Tight Dimension Independent Lower Bound on Optimal Expected Convergence Rate for Diminishing Step Sizes in SGD

Phuong Ha Nguyen   Lam M. Nguyen   Marten van Dijk

October 12, 2018

Abstract

We study convergence of Stochastic Gradient Descent (SGD) for strongly convex and smooth objective functions F.

We prove a lower bound on the expected convergence rate which holds for any sequence of diminishing stepsizes that are designed based on only global knowledge such as the fact that F is smooth and strongly convex, the component functions are smooth and convex, together with more additional information. Our lower bound meets the expected convergence rate of a recently proposed sequence of stepsizes at ICML 2018, which is based on such knowledge, within a factor 32. This shows that the stepsizes as proposed in the ICML paper are close to optimal. Furthermore, we conclude that in order to be able to construct stepsizes that beat our lower bound, more detailed information about F must be known.

Our work significantly improves over the state-of-the-art lower bound which we show is another factor $643 \cdot d$ worse, where d is the dimension. We are the first to prove a lower bound that comes within a small constant – independent from any other problem specific parameters – from an optimal solution.

1   Introduction

We are interested in solving the following stochastic optimization problem

$$\min_{w \in \mathbb{R}^d} \{F(w) = \mathbb{E}[f(w; \xi)]\}, \tag{1}$$

where ξ is a random variable obeying some distribution $g(\xi)$. In the case of empirical risk minimization with a training set $\{(x_i, y_i)\}_{i=1}^n$, ξ_i is a random variable that is defined by a single random sample (x, y)

Phuong Ha Nguyen, Department of Electrical and Computer Engineering, University of Connecticut, Storrs, CT, USA. Email: phuongha.ntu@gmail.com

Lam M. Nguyen, IBM Thomas J. Watson Research Center, Yorktown Heights, NY, USA. Email: LamNguyen.MLTD@ibm.com

Marten van Dijk, Department of Electrical and Computer Engineering, University of Connecticut, Storrs, CT, USA. Email: marten.van_dijk@uconn.edu

1

pulled uniformly from the training set. Then, by defining $f_i(w) := f(w; \xi_i)$, empirical risk minimization reduces to

$$\min_{w \in \mathbb{R}^d} \left\{ F(w) = \frac{1}{n} \sum_{i=1}^{n} f_i(w) \right\}. \tag{2}$$

Problems of this type arise frequently in supervised learning applications [7]. The classic first-order methods to solve problem (2) are gradient descent (GD) [17] and stochastic gradient descent (SGD) [1] [19] algorithms. GD is a standard deterministic gradient method, which updates iterates along the negative full gradient with learning η_t as follows

$$w_{t+1} = w_t - \eta_t \nabla F(w_t) = w_t - \frac{\eta_t}{n} \sum_{i=1}^{n} \nabla f_i(w_t) \,, \ t \geq 0.$$

We can choose $\eta_t = \eta = \mathcal{O}(1/L)$ and achieve a linear convergence rate for the strongly convex case [14]. The upper bound of the convergence rate of GD and SGD has been studied in [2, 4, 14, 20]. However, GD requires evaluation of n derivatives at each step, which is very expensive and therefore avoided in large-scale optimization. To reduce the computational cost for solving (2), a class of variance reduction methods [9, 6, 8, 15] has been proposed. The difference between GD and variance reduction methods is that GD needs to compute the full gradient at each step, while the variance reduction methods will compute the full gradient after a certain number of steps. In this way, variance reduction methods have less computational cost compared to GD. To avoid evaluating the full gradient at all, SGD generates an unbiased random variable ξ_t such as

$$\mathbb{E}_{\xi_t}[\nabla f(w_t; \xi_t)] = \nabla F(w_t),$$

and then evaluates gradient $\nabla f(w_t; \xi_t)$ for ξ_t drawn from distribution $g(\xi)$. After this, w_t is updated as follows

$$w_{t+1} = w_t - \eta_t \nabla f(w_t; \xi_t). \tag{3}$$

Algorithm 1 provides a detailed description. Obviously, the computational cost of SGD is n times cheaper than that of GD. However, as has been shown in literature we need to choose $\eta_t = \mathcal{O}(1/t)$ and the convergence rate of SGD is slowed down to $\mathcal{O}(1/t)$ [3], which is a sublinear convergence rate.

Algorithm 1 Stochastic Gradient Descent (SGD) Method
___
    Initialize: w_0
    Iterate:
    for $t = 0, 1, \ldots$ **do**
        Choose a step size (i.e., learning rate) $\eta_t > 0$.
        Generate a random variable ξ_t.
        Compute a stochastic gradient $\nabla f(w_t; \xi_t)$.
        Update the new iterate $w_{t+1} = w_t - \eta_t \nabla f(w_t; \xi_t)$.
    end for
___

In this paper we focus on the general problem (1) where F is strongly convex. Since F is strongly convex, a unique optimal solution of (1) exists and throughout the paper we denote this optimal solution by w_*. The

---

[1]We notice that even though stochastic gradient is referred to as SG in literature, the term stochastic gradient descent (SGD) has been widely used in many important works of large-scale learning.

starting point for analysis is the recurrence

$$\mathbb{E}[\|w_{t+1} - w_*\|^2] \le (1 - \mu\eta_t)\mathbb{E}[\|w_t - w_*\|^2] + \eta_t^2 N, \tag{4}$$

where

$$N = 2\mathbb{E}[\|\nabla f(w_*; \xi)\|^2]$$

and η_t is upper bounded by $\frac{1}{2L}$; the recurrence has been shown to hold if we assume (1) N is finite, (2) $F(.)$ is μ-strongly convex, (3) $f(w; \xi)$ is L-smooth and (4) convex [16, 10]; we detail these assumptions below:

Assumption 1 (μ-strongly convex)**.** *The objective function* $F : \mathbb{R}^d \to \mathbb{R}$ *is μ-strongly convex, i.e., there exists a constant $\mu > 0$ such that $\forall w, w' \in \mathbb{R}^d$,*

$$F(w) - F(w') \ge \langle \nabla F(w'), (w - w') \rangle + \frac{\mu}{2}\|w - w'\|^2. \tag{5}$$

As shown in [14, 3], Assumption 1 implies

$$2\mu[F(w) - F(w_*)] \le \|\nabla F(w)\|^2 , \ \forall w \in \mathbb{R}^d. \tag{6}$$

Assumption 2 (L-smooth)**.** *$f(w; \xi)$ is L-smooth for every realization of ξ, i.e., there exists a constant $L > 0$ such that, $\forall w, w' \in \mathbb{R}^d$,*

$$\|\nabla f(w; \xi) - \nabla f(w'; \xi)\| \le L\|w - w'\|. \tag{7}$$

Assumption 2 implies that F is also L-smooth.

Assumption 3. *$f(w; \xi)$ is convex for every realization of ξ, i.e., $\forall w, w' \in \mathbb{R}^d$,*

$$f(w; \xi) - f(w'; \xi) \ge \langle \nabla f(w'; \xi), (w - w') \rangle.$$

We notice that the earlier established recurrence in [12] under the same set of assumptions

$$\mathbb{E}[\|w_{t+1} - w_*\|^2] \le (1 - 2\mu\eta_t + 2L^2\eta_t^2)\mathbb{E}[\|w_t - w_*\|^2] + \eta_t^2 N$$

is similar, but worse than (4) as it only holds for $\eta_t < \frac{\mu}{L^2}$ where (4) holds for $\eta_t \le \frac{1}{2L}$. Only for step sizes $\eta_t < \frac{\mu}{2L^2}$ the above recurrence provides a better bound than (4), i.e., $1 - 2\mu\eta_t + 2L^2\eta_t^2 \le 1 - \mu\eta_t$. In practical settings such as logistic regression $\mu = \mathcal{O}(1/n)$, $L = \mathcal{O}(1)$, and $t = \mathcal{O}(n)$ (i.e. t is at most a relatively small constant number of epochs, where a single epoch represents n iterations resembling the complexity of a single GD computation). As we will show, for this parameter setting the optimally chosen step sizes are $\gg \frac{\mu}{L^2}$. This is the reason we focus in this paper on analysing recurrence (4): For $\eta_t \le \frac{1}{2L}$,

$$Y_{t+1} \le (1 - \mu\eta_t)Y_t + \eta_t^2 N,$$

where $Y_t = \mathbb{E}[\|w_t - w_*\|^2]$.

