}[\|\hat{\mathbf{g}}_t\|^2] \leq d + 63$ which is a bounded quantity for fixed d.

Theorem 4. *Consider the strongly convex stochastic optimization problem presented above. If SGD is initialized at any point $\mathbf{w}_1$ 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*

$$\mathbb{E}\left[F(\bar{\mathbf{w}}_T) - F(\mathbf{w}^*)\right] \ \geq\ \frac{3c}{16T}\sum_{t=T_0+2}^{T}\left(\frac{1}{t}\right) - \frac{T_0}{T}.$$

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

We note that the requirement of $w_{1,1} \geq 0$ is just for convenience, and the analysis also carries through, with some second-order factors, if we let $w_{1,1} < 0$.

5. Recovering an $\mathcal{O}(1/T)$ Rate for SGD with α-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 $\mathcal{O}(1/T)$ rate for any F, we might turn to the algorithms of (Hazan & Kale, 2011) and (Juditsky & Nesterov, 2010). However, these algorithms constitute a significant departure from standard SGD. In this section, we show that it is actually possible to get an $\mathcal{O}(1/T)$ rate using a much simpler modification of the algorithm: given the sequence of points $\mathbf{w}_1, \ldots, \mathbf{w}_T$ provided by SGD, instead of returning the average $\bar{\mathbf{w}}_T = (\mathbf{w}_1 + \ldots + \mathbf{w}_T)/T$, we average and return just a suffix, namely

$$\bar{\mathbf{w}}_T^\alpha = \frac{\mathbf{w}_{(1-\alpha)T+1} + \ldots + \mathbf{w}_T}{\alpha T},$$

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

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

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

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

Note that for any constant $\alpha \in (0,1)$, the bound above is $\mathcal{O}(G^2/\lambda T)$. This applies to any relevant step size $c/\lambda t$, and matches the optimal guarantees in (Hazan & Kale, 2011) up to constant factors. However, this is shown for standard SGD, as opposed to the more specialized algorithm of (Hazan & Kale, 2011). Finally, we note that it might be tempting to use Thm. 5 as a guide to choose the averaging window, by optimizing the bound for α (for instance, for $c = 1$, the optimum is achieved around $\alpha \approx 0.65$). However, we note that the optimal value of α is dependent on the constants in the bound, which may not be the tightest or most "correct" ones.

Proof Sketch. The proof combines the analysis of online gradient descent (Hazan et al., 2007) and Lemma 1. In particular, starting as in the proof of Lemma 1, and extracting the inner products, we get

$$\sum_{t=(1-\alpha)T+1}^{T} \mathbb{E}[\langle \mathbf{g}_t, \mathbf{w}_t - \mathbf{w}^* \rangle] \leq \sum_{t=(1-\alpha)T+1}^{T} \frac{\eta_t G^2}{2} + \sum_{t=(1-\alpha)T+1}^{T} \left(\frac{\mathbb{E}[\|\mathbf{w}_t - \mathbf{w}^*\|^2]}{2\eta_t} - \frac{\mathbb{E}[\|\mathbf{w}_{t+1} - \mathbf{w}^*\|^2]}{2\eta_t} \right). \tag{3}$$

Rearranging the r.h.s., and using the convexity of F to relate the l.h.s. to $\mathbb{E}[F(\bar{\mathbf{w}}_T^\alpha) - F(\mathbf{w}^*)]$, we get a convergence upper bound of

$$\frac{1}{2\alpha T} \left(\frac{\mathbb{E}[\|\mathbf{w}_{(1-\alpha)T+1} - \mathbf{w}^*\|^2]}{\eta_{(1-\alpha)T+1}} + G^2 \sum_{t=(1-\alpha)T+1}^{T} \eta_t + \sum_{t=(1-\alpha)T+1}^{T} \mathbb{E}[\|\mathbf{w}_t - \mathbf{w}^*\|^2] \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} \right) \right).$$

