HIGH PROBABILITY ERROR BOUNDS OF SGD IN UNBOUNDED DOMAIN

Anonymous authors
Paper under double-blind review

ABSTRACT

This paper studies the high probability convergence behaviour of the stochastic gradient descent (SGD) method applied to convex problems. The existing tail-bound analysis of SGD relies crucially on assuming the domain of the problem to be bounded. In this work, we show that the bounded domain assumption can be removed for free. That is, we prove SGD in an unbounded domain enjoys the same high probability error bound as the bound established in the bounded domain; SGD converges with rate $\mathcal{O}(\log(1/\delta)/\epsilon^2)$ no matter the problem domain is bounded or not. As a by-product, we also prove that the trajectory of SGD is guaranteed to stay in a neighbourhood of the initialization with almost bounded diameter. As simple extensions of our analysis, we further establish the high probability error bounds of the last iterate of SGD and SGD with momentum, respectively.

1 INTRODUCTION

Stochastic gradient descent (SGD) is a simple and effective optimization algorithm for solving big data problems. Indeed, SGD and its variants, such as AdaGrad (Duchi et al., 2011) and ADAM (Kingma & Ba, 2015) are the dominant algorithms for training today's complex machine learning models with massive training data. Due to the empirical success of SGD, its convergence theory has been extensively studied in the literature (Robbins & Monro, 1951; Polyak & Juditsky, 1992; Zhang, 2004; Shalev-Shwartz et al., 2007; Nemirovski et al., 2009; Ghadimi & Lan, 2013; Shamir & Zhang, 2013; Bubeck, 2015).

Classic convergence analysis of SGD mostly focused on the error bounds in terms of expectation. It is well-established that the averaged iterate of SGD enjoys an expected error bound $\mathcal{O}(1/\sqrt{T})$ for convex problems and $\mathcal{O}(1/T)$ (Shalev-Shwartz et al., 2007; Nemirovski et al., 2009; Lacoste-Julien et al., 2012) for strongly convex problems. High probability error bounds of SGD exist (Kakade & Tewari, 2008; Rakhlin et al., 2012; Hazan & Kale, 2014; Harvey et al., 2019) but are less-studied compared with the convergence analysis on expectation. Consequently, there are still some missing pieces for a thorough understanding of the high probability error bounds of SGD. We aim to fill some of the missing pieces in this work. In particular, we study the high probability error bounds of SGD for the constrained convex problem with *unbounded domain*, that is, the constraint set may not have bounded diameter; for example, consider the constraint set $\mathbb{R}^d$, where d is the dimension of variables. Formally, we summarize our contributions as follows.

- For constrained convex problems that are Lipschitz continuous (or smooth) but with unbounded domain, we show that the averaged iterate of SGD enjoys a $\mathcal{O}(\log(1/\delta)/\sqrt{T})$ error bound (Theorem 5.4 and Theorem 6.2), where T is the number of iterations and δ is the confidence level. To our knowledge, this is the first high probability error bound in the form of $\mathcal{O}(\log(1/\delta)/\sqrt{T})$ without assuming the constraint set to have bounded diameter.
- As a by-product of our analysis, we show that the iterates generated from SGD are guaranteed to stay in a bounded region with high probability even if the constraint set may not be bounded (Theorem 5.5 and Theorem 6.3). We note that this result can potentially relax the bounded domain assumption for the analysis SGD in various settings.
- As simple extensions of our analysis in Section 5 and Section 6, we show that the high probability error bounds of the last iterate of SGD and SGD with momentum (mSGD) can easily be established.

The theoretical analysis in this work is arguably concise. Our proof relies on standard techniques in the analysis of first-order methods and a recent martingale concentration inequality called the generalized Freedman's inequality due to Harvey et al. (2019). Although our proof techniques are simple, we stress that the conclusions established in this work are non-trivial.

2 Related work

The stochastic gradient descent (SGD) algorithm is an old algorithm that can be traced back to 1951 (Robbins & Monro, 1951). Although simple, SGD is particularly effective in handling today's large-scale datasets and complex machine learning models because its convergence rate is independent of the number of data and model size. SGD is now a fundamental learning algorithm in the machine learning community. Given the empirical success and popularity of SGD, a huge line of works have been devoted to (i) designing new variants of SGD to improve its convergence either theoretically or empirically (Shalev-Shwartz et al., 2007; Ghadimi & Lan, 2013; Johnson & Zhang, 2013; Defazio et al., 2014; Kingma & Ba, 2015; Schmidt et al., 2017; Allen-Zhu, 2017a;b); (ii) refining the classic analysis of SGD in various settings and contributing to a better understanding of the classic SGD algorithm (Zhang, 2004; Rakhlin et al., 2012; Shamir & Zhang, 2013; Schmidt & Le Roux, 2013; Ma et al., 2018; Vaswani et al., 2019; Fang et al., 2021). This work belongs to the second category.

Most classic convergence rates of SGD are characterized in terms of expectation (Nemirovski et al., 2009; Ghadimi & Lan, 2013). The high probability error bounds of SGD is scarce in early works but have received substantial attention in recent years (Kakade & Tewari, 2008; Rakhlin et al., 2012; Hazan & Kale, 2014; Harvey et al., 2019; Feldman & Vondrák, 2019; Li & Orabona, 2020; Jain et al., 2021; Varre et al., 2021; Cutkosky & Mehta, 2021; Liu & Lu, 2021; Davis et al., 2021; Zhu et al., 2022). However, to our knowledge, high probability error bound in form $\mathcal{O}(\log(1/\delta)/\sqrt{T})$ has not been established for a single run of SGD in the convex and unbounded domain scenario. Our theoretical analysis owes to Harvey et al. (2019) as the main technical tool used in our analysis is the generalized Freedman's inequality proposed by Harvey et al. (2019). We note that Harvey et al. (2019) mainly focused on strongly convex problems, and it is already known that the high probability error bound of SGD for strongly convex problems holds regardless of the bounded domain assumption (Remark 4.2); for non-strongly convex problems, Harvey et al. (2019) assumed the domain to be bounded. Therefore the conclusions established in this work do not overlap with Harvey et al. (2019).

Notations Throughout the paper, we denote $[n] \coloneqq \{1, 2, \ldots, n\}$ for any positive integer n. We denote $\|\cdot\|$ as the Euclidean norm and use $\widetilde{\mathcal{O}}(\cdot)$ to hide poly-logarithmic terms.

3 Preliminaries

We consider the problem

$$\min_{x \in \mathcal{X}} f(x), \tag{P}$$

where $f(x)$ is a convex function and $\mathcal{X}$ is a closed and convex constraint set. We denote $f^* \coloneqq \inf_{x \in \mathcal{X}} f(x)$. For simplicity, we assume that the minimum of f is attainable and denote x^* as a solution of (P); when the minimum is not attainable, we can instead fix x^* as an approximate solution whose objective is close to f^*. Throughout this paper, we assume that f is bounded below, e.g., $f^* > -\infty$. The projected SGD algorithm is given in Algorithm 1, where $\Pi_{\mathcal{X}}(\cdot)$ denotes the projection operator onto the set $\mathcal{X}$, i.e., $\Pi_{\mathcal{X}}(x) = \arg\min_{y \in \mathcal{X}} \|x - y\|$ for any $x \in \mathbb{R}^d$. For convenience, we use SGD as a shortcut of projected SGD in the following content. We impose the following assumption on the gradient noises of Algorithm 1.

Assumption 3.1 (Bounded and unbiased gradient noise)**.** *There exist $M > 0$ such that*

$$\mathbb{E}[\xi^{(t)} \mid x^{(t)}, \ldots, x^{(1)}] = 0 \qquad \textit{and} \qquad \|\xi^{(t)}\| \leq M \quad \text{a.s.} \quad \forall t \in \mathbb{N},$$

where $\xi^{(t)}$'s are the gradient noises of Algorithm 1.

The above assumption ensures that our estimation for the gradient is unbiased and the error is bounded by some constant in all iterations. Note that it is possible to relax the bounded noise assumption to the sub-Gaussian (light-tail) noise assumption, and all conclusions established in this work still hold but with some additional poly-logarithmic terms. For the simplicity of analysis and cleanness of presentation, we assume the noise is bounded almost sure.

Algorithm 1 Stochastic gradient descent (SGD)

Input: number of iteration $T \in \mathbb{N}_+$, initial iterate $x^{(1)}$, learning rate $\eta > 0$.
for $t \leftarrow 1, \dots, T-1$ **do**
  $\hat{g}^{(t)} = g^{(t)} + \xi^{(t)},\ g^{(t)} \in \partial f(x^{(t)})$ ▷ get an unbiased estimation for gradient
  $x^{(t+1)} = \Pi_{\mathcal{X}}(x^{(t)} - \eta \hat{g}^{(t)})$ ▷ update iterate
end for
Output: averaged iterate $x_{\text{out}} = \frac{1}{T}\sum_{t=1}^{T} x^{(t)}$.

We restate the generalized Freedman's inequality from Harvey et al. (2019), which serves as the key technical tool for our analysis.

Lemma 3.2 (Harvey et al., 2019, Theorem 3.3)**.** *Let $\{d_i, \mathcal{F}_i\}_{i=1}^T$ be a martingale difference sequence. Suppose $v_{i-1} \geq 0, \forall i \in [T]$ are $\mathcal{F}_{i-1}$-measurable random variables such that $\mathbb{E}[\exp(\lambda d_i) \mid \mathcal{F}_{i-1}] \leq \exp(\frac{\lambda^2}{2} v_{i-1})$ for all $i \in [T], \lambda > 0$. Let $S_t = \sum_{i=1}^t d_i$ and $V_t = \sum_{i=1}^t v_{i-1}$. Let $\delta \in (0,1)$ and suppose there are positive values $R(\delta) > 0$ and non-negative values $\{\alpha_i\}_{i=1}^T$ such that $\Pr\left[V_T \leq \sum_{i=1}^T \alpha_i d_i + R(\delta)\right] \geq 1 - \delta$. Then*

$$\Pr\left[S_T \geq x\right] \ \leq\ \delta + \exp\left(-\frac{x^2}{4\max_{i\in[T]} \alpha_i x + 8R(\delta)}\right) \quad \forall x > 0.$$

4 Classic convergence analysis

We review the classic convergence analysis of SGD for convex problems and describe why it is challenge to derive a high probability error bound in the form of $\mathcal{O}(\text{polylog}(1/\delta)/\sqrt{T})$, where $\delta \in (0,1)$ is the confidence level. We assume that f is convex and G-Lipschitz continuous for some $G > 0$ in this section.