Problem Statement: It is well-known that based on the above assumptions (*without* the so-called bounded gradient assumption) and knowledge of only μ and L a sequence of stepsizes η_t can be constructed such that Y_t is smaller than $\mathcal{O}(1/t)$ [16]; more explicitly, $Y_t \le \frac{16N}{\mu} \frac{1}{\mu(t - \frac{20L}{\mu}) + 4L}$. Knowing a tight lower bound on Y_t is important because of the following reasons: (1) It helps us understand into what extend a given

sequence of stepsizes η_t leads to an optimal expected convergence rate. (2) The lower bound tells us that a sequence of step sizes as a function of only μ and L cannot beat an expected convergence rate of $O(1/t)$. More information is needed in a construction of η_t if we want to achieve a better expected convergence rate $Y_t = \mathcal{O}(1/t^p)$ where $p > 1$.

Related Work and Contribution: The authors of [13] proposed the first formal study about lower bounding the expected convergence rate for SGD. The authors of [1] and [18] independently studied this lower bound using information theory and were able to improve it.

As in this paper, the derivation in [1] is for SGD where the sequence of step sizes is a-priori fixed based on global information regarding assumed stochastic parameters concerning the objective function F. Their proof uses the following three assumptions (in this paper we assume a different set of assumptions as listed above):

1. The assumption of a strongly convex objective function, i.e., Assumption 1 (see Definition 3 in [1]).

2. There exists a bounded convex set $\mathcal{S} \subset \mathbb{R}^d$ such that

$$\mathbb{E}[\|\nabla f(w; \xi)\|^2] \le \sigma^2$$

   for all $w \in \mathcal{S} \subset \mathbb{R}^d$ (see Definition 1 in [1]). Notice that this is not the same as the bounded gradient assumption where $\mathcal{S} = \mathbb{R}^d$ is unbounded.[2]

3. The objective function F is a convex Lipschitz function, i.e., there exists a positive number K such that
$$\|F(w) - F(w')\| \le K\|w - w'\|, \forall w, w' \in \mathcal{S} \subset \mathbb{R}^d.$$

   We notice that this assumption actually implies the assumption on bounded gradients as stated above.

To prove the lower bound for strongly convex and Lipschitz objective functions, the authors constructed a class of objective functions and showed that the lower bound for this class is, in terms of the notation used in this paper,

$$\frac{\log(2/\sqrt{e})}{108d} \frac{N}{\mu^2 t}. \tag{8}$$

We revisit their derivation in supplementary material A where we show how their lower bound transforms into (8). Notice that their lower bound depends on dimension d.

In this paper we prove for strongly convex and smooth objective functions the lower bound

$$\approx \frac{1}{2} \frac{N}{\mu^2 t}.$$

Our lower bound is independent from d and, in fact, it meets the expected convergence rate for a specifically constructed sequence of step sizes (based on only the parameters μ for strong convexity and L for smoothness) within a factor 32. This proves that this sequence of step sizes leads to an optimal expected convergence rate within the small factor of 32 and proves that our lower bound is tight within a factor of 32. Notice that we significantly improve over state of the art since (8) is a factor $643 \cdot d$ larger than our lower bound, and more important, our lower bound is independent of d.

---

[2]The bounded gradient assumption, where $\mathcal{S}$ is unbounded, is in conflict with assuming strong convexity as explained in [16].

The specifically constructed sequence of step sizes mentioned above is from [16] and is given by $\eta_t = \frac{2}{\mu t + 4L}$ and yields expected convergence rate $Y_t \approx \frac{16N}{\mu^2 t}$. This explains the 32 factor difference.

In [19], the authors proved that in order to make SGD converge, stepsizes η_t should satisfy conditions

$$\sum_{t=0}^{\infty} \eta_t = \infty \text{ and } \sum_{t=0}^{\infty} \eta_t^2 < \infty.$$

In [12], the authors studied the expected convergence rates for another class of step sizes of $\mathcal{O}(1/t^p)$ where $0 < p \leq 1$. However, the authors of both [19] and [12] do not discuss about the optimal stepsizes among all proposed stepsizes which is what is done in this paper.

Outline: The paper is organized as follows. Section 2 describes a class of strongly convex and smooth objective functions which is used to derive the lower bound. We verify our theory by experiments in Section 3. Suplementary material A comprehensively studies the work in [1]. Section 4 concludes the paper.

2   Lower Bound and Optimal Stepsize for SGD

In this paper, we consider the following extended problem of SGD: When constructing a sequence of stepsizes η_t, we do not only have access to μ and L; in addition we also have access to N, access to the full gradient $\nabla F(w_t)$ in the t-th iteration, and access to an oracle that knows Y_t, and in the t-th iteration. Notice that this allows adaptively constructed η_t into some extend and our lower bound will hold for this more general case.

Note that the construction of η_t as analyzed in this paper does **not** depend on knowledge of the stochastic gradient $\nabla f(w_t; \xi_t)$. So, we do not consider step sizes that are adaptively computed based on $\nabla f(w_t; \xi_t)$.

We study the best lower bound on the expected convergence rate for any possible sequence of stepsizes η_t that satisfy the requirements given above in the extended SGD setting.

In order to prove a lower bound we propose a specific class of strongly convex and smooth objective function F and we show in the extended SGD setting how to compute the *optimal* step size η_t as a function of $\mu, L, N, \nabla F(w_t)$, and an oracle with access to Y_t. We will show that the optimal stepsize η_t is based on $\mu, L, N, Y_0, \ldots, Y_t$. For completeness, as in Algorithm 1, the next w_{t+1} is defined as $w_{t+1} = w_t - \eta_t \nabla f(w_t; \xi_t)$.

We consider the following class of objective functions F: We consider a multivariate normal distribution of a d-dimensional random vector ξ, i.e., $\xi \sim \mathcal{N}(m, \Sigma)$, where $m = \mathbb{E}[\xi]$ and $\Sigma = \mathbb{E}[(\xi - m)(\xi - m)^{\mathrm{T}}]$ is the (symmetric positive semi-definite) covariance matrix. The density function of ξ is chosen as

$$g(\xi) = \frac{\exp\left(\frac{-(\xi - m)^{\mathrm{T}}\Sigma^{-1}(\xi - m)}{2}\right)}{\sqrt{(2\pi)^d |\Sigma|}}.$$

We select component functions

$$f(w; \xi) = s(\xi)\frac{\|w - \xi\|^2}{2},$$

where function $s(\xi)$ is constructed *a-priori* according to the following random process:

5

- With probability $1 - \mu/L$, we draw $s(\xi)$ from the uniform distribution over interval $[0, \mu/(1 - \mu/L)]$.

- With probability μ/L, we draw $s(\xi)$ from the uniform distribution over interval $[0, L]$.

The following theorem analyses the sequence of optimal step sizes for our class of objective functions and gives a lower bound on the corresponding expected convergence rates. The theorem states that we cannot find a better sequence of step sizes. In other words without any more additional information about the objective function (beyond $\mu, L, N, Y_0, \ldots, Y_t$ for computing η_t), we can at best prove a general upper bound which is at least the lower bound as stated in the theorem. As explained in the introduction an upper bound which is only a factor 32 larger than the theorem's lower bound exists.

As a disclaimer we notice that for some objective functions the expected convergence rate can be much better than what is stated in the theorem: This is due to the specific nature of the objective function itself. However, without knowledge about this nature, one can only prove a general upper bound on the expected convergence rate Y_t and any such upper bound must be at least the lower bound as proven in the next theorem. Therefore, as a conclusion of the theorem we infer that only if more/other information is used for adaptively computing η_t, then it may be possible to derive stronger upper bounds that beat the lower bound of the theorem.