Lemma 1 tells us that with any strongly convex F, even non-smooth, we have $\mathbb{E}[\|\mathbf{w}_t - \mathbf{w}^*\|^2] \leq \mathcal{O}(1/t)$. Plugging this in and performing a few more manipulations, the result follows. □

One potential disadvantage of suffix averaging is that if we cannot store all the iterates $\mathbf{w}_t$ in memory, then we need to know from which iterate αT to start computing the suffix average (in contrast, standard averaging can be computed "on-the-fly" without knowing the stopping time T in advance). However, even if T is not known, this can be easily addressed in several ways. For example, since our results are robust to the value of α, it is really enough to guess when we passed some "constant" portion of all iterates. Alternatively, one can divide the rounds into exponentially increasing epochs, and maintain the average just of the current epoch. Such an average would always correspond to a constant-portion suffix of all iterates.

6. High-Probability Bounds

All our previous bounds were on the *expected* suboptimality $\mathbb{E}[F(\mathbf{w}) - F(\mathbf{w}^*)]$ of an appropriate predictor $\mathbf{w}$. We now outline how these results can be strengthened to bounds on $F(\mathbf{w}_t) - F(\mathbf{w}^*)$ which hold with arbitrarily high probability $1 - \delta$, with the bound depending logarithmically on δ. They are slightly worse than our in-expectation bounds by having worse dependence on the step size parameter c and an additional $\log(\log(T))$ factor (interestingly, a similar factor also appears in the analysis of (Hazan & Kale, 2011), and we do not know if it is necessary). The key result is the following strengthening of Lemma 1, under slightly stronger technical conditions.

Lemma 2. *Let $\delta \in (0, 1/e)$ and $T \geq 4$. Suppose F is λ-strongly convex over a convex set $\mathcal{W}$, and that $\|\hat{\mathbf{g}}_t\|^2 \leq G^2$ with probability 1. Then if we pick $\eta_t = c/\lambda t$ for some constant $c > 1/2$, such that $2c$ is a whole number, it holds with probability at least $1-\delta$ that for any $t \in \{4c^2+4c, \ldots, T-1, T\}$ that*

$$\|\mathbf{w}_t - \mathbf{w}^*\|^2 \leq \frac{12c^2G^2}{\lambda^2 t} + 8(121G+1)G\frac{c\log(\log(t)/\delta)}{\lambda t}.$$

We note that the assumptions on $2c$ and t are only for simplifying the result. To obtain high probability versions of Thm. 1, Thm. 2, and Thm. 5, we simply need to plug in this lemma in lieu of Lemma 1 in their proofs. This leads overall to rates of the form $\mathcal{O}(\log(\log(T)/\delta)/T)$ which hold with probability $1-\delta$.

7. Experiments

We now turn to empirically study how the algorithms behave, and compare it to our theoretical findings.

We studied the following four algorithms:

1. Sgd-A: Performing SGD and then returning the average point over all T rounds.
2. Sgd-α: Performing SGD with α-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.
4. Epoch-Gd: The optimal algorithm of (Hazan & Kale, 2011) for strongly convex stochastic optimization.

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. We define $\mathcal{W} = [-1, 1]^5$, and $F(\mathbf{w}) = \|\mathbf{w}\|^2$. The stochastic gradient oracle, given a point $\mathbf{w}$, returns the stochastic gradient $\mathbf{w}+\mathbf{z}$ where $\mathbf{z}$ is uniformly distributed in $[-1, 1]^5$. Clearly, this is an unbiased estimate of the gradient of F at $\mathbf{w}$. The initial point $\mathbf{w}_1$ of all 4 algorithms was chosen uniformly at random from $\mathcal{W}$. The results are presented in Fig. 1, and it is clear that all 4 algorithms indeed achieve 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.

