On the Last-Iterate Convergence of Clipped Gradient Methods

Zijian Liu
New York University
zl3067@stern.nyu.edu

Abstract

In stochastic (strongly) convex optimization, establishing last-iterate convergence guarantees for first-order methods has recently garnered increasing attention, as this not only deepens the theoretical understanding of these algorithms but also aligns more closely with practice. Notably, people have demonstrated that one of the most famous, simple, and popular methods, Stochastic Gradient Descent (SGD), provably achieves last-iterate convergence. However, the existing results may have limited applicability, since they typically rely on strong conditions on the gradient noise (e.g., an exponentially decaying tail) rather than the more realistic heavy-tailed noise. Unfortunately, when faced with heavy-tailed noise, SGD is known to exhibit undesirable behavior or even fail to converge. To address the challenge of heavy-tailed noise, people have proposed Clipped-SGD, an algorithm that combines SGD with a simple mechanism, gradient clipping. Although Clipped-SGD performs well in practice, its last-iterate convergence property remains highly underexplored. In this work, we prove the first optimal last-iterate rates in high probability for Clipped-SGD under heavy-tailed noise, thereby closing a gap in the literature.

1 Introduction[1]

In stochastic (strongly) convex optimization, establishing last-iterate convergence guarantees[2] for first-order methods has attracted much attention in recent years. This topic not only deepens our theoretical understanding of different algorithms but also aligns more closely with reality, given that the last iterate is often used as the final output in practice. Notably, a particularly exciting line of work (see, e.g., [3, 14, 16, 20, 26, 28, 33, 36, 39]) has demonstrated that one of the most famous, simple, and popular methods, Stochastic Gradient Descent (SGD) [37], which serves as the prototype for many follow-up first-order algorithms, provably achieves last-iterate convergence rates, even optimal ones, for both Lipschitz and smooth objectives.

However, upon closer examination, one may realize that the existing last-iterate convergence results for SGD may have limited applicability, since they typically rely on strong conditions on the gradient noise (which refers to the difference between the stochastic and the true gradients). Especially, two commonly considered conditions, the finite-variance noise assumption for convergence in expectation and the exponentially decaying tail noise assumption (e.g., sub-Gaussian noise) for convergence in high probability, do not accurately reflect practice, as they are observed to be violated in many different scenarios [1, 4, 13, 40, 43]. Given this, a more realistic assumption, known as heavy-tailed noise, has been proposed to model the behavior of gradient noise. Specifically, under the assumption

[1] To save space, we only provide a simplified leading term in the rate when discussing convergence results in this section.

[2] Throughout the paper, given an objective function F and a point $\boldsymbol{x}$ in the optimization domain, we measure the convergence by the function value gap, i.e., $F(\boldsymbol{x}) - F_\star$, where $F_\star = F(\boldsymbol{x}_\star)$ denotes the function value at the optimal solution $\boldsymbol{x}_\star$. Note that for μ-strongly convex F (not necessarily smooth), the function value gap always performs as a stronger convergence criterion than the commonly used metric $\mu \left\|\boldsymbol{x} - \boldsymbol{x}_\star\right\|^2$, as discussed in the literature (see, e.g., [26]).

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

of heavy-tailed noise, the gradient noise is considered to have only a finite $\mathtt{p}$-th moment, where $\mathtt{p} \in (1, 2]$ is called the tail index.

Unfortunately, when faced with heavy-tailed noise, SGD is known to exhibit undesirable behavior, such as an inevitable $\mathrm{poly}(1/\delta)$ dependence on the failure probability $\delta \in (0, 1]$ in high-probability convergence [24, 38], or even to fail to converge in expectation [43]. To address this challenge, people have proposed Clipped-SGD, an algorithm that combines SGD with a simple mechanism known as gradient clipping. Concretely, in contrast to SGD, which directly applies the update rule to the stochastic gradient $\boldsymbol{g}$, Clipped-SGD first preprocesses $\boldsymbol{g}$ via the clipping function to obtain its clipped counterpart, $\mathtt{clip}_\tau(\boldsymbol{g}) \triangleq \min\{\tau, \|\boldsymbol{g}\|\} \frac{\boldsymbol{g}}{\|\boldsymbol{g}\|}$, where τ is called the clipping threshold, and then uses the clipped stochastic gradient in the following procedure.

Although Clipped-SGD performs well in practice and has provable convergence guarantees in various settings, its last-iterate convergence property remains highly underexplored. As far as we are aware, only two studies have established last-iterate rates under heavy-tailed noise. The first is [34], which showed a last-iterate rate of $\mathcal{O}(1/T^{1-\frac{1}{\mathtt{p}}})$ in expectation for Lipschitz convex objectives; this rate is known to be optimal in T [23, 30, 41]. The second is [7], in which the authors developed a last-iterate rate of $\mathcal{O}(\ln^2(T^2/\delta)/T^{\frac{2}{3}(1-\frac{1}{\mathtt{p}})})$ in high probability for smooth convex objectives. While, to the best of our knowledge, no high-probability lower bounds in the corresponding setting are known in the literature, the rate in [7] appears to be suboptimal for the following two reasons. First, it is worse than the best-known high-probability rate $\mathcal{O}(\ln^{1-\frac{1}{\mathtt{p}}}(1/\delta)/T^{1-\frac{1}{\mathtt{p}}})$ for the average iterate under the same setting [32]. Second, when $\mathtt{p} = 2$, the rate in [7] reduces to only $\mathcal{O}(\ln^2(T^2/\delta)/T^{\frac{1}{3}})$, which is slower than the best-known in-expectation rate $\mathcal{O}(1/\sqrt{T})$ for the last iterate of SGD [26].

Given the above discussion, an important research question remains to be investigated:

Under heavy-tailed noise, can we establish high-probability rates for the last iterate of Clipped-SGD *that match its best-known high-probability bounds for the average iterate?*

1.1 Our Contributions

This work answers the above question affirmatively by proving the following results, thereby closing a gap in the literature.

- For general convex optimization, we prove that the last iterate of Clipped-SGD converges in high probability at a rate of $\mathcal{O}(\ln^{1-\frac{1}{\mathtt{p}}}(1/\delta)/T^{1-\frac{1}{\mathtt{p}}})$ for both Lipschitz and smooth objectives (cf. Theorems 1 and 2), matching the best-known average-iterate results in [23] for Lipschitz functions and [32] for smooth functions.
- For strongly convex optimization[3], we prove that the last iterate of Clipped-SGD converges in high probability at a rate of $\mathcal{O}(\ln^{2-\frac{2}{\mathtt{p}}}(1/\delta)/T^{2-\frac{2}{\mathtt{p}}})$ for both Lipschitz and smooth objectives (cf. Theorems 3 and 4), matching the best-known average-iterate result in [23] for Lipschitz functions and improving over the result in [38] for smooth functions by removing extra logarithmic factors.
- For Lipschitz objectives, our rates are optimal in all parameters (including other problem-dependent parameters currently hidden in the big-$\mathcal{O}$ notation) under the classical heavy-tailed noise assumption, as implied by the high-probability minimax lower bounds in [23]. For smooth objectives, although no high-probability minimax lower bounds are known, we believe that our rates are also optimal under the classical heavy-tailed noise assumption.
- Notably, our rates are derived under the fine-grained heavy-tailed noise condition (cf. Assumption 5) recently introduced in [23] and exhibit a refined dependence on the noise parameter, as in [23].
- We highlight that our analysis contains an interesting idea. Specifically, we develop a new argument that maintains the optimal $\mathrm{poly}(\ln(1/\delta))$ dependence on δ. Our analysis is inspired by induction-based proofs in the literature, which, in contrast, typically lead to a suboptimal $\mathrm{poly}(\ln(T/\delta))$ dependence. For more details, we refer the interested reader to Section 4.

[3] In fact, our results extend beyond strongly convex functions. See Assumption 2 for details.

1.2 Related Work

First, we review the literature on high-probability convergence results for `Clipped-SGD` in stochastic (strongly) convex optimization under heavy-tailed noise. For high-probability convergence in stochastic nonconvex optimization under heavy-tailed noise (not limited to the `Clipped-SGD` algorithm), the interested reader could refer to, for example, [2, 8, 11, 18, 22, 27, 32, 38], for recent progress.

Upper bounds. For convex problems, the best-known high-probability rate for the average iterate is $\mathcal{O}(\ln^{1-\frac{1}{\mathtt{p}}}(1/\delta)/T^{1-\frac{1}{\mathtt{p}}})$, for both Lipschitz functions [23] and smooth functions [32]. For strongly convex problems, the best-known high-probability bounds are $\mathcal{O}(\ln^{2-\frac{2}{\mathtt{p}}}(1/\delta)/T^{2-\frac{2}{\mathtt{p}}})$ for the weighted average iterate and the squared distance $\|\boldsymbol{x}_{T+1}-\boldsymbol{x}_\star\|^2$ in the Lipschitz setting [23] and $\widetilde{\mathcal{O}}(\ln^{2-\frac{2}{\mathtt{p}}}(T/\delta)/T^{2-\frac{2}{\mathtt{p}}})$ for $\|\boldsymbol{x}_{T+1}-\boldsymbol{x}_\star\|^2$ in the smooth setting [38].

Regarding last-iterate convergence, the only work we are aware of is [7], which proved the first and only rate in the order of $\mathcal{O}(\ln^2(T^2/\delta)/T^{\frac{2}{3}(1-\frac{1}{\mathtt{p}})})$ for smooth convex objectives. However, such a bound is worse than the best-known average-iterate rate $\mathcal{O}(\ln^{1-\frac{1}{\mathtt{p}}}(1/\delta)/T^{1-\frac{1}{\mathtt{p}}})$ mentioned above for the same setting. Note that in the strongly convex case, as discussed in Footnote 2, the existing results for $\|\boldsymbol{x}-\boldsymbol{x}_\star\|^2$ do not align with the metric $F(\boldsymbol{x})-F_\star$ considered in this work.

Second, we discuss lower bounds for stochastic (strongly) convex optimization under heavy-tailed noise (or similar conditions). Given the limited number of results in the literature, we include both high-probability and in-expectation lower bounds.

Lower bounds. To the best of our knowledge, only a few results of high-probability lower bounds are known, and all of them are established for Lipschitz functions. In detail, [35] proved the first high-probability minimax lower bound $\Omega(1/T^{1-\frac{1}{\mathtt{p}}})$ for convex optimization. However, it considered the heavy-tailed stochastic gradient assumption rather than the heavy-tailed noise assumption, and, more critically, did not provide a tight dependence on parameters other than T, such as the failure probability δ. Under the same heavy-tailed stochastic gradient assumption, but only for $\mathtt{p}=2$, [6] showed an improved minimax lower bound $\Omega(\sqrt{\ln(1/\delta)/T})$; for general $\mathtt{p}\in(1,2]$, [10] gave an algorithm-dependent lower bound $\Omega((1/\delta)^{\frac{1}{\mathtt{p}}}/T^{1-\frac{1}{\mathtt{p}}})$. Under the heavy-tailed noise assumption, [23] recently characterized the first high-probability minimax lower bounds, namely, $\Omega(\ln^{1-\frac{1}{\mathtt{p}}}(1/\delta)/T^{1-\frac{1}{\mathtt{p}}})$ for convex optimization and $\Omega(\ln^{2-\frac{2}{\mathtt{p}}}(1/\delta)/T^{2-\frac{2}{\mathtt{p}}})$ for strongly convex optimization.