Now we present the classic convergence analysis of SGD. Denote $\{x^{(t)}\}_{t=1}^T$ as the iterates generated from Algorithm 1. For any $t \in \mathbb{N}$, we have

$$\begin{aligned}
\|x^{(t+1)} - x^*\|^2 \ &= \ \|\Pi_{\mathcal{X}}\big(x^{(t)} - \eta(g^{(t)} + \xi^{(t)})\big) - x^*\|^2 \\
&\overset{\text{(i)}}{\leq}\ \|x^{(t)} - \eta(g^{(t)} + \xi^{(t)}) - x^*\|^2 \\
&=\ \|x^{(t)} - x^*\|^2 - 2\eta\langle g^{(t)}, x^{(t)} - x^*\rangle - 2\eta\langle \xi^{(t)}, x^{(t)} - x^*\rangle + \eta^2\|g^{(t)} + \xi^{(t)}\|^2 \\
&\overset{\text{(ii)}}{\leq}\ \|x^{(t)} - x^*\|^2 - 2\eta(f(x^{(t)}) - f^*) - 2\eta\langle \xi^{(t)}, x^{(t)} - x^*\rangle + 2\eta^2(G^2 + M^2),
\end{aligned} \tag{1}$$

where (i) is by the non-expansiveness of the projection operator and (ii) comes from the convexity of f. Rearranging the above inequality leads to

$$2\eta(f(x^{(t)}) - f^*) \ \leq\ \|x^{(t)} - x^*\|^2 - \|x^{(t+1)} - x^*\|^2 - 2\eta\langle \xi^{(t)}, x^{(t)} - x^*\rangle + 2\eta^2(G^2 + M^2).$$

Summing the above inequality over $t \in \{1, \dots, T\}$ and divide both sides by $2\eta T$. We obtain

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}(f(x^{(t)}) - f^*) \ \leq\ \frac{\|x^{(1)} - x^*\|^2}{2\eta T} - \frac{1}{T}\sum_{t=1}^{T}\langle \xi^{(t)}, x^{(t)} - x^*\rangle + \eta(G^2 + M^2) \\
\overset{\text{(i)}}{\Longrightarrow}\ &f(x_{\text{out}}) - f^* \ \leq\ \underbrace{\frac{\|x^{(1)} - x^*\|^2}{2\eta T} + \eta(G^2 + M^2)}_{:=X} + \underbrace{\frac{1}{T}\sum_{t=1}^{T}\langle \xi^{(t)}, x^* - x^{(t)}\rangle}_{:=Z},
\end{aligned} \tag{2}$$

where (i) is by the convexity of f. In order to prove $f(x_{\text{out}}) - f^* = \mathcal{O}\big(\text{polylog}(1/\delta)/\sqrt{T}\big)$ with probability at least $1 - \delta$, we need to bound X and Z with $\mathcal{O}\big(\text{polylog}(1/\delta)/\sqrt{T}\big)$ respectively.

- Bounding the term X is easy. It is obvious that $X = \mathcal{O}(1/\sqrt{T})$ by setting $\eta \propto 1/\sqrt{T}$.

- The challenge comes from the tail bound of the term Z. When the domain $\mathcal{X}$ has bounded diameter, i.e., $\sup_{x,y\in\mathcal{X}} \|x-y\| \leq R$ for some $R>0$, we can apply the Azuma's inequality or the Freedman's inequality and obtain $Z = \mathcal{O}(\sqrt{\log(1/\delta)}/\sqrt{T})$. However, when $\mathcal{X}$ is unbounded, for example $\mathcal{X}=\mathbb{R}^d$, it is unclear if $\max_{t\in[T]} \|x^{(t)}-x^*\|$ can be bounded by some constant, and this technical issue prevent us from using the Azuma's inequality or the Freedman's inequality.

We list several remarks for the above analysis before proceeding to our results.

Remark 4.1. When analyzing the convergence of SGD on expectation, the term Z in eq. (2) can be ignored since $\mathbb{E}[Z]=0$. Because of the term Z, the analysis of high probability error bounds is usually more complicated than the analysis of expected error bounds.

Remark 4.2. When f is μ-strongly convex for some $\mu>0$, it is already well-established that the distance between iterates and the solution ($\|x^{(t)}-x^*\|$) is uniformly bounded (Rakhlin et al., 2012). Therefore the convergence of SGD for strongly convex problems holds no matter the domain is bounded or not. Consequently, strong convexity will trivialize our analysis, and we do not include strongly convex problems in our discussion.

Remark 4.3. When f is non-convex, without further assumptions, classic analyses usually resort to the convergence to stationary points instead of the global minimum. Interestingly, in the analysis of the convergence to stationary points, the term $\langle \xi^{(t)}, x^*-x^{(t)}\rangle$ is replaced by $\langle \xi^{(t)}, \nabla f(x^{(t)})\rangle$ and the tail bound of the latter term is easy to derive; see Appendix E for a brief discussion. Consequently, the high probability error bound of SGD for non-convex problems holds regardless of the bounded domain assumption, and we exclude non-convex problems from our discussion.

5 High probability error bound of SGD for convex and Lipschitz functions

In this section, we derive a high probability error bound of SGD for Lipschitz continuous problems without the bounded domain assumption step by step. All missing proofs are placed in Appendix. We begin with a lemma that upper bounds the distance between the iterates generated by Algorithm 1 and the solution, i.e., $\|x^{(t)}-x^*\|^2$.

Lemma 5.1. *Suppose that f is convex and G-Lipschitz for some $G>0$ and Assumption 3.1 holds. Denote $\{x^{(t)}\}_{t=1}^T$ as the iterates from Algorithm 1 with learning rate $\eta>0$. Then for any $t\in[T]$,*

$$\|x^{(t)}-x^*\|^2 \leq \|x^{(1)}-x^*\|^2 + 2\eta\sum_{i=1}^{t-1}\langle \xi^{(i)}, x^*-x^{(i)}\rangle + 2\eta^2(t-1)(G^2+M^2).$$

As discussed in the previous section, the tail bound of the martingale sequence Z in eq. (2) is the key to establishing the high probability error bound of SGD. When deriving the tail bound of a martingale sequence, it is natural to begin with its total conditional variance (TCV). Based on Lemma 5.1, we can characterize the total conditional variance (TCV) of $\frac{1}{T}\sum_{i=1}^t\langle \xi^{(i)}, x^*-x^{(i)}\rangle$ for any $t\in[T]$ in the following proposition. Proposition 5.2 serves as a technical preparation for applying the generalized Freedman's inequality.

Proposition 5.2. *Suppose that f is convex and G-Lipschitz for some $G>0$ and Assumption 3.1 holds. Denote $\{x^{(t)}\}_{t=1}^T$ as the iterates from Algorithm 1 with learning rate $\eta>0$. For any $t\in[T]$, let $\mathcal{F}_{t-1}$ be the σ-algebra generated from $\{x^{(1)},\ldots,x^{(t)}\}$, $d_t \coloneqq \frac{1}{T}\langle \xi^{(t)}, x^*-x^{(t)}\rangle$, $v_{t-1} \coloneqq \frac{M^2}{T^2}\|x^{(t)}-x^*\|^2$ and define $V_t=\sum_{i=1}^t v_{i-1}$. Then for any $t\in[T]$,*

$$\mathbb{E}[\exp(\lambda d_t)\mid \mathcal{F}_{t-1}] \leq \exp\left(\frac{\lambda^2}{2}v_{t-1}\right) \qquad \forall \lambda\in\mathbb{R}$$

and

$$V_t \leq \frac{M^2\|x^{(1)}-x^*\|^2}{T} + \frac{2M^2\eta}{T}\sum_{i=1}^{t-1}(t-i)d_i + 2M^2(G^2+M^2)\eta^2. \tag{3}$$

Equation (3) suggests that the total conditional variance of the martingale sequence $\sum_{i=1}^{t} d_i$ is bounded by a linear transformation of the martingale sequence itself, which is referred as the "chicken and egg" phenomenon by Harvey et al. (2019). With Proposition 5.2, we are now ready to apply the generalized Freedman's inequality (Lemma 3.2) and derive a high probability upper bound of the term $\frac{1}{T}\sum_{t=1}^{T}\langle \xi^{(t)}, x^* - x^{(t)}\rangle$.

Proposition 5.3. *Suppose f is convex and G-Lipschitz for some $G > 0$ and Assumption 3.1 holds. Denote $\{x^{(t)}\}_{t=1}^{T}$ as the iterates generated from Algorithm 1 with learning rate $\eta > 0$ and let $D_X \coloneqq \|x^{(1)} - x^*\|$. Then for any $t \in [T]$ and $\delta \in (0,1)$,*

$$\frac{1}{T}\sum_{i=1}^{t}\langle \xi^{(i)}, x^* - x^{(i)}\rangle \;\le\; 16M^2\eta\log(1/\delta) + 4\left(\frac{MD_X}{\sqrt{T}} + \sqrt{2}M\eta\sqrt{G^2+M^2}\right)\sqrt{\log(1/\delta)}$$

with probability at least $1-\delta$.

Setting $\eta \propto 1/\sqrt{T}$, Proposition 5.3 immediately gives a $\mathcal{O}(\log(1/\delta)/\sqrt{T})$ tail bound of the term Z and therefore further leads to a $\mathcal{O}(\log(1/\delta)/\sqrt{T})$ high probability error bound of Algorithm 1. Formally, the following theorem is established.