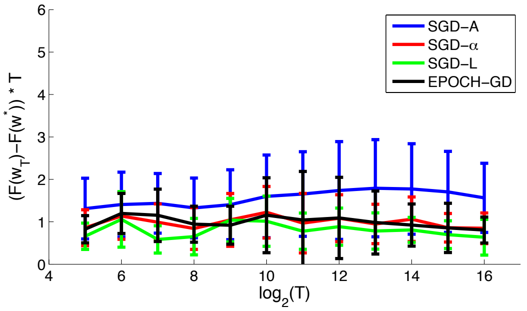

Figure 1. Results for smooth strongly convex stochastic optimization problem. The experiment was repeated 10 times, and we report the mean and standard deviation for each choice of T. The X-axis is the log-number of rounds $\log(T)$, and the Y-axis is $(F(\mathbf{w}_T)-F(\mathbf{w}^*))*T$. The scaling by T means that a roughly constant graph corresponds to a $\Theta(1/T)$ rate, whereas a linearly increasing graph corresponds to a $\Theta(\log(T)/T)$ rate.

Second, as another simple experiment, we measured the performance of the algorithms on the non-smooth, strongly convex problem described in the proof of Thm. 4. In particular, we simulated this problem with $d = 5$, and picked $\mathbf{w}_1$ uniformly at random from $\mathcal{W}$. The results are presented in Fig. 2. 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-α seem to perform somewhat better than Epoch-Gd. Also, while the average performance of Sgd-L and Sgd-α are similar, Sgd-α has less variance. This is reasonable, considering the fact that Sgd-α 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 (Shalev-Shwartz et al., 2011) and (Joachims, 2006), 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, $\{\mathbf{x}_i, y_i\}_{i=1}^m$, we defined F to be the standard (non-smooth) objective function of Support Vector Machines, namely

$$F(\mathbf{w}) = \frac{\lambda}{2}\|\mathbf{w}\|^2 + \frac{1}{m}\sum_{i=1}^{m}\max\{0, 1-y_i\langle\mathbf{x}_i, \mathbf{w}\rangle\}. \quad (4)$$

Following (Shalev-Shwartz et al., 2011) and (Joachims, 2006), we took $\lambda = 10^{-4}$ for ccat, $\lambda = 10^{-6}$ for cov1, and $\lambda = 5 \times 10^{-5}$ for astro-ph. The stochastic gra-

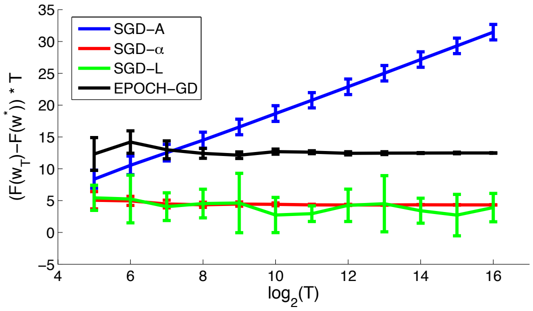

Figure 2. Results for the non-smooth strongly convex stochastic optimization problem. The experiment was repeated 10 times, and we report the mean and standard deviation for each choice of T. The X-axis is the log-number of rounds $\log(T)$, and the Y-axis is $(F(\mathbf{w}_T) - F(\mathbf{w}^*)) * T$. The scaling by T means that a roughly constant graph corresponds to a $\Theta(1/T)$ rate, whereas a linearly increasing graph corresponds to a $\Theta(\log(T)/T)$ rate.

dient given $\mathbf{w}_t$ was computed by taking a single randomly drawn training example $(\mathbf{x}_i, y_i)$, and computing the gradient with respect to that example, namely

$$\hat{\mathbf{g}}_t = \lambda \mathbf{w}_t - \mathbf{1}_{y_i \langle \mathbf{x}_i, \mathbf{w}_t \rangle \leq 1} y_i \mathbf{x}_i.$$

Each dataset comes with a separate test set, and we also report the objective function value with respect to that set (as in Eq. (4), this time with $\{\mathbf{x}_i, y_i\}$ representing the test set examples). All algorithms were initialized at $\mathbf{w}_1 = 0$, with $\mathcal{W} = \mathbb{R}^d$ (i.e. no projections were performed - see the discussion in Sec. 2).