Theorem 1. *We assume that component functions $f(w; \xi)$ are constructed according to the recipe described above. Then, the corresponding objective function is μ-strongly convex and the component functions are L-smooth and convex with $\mu < L$.*

If we run Algorithm 1 and assume that an oracle accessing $Y_t = \mathbb{E}[\|w_t - w_*\|^2]$ is given at the t-th iteration (our extended SGD problem setting), then an exact expression for the optimal sequence of stepsizes η_t (see (12)) based on $\mu, L, N, Y_0, \ldots, Y_t$ can be given. For this sequence of stepsizes,

$$Y_t \geq \frac{N}{2\mu} \frac{1}{W}, \tag{9}$$

and for $t \geq T' = \frac{20L}{\mu}$,

$$Y_t \leq \frac{16N}{\mu} \frac{1}{\mu(t - \frac{20L}{\mu}) + 4L}, \tag{10}$$

where

$$W = \mu t + \frac{\mu(L + \mu/(1 - \mu/L))}{12} + \mu[\ln(t + L\frac{1 + \mu/(L(1 - \mu/L))}{12}) - \ln(L\frac{1 + \mu/(L(1 - \mu/L))}{12})].$$

Proof. Clearly, $f(w; \xi)$ is $s(\xi)$-smooth where the maximum value of $s(\xi)$ is equal to L. That is, all functions $f(w; \xi)$ are L-smooth (and we cannot claim a smaller smoothness parameter). We notice that

$$\mathbb{E}_\xi[s(\xi)] = (1 - \mu/L)\frac{\mu/(1 - \mu/L)}{2} + (\mu/L)\frac{L}{2} = \mu$$

and

$$\mathbb{E}_\xi[s(\xi)^2] = (1 - \mu/L)\frac{(\mu/(1 - \mu/L))^2}{12} + (\mu/L)\frac{L^2}{12} = \frac{\mu(L + \frac{\mu}{1-\mu/L})}{12}.$$

6

With respect to $f(w; \xi)$ and distribution $g(\xi)$ we define

$$F(w) = \mathbb{E}_\xi[f(w; \xi)] = \mathbb{E}_\xi[s(\xi)\frac{\|w - \xi\|^2}{2}].$$

Since $s(\xi)$ only assigns a random variable to ξ which is drawn from a distribution whose description is not a function of ξ, random variables $s(\xi)$ and ξ are statistically independent. Therefore,

$$F(w) = \mathbb{E}_\xi[s(\xi)\frac{\|w - \xi\|^2}{2}] = \mathbb{E}_\xi[s(\xi)]\mathbb{E}_\xi[\frac{\|w - \xi\|^2}{2}]$$
$$= \mu\mathbb{E}_\xi[\frac{\|w - \xi\|^2}{2}]$$

Notice:

1. $\|w - \xi\|^2 = \|(w - m) + (m - \xi)\|^2 = \|w - m\|^2 + 2\langle w - m, m - \xi\rangle + \|m - \xi\|^2$.

2. Since $m = \mathbb{E}[\xi]$, we have $\mathbb{E}[m - \xi] = 0$.

3. $\mathbb{E}[\|m - \xi\|^2] = \sum_{i=1}^d \mathbb{E}[(m_i - \xi_i)^2] = \sum_{i=1}^d \Sigma_{i,i} = \text{Tr}(\Sigma)$.

Therefore,

$$F(w) = \mu\mathbb{E}_\xi[\frac{\|w - \xi\|^2}{2}] = \mu\frac{\|w - m\|^2}{2} + \mu\frac{\text{Tr}(\Sigma)}{2},$$

and this shows F is μ-strongly convex and has minimum

$$w_* = m.$$

Since

$$\nabla_w[\|w - \xi\|^2] = \nabla_w[\langle w, w\rangle - 2\langle w, \xi\rangle + \langle \xi, \xi\rangle]$$
$$= 2w - 2\xi = 2(w - \xi),$$

we have

$$\nabla_w f(w; \xi) = s(\xi)(w - \xi).$$

In our notation

$$N = 2\mathbb{E}_\xi[\|\nabla f(w_*; \xi)\|^2] = 2\mathbb{E}_\xi[s(\xi)^2\|w_* - \xi\|^2].$$

By using similar arguments as used above we can split the expectation and obtain

$$N = 2\mathbb{E}_\xi[s(\xi)^2\|w_* - \xi\|^2] = 2\mathbb{E}_\xi[s(\xi)^2]\mathbb{E}_\xi[\|w_* - \xi\|^2].$$

We already calculated

$$\mathbb{E}_\xi[\|w_* - \xi\|^2] = \|w_* - m\|^2 + \text{Tr}(\Sigma) = \text{Tr}(\Sigma)$$

and we know

$$\mathbb{E}_\xi[s(\xi)^2] = \frac{\mu(L + \frac{\mu}{1 - \mu/L})}{12}.$$

This yields

$$N = 2\mathbb{E}_\xi[s(\xi)^2]\mathbb{E}_\xi[\|w_* - \xi\|^2] = \frac{\mu(L + \frac{\mu}{1-\mu/L})}{6}\mathrm{Tr}(\Sigma).$$

In the SGD algorithm we compute

$$\begin{aligned}
w_{t+1} &= w_t - \eta_t \nabla f(w_t; \xi_t) \\
&= w_t - \eta_t s(\xi_t)(w_t - \xi_t) \\
&= (1 - \eta_t s(\xi_t))w_t + \eta_t s(\xi_t)\xi_t.
\end{aligned}$$

We choose w_0 according to the following computation: We draw ξ from its distribution and apply full gradient descent in order to find w_0 which minimizes $f(w; \xi)$ for w. Since

$$f(w; \xi) = s(\xi)\frac{\|w - \xi\|^2}{2},$$

the minimum is achieved by $w_0 = \xi$. Therefore,

$$Y_0 = \mathbb{E}[\|w_0 - w_*\|^2] = \mathbb{E}[\|\xi - w_*\|^2] = \mathrm{Tr}(\Sigma).$$

Let $\mathcal{F}_t = \sigma(w_0, \xi_0, \dots, \xi_{t-1})$ be the σ-algebra generated by $w_0, \xi_0, \dots, \xi_{t-1}$. We derive

$$\mathbb{E}[\|w_{t+1} - w_*\|^2|\mathcal{F}_t] = \mathbb{E}[\|(1 - \eta_t s(\xi_t))(w_t - w_*) + \eta_t s(\xi_t)(\xi_t - w_*)\|^2|\mathcal{F}_t]$$

which is equal to

$$\mathbb{E}[(1 - \eta_t s(\xi_t))^2\|w_t - w_*\|^2 + 2\eta_t s(\xi_t)(1 - \eta_t s(\xi_t))\langle w_t - w_*, \xi_t - w_*\rangle + \eta_t^2 s(\xi_t)^2\|\xi_t - w_*\|^2|\mathcal{F}_t]. \tag{11}$$

Given $\mathcal{F}_t$, w_t is not a random variable. Furthermore, we can use linearity of taking expectations and as above split expectations:

$$\mathbb{E}[(1 - \eta_t s(\xi_t))^2]\|w_t - w_*\|^2 + \mathbb{E}[2\eta_t s(\xi_t)(1 - \eta_t s(\xi_t))]\langle w_t - w_*, \mathbb{E}[\xi_t - w_*]\rangle + \mathbb{E}[\eta_t^2 s(\xi_t)^2]\mathbb{E}[\|\xi_t - w_*\|^2].$$

Again notice that $\mathbb{E}[\xi_t - w_*] = 0$ and $\mathbb{E}[\|\xi_t - w_*\|^2] = \mathrm{Tr}(\Sigma)$. So, $\mathbb{E}[\|w_{t+1} - w_*\|^2|\mathcal{F}_t]$ is equal to

$$\begin{aligned}
&\mathbb{E}[(1 - \eta_t s(\xi_t))^2]\|w_t - w_*\|^2 + \eta_t^2\frac{N}{2} \\
= \quad &(1 - 2\eta_t\mu + \eta_t^2\frac{\mu(L + \frac{\mu}{1-\mu/L})}{12})\|w_t - w_*\|^2 + \eta_t^2\frac{N}{2} \\
= \quad &(1 - \mu\eta_t(2 - \eta_t(L + \frac{\mu}{1-\mu/L})/12))\|w_t - w_*\|^2 + \eta_t^2\frac{N}{2}.
\end{aligned}$$

In terms of $Y_t = \mathbb{E}[\|w_t - w_*\|^2]$, by taking the full expectation (also over $\mathcal{F}_t$) we get

$$Y_{t+1} = (1 - \mu\eta_t(2 - \eta_t(L + \frac{\mu}{1-\frac{\mu}{L}})/12))Y_t + \eta_t^2\frac{N}{2}.$$

This is very close to recurrence (4).