Theorem 5.4. *Suppose f is convex and G-Lipschitz for some $G > 0$ and Assumption 3.1 holds. Denote x_{out} as the output of Algorithm 1 with learning rate $\eta = \min\{1/(M\sqrt{T}), 1/\sqrt{T}\}$. Then*

$$f(x_{\text{out}}) - f^* \;=\; \mathcal{O}\left(\frac{\log(1/\delta)}{\sqrt{T}}\right)$$

with probability at least $1-\delta$ for any $\delta \in (0,1)$.

Note that we directly bounded the term $\frac{1}{T}\sum_{t=1}^{T}\langle \xi^{(t)}, x^* - x^{(t)}\rangle$ in the above analysis, and we have not prove if the distance between iterates and solution, e.g., $\|x^{(t)} - x^*\|$ can be bounded during the optimization procedure. Interestingly, Lemma 5.1 allows us to further translate the upper bound of $\frac{1}{T}\sum_{i=1}^{t}\langle \xi^{(i)}, x^* - x^{(i)}\rangle$ to $\|x^{(t)} - x^*\|$. The following Theorem makes this precise.

Theorem 5.5. *Suppose f is convex and G-Lipschitz for some $G > 0$ and Assumption 3.1 holds. Denote $\{x^{(t)}\}_{t=1}^{T}$ as the iterates generated from Algorithm 1 with learning rate $\eta \le \min\{1/(M\sqrt{T}), 1/\sqrt{T}\}$. Then*

$$\max_{t\in[T]}\|x^{(t)} - x^*\| \;\le\; D_X + \sqrt{2(G^2+M^2)} + 4\sqrt{2\log(T/\delta)}$$

with probability at least $1-\delta$ for any $\delta \in (0,1)$, where $D_X \coloneqq \|x^{(1)} - x^*\|$.

Theorem 5.5 is a useful result as it demonstrates that the iterates generated from the SGD algorithm are guaranteed to stay in a neighbourhood of the initialization whose diameter only logarithmically depends on the number of iteration T. This convenient property of SGD allows us to relax the bounded domain assumption used in existing convergence analysis of SGD almost for free; the cost is just some logarithmic terms. We show some direct applications of Theorem 5.5 in Section 7.1.

Remark 5.1. All results in this section are based on the fixed learning rate scheduling $\eta = 1/\sqrt{T}$, which requires us to know the total number of iterations T as a priori. When T is not known in advance, dynamic learning rate scheduling $\eta_t \propto 1/\sqrt{t}$ is usually adopted. Following exactly the same proof template and some tedious calculation, it is easy to show that all results in this section also hold for the dynamic learning rate scheduling $\eta_t \propto 1/\sqrt{t}$. For the simplicity of the presentation, we omit the error bounds with the dynamic learning rate.

6 HIGH PROBABILITY ERROR BOUND OF SGD FOR CONVEX AND SMOOTH FUNCTIONS

In this section, we focus on the high probability error bound of SGD for smooth and convex functions that may not be globally Lipschitz continuous. This class of functions is quite common in practice, for example the unconstrained least-square problem.

For convex functions, it is well-known that SGD applies to the Lipschitz continuous and smooth functions enjoys similar convergence rate on expectation (Ghadimi & Lan, 2013). Therefore it is natural to conjecture that the same conclusion also holds for the high probability error bound. In the following content of this section, we show that this is indeed the case; with some minor modifications to the proofs, all the propositions and theorems established in Section 5 also hold for smooth and convex functions. To avoid tedious repetition, we only state the main conclusions in this section and place the auxiliary propositions and detailed proofs in Appendix.

Lemma 6.1. *Suppose f is convex and L-smooth for some $L > 0$ and Assumption 3.1 holds. Denote $\{x^{(t)}\}_{t=1}^{T}$ as the iterates from Algorithm 1 with the learning rate $\eta \in (0, 1/(2L)]$. Then for any $t \in [T]$,*

$$\|x^{(t)} - x^*\|^2 \;\leq\; \|x^{(1)} - x^*\|^2 + 2\eta \sum_{i=1}^{t-1} \langle \xi^{(i)}, x^* - x^{(i)} \rangle + 2(t-1)\eta^2 (2\|\nabla f(x^*)\|^2 + M^2).$$

Lemma 6.1 is very similar to Lemma 5.1, the minor differences are: (i) we need to require $\eta \in (0, 1/(2L)]$ in Lemma 6.1 instead of $\eta > 0$ in Lemma 5.1; (ii) the Lipschitz parameter G^2 in Lemma 5.1 is replaced by the gradient norm at solution, e.g., $2\|\nabla f(x^*)\|^2$ in Lemma 6.1. Then it is straightforward to see that Proposition 5.2 and Proposition 5.3 should also hold for convex and smooth objectives with some minor modifications. We directly state the convergence result and describe the modified versions of Proposition 5.2 and Proposition 5.3 in the Appendix.

Theorem 6.2. *Suppose f is convex and L-smooth for some $L > 0$ and Assumption 3.1 holds. Denote x_{out} as the output of Algorithm 1 with learning rate $\eta = \min\{1/(2L), 1/(M\sqrt{T}), 1/\sqrt{T}\}$. Then*

$$f(x_{\text{out}}) - f^* \;=\; \mathcal{O}\left(\frac{\log(1/\delta)}{\sqrt{T}}\right)$$

with probability at least $1 - \delta$ for any $\delta \in (0, 1)$.

Similar to Theorem 5.5, we can also develop an upper bound on the distance to solution for the trajectory of SGD applies to convex and smooth objectives. The only difference is that we need to use Lemma 6.1 instead of Lemma 5.1 during the proof.

Theorem 6.3. *Suppose f is convex and L-smooth for some $L > 0$ and Assumption 3.1 holds. Denote $\{x^{(t)}\}_{t=1}^{T}$ as the iterates generated from Algorithm 1 with learning rate $\eta \leq \min\{1/(2L), 1/(M\sqrt{T}), 1/\sqrt{T}\}$. Then*

$$\max_{t \in [T]} \|x^{(t)} - x^*\| \;\leq\; D_X + \sqrt{4\|\nabla f(x^*)\|^2 + 2M^2} + 4\sqrt{2\log(T/\delta)}$$

with probability at least $1 - \delta$ for any $\delta \in (0, 1)$.

As an immediate consequence of Theorem 6.3, we can show that the iterates of SGD has bounded gradient norm.

Corollary 6.4. *Under the same conditions as Theorem 6.3. For any $x \in \text{conv}\{x^{(1)}, \ldots, x^{(T)}\}$,*

$$\|\nabla f(x)\| \;\leq\; \|\nabla f(x^*)\| + LD_X + L\sqrt{4\|\nabla f(x^*)\|^2 + 2M^2} + 4L\sqrt{2\log(T/\delta)}$$

with probability at least $1 - \delta$ for any $\delta \in (0, 1)$.

Remark 6.1. Although smooth functions are not necessarily globally Lipschitz continuous in its domain, Corollary 6.4 indicates that the trajectory of SGD is guaranteed to enjoy a finite Lipschitz constant with high probability. Therefore when analyzing the theoretical properties of SGD, Corollary 6.4 suggests that the Lipschitz continuity assumption can be *removed almost for free* if smoothness is assumed; the cost is just an additional $\sqrt{\log(T/\delta)}$ term.

7 Extensions

We present some extensions of the analysis in previous sections. In particular, we show that the high probability error bound of the last iterate of SGD and SGD with momentum can be established without the bounded domain assumption. We also briefly discuss differential-private SGD (DP-SGD) as another potential application of Theorem 5.5 and Theorem 6.3.

7.1 Convergence of the last iterate

The high probability error bounds established in Section 5 and Section 6 are based on the averaged iterate, that is $x_{\text{out}} = \frac{1}{T}\sum_{t=1}^{T} x^{(t)}$. In fact, most classic convergence analyses of SGD did focus on the averaged iterate because the averaged iterate is easy to analyze and can usually yield clear proofs.

However, the averaged iterate is rarely used in practice, and practitioners usually prefer to use the last iterate as the output of SGD. This gap between theory and practice has received substantial interest in recent years, and both convergences on expectation and high probability convergence rates of the last iterate of SGD have been established nowadays (Zhang, 2004; Rakhlin et al., 2012; Shamir & Zhang, 2013; Harvey et al., 2019). However, to our knowledge, all existing analyses require the domain $\mathcal{X}$ to have a bounded diameter, which is a rather restricted setting. In this section, as simple corollaries of Theorem 5.5 and Theorem 6.3, we show that the bounded domain assumption can be relaxed almost for free.

Before proceeding to the conclusion, we first review an existing high probability error bound of the last iterate of SGD with the bounded domain assumption, which is given by Harvey et al. (2019).

Theorem 7.1 (Harvey et al., 2019, Theorem 3.2)**.** *Assume f is convex and G-Lipschitz for some $G > 0$, $\operatorname{diam}(\mathcal{X}) \le R$ for some $R > 0$ and Assumption 3.1 holds. Denote $\{x^{(t)}\}_{t=1}^{T}$ as the iterates generated by Algorithm 1 with $\eta = 1/\sqrt{T}$. Then*

$$f(x^{(T)}) - f^* \;=\; \mathcal{O}\left(\frac{\log(T)\log(1/\delta)}{\sqrt{T}}\right)$$

with probability at least $1-\delta$.

Note that Theorem 7.1 is slightly different from Harvey et al., 2019, Theorem 3.2 where we adopt constant learning rate $\eta = 1/\sqrt{T}$ instead of the dynamic learning rate $\eta_t = 1/\sqrt{t}$ used by Harvey et al. (2019). A careful examination of the proof of Harvey et al., 2019, Theorem 3.2 concludes that it is sufficient to assume $\max_{t\in[T]} \|x^{(t)} - x^*\| \le R$ instead of assuming $\operatorname{diam}(\mathcal{X}) \le R$ for the derivation. Therefore, combining Theorem 7.1 and Theorem 5.5 immediately leads to the following corollary.

Corollary 7.2. *Assume f is convex and G-Lipschitz for some $G > 0$ (or L-smooth for some $L > 0$) and Assumption 3.1 holds. Denote $\{x^{(t)}\}_{t=1}^{T}$ as the iterates generated by Algorithm 1 with $\eta = \min\{1/(M\sqrt{T}), 1/\sqrt{T}\}$. Then*

$$f(x^{(T)}) - f^* \;=\; \widetilde{\mathcal{O}}\left(\frac{1}{\sqrt{T}}\right)$$

with probability at least $1-\delta$.