Next, we turn our attention to in-expectation lower bounds. Under the heavy-tailed stochastic gradient assumption, [30] and [41] provided a minimax lower bound $\Omega(1/T^{1-\frac{1}{\mathtt{p}}})$ for Lipschitz convex functions, and [43] established the first $\Omega(1/T^{2-\frac{2}{\mathtt{p}}})$ result for smooth strongly convex objectives. Under the heavy-tailed noise assumption, [23] recently showed similar lower bounds $\Omega(1/T^{1-\frac{1}{\mathtt{p}}})$ and $\Omega(1/T^{2-\frac{2}{\mathtt{p}}})$, for Lipschitz convex and strongly convex optimization, respectively.

Finally, we remark that all lower bounds in [23] can apply to the fine-grained heavy-tailed noise assumption beyond the classical one. We will elaborate further on this point later when discussing our theorems in Section 3.

2 Preliminary

Notation. $\mathbb{N}$ is the set of natural numbers (excluding 0). $[T]\triangleq\{1,\dots,T\}$, $\forall T\in\mathbb{N}$. $\lceil\cdot\rceil$ and $\lfloor\cdot\rfloor$ denote the ceiling and floor functions, respectively. $\langle\cdot,\cdot\rangle$ is the standard Euclidean inner product, and $\|\cdot\|\triangleq\sqrt{\langle\cdot,\cdot\rangle}$ is the Euclidean norm. $\|\cdot\|_{\mathsf{op}}$ denotes the operator norm. $\mathbb{S}^{d-1}$ is the unit sphere in $\mathbb{R}^d$. $\mathbb{B}^d(\boldsymbol{x},r)$ is the Euclidean ball centered at $\boldsymbol{x}\in\mathbb{R}^d$ with radius $r>0$. Given a convex $h:\mathbb{R}^d\to\mathbb{R}$, $\nabla h(\boldsymbol{x})$ denotes an arbitrary element in $\partial h(\boldsymbol{x})$, where $\partial h(\boldsymbol{x})$ is the subgradient set of h at $\boldsymbol{x}$.

Objective. We study the composite optimization problem in the form of

$$\inf_{\boldsymbol{x}\in\mathcal{X}} F(\boldsymbol{x})\triangleq f(\boldsymbol{x})+r(\boldsymbol{x}),$$

where $\mathcal{X}\subseteq\mathbb{R}^d$ is a nonempty closed convex set.

Assumptions. Our analysis relies on the following assumptions.

Assumption 1. *There exists* $\boldsymbol{x}_\star \in \mathcal{X}$ *such that* $F_\star \triangleq F(\boldsymbol{x}_\star) = \inf_{\boldsymbol{x}\in\mathcal{X}} F(\boldsymbol{x})$.

Assumption 2. *Both* $f : \mathbb{R}^d \to \mathbb{R}$ *and* $r : \mathbb{R}^d \to \mathbb{R}$ *are convex. Moreover,* F *is* μ*-strongly star-convex on* $\mathcal{X}$ *for some* $\mu \geq 0$*, i.e.,* $F(\lambda\boldsymbol{x}_\star + (1-\lambda)\boldsymbol{x}) + \frac{\mu\lambda(1-\lambda)}{2}\|\boldsymbol{x}_\star - \boldsymbol{x}\|^2 \leq \lambda F_\star + (1-\lambda)F(\boldsymbol{x}), \forall \boldsymbol{x} \in \mathcal{X}, \lambda \in [0,1]$*, where* $\boldsymbol{x}_\star$ *is the minimizer specified in Assumption 1.*

Assumption 3. f *is* G*-Lipschitz on* $\mathcal{X}$*, i.e.,* $\|\nabla f(\boldsymbol{x})\| \leq G, \forall \boldsymbol{x} \in \mathcal{X}$.

Assumption 4. f *is* L*-smooth on* $\mathcal{X}$*, i.e.,* $\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\| \leq L\|\boldsymbol{x} - \boldsymbol{y}\|, \forall \boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}$.

Assumptions 1, 3, and 4 are standard in the optimization literature [5, 21, 31]. The convexity requirement in Assumption 2 is also fundamental. We remark that the μ-strong star-convexity condition [29] in Assumption 2 is strictly weaker than the commonly assumed μ-strong convexity condition in the literature.

Before moving on, we recall a useful property of μ-strongly star-convex functions (see, e.g., [17]) in the following Lemma 1. For completeness, we reproduce its proof in Appendix A.

Lemma 1. *Under Assumptions 1 and 2, we have* $\frac{\mu}{2}\|\boldsymbol{x} - \boldsymbol{x}_\star\|^2 \leq F(\boldsymbol{x}) - F_\star \leq \frac{\|\nabla F(\boldsymbol{x})\|^2}{2\mu}, \forall \boldsymbol{x} \in \mathcal{X}$.

Next, we consider a fine-grained heavy-tailed noise assumption recently introduced in [23]. For the plausibility of Assumption 5, we refer the reader to the discussions in Section 2 of [23].

Assumption 5. *There exists a function* $\boldsymbol{g} : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}^d$ *and a probability distribution* $\mathcal{D}$ *on* $\mathcal{Z}$ *such that* $\mathbb{E}_{Z\sim\mathcal{D}}[\boldsymbol{g}(\boldsymbol{x}; Z)] = \nabla f(\boldsymbol{x}), \forall \boldsymbol{x} \in \mathcal{X}$*. In addition, for some* $\mathtt{p} \in (1, 2]$*, we have*

$$\mathbb{E}_{Z\sim\mathcal{D}}\left[|\langle \boldsymbol{e}, \boldsymbol{g}(\boldsymbol{x}; Z) - \nabla f(\boldsymbol{x})\rangle|^{\mathtt{p}}\right] \leq \sigma_{\mathtt{s}}^{\mathtt{p}}, \quad \mathbb{E}_{Z\sim\mathcal{D}}\left[\|\boldsymbol{g}(\boldsymbol{x}; Z) - \nabla f(\boldsymbol{x})\|^{\mathtt{p}}\right] \leq \sigma_{\mathtt{l}}^{\mathtt{p}}, \quad \forall \boldsymbol{e} \in \mathbb{S}^{d-1}, \boldsymbol{x} \in \mathcal{X},$$

where $\sigma_{\mathtt{s}}$ *and* $\sigma_{\mathtt{l}}$ *are two constants satisfying* $0 \leq \sigma_{\mathtt{s}} \leq \sigma_{\mathtt{l}} \leq \sqrt{\pi d/2}\sigma_{\mathtt{s}}$.

Remark 1. Hereafter, we omit $Z \sim \mathcal{D}$ from $\mathbb{E}_{Z\sim\mathcal{D}}$ when the context is clear.

Remark 2. Assumption 5 recovers the classical heavy-tailed noise assumption by setting both $\sigma_{\mathtt{s}}$ and $\sigma_{\mathtt{l}}$ equal to a common parameter σ (or by simply ignoring the existence of $\sigma_{\mathtt{s}}$ and viewing $\sigma_{\mathtt{l}}$ as the single parameter of gradient noise).

As in [23], we also define the generalized effective dimension $d_{\mathtt{eff}}$ as follows:

$$d_{\mathtt{eff}} \triangleq \sigma_{\mathtt{l}}^2/\sigma_{\mathtt{s}}^2 \in \{0\} \cup [1, \pi d/2], \tag{1}$$

where $d_{\mathtt{eff}} = 0$ if and only if $\sigma_{\mathtt{s}} = \sigma_{\mathtt{l}} = 0$, i.e., the deterministic case. Note that $d_{\mathtt{eff}}$ can attain order $\Theta(d)$ [23] and recovers the effective dimension used in [9] when $\mathtt{p} = 2$.

3 Last-Iterate Convergence of Clipped Stochastic Gradient Descent

Algorithm 1 Clipped Stochastic Gradient Descent (`Clipped-SGD`)

Input: initial point $\boldsymbol{x}_1 \in \mathcal{X}$, time horizon $T \in \mathbb{N}$
for $t = 1$ **to** T **do**
  Decide a clipping threshold $\tau_t > 0$
  $\boldsymbol{g}_t^{\mathrm{c}} = \mathtt{clip}_{\tau_t}(\boldsymbol{g}_t)$ where $\boldsymbol{g}_t = \boldsymbol{g}(\boldsymbol{x}_t; Z_t)$ and $Z_t \sim \mathcal{D}$ is independent from the history
  Decide a stepsize $\eta_t > 0$
  $\boldsymbol{x}_{t+1} = \mathrm{argmin}_{\boldsymbol{x}\in\mathcal{X}} r(\boldsymbol{x}) + \langle \boldsymbol{g}_t^{\mathrm{c}}, \boldsymbol{x}\rangle + \frac{\|\boldsymbol{x} - \boldsymbol{x}_t\|^2}{2\eta_t}$
end for
Output: $\boldsymbol{x}_{T+1}$

Remark 3. In the same manner as [23], we refer to Algorithm 1 simply as `Clipped-SGD`, rather than `Proximal-Clipped-SGD`.

Remark 4. We treat the time horizon T in Algorithm 1 as known to avoid any potential logarithmic dependence on T.

This work studies the Clipped Stochastic Gradient Descent (Clipped-SGD) method, as given in Algorithm 1, and will establish the first optimal last-iterate convergence rates in high probability.

To ease the notation in the upcoming theorems, we denote by $D \triangleq \|\boldsymbol{x}_\star - \boldsymbol{x}_1\|$ the distance between the optimal solution $\boldsymbol{x}_\star$ and the initial point $\boldsymbol{x}_1$. Given a failure probability $\delta \in (0,1]$, we adopt an important quantity $\tau_\star$ proposed in [23], defined as

$$\tau_\star \triangleq \left(\min\left\{\frac{\sigma_{\mathtt{s}}\sigma_{\mathtt{l}}^{\mathtt{p}-1}}{\ln\frac{3}{\delta}}, \frac{\sigma_{\mathtt{s}}^{\mathtt{p}}}{\sigma_{\mathtt{l}}^{\mathtt{p}-1}\mathbb{1}\left[\mathtt{p}<2\right]}\right\}\right)^{\frac{1}{\mathtt{p}}}. \tag{2}$$

By the definition of the generalized effective dimension $d_{\mathtt{eff}} \overset{(1)}{=} \sigma_{\mathtt{l}}^2/\sigma_{\mathtt{s}}^2$, $\tau_\star$ can also be written as

$$\tau_\star = \sigma_{\mathtt{l}}/\varphi_\star^{1/\mathtt{p}} \quad \text{where} \quad \varphi_\star \triangleq \max\left\{\sqrt{d_{\mathtt{eff}}}\ln\frac{3}{\delta}, d_{\mathtt{eff}}\mathbb{1}\left[\mathtt{p}<2\right]\right\}. \tag{3}$$

3.1 General Convex Optimization

This subsection focuses on general convex optimization, i.e., $\mu = 0$ in Assumption 2.