The optimal step size in this case is solved by taking the derivative with respect to η_t. The derivative is equal to

$$-2\mu Y_t + 2\eta_t \mu (L + \frac{\mu}{1 - \frac{\mu}{L}})Y_t/12 + N\eta_t \tag{12}$$

which shows the minimum is achieved for

$$\eta_t = \frac{2\mu Y_t}{N + \mu(L + \frac{\mu}{1 - \frac{\mu}{L}})Y_t/6} \tag{13}$$

giving

$$Y_{t+1} = Y_t - \frac{2\mu^2 Y_t^2}{N + \mu(L + \frac{\mu}{1 - \frac{\mu}{L}})Y_t/6}$$

$$= Y_t - \frac{2\mu^2 Y_t^2}{N(1 + Y_t/\text{Tr}(\Sigma))}. \tag{14}$$

We note that $Y_{t+1} \leq Y_t$ for any $t \geq 0$. We proceed by proving a lower bound on Y_t. Clearly,

$$Y_{t+1} \geq Y_t - \frac{2\mu^2 Y_t^2}{N} \tag{15}$$

Let us define $\gamma = 2\mu^2/N$. We can rewrite (15) as follows:

$$\gamma Y_{t+1} \geq \gamma Y_t (1 - \gamma Y_t) \text{ or}$$

$$(\gamma Y_{t+1})^{-1} \leq 1 + (\gamma Y_t)^{-1} + \frac{1}{(\gamma Y_t)^{-1} - 1}.$$

In order to make the inequality above correct, we require $1 - \gamma Y_t > 0$ for any $t \geq 0$. Since $Y_{t+1} \leq Y_t$, we only need $Y_0 < \frac{1}{\gamma}$. This means

$$\text{Tr}(\Sigma) < \frac{N}{2\mu^2} \text{ or } \text{Tr}(\Sigma) < \frac{\mu(L + \frac{\mu}{1 - \mu/L})}{6} \frac{\text{Tr}(\Sigma)}{2\mu^2}.$$

This is equivalent to $\mu < 3(L + \frac{\mu}{1 - \frac{\mu}{L}})$ which is obviously true since $\mu < L$.

This implies

$$(\gamma Y_{t+1})^{-1} \leq (t+1) + (\gamma Y_0)^{-1} + \sum_{i=0}^{t} \frac{1}{(\gamma Y_i)^{-1} - 1}.$$

Since

$$\sum_{i=0}^{t} \frac{1}{(\gamma Y_i)^{-1} - 1} \leq \ln(t + 1 + (\gamma Y_0)^{-1}) - \ln((\gamma Y_0)^{-1}),$$

we have the following inequality:

$$(\gamma Y_{t+1})^{-1} \leq (t+1) + (\gamma Y_0)^{-1} + \ln(t + 1 + (\gamma Y_0)^{-1}) - \ln((\gamma Y_0)^{-1}.$$

Reordering, substituting $\gamma = 2\mu^2/N$, and replacing $t+1$ by t yields, for $t \geq 0$, the lower bound

$$
\begin{aligned}
Y_t &\geq \frac{N}{2\mu} \frac{1}{\mu t + N/(2Y_0) + \mu[\ln(t + \frac{N}{2\mu Y_0}) - \ln(\frac{N}{2\mu Y_0})]} \\
&= \frac{N}{2\mu} \frac{1}{W},
\end{aligned}
$$

where

$$
W = \mu t + \frac{\mu(L + \mu/(1 - \mu/L))}{12} + \mu[\ln(t + L\frac{1 + \mu/(L(1 - \mu/L))}{12}) - \ln(L\frac{1 + \mu/(L(1 - \mu/L))}{12})].
$$

The upper bound of Y_t comes from the following fact. If we run Algorithm 1 with stepsize $\eta_t' = \frac{2}{\mu t + 4L}$ for $t \geq 0$ in [16], then we have from [16] an expected convergence rate

$$
Y_t' \leq \frac{16N}{\mu} \frac{1}{\mu(t - T') + 4L},
$$

where

$$
T' = \frac{4L}{\mu} \max\{\frac{L\mu Y_0}{N}, 1\} - \frac{4L}{\mu}.
$$

Substituting

$$
N = \frac{\mu(L + \frac{\mu}{1 - \mu/L})}{6} \mathrm{Tr}(\Sigma) \text{ and } Y_0 = \mathrm{Tr}(\Sigma)
$$

yields $T' = \frac{20L}{\mu}$. Since η_t is the most optimal stepsize and η_t' is not, $Y_t \leq Y_t'$. I.e., we have

$$
Y_t \leq \frac{16N}{\mu} \frac{1}{\mu(t - \frac{20L}{\mu}) + 4L}.
$$

$\square$

Corollary 1. *Given the class of objective functions analyzed in Theorem 1, we run Algorithm 1 and assume an oracle with access to $Y_t = \mathbb{E}[\|w_t - w_*\|^2]$ as well as the **full gradient** $\nabla F(w_t)$ at the t-th iteration. An exact expression for the optimal sequence of stepsizes η_t based on μ, L, N and this extended oracle can be given. For this sequence of stepsizes, the same lower and upper bounds on the expected convergence rate Y_t as in Theorem 1 hold.*

Proof. The proof of this corollary is directly derived from the reason why we are allowed to transform (11) into (12), i.e., η_t and ξ_t must be independent to get (12) from (11). If the construction of η_t does not depend on ξ_t (or $\nabla f(w_t; \xi_t)$), then only Y_t is required to construct the optimal stepsize η_t. It implies that the information of $\nabla F(w_t)$ is not useful and we can borrow the proof of Theorem 1 to arrive at the result of this corollary. $\square$

Let us consider the set of all possible objective functions $\mathcal{F}$ which are μ-strongly convex and L-smooth. For an objective function $F \in \mathcal{F}$, let $\gamma_t^F(\mu, L, \mathcal{U})$ be defined as the smallest expected convergence rate Y_t that can be achieved by a stepsize construction $\{\eta_t\}$, where $\{\eta_t\}$ is computed as a function h of μ, L and oracle $\mathcal{U}$ with access to N, Y_t and $\nabla F(w_t)$ at the t-th iteration. That is,

$$
\gamma_t^F(\mu, L, \mathcal{U}) = \inf_{\{\eta_t\} \in h(\mu, L, \mathcal{U})} Y_t(F, \{\eta_t\}),
$$

where Y_t is explicitly shown as a function of the objective function and sequence of step sizes.

Among the objective functions $F \in \mathcal{F}$, we consider objective function F which has the **worst** expected convergence rate at the t-th iteration. Let us denote the expected convergence rate Y_t that corresponds to the worst objective function as $\gamma_t(\mu, L, \mathcal{U})$. Precisely,

$$\gamma_t(\mu, L, \mathcal{U}) = \sup_{F \in \mathcal{F}} \gamma_t^F(\mu, L, \mathcal{U})$$
$$= \sup_{F \in \mathcal{F}} \inf_{\{\eta\} \in h(\mu, L, \mathcal{U})} Y_t[F, \eta].$$

The lower bound and upper bound of $\gamma_t(\mu, L, \mathcal{U})$ is stated in Corollary 2.

Corollary 2. *Given μ, L and oracle $\mathcal{U}$ with access to N, Y_t and $\nabla F(w_t)$ at the t-th iteration, the convergence rate of the worst strongly convex and smooth objective function F with optimal stepsize η_t based on μ, L and $\mathcal{U}$ is $\gamma_t(\mu, L, \mathcal{U})$. The expected convergence rate $\gamma_t(\mu, L, \mathcal{U})$ satisfies the same lower bounds on the expected convergence rate as in Theorem 1 where Y_t is substituted by $\gamma_t(\mu, L, \mathcal{U})$. As an upper bound we have*

$$\gamma_t(\mu, L, \mathcal{U}) \leq \frac{16N}{\mu} \frac{1}{\mu(t - T') + 4L},$$

where

$$T' = \frac{4L}{\mu} \max\{\frac{L\mu Y_0}{N}, 1\} - \frac{4L}{\mu}$$

for

$$h(\mu, L, \mathcal{U}) = \{\eta_t = \frac{2}{\mu t + 4L}\}.$$

Notice that scheme h for constructing step sizes is independent of oracle $\mathcal{U}$, in other words its knowledge is not needed.