Corollary 7.2 achieves essentially the same error bound as Theorem 7.1 up to poly-logarithmic terms but without assuming $\mathcal{X}$ to have bounded diameter, e.g., $\operatorname{diam}(\mathcal{X}) < \infty$.

7.2 SGD with momentum

We consider the SGD with momentum (mSGD), also known as the heavy-ball method. The detailed algorithm is shown in Algorithm 2. Note that we present mSGD in the iterate-moving-average form for the ease of analysis; mSGD may be equivalently formulated as other forms (Gower, 2022).

Algorithm 2 Stochastic gradient descent with momentum (mSGD)

Input: number of iterations $T \in \mathbb{N}_+$, initial iterate $x^{(1)}$, $\eta > 0$, $\lambda_t : \mathbb{N} \to \mathbb{R}_+$.
$z^{(1)} = x^{(1)}$
for $t \leftarrow 1, \dots, T-1$ **do**
  $\hat{g}^{(t)} = g^{(t)} + \xi^{(t)},\ g^{(t)} \in \partial f(x^{(t)})$ ▷ get an unbiased estimation for gradient
  $z^{(t+1)} = \Pi_{\mathcal{X}}\big(z^{(t)} - \eta_t \hat{g}^{(t)}\big)$ ▷ update the iterate z
  $x^{(t+1)} = \dfrac{\lambda_t}{\lambda_t + 1} x^{(t)} + \dfrac{1}{\lambda_t + 1} z^{(t+1)}$ ▷ update the iterate x
end for
Output: the last iterate $x_{\text{out}} = x^{(T)}$.

Following almost the same proof templates described in Section 5 and Section 6, we can obtain the following high probability error bound of mSGD.

Theorem 7.3. *Suppose f is convex and G-Lipschitz for some $G > 0$ and Assumption 3.1 holds. Denote $\{x^{(t)}\}_{t=1}^T$ as the iterates generated from Algorithm 2 with learning rate $\eta = 1/\sqrt{T}$ and $\lambda_t = t$. Then*

$$f(x^{(T)}) - f^* \ = \ \mathcal{O}\left(\frac{\log(1/\delta)}{\sqrt{T}}\right)$$

with probability at least $1 - \delta$ for any $\delta \in (0, 1)$.

Again, following the same proof template, we can also develop a high probability bound on the distance between iterates and solution.

Theorem 7.4. *Suppose f is convex and G-Lipschitz for some $G > 0$ and Assumption 3.1 holds. Denote $\{x^{(t)}\}_{t=1}^T$ and $\{z^{(t)}\}_{t=1}^T$ as the iterates generated from Algorithm 2 with learning rate $\eta \le \min\{1/(M\sqrt{T}), 1/\sqrt{T}\}$ and $\lambda_t = t$. Then*

$$\max_{t \in [T]} \|z^{(t)} - x^*\| \ \le \ D_X + \sqrt{2(G^2 + M^2)} + 4\sqrt{2\log(T/\delta)}$$

with probability at least $1 - \delta$ for any $\delta \in (0, 1)$.

Note that the above distance to solution is based on $\{z^{(t)}\}_{t=1}^T$, it is straightforward that similar result also holds for $\{x^{(t)}\}_{t=1}^T$ by noticing $x^{(t+1)}$ is a convex combination of $x^{(t)}$ and $z^{(t+1)}$. We omit the details.

7.3 Differential-private SGD

Training models without leaking information of the training data is an emerging research topic in machine learning. The differential-private SGD (DP-SGD) is a simple and effective algorithm that can approximately solve the empirical-risk minimization problem with a privacy guarantee. Due to the popularity of DP-SGD, the convergence of DP-SGD in various scenarios has been extensively studied in recent years (Song et al., 2013; Bassily et al., 2014; Wang et al., 2017; Bassily et al., 2019; Feldman et al., 2020; Asi et al., 2021).

Lipschitz continuity plays a crucial role in the analysis of DP-SGD. When the objective is smooth but not globally Lipschitz continuous, for example, the unconstrained least square problem, bounded domain assumption is usually required. There have been some recent discussions on removing the bounded domain assumption used in the analysis of DP-SGD; see for example Wang et al., 2022, Remark 5. Given that the convergence analysis of DP-SGD in the literature of differential-privacy usually follows the same proof template as the convergence of SGD (Bassily et al., 2014), thus our Theorem 5.5, Theorem 6.3 and Corollary 6.4 have the potential to relax the bounded domain assumption used in the existing convergence analysis of DP-SGD. We conjecture that the bounded domain assumption can be removed when analyzing DP-SGD; the costs are some additional poly-logarithmic terms on the error bound and the light-tail-noise assumption Assumption 3.1. Exploring this direction in depth would deviate too far from the main purpose of this paper, and we left this tiny conjecture as a future research topic.

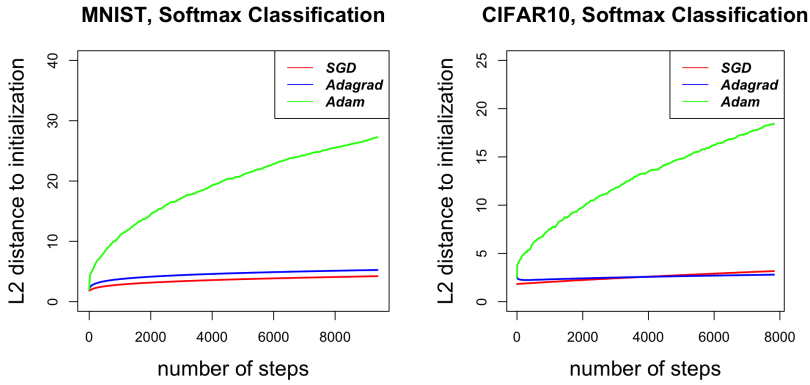

Figure 1: The evolution of the distance between iterates and initialization ($\|x^{(t)} - x^{(1)}\|$) on the MNIST and CIFAR datasets with softmax classification.

8 Numerical Study

We conduct some numerical experiments to complement our analysis in previous sections. We perform softmax classification on two standard image classification datasets: the MNIST (LeCun & Cortes, 2011) and CIFAR10 (Krizhevsky, 2009) datasets. We use SGD, Adagrad, and Adam with constant learning rates and the same initialization to train the model, respectively. We set the learning rate to be 0.01 and the batch size to be 128 for all experiments. All experiments are conducted on a server with 64GB memory, 32 CPUs, and 2 NVIDIA 3090 GPUs.

For each training algorithm, we plot the evolution of the 2-norm distance between iterates and the initialization, i.e., $\|x^{(t)} - x^{(1)}\|$. The results are shown in Figure 1. From Figure 1, we can observe that the iterate-initialization-distance is almost uniformly bounded among all training iterations for the SGD and Adagrad optimizers. Given that $\max_{t\in[T]} \|x^{(t)} - x^{(1)}\| \leq C$ for some $C > 0$ implies $\max_{i,j\in[T]} \|x^{(i)} - x^{(j)}\| \leq 2C$. Figure 1 indicates that the iterates generated from the SGD and Adagrad optimizers tend to stay in a bounded region, which is consistent with our theoretical analysis on SGD in Section 5 and Section 6. However, for the Adam optimizer, the distance between iterates and initialization tends to grow quickly as the number of iterations increases. This empirical observation suggests the iterates generated from Adam may not stay in a bounded region.

9 Conclusion

In this paper, we studied the high probability error bounds of SGD under an unbounded domain. We developed concise and flexible proof templates showing that SGD applied to convex problems with possibly unbounded domain attains the $\mathcal{O}(\log(1/\delta)/\sqrt{T})$ error bound. Our theoretical results also indicate that the iterates generated from SGD will stay in a bounded region with a high probability, which further suggests that the bounded domain assumption can be removed almost for free when analyzing the high probability error bound of SGD. Simple corollaries of our analysis show that the high probability error bound on the last iterate of SGD and mSGD can also be established.

Future directions remain. The classic convergence analysis of Adagrad algorithm also relies crucially on the bounded domain assumption (Duchi et al., 2011). Different from the SGD algorithm, the analysis of Adagrad requires the bounded domain assumption even when analyzing the convergence in terms of expectation. Therefore, whether it is possible to remove the bounded domain assumption used in the convergence analysis of Adagrad is an important research question to be answered. Another possible direction is to extend the results in this paper to composite problems in the form $f + g$, where g is some nonsmooth convex regularizers. More specifically, one can study the high probability error bound of proximal-SGD for composite problems based on the proof techniques developed by this paper. Finally, as mentioned in Section 7.3, the proof template in this paper can potentially improve the analysis of DP-SGD, which may be of interest to the DP community.

REFERENCES

Zeyuan Allen-Zhu. Katyusha: The first direct acceleration of stochastic gradient methods. *Journal of Machine Learning Research*, 18:221:1–221:51, 2017a.

Zeyuan Allen-Zhu. Natasha: Faster non-convex stochastic optimization via strongly non-convex parameter. In *Proceedings of ICML*, volume 70, pp. 89–97, 2017b.

Hilal Asi, John C. Duchi, Alireza Fallah, Omid Javidbakht, and Kunal Talwar. Private adaptive gradient methods for convex optimization. In *Proceedings of the International Conference on Machine Learning, ICML*, 2021.

Raef Bassily, Adam D. Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *IEEE Annual Symposium on Foundations of Computer Science, FOCS*, pp. 464–473, 2014.

Raef Bassily, Vitaly Feldman, Kunal Talwar, and Abhradeep Guha Thakurta. Private stochastic convex optimization with optimal rates. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3-4):231–357, 2015.

Ashok Cutkosky and Harsh Mehta. High-probability bounds for non-convex stochastic optimization with heavy tails. In *Advances in Neural Information Processing Systems, NeurIPS*, pp. 4883–4895, 2021.

