Open Problem: Tight Convergence of SGD in Constant Dimension

Tomer Koren TKOREN@TAUEX.TAU.AC.IL
Shahar Segal SHAHARSEGAL1@MAIL.TAU.AC.IL
School of Computer Science, Tel Aviv University, Tel Aviv, Israel

Editors: Jacob Abernethy and Shivani Agarwal

Abstract

Stochastic Gradient Descent (SGD) is one of the most popular optimization methods in machine learning and has been studied extensively since the early 50's. However, our understanding of this fundamental algorithm is still lacking in certain aspects. We point out to a gap that remains between the known upper and lower bounds for the expected suboptimality of the last SGD point whenever the dimension is a constant independent of the number of SGD iterations T, and in particular, that the gap is still unaddressed even in the one dimensional case. For the latter, we provide evidence that the correct rate is $\Theta(1/\sqrt{T})$ and conjecture that the same applies in any (constant) dimension.

1. Background

We consider the problem of minimizing a convex and Lipschitz objective F over a convex domain $\mathcal{W} \subseteq \mathbb{R}^d$ using Stochastic Gradient Descent (SGD). Formally, given a stochastic gradient oracle that for an input $w_t \in \mathcal{W}$ returns a random vector $\widehat{g}_t$ whose expectation is a subgradient of F at w_t, SGD produces a sequence of iterates $w_1, \ldots, w_T \in \mathcal{W}$ over T iterations according to $w_{t+1} = \Pi_{\mathcal{W}}[w_t - \eta_t \widehat{g}_t]$, where $\Pi_{\mathcal{W}}$ denotes projection onto $\mathcal{W}$. Here $\eta_1, \ldots, \eta_T$ is a sequence of step sizes; standard choices are a fixed setting of $\eta_t = \eta = \Theta(1/\sqrt{T})$, or a decreasing schedule $\eta_t = \Theta(1/\sqrt{t})$ for all t. In both case, it is well known that the average iterate $\overline{w}_T = \frac{1}{T}\sum_{t=1}^{T} w_t$ attains the tight $\Theta(1/\sqrt{T})$ convergence rate for the expected suboptimality $\mathbf{E}[F(\overline{w}_T)] - F(w^*)$, where $w^* = \arg\min_{w \in \mathcal{W}} F(w)$. (More recently, the same rate was shown for various other forms of averaging; see Shamir and Zhang, 2013.)

A more recent line of work has focused on nailing down the tight convergence rate for the expected suboptimality of the *last iterate* w_T itself, which is very often a more preferred choice when using SGD (and variants) in practice. Shamir and Zhang (2013) established an $O(\log(T)/\sqrt{T})$ convergence rate for the last iterate with a standard step size sequence of $\eta_t = \Theta(1/\sqrt{t})$. A few years later, Harvey et al. (2019) showed that with the same step size schedule, this rate cannot be improved and demonstrated a function F for which the suboptimality of the last iterate is $\Omega(\log(T)/\sqrt{T})$.[1] In a related effort, Jain et al. (2019) gave an $O(1/\sqrt{T})$ bound for the last iterate, closing the gap between the average and the last iterate, but used a very different (and rather non-standard) step size sequence.

This is however not the end of the story. A closer look at Harvey et al. (2019)'s lower bound construction reveals that it applies when the dimensionality of the problem is $\Omega(T)$, and in particular, it does not address optimization in constant dimension. It is unclear if and how can an analogous construction be designed for when the dimension is constant and cannot grow with T. To the best of

1. Their construction is in fact deterministic (i.e., the gradient oracle is non-stochastic), which implies that the same lower bound applies even for non-stochastic (non-smooth) optimization with (sub)gradient descent.

our knowledge, even in dimension one the best known lower bound is the classical $\Omega(1/\sqrt{T})$, while the best known upper bound is the $O(\log(T)/\sqrt{T})$ due to Shamir and Zhang (2013).

We note that in the non-stochastic one-dimensional case, obtaining an improved $O(1/\sqrt{T})$ upper bound is actually straightforward: it is not hard to see that in dimension one, upon reaching a good point with $F(w_\tau) - F(w^*) = O(1/\sqrt{T})$, subsequent subgradient steps can only increase this gap by at most $O(\eta)$ (assuming the objective is Lipschitz); combining this observation with the standard result that for $\eta = \Theta(1/\sqrt{T})$, the average suboptimality $\frac{1}{T}\sum_{t=1}^{T}(F(w_t) - F(w^*))$ is $O(1/\sqrt{T})$ (and so there must exist such a good iterate w_τ) implies that the bound for the last iterate is also $O(1/\sqrt{T})$. As we discuss below, adapting this simple argument to stochastic optimization with SGD (and to higher dimensions) appears surprisingly challenging.