3.1.1 Lipschitz Objectives

We first consider the case where f is G-Lipschitz on $\mathcal{X}$.

Theorem 1. *Under Assumptions 1, 2 (with $\mu = 0$), 3, and 5, for any $T \in \mathbb{N}$ and $\delta \in (0,1]$, setting $\eta_t = \eta_\star(T-t+1), \tau_t = \max\left\{2G, \tau_\star T^{\frac{1}{\mathtt{p}}}\right\}, \forall t \in [T]$, where $\eta_\star$ is a properly picked parameter defined in (13), then* Clipped-SGD *(Algorithm 1) guarantees that, with probability at least $1-\delta$,*

$$F(\boldsymbol{x}_{T+1}) - F_\star = \mathcal{O}\left(\frac{(\varphi_{\mathrm{lip}} + \ln\frac{3}{\delta})GD}{T} + \frac{GD}{\sqrt{T}} + \frac{(\sigma_{\mathtt{s}}^{\frac{2}{\mathtt{p}}-1}\sigma_{\mathtt{l}}^{2-\frac{2}{\mathtt{p}}} + \sigma_{\mathtt{s}}^{\frac{1}{\mathtt{p}}}\sigma_{\mathtt{l}}^{1-\frac{1}{\mathtt{p}}}\ln^{1-\frac{1}{\mathtt{p}}}\frac{3}{\delta})D}{T^{1-\frac{1}{\mathtt{p}}}}\right),$$

in which $\varphi_{\mathrm{lip}} \le \varphi_\star$ is a constant defined in (11) and equals $\varphi_\star$ when $T = \Omega\left(\frac{G^{\mathtt{p}}}{\sigma_{\mathtt{l}}^{\mathtt{p}}}\varphi_\star\right)$.

To the best of our knowledge, for Lipschitz objectives, Theorem 1 provides the first and only high-probability convergence rate for the last iterate of Clipped-SGD. Even in the special case $\mathtt{p} = 2$, i.e., finite-variance noise, no such high-probability last-iterate convergence results exist in the literature.

We highlight that our Theorem 1 exactly matches the best-known average-iterate convergence rate of Clipped-SGD in high probability [23] and hence has three implications. First, under the classical heavy-tailed noise condition in the literature (i.e., $\sigma_{\mathtt{s}} = \sigma_{\mathtt{l}} = \sigma$ for some $\sigma \ge 0$), the rate in Theorem 1 reduces to $\mathcal{O}\left(\frac{GD\ln\frac{1}{\delta}}{T} + \frac{GD}{\sqrt{T}} + \frac{\sigma D\ln^{1-\frac{1}{\mathtt{p}}}\frac{1}{\delta}}{T^{1-\frac{1}{\mathtt{p}}}}\right)$ by noting that $d_{\mathtt{eff}} \overset{(1)}{=} \sigma_{\mathtt{l}}^2/\sigma_{\mathtt{s}}^2 = 1 \Rightarrow \varphi_{\mathrm{lip}} \le \varphi_\star \overset{(3)}{=} \ln\frac{3}{\delta}$. Remarkably, the leading term $\mathcal{O}\left(\frac{\sigma D\ln^{1-\frac{1}{\mathtt{p}}}\frac{1}{\delta}}{T^{1-\frac{1}{\mathtt{p}}}}\right)$ in the resulting bound is known to be fully optimal [23]. Second, under the fine-grained heavy-tailed noise assumption (i.e., $\sigma_{\mathtt{s}}$ and $\sigma_{\mathtt{l}}$ may differ), the rate established in Theorem 1 is optimal in T, yet may be suboptimal in terms of the problem-dependent parameters $\sigma_{\mathtt{s}}$ and $\sigma_{\mathtt{l}}$. This is because, under Assumption 5, the only existing lower bound is $\Omega\left(\frac{(\sigma_{\mathtt{s}}^{\frac{2}{\mathtt{p}}-1}\sigma_{\mathtt{l}}^{2-\frac{2}{\mathtt{p}}} + \sigma_{\mathtt{s}}\ln^{1-\frac{1}{\mathtt{p}}}\frac{1}{\delta})D}{T^{1-\frac{1}{\mathtt{p}}}}\right)$ [23], which suggests that there could still be room to further improve the dependence on $\sigma_{\mathtt{s}}$ and $\sigma_{\mathtt{l}}$. Third, when $\sigma_{\mathtt{l}}$, equivalently $\sigma_{\mathtt{s}}$, tends to 0, Theorem 1 recovers the standard $\mathcal{O}\left(\frac{GD}{\sqrt{T}}\right)$ rate for deterministic Lipschitz convex optimization.

Moreover, we remark that the linearly decaying stepsize employed in Theorem 1 is critical to our proof. This kind of stepsize was originally introduced in [42] to achieve an optimal convergence rate for Gradient Descent on Lipschitz objectives in deterministic convex optimization. Subsequently, [26] demonstrated that it can also be applied to stochastic (strongly) convex optimization and obtained optimal last-iterate convergence rates in various settings. Our work further investigates the potential of such a stepsize in Clipped-SGD and, consequently, establishes the first optimal last-iterate convergence rates in high probability under heavy-tailed noise.

3.1.2 Smooth Objectives

Next, we consider the setting where f is L-smooth on $\mathcal{X}$. In this case, we need an extra assumption.

Assumption 6. *We have access to a constant G_1 satisfying $G_1 \geq \|\nabla f(\boldsymbol{x}_1)\|$.*

Discussion on Assumption 6. We note that the same assumption was also imposed before in [32] under the setting $r = 0$ (see their Assumption (5)). As pointed out by [32], this is a mild condition, since there are several ways to estimate G_1. For example, when $\mathcal{X} = \mathbb{R}^d$ and $r = 0$ as in most prior work, we have $\nabla f(\boldsymbol{x}_\star) = \mathbf{0}$ and hence can set $G_1 = LD$ by noticing that $\|\nabla f(\boldsymbol{x}_1)\| = \|\nabla f(\boldsymbol{x}_1) - \nabla f(\boldsymbol{x}_\star)\| \leq LD$. For more details on the justification of Assumption 6, we kindly refer the interested reader to [32].

Equipped with Assumption 6, we can prove the following Theorem 2.

Theorem 2. *Under Assumptions 1, 2 (with $\mu = 0$), 4, 5, and 6, for any $T \in \mathbb{N}$ and $\delta \in \left(0, \frac{1}{2}\right]$, setting $\eta_t = \eta_\star(T - t + 1), \tau_t = \max\left\{2L\|\boldsymbol{x}_t - \boldsymbol{x}_1\| + 2G_1, \tau_\star T^{\frac{1}{\mathtt{p}}}\right\}, \forall t \in [T]$, where $\eta_\star$ is a properly picked parameter defined in (19), then* `Clipped-SGD` *(Algorithm 1) guarantees that, with probability at least $1 - 2\delta$,*

$$F(\boldsymbol{x}_{T+1}) - F_\star = \mathcal{O}\left(\frac{(\varphi_{\mathrm{smo}} + \ln\frac{3}{\delta})(LD + G_1)D}{T} + \frac{(\sigma_{\mathtt{s}}^{\frac{2}{\mathtt{p}}-1}\sigma_{\mathtt{1}}^{2-\frac{2}{\mathtt{p}}} + \sigma_{\mathtt{s}}^{\frac{1}{\mathtt{p}}}\sigma_{\mathtt{1}}^{1-\frac{1}{\mathtt{p}}}\ln^{1-\frac{1}{\mathtt{p}}}\frac{3}{\delta})D}{T^{1-\frac{1}{\mathtt{p}}}}\right),$$

in which $\varphi_{\mathrm{smo}} \leq \varphi_\star$ is a constant defined in (12) and equals $\varphi_\star$ when $T = \Omega\left(\frac{(LD+G_1)^{\mathtt{p}}}{\sigma_{\mathtt{1}}^{\mathtt{p}}}\varphi_\star\right)$.

As far as we know, for smooth objectives, a recent work [7] obtained the first and currently only high-probability convergence result for the last iterate of `Clipped-SGD` in the setting where $r = 0$, $\mathcal{X} = \mathbb{R}^d$, and $\sigma_{\mathtt{s}} = \sigma_{\mathtt{1}} = \sigma$ (i.e., the classical heavy-tailed noise assumption) with a rate $\mathcal{O}\left(\frac{LD^2\ln^2\frac{T^2}{\delta}}{T} + \frac{\max\left\{\sigma D, L^{\frac{\mathtt{p}-1}{3\mathtt{p}-1}}\sigma^{\frac{2\mathtt{p}}{3\mathtt{p}-1}}D^{\frac{4\mathtt{p}-2}{3\mathtt{p}-1}}, L^{\frac{1}{3}}\sigma^{\frac{2}{3}}D^{\frac{4}{3}}\right\}\ln^2\frac{T^2}{\delta}}{T^{\frac{2}{3}\left(1-\frac{1}{\mathtt{p}}\right)}}\right)$. However, this rate is much worse than the best-known average-iterate rate $\mathcal{O}\left(\frac{LD^2\ln\frac{1}{\delta}}{T} + \frac{\sigma D\ln^{1-\frac{1}{\mathtt{p}}}\frac{1}{\delta}}{T^{1-\frac{1}{\mathtt{p}}}}\right)$ [32]. In contrast, by taking $G_1 = LD$ and using $d_{\mathtt{eff}} \overset{(1)}{=} \sigma_{\mathtt{1}}^2/\sigma_{\mathtt{s}}^2 = 1 \Rightarrow \varphi_{\mathrm{smo}} \leq \varphi_\star \overset{(3)}{=} \ln\frac{3}{\delta}$ in this special case, the last-iterate rate in Theorem 2 can be simplified into $\mathcal{O}\left(\frac{LD^2\ln\frac{1}{\delta}}{T} + \frac{\sigma D\ln^{1-\frac{1}{\mathtt{p}}}\frac{1}{\delta}}{T^{1-\frac{1}{\mathtt{p}}}}\right)$, which matches the bound in [32] and substantially improves over [7]. Moreover, we believe that the leading term in this simplified rate is optimal, although, as previously noted, no high-probability lower bound exists in this setting.

Notably, our Theorem 2 holds in a more general setting. In particular, it is the first provable result applicable to smooth objectives under the fine-grained Assumption 5. Although it is currently unclear whether the bound is tight or not (since the lower bounds in [23] were proved only for Lipschitz f), we believe that a conclusion similar to the previous one for the Lipschitz case holds, that is, the leading term in Theorem 2 is tight in T, whereas the term $\sigma_{\mathtt{s}}^{1/\mathtt{p}}\sigma_{\mathtt{1}}^{1-1/\mathtt{p}}$ may not be optimal (and we conjecture that the optimal dependence should be $\sigma_{\mathtt{s}}$).

In addition, we would like to briefly discuss the first component, $2L\|\boldsymbol{x}_t - \boldsymbol{x}_1\| + 2G_1$, in our clipping threshold τ_t. The reader familiar with the literature may immediately recognize a similar choice, $2L\max_{s\in[t]}\|\boldsymbol{x}_s - \boldsymbol{x}_1\| + 2G_1$, proposed in [32], which directly inspired our choice. However, as one can see, ours made a simplification by replacing $\max_{s\in[t]}\|\boldsymbol{x}_s - \boldsymbol{x}_1\|$ with $\|\boldsymbol{x}_t - \boldsymbol{x}_1\|$.

