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

Phuong Ha Nguyen
Electrical and Computer Engineering
University of Connecticut, USA
phuongha.ntu@gmail.com

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

Marten van Dijk
Electrical and Computer Engineering
University of Connecticut, USA
marten.van_dijk@uconn.edu

Abstract

We study the convergence of Stochastic Gradient Descent (SGD) for strongly convex objective functions. We prove for all t a lower bound on the expected convergence rate after the t-th SGD iteration; the lower bound is over all possible sequences of diminishing step sizes. It implies that recently proposed sequences of step sizes at ICML 2018 and ICML 2019 are *universally* close to optimal in that the expected convergence rate after *each* iteration is within a factor 32 of our lower bound. This factor is independent of dimension d. We offer a framework for comparing with lower bounds in state-of-the-art literature and when applied to SGD for strongly convex objective functions our lower bound is a significant factor $775 \cdot d$ larger compared to existing work.

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) 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 [9]. The classic first-order methods to solve problem (2) are gradient descent (GD) [21] and stochastic gradient descent (SGD)[1] [23] algorithms. GD is a standard deterministic gradient method, which updates iterates along the negative full gradient with learning rate η_t as follows

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

[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.

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

We can choose $\eta_t = \eta = \mathcal{O}(1/L)$ and achieve a linear convergence rate for the strongly convex case [17]. The upper bound of the convergence rate of GD and SGD has been studied in [2, 4, 17, 24, 19, 18, 8].

The disadvantage of GD is that it 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 [12, 6, 10, 20] 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 satisfying

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

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

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

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_* and are interested in studying the expected convergence rate

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

Algorithm 1 provides a detailed description of SGD. Obviously, the computational cost of a single iteration in SGD is n times cheaper than that of a single iteration in GD. However, as has been shown in literature we need to choose $\eta_t = \mathcal{O}(1/t)$ and the expected 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 with probability density $g(\xi_t)$.
  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

Problem Statement and Contributions: We seek to find a tight lower bound on the expected convergence rate Y_t with the purpose of showing that the stepsize sequences of [19] and [8] for classical SGD is optimal for μ-strongly convex and L-smooth respectively expected L-smooth objective functions within a *small dimension independent constant factor*. This is important because of the following reasons:

1. The lower bound tells us that a sequence of stepsizes as a function of only μ and L cannot beat an expected convergence rate of $O(1/t)$ – this is known general knowledge and was already proven in [1], where a *dimension dependent* lower bound for a larger class of algorithms that includes SGD was proven. For the class of SGD with diminishing stepsizes as a function of only global parameters μ and L we show a *dimension independent* lower bound which is a factor $775 \cdot d$ larger.
2. We now understand into what extent the sequence of stepsizes of [19] and [8] are optimal in that it leads to minimal expected convergence rates Y_t for *all* t: For each t we will show a *dimension independent* lower bound on Y_t over *all possible* stepsize sequences. This includes the *best possible* stepsize sequence which minimizes Y_t for a *given* t. Our lower bound achieves the upper bound on Y_t for the stepsize sequences of [19] and [8] within a factor 32 for *all* t. This implies that these stepsize sequences universally minimizes each Y_t within factor 32.

3. As a consequence, in order to attain a better expected convergence rate, we need to *either* assume more specific knowledge about the objective function F so that we can construct a better stepsize sequence for SGD based on this additional knowledge *or* we need to step away from SGD and use a different kind of algorithm. For example, the larger class of algorithms in [1] may contain a non-SGD algorithm which may get close to the lower bound proved in [1] which is a factor $775 \cdot d$ smaller. Since the larger class of algorithms in [1] contains algorithms such as Adam [11], AdaGrad [7], SGD-Momentum [25], RMSProp [27] we now know that these practical algorithms will at most improve a factor $32 \cdot 775 \cdot d$ over SGD for strongly convex optimization – this can be significant as this can lead to orders of magnitude less gradient computations. We are the first to make such quantification.

Outline: Section 2 discusses background: First, we discuss the recurrence on Y_t used in [19] for proving their upper bound on Y_t – this recurrence plays a central role in proving our lower bound. We discuss the upper bounds of both [19] and [8] – the latter holding for a larger class of algorithms. Second, we explain the lower bound of [1] in detail in order to be able to properly compare with our lower bound. Section 3 introduces a framework for comparing bounds and explains the consequences of our lower bound in detail. Section 4 describes a class of strongly convex and smooth objective functions which is used to derive our lower bound. We also verify our theory by experiments in supplemental material B. Section 5 concludes the paper.

2 Background

We explain the upper bound of [19, 8], and lower bound of [1] respectively.

2.1 Upper Bound for Strongly Convex and Smooth Objective Functions

The starting point for analysis is the recurrence first introduced in [19, 13]

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

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, see [19, 13], if we assume

1. $F(.)$ is μ-strongly convex,
2. $f(w; \xi)$ is L-smooth,
3. $f(w; \xi)$ is convex, and
4. N is finite;

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 $\forall w, w' \in \mathbb{R}^d$,*

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

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, $\forall w, w' \in \mathbb{R}^d$,*

$$\|\nabla f(w; \xi) - \nabla f(w'; \xi)\| \leq L\|w - w'\|. \tag{6}$$

Assumption 2 implies that F is also L-smooth.

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

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

Assumption 4. *$N = 2\mathbb{E}[\|\nabla f(w_*; \xi)\|^2]$ is finite.*

We denote the set of strongly convex objective functions by $\mathcal{F}_{str}$ and denote the subset of $\mathcal{F}_{str}$ satisfying Assumptions 1, 2, 3, and 4 by $\mathcal{F}_{sm}$.