2. Open Problems and Conjectures

Our main open question can thus be summarized as follows:

Open Problem 1. *What is the (expected) convergence rate of the last point of SGD with a fixed step size $\eta = \Theta(1/\sqrt{T})$ for a convex function in **constant** dimension?*

As noted above, even the one dimensional case is open: while for deterministic optimization establishing a $\Theta(1/\sqrt{T})$ rate is straightforward, proving the same for SGD appears more challenging, even just in expectation. We conjecture that the right rate of SGD in the one dimensional case is indeed $\Theta(1/\sqrt{T})$; the next section gives more details and some preliminary supporting evidence. For dimension $d > 1$, a natural conjecture is that the right convergence rate is $\Theta(\log(d)/\sqrt{T})$, but we have no indication to corroborate this.

We also state an analogous question for non-smooth GD in constant dimension $d > 1$:

Open Problem 2. *What is the convergence rate of the last point of GD with a fixed step size $\eta = \Theta(1/\sqrt{T})$ for a (non-smooth) convex function in **constant** dimension $d > 1$?*

We remark that analogous questions can be posed for SGD with a decreasing step size sequence of the form $\eta_t = \Theta(1/\sqrt{t})$; however, we argue that the fixed step-size versions are more basic and should be resolved before these. (Both Shamir and Zhang, 2013 and Harvey et al., 2019 study the decreasing step-size version of SGD but their arguments can be easily adjusted for the simpler case of a fixed step size.) The same goes for the strongly convex versions of the questions, which are also open to the best of our knowledge.

3. Preliminary Evidence in One Dimension

We present a simple one-dimensional example to illustrate the behavior we conjecture for SGD in dimension one. Consider a scaled absolute value function $F(w) = \varepsilon|w|$, where $0 \le \varepsilon \le 1$, paired with the following stochastic gradient oracle: at any $w \neq 0$, it returns $\operatorname{sign}(w)$ with probability $\frac{1+\varepsilon}{2}$, and $-\operatorname{sign}(w)$ with the remaining probability (with an expectation of $\varepsilon \operatorname{sign}(w) = \nabla F(w)$); at $w = 0$ it returns ± 1 with probability $\frac{1-\varepsilon}{4}$ each, and 0 otherwise (thus the expected subgradient is $0 \in \partial F(0)$).[2]

We consider SGD iterations with a fixed step size $\eta > 0$ initialized at the optimum $w^* = 0$, giving rise to a stochastic process $W_0 = 0, W_1, \ldots, W_T$. We aim to show that $\mathbf{E}[|W_T|]$ is at most $O(\eta/\varepsilon)$, as this would imply that $\mathbf{E}[F(W_T)] = \varepsilon\mathbf{E}[|W_T|] = O(\eta)$ and having $\eta = \Theta(1/\sqrt{T})$ would give the desired bound over the expected suboptimality of the last iterate.

2. The exact behavior at $w = 0$ is not crucial but somewhat simplifies the calculations below.

Reduction to a one-sided random walk. The absolute location of the SGD iterates can be described as a one-sided random walk $\{X_t\}_{t\geq 0}$ on the nonnegative integers, such that $|W_t| = \eta X_t$ for all $t \geq 0$ (see Fig. 1). At every step, the probability to move up (by +1) is $\frac{1-\varepsilon}{2}$. If the current location is 0, we stay at 0 with the remaining probability, and otherwise we move down (by -1). Our goal is to then to show that the expected location of the walk after T steps is bounded as $\mathbf{E}[X_T] = O(1/\varepsilon)$.

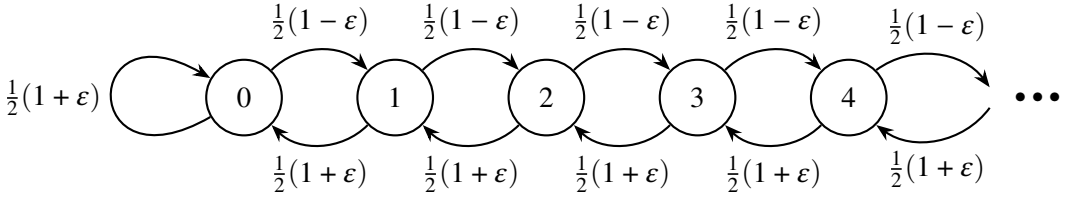

Figure 1: X_t is a one-sided random walk on the non-negative integers.