Proof. Due to the definition of $\gamma_t(\mu, L, \mathcal{U})$, it is always larger than $\gamma_t^F(\mu, L, \mathcal{U})$ for all $F \in \mathcal{F}$. From Corollary 1 we infer that $\gamma_t^F(\mu, L, \mathcal{U})$ is larger than the lower bound $\frac{N}{2\mu} \frac{1}{W}$ as specified in Theorem 1. Since this holds for all $F \in \mathcal{F}$, it also holds for the supremum over $F \in \mathcal{F}$.

The upper bound follows from the result in [16], i.e., for any given F and $\eta_t = \frac{2}{\mu t + 4L}$, we have

$$Y_t \leq \frac{16N}{\mu} \frac{1}{\mu(t - T') + 4L},$$

where

$$T' = \frac{4L}{\mu} \max\{\frac{L\mu Y_0}{N}, 1\} - \frac{4L}{\mu}.$$

$\square$

The importance of Corollary 2 is that, for the **worst** objective function, we can now compute the gap between the lower bound and the upper bound, i.e., they are separated by a factor of 32. This implies that no scheme for constructing a sequence of stepsizes η_t that is based on μ, L, and oracle $\mathcal{U}$ can achieve a better expected convergence rate than $\mathcal{O}(1/t)$. The only way to achieve a better expected convergence rate is to use a scheme that has access to information beyond what is given by μ, L, and oracle $\mathcal{U}$. For example, if the

construction of η_t must depend on the information of the full gradient $\nabla F(w_t)$ as well as the stochastic gradient $\nabla f(w_t, \xi_t)$, or we have to develop a new updating form to replace the updating form of SGD (i.e. $w_{t+1} = w_t - \eta_t \nabla f(w_t; \xi_t)$).

Corollary 2 shows that the lower bound and the upper bound of Y_t are $\mathcal{O}(1/t)$ (see (9)). Furthermore, it offers a general strategy $\eta_t = \frac{2}{\mu t + 4L}$ for computing step sizes which only depends on μ and L in order to realize the upper bound (which comes within a factor 32 of the lower bound). This means that we can finally conclude that there does not exist a significantly better construction for step sizes than $\eta_t = \frac{2}{\mu t + 4L}$ for classical SGD (not our extended SGD problem).

3 Numerical Experiments

We verify our theory by considering simulations with different values of sample size n (1000, 10000, and 100000) and vector size d (10, 100, and 1000). First, we generate n vectors of size d with mean m and positive definite covariance matrix Σ. To be simple, we generate $m \in \mathbb{R}^d$ and diagonal matrix $\Sigma \in \mathbb{R}^{d \times d}$ with uniform at random in $[0, 1]$ for each element in m and each element in the diagonal of Σ. We experimented with 10 runs and reported the average results.
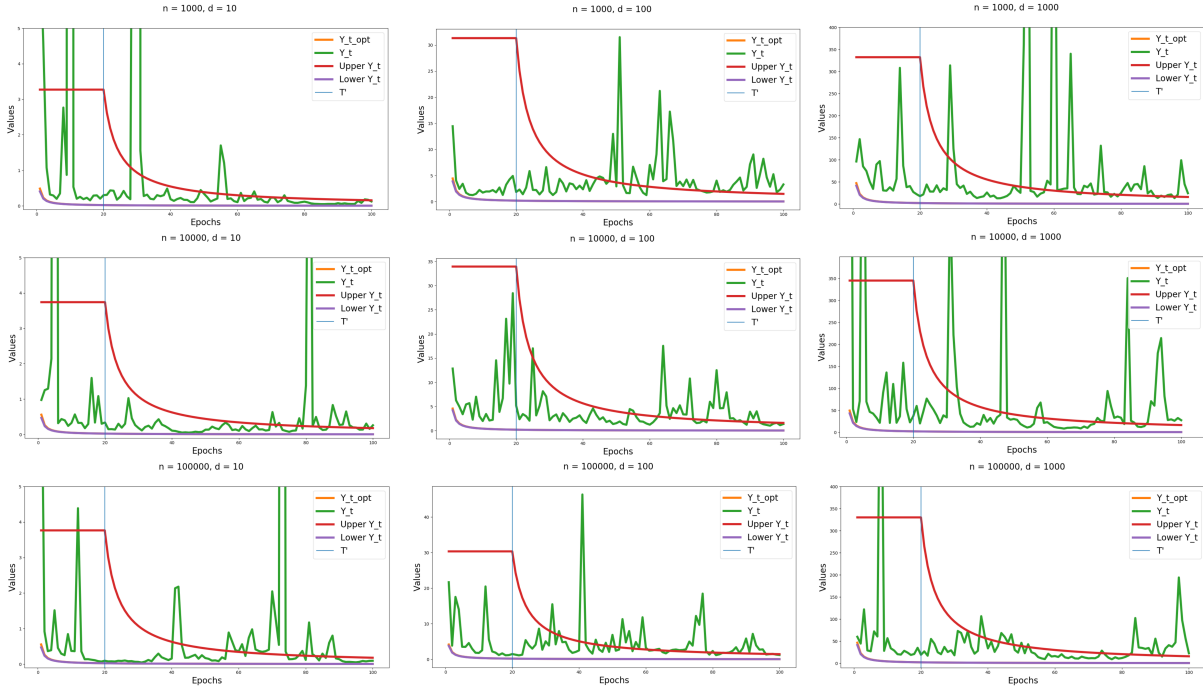

Figure 1: Y_t and its upper and lower bounds

We denote the labels "Upper Y_t" (red line) "Lower Y_t" (violet line) in Figure 1 as the upper and lower bounds of Y_t in (10) and (9), respectively; "Y_t_opt" (orange line) as Y_t defined in Theorem 1 with the given information from the oracle; "Y_t" (green) as the squared norm of the difference between w_t and w_*, where w_t generated from Algorithm 1 with learning rate in (13). We note that "Lower Y_t" and "Y_t_opt" are very close to each other in Figure 1 and the difference between them is shown in Figure 2. Note that Y_t in Figure 1 is computed as average of 10 runs of $\|w_t - w_*\|^2$ (not exactly $\mathbb{E}[\|w_t - w_*\|^2]$).
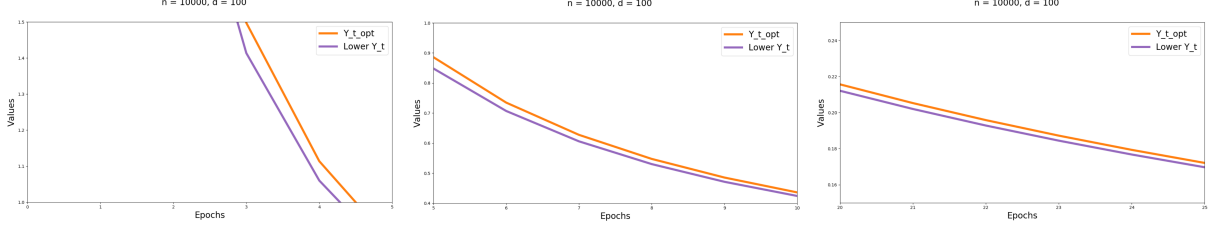
Figure 2: The difference between "Lower Y_t" and "Y_t_opt" ($n = 10000$, $d = 100$)

Discussion: We have a vertical line at epoch 20 because we expect to see the upper bound in (10) to take effect when $t \geq T' = \frac{20L}{\mu}$. The "Upper Y_t" (red line), "Lower Y_t" (violet line) and "Y_t_opt" (orange line) do not oscillate because they can be correctly computed using formulas (10), (9) and (14), respectively, i.e., all the lines do not have variation. The green line "Y_t" for stepsize $\eta_t = \frac{2}{\mu t + 4L}$ in Figure 1 oscillates because in our analysis we do not consider the variance of Y_t. As shown in (4), we have

$$\mathbb{E}[\|w_{t+1} - w_*\|^2] \leq (1 - \mu\eta_t)\mathbb{E}[\|w_t - w_*\|^2] + \eta_t^2 N.$$

It is clear that a decrease in η_t leads to a decrease of the variance of Y_t (i.e. $\eta_t^2 N$). This fact is reflected in all subfigures in Figure 1. We expect that increasing d and n (the number of dimensions in data and the number of data points) would increase the variance. Hence, we see that it requires larger t to make the variance approach 0 as shown in Figure 1. We can see that when t is sufficiently large, then optimality of $\eta_t = \frac{2}{\mu t + 4L}$ is clearly shown in Figure 1 when $n = 1000$ and $d = 10$, i.e., the green line is in between red line (upper bound) and violet line (lower bound). Moreover, these two bounds are pretty close to each other when t is sufficient large.