Damek Davis, Dmitriy Drusvyatskiy, Lin Xiao, and Junyu Zhang. From low probability to high confidence in stochastic convex optimization. *J. Mach. Learn. Res.*, 22:49:1–49:38, 2021.

Aaron Defazio, Francis R. Bach, and Simon Lacoste-Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems, NeurIPS*, pp. 1646–1654, 2014.

John C. Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.*, 12:2121–2159, 2011.

Huang Fang, Zhenan Fan, and Michael Friedlander. Fast convergence of stochastic subgradient method under interpolation. In *Proceedings of ICLR*, 2021.

Vitaly Feldman and Jan Vondrák. High probability generalization bounds for uniformly stable algorithms with nearly optimal rate. In *Conference on Learning Theory, COLT*, 2019.

Vitaly Feldman, Tomer Koren, and Kunal Talwar. Private stochastic convex optimization: optimal rates in linear time. In *Proccedings of ACM Symposium on Theory of Computing, STOC*, 2020.

Saeed Ghadimi and Guanghui Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

Robert M. Gower. Convergence theorems for gradient descent. *Technical Report*, 2022. URL https://gowerrobert.github.io/pdf/M2_statistique_optimisation/grad_conv.pdf.

Nicholas J. A. Harvey, Christopher Liaw, Yaniv Plan, and Sikander Randhawa. Tight analyses for non-smooth stochastic gradient descent. In *Conference on Learning Theory, COLT*, 2019.

Elad Hazan and Satyen Kale. Beyond the regret minimization barrier: optimal algorithms for stochastic strongly-convex optimization. *J. Mach. Learn. Res.*, 15(1):2489–2512, 2014.

Prateek Jain, Dheeraj M. Nagaraj, and Praneeth Netrapalli. Making the last iterate of SGD information theoretically optimal. *SIAM J. Optim.*, 31(2):1108–1130, 2021.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems, NeurIPS*, volume 26, pp. 315–323, 2013.

Sham M. Kakade and Ambuj Tewari. On the generalization ability of online strongly convex programming algorithms. In *Advances in Neural Information Processing Systems, NeurIPS*, 2008.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations, ICLR*, 2015.

Alex Krizhevsky. Learning multiple layers of features from tiny images, 2009.

Simon Lacoste-Julien, Mark Schmidt, and Francis Bach. A simpler approach to obtaining an O(1/t) convergence rate for the projected stochastic subgradient method. *arXiv e-prints*, art. arXiv:1212.2002, 2012.

Yann LeCun and Corinna Cortes. The MNIST database. http://yann.lecun.com/exdb/mnist/, 2011.

Xiaoyu Li and Francesco Orabona. A high probability analysis of adaptive sgd with momentum. *arXiv e-prints*, art. arXiv:2007.14294, July 2020.

Daogao Liu and Zhou Lu. The convergence rate of sgd's final iterate: Analysis on dimension dependence. *arXiv e-prints*, art. arXiv:2106.14588, June 2021.

Siyuan Ma, Raef Bassily, and Mikhail Belkin. The power of interpolation: Understanding the effectiveness of SGD in modern over-parametrized learning. In *Proceedings of the International Conference on Machine Learningz, ICML*, pp. 3331–3340, 2018.

Arkadi Nemirovski, Anatoli B. Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

Y. Nesterov. *Introductory Lectures on Convex Optimization*. Kluwer Academic, Dordrecht, The Netherlands, 2004.

B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.

Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. In *Proceedings of the International Conference on Machine Learningz, ICML*, 2012.

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

Mark Schmidt and Nicolas Le Roux. Fast convergence of stochastic gradient descent under a strong growth condition. *arXiv e-prints*, art. arXiv:1308.6370, Aug 2013.

Mark Schmidt, Nicolas Le Roux, and Francis R. Bach. Minimizing finite sums with the stochastic average gradient. *Math. Program.*, 162(1-2):83–112, 2017.

Shai Shalev-Shwartz, Yoram Singer, and Nathan Srebro. Pegasos: Primal estimated sub-gradient solver for svm. In *Proceedings of the International Conference on Machine Learningz, ICML*, pp. 807–814, 2007.

Ohad Shamir and Tong Zhang. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In *Proceedings of the International Conference on Machine Learning, ICML*, 2013.

Shuang Song, Kamalika Chaudhuri, and Anand D. Sarwate. Stochastic gradient descent with differentially private updates. In *IEEE Global Conference on Signal and Information Processing, GlobalSIP*, pp. 245–248, 2013.

Aditya Vardhan Varre, Loucas Pillaud-Vivien, and Nicolas Flammarion. Last iterate convergence of SGD for least-squares in the interpolation regime. In *Advances in Neural Information Processing Systems, NeurIPS*, 2021.

Sharan Vaswani, Aaron Mishkin, Issam H. Laradji, Mark Schmidt, Gauthier Gidel, and Simon Lacoste-Julien. Painless stochastic gradient: Interpolation, line-search, and convergence rates. In *Advances in Neural Information Processing Systems, NeurIPS*, 2019.

Di Wang, Minwei Ye, and Jinhui Xu. Differentially private empirical risk minimization revisited: Faster and more general. In *Proceedings of the Advances in Neural Information Processing Systems*, pp. 2722–2731, 2017.

Puyu Wang, Yunwen Lei, Yiming Ying, and Hai Zhang. Differentially private sgd with non-smooth losses. *Applied and Computational Harmonic Analysis*, 56:306–336, 2022.

Tong Zhang. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *Proceedings of the International Conference on Machine Learning, ICML*, 2004.

Wanrong Zhu, Zhipeng Lou, and Wei Biao Wu. Beyond sub-gaussian noises: Sharp concentration analysis for stochastic gradient descent. *J. Mach. Learn. Res.*, 23:46:1–46:22, 2022.

APPENDIX

A FACTS AND LEMMAS

Fact A.1. *For any $n \in \mathbb{N}_+$, $\sum_{i=1}^{n} 1/i \leq 1 + \log(n)$.*

Lemma A.2 (Azuma's inequality)**.** *Let $\{X_t\}_{t=1}^{\infty}$ be a martingale difference sequence such that $|X_{t+1} - X_t| \leq c_t$ with probability 1 for any $t \in \mathbb{N}_+$. Let $S_t = \sum_{i=1}^{t} X_i \; \forall t \in \mathbb{N}_+$. Then*

$$\Pr\left[S_T \geq \epsilon\right] \;\leq\; 2\exp\left(-\frac{\epsilon^2}{2\sum_{t=1}^{T} c_t^2}\right)$$

for any $\epsilon > 0$ and $T \in \mathbb{N}_+$.

Lemma A.3. *Suppose that f is convex and L-smooth on $\mathcal{X}$ for some $L > 0$. Then*

$$\|\nabla f(x)\|^2 \;\leq\; 2L(f(x) - f^*) + 2\|\nabla f(x^*)\|^2 \qquad \forall x \in \mathcal{X}.$$

Proof. By Nesterov (2004, Theorem 2.1.5), we know that

$$f(x) - f(y) - \langle \nabla f(y), x - y\rangle \;\geq\; \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|^2 \qquad \forall x, y \in \mathcal{X}.$$

Making the identification $x = x, y = x^*$, we obtain that

$$\begin{aligned}
&f(x) - f(x^*) - \langle \nabla f(x^*), x - x^*\rangle \;\geq\; \frac{1}{2L}\|\nabla f(x) - \nabla f(x^*)\|^2 \\
\overset{\text{(i)}}{\Longrightarrow}\; &f(x) - f(x^*) \;\geq\; \frac{1}{2L}\|\nabla f(x) - \nabla f(x^*)\|^2 \\
\Longrightarrow\; &\|\nabla f(x)\|^2 \;\leq\; 2\|\nabla f(x) - \nabla f(x^*)\|^2 + 2\|\nabla f(x^*)\|^2 \;\leq\; 2L(f(x) - f^*) + 2\|\nabla f(x^*)\|^2,
\end{aligned}$$

where (i) is by the first-order optimality condition, e.g., $\langle \nabla f(x^*), x - x^*\rangle \geq 0 \; \forall x \in \mathcal{X}$. □

B MISSING PROOFS FOR SECTION 5

Lemma 5.1. *Suppose that f is convex and G-Lipschitz for some $G > 0$ and Assumption 3.1 holds. Denote $\{x^{(t)}\}_{t=1}^{T}$ as the iterates from Algorithm 1 with learning rate $\eta > 0$. Then for any $t \in [T]$,*

$$\|x^{(t)} - x^*\|^2 \;\leq\; \|x^{(1)} - x^*\|^2 + 2\eta\sum_{i=1}^{t-1}\langle \xi^{(i)}, x^* - x^{(i)}\rangle + 2\eta^2(t-1)(G^2 + M^2).$$

Proof. For any $t \in \{1, 2, \ldots, T-1\}$

$$\begin{aligned}
\|x^{(t+1)} - x^*\|^2 &\overset{\text{(i)}}{\leq} \|x^{(t)} - x^*\|^2 - 2\eta(f(x^{(t)}) - f^*) - 2\eta\langle \xi^{(t)}, x^{(t)} - x^*\rangle + 2\eta^2(G^2 + M^2) \\
&\overset{\text{(ii)}}{\leq} \|x^{(t)} - x^*\|^2 + 2\eta\langle \xi^{(t)}, x^* - x^{(t)}\rangle + 2\eta^2(G^2 + M^2),
\end{aligned}$$

where (i) comes from eq. (1) and (ii) is true by noticing $\eta > 0$ and $f(x^{(t)}) \geq f^*$.