Denote by $p_{t,n} = \Pr[X_t = n]$ the probability to be at location n at time t. Since the walk is biased towards zero, it is immediate that the probabilities $p_{t,n}$ decrease exponentially fast with n:

Claim 3. *For all t, n, it holds that $p_{t,n} \leq (\frac{1-\varepsilon}{1+\varepsilon})^n p_{t,0}$. In particular, this implies $p_{t,0} \geq \frac{2\varepsilon}{1+\varepsilon}$ for all t.*

Proof. By induction on t; for $t = 0$, this is immediate since we begin the walk at 0, so $p_{t,0} = 1$ and $p_{t,n} = 0$ for $n > 0$. Inductively, for any $t > 0, n > 1$:

$$p_{t,n} = \frac{1-\varepsilon}{2} p_{t-1,n-1} + \frac{1+\varepsilon}{2} p_{t-1,n+1} \leq \frac{1-\varepsilon}{1+\varepsilon}\left(\frac{1-\varepsilon}{2} p_{t-1,n-2} + \frac{1+\varepsilon}{2} p_{t-1,n}\right) = \frac{1-\varepsilon}{1+\varepsilon} p_{t,n-1}.$$

For $n = 1$ we can use a similar argument. This exponential decrease implies that:

$$p_{t,0} = 1 - \sum_{i=1}^{t} p_{t,i} \geq 1 - \sum_{i=1}^{\infty}\left(\frac{1-\varepsilon}{1+\varepsilon}\right)^i p_{t,0} \implies p_{t,0} \geq \left(\sum_{i=0}^{\infty}\left(\frac{1-\varepsilon}{1+\varepsilon}\right)^i\right)^{-1} = 1 - \frac{1-\varepsilon}{1+\varepsilon} = \frac{2\varepsilon}{1+\varepsilon}. \quad \blacksquare$$

This simple claim already implies our desired bound in two extreme cases: if ε is large (a positive constant) then the exponential decay implies that $\mathbf{E}[X_T] = O(1)$, whence $\mathbf{E}[F(W_T)] = O(\eta)$; and if $\varepsilon = 0$ then the objective function is flat and we do not care about the deviation of X_T (even though it is fairly large in that case: when $\varepsilon = 0$ this is just the expected absolute deviation of a simple *unbiased* walk, and it is a standard fact that $\mathbf{E}[X_T] = \Theta(\sqrt{T})$). For intermediate values of ε, however, the situation appears to be a bit more subtle, as we show next.

A loose bound from simple concentration. As a straightforward consequence of the exponential decay in Claim 3, it is not hard to show that $\mathbf{E}[X_T] = O(\frac{1}{\varepsilon}\log\frac{1}{\varepsilon})$. This bound is too weak for our purposes; however, it turns out that such an argument falls short of giving a better bound:

Claim 4. *There exists a distribution $\{p_n\}_{n\geq 0}$ over the nonnegative integers that satisfies the condition $p_n \leq (\frac{1-\varepsilon}{1+\varepsilon})^n p_0$ for all n, yet its expectation is $\Omega(\frac{1}{\varepsilon}\log\frac{1}{\varepsilon})$.*

Proof. Denote $\gamma = 1 - \frac{1-\varepsilon}{1+\varepsilon}$ (we will assume for simplicity that $\gamma < \frac{1}{2}$) and consider a distribution where the probability of each integer n between $n_1 = \frac{1}{2\gamma}\log\frac{1}{\gamma}$ and $n_2 = n_1 + \frac{1}{\gamma}$ is $p_n = \frac{1}{2}\gamma$, and the remaining mass is assigned to $n = 0$. We then have $p_0 \geq \frac{1}{2}$ and for each $n_1 \leq n \leq n_2$ it indeed holds that $(\frac{1-\varepsilon}{1+\varepsilon})^n p_0 = p_0(1-\gamma)^n \geq \frac{1}{2}e^{-2\gamma n} \geq \frac{1}{2}\gamma = p_n$, However, the expectation of this distribution is at least $\Omega(\frac{1}{\gamma}\log\frac{1}{\gamma})$ as a constant fraction of the probability mass is placed on integers $\geq n_1$. $\blacksquare$

A tight bound via generating functions. We give a more delicate argument that yields a tight $\mathbf{E}[X_T] = O(1/\varepsilon)$ bound, which makes use of a generating function associated with the one-sided random walk. Let us denote the step taken at time t by Y_t, which can be $+1, -1$ or 0 with probabilities depending on whether $X_t = 0$. (Note that the Y_t are *not* independent). By definition $X_T = \sum_{t=0}^{T-1} Y_t$ and the expectation be can written as:

$$\mathbf{E}[X_T] = \sum_{t=0}^{T-1} \mathbf{E}[Y_t] = \sum_{t=0}^{T-1} \left(-\varepsilon \Pr[X_t \neq 0] + \frac{1-\varepsilon}{2} \Pr[X_t = 0] \right) = \frac{1+\varepsilon}{2} \sum_{t=0}^{T-1} \left(p_{t,0} - \frac{2\varepsilon}{1+\varepsilon} \right). \quad (1)$$

We thus wish to bound the sum of the $p_{t,0}$. To do so, we will use the associated generating function:

$$\mathcal{H}(z) = \sum_{t=0}^{\infty} p_{t,0}\, z^t.$$

Following ideas found in Feller (1971) (see also Drmota, 2003), we can show the following.