3.2 Strongly Star-Convex Optimization

This subsection focuses on strongly star-convex optimization, i.e., $\mu > 0$ in Assumption 2. To avoid algebraic issues in the analysis, we consider only $T \geq 2$.

3.2.1 Lipschitz Objectives

As before, we begin with the case where f is G-Lipschitz on $\mathcal{X}$.

Theorem 3. *Under Assumptions 1, 2 (with $\mu > 0$), 3, and 5, for any $T \geq 2$ and $\delta \in \left(0, \frac{1}{2}\right]$, setting* $\eta_t = \begin{cases} \frac{6}{\mu t} & t \leq S \\ \frac{T-t+1}{\mu T^2} & t > S \end{cases}, \tau_t = \begin{cases} \max\left\{2G, \tau_\star t^{\frac{1}{\mathtt{p}}}\right\} & t \leq S \\ \max\left\{2G, \tau_\star \left\lfloor \frac{T}{2} \right\rfloor^{\frac{1}{\mathtt{p}}}\right\} & t > S \end{cases}, \forall t \in [T]$, *where $S \triangleq \left\lceil \frac{T}{2} \right\rceil$, then* `Clipped-SGD` *(Algorithm 1) guarantees that, with probability at least $1 - 2\delta$,*

$$F(\boldsymbol{x}_{T+1}) - F_\star = \mathcal{O}\left(\frac{\mu D^2}{T^3} + \frac{(\varphi_{\mathrm{lip}}^2 + \ln^2 \frac{3}{\delta})G^2}{\mu T^2} + \frac{G^2}{\mu T} + \frac{\sigma_{\mathtt{s}}^{\frac{4}{\mathtt{p}}-2} \sigma_{\mathtt{1}}^{4-\frac{4}{\mathtt{p}}} + \sigma_{\mathtt{s}}^{\frac{2}{\mathtt{p}}} \sigma_{\mathtt{1}}^{2-\frac{2}{\mathtt{p}}} \ln^{2-\frac{2}{\mathtt{p}}} \frac{3}{\delta}}{\mu T^{2-\frac{2}{\mathtt{p}}}}\right),$$

in which $\varphi_{\mathrm{lip}} \leq \varphi_\star$ is a constant defined in (11) and equals $\varphi_\star$ when $T = \Omega\left(\frac{G^{\mathtt{p}}}{\sigma_{\mathtt{1}}^{\mathtt{p}}} \varphi_\star\right)$.

Theorem 3 presents the first last-iterate convergence rate for `Clipped-SGD` in stochastic strongly star-convex optimization under heavy-tailed noise, which fills out the missing pieces in the literature, as no such result was known even in the simple case $\mathtt{p} = 2$ and $\sigma_{\mathtt{s}} = \sigma_{\mathtt{1}} = \sigma$, i.e., classical finite-variance noise with noise level $\sigma \geq 0$.

As one can check, Theorem 3 matches the previously fastest rate proved for $F(\bar{\boldsymbol{x}}_{T+1}) - F_\star$ (where $\bar{\boldsymbol{x}}_{T+1}$ is a weighted average of $\{\boldsymbol{x}_t\}_{t=2}^{T+1}$) and $\mu \left\|\boldsymbol{x}_{T+1} - \boldsymbol{x}_\star\right\|^2$ [23]. We remind the reader that, as mentioned earlier in Footnote 2, $F(\boldsymbol{x}_{T+1}) - F_\star$ always serves as a stronger convergence criterion than $\mu \left\|\boldsymbol{x}_{T+1} - \boldsymbol{x}_\star\right\|^2$, implying that our new result is more favorable than [23]. However, similar to the case of $\mu = 0$ discussed previously, there is also a gap between the leading term $\mathcal{O}\left(\frac{\sigma_{\mathtt{s}}^{\frac{4}{\mathtt{p}}-2} \sigma_{\mathtt{1}}^{4-\frac{4}{\mathtt{p}}} + \sigma_{\mathtt{s}}^{\frac{2}{\mathtt{p}}} \sigma_{\mathtt{1}}^{2-\frac{2}{\mathtt{p}}} \ln^{2-\frac{2}{\mathtt{p}}} \frac{3}{\delta}}{\mu T^{2-\frac{2}{\mathtt{p}}}}\right)$ in Theorem 3 and the lower bound $\Omega\left(\frac{\sigma_{\mathtt{s}}^{\frac{4}{\mathtt{p}}-2} \sigma_{\mathtt{1}}^{4-\frac{4}{\mathtt{p}}} + \sigma_{\mathtt{s}}^{2} \ln^{2-\frac{2}{\mathtt{p}}} \frac{1}{\delta}}{\mu T^{2-\frac{2}{\mathtt{p}}}}\right)$ proved by [23], closing which could be an interesting task that we leave in the future.

One may observe that our choices of $\{\eta_t\}_{t=1}^T$ and $\{\tau_t\}_{t=1}^T$ have a two-stage structure, and especially, they coincide with those chosen in [23] during the first stage, i.e., when $t \leq \left\lceil \frac{T}{2} \right\rceil$. We note that this is done intentionally as part of our proof. Intuitively, we use the first stage to establish convergence in squared distance, which is finally converted into convergence in function value. The interested reader could refer to the proof of Theorem 7 for details.

3.2.2 Smooth Objectives

Lastly, we turn to the setting where f is L-smooth on $\mathcal{X}$. In this case, we require another assumption different from the previous Assumption 6 for smooth convex optimization.

Assumption 7. *We have access to a constant G_1 satisfying $G_1 \geq \max\left\{\left\|\nabla F(\boldsymbol{x}_1)\right\|, \left\|\nabla f(\boldsymbol{x}_1)\right\|\right\}$. In addition, $\mathcal{X}$ is compact with diameter at most $2G_1/\mu$.*

Remark 5. Strictly speaking, $\|\nabla F(\boldsymbol{x}_1)\|$ should be replaced by $\max_{\boldsymbol{g} \in \partial r(\boldsymbol{x}_1)} \|\nabla f(\boldsymbol{x}_1) + \boldsymbol{g}\|$, given that r may be non-differentiable, e.g., $r(\boldsymbol{x}) = \|\boldsymbol{x}\|_1$. However, we simply write it as $\|\nabla F(\boldsymbol{x}_1)\|$ to simplify the expression.

Discussion on Assumption 7. One might consider the compactness requirement on $\mathcal{X}$ too restrictive. However, we would like to argue that this entails no loss of generality due to the following Fact 1.

Fact 1. *Under Assumptions 1 and 2 (with $\mu > 0$), we have $\boldsymbol{x}_\star \in \mathbb{B}^d\left(\boldsymbol{x}_1, \left\|\nabla F(\boldsymbol{x}_1)\right\|/\mu\right)$.*

Proof. By Lemma 1, $\frac{\mu}{2}\left\|\boldsymbol{x}_1 - \boldsymbol{x}_\star\right\|^2 \leq F(\boldsymbol{x}_1) - F_\star \leq \frac{\|\nabla F(\boldsymbol{x}_1)\|^2}{2\mu} \Rightarrow \boldsymbol{x}_\star \in \mathbb{B}^d\left(\boldsymbol{x}_1, \left\|\nabla F(\boldsymbol{x}_1)\right\|/\mu\right)$. □

By Fact 1 and the definition of G_1 in Assumption 7, we know $\boldsymbol{x}_\star \in \mathbb{B}^d\left(\boldsymbol{x}_1, G_1/\mu\right)$, which further implies that

$$F_\star = F(\boldsymbol{x}_\star) = \inf_{\boldsymbol{x} \in \mathcal{X}} F(\boldsymbol{x}) = \inf_{\boldsymbol{x} \in \mathcal{X} \cap \mathbb{B}^d(\boldsymbol{x}_1, G_1/\mu)} F(\boldsymbol{x}).$$

Hence, even if $\mathcal{X}$ is not compact, we can replace $\mathcal{X}$ with $\mathcal{X} \cap \mathbb{B}^d\left(\boldsymbol{x}_1, G_1/\mu\right)$, which preserves the same optimal solution and restricts the problem to a compact domain with diameter at most $2G_1/\mu$.

With Assumption 7 in hand, we can prove the following Theorem 4.

Theorem 4. *Under Assumptions 1, 2 (with $\mu > 0$), 4, 5, and 7, for any $T \geq 2$ and $\delta \in \left(0, \frac{1}{2}\right]$, setting* $\eta_t = \begin{cases} \frac{6}{\mu(t+12\kappa)} & t \leq S \\ \frac{T-t+1}{\mu T(T+\kappa)} & t > S \end{cases}, \tau_t = \begin{cases} \max\left\{2L\left\|\boldsymbol{x}_t - \boldsymbol{x}_1\right\| + 2G_1, \tau_\star \left[T(T+\kappa)\right]^{\frac{1}{2\mathrm{p}}}\right\} & t \leq S \\ \max\left\{2(1+2\kappa)G_1, \tau_\star \left\lfloor \frac{T}{2} \right\rfloor^{\frac{1}{\mathrm{p}}}\right\} & t > S \end{cases}, \forall t \in [T]$*, where $S \triangleq \left\lceil \frac{T}{2} \right\rceil$ and $\kappa \triangleq \frac{L}{\mu}$, then* `Clipped-SGD` *(Algorithm 1) guarantees that, with probability at least $1 - 2\delta$,*

$$F(\boldsymbol{x}_{T+1}) - F_\star = \mathcal{O}\left(\frac{(\varphi_{\mathrm{smo}}^2 + \ln^2\frac{3}{\delta})(1+\kappa)^2 G_1^2}{\mu T^2} + \frac{\sigma_{\mathtt{s}}^{\frac{4}{\mathrm{p}}-2}\sigma_{\mathtt{1}}^{4-\frac{4}{\mathrm{p}}} + \sigma_{\mathtt{s}}^{\frac{2}{\mathrm{p}}}\sigma_{\mathtt{1}}^{2-\frac{2}{\mathrm{p}}}\ln^{2-\frac{2}{\mathrm{p}}}\frac{3}{\delta}}{\mu\left[T(T+\kappa)\right]^{1-\frac{1}{\mathrm{p}}}}\right),$$

in which $\varphi_{\mathrm{smo}} \leq \varphi_\star$ is a constant defined in (12) and equals $\varphi_\star$ when $T = \Omega\left(\frac{(1+\kappa)^{\mathrm{p}} G_{\mathtt{1}}^{\mathrm{p}}}{\sigma_{\mathtt{1}}^{\mathrm{p}}}\varphi_\star\right)$.

Theorem 4 gives not only the first (and, in fact, the only) high-probability convergence result in terms of $F(\boldsymbol{x}_{T+1}) - F_\star$ for `Clipped-SGD` in stochastic strongly star-convex optimization, but also the first upper bound for smooth objectives under the fine-grained Assumption 5, exhibiting a finer dependence on the noise parameter similar to Theorem 3.