Apply the above inequality recursively, we obtain that

$$\|x^{(t)} - x^*\|^2 \;\leq\; \|x^{(1)} - x^*\|^2 + 2\eta\sum_{i=1}^{t-1}\langle \xi^{(i)}, x^* - x^{(i)}\rangle + 2\eta^2(t-1)(G^2 + M^2)$$

for any $t \in [T]$. □

Proposition 5.2. *Suppose that f is convex and G-Lipschitz for some $G > 0$ and Assumption 3.1 holds. Denote $\{x^{(t)}\}_{t=1}^{T}$ as the iterates from Algorithm 1 with learning rate $\eta > 0$. For any $t \in [T]$, let $\mathcal{F}_{t-1}$ be the σ-algebra generated from $\{x^{(1)}, \ldots, x^{(t)}\}$, $d_t \coloneqq \frac{1}{T}\langle \xi^{(t)}, x^* - x^{(t)}\rangle$, $v_{t-1} \coloneqq \frac{M^2}{T^2}\|x^{(t)} - x^*\|^2$ and define $V_t = \sum_{i=1}^{t} v_{i-1}$. Then for any $t \in [T]$,*

$$\mathbb{E}[\exp(\lambda d_t) \mid \mathcal{F}_{t-1}] \;\leq\; \exp\left(\frac{\lambda^2}{2}v_{t-1}\right) \qquad \forall \lambda \in \mathbb{R}$$

and

$$V_t \;\le\; \frac{M^2\|x^{(1)}-x^*\|^2}{T} + \frac{2M^2\eta}{T}\sum_{i=1}^{t-1}(t-i)d_i + 2M^2(G^2+M^2)\eta^2. \tag{3}$$

Proof. By the definition of d_i and Assumption 3.1, we have that

$$d_t^2 \overset{\text{(i)}}{\le} \frac{M^2}{T^2}\|x^{(t)}-x^*\|^2 \implies \mathbb{E}[\exp(\lambda d_t)\mid \mathcal{F}_{t-1}] \;\le\; \exp\left(\frac{\lambda^2}{2}\frac{M^2}{T^2}\|x^{(t)}-x^*\|^2\right) \qquad \forall\lambda\in\mathbb{R},$$

where (i) is by the fact that $d_t^2 \le \|\xi^{(t)}\|^2\|x^{(t)}-x^*\|^2/T^2$. Then we derive the upper bound of V_t,

$$\begin{aligned}
V_t \;&=\; \sum_{i=1}^{t} v_{i-1} \\
&=\; \frac{M^2}{T^2}\sum_{i=1}^{t}\|x^{(i)}-x^*\|^2 \\
&\overset{\text{(i)}}{\le}\; \frac{M^2}{T^2}\sum_{i=1}^{t}\left(\|x^{(1)}-x^*\|^2 + 2\eta\sum_{j=1}^{i-1}\langle\xi^{(j)}, x^*-x^{(j)}\rangle + 2\eta^2(i-1)(G^2+M^2)\right) \\
&\le\; \frac{M^2\|x^{(1)}-x^*\|^2}{T} + \frac{2M^2\eta}{T}\sum_{i=1}^{t}\sum_{j=1}^{i-1} d_j + 2M^2(G^2+M^2)\eta^2 \\
&\le\; \frac{M^2\|x^{(1)}-x^*\|^2}{T} + \frac{2M^2\eta}{T}\sum_{i=1}^{t-1}(t-i)d_i + 2M^2(G^2+M^2)\eta^2,
\end{aligned} \tag{4}$$

where (i) is by Lemma 5.1. This finishes the proof. □

Proposition 5.3. *Suppose f is convex and G-Lipschitz for some $G>0$ and Assumption 3.1 holds. Denote $\{x^{(t)}\}_{t=1}^T$ as the iterates generated from Algorithm 1 with learning rate $\eta>0$ and let $D_X := \|x^{(1)}-x^*\|$. Then for any $t\in[T]$ and $\delta\in(0,1)$,*

$$\frac{1}{T}\sum_{i=1}^{t}\langle\xi^{(i)}, x^*-x^{(i)}\rangle \;\le\; 16M^2\eta\log(1/\delta) + 4\left(\frac{MD_X}{\sqrt{T}} + \sqrt{2}M\eta\sqrt{G^2+M^2}\right)\sqrt{\log(1/\delta)}$$

with probability at least $1-\delta$.

Proof. Let

$$d_t := \frac{1}{T}\langle\xi^{(t)}, x^*-x^{(t)}\rangle, \quad S_t := \sum_{i=1}^{t} d_i, \quad v_{t-1} := \frac{M^2\|x^{(t)}-x^*\|^2}{T^2}, \quad V_t := \sum_{i=1}^{t} v_i$$

for any $t\in[T]$. Given $t\in[T]$, apply Proposition 5.2 and Lemma 3.2 and making the identification

$$\alpha_i = \frac{2M^2\eta(t-i)}{T} \quad \forall i\in[T], \qquad R := \frac{M^2\|x^{(1)}-x^*\|^2}{T} + 2M^2\eta^2(G^2+M^2).$$

We obtain

$$\Pr[S_t\ge x] \;\le\; \exp\left(-\frac{x^2}{4\max_{i\in[t]}\alpha_i x + 8R}\right) \qquad \forall x>0.$$

Let $\alpha = \max_{i\in[t]}\alpha_i$ and set $x = \max\{8\alpha\log(1/\delta), 4\sqrt{R\log(1/\delta)}\}$ gives

$$\Pr\left[S_t \ge \max\{8\alpha\log(1/\delta), 4\sqrt{R\log(1/\delta)}\}\right] \le \delta \qquad \forall\delta\in(0,1).$$

Plug in α and $R(\delta)$, we obtain

$$\begin{aligned}
&\Pr\left[S_t \geq \max\{8\alpha\log(1/\delta), 4\sqrt{R\log(1/\delta)}\}\right] \leq \delta \\
\implies &\Pr\left[S_t \geq 8\alpha\log(1/\delta) + 4\sqrt{R\log(1/\delta)}\right] \leq \delta \\
\implies &\Pr\left[S_t \geq 16M^2\eta\log(1/\delta) + 4\sqrt{\log(1/\delta)}\left(\frac{M\|x^{(1)} - x^*\|}{\sqrt{T}} + \sqrt{2}M\eta\sqrt{G^2+M^2}\right)\right] \leq \delta.
\end{aligned} \tag{5}$$

Noticing $S_t = \frac{1}{T}\sum_{i=1}^{t}\langle\xi^{(i)}, x^* - x^{(i)}\rangle$ and the above proof template holds for all $t \in [T]$, the proof is finished.

□

Theorem 5.4. *Suppose f is convex and G-Lipschitz for some $G > 0$ and Assumption 3.1 holds. Denote x_{out} as the output of Algorithm 1 with learning rate $\eta = \min\{1/(M\sqrt{T}), 1/\sqrt{T}\}$. Then*

$$f(x_{\text{out}}) - f^* \;=\; \mathcal{O}\left(\frac{\log(1/\delta)}{\sqrt{T}}\right)$$

with probability at least $1-\delta$ for any $\delta \in (0,1)$.

Proof. We start with eq. (2). By setting $\eta = \min\{1/(M\sqrt{T}), 1/\sqrt{T}\}$,

$$X \;=\; \frac{\|x^{(1)} - x^*\|^2(M+1)}{2\sqrt{T}} + \frac{G^2+M^2}{\sqrt{T}}. \tag{6}$$

Then we apply Proposition 5.3 to bound the term Z. To sum up,

$$f(x_{\text{out}}) - f^* \;\leq\; X + \frac{16M\log(1/\delta)}{\sqrt{T}} + 4\left(\frac{M\|x^{(1)} - x^*\| + \sqrt{2(G^2+M^2)}}{\sqrt{T}}\right)\sqrt{\log(1/\delta)}$$

with probability at least $1-\delta$. Substitute X with eq. (6) finishes the proof. □

Theorem 5.5. *Suppose f is convex and G-Lipschitz for some $G > 0$ and Assumption 3.1 holds. Denote $\{x^{(t)}\}_{t=1}^{T}$ as the iterates generated from Algorithm 1 with learning rate $\eta \leq \min\{1/(M\sqrt{T}), 1/\sqrt{T}\}$. Then*

$$\max_{t\in[T]}\|x^{(t)} - x^*\| \;\leq\; D_X + \sqrt{2(G^2+M^2)} + 4\sqrt{2\log(T/\delta)}$$

with probability at least $1-\delta$ for any $\delta \in (0,1)$, where $D_X \coloneqq \|x^{(1)} - x^*\|$.

Proof. Let $S_t = \frac{1}{T}\sum_{i=1}^{t}\langle\xi^{(i)}, x^* - x^{(i)}\rangle \;\forall t \in [T]$. Then by Lemma 5.1, we know that for all $t \in [T]$,

$$\begin{aligned}
\|x^{(t)} - x^*\|^2 \;&\leq\; \|x^{(1)} - x^*\|^2 + 2\eta T S_{t-1} + 2\eta^2(t-1)(G^2+M^2) \\
&\overset{\text{(i)}}{\leq}\; \|x^{(1)} - x^*\|^2 + 2\eta T S_{t-1} + 2(G^2+M^2),
\end{aligned} \tag{7}$$

where (i) is by the definition of η. By Proposition 5.3, we further know that

$$\begin{aligned}
2\eta T S_{t-1} \;&\leq\; 32M^2\eta^2 T\log(1/\delta) + 8\sqrt{\log(1/\delta)}\left(\frac{M\|x^{(1)} - x^*\|\eta T}{\sqrt{T}} + \sqrt{2}M\eta^2 T\sqrt{G^2+M^2}\right) \\
&\overset{\text{(i)}}{\leq}\; 32\log(1/\delta) + 8\left(\|x^{(1)} - x^*\| + \sqrt{2}\sqrt{G^2+M^2}\right)\sqrt{\log(1/\delta)},
\end{aligned}$$

with probability at least $1-\delta$, where (i) is by the definition of η. Plug the above inequality into eq. (7), we obtain that

$$\|x^{(t)} - x^*\|^2 \;\leq\; D_X^2 + 2(G^2+M^2) + 32\log(1/\delta) + 8\left(D_X + \sqrt{2(G^2+M^2)}\right)\sqrt{\log(1/\delta)}, \tag{8}$$

with probability at least $1-\delta$. Noticing that eq. (8) holds for all $t \in [T]$. By union bound, we can further obtain

$$\max_{t\in[T]} \|x^{(t)} - x^*\|^2 \;\le\; D_X^2 + 2(G^2+M^2) + 32\log(1/\delta) + 8\left(D_X + \sqrt{2(G^2+M^2)}\right)\sqrt{\log(1/\delta)}$$

$$\implies \max_{t\in[T]} \|x^{(t)} - x^*\| \;\le\; D_X + \sqrt{2(G^2+M^2)} + 4\sqrt{2\log(1/\delta)}$$

with probability at least $1-T\delta$. Substitute δ with δ/T finishes the proof. □

