JMLR: Workshop and Conference Proceedings vol (2012) 1–3

Is Averaging Needed for Strongly Convex Stochastic Gradient Descent?

Ohad Shamir OHADSH@MICROSOFT.COM
Microsoft Research New England
One Memorial Drive, Cambridge MA 02142, USA

Abstract

Stochastic gradient descent (SGD) is a simple and very popular iterative method to solve stochastic optimization problems which arise in machine learning. A common practice is to return the average of the SGD iterates. While the utility of this is well-understood for general convex problems, the situation is much less clear for strongly convex problems (such as solving SVM). Although the standard analysis in the strongly convex case requires averaging, it was recently shown that this actually degrades the convergence rate, and a better rate is obtainable by averaging just a suffix of the iterates. The question we pose is whether averaging is needed at all to get optimal rates.

We consider the problem of stochastically optimizing a convex function F over a convex domain $\mathcal{W}$ using stochastic gradient descent (SGD). The algorithm makes use of an oracle, which given some $\mathbf{w} \in \mathcal{W}$, returns a random vector $\hat{\mathbf{g}}$ whose expectation is a subgradient of $F(\mathbf{w})$. For example, consider the linear SVM optimization problem over a training set $\{(\mathbf{x}_i, y_i)\}_{i=1}^m$,

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

Given some $\mathbf{w}$, we can easily compute an unbiased estimate of its gradient, by picking a single example $(\mathbf{x}_i, y_i)$ uniformly at random, and returning a subgradient of $\frac{\lambda}{2}\|\mathbf{w}\|^2 + \max\{0, 1 - y_i\langle \mathbf{x}_i, \mathbf{w}\rangle\}$.

SGD is parameterized by step sizes $\eta_1, \ldots, \eta_T$, and is defined as follows: we initialize $\mathbf{w}_1 \in \mathcal{W}$ arbitrarily. At each round t, we obtain an unbiased estimate $\hat{\mathbf{g}}_t$ of a subgradient of $F(\mathbf{w}_t)$, and 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 produces a sequence of iterates $\mathbf{w}_1, \ldots, \mathbf{w}_T$. For general convex problems, it is well-known that if we pick $\eta_t = \Theta(1/\sqrt{T})$, and return the *average* of the iterates $\bar{\mathbf{w}}_T = (\mathbf{w}_1 + \ldots, \mathbf{w}_T)/T$, then under mild conditions, $F(\bar{\mathbf{w}}_T) - \inf_{\mathbf{w}\in\mathcal{W}} F(\mathbf{w}) \leq \mathcal{O}(1/\sqrt{T})$, both in expectation and in high probability $1 - \delta$ (with logarithmic dependence on δ).

We focus here on cases where F is *strongly convex* - roughly speaking, that it can be lower bounded at any point in $\mathcal{W}$ by a quadratic function (as in the case of SVM optimization). When F is strongly convex, one can obtain faster convergence rates. For example, by picking $\eta_t = 1/t$, the expected suboptimality of $\bar{\mathbf{w}}_T$ is $\mathcal{O}(\log(T)/T)$ (Hazan et al. (2007); Shalev-Shwartz et al. (2011)).

Recently, Rakhlin et al. (2012) showed that in fact, one can get rid of the $\log(T)$ factor, and get an optimal $1/T$ rate for step sizes $\eta_t = \Theta(1/T)$, by replacing the simple average $\bar{\mathbf{w}}_T$

by *suffix averaging* - namely, the average of the last αT iterates for some constant $\alpha \in (0,1)$. The resulting expected suboptimality bound is of the form $\mathcal{O}\left(\frac{\log(1/(1-\alpha))}{\alpha}\frac{1}{T}\right)$. Moreover, this change is significant: whereas we can always get $1/T$ rate with suffix averaging, there are cases where the rate of $\bar{\mathbf{w}}_T$ is no better than $\log(T)/T$. This disparity in performance was also demonstrated empirically on real-world datasets.

However, this is not the end of the story. While the theoretical result above suggests we should average some constant portion of the iterates, experimental studies suggest that a simpler approach of averaging just the last few iterates, or even returning the last iterate $\mathbf{w}_T$, can work very well in practice (Shalev-Shwartz et al. (2011); Rakhlin et al. (2012)). Thus, we currently do not have a satisfactory understanding of which averaging scheme is best, or whether averaging is needed at all, leading to the question posed in the title.

It is important to note that when F is also *smooth* at the optimum $\mathbf{w}^*$ (namely, for some $\mu > 0$, $F(\mathbf{w}) - F(\mathbf{w}^*) \leq \mu\|\mathbf{w} - \mathbf{w}^*\|^2$ for any $\mathbf{w} \in \mathcal{W}$), then one can show that any kind of suffix averaging, and even just the last iterate, enjoys a $1/T$ convergence rate (see more details below). However, many important problems in practice, such as SVM optimization, are not smooth. Therefore, we can formulate the following specific question, with a 50$ monetary reward.

- ($50) **What is the expected suboptimality $\mathbb{E}[F(\mathbf{w}_T) - F(\mathbf{w}^*)]$ of the last iterate returned by SGD, for general (possibly non-smooth) strongly convex problems?**

The bound should hold for some fixed choice of the step-sizes, and fixed strong convexity parameter, for any stochastic subgradients with bounded norm, and any bounded $\mathcal{W}$. The full reward will be given if the rate is shown to be tight up to universal constants.