We notice that the earlier established recurrence in [15] under the same set of assumptions

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

is similar, but worse than (4) as it only holds for $\eta_t < \frac{\mu}{L^2}$ where (4) holds for $\eta_t \leq \frac{1}{2L}$. Only for step sizes $\eta_t < \frac{\mu}{2L^2}$ the above recurrence provides a better bound than (4), i.e., $1-2\mu\eta_t+2L^2\eta_t^2 \leq 1-\mu\eta_t$. 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). See (8) below, 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 analyzing recurrence (4) in order to prove our lower bound: For $\eta_t \leq \frac{1}{2L}$,

$$Y_{t+1} \leq (1 - \mu\eta_t)Y_t + \eta_t^2 N, \tag{7}$$

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

Based on the above assumptions (*without* the so-called bounded gradient assumption) and knowledge of only μ and L a sequence of step sizes η_t can be constructed such that Y_t is smaller than $\mathcal{O}(1/t)$ [19]; more explicitly, for the sequence of step sizes

$$\eta_t = \frac{2}{\mu t + 4L} \tag{8}$$

we have for all objective functions in $\mathcal{F}_{sm}$ the upper bound

$$Y_t \leq \frac{16N}{\mu}\frac{1}{\mu(t - T') + 4L} = \frac{16N}{\mu^2 t}(1 + \mathcal{O}(1/t)), \tag{9}$$

where

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

We notice that [8] studies the larger class, which we denote $\mathcal{F}_{esm}$, which is defined as $\mathcal{F}_{sm}$ where expected smoothness is assumed in stead of smoothness and convexity of component functions. We rephrase their assumption for classical SGD as studied in this paper.[2]

Assumption 5. *(L-smooth in expectation) The objective function $F : \mathbb{R}^d \to \mathbb{R}$ is L-smooth in expectation if there exists a constant $L > 0$ such that, $\forall w \in \mathbb{R}^d$,*

$$\mathbb{E}[\|\nabla f(w;\xi) - \nabla f(w_*;\xi)\|^2] \leq 2L\|F(w) - F(w_*)\|. \tag{10}$$

The results in [8] assume the above assumption for empirical risk minimization (2). L-smoothness, see [17], implies Lipschitz continuity (i.e., $\forall w, w' \in \mathbb{R}^d$,

$$f(w,\xi) \leq f(w',\xi) + \langle \nabla f(w',\xi), (w - w')\rangle + \frac{L}{2}\|w - w'\|^2$$

) and together with Proposition A.1 in [8] this implies L-smooth in expectation. This shows that $\mathcal{F}_{esm}$ defined by Assumptions 1, 4, and 5 is indeed a superset of $\mathcal{F}_{sm}$.

The step sizes (8) from [19] for $\mathcal{F}_{sm} \subseteq \mathcal{F}_{esm}$ and

$$\eta_t = \frac{2t+1}{(t+1)^2\mu} \text{ for } t > \frac{4L}{\mu} \text{ and } \eta_t = \frac{1}{2L} \text{ for } t \leq \frac{4L}{\mu} \tag{11}$$

developed for $\mathcal{F}_{esm}$ in [8] and [19] are equivalent in that they are both $\approx \frac{2}{\mu t}$ for t large enough. Both step size sequences give exactly the same asymptotic upper bound (9) on Y_t (in our notation).

In [23], the authors proved the convergence of SGD for the step size sequence $\{\eta_t\}$ satisfying conditions $\sum_{t=0}^{\infty}\eta_t = \infty$ and $\sum_{t=0}^{\infty}\eta_t^2 < \infty$. In [15], 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 [23] and [15] do *not* discuss about the optimal step sizes among all proposed step sizes which is what is done in this paper.

[2]This means that distribution $\mathcal{D}$ in [8] must be over unit vectors $v \in [0,\infty)^n$, where n is the number of component functions, i.e., n possible values for ξ. Arbitrary distributions $\mathcal{D}$ correspond to SGD with mini-batches where each component function indexed by ξ is weighted with v_ξ.

2.2 Lower Bound for First Order Stochastic Oracles

The authors of [16] proposed the first formal study on lower bounding the expected convergence rate for a large class of algorithms which includes SGD. The authors of [1] and [22] independently studied this lower bound using information theory and were able to improve it.

The derivation in [1] is for algorithms including SGD where the sequence of stepsizes is a-priori fixed based on global information regarding assumed stochastic parameters concerning the objective function F. Their proof uses the following set of assumptions: First, The assumption of a strongly convex objective function, i.e., Assumption 1 (see Definition 3 in [1]). Second, the objective function is convex Lipschitz:

Assumption 6. *(convex Lipschitz) The objective function F is a convex Lipschitz function, i.e., there exists a bounded convex set $\mathcal{S} \subset \mathbb{R}^d$ and a positive number K such that $\forall w, w' \in \mathcal{S} \subset \mathbb{R}^d$*

$$\|F(w) - F(w')\| \leq K\|w - w'\|.$$

We notice that this assumption implies the assumption on bounded gradients as stated here (and explicitly mentioned in Definition 1 in [1]): There exists a bounded convex set $\mathcal{S} \subset \mathbb{R}^d$ and a positive number σ such that

$$\mathbb{E}[\|\nabla f(w;\xi)\|^2] \leq \sigma^2 \tag{12}$$

for all $w \in \mathcal{S} \subset \mathbb{R}^d$. This is not the same as the bounded gradient assumption where $\mathcal{S} = \mathbb{R}^d$ is unbounded.[3] Clearly, for w_*, (12) implies a finite $N \leq 2\sigma^2$.

We define $\mathcal{F}_{lip}$ as the set of strongly convex objective functions that satisfy Assumption 6. Classes $\mathcal{F}_{esm}$ and $\mathcal{F}_{lip}$ are both subsets of $\mathcal{F}_{str}$ and differ (are not subclasses of each other) in that they assume expected smoothness and convex Lipschitz respectively.

To prove a lower bound of Y_t for $\mathcal{F}_{lip}$, the authors constructed a class of objective functions $\subseteq \mathcal{F}_{lip}$ and showed a lower bound of Y_t for this class; in terms of the notation used in this paper,

$$\frac{\log(2/\sqrt{e})}{432 \cdot d} \frac{N}{\mu^2 t}. \tag{13}$$

The authors of [1] prove lower bound (13) for the class $\mathcal{A}_{stoch}$ of *stochastic first order algorithms* that can be understood as operating based on information provided by a stochastic first-order oracle, i.e., any algorithm which bases its computation in the t-th iteration on μ, K or L, d, and access to an oracle that provides $f(w_t;\xi_t)$ and $\nabla f(w_t;\xi_t)$. This class includes $\mathcal{A}_{SGD}$ defined as SGD with some sequence of diminishing step sizes as a function of global parameters such as μ and L or μ and K, see Algorithm 1. We notice that $\mathcal{A}_{stoch}$ also includes practical algorithms such as Adam [11], etc. We revisit their derivation in supplementary material C where we show[4] how their lower bound transforms into (13). Notice that their lower bound depends on dimension d.