Under the classical heavy-tailed noise assumption with noise level σ, Theorem 4 results in a leading term $\mathcal{O}\left(\frac{\sigma^2 \ln^{2-\frac{2}{\mathrm{p}}}\frac{1}{\delta}}{\mu[T(T+\kappa)]^{1-\frac{1}{\mathrm{p}}}}\right)$ for $F(\boldsymbol{x}_{T+1}) - F_\star$, which improves over the best-known but suboptimal one $\widetilde{\mathcal{O}}\left(\frac{\sigma^2 \ln^{2-\frac{2}{\mathrm{p}}}\frac{T}{\delta}}{\mu T^{2-\frac{2}{\mathrm{p}}}}\right)$ for $\mu\left\|\boldsymbol{x}_{T+1} - \boldsymbol{x}_\star\right\|^2$ in the same setting [38] (which additionally requires $r = 0$). However, we note that the higher-order term in [38] decays exponentially and is therefore better than ours, which decays only polynomially. It will be interesting to investigate whether a more sophisticated design of $\{\eta_t\}_{t=1}^T$ and $\{\tau_t\}_{t=1}^T$, such as a three-stage schedule [26], could refine Theorem 4 and yield an exponentially decaying higher-order term.

Regarding the optimality of Theorem 4, even though no high-probability lower bounds currently exist under either the classical or the fine-grained heavy-tailed noise assumption, we believe that the leading term in Theorem 4 is always optimal in T, δ, and μ, while the term $\sigma_{\mathtt{s}}^{2/\mathrm{p}}\sigma_{\mathtt{1}}^{2-2/\mathrm{p}}$ may be suboptimal under the fine-grained heavy-tailed noise assumption.

Finally, as in Theorem 3, Theorem 4 also adopts a two-stage structure for $\{\eta_t\}_{t=1}^T$ and $\{\tau_t\}_{t=1}^T$. The reason for this choice will become clear in the proof of Theorem 8.

4 New Idea in Theoretical Analysis

In this section, we describe an interesting idea developed in our theoretical analysis. Before beginning, we introduce some notation that will be used later. In the following, let $\mathcal{F}_t \triangleq \sigma\left(Z_1, \ldots, Z_t\right)$ denote the natural filtration, and $\mathbb{E}_t\left[\cdot\right] \triangleq \mathbb{E}\left[\cdot \mid \mathcal{F}_t\right]$ denote the conditional expectation.

Our contribution is a novel argument that goes beyond standard induction while preserving the optimal $\mathrm{poly}(\ln(1/\delta))$ dependence on δ. To illustrate this point, we use smooth convex optimization as an example. In the discussion below, we assume that both $\{\eta_t\}_{t=1}^T$ and $\{\tau_t\}_{t=1}^T$ are deterministic to simplify the presentation.

To prove convergence of `Clipped-SGD` in this case (even for the average iterate), one typically needs to show that $\left\|\boldsymbol{x}_\star - \boldsymbol{x}_{t+1}\right\|^2 = \mathcal{O}(D^2), \forall t \in [T]$ holds in high probability. A common strategy in prior work is induction (e.g., [15, 38]), which consists of the following two parts.

Part I. One establishes a collection of inequalities of the form $\left\|\boldsymbol{x}_\star - \boldsymbol{x}_{t+1}\right\|^2 = \mathcal{O}(D^2 + (\sum_{s=1}^t X_s)^2 + \sum_{s=1}^t Y_s^2), \forall t \in [T]$, where $\{X_t\}_{t=1}^T$ is a martingale difference sequence (MDS) adapted to the filtration $\{\mathcal{F}_t\}_{t=0}^T$, and $\{Y_t\}_{t=1}^T$ is a nonnegative predictable sequence with respect to $\{\mathcal{F}_t\}_{t=0}^T$. In particular, they satisfy $|X_t| \leq h_1(\left\|\boldsymbol{x}_\star - \boldsymbol{x}_t\right\|, \eta_t, \tau_t)$, $\mathbb{E}_{t-1}\left[X_t^2\right] \leq h_2(\left\|\boldsymbol{x}_\star - \boldsymbol{x}_t\right\|, \eta_t, \tau_t)$, and $Y_t \leq h_3(\left\|\boldsymbol{x}_\star - \boldsymbol{x}_t\right\|, \eta_t, \tau_t)$, for three nonnegative functions h_1, h_2, and h_3, each of which is nondecreasing in its first argument.

Part II. Let $\zeta_t \triangleq \mathbb{1}\left[\|\boldsymbol{x}_\star - \boldsymbol{x}_t\| = \mathcal{O}(D)\right], \forall t \in [T+1]$ and $E_t \triangleq \{\zeta_s = 1, \forall s \leq t\}, \forall t \in [T+1]$. Existing induction-based analysis assumes that, for a given $t \in [T]$, $\Pr[E_t] \geq 1 - \frac{t-1}{T}\delta$. Then, on the event E_t, we have $\|\boldsymbol{x}_\star - \boldsymbol{x}_{t+1}\|^2 = \mathcal{O}(D^2 + (\sum_{s=1}^t X_s\zeta_s)^2 + \sum_{s=1}^t Y_s^2\zeta_s^2)$. Since $\zeta_t \in \mathcal{F}_{t-1}$, we know $|X_t\zeta_t| \leq h_{1,t} = \mathcal{O}(h_1(D, \eta_t, \tau_t))$, $\mathbb{E}_{t-1}\left[X_t^2\zeta_t^2\right] \leq h_{2,t} = \mathcal{O}(h_2(D, \eta_t, \tau_t))$, and $Y_t\zeta_t \leq h_{3,t} = \mathcal{O}(h_3(D, \eta_t, \tau_t))$, which further implies $\Pr[e_t] \geq 1 - \frac{\delta}{T}$ by Freedman's inequality (cf. Lemma 11), where $e_t \triangleq \left\{\left|\sum_{s=1}^t X_s\zeta_s\right| \leq \mathcal{O}\left(\max_{s\in[t]} h_{1,s} \ln\frac{T}{\delta} + \sqrt{\sum_{s=1}^t h_{2,s}\ln\frac{T}{\delta}}\right)\right\}, \forall t \in [T]$. Note that, on the event $e_t \cap E_t$, $\|\boldsymbol{x}_\star - \boldsymbol{x}_{t+1}\|^2 = \mathcal{O}(D^2 + \max_{s\in[t]} h_{1,s}^2 \ln^2\frac{T}{\delta} + \sum_{s=1}^t h_{2,s}\ln\frac{T}{\delta} + \sum_{s=1}^t h_{3,s}^2) \overset{\text{proper } \{\eta_t\}_{t=1}^T \text{ and } \{\tau_t\}_{t=1}^T}{=} \mathcal{O}(D^2) \Rightarrow \zeta_{t+1} = 1$. Thus, $E_{t+1} = E_t \cap \{\zeta_{t+1} = 1\} \supseteq e_t \cap E_t \cap \{\zeta_{t+1} = 1\} = e_t \cap E_t$ and hence $\Pr[E_{t+1}] \geq 1 - \frac{t}{T}\delta$. This completes the induction.

As can be seen, the induction argument in **Part II** inevitably introduces a suboptimal $\mathrm{poly}(\ln(T/\delta))$ factor into the choices of $\{\eta_t\}_{t=1}^T$ and $\{\tau_t\}_{t=1}^T$, and consequently into the final convergence rate. However, we observe that this issue can be solved by a new refined argument, introduced below.

Refined Part II (Ours). We observe that $\Pr[E] \geq 1 - \delta$ by Freedman's inequality (cf. Lemma 11), where $E \triangleq \left\{\max_{t\in[T]}\left|\sum_{s=1}^t X_s\zeta_s\right| \leq \mathcal{O}\left(\max_{t\in[T]} h_{1,t}\ln\frac{1}{\delta} + \sqrt{\sum_{t=1}^T h_{2,t}\ln\frac{1}{\delta}}\right)\right\}$. Suppose we pick $\{\eta_t\}_{t=1}^T$ and $\{\tau_t\}_{t=1}^T$ to satisfy $\mathcal{O}(D^2 + \max_{t\in[T]} h_{1,t}^2\ln^2\frac{1}{\delta} + \sum_{t=1}^T h_{2,t}\ln\frac{1}{\delta} + \sum_{t=1}^T h_{3,t}^2) = \mathcal{O}(D^2)$, then we claim $\{\zeta_{t+1} = 1\} \supseteq E \cap E_t, \forall t \in [T]$. Given $t \in [T]$, under $E \cap E_t$, we observe

$$\begin{aligned}\|\boldsymbol{x}_\star - \boldsymbol{x}_{t+1}\|^2 &= \mathcal{O}(D^2 + (\sum_{s=1}^t X_s\zeta_s)^2 + \sum_{s=1}^t Y_s^2\zeta_s^2) \leq \mathcal{O}(D^2 + \max_{t\in[T]}(\sum_{s=1}^t X_s\zeta_s)^2 + \sum_{t=1}^T Y_t^2\zeta_t^2)\\ &= \mathcal{O}(D^2 + \max_{t\in[T]} h_{1,t}^2\ln^2\frac{1}{\delta} + \sum_{t=1}^T h_{2,t}\ln\frac{1}{\delta} + \sum_{t=1}^T h_{3,t}^2) = \mathcal{O}(D^2) \Rightarrow \zeta_{t+1} = 1.\end{aligned}$$

Hence, our claim holds, which further implies that $E \cap E_{t+1} = E \cap E_t \cap \{\zeta_{t+1} = 1\} = E \cap E_t, \forall t \in [T] \Rightarrow E_{T+1} \supseteq E \cap E_{T+1} = E \cap E_1 = E \Rightarrow \Pr[E_{T+1}] \geq 1 - \delta$. As one can see, our new argument maintains the optimal $\mathrm{poly}(\ln(1/\delta))$ dependence on δ. For the formal application of the above idea, we refer the reader to Lemma 7.

5 Conclusion and Limitations

This work proves the first last-iterate convergence rates optimal in the time horizon T for `Clipped-SGD` in high probability under heavy-tailed noise across various settings, thereby closing an existing gap in the literature. Specifically, we provide last-iterate convergence guarantees for both convex and strongly star-convex objectives, covering smooth and nonsmooth settings. Under the classical heavy-tailed noise assumption (i.e., $\sigma_{\mathtt{s}} = \sigma_{\mathtt{1}} = \sigma$ in Assumption 5), these are the first minimax-optimal high-probability last-iterate bounds for Lipschitz objectives; matching lower bounds for smooth objectives remain open. Notably, our analysis also accommodates composite objectives and constrained optimization, and captures a refined dependence on noise parameters under the fine-grained Assumption 5.

Our work nevertheless has two major limitations. First, as in prior work, the stepsize and clipping threshold depend on knowledge of the tail index $\mathtt{p}$, which is generally unknown and difficult to estimate reliably in practice. Removing this requirement would broaden the applicability of our results. Second, our analysis assumes that the time horizon T is known in advance, and thus our current guarantees do not directly extend to the anytime setting. Developing methods that adapt to an unknown tail index while providing anytime last-iterate guarantees is therefore an interesting direction for future research.

Acknowledgments

The author thanks the anonymous reviewers for their valuable feedback.