4   Conclusion

In this paper, we study the convergence of SGD. We show that for **any** given stepsize η_t constructed based on μ, L, N, and an oracle with access to Y_t and $\nabla F(w_t)$ at the t-th iteration, the best possible lower bound of the convergence is $\mathcal{O}(1/t)$. Note that this extends classical SGD where only μ and L are given for construction of η_t. This result implies that the best possible lower bound of the convergence rate for any possible stepsize η_t based on μ and L is $\mathcal{O}(1/t)$. This result confirms the optimality of the proposed stepsize $\eta_t = \frac{2}{\mu t + 4L}$ for $t \geq 0$ in [16]. Compared to the result in [1], our proposed class of objective functions is simple and does not require many assumptions for the sake of proof. Also our lower bound is orders of magnitude more tight as it is the first lower bound to be independent of dimension d. In addition, [1] does not study the lower bound of the extended problem of SGD.

References

[1] Alekh Agarwal, Peter L Bartlett, Pradeep Ravikumar, and Martin J Wainwright. Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization. 2010.

[2] D.P. Bertsekas. *Nonlinear Programming*. Athena Scientific, 1999.

[3] Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *arXiv:1606.04838*, 2016.

[4] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[5] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, New York, NY, USA, 1991.

[6] Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *NIPS*, pages 1646–1654, 2014.

[7] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics, 2nd edition, 2009.

[8] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *NIPS*, pages 315–323, 2013.

[9] Nicolas Le Roux, Mark Schmidt, and Francis Bach. A stochastic gradient method with an exponential convergence rate for finite training sets. In *NIPS*, pages 2663–2671, 2012.

[10] Rémi Leblond, Fabian Pederegosa, and Simon Lacoste-Julien. Improved asynchronous parallel optimization analysis for stochastic incremental methods. *arXiv preprint arXiv:1801.03749*, 2018.

[11] Lucien LeCam et al. Convergence of estimates under dimensionality restrictions. *The Annals of Statistics*, 1(1):38–53, 1973.

[12] Eric Moulines and Francis R Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems*, pages 451–459, 2011.

[13] Arkadii Semenovich Nemirovsky and David Borisovich Yudin. Problem complexity and method efficiency in optimization. 1983.

[14] Yurii Nesterov. *Introductory lectures on convex optimization : a basic course*. Applied optimization. Kluwer Academic Publ., Boston, Dordrecht, London, 2004.

[15] Lam M. Nguyen, Jie Liu, Katya Scheinberg, and Martin Takáč. SARAH: A novel method for machine learning problems using stochastic recursive gradient. In *ICML*, 2017.

[16] Lam M. Nguyen, Phuong Ha Nguyen, Marten van Dijk, Peter Richtarik, Katya Scheinberg, and Martin Takac. SGD and hogwild! Convergence without the bounded gradients assumption. In *ICML*, 2018.

[17] Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer, New York, 2nd edition, 2006.

[18] Maxim Raginsky and Alexander Rakhlin. Information-Based Complexity, Feedback and Dynamics in Convex Programming. *IEEE Trans. Information Theory*, 57(10):7036–7056, 2011.

[19] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951.

[20] Marten van Dijk, Lam Nguyen, Phuong Ha Nguyen, and Dzung Phan. Characterization of Convex Objective Functions and Optimal Expected Convergence Rates for SGD. *arXiv preprint*, 2018.

[21] Bin Yu. Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer, 1997.

A Related Work

In [1], the authors showed that the lower bound of Y_t is $\mathcal{O}(1/t)$ *with* bounded gradient assumption for objective function F over a convex set $\mathcal{S}$. To show the lower bound, the authors use the following three assumptions for the objective function F:

1. The assumption of a strongly convex objective function, i.e., Assumption 1 (see Definition 3 in [1]).

2. There exists a bounded convex set $\mathcal{S} \subset \mathbb{R}^d$ such that

$$\mathbb{E}[\|\nabla f(w;\xi)\|^2] \leq \sigma^2$$

   for all $w \in \mathcal{S} \subset \mathbb{R}^d$ (see Definition 1 in [1]). Notice that this is not the same as the bounded gradient assumption where $\mathcal{S} = \mathbb{R}^d$ is unbounded.

3. The objective function F is a convex Lipschitz function, i.e., there exists a positive number K such that

$$\|F(w) - F(w')\| \leq K\|w - w'\|, \forall w, w' \in \mathcal{S} \subset \mathbb{R}^d.$$

   We notice that this assumption actually implies the assumption on bounded gradients as stated above.

On the existence of the assumption of bounded convex set $\mathcal{S} \subset \mathbb{R}^d$ where SGD converges: let us restate the example in [16], i.e. $F(w) = \frac{1}{2}(f_1(w) + f_2(w))$ where $f_1(w) = \frac{1}{2}w^2$ and $f_2(w) = w$. It is obvious that F is strongly convex but f_1 and f_2 are not. Let $w_0 = 0 \in \mathcal{S}$, for any number $t \geq 0$, with probability $\frac{1}{2^t}$, the steps of SGD algorithm for all $i < t$ are $w_{i+1} = w_i - \eta_i$. This implies that $w_t = -\sum_{i=1} \eta_i$. Since $\sum_{i=1} \eta_i = \infty$, w_t will escape the set $\mathcal{S}$ when t is sufficiently large. Hence, we have the following conclusion $\mathcal{S}$ must be unbounded, e.g., $\mathcal{S} = \mathbb{R}^d$ if there is at least one component function f of F which is not strongly convex. This result implies that $\mathcal{S}$ is a pure subset of $\mathbb{R}^d$ if and only if all component functions f of F are strongly convex.

If $\mathcal{S}$ is $\mathbb{R}^d$, we have the following results:

On the non-coexistence of the assumption of a bounded gradient over $\mathbb{R}^d$ and assumption of having strong convexity: As pointed out in [16], the assumption of bounded gradient does not co-exist with strongly convex assumption. As shown in [16], for any $w \in \mathbb{R}^d$, we have

$$2\mu[F(w) - F(w_*)] \overset{(6)}{\leq} \|\nabla F(w)\|^2 = \|\mathbb{E}[\nabla f(w;\xi)]\|^2$$
$$\leq \mathbb{E}[\|\nabla f(w;\xi)\|^2] \leq \sigma^2.$$

Therefore,

$$F(w) \leq \frac{\sigma^2}{2\mu} + F(w_*), \forall w \in \mathbb{R}^d.$$

Note that, the from Assumption 1 and $\nabla F(w_*) = 0$, we have

$$F(w) \geq \mu\|w - w_*\|^2 + F(w_*), \forall w \in \mathbb{R}^d.$$

Clearly, the two last inequalities contradict to each other for sufficiently large $\|w - w_*\|^2$. Precisely, only when σ is equal to ∞, then the assumption of bounded gradient and the assumption of strongly convexity

of F can co-exist. However, σ cannot be ∞ and this result implies that there does not exist any objective function F satisfies the assumption of bounded gradients over $\mathbb{R}^d$ and the assumption of having a strongly convex objective function at the same time.

On the non-coexistence of the assumption of being convex Lipschitz over $\mathbb{R}^d$ and assumption of being strongly convex: Moreover, we can also show that the assumption of convex Lipschitz function does not co-exist with the assumption of being strongly convex. As shown in Section 2.3 in [1], the assumption of Lipschitz function implies that $\|\nabla F(w)\| \leq K, \forall w \in R^d$. Hence, by using the same argument from the analysis of the non-coexistence of bounded gradient assumption and assumption of strongly convex, we can conclude that these two assumptions cannot co-exist. In other words, there does not exist an objective function F which satisfies the assumption of convex Lipschitz function and assumption of being strongly convex at the same time.