3 Framework for Upper and Lower Bounds

Let $par(F)$ denote the concrete values of the global parameters of an objective function F such as the values for μ and L corresponding to objective functions F in $\mathcal{F}_{sm}$ and $\mathcal{F}_{esm}$ or μ and K corresponding to objective functions F in $\mathcal{F}_{lip}$. When defining a class $\mathcal{F}$ of objective functions, we also need to explain how $\mathcal{F}$ defines a corresponding $par(.)$ function. We will use the notation $\mathcal{F}[p]$ to stand for the subclass $\{F \in \mathcal{F} \ : \ p = par(F)\} \subseteq \mathcal{F}$, i.e., the subclass of objective functions of $\mathcal{F}$ with the same parameters p. We assume that parameters of a class are included in the parameters of a smaller subclass: For example, $\mathcal{F}_{sm}$ is a subset of the class of strongly convex objective functions $\mathcal{F}_{str}$ with only global parameter μ. This means that for concrete values μ and L we have $\mathcal{F}_{sm}[\mu, L] \subseteq \mathcal{F}_{str}[\mu]$.

For a given objective function F, we are interested in the best possible expected convergence rate after the t-th iteration among all possible algorithms A in a larger class of algorithms $\mathcal{A}$. Here, we

[3] The bounded gradient assumption, where $\mathcal{S}$ is unbounded, is in conflict with assuming strong convexity as explained in [19].

[4] We also discuss the underlying assumption of convex Lipschitz and show that in order for the analysis in [1] to follow through one – likely tedious but believable – statement still needs a formal proof.

assume that $\mathcal{A}$ is a subclass of the larger class $\mathcal{A}_{stoch,\mathcal{U}}$ of stochastic first order algorithms where the computation in the t-th iteration not only has access to $par(F)$ and access to an oracle that provides $f(w_t;\xi_t)$ and $\nabla f(w_t;\xi_t)$ but also access to possibly another oracle $\mathcal{U}$ providing even more information. Notice that $\mathcal{A} \subseteq \mathcal{A}_{stoch} \subseteq \mathcal{A}_{stoch,\mathcal{U}}$ for any oracle $\mathcal{U}$. With respect to the expected convergence rate, we want to know which algorithm A in $\mathcal{A}$ minimizes Y_t the most. Notice that for different t this may be a different algorithm A. We define for $F \in \mathcal{F}$ (with associated $par(.)$)

$$\gamma_t^F(\mathcal{A}) = \inf_{A\in\mathcal{A}} Y_t(F,A),$$

where Y_t is explicitly shown as a function of the objective function F and choice of algorithm A.

Among the objective functions $F \in \mathcal{F}$ with same global parameters $p = par(F)$ (i.e., $F \in \mathcal{F}[p]$), we consider the objective function F which has the *worst* expected convergence rate at the t-th iteration. This is of interest to us because algorithms A only have access to $p = par(F)$ as the sole information about objective function F, hence, if we prove an upper bound on the expected convergence rate for algorithm A, then this upper bound must hold for all $F \in \mathcal{F}$ with the same parameters $p = par(F)$. In other words such an upper bound must be at least

$$\gamma_t(\mathcal{F}[p],\mathcal{A}) = \sup_{F\in\mathcal{F}[p]} \gamma_t^F(\mathcal{A}) = \sup_{F\in\mathcal{F}[p]} \inf_{A\in\mathcal{A}} Y_t(F,A).$$

So, any lower bound on $\gamma_t(\mathcal{F}[p],\mathcal{A})$ gives us a lower bound on the best possible upper bound on Y_t that can be achieved. Such a lower bound tells us into what extent the expected convergence rate Y_t cannot be improved.

The lower bound (13) and upper bound (9) are not only a function of μ in $p = par(F)$ but also a function of N which is outside $p = par(F)$ for $F \in \mathcal{F}_{lip}$ or $F \in \mathcal{F}_{esm}$. We are really interested in such more fine-grained bounds that are a function of N. For this reason we need to consider the subclass of objective functions F in $\mathcal{F}[p]$ that all have the same N. We implicitly understand that N is an auxiliary parameter of an objective function F and we denote this as a function of F as $N(F)$. We define $\mathcal{F}^a[p] = \{F \in \mathcal{F}[p] \ : \ a = aux(F)\}$ where $aux(.)$ represents for example $N(.)$. This leads to notation like $\mathcal{F}^N_{lip}[\mu,K,d]$. Notice that $p = par(F)$ can be used by an algorithm $A \in \mathcal{A}$ while $a = aux(F)$ is not available to A through $p = par(F)$ (but may be available through access to an oracle).

If we find a tight lower bound with upper bound up to a constant factor, as in this paper, then we know that the algorithm that achieves the upper bound is close to optimal in that the expected convergence rate cannot be further minimized/improved in a significant way. In practice we are only interested in upper bounds on Y_t that can be realized by the *same* algorithm A (if not, then we need to know a-priori the exact number of iterations t we want to run an algorithm and then choose the best one for that t). In this paper we consider the algorithm A for F in $\mathcal{F}_{sm}$ resp. $\mathcal{F}_{esm}$ defined as SGD with diminishing step sizes (8) resp. (11) as a function of $par(F) = (\mu, L)$ giving upper bound (9) on expected convergence rate $Y_t(F,A)$. We show that A is close to optimal.