Claim 5. *We have*

$$\mathcal{H}(z) = \frac{2}{1-(1+\varepsilon)z + \sqrt{1-(1-\varepsilon^2)z^2}} = \frac{1-(1+\varepsilon)z - \sqrt{1-(1-\varepsilon^2)z^2}}{(1+\varepsilon)z(z-1)}.$$

Proof. Consider another generating function $\mathcal{M}(z) = \sum_{t=0}^{\infty} q_t z^t$ where q_t is the probability of the following event: the walk reached $X_t = 0$ without ever crossing the self-loop at zero. We show that:

$$\mathcal{H}(z) = 1 + \frac{1+\varepsilon}{2} z \cdot \mathcal{H}(z) + \frac{1-\varepsilon}{2} z \cdot \mathcal{M}(z) \cdot \frac{1+\varepsilon}{2} z \cdot \mathcal{H}(z) \quad (2)$$

by examining the first step taken by the walk. If that step is staying at 0 then the contribution is $\frac{1+\varepsilon}{2} z \cdot \mathcal{H}(z)$. Otherwise, it steps to 1 (with a contribution of $\frac{1-\varepsilon}{2} z$), then takes a walk on the positive integers and reaches 1 back (this is equivalent to $\mathcal{M}(z)$), and right after that takes a step back to zero ($\frac{1+\varepsilon}{2} z$), and start again from zero ($\mathcal{H}(z)$). Using similar arguments, we can show that:

$$\mathcal{M}(z) = 1 + \frac{1-\varepsilon}{2} z \cdot \mathcal{M}(z) \cdot \frac{1+\varepsilon}{2} z \cdot \mathcal{M}(z).$$

Solving this quadratic for $\mathcal{M}(z)$, plugging into Eq. (2) and solving for $\mathcal{H}(z)$ proves the claim. ∎

From Claim 3 we know that $p_{t,0} \geq \frac{2\varepsilon}{1+\varepsilon}$ for all t, thus from Eq. (1) we obtain

$$\mathbf{E}[X_T] \leq \frac{1+\varepsilon}{2} \sum_{t=0}^{\infty} \left(p_{t,0} - \frac{2\varepsilon}{1+\varepsilon} \right) = \lim_{z \to 1} \frac{1+\varepsilon}{2} \left(\mathcal{H}(z) - \sum_{t=0}^{\infty} \frac{2\varepsilon}{1+\varepsilon} z^t \right).$$

To compute the limit on the right-hand side, we simplify using $\sum_{t=0}^{\infty} z^t = 1/(1-z)$:

$$\mathcal{H}(z) - \frac{2\varepsilon}{1+\varepsilon} \sum_{t=0}^{\infty} z^t = \frac{1-\varepsilon}{1+\varepsilon} \cdot \frac{2}{1-(1-\varepsilon)z + \sqrt{1-(1-\varepsilon^2)z^2}}.$$

For $z \to 1$, this expression becomes $\frac{1}{\varepsilon} \frac{1-\varepsilon}{1+\varepsilon}$, which gives us the bound $\mathbf{E}[X_T] \leq \frac{1-\varepsilon}{2\varepsilon}$. In terms of the original SGD with step size η, this implies $\mathbf{E}[F(W_T)] \leq \frac{1-\varepsilon}{2\varepsilon} \cdot \eta\varepsilon = O(\eta)$, as we set out to prove.

References

Michael Drmota. Discrete random walks on one-sided "periodic" graphs. In *Discrete Random Walks (DRW'03)*, DMTCS Proceedings, pages 83–94. DMTCS, 2003.

William Feller. An introduction to probability theory and its applications. *Wiley Series in Probability and Mathematical Statistics, New York: Wiley, 1971, 3rd ed.*, 1971.

Nicholas JA Harvey, Christopher Liaw, Yaniv Plan, and Sikander Randhawa. Tight analyses for non-smooth stochastic gradient descent. In *COLT*, pages 1579–1613, 2019.

Prateek Jain, Dheeraj Nagaraj, and Praneeth Netrapalli. Making the last iterate of SGD information theoretically optimal. In *COLT*, pages 1752–1755, 2019.

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