A.1   Discussion on the usage of Assumptions in [1]

As stated in Section 3 and Section 4.1.1 in [1], the authors construct a class of strongly convex Lipschitz objective function F which has $K = \sigma$. The authors showed that the problem of convex optimization for the constructed class of objective functions F is at least as hard as estimating the biases of d independent coins (i.e., the problem of estimating parameters of Bernoulli variables). As one additional important assumption to prove the lower bound of the SGD algorithm, the authors assume the **existence** of stepsizes η_t which make the SGD algorithm converge for a given objective function F under the three aforementioned assumptions (see Lemma 2 in [1]). Note that the proof of the lower bound of Y_t of SGD is described in Theorem 2 in [1] and Theorem 2 uses their Lemma 2. If their Lemma 2 is not valid, then the proof of the lower bound of Y_t in Theorem 2 is also not valid.

Actually, the authors in [1] do not require all the component function of objective function F to be strongly convex. Given the proof strategy of the convergence of SGD, one may require that the convex set $\mathcal{S}$ where F has all these nice properties must be $\mathbb{R}^d$ as explained above. This, however, will lead to the non-coexistence of bounded gradient assumption and strongly convex assumption and the non-coexistence of Lipschitz function assumption and strongly convex assumption as discussed above. In this case, their Lemma 2 is not valid because of non-existence of an objective function F, in which case the proof of lower bound of Y_t in Theorem 2 is not correct.

However, we explain why the setup as proposed in [1] is still acceptable and lead to a proper lower bound: The paper assumes that we only restrict the analysis of SGD in a bounded convex set $\mathcal{S}$ which is not $\mathbb{R}^d$, and only in this bounded set $\mathcal{S}$ we assume that objective function acts like a Lipschitz function (implying bounded gradients in $\mathcal{S}$).

There are two possible cases at the t-th iteration of SGD algorithm, the algorithm diverges or converges. Let us define $p = Pr(w_t \notin \mathcal{S})$. Hence, $Pr(w_t \in \mathcal{S}) = 1 - p$. Let

$$Y_t^{conv} = \mathbb{E}[\|w_t - w_*\|^2 | w_t \in \mathcal{S}]$$

and

$$Y_t^{div} = \mathbb{E}[\|w_t - w_*\|^2 | w_t \notin \mathcal{S}].$$

Since $Y_t = \mathbb{E}[\|w_t - w_*\|^2$, Y_t is equal to

$$
\begin{aligned}
Y_t &= p \cdot Y_t^{div} + (1-p) \cdot Y_t^{conv} \\
&\geq p \cdot Y_t^{conv} + (1-p) \cdot Y_t^{conv} \\
&\geq Y_t^{conv} \\
&\geq \text{ lower bound in [1].}
\end{aligned}
$$

The above derivation hinges on the first inequality where we assume $Y_t^{div} \geq Y_t^{conv}$. Typically, it is always true that $Y_t^{div} \geq Y_t^{conv}$ because w_t gets far from w_* for the divergence case and it gets close to w_* for the convergence case. Of course a proper proof of this property is still needed in order to rigously complete the argument leading the the lower bound in [1].

The above result is interesting because now we **only** need to prove the convergence of SGD in a certain convex set $\mathcal{S}$ with a certain probability p. This is completely different from the proof of convergence of SGD in the general case as in [12] and [16] where we need to prove it with probability of 1.

We describe the setup of the class of strong convex functions proposed in [1] and then we show that our result is much more tight when compared to their result.

A.2 Setup

As shown in Section 4.1.1 [1], the following two sets are required.

1. Subset $\mathcal{V} \subset \{-1, +1\}^d$ and $\mathcal{V} = \{\alpha^1, \ldots, \alpha^M\}$ with $\Delta_H(\alpha^j, \alpha^k) \geq \frac{d}{4}$ for all $j \neq k$, where Δ_H denotes the Hamming metric, i.e $\Delta_H(\alpha, \beta) := \sum_{i=1}^d \mathbb{I}[\alpha_i \neq \beta_i]$. As discussed by the authors, $|\mathcal{V}| = M \geq (2/\sqrt{e})^{\frac{d}{2}}$.

2. Subset $\mathcal{F}_{base} = \{f_i^+, f_i^-, i = 1, \ldots, d\}$ where f_i^+, f_i^- will be designed depending on the problem at hand.

Given $\mathcal{V}, \mathcal{F}_{base}$ and a constant $\delta \in (0, \frac{1}{4}]$, we define the function class $\mathcal{F}(\delta) := \{F_\alpha, \alpha \in \mathcal{V}\}$ where

$$
F_\alpha(w) := \frac{c}{d} \sum_{i=1}^d \{(1/2 + \alpha_i \delta) f_i^+(w) + (1/2 - \alpha_i \delta) f_i^-(w)\}. \tag{16}
$$

The $\mathcal{F}_{base}$ and constant c are chosen in such a way that $\mathcal{F}(\delta) \subset \mathcal{F}$ where $\mathcal{F}$ is the class of strongly convex objective functions defined over set $\mathcal{S}$ and satisfies all the assumptions as mentioned before. In case $\mathcal{F}$ is the class of strongly convex functions, the key idea to compute the lower bound of SGD proposed in [1] by applying Fano's inequality [21] and Le Cam's bound [5, 11] is as follows: If an SGD algorithm $\mathcal{M}_t$ works well for optimizing a given function $F_{\alpha^*}, \alpha^* \in \mathcal{V}$ with a given oracle $\mathcal{U}$, then there exists a hypothesis test finding $\hat{\alpha}$ such that:

$$
\frac{1}{3} \geq \Pr_{\mathcal{U}}[\hat{\alpha}(\mathcal{M}_t) \neq \alpha] \geq 1 - 2\frac{16dt\delta^2 + \log(2)}{d \log(2/\sqrt{e})}. \tag{17}
$$

17

From (17), we have

$$\frac{16dt\delta^2 + \log(2)}{d\log(2/\sqrt{e})} \approx \frac{16dt\delta^2}{d\log(2/\sqrt{e})} \geq 2/3.$$

Hence,

$$t \geq \frac{\log(2/\sqrt{e})}{48}\frac{1}{\delta^2}. \tag{18}$$

As shown in Section 4.3 [1], to proceed the proof, we set $Y_t = \frac{c\delta^2 r^2}{18(1-\theta)}$. Pluging this ϵ into (18) yields

$$Y_t \geq \frac{1}{t}\frac{\log(2/\sqrt{e})}{864}\frac{cr^2}{1-\theta}. \tag{19}$$

In addition to the proof of the lower bound, we also need to set $c = \frac{Ld}{rd^{1/p}}$ and $\mu^2 = \frac{L}{rd^{1/p}}(1-\theta)$ where $\mathcal{S} = B_\infty(r)$. By substituting c and μ^2 into (19), we obtain:

$$Y_t \geq \frac{1}{t}\frac{\log(2/\sqrt{e})}{864d}\frac{1}{\mu^2}c^2 r^2. \tag{20}$$

To complete the description of the setup in [1], we briefly describe the proposed oracle $\mathcal{U}$ which outputs some information to the SGD algorithm at each iteration for constructing the stepsize η_t. There are two types of oracle $\mathcal{U}$ defined as follows.