Given the above definitions we have

$$\gamma_t(\mathcal{F}[p],\mathcal{A}) \leq \gamma_t(\mathcal{F}'[p'],\mathcal{A}') \tag{14}$$

for $\mathcal{F}[p] \subseteq \mathcal{F}'[p']$ and $\mathcal{A}' \subseteq \mathcal{A}$, i.e., the worst objective function in a larger class of objective functions is worse than the worst objective function in a smaller class of objective functions (see the supremum used in defining γ_t) and the best algorithm from a larger class of algorithms is better than the best algorithm from a smaller class of algorithms (see the infinum used in defining γ_t). This implies

$$\gamma_t(\mathcal{F}^N_{lip}[\mu,K,d],\mathcal{A}_{stoch}) \quad \leq \quad \gamma_t(\mathcal{F}^N_{str}[\mu],\mathcal{A}_{SGD}), \tag{15}$$

$$\gamma_t(\mathcal{F}^N_{sm}[\mu,L],\mathcal{A}_{ExtSGD}) \quad \leq \quad \gamma_t(\mathcal{F}^N_{esm}[\mu,L],\mathcal{A}_{SGD}) \leq \gamma_t(\mathcal{F}^N_{str}[\mu],\mathcal{A}_{SGD}), \tag{16}$$

where $\mathcal{A}_{SGD} \subseteq \mathcal{A}_{ExtSGD}$ is defined as follows:

In our framework we introduce *extended SGD* as the class $\mathcal{A}_{ExtSGD}$ of SGD algorithms where the stepsize in the t-th iteration can be computed based on global parameters μ, L, and access to an oracle $\mathcal{U}$ that provides additional information N, $\nabla F(w_t)$, and Y_t. This class also includes SGD with

diminishing stepsizes as defined in Algorithm 1, i.e., $\mathcal{A}_{SGD} \subseteq \mathcal{A}_{ExtSGD}$. The reason for introducing the larger class $\mathcal{A}_{ExtSGD}$ is not because it contains practical algorithms different than SGD, on the contrary. The only reason is that it allows us to define *one single algorithm* $A \in \mathcal{A}_{ExtSGD}$ which realizes $\gamma_t^F(\mathcal{A}_{ExtSGD})$ *for all* t for all F in a to be constructed subclass $\mathcal{F} \subseteq \mathcal{F}_{sm}$ – the topic of the next section. This property allows a rather straightforward calculus based proof without needing to use more advanced concepts from information and probability theory as required in the proof of [1]. Looking ahead, we will prove in Theorem 1

$$\frac{1}{2}\frac{N}{\mu^2 t}(1 - \mathcal{O}((\ln t)/t)) \leq \gamma_t(\mathcal{F}_{sm}^N[\mu, L], \mathcal{A}_{ExtSGD}). \tag{17}$$

Notice that the construction of η_t for algorithms in $\mathcal{A}_{ExtSGD}$ 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 $\nabla f(w_t; \xi_t)$.

As a disclaimer we notice that for some objective functions $F \in \mathcal{F}_{sm}^N[\mu, L]$ the expected convergence rate can be much better than what is stated in (17); this is because $\gamma_t(\{F\}, \mathcal{A}_{ExtSGD})$ can be much smaller than $\gamma_t(\mathcal{F}_{sm}^N[\mu, L], \mathcal{A}_{ExtSGD})$, see (14). This is due to the specific nature of the objective function F 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 (17).

Results (13) and (9) of the previous section combined with (15), (16), and (17) yield

$$\frac{\log(2/\sqrt{e})}{432 \cdot d}\frac{N}{\mu^2 t} \leq \gamma_t(\mathcal{F}_{lip}^N[\mu, K, d], \mathcal{A}_{stoch}) \quad \leq \quad \gamma_t(\mathcal{F}_{str}^N[\mu], \mathcal{A}_{SGD}), \tag{18}$$

$$\frac{1}{2}\frac{N}{\mu^2 t}(1 - \mathcal{O}((\ln t)/t)) \leq \gamma_t(\mathcal{F}_{esm}^N[\mu, L], \mathcal{A}_{ExtSGD}) \quad \leq \quad \gamma_t(\mathcal{F}_{str}^N[\mu], \mathcal{A}_{SGD}), \tag{19}$$

$$\begin{aligned} \frac{1}{2}\frac{N}{\mu^2 t}(1 - \mathcal{O}((\ln t)/t)) \leq \gamma_t(\mathcal{F}_{sm}^N[\mu, L], \mathcal{A}_{ExtSGD}) &\leq \gamma_t(\mathcal{F}_{esm}^N[\mu, L], \mathcal{A}_{SGD}) \\ &\leq \frac{16N}{\mu^2 t}(1 + \mathcal{O}(1/t)). \end{aligned} \tag{20}$$

We conclude the following observations (our contributions):

1. The first inequality (18) is from [1]. Comparing (19) to (18) shows that as a lower bound for $\gamma_t(\mathcal{F}_{str}^N[\mu], \mathcal{A}_{SGD})$ (SGD for the class of strongly convex objective functions) our lower bound (17) is dimension independent and improves the lower bound (13) of [1] by a factor $775 \cdot d$. This is a significant improvement.
2. However, our lower bound does not hold for the larger class $\mathcal{A}_{stoch}$. This teaches us that if we wish to reach smaller (better) expected convergence rates, then one approach is to step beyond SGD where our lower bound does not hold implying that within $\mathcal{A}_{stoch}$ there may be an opportunity to find an algorithm leading to at most a factor $32 \cdot 775 \cdot d$ smaller expected convergence rate compared to upper bound (20). This is the first exact quantification into what extent a better (practical) algorithm when compared to classical SGD can be found. E.g., Adam [11], AdaGrad [7], SGD-Momentum [25], RMSProp [27] are all in $\mathcal{A}_{stoch}$ and can beat classical SGD by at most a factor $32 \cdot 775 \cdot d$.
3. When searching for a better algorithm in $\mathcal{A}_{stoch}$ which significantly improves over SGD, it does not help to take an SGD-like algorithm which uses step sizes that are a function of iteratively computed estimates of $\nabla F(w_t)$ and Y_t as this would keep such an algorithm in $\mathcal{A}_{ExtSGD}$ for which our lower bound is tight.
4. Another approach to reach smaller expected convergence rates is to stick with SGD but consider a smaller restricted class of objective functions for which more/other information in the form of extra global parameters is available for adaptively computing η_t.
5. For strongly convex and smooth, respectively expected smooth, objective functions the algorithm $A \in \mathcal{A}_{SGD}$ with stepsizes $\eta_t = \frac{2}{\mu t + 4L}$, respectively $\eta_t = \frac{2t+1}{(t+1)^2\mu}$ for $t > \frac{4L}{\mu}$ and $\eta_t = \frac{1}{2L}$ for $t \leq \frac{4L}{\mu}$, realizes the upper bound in (20) for all t. Inequalities (20) show that this algorithm is close to optimal: For each t, the best sequence of diminishing step sizes which minimizes Y_t can at most achieve a constant (dimension independent) factor 32 smaller expected convergence rate.