C Proofs for Section 6

Lemma 6.1. *Suppose f is convex and L-smooth for some $L>0$ and Assumption 3.1 holds. Denote $\{x^{(t)}\}_{t=1}^T$ as the iterates from Algorithm 1 with the learning rate $\eta \in (0, 1/(2L)]$. Then for any $t \in [T]$,*

$$\|x^{(t)} - x^*\|^2 \;\le\; \|x^{(1)} - x^*\|^2 + 2\eta\sum_{i=1}^{t-1}\langle \xi^{(i)}, x^* - x^{(i)}\rangle + 2(t-1)\eta^2(2\|\nabla f(x^*)\|^2 + M^2).$$

Proof. We start with eq. (1) (note that we need to substitute G with $\|\nabla f(x^{(t)})\|$ since we are not assuming Lipschitz continuous now).

$$\begin{aligned}
&\|x^{(t+1)} - x^*\|^2 \\
\le\;& \|x^{(t)} - x^*\|^2 - 2\eta(f(x^{(t)}) - f^*) - 2\eta\langle \xi^{(t)}, x^{(t)} - x^*\rangle + 2\eta^2\|\nabla f(x^{(t)})\|^2 + 2\eta^2 M^2 \\
\overset{\text{(i)}}{\le}\;& \|x^{(t)} - x^*\|^2 - 2(\eta - 2L\eta^2)(f(x^{(t)}) - f^*) - 2\eta\langle \xi^{(t)}, x^{(t)} - x^*\rangle + 4\eta^2\|\nabla f(x^*)\|^2 + 2\eta^2 M^2 \\
\overset{\text{(ii)}}{\le}\;& \|x^{(t)} - x^*\|^2 - 2\eta\langle \xi^{(t)}, x^{(t)} - x^*\rangle + 2\eta^2(2\|\nabla f(x^*)\|^2 + M^2),
\end{aligned}$$

where (i) is by Lemma A.3 and (ii) is true since $\eta - 2L\eta \ge 0$ by assuming $\eta \in (0, 1/(2L)]$. Applying the above inequality recursively yields the desired result. □

Proposition C.1. *Suppose that f is convex and L-smooth for some $L>0$ and Assumption 3.1 holds. Denote $\{x^{(t)}\}_{t=1}^T$ as the iterates generated from Algorithm 1 with learning rate $\eta \in (0, 1/(2L)]$. For any $t \in [T]$, let $\mathcal{F}_{t-1}$ be the σ-algebra generated from $\{x^{(1)}, \dots, x^{(t)}\}$, $d_t := \frac{1}{T}\langle \xi^{(t)}, x^* - x^{(t)}\rangle$, $v_{t-1} := \frac{M^2}{T^2}\|x^{(t)} - x^*\|^2$ and define $V_t = \sum_{i=1}^t v_i$. Then*

$$\mathbb{E}[\exp(\lambda d_t) \mid \mathcal{F}_{t-1}] \;\le\; \exp\left(\frac{\lambda^2}{2} v_{t-1}\right) \qquad \forall \lambda \in \mathbb{R} \tag{9}$$

and

$$V_t \;\le\; \frac{M^2\|x^{(1)} - x^*\|^2}{T} + \frac{2M^2\eta}{T}\sum_{i=1}^{t-1}(t-i)d_i + 2\eta^2 M^2(2\|\nabla f(x^*)\|^2 + M^2)$$

for any $t \in [T]$.

Proof. The proof for eq. (9) is exactly the same as in Proposition 5.2 and omit the details.

$$\begin{aligned}
V_t \;&=\; \sum_{i=1}^t v_{i-1} \\
&=\; \frac{M^2}{T^2}\sum_{i=1}^t \|x^{(i)} - x^*\|^2 \\
&\overset{\text{(i)}}{\le}\; \frac{M^2}{T^2}\sum_{i=1}^t \left(\|x^{(1)} - x^*\|^2 + 2\eta\sum_{j=1}^{i-1}\langle \xi^{(j)}, x^* - x^{(j)}\rangle + 2\eta^2(i-1)(2\|\nabla f(x^*)\|^2 + M^2)\right) \\
&\le\; \frac{M^2\|x^{(1)} - x^*\|^2}{T} + \frac{2M^2\eta}{T}\sum_{i=1}^t\sum_{j=1}^{i-1} d_j + 2\eta^2 M^2(2\|\nabla f(x^*)\|^2 + M^2)
\end{aligned}$$

$$\leq \frac{M^2\|x^{(1)}-x^*\|^2}{T} + \frac{2M^2\eta}{T}\sum_{i=1}^{t-1}(t-i)d_i + 2\eta^2M^2(2\|\nabla f(x^*)\|^2+M^2),$$

□

Proposition C.2. *Suppose f is convex and L-smooth for some $L>0$ and Assumption 3.1 holds. Denote $\{x^{(t)}\}_{t=1}^T$ as the iterates generated from Algorithm 1 with learning rate $\eta \in (0, 1/(2L)]$. Then for any $t \in [T]$ and $\delta \in (0,1)$,*

$$\frac{1}{T}\sum_{i=1}^{t}\langle \xi^{(i)}, x^* - x^{(i)}\rangle \leq c_3\sigma^2\eta\log(2/\delta) + 4c_4\sqrt{\log(2/\delta)}\left(\frac{\sigma\|x^{(1)}-x^*\|}{\sqrt{T}} + 2\sigma\eta\|\nabla f(x^*)\|\right)$$

with probability at least $1-\delta$, where c_3, c_4 are some absolute positive constants.

Proof. Let

$$d_t \coloneqq \frac{1}{T}\langle \xi^{(t)}, x^* - x^{(t)}\rangle, \quad S_t \coloneqq \sum_{i=1}^{t} d_i, \quad v_{t-1} \coloneqq \frac{M^2\|x^{(t)}-x^*\|^2}{T^2}, \quad V_t \coloneqq \sum_{i=1}^{t} v_i$$

for any $t \in [T]$.

Given $t \in [T]$, apply Proposition C.1 and Lemma 3.2 and making the identification

$$\alpha_i = \frac{2M^2\eta(t-i)}{T} \quad \forall i \in [T], \qquad R \coloneqq \frac{M^2\|x^{(1)}-x^*\|^2}{T} + 2\eta^2M^2(2\|\nabla f(x^*)\|^2+M^2).$$

We obtain

$$\Pr[S_t \geq x] \leq \exp\left(-\frac{x^2}{4\max_{i\in[t]}\alpha_i x + 8R}\right) \qquad \forall x > 0.$$

Let $\alpha = \max_{i\in[t]}\alpha_i$ and set $x = \max\{8\alpha\log(1/\delta), 4\sqrt{R\log(1/\delta)}\}$ gives

$$\Pr\left[S_t \geq \max\{8\alpha\log(1/\delta), 4\sqrt{R\log(1/\delta)}\}\right] \leq \delta \qquad \forall \delta \in (0,1).$$

Plug in α and R, we obtain

$$\begin{aligned}
&\Pr\left[S_t \geq \max\{8\alpha\log(1/\delta), 4\sqrt{R\log(1/\delta)}\}\right] \leq \delta \\
\implies &\Pr\left[S_t \geq 8\alpha\log(1/\delta) + 4\sqrt{R\log(1/\delta)}\right] \leq \delta \\
\implies &\Pr\left[S_t \geq 16M^2\eta\log(1/\delta) + 4\sqrt{\log(1/\delta)}\left(\frac{M\|x^{(1)}-x^*\|}{\sqrt{T}} + M\eta\sqrt{4\|\nabla f(x^*)\|^2+2M^2}\right)\right] \leq \delta.
\end{aligned} \tag{10}$$

Noticing that $S_t = \frac{1}{T}\sum_{i=1}^{t}\langle \xi^{(i)}, x^* - x^{(i)}\rangle$ and the above proof template holds for all $t \in [T]$. The proof is finished. □

Theorem 6.2. *Suppose f is convex and L-smooth for some $L>0$ and Assumption 3.1 holds. Denote x_{out} as the output of Algorithm 1 with learning rate $\eta = \min\{1/(2L), 1/(M\sqrt{T}), 1/\sqrt{T}\}$. Then*

$$f(x_{\text{out}}) - f^* = \mathcal{O}\left(\frac{\log(1/\delta)}{\sqrt{T}}\right)$$

with probability at least $1-\delta$ for any $\delta \in (0,1)$.

Proof. We start with eq. (2) (substitute G^2 with $2\|f(x^*)\|^2$). By setting $\eta = \min\{1/(2L), 1/(M\sqrt{T}), 1/\sqrt{T}\}$,

$$X = \frac{L\|x^{(1)}-x^*\|^2}{T} + \frac{\|x^{(1)}-x^*\|^2(M+1)}{2\sqrt{T}} + \frac{2\|\nabla f(x^*)\|^2+M^2}{\sqrt{T}}.$$

Then we apply Proposition C.2 to bound the term Z. To sum up,

$$f(x_{\text{out}}) - f^* \leq \frac{L\|x^{(1)} - x^*\|^2}{T} + \frac{\|x^{(1)} - x^*\|^2(M+1)}{2\sqrt{T}} + \frac{2\|\nabla f(x^*)\|^2 + M^2}{\sqrt{T}} + \frac{16M\log(1/\delta)}{\sqrt{T}} + 4\left(\frac{M\|x^{(1)} - x^*\| + M\sqrt{4\|\nabla f(x^*)\|^2 + 2M^2}}{\sqrt{T}}\right)\sqrt{\log(1/\delta)}$$

with probability at least $1-\delta$, which finishes the proof. □

Theorem 6.3. *Suppose f is convex and L-smooth for some $L > 0$ and Assumption 3.1 holds. Denote $\{x^{(t)}\}_{t=1}^T$ as the iterates generated from Algorithm 1 with learning rate $\eta \leq \min\{1/(2L), 1/(M\sqrt{T}), 1/\sqrt{T}\}$. Then*

$$\max_{t\in[T]} \|x^{(t)} - x^*\| \leq D_X + \sqrt{4\|\nabla f(x^*)\|^2 + 2M^2} + 4\sqrt{2\log(T/\delta)}$$

with probability at least $1-\delta$ for any $\delta \in (0,1)$.