The results of the experiments are presented in Fig. 3,Fig. 4 and Fig. 5. In all experiments, Sgd-A performed the worst. The other 3 algorithms performed rather similarly, with Sgd-α being slightly better on the Cov1 dataset, and Sgd-L being slightly better on the other 2 datasets.

In summary, our experiments indicate the following:

- Sgd-A, which averages over all T predictors, is worse than the other approaches. This accords with our theory, as well as the results reported in (Shalev-Shwartz et al., 2011).
- The Epoch-Gd algorithm does have better performance than Sgd-A, but a similar or better performance was obtained using the simpler approaches of α-suffix averaging (Sgd-α) or even just returning the last predictor (Sgd-L). The good performance of Sgd-α is supported by our theoretical results, and so does the performance of Sgd-L in the strongly convex and smooth case.

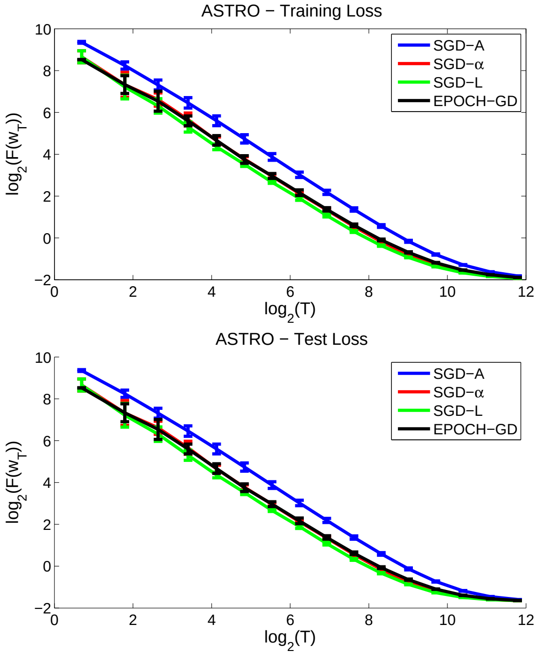

Figure 3. Results for the ASTRO-PH dataset. The left row refers to the average loss on the training data, and the right row refers to the average loss on the test data. Each experiment was repeated 10 times, and we report the mean and standard deviation for each choice of T. The X-axis is the log-number of rounds $\log(T)$, and the Y-axis is the log of the objective function $\log(F(\mathbf{w}_T))$.

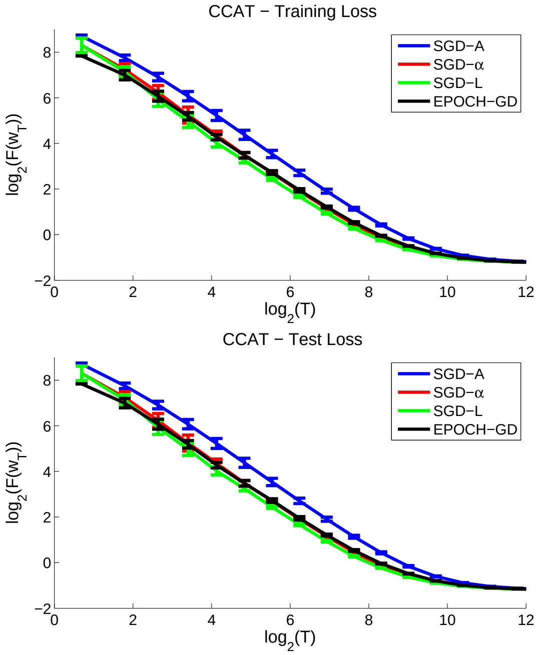

Figure 4. Results for the CCAT dataset. See Fig. 3 caption for details.