4 Lower Bound for Extended SGD

In order to prove a lower bound we propose a specific subclass of strongly convex and smooth objective functions F and we show in the extended SGD setting how, based on recurrence (7), to compute the *optimal* step size η_t as a function of μ and L and an oracle $\mathcal{U}$ with access to N, $\nabla F(w_t)$, and Y_t, i.e., this step size achieves the smallest Y_{t+1} at the t-th iteration.

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

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

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:

- 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. The proof of the lower bound is presented in supplemental material A):

Theorem 1. *We assume that component functions $f(w;\xi)$ are constructed according to the recipe described above with $\mu < L/18$. Then, the corresponding objective function is μ-strongly convex and the component functions are L-smooth and convex.*

If we run Algorithm 1 and assume that access to an oracle $\mathcal{U}$ with access to N, $\nabla F(w_t)$, and Y_t is given at the t-th iteration (our extended SGD problem setting), then an exact expression for the optimal sequence of stepsizes η_t based on $\mu, L, N, Y_0, \ldots, Y_t$ can be given, i.e., this sequence of stepsizes achieves the smallest possible Y_{t+1} at the t-th iteration for all t. For this sequence of stepsizes,

$$Y_t \geq \frac{N}{2\mu}\frac{1}{\mu t + 2\mu\ln(t+1) + W}, \tag{21}$$

where

$$W = \frac{L^2}{12(L-\mu)}.$$

In supplemental material B we show numerical experiments in agreement with the presented theorem.

5 Conclusion

We have studied the convergence of SGD by introducing a framework for comparing upper bounds and lower bounds and by proving a new lower bound based on straightforward calculus. The new lower bound is dimension independent and improves a factor $775 \cdot d$ over previous work [1] applied to SGD, shows the optimality of step sizes in [19, 8], and shows that practical algorithms like Adam [11], AdaGrad [7], SGD-Momentum [25], RMSProp [27] for strongly convex objective functions can at most achieve a factor $32 \cdot 775 \cdot d$ smaller expected convergence rate compared to classical SGD.

Acknowledgement

We thank the reviewers for useful suggestions to improve the paper. Phuong Ha Nguyen and Marten van Dijk were supported in part by AFOSR MURI under award number FA9550-14-1-0351.

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] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159, 2011.

[8] Robert Mansel Gower, Nicolas Loizou, Xun Qian, Alibek Sailanbayev, Egor Shulgin, and Peter Richtarik. Sgd: General analysis and improved rates. *arXiv preprint arXiv:1901.09401*, 2019.

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

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

[11] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[12] 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.

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

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

[15] 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.

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

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

[18] Lam Nguyen, Phuong Ha Nguyen, Peter Richtarik, Katya Scheinberg, Martin Takac, and Marten van Dijk. New convergence aspects of stochastic gradient algorithms. *arXiv preprint arXiv:1811.12403*, 2018.

[19] Lam 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.

[20] 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.

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

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

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

[24] Mark Schmidt and Nicolas Le Roux. Fast convergence of stochastic gradient descent under a strong growth condition. *arXiv preprint arXiv:1308.6370*, 2013.

[25] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pages 1139–1147, 2013.

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

[27] Fangyu Zou, Li Shen, Zequn Jie, Weizhong Zhang, and Wei Liu. A Sufficient Condition for Convergences of Adam and RMSProp. *arXiv preprint arXiv:1811.09358*, 2018.

Supplementary Material

A Proof

We extend Theorem 1 with an upper bound used in our numerical experiments.

Theorem 1 *We assume that component functions $f(w;\xi)$ are constructed according to the recipe described in Section 4 with $\mu < L/18$. Then, the corresponding objective function is μ-strongly convex and the component functions are L-smooth and convex.*

If we run Algorithm 1 and assume that access to an oracle $\mathcal{U}$ with access to N, $\nabla F(w_t)$, and Y_t is given at the t-th iteration (our extended SGD problem setting), then an exact expression for the optimal sequence of stepsizes η_t based on $\mu, L, N, Y_0, \ldots, Y_t$ can be given, i.e., this sequence of stepsizes achieves the smallest possible Y_{t+1} at the t-th iteration for all t. For this sequence of stepsizes,

$$Y_t \geq \frac{N}{2\mu} \frac{1}{\mu t + 2\mu \ln(t+1) + W},$$

where $W = \frac{L^2}{12(L-\mu)}$ and for $t \geq T' = \frac{20L}{\mu}$,

$$Y_t \leq \frac{16N}{\mu} \frac{1}{\mu t - 16L}. \tag{22}$$

Proof. We first restrict oracle $\mathcal{U}$ to only supply information about N and Y_t at the t-th iteration. At the end of this proof we show that our arguments generalize to the more powerful oracle $\mathcal{U}$ which also provides the full gradient $\nabla F(w_t)$ at the t-th iteration.

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

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

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

Therefore, $F(w) = \mu\mathbb{E}_\xi[\frac{\|w-\xi\|^2}{2}] = \mu\frac{\|w-m\|^2}{2} + \mu\frac{\mathrm{Tr}(\Sigma)}{2}$, and this shows F is μ-strongly convex and has minimum $w_* = m$.