Another relevant issue is obtaining a bound which holds not just in expectation, but also in arbitrarily high probability $1 - \delta$ with logarithmic dependence on δ. Such a high probability result may be important for quantifying the reliability of the last iterate, and how much its suboptimality can fluctuate. Thus, an extra $20 will be awarded for proving a tight bound (up to constants) on the suboptimality of $\mathbf{w}_T$ which holds in high probability.

Partial Results

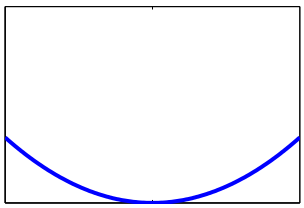
In the analysis of SGD in a stochastic setting, it is common to analyze the evolution of the quantity $\|\mathbf{w}_t - \mathbf{w}^*\|^2$. By definition of SGD and assuming (w.l.o.g.) that F is 1-strongly convex, it is not hard to show that

$$\mathbb{E}[\|\mathbf{w}_{t+1} - \mathbf{w}^*\|^2] \leq (1 - 2\eta_t)\mathbb{E}[\|\mathbf{w}_t - \mathbf{w}^*\|^2] + \mathcal{O}(\eta_t^2). \tag{1}$$

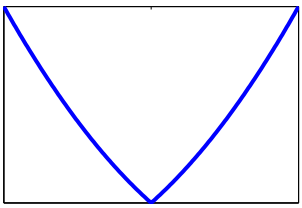
By choosing $\eta_t = \Theta(1/t)$ appropriately and solving the recursion, we get

$$\mathbb{E}[\|\mathbf{w}_T - \mathbf{w}^*\|^2] \leq \mathcal{O}\left(\frac{1}{T}\right). \tag{2}$$

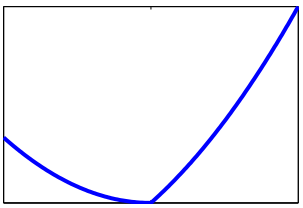
When F is also smooth, this immediately leads to the $\mathbb{E}[F(\mathbf{w}_T) - F(\mathbf{w}^*)] \leq \mathcal{O}(1/T)$ bound mentioned earlier. However, the question here is what can we say in the non-smooth case. Unfortunately, Eq. (2) is no longer good enough. Using Jensen's inequality and



$F(w) = \frac{1}{2}w^2$



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



$F(w) = \frac{1}{2}w^2 + \max\{w, 0\}$

an assumption that F is Lipschitz, it only allows us to bound $F(\mathbf{w}_T) - F(\mathbf{w}^*)$ by the disappointing rate $\mathcal{O}(1/\sqrt{T})$, and this is the best one can hope for. Thus, to get a better rate in the non-smooth case, it seems one must prove a stronger version of Eq. (2).

We conjecture that this should be possible. For example, consider the scalar function $F(w) = \frac{1}{2}w^2 + |w|$ (so that $w^* = 0$ - see figure above), and suppose we take $\eta_t = 1/t$. A straightforward calculation reveals that

$$\mathbb{E}[w_{t+1}^2] \leq (1 - 2\eta_t)\,\mathbb{E}[w_t^2] - 2\eta_t \mathbb{E}[|w_t|] + \mathcal{O}\left(\eta_t^2\right).$$

This is stronger than Eq. (1), due to the $-\mathbb{E}[|w_t|]$ component. Intuitively, the fact that F is non-smooth and "pointy" around the optimum makes w_t converge faster. This seems to compensate for the fact that $F(w) - F(w^*)$ now scales down like $|w|$ rather than w^2 (as in the smooth case). Indeed, if we ignore the fact that we deal with expectations, and consider a recursive inequality of the form

$$x_{t+1}^2 \leq (1 - 2\eta_t)\,x_t^2 - 2\eta_t x_t + \mathcal{O}\left(\eta_t^2\right),$$

and $\eta_t = \Theta(1/t)$, it can be shown that x_t^2 scales down as $1/t^2$. This seems to suggest that $\mathbb{E}[w_T^2]$ indeed converges as $\mathcal{O}(1/T^2)$ and not just $\mathcal{O}(1/T)$, leading to $\mathbb{E}[F(w_T) - F(w^*)] \leq \mathcal{O}(1/T)$ as required.

Even if such a result can be proven formally, we don't know how to generalize this approach to more complex functions. For example, consider the function $F(w) = \frac{1}{2}w^2 + \max\{w, 0\}$ (see figure above). This function is simultaneously smooth around $w^* = 0$ in the negative direction, and non-smooth in the positive direction. In this case, any bound on $\mathbb{E}[w_T^2]$ or even $\mathbb{E}[|w_T|]$ will not be enough, as it would correspond to different suboptimality rates in the positive and negative direction. In this particular example, a delicate case analysis might be possible, but it's not clear how to make things work with multi-dimensional functions, where the amount of "non-smoothness" can vary continuously in different directions.

References

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

A. Rakhlin, O. Shamir, and K. Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. In *ICML*, 2012.

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