References

[1] Kwangjun Ahn, Xiang Cheng, Minhak Song, Chulhee Yun, Ali Jadbabaie, and Suvrit Sra. Linear attention is (maybe) all you need (to understand transformer optimization). In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun, editors, *International Conference on Learning Representations*, volume 2024, pages 16193–16205, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/4661b55200c03a8c4bb9c2974b4fb12d-Paper-Conference.pdf.

[2] Aleksandar Armacki, Shuhua Yu, Pranay Sharma, Gauri Joshi, Dragana Bajovic, Dusan Jakovetic, and Soummya Kar. High-probability convergence bounds for online nonlinear stochastic gradient descent under heavy-tailed noise. In Yingzhen Li, Stephan Mandt, Shipra Agrawal, and Emtiyaz Khan, editors, *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pages 1774–1782. PMLR, 03–05 May 2025. URL https://proceedings.mlr.press/v258/armacki25a.html.

[3] Amit Attia, Matan Schliserman, Uri Sherman, and Tomer Koren. Fast last-iterate convergence of sgd in the smooth interpolation regime. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen, editors, *Advances in Neural Information Processing Systems*, volume 38, pages 104951–104987. Curran Associates, Inc., 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/97335d45cab57ecb52727c4380ca14a4-Paper-Conference.pdf.

[4] Barak Battash, Lior Wolf, and Ofir Lindenbaum. Revisiting the noise model of stochastic gradient descent. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li, editors, *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 4780–4788. PMLR, 02–04 May 2024. URL https://proceedings.mlr.press/v238/battash24a.html.

[5] Léon Bottou, Frank E. Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018. doi: 10.1137/16M1080173. URL https://doi.org/10.1137/16M1080173.

[6] Yair Carmon and Oliver Hinder. The price of adaptivity in stochastic convex optimization. In Shipra Agrawal and Aaron Roth, editors, *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pages 772–774. PMLR, 30 Jun–03 Jul 2024. URL https://proceedings.mlr.press/v247/carmon24a.html.

[7] Savelii Chezhegov, Daniela Angela Parletta, Andrea Paudice, and Eduard Gorbunov. High-probability bounds for the last iterate of clipped sgd. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust, editors, *International Conference on Learning Representations*, volume 2026, pages 1280–1310, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/03382840834c7fbdc1e882fac11904d6-Paper-Conference.pdf.

[8] Ashok Cutkosky and Harsh Mehta. High-probability bounds for non-convex stochastic optimization with heavy tails. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 4883–4895. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/26901debb30ea03f0aa833c9de6b81e9-Paper.pdf.

[9] Aniket Das, Dheeraj Nagaraj, Soumyabrata Pal, Arun Sai Suggala, and Prateek Varshney. Near-optimal streaming heavy-tailed statistical estimation with clipped sgd. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 8834–8900. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/10bf96894abaf4c293b205709a98fc74-Paper-Conference.pdf.

[10] Ilyas Fatkhullin, Florian Hübler, and Guanghui Lan. Can sgd handle heavy-tailed noise? *arXiv preprint arXiv:2508.04860*, 2025.

[11] Adrien Fradin, Abdurakhmon Sadiev, Laurent Condat, and Peter Richtárik. Tight lower bounds and optimal algorithms for stochastic nonconvex optimization with heavy-tailed noise. In Emtiyaz Khan, Yingzhen Li, Arno Solin, and Aaditya Ramdas, editors, *Proceedings of The 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pages 2548–2556. PMLR, 02–05 May 2026. URL https://proceedings.mlr.press/v300/fradin26a.html.

[12] David A. Freedman. On Tail Probabilities for Martingales. *The Annals of Probability*, 3(1): 100 – 118, 1975. doi: 10.1214/aop/1176996452. URL https://doi.org/10.1214/aop/1176996452.

[13] Saurabh Garg, Joshua Zhanson, Emilio Parisotto, Adarsh Prasad, Zico Kolter, Zachary Lipton, Sivaraman Balakrishnan, Ruslan Salakhutdinov, and Pradeep Ravikumar. On proximal policy optimization's heavy-tailed gradients. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 3610–3619. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/garg21b.html.

[14] Guillaume Garrigos, Daniel Cortild, Lucas Ketels, and Juan Peypouquet. Last-iterate complexity of sgd for convex and smooth stochastic problems. *arXiv preprint arXiv:2507.14122*, 2025.

[15] Eduard Gorbunov, Abdurakhmon Sadiev, Marina Danilova, Samuel Horváth, Gauthier Gidel, Pavel Dvurechensky, Alexander Gasnikov, and Peter Richtárik. High-probability convergence for composite and distributed stochastic minimization and variational inequalities with heavy-tailed noise. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 15951–16070. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/gorbunov24a.html.

[16] Nicholas J. A. Harvey, Christopher Liaw, Yaniv Plan, and Sikander Randhawa. Tight analyses for non-smooth stochastic gradient descent. In Alina Beygelzimer and Daniel Hsu, editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 1579–1613. PMLR, 25–28 Jun 2019. URL https://proceedings.mlr.press/v99/harvey19a.html.

[17] Oliver Hinder, Aaron Sidford, and Nimit Sohoni. Near-optimal methods for minimizing star-convex functions and beyond. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 1894–1938. PMLR, 09–12 Jul 2020. URL https://proceedings.mlr.press/v125/hinder20a.html.

[18] Florian Hübler, Ilyas Fatkhullin, and Niao He. From gradient clipping to normalization for heavy tailed sgd. In Yingzhen Li, Stephan Mandt, Shipra Agrawal, and Emtiyaz Khan, editors, *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pages 2413–2421. PMLR, 03–05 May 2025. URL https://proceedings.mlr.press/v258/hubler25a.html.

[19] Maor Ivgi, Oliver Hinder, and Yair Carmon. DoG is SGD's best friend: A parameter-free dynamic step size schedule. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 14465–14499. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/ivgi23a.html.

[20] Prateek Jain, Dheeraj M. Nagaraj, and Praneeth Netrapalli. Making the last iterate of sgd information theoretically optimal. *SIAM Journal on Optimization*, 31(2):1108–1130, 2021. doi: 10.1137/19M128908X. URL https://doi.org/10.1137/19M128908X.

[21] Guanghui Lan. *First-order and stochastic optimization methods for machine learning*. Springer, 2020.

[22] Langqi Liu, Yibo Wang, and Lijun Zhang. High-probability bound for non-smooth non-convex stochastic optimization with heavy tails. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 32122–32138. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/liu24bo.html.

[23] Zijian Liu. Clipped gradient methods for nonsmooth convex optimization under heavy-tailed noise: A refined analysis. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust, editors, *International Conference on Learning Representations*, volume 2026, pages 116560–116608, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/bd5b846a033439d079ed65169e660078-Paper-Conference.pdf.

[24] Zijian Liu. Online convex optimization with heavy tails: Old algorithms, new regrets, and applications. In Matus Telgarsky and Jonathan Ullman, editors, *Proceedings of The 37th International Conference on Algorithmic Learning Theory*, volume 313 of *Proceedings of Machine Learning Research*, pages 1–47. PMLR, 23–26 Feb 2026. URL https://proceedings.mlr.press/v313/liu26a.html.

[25] Zijian Liu and Zhengyuan Zhou. Stochastic nonsmooth convex optimization with heavy-tailed noises: High-probability bound, in-expectation rate and initial distance adaptation. *arXiv preprint arXiv:2303.12277*, 2023.

[26] Zijian Liu and Zhengyuan Zhou. Revisiting the last-iterate convergence of stochastic gradient methods. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun, editors, *International Conference on Learning Representations*, volume 2024, pages 40394–40428, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/b1239ceb879da48b4bac7aa91ea53162-Paper-Conference.pdf.

[27] Zijian Liu, Jiawei Zhang, and Zhengyuan Zhou. Breaking the lower bound with (little) structure: Acceleration in non-convex stochastic optimization with heavy-tailed noise. In Gergely Neu and Lorenzo Rosasco, editors, *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 2266–2290. PMLR, 12–15 Jul 2023. URL https://proceedings.mlr.press/v195/liu23c.html.

[28] Eric Moulines and Francis Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011. URL https://proceedings.neurips.cc/paper_files/paper/2011/file/40008b9a5380fcacce3976bf7c08af5b-Paper.pdf.

[29] Ion Necoara, Yu Nesterov, and Francois Glineur. Linear convergence of first order methods for non-strongly convex optimization. *Mathematical programming*, 175(1):69–107, 2019.

[30] Arkadi Nemirovski and David Yudin. Problem complexity and method efficiency in optimization. *Wiley-Interscience*, 1983.

[31] Yurii Nesterov. *Lectures on convex optimization*, volume 137. Springer, 2018.

[32] Ta Duy Nguyen, Thien H Nguyen, Alina Ene, and Huy Nguyen. Improved convergence in high probability of clipped gradient methods with heavy tailed noise. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 24191–24222. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/4c454d34f3a4c8d6b4ca85a918e5d7ba-Paper-Conference.pdf.

[33] Francesco Orabona. Last iterate of sgd converges (even in unbounded domains). 2020. URL https://parameterfree.com/2020/08/07/last-iterate-of-sgd-converges-even-in-unbounded-domains/.

[34] Daniela Angela Parletta, Andrea Paudice, and Saverio Salzo. An improved analysis of the clipped stochastic subgradient method under heavy-tailed noise. *arXiv preprint arXiv:2410.00573*, 2024.

[35] Maxim Raginsky and Alexander Rakhlin. Information complexity of black-box convex optimization: A new look via feedback information theory. In *2009 47th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 803–810, 2009. doi: 10.1109/ALLERTON.2009.5394945.

[36] Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. *arXiv preprint arXiv:1109.5647*, 2011.

[37] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[38] Abdurakhmon Sadiev, Marina Danilova, Eduard Gorbunov, Samuel Horváth, Gauthier Gidel, Pavel Dvurechensky, Alexander Gasnikov, and Peter Richtárik. High-probability bounds for stochastic optimization and variational inequalities: the case of unbounded variance. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 29563–29648. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/sadiev23a.html.

[39] Ohad Shamir and Tong Zhang. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 71–79, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL https://proceedings.mlr.press/v28/shamir13.html.

[40] Umut Simsekli, Levent Sagun, and Mert Gurbuzbalaban. A tail-index analysis of stochastic gradient noise in deep neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5827–5837. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/simsekli19a.html.

[41] Nuri Mert Vural, Lu Yu, Krishna Balasubramanian, Stanislav Volgushev, and Murat A Erdogdu. Mirror descent strikes again: Optimal stochastic convex optimization under infinite noise variance. In Po-Ling Loh and Maxim Raginsky, editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 65–102. PMLR, 02–05 Jul 2022. URL https://proceedings.mlr.press/v178/vural22a.html.

[42] Moslem Zamani and François Glineur. Exact convergence rate of the last iterate in subgradient methods. *SIAM Journal on Optimization*, 35(3):2182–2201, 2025. doi: 10.1137/24M1717762. URL https://doi.org/10.1137/24M1717762.