Since

$$\begin{aligned}\nabla_w[\|w-\xi\|^2] &= \nabla_w[\langle w,w\rangle - 2\langle w,\xi\rangle + \langle \xi,\xi\rangle] \\ &= 2w - 2\xi = 2(w-\xi),\end{aligned}$$

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 ($w_* = m$)

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

and we know

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

This yields

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

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 draw ξ from its distribution and set $w_0 = \xi$. Therefore,

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

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

$$\begin{aligned}&\mathbb{E}[(1-\eta_t s(\xi_t))^2\|w_t-w_*\|^2 \\ &+ 2\eta_t s(\xi_t)(1-\eta_t s(\xi_t))\langle w_t - w_*, \xi_t - w_*\rangle \\ &+ \eta_t^2 s(\xi_t)^2\|\xi_t - w_*\|^2|\mathcal{F}_t].\end{aligned} \tag{23}$$

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

$$\begin{aligned}&\mathbb{E}[(1-\eta_t s(\xi_t))^2]\|w_t-w_*\|^2 \\ &+ \mathbb{E}[2\eta_t s(\xi_t)(1-\eta_t s(\xi_t))]\langle w_t - w_*, \mathbb{E}[\xi_t - w_*]\rangle \\ &+ \mathbb{E}[\eta_t^2 s(\xi_t)^2]\mathbb{E}[\|\xi_t - w_*\|^2].\end{aligned} \tag{24}$$

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

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

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

$$Y_{t+1} = (1 - \mu\eta_t(2 - \frac{\eta_t}{12}\frac{L^2}{L-\mu}))Y_t + \eta_t^2\frac{N}{2}. \tag{25}$$

This is very close to recurrence (4).

Equation (25) expresses Y_{t+1} as a function $Y_{t+1}(\eta_t, Y_t)$ of η_t and Y_t. Given Y_0, we want to minimize Y_{t+1} with respect to the step sizes $\eta_t, \eta_{t-1}, \ldots, \eta_0$. For $i < t$ we derive

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

and for $i = t$ we derive

$$\frac{\partial Y_{t+1}}{\partial \eta_i} = -2\mu Y_t + 2\mu\frac{\eta_t}{12}\frac{L^2}{L-\mu}Y_t + N\eta_t. \tag{26}$$

We reach a stationary point for Y_{t+1} as a function of step sizes $\eta_t, \eta_{t-1}, \ldots, \eta_0$ if each of the partial derivatives with respect to η_i is equal to 0. We notice that if for all t

$$1 - \mu\eta_t(2 - \frac{\eta_t}{12}\frac{L^2}{L-\mu}) > 0, \tag{27}$$

then, for $i < t$, $\frac{\partial Y_{t+1}}{\partial \eta_i} = 0$ if and only if $\frac{\partial Y_t}{\partial \eta_i} = 0$. This implies that Y_{t+1} has a stationary point if and only if

$$\forall_{0 \le i \le t}\ \frac{\partial Y_{i+1}}{\partial \eta_i} = 0.$$

Hence, if a step size sequence satisfies this for all t, then it leads to stationary points for all Y_{t+1} as function of $\eta_t, \eta_{t-1}, \ldots, \eta_0$. So, such a sequence of step sizes simultaneously achieves stationary points for all Y_{t+1}.

For the argument to hold, we need to prove (27). The left hand side of (27) achieves its minimum value

$$1 - 12\mu\frac{L-\mu}{L^2}$$

for $\eta_t = 12\frac{L-\mu}{L^2}$. For $\mu < \frac{L}{12}$, $12\mu(L-\mu) < 12\mu L < L^2$ implying that this minimum value is larger than zero.

As explained above the optimal step size η_t in a sequence of optimal step sizes that minimizes all expected convergence rates Y_t is computed by taking the derivative of Y_{t+1} with respect to η_t. This derivative is equal to (26) and shows that the minimum is achieved for

$$\eta_t = \frac{2\mu Y_t}{N + \frac{\mu L^2}{6(L-\mu)}Y_t} \tag{28}$$

giving, see (25),

$$\begin{aligned} Y_{t+1} &= Y_t - \frac{2\mu^2 Y_t^2}{N + \frac{\mu L^2}{6(L-\mu)}Y_t} \\ &= Y_t - \frac{2\mu^2 Y_t^2}{N(1 + Y_t/\mathrm{Tr}(\Sigma))}. \end{aligned} \tag{29}$$

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

$$Y_{t+1} \ge Y_t - \frac{2\mu^2 Y_t^2}{N} \tag{30}$$

Let us define $\gamma = 2\mu^2/N$. We can rewrite (30) as follows:

$$\begin{aligned} \gamma Y_{t+1} &\ge \gamma Y_t(1 - \gamma Y_t) \text{ or} \\ (\gamma Y_{t+1})^{-1} &\le 1 + (\gamma Y_t)^{-1} + \frac{1}{(\gamma Y_t)^{-1} - 1}. \end{aligned} \tag{31}$$

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 is implied by $Y_0 = \mathrm{Tr}(\Sigma) < \frac{2}{3\gamma}$, a condition which is needed in the next sequence of arguments. This stronger condition means that we need

$$\mathrm{Tr}(\Sigma) < \frac{N}{3\mu^2}, \text{ i.e., } \mathrm{Tr}(\Sigma) < \frac{\mu L^2}{6(L-\mu)} \frac{\mathrm{Tr}(\Sigma)}{3\mu^2}$$

after substituting N. This is equivalent to $\mu < \frac{L^2}{18(L-\mu)}$ which is true for $\mu < L/18$.

By using induction on t, upper bound (31) implies

$$(\gamma Y_{t+1})^{-1} \leq (t+1) + (\gamma Y_0)^{-1} + \sum_{i=0}^{t} \frac{1}{(\gamma Y_i)^{-1} - 1}. \tag{32}$$

In order to further upper bound the sum in the right hand side, we first find a lower bound on $(\gamma Y_i)^{-1}$. We rewrite equation (29) as

$$(\gamma Y_{t+1}) = (\gamma Y_t)(1 - \frac{(\gamma Y_t)}{1 + Y_t/\mathrm{Tr}(\Sigma)}).$$

Since $Y_t \leq Y_0 = \mathrm{Tr}(\Sigma)$, we have

$$(\gamma Y_{t+1}) \leq (\gamma Y_t)(1 - \frac{(\gamma Y_t)}{2}).$$

This translates into

$$\begin{aligned}
(\gamma Y_{t+1})^{-1} &\geq \frac{(\gamma Y_t)^{-1}}{1 - (\gamma Y_t)/2} = \frac{(\gamma Y_t)^{-2}}{(\gamma Y_t)^{-1} - 1/2} \\
&= \frac{1}{2} + (\gamma Y_t)^{-1} + \frac{1}{4(\gamma Y_t)^{-1} - 2} \\
&\geq \frac{1}{2} + (\gamma Y_t)^{-1},
\end{aligned}$$

where the last inequality follows from $(\gamma Y_t)^{-1} > (\gamma Y_0)^{-1} = (\gamma \mathrm{Tr}(\Sigma))^{-1} > 1$ making $4(\gamma Y_t)^{-1} - 2$ positive.