Proof. The proof is basically the same as the proof of Theorem 5.5. We omit the details. □

D Proofs for Section 7

First, we develop some technical lemmas and propositions. Lemma D.1, Proposition D.2 and Proposition D.3 are analogies of Lemma 5.1, Proposition 5.2 and Proposition 5.3 to the mSGD algorithm.

Lemma D.1. *Suppose that f is convex and G-Lipschitz for some $G > 0$ and Assumption 3.1 holds. Denote $\{x^{(t)}\}_{t=1}^T$ and $\{z^{(t)}\}_{t=1}^T$ as the iterates generated from Algorithm 2 with learning rate $\eta > 0$. Then for any $t \in [T]$,*

$$\|z^{(t)} - x^*\|^2 \leq \|x^{(1)} - x^*\|^2 + 2(t-1)\eta^2(G^2 + M^2) + 2\eta\sum_{i=1}^{t-1}\langle \xi^{(t)}, x^* - z^{(t)}\rangle.$$

Proof. Given any $t \in [T]$,

$$\begin{aligned}
&\|z^{(t+1)} - x^*\|^2 \\
\leq\ &\|z^{(t)} - x^* - \eta\nabla f(x^{(t)}) - \eta\xi^{(t)}\|^2 \\
=\ &\|z^{(t)} - x^*\|^2 - 2\eta\langle \nabla f(x^{(t)}), z^{(t)} - x^*\rangle - 2\eta\langle \xi^{(t)}, z^{(t)} - x^*\rangle + 2\eta^2\left(G^2 + M^2\right) \\
=\ &\|z^{(t)} - x^*\|^2 - 2\eta\langle \nabla f(x^{(t)}), x^{(t)} - x^*\rangle - 2\eta\lambda_{t-1}\langle \nabla f(x^{(t)}), x^{(t)} - x^{(t-1)}\rangle \\
&\quad - 2\eta\langle \xi^{(t)}, z^{(t)} - x^*\rangle + 2\eta^2\left(G^2 + M^2\right) \\
\leq\ &\|z^{(t)} - x^*\|^2 - 2\eta(f(x^{(t)}) - f^*) - 2\eta\lambda_{t-1}(f(x^{(t)}) - f(x^{(t-1)})) \\
&\quad - 2\eta\langle \xi^{(t)}, z^{(t)} - x^*\rangle + 2\eta^2\left(G^2 + M^2\right) \\
=\ &\|z^{(t)} - x^*\|^2 - 2\eta(1+\lambda_{t-1})(f(x^{(t)}) - f^*) + 2\eta\lambda_{t-1}(f(x^{(t-1)}) - f^*) \\
&\quad - 2\eta\langle \xi^{(t)}, z^{(t)} - x^*\rangle + 2\eta^2\left(G^2 + M^2\right).
\end{aligned}$$

By the definition of λ_t, we know that $\lambda_t = 1 + \lambda_{t-1}$. Rearranging the above inequality gives

$$\begin{aligned}
&\|z^{(t+1)} - x^*\|^2 + 2\eta\lambda_t(f(x^{(t)}) - f^*) \\
\leq\ &\|z^{(t)} - x^*\|^2 + 2\eta\lambda_{t-1}(f(x^{(t-1)}) - f^*) + 2\eta^2\left(G^2 + M^2\right) + 2\eta\langle \xi^{(t)}, x^* - z^{(t)}\rangle.
\end{aligned} \tag{11}$$

Summing over the above inequality over $\{1, 2, \ldots, t-1\}$ and noticing $f(x^{(t-1)}) - f^* \geq 0$, we obtain the desired result. □

Proposition D.2. *Suppose that f is convex and G-Lipschitz for some $G > 0$ and Assumption 3.1 holds. Denote $\{x^{(t)}, z^{(t)}\}_{t=1}^T$ as the iterates from Algorithm 1 with learning rate $\eta > 0$. For any $t \in [T]$, let $\mathcal{F}_{t-1}$ be the σ-algebra generated from $\{\xi^{(1)}, \dots, \xi^{(t-1)}\}$, $d_t := \frac{1}{T}\langle \xi^{(t)}, x^* - z^{(t)} \rangle$, $v_{t-1} := \frac{M^2}{T^2}\|z^{(t)} - x^*\|^2$ and define $V_t = \sum_{i=1}^t v_{i-1}$. Then for any $t \in [T]$,*

$$\mathbb{E}[\exp(\lambda d_t) \mid \mathcal{F}_{t-1}] \;\le\; \exp\left(\frac{\lambda^2}{2} v_{t-1}\right) \qquad \forall \lambda \in \mathbb{R}$$

and

$$V_t \;\le\; \frac{M^2\|x^{(1)} - x^*\|^2}{T} + \frac{2M^2\eta}{T}\sum_{i=1}^{t-1}(t-i)d_i + 2M^2(G^2 + M^2)\eta^2.$$

Proof. The proof follows exactly the same as Proposition 5.2, the only difference is that we apply Lemma D.1 instead of Lemma 5.1. □

Proposition D.3. *Suppose f is convex and G-Lipschitz for some $G > 0$ and Assumption 3.1 holds. Denote $\{x^{(t)}, z^{(t)}\}_{t=1}^T$ as the iterates generated from Algorithm 1 with learning rate $\eta > 0$ and let $D_X := \|x^{(1)} - x^*\|$. Then for any $t \in [T]$ and $\delta \in (0, 1)$,*

$$\frac{1}{T}\sum_{i=1}^t \langle \xi^{(i)}, x^* - z^{(i)} \rangle \;\le\; 16M^2\eta\log(1/\delta) + 4\left(\frac{MD_X}{\sqrt{T}} + \sqrt{2}M\eta\sqrt{G^2 + M^2}\right)\sqrt{\log(1/\delta)}$$

with probability at least $1 - \delta$.

Proof. The proof follows exactly the same as Proposition 5.3. □

Theorem 7.3. *Suppose f is convex and G-Lipschitz for some $G > 0$ and Assumption 3.1 holds. Denote $\{x^{(t)}\}_{t=1}^T$ as the iterates generated from Algorithm 2 with learning rate $\eta = 1/\sqrt{T}$ and $\lambda_t = t$. Then*

$$f(x^{(T)}) - f^* \;=\; \mathcal{O}\left(\frac{\log(1/\delta)}{\sqrt{T}}\right)$$

with probability at least $1 - \delta$ for any $\delta \in (0, 1)$.

Proof. We start with eq. (11). Summing eq. (11) over $\{1, 2, \dots, T\}$ and rearranging, we obtain that

$$\begin{aligned} f(x^{(T)}) - f^* \;&\le\; \frac{\|z^{(1)} - x^*\|^2}{2\eta\lambda_T} + \frac{2T\eta}{\lambda_T}(G^2 + M^2) + \frac{1}{\lambda_T}\langle \xi^{(t)}, x^* - z^{(t)} \rangle \\ &\le\; \frac{\|z^{(1)} - x^*\|^2}{2\eta T} + 2\eta(G^2 + M^2) + \frac{1}{T}\langle \xi^{(t)}, x^* - z^{(t)} \rangle, \end{aligned}$$

where the second line is by the definition of $\lambda_t = t$. By setting $\eta = 1/\sqrt{T}$ and combining Proposition D.3, we obtain the desried result. □

Theorem 7.4. *Suppose f is convex and G-Lipschitz for some $G > 0$ and Assumption 3.1 holds. Denote $\{x^{(t)}\}_{t=1}^T$ and $\{z^{(t)}\}_{t=1}^T$ as the iterates generated from Algorithm 2 with learning rate $\eta \le \min\{1/(M\sqrt{T}), 1/\sqrt{T}\}$ and $\lambda_t = t$. Then*

$$\max_{t \in [T]} \|z^{(t)} - x^*\| \;\le\; D_X + \sqrt{2(G^2 + M^2)} + 4\sqrt{2\log(T/\delta)}$$

with probability at least $1 - \delta$ for any $\delta \in (0, 1)$.

Proof. The proof follows exactly the same as Theorem 5.5. We omit the details. □

E OTHER DISCUSSIONS

For nonconvex problems, following the standard proof template, we start the derivation with the descent lemma.

$$\begin{aligned} f(x^{(t+1)}) &\leq f(x^{(t)}) + \langle \nabla f(x^{(t)}), x^{(t+1)} - x^{(t)} \rangle + \frac{L}{2}\|x^{(t+1)} - x^{(t)}\|^2 \\ &\leq f(x^{(t)}) - \eta \|\nabla f(x^{(t)})\|^2 - \eta \langle \nabla f(x^{(t)}), \xi^{(t)} \rangle + L\eta^2 \|\nabla f(x^{(t)})\|^2 + L\eta^2 \|\xi^{(t)}\|^2. \end{aligned}$$

Rearranging gives

$$(\eta - L\eta^2)\|\nabla f(x^{(t)})\|^2 \leq f(x^{(t)}) - f(x^{(t+1)}) - \eta \langle \nabla f(x^{(t)}), \xi^{(t)} \rangle + L\eta^2 M^2.$$

Different from the convex case, the term $\langle \xi^{(t)}, x^{(t)} - x^* \rangle$ is replaced by $\langle \nabla f(x^{(t)}), \xi^{(t)} \rangle$. The latter is easier the analyze since it does not require us to bound $\|x^{(t)} - x^*\|$. When f is Lipschitz continuous, one can simply bound $\sum_{t=1}^{T} \langle \nabla f(x^{(t)}), \xi^{(t)} \rangle$ by the Azuma's inequality. Then the standard analysis (Ghadimi & Lan, 2013) will gives a $\mathcal{O}(\mathrm{polylog}(1/\delta)/\sqrt{T})$ convergence rate to stationary point.