[43] Jingzhao Zhang, Sai Praneeth Karimireddy, Andreas Veit, Seungyeon Kim, Sashank Reddi, Sanjiv Kumar, and Suvrit Sra. Why are adaptive methods good for attention models? In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 15383–15393. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/b05b57f6add810d3b7490866d74c0053-Paper.pdf.

[44] Jiujia Zhang and Ashok Cutkosky. Parameter-free regret in high probability with heavy tails. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 8000–8012. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/349956dee974cfdcbbb2d06afad5dd4a-Paper-Conference.pdf.

A Proof of Lemma 1

Proof. If $\mu = 0$, the inequality holds automatically. Thus, we can focus only on the case $\mu > 0$. By Assumption 2, for any $\boldsymbol{x} \in \mathcal{X}$ and $\lambda \in [0, 1]$,

$$F(\lambda \boldsymbol{x}_\star + (1-\lambda)\boldsymbol{x}) + \frac{\mu\lambda(1-\lambda)}{2} \|\boldsymbol{x}_\star - \boldsymbol{x}\|^2 \leq \lambda F_\star + (1-\lambda)F(\boldsymbol{x}). \tag{4}$$

Combining $F(\lambda \boldsymbol{x}_\star + (1-\lambda)\boldsymbol{x}) \geq F_\star$ and (4), we obtain that

$$F_\star + \frac{\mu\lambda(1-\lambda)}{2} \|\boldsymbol{x}_\star - \boldsymbol{x}\|^2 \leq \lambda F_\star + (1-\lambda)F(\boldsymbol{x}), \forall \boldsymbol{x} \in \mathcal{X}, \lambda \in [0,1]\,.$$

Therefore,

$$\frac{\mu\lambda}{2} \|\boldsymbol{x}_\star - \boldsymbol{x}\|^2 \leq F(\boldsymbol{x}) - F_\star, \forall \boldsymbol{x} \in \mathcal{X}, \lambda \in [0,1) \Rightarrow \frac{\mu}{2} \|\boldsymbol{x} - \boldsymbol{x}_\star\|^2 \leq F(\boldsymbol{x}) - F_\star, \forall \boldsymbol{x} \in \mathcal{X}. \tag{5}$$

Combining $F(\lambda \boldsymbol{x}_\star + (1-\lambda)\boldsymbol{x}) \geq F(\boldsymbol{x}) + \langle \nabla F(\boldsymbol{x}), \lambda(\boldsymbol{x}_\star - \boldsymbol{x})\rangle$, due to the convexity of F, and (4), we obtain that

$$F(\boldsymbol{x}) + \langle \nabla F(\boldsymbol{x}), \lambda(\boldsymbol{x}_\star - \boldsymbol{x})\rangle + \frac{\mu\lambda(1-\lambda)}{2} \|\boldsymbol{x}_\star - \boldsymbol{x}\|^2 \leq \lambda F_\star + (1-\lambda)F(\boldsymbol{x}), \forall \boldsymbol{x} \in \mathcal{X}, \lambda \in [0,1]\,.$$

Therefore,

$$F(\boldsymbol{x}) + \langle \nabla F(\boldsymbol{x}), \boldsymbol{x}_\star - \boldsymbol{x}\rangle + \frac{\mu(1-\lambda)}{2} \|\boldsymbol{x}_\star - \boldsymbol{x}\|^2 \leq F_\star, \forall \boldsymbol{x} \in \mathcal{X}, \lambda \in (0,1]\,,$$

which implies that

$$F(\boldsymbol{x}) - F_\star \leq \langle \nabla F(\boldsymbol{x}), \boldsymbol{x} - \boldsymbol{x}_\star\rangle - \frac{\mu}{2} \|\boldsymbol{x}_\star - \boldsymbol{x}\|^2, \forall \boldsymbol{x} \in \mathcal{X}.$$

Since

$$\langle \nabla F(\boldsymbol{x}), \boldsymbol{x} - \boldsymbol{x}_\star\rangle \leq \|\nabla F(\boldsymbol{x})\| \|\boldsymbol{x}_\star - \boldsymbol{x}\| \leq \frac{\|\nabla F(\boldsymbol{x})\|^2}{2\mu} + \frac{\mu}{2} \|\boldsymbol{x}_\star - \boldsymbol{x}\|^2\,,$$

we conclude that

$$F(\boldsymbol{x}) - F_\star \leq \frac{\|\nabla F(\boldsymbol{x})\|^2}{2\mu}, \forall \boldsymbol{x} \in \mathcal{X}. \tag{6}$$

We combine (5) and (6) to complete the proof. □

B Missing Proofs for Theorems

This section provides the missing proofs for theorems presented in the main paper. All intermediate lemmas employed in the analysis will be proved in Appendix C.

Before we begin, let us review the notation introduced previously and also define some new ones. We first recall the term generalized effective dimension $d_{\mathtt{eff}}$, defined in (1) before, as follows:

$$d_{\mathtt{eff}} = \sigma_{\mathtt{1}}^2/\sigma_{\mathtt{s}}^2 \in \{0\} \cup [1, \pi d/2]\,, \tag{7}$$

in which $d_{\mathtt{eff}} = 0$ if and only if $\sigma_{\mathtt{s}} = \sigma_{\mathtt{1}} = 0$, where $\sigma_{\mathtt{s}}$ and $\sigma_{\mathtt{1}}$ are two problem-dependent parameters introduced in Assumption 5. Next, a particularly useful quantity given in (2) is

$$\tau_\star = \left(\min \left\{ \frac{\sigma_{\mathtt{s}} \sigma_{\mathtt{1}}^{\mathtt{p}-1}}{\ln \frac{3}{\delta}}, \frac{\sigma_{\mathtt{s}}^2}{\sigma_{\mathtt{1}}^{2-\mathtt{p}} \mathbb{1}\left[\mathtt{p} < 2\right]} \right\} \right)^{\frac{1}{\mathtt{p}}}, \tag{8}$$

where $\delta \in (0, 1]$ is the failure probability and $\mathtt{p} \in (1, 2]$ is the tail index (see also Assumption 5). Another useful term mentioned before in (3) is

$$\varphi_\star = \max \left\{ \sqrt{d_{\mathtt{eff}}} \ln \frac{3}{\delta}, d_{\mathtt{eff}} \mathbb{1}\left[\mathtt{p} < 2\right] \right\}. \tag{9}$$

In particular, by (7), (8), and (9), the following equation holds

$$\varphi_\star = \frac{\sigma_{\mathtt{1}}^{\mathtt{p}}}{\tau_\star^{\mathtt{p}}}. \tag{10}$$

Moreover, $D = \|\boldsymbol{x}_\star - \boldsymbol{x}_1\|$ always denotes the distance between the optimal solution $\boldsymbol{x}_\star$ and the initial point $\boldsymbol{x}_1$. Finally, we introduce two new quantities in the following.

- For Lipschitz objectives (i.e., Assumption 3), let

$$\varphi_{\mathrm{lip}} \triangleq \min_{\beta \in \left[0, \frac{1}{2}\right]} \frac{\sigma_1^{\mathtt{p}} T^{\beta}}{G^{\beta \mathtt{p}} \tau_\star^{(1-\beta)\mathtt{p}}} \overset{(10)}{=} \min\left\{\varphi_\star, \sqrt{\frac{\sigma_1^{\mathtt{p}}}{G^{\mathtt{p}}} \varphi_\star T}\right\}. \tag{11}$$

  Note that we have $\varphi_{\mathrm{lip}} = \varphi_\star$ when $T = \Omega\left(\frac{G^{\mathtt{p}}}{\sigma_1^{\mathtt{p}}} \varphi_\star\right)$.

- For smooth objectives (i.e., Assumption 4), let

$$\varphi_{\mathrm{smo}} \triangleq \min_{\beta \in \left[0, \frac{1}{2}\right]} \frac{\sigma_1^{\mathtt{p}} T^{\beta}}{\bar{G}^{\beta \mathtt{p}} \tau_\star^{(1-\beta)\mathtt{p}}} \overset{(10)}{=} \min\left\{\varphi_\star, \sqrt{\frac{\sigma_1^{\mathtt{p}}}{\bar{G}^{\mathtt{p}}} \varphi_\star T}\right\}, \tag{12}$$

  in which $\bar{G} \triangleq LD + G_1$ (resp. $\bar{G} \triangleq (1+\kappa)G_1$) and G_1 is defined in Assumption 6 (resp. Assumption 7) for general convex optimization (resp. strongly star-convex optimization). Note that we have $\varphi_{\mathrm{smo}} = \varphi_\star$ when $T = \Omega\left(\frac{\bar{G}^{\mathtt{p}}}{\sigma_1^{\mathtt{p}}} \varphi_\star\right)$.

B.1 General Convex Optimization

We first consider general convex optimization, i.e., $\mu = 0$ in Assumption 2.

B.1.1 Lipschitz Objectives

Theorem 5 (Full statement of Theorem 1)**.** *Under Assumptions 1, 2 (with $\mu = 0$), 3, and 5, for any $T \in \mathbb{N}$ and $\delta \in (0, 1]$, setting $\eta_t = \eta_\star (T - t + 1), \tau_t = \max\left\{2G, \tau_\star T^{\frac{1}{\mathtt{p}}}\right\}, \forall t \in [T]$, where*

$$\eta_\star = \min\left\{\frac{D}{\left(\varphi_{\mathrm{lip}} + \ln\frac{3}{\delta}\right) GT}, \frac{D}{GT^{\frac{3}{2}}}, \frac{D}{\left(\sigma_{\mathtt{s}}^{\frac{2}{\mathtt{p}}-1} \sigma_{\mathtt{l}}^{2-\frac{2}{\mathtt{p}}} + \sigma_{\mathtt{s}}^{\frac{1}{\mathtt{p}}} \sigma_{\mathtt{l}}^{1-\frac{1}{\mathtt{p}}} \ln^{1-\frac{1}{\mathtt{p}}} \frac{3}{\delta}\right) T^{1+\frac{1}{\mathtt{p}}}}\right\}, \tag{13}$$

and $\varphi_{\mathrm{lip}} \leq \varphi_\star$ is a constant defined in (11), then `Clipped-SGD` *(Algorithm 1) guarantees that, with probability at least $1 - \delta$,*

$$F(\boldsymbol{x}_{T+1}) - F_\star = \mathcal{O}\left(\frac{\left(\varphi_{\mathrm{lip}} + \ln\frac{3}{\delta}\right) GD}{T} + \frac{GD}{\sqrt{T}} + \frac{\left(\sigma_{\mathtt{s}}^{\frac{2}{\mathtt{p}}-1} \sigma_{\mathtt{l}}^{2-\frac{2}{\mathtt{p}}} + \sigma_{\mathtt{s}}^{\frac{1}{\mathtt{p}}} \sigma_{\mathtt{l}}^{1-\frac{1}{\mathtt{p}}} \ln^{1-\frac{1}{\mathtt{p}}} \frac{3}{\delta}\right) D}{T^{1-\frac{1}{\mathtt{p}}}}\right).$$