The resulting inequality leads to a recurrence and by using induction on t we obtain

$$(\gamma Y_{t+1})^{-1} \geq (t+1)/2 + (\gamma Y_0)^{-1}.$$

Now we are able to upper bound

$$\begin{aligned}
\sum_{i=0}^{t} \frac{1}{(\gamma Y_i)^{-1} - 1} &\leq \sum_{i=0}^{t} \frac{1}{i/2 + (\gamma Y_0)^{-1} - 1} \\
&= 2 \sum_{i=0}^{t} \frac{1}{i + 2((\gamma Y_0)^{-1} - 1)}.
\end{aligned}$$

We showed earlier that $\mu < L/18$ implies $Y_0 < \frac{2}{3\gamma}$. Substituting this upper bound in our derivation leads to

$$\sum_{i=0}^{t} \frac{1}{(\gamma Y_i)^{-1} - 1} \leq 2 \sum_{i=0}^{t} \frac{1}{i+1} \leq 2 \ln(t+2).$$

Combining with (32) we have the following inequality:

$$(\gamma Y_{t+1})^{-1} \leq (t+1) + (\gamma Y_0)^{-1} + 2 \ln(t+2).$$

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

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

where

$$W = N/(2\mu Y_0) = \frac{L^2}{12(L-\mu)}.$$

We now extend oracle $\mathcal{U}$ to also provide information about *full gradient* $\nabla F(w_t)$ at the t-th iteration. The above proof generalizes to this more powerful oracle. This is because of the reason why we are allowed to transform (23) into (24), i.e., η_t and ξ_t must be independent to get (24) from (23). 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 above proof to arrive at the lower bound of this theorem.

The upper bound for Y_t comes from the following fact. If we run Algorithm 1 with step size $\eta'_t = \frac{2}{\mu t + 4L}$ for $t \geq 0$ in [19], then we have from [19] an expected convergence rate

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

for $t \geq T'$, where

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

Substituting

$$N = \frac{\mu L^2}{6(L-\mu)} \mathrm{Tr}(\Sigma) \text{ and } Y_0 = \mathrm{Tr}(\Sigma)$$

yields $T' \leq \frac{20L}{\mu}$. Since η_t is the most optimal step size and η'_t is not, $Y_t \leq Y'_t$. I.e., we have for $t \geq \frac{20L}{\mu} \geq T'$,

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

□

B 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). We generate $m \in \mathbb{R}^d$ and a diagonal matrix $\Sigma \in \mathbb{R}^{d \times d}$ by drawing each element in m and each element on the diagonal of Σ at random from a uniform distribution over $[0, 1]$. We have $L = 1$ and $\mu = 1/n$ where n is the number of samples. Hence the condition number L/μ is equal to n and represents the number of SGD iterations in a single epoch. We experimented with 10 runs and reported the average results.

We denote the labels "Upper Y_t" (red line) and "Lower Y_t" (violet line) in Figure 1 as the upper and lower bounds of Y_t in (22) and (21) respectively (with a vertical line at epoch 20 because we expect to see the upper bound take effect when $t \geq T' = 20L/\mu$, see supplemental material A); "Y_t_opt" (orange line) as Y_t defined in Theorem 1 computed by using information from oracle $\mathcal{U}$; "Y_t" (green line) as the squared norm of the difference between w_t and w_*, where w_t is generated from Algorithm 1 with learning rate (28). Note that Y_t in Figure 1 is computed as average of 10 runs of $\|w_t - w_*\|^2$ (not exactly $\mathbb{E}[\|w_t - w_*\|^2]$).

"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 (22), (21), and (29), respectively, i.e., they have no variation. The green line "Y_t" for stepsize $\eta_t = \frac{2}{\mu t + 4L}$ in Figure 1 oscillates because our analysis does not consider the variance of $\|w_t - w_*\|^2$. From (4) we infer that a decrease in η_t leads to a decrease of the variance of $\|w_t - w_*\|^2$. 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) will increase the variance. Hence, it requires larger t to make the variance approach 0 as shown in Figure 1. For sufficiently large t, the 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). 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.