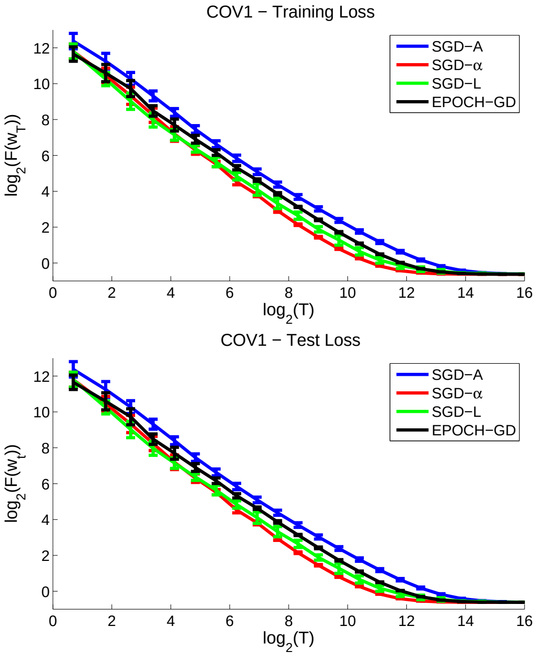

Figure 5. Results for the CCAT dataset. See Fig. 3 caption for details.

- SGD-L also performed rather well (with what seems like a $\Theta(1/T)$ rate) on the non-smooth problem reported in Fig. 2, although with a larger variance than SGD-α. Our current theory does not cover the convergence of the last predictor in non-smooth problems - see the discussion below.

8. Discussion

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 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. Our experiments seem to support this conclusion.

There are several open issues remaining. In particular, the $\mathcal{O}(1/T)$ rate in the non-smooth case still requires some sort of averaging. However, in our experiments and other studies (e.g. (Shalev-Shwartz et al., 2011)), returning the last iterate $\mathbf{w}_T$ also seems to perform quite well. Our current theory does not cover this - at best, one can use Lemma 1 and Jensen's inequality to argue that the last iterate has a $\mathcal{O}(1/\sqrt{T})$ rate, but the behavior in practice is clearly much better. Does SGD, *without* averaging, obtain an $\mathcal{O}(1/T)$ rate for general strongly convex problems? Also, a fuller empirical study is warranted of whether and which averaging scheme is best in practice.

Acknowledgements: We thank Elad Hazan and Satyen Kale for helpful comments on an earlier version of this paper.

References

Bach, F. and Moulines, E. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *NIPS*, 2011.

Hazan, E. and Kale, S. Beyond the regret minimization barrier: An optimal algorithm for stochastic strongly-convex optimization. In *COLT*, 2011.

Hazan, E., Agarwal, A., and Kale, S. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2-3):169–192, 2007.

Joachims, T. Training linear SVMs in linear time. In *KDD*, 2006.

Juditsky, A. and Nesterov, Y. Primal-dual subgradient methods for minimizing uniformly convex functions. Technical Report (August 2010), available at `http://hal.archives-ouvertes.fr/docs/00/50/89/33/PDF/Strong-hal.pdf`, 2010.

Kushner, H. and Yin, G. *Stochastic Approximation and Recursive Algorithms and Applications*. Springer, 2nd edition, 2003.

Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A. Robust stochastic approximation approach to stochastic programming. *SIAM J. Optim.*, 19(4): 1574–1609, 2009.

Rakhlin, A., Shamir, O., and Sridharan, K. Making gradient descent optimal for strongly convex stochastic optimization. ArXiv Technical Report, arXiv:1109.5647, 2011.

Shalev-Shwartz, S., Shamir, O., Srebro, N., and Sridharan, K. Stochastic convex optimization. In *COLT*, 2009.

Shalev-Shwartz, S., Singer, Y., Srebro, N., and Cotter, A. Pegasos: primal estimated sub-gradient solver for svm. *Mathematical Programming*, 127(1):3–30, 2011.