1. Oracle $\mathcal{U}_A$: 1-dimensional unbiased gradients

   (a) Pick an index $i \in 1, \ldots, d$ uniformly at random.
   (b) Draw $b_i \in \{0, 1\}$ according to a Bernoulli distribution with parameter $1/2 + \alpha_i\delta$.
   (c) For the given input $x \in \mathcal{S}$, return the value f_i and subgradient ∇f_i of the function

   $$f_{i,A} := c[b_i f_i^+ + (1 - b_i)f_i^-].$$

2. Oracle $\mathcal{U}_B$: d-dimensional unbiased gradients.

   - For $i = 1, \ldots, d$, draw $b_i \in \{0, 1\}$ according to a Bernoulli distribution with parameter $1/2 + \alpha_i\delta$.
   - For the given input $x \in \mathcal{S}$, return the value f_i and subgradient ∇f_i of the function

   $$f_{i,B} := \frac{c}{d}\sum_{i=1}^{d}[b_i f_i^+ + (1 - b_i)f_i^-].$$

A.3   Analysis and Comparison

In this section, we want to compare our lower bound in Corollary 2 (i.e., $\frac{N}{2\mu^2 t}$) with the one in (20) when t is sufficiently large. In order to do this, we need to compute $N = 2\mathbb{E}[\|\nabla f(w^*;\xi)\|^2]$ for the strongly

convex function class proposed in [1]. For the strongly convex case, the authors defined the base functions as follows. Given a parameter $\theta \in [0, 1)$, we have

$$f_i^+(w) = r\theta|w_i + r| + \frac{1-\theta}{4}(w_i + r)^2,$$

$$f_i^-(w) = r\theta|w_i - r| + \frac{1-\theta}{4}(w_i - r)^2,$$

where $w = (w_1, \ldots, w_d)$. Let e_i be $(1/2 + \alpha_i\delta)$. Substituting e_i in (16) yields $F_\alpha(w) = \frac{1}{d}[\sum_{i=1}^d f_{\alpha,i}(w)]$ where $f_{\alpha,i}(w) = c[e_i f_i^+(w) + (1 - e_i)f_i^-(w)]$. Due to the construction of F_α, the definition of $f_{\alpha,i}(w)$ and the construction of oracle $\mathcal{U}_A$ or oracle $\mathcal{U}_B$, w^* of F_α can be found by finding each w_i^* for each $f_{\alpha,i}(w)$ first. Precisely, we have the following cases:

1. $w_i < -r$: we have
   - $f_{\alpha,i}(w) = -r\theta(w_i + r)e_i + \frac{1-\theta}{4}(w_i + r)^2 e_i - r\theta(w_i - r)(1 - e_i) + \frac{1-\theta}{4}(w_i - r)^2(1 - e_i)$.
   - $\nabla f_{\alpha,i}(w) = (1 - \theta)e_i r - \frac{1+\theta}{2}r + \frac{1-\theta}{2}w_i$.
   - $\nabla f_{\alpha,i}(w) = 0$ at $w_i^{-r} = r[1 - 2e_i + \frac{2\theta}{1-\theta}]$.

2. $-r \leq w_i \leq r$: we have
   - $f_{\alpha,i}(w) = r\theta(w_i + r)e_i + \frac{1-\theta}{4}(w_i + r)^2 e_i - r\theta(w_i - r)(1 - e_i) + \frac{1-\theta}{4}(w_i - r)^2(1 - e_i)$.
   - $\nabla f_{\alpha,i}(w) = (1 + \theta)e_i r - \frac{1+\theta}{2}r + \frac{1-\theta}{2}w_i$.
   - $\nabla f_{\alpha,i}(w) = 0$ at $w_i^{[-r,r]} = r\frac{1+\theta}{1-\theta}(1 - 2e_i)$.

3. $r \leq w_i \leq \infty$: we have
   - $f_{\alpha,i}(w) = r\theta(w_i + r)e_i + \frac{1-\theta}{4}(w_i + r)^2 e_i + r\theta(w_i - r)(1 - e_i) + \frac{1-\theta}{4}(w_i - r)^2(1 - e_i)$.
   - $\nabla f_{\alpha,i}(w) = (1 - \theta)e_i r + \frac{3\theta-1}{2}r + \frac{1-\theta}{2}w_i$.
   - $\nabla f_{\alpha,i}(w) = 0$ at $w_i^r = r[1 - 2e_i - 2\frac{\theta}{1-\theta}]$.

Now, we have five important points $w_i^{-r}, w_i^{[-r,r]}, w_i^r, -r$ and r and at these points F_α can be minimum. We consider the following cases

1. $\alpha_i = -1$ and then $e_i = \frac{1}{2} + \alpha_i\delta = \frac{1}{2} - \delta$ where $\delta \in [0, 1/4)$, we have
   - $w_i^{-r} = r[\frac{2\theta}{1-\theta} + 2\delta] > -r$.
   - $w_i^{[-r,r]} = r\frac{1+\theta}{1-\theta}(2\delta)$. In this case $w_i^{[-r,r]}$ may belong $[-r, r]$ or it may be greater than r.
   - $w_i^r = r(2\delta - \frac{2\theta}{1-\theta}) < r$.

   This result implies F_α is minimum at $w_i^* = r$ and $\nabla f_{\alpha,i}(w^*) = cr[(1-\theta)e_i + \theta] = cr[(1-\theta)(1/2 - \delta) + \theta]$. Or it can be minimum at $w_i^{[-r,r]}$ if $w_i^{[-r,r]} \in [-r, r]$ and $\nabla f_{\alpha,i}(w^*) = 0$.

2. $\alpha_i = +1$ and then $e_i = \frac{1}{2} + \alpha_i\delta = \frac{1}{2} + \delta$ where $\delta \in [0, 1/4)$, we have
   - $w_i^{-r} = r[\frac{2\theta}{1-\theta} - 2\delta]$. Since $\frac{2\theta}{1-\theta} - 2\delta > -1$ when $\delta \in [0, 1/4)$ and $\theta \in [0, 1)$. Hence $w_i^{-r} > -r$.

- $w_i^{[-r,r]} = r\frac{1+\theta}{1-\theta}(-2\delta) < 0$. In this case $w_i^{[-r,r]}$ may belong $[-r, r]$ or it may be smaller than $-r$.

- $w_i^r = r(-2\delta - \frac{2\theta}{1-\theta}) < r$.

This result implies F_α is minimum at $w_i^* = -r$ and $\nabla f_{\alpha,i}(w^*) = cr[(1-\theta)e_i - 1] = cr[(1-\theta)(1/2 + \delta) - 1]$. Or it can be minimum at $w_i^{[-r,r]}$ if $w_i^{[-r,r]} \in [-r, r]$ and $\nabla f_{\alpha,i}(w^*) = 0$.

By definition, we have

$$N = 2\mathbb{E}[\|\nabla f_i(w^*)\|^2] = 2\frac{1}{d}\sum_{i=1}^d [e_i\|c\nabla f_i^+(w^*)\|^2 + (1 - e_i)\|c\nabla f_i^-(w^*)\|^2]$$

$$\geq 2\min_i[e_i\|c\nabla f_i^+(w^*)\|^2 + (1 - e_i)\|c\nabla f_i^-(w^*)\|^2]$$

From the analysis above, we have four possible w_i^*, i.e., $-r, r, r\frac{1+\theta}{1-\theta}(-2\delta)$ and $r\frac{1+\theta}{1-\theta}(2\delta)$. If we plug w^* which has $w_i^* = -r$ or $w_i^* = r$, then we have $[e_i\|c\nabla f_i^+(w^*)\|^2 + (1 - e_i)\|c\nabla f_i^-(w^*)\|^2] = (1/2 - \delta)c^2r^2$. For w_i^* which has $w_i^* = r\frac{1+\theta}{1-\theta}(-2\delta)$ or $r\frac{1+\theta}{1-\theta}(2\delta)$, we have $[e_i\|c\nabla f_i^+(w^*)\|^2 + (1 - e_i)\|c\nabla f_i^-(w^*)\|^2] = (1/4 - \delta^2)(1 + \theta)^2c^2r^2$. In all case, it is obvious that

$$N \geq 2\min_i[e_i\|c\nabla f_i^+(w^*)\|^2 + (1 - e_i)\|c\nabla f_i^-(w^*)\|^2] \geq \frac{1}{8}c^2r^2, \forall \delta \in [0, 1/4), \theta \in [0, 1).$$

Since we want to compare our lower bound and the lower bound in [1], we need to compare our lower bound with the possibly largest bound of the lower bound in [1]. So, we give an advantage to the lower bound in [1] by assuming $N = \frac{1}{8}c^2r^2$. Hence, substituting $N = \frac{1}{8}c^2r^2$ into (20) yields

$$Y_t \geq \frac{\log(2/\sqrt{e})}{108d}\frac{N}{\mu^2 t}. \tag{21}$$

Clearly, the lower bound in (21) is much smaller than the one of ours in Corollary 2, i.e., $\frac{N}{2\mu^2 t}$ when t is sufficiently large. Moreover, this lower bound depends on $1/d$ and it becomes smaller when d increases. It implies our lower bound in this paper is much better than the one in (21), i.e. it is much more tight.