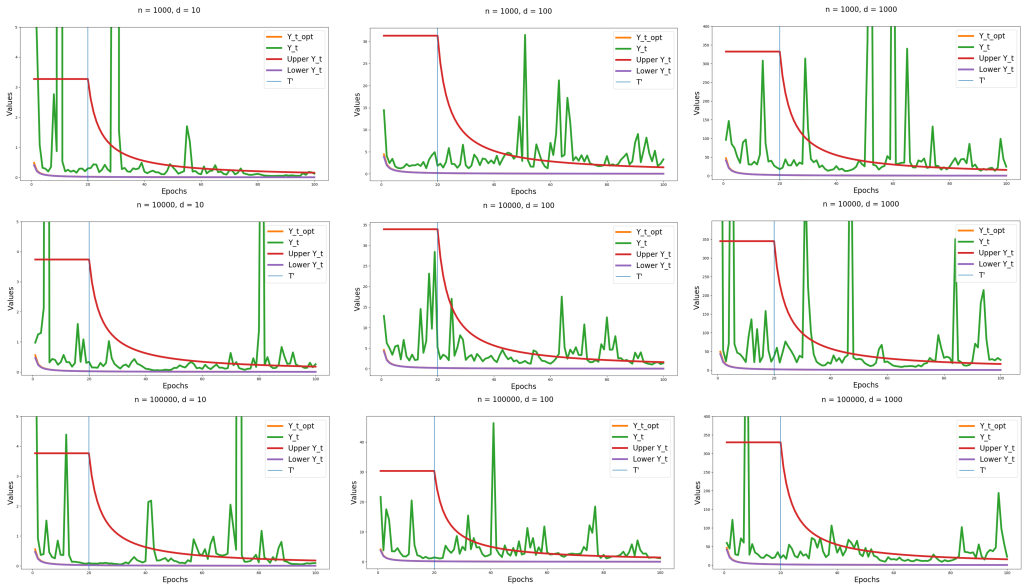

Figure 1: Y_t and its upper and lower bounds

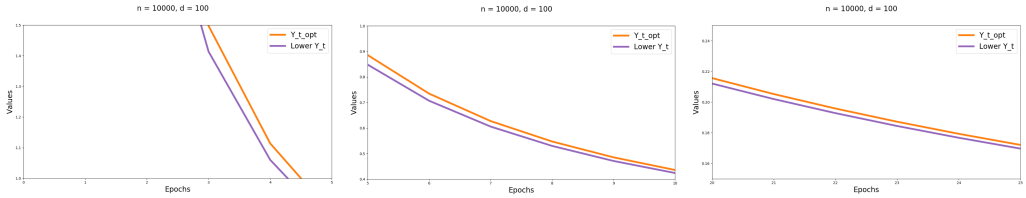

Figure 2: The difference between "Lower Y_t" and "Y_t_opt" ($n = 10000$, $d = 100$)

C 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 [19], 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. We conclude that in $\mathcal{F}_{str}$ there are objective functions that can escape any bounded set $\mathcal{S}$ with non-zero probability.

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 [19], the assumption of bounded gradient does not co-exist with strongly convex assumption. As shown in [17, 3], Assumption 1 on strong convexity implies

$$2\mu[F(w) - F(w_*)] \leq \|\nabla F(w)\|^2 \ , \ \forall w \in \mathbb{R}^d. \tag{33}$$

As shown in [19], for any $w \in \mathbb{R}^d$, we have

$$\begin{aligned} 2\mu[F(w) - F(w_*)] &\overset{(33)}{\leq} \|\nabla F(w)\|^2 = \|\mathbb{E}[\nabla f(w;\xi)]\|^2 \\ &\leq \mathbb{E}[\|\nabla f(w;\xi)\|^2] \leq \sigma^2. \end{aligned}$$

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.

C.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 a first order stochastic algorithm, the authors assume the **existence** of stepsizes η_t which make an first order stochastic 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 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.

Given the proof strategy in [1] of the convergence of a first order stochastic algorithm, 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] may still be 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 a first order stochastic algorithm, the algorithm diverges or converges. Let us define $p_t = Pr(w_t \notin \mathcal{S})$. Hence, $Pr(w_t \in \mathcal{S}) = 1 - p_t$. 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, for strongly convex objective functions and $w_*, w_0 \in \mathcal{S}$), it seems 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 rigorously complete the argument leading to the lower bound in [1]. In fact this remains an open problem (one can invent strange corner cases that need extra thought/proof).

The above result is interesting because now we **only** need to prove the convergence of a first order stochastic algorithm in a certain convex set $\mathcal{S}$ with a certain probability p. This is completely different from the proof of convergence of e.g. SGD in the general case as in [15] and [19, 8] where we need to prove it with probability 1.

C.2 Setup

We describe the setup of the class of strong convex functions proposed in [1].

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

$$F_\alpha(w) := \frac{c}{d} \sum_{i=1}^d \{(1/2 + \alpha_i \delta) f_i^+(w) + (1/2 - \alpha_i \delta) f_i^-(w)\}. \tag{34}$$

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 [26] and Le Cam's bound [5, 14] 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:

$$\frac{1}{3} \geq \Pr_{\mathcal{U}}[\hat{\alpha}(\mathcal{M}_t) \neq \alpha] \geq 1 - 2\frac{16dt\delta^2 + \log(2)}{d\log(2/\sqrt{e})}. \tag{35}$$

From (35), 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,

$$t \geq \frac{\log(2/\sqrt{e})}{48} \frac{1}{\delta^2}. \tag{36}$$

As shown in Section 4.3 [1], to proceed the proof, we set $Y_t = \frac{c\delta^2 r^2}{18(1-\theta)}$. Combining with (36) yields

$$Y_t \geq \frac{1}{t}\frac{\log(2/\sqrt{e})}{864}\frac{cr^2}{1-\theta}. \tag{37}$$

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 $\mathcal{S} = \mathrm{B}_\infty(r)$. By substituting c and μ^2 into (37), we obtain:

$$Y_t \geq \frac{1}{t}\frac{\log(2/\sqrt{e})}{864d}\frac{1}{\mu^2}c^2r^2. \tag{38}$$

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

C.3 Analysis and Comparison

In this section, we want to compare our lower bound ($\approx \frac{N}{2\mu^2 t}$) with the one in (38) 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 (34) 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

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

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 $[e_i\|c\nabla f_i^+(w^*)\|^2 + (1-e_i)\|c\nabla f_i^-(w^*)\|^2] = (1/4-\delta^2)(1+\theta)^2c^2r^2$. This proves that

$$N = 2\beta c^2 r^2$$

with β somewhere in the range

$$[(\frac{1}{2}-\delta), (\frac{1}{4}-\delta^2)(1+\theta)^2] \text{ or } [(\frac{1}{4}-\delta^2)(1+\theta)^2, (\frac{1}{2}-\delta)],$$

where $\delta \in [0, 1/4)$ and $\theta \in [0, 1)$.

Substituting $N = 2\beta c^2 r^2$ into (38) yields

$$Y_t \geq \frac{\log(2/\sqrt{e})}{(864\cdot d)(2\beta)}\frac{N}{\mu^2 t}, \tag{39}$$

which is further minimized by taking

$$\beta = \max\{(\frac{1}{2}-\delta), (\frac{1}{4}-\delta^2)(1+\theta)^2\}.$$

Notice that, given our freedom in choosing δ and θ, we can minimize β as a function of δ and θ in order to maximize the lower bound in (39). This gives (in the limit) $\delta = 1/4$ with $\theta \leq 2/\sqrt{3} - 1 = 0.155$ leading to $\beta = 1/4$. This leads to the final lower bound

$$Y_t \geq \frac{\log(2/\sqrt{e})}{432 \cdot d} \frac{N}{\mu^2 t}.$$

Clearly, the lower bound in is much smaller than our lower bound of $\approx \frac{N}{2\mu^2 t}$ when t is sufficiently large. Moreover, this lower bound depends on $1/d$ and it becomes smaller when d increases.