Proof. Given $T \in \mathbb{N}$ and $\delta \in (0, 1]$, note that $\{\eta_t\}_{t=1}^T$ and $\{\tau_t\}_{t=1}^T$ satisfy Conditions 1 and 2 in Lemma 5. Therefore, Lemmas 4 ($\omega = 1$ now) and 5 together yield that, with probability at least $1 - \delta$,

$$F(\boldsymbol{x}_{T+1}) - F_\star \leq \frac{v_0^2 D^2 + 2U_T^{\mathrm{cvx}} + 8\sum_{t=1}^T v_t^2 \eta_t^2 G^2}{\eta_T}, \tag{14}$$

where U_T^{cvx} is a constant satisfying

$$U_T^{\mathrm{cvx}} = \mathcal{O}\left(\max_{t \in [T]} v_t^2 \eta_t^2 \tau_t^2 \ln^2\frac{3}{\delta} + \sum_{t=1}^T \sigma_{\mathtt{l}}^{\mathtt{p}} v_t^2 \eta_t^2 \tau_t^{2-\mathtt{p}} + \left(\sum_{t=1}^T \frac{\sigma_{\mathtt{s}} \sigma_{\mathtt{l}}^{\mathtt{p}-1} v_t \eta_t}{\tau_t^{\mathtt{p}-1}} + \frac{\sigma_{\mathtt{l}}^{\mathtt{p}} G v_t \eta_t}{\tau_t^{\mathtt{p}}}\right)^2\right), \tag{15}$$

and $\left\{v_t^2\right\}_{t=0}^T$ is a sequence defined in (43) as

$$v_t^2 \triangleq \frac{\eta_T}{\sum_{s=t}^T \eta_s}, \forall t \in [T] \quad \text{and} \quad v_0^2 \triangleq v_1^2.$$

We observe that $v_t^2 = \frac{2}{(T-t+2)(T-t+1)}, \forall t \in [T]$ under the stepsize $\eta_t = \eta(T-t+1), \forall t \in [T]$ for any $\eta > 0$. As such, we can simplify (14) and (15) into

$$F(\boldsymbol{x}_{T+1}) - F_\star \leq \frac{2D^2}{\eta T^2} + \frac{2U_T^{\text{cvx}}}{\eta} + 16\eta G^2 T, \tag{16}$$

$$U_T^{\text{cvx}} = \mathcal{O}\left(\eta^2 \left[\max_{t\in[T]} \tau_t^2 \ln^2 \frac{3}{\delta} + \sum_{t=1}^{T} \sigma_{\mathtt{l}}^{\mathtt{p}} \tau_t^{2-\mathtt{p}} + \left(\sum_{t=1}^{T} \frac{\sigma_{\mathtt{s}} \sigma_{\mathtt{l}}^{\mathtt{p}-1}}{\tau_t^{\mathtt{p}-1}} + \frac{\sigma_{\mathtt{l}}^{\mathtt{p}} G}{\tau_t^{\mathtt{p}}}\right)^2\right]\right).$$

Furthermore, when $\tau_t = \max\{2G, \tau\}, \forall t \in [T]$ for $\tau \geq 0$, U_T^{cvx} can be further reduced into

$$\begin{aligned} U_T^{\text{cvx}} &= \mathcal{O}\left(\eta^2 \left[\left(\frac{\sigma_{\mathtt{l}}^{2\mathtt{p}} T^2}{\max\{2G, \tau\}^{2\mathtt{p}}} + \ln^2 \frac{3}{\delta}\right) G^2 + \tau^2 \ln^2 \frac{3}{\delta} + \sigma_{\mathtt{l}}^{\mathtt{p}} \tau^{2-\mathtt{p}} T + \frac{\sigma_{\mathtt{s}}^2 \sigma_{\mathtt{l}}^{2\mathtt{p}-2}}{\tau^{2\mathtt{p}-2}} T^2\right]\right) \\ &= \mathcal{O}\left(\eta^2 \left[\left(\min_{\beta \in \left[0, \frac{1}{2}\right]} \frac{\sigma_{\mathtt{l}}^{2\mathtt{p}} T^2}{G^{2\beta\mathtt{p}} \tau^{2(1-\beta)\mathtt{p}}} + \ln^2 \frac{3}{\delta}\right) G^2 + \tau^2 \ln^2 \frac{3}{\delta} + \sigma_{\mathtt{l}}^{\mathtt{p}} \tau^{2-\mathtt{p}} T + \frac{\sigma_{\mathtt{s}}^2 \sigma_{\mathtt{l}}^{2\mathtt{p}-2}}{\tau^{2\mathtt{p}-2}} T^2\right]\right). \end{aligned}$$

One more step, we replace τ with $\tau T^{\frac{1}{\mathtt{p}}}$ to have

$$U_T^{\text{cvx}} = \mathcal{O}\left(\eta^2 \left[\left(\min_{\beta \in \left[0, \frac{1}{2}\right]} \frac{\sigma_{\mathtt{l}}^{2\mathtt{p}} T^{2\beta}}{G^{2\beta\mathtt{p}} \tau^{2(1-\beta)\mathtt{p}}} + \ln^2 \frac{3}{\delta}\right) G^2 + \left(\tau^2 \ln^2 \frac{3}{\delta} + \sigma_{\mathtt{l}}^{\mathtt{p}} \tau^{2-\mathtt{p}} + \frac{\sigma_{\mathtt{s}}^2 \sigma_{\mathtt{l}}^{2\mathtt{p}-2}}{\tau^{2\mathtt{p}-2}}\right) T^{\frac{2}{\mathtt{p}}}\right]\right). \tag{17}$$

Finally, we plug in $\tau = \tau_\star$ (see (8)) into (17) and use the definition of φ_{lip} (see (11)) to have

$$U_T^{\text{cvx}} = \mathcal{O}\left(\eta^2 \left[\left(\varphi_{\text{lip}}^2 + \ln^2 \frac{3}{\delta}\right) G^2 + \left(\sigma_{\mathtt{s}}^{\frac{4}{\mathtt{p}}-2} \sigma_{\mathtt{l}}^{4-\frac{4}{\mathtt{p}}} + \sigma_{\mathtt{s}}^{\frac{2}{\mathtt{p}}} \sigma_{\mathtt{l}}^{2-\frac{2}{\mathtt{p}}} \ln^{2-\frac{2}{\mathtt{p}}} \frac{3}{\delta}\right) T^{\frac{2}{\mathtt{p}}}\right]\right).$$

Combine the above result and (16) to obtain

$$\begin{aligned} &F(\boldsymbol{x}_{T+1}) - F_\star \\ =&\mathcal{O}\left(\frac{D^2}{\eta T^2} + \eta\left(\varphi_{\text{lip}}^2 + \ln^2 \frac{3}{\delta}\right) G^2 + \eta G^2 T + \eta\left(\sigma_{\mathtt{s}}^{\frac{4}{\mathtt{p}}-2} \sigma_{\mathtt{l}}^{4-\frac{4}{\mathtt{p}}} + \sigma_{\mathtt{s}}^{\frac{2}{\mathtt{p}}} \sigma_{\mathtt{l}}^{2-\frac{2}{\mathtt{p}}} \ln^{2-\frac{2}{\mathtt{p}}} \frac{3}{\delta}\right) T^{\frac{2}{\mathtt{p}}}\right). \end{aligned} \tag{18}$$

Plug in $\eta = \eta_\star$ (see (13)) to conclude that

$$F(\boldsymbol{x}_{T+1}) - F_\star = \mathcal{O}\left(\frac{\left(\varphi_{\text{lip}} + \ln \frac{3}{\delta}\right) GD}{T} + \frac{GD}{\sqrt{T}} + \frac{\left(\sigma_{\mathtt{s}}^{\frac{2}{\mathtt{p}}-1} \sigma_{\mathtt{l}}^{2-\frac{2}{\mathtt{p}}} + \sigma_{\mathtt{s}}^{\frac{1}{\mathtt{p}}} \sigma_{\mathtt{l}}^{1-\frac{1}{\mathtt{p}}} \ln^{1-\frac{1}{\mathtt{p}}} \frac{3}{\delta}\right) D}{T^{1-\frac{1}{\mathtt{p}}}}\right).$$

□

B.1.2 Smooth Objectives

Theorem 6 (Full statement of Theorem 2)**.** *Under Assumptions 1, 2 (with $\mu = 0$), 4, 5, and 6, for any $T \in \mathbb{N}$ and $\delta \in \left(0, \frac{1}{2}\right]$, setting $\eta_t = \eta_\star(T-t+1), \tau_t = \max\left\{2L\|\boldsymbol{x}_t - \boldsymbol{x}_1\| + 2G_1, \tau_\star T^{\frac{1}{\mathtt{p}}}\right\}, \forall t \in [T]$, where*

$$\eta_\star = \min\left\{\frac{D}{8\sqrt{\mathtt{C}}\left(\varphi_{\text{smo}} + \ln \frac{3}{\delta}\right)(LD + G_1)T}, \frac{D}{\left(\sigma_{\mathtt{s}}^{\frac{2}{\mathtt{p}}-1} \sigma_{\mathtt{l}}^{2-\frac{2}{\mathtt{p}}} + \sigma_{\mathtt{s}}^{\frac{1}{\mathtt{p}}} \sigma_{\mathtt{l}}^{1-\frac{1}{\mathtt{p}}} \ln^{1-\frac{1}{\mathtt{p}}} \frac{3}{\delta}\right) T^{1+\frac{1}{\mathtt{p}}}}\right\}, \tag{19}$$

$\mathtt{C} \geq 1$ is a numerical constant, and $\varphi_{\text{smo}} \leq \varphi_\star$ is a constant defined in (12), then `Clipped-SGD` *(Algorithm 1) guarantees that, with probability at least $1 - 2\delta$,*

$$F(\boldsymbol{x}_{T+1}) - F_\star = \mathcal{O}\left(\frac{\left(\varphi_{\text{smo}} + \ln \frac{3}{\delta}\right)(LD + G_1)D}{T} + \frac{\left(\sigma_{\mathtt{s}}^{\frac{2}{\mathtt{p}}-1} \sigma_{\mathtt{l}}^{2-\frac{2}{\mathtt{p}}} + \sigma_{\mathtt{s}}^{\frac{1}{\mathtt{p}}} \sigma_{\mathtt{l}}^{1-\frac{1}{\mathtt{p}}} \ln^{1-\frac{1}{\mathtt{p}}} \frac{3}{\delta}\right) D}{T^{1-\frac{1}{\mathtt{p}}}}\right).$$

Proof. Given $T \in \mathbb{N}$ and $\delta \in \left(0, \frac{1}{2}\right]$, note that $\{\eta_t\}_{t=1}^T$ and $\{\tau_t\}_{t=1}^T$ satisfy Conditions 1 and 2 in Lemma 7. Therefore, by Lemma 7, for any $\bar{D} > 0$ such that (where $\tau = \tau_\star T^{\frac{1}{\mathtt{p}}}$, $\bar{G} = L(\bar{D}+D)+G_1$, $\rho = \max\left\{2\bar{G}, \tau\right\}$, and $\mathtt{C} \geq 1$ is a numerical constant)

$$2D^2 + 4\mathtt{C}\left[\max_{t\in[T]} \eta_t^2\rho^2 \ln^2\frac{3}{\delta} + \sum_{t=1}^T \sigma_1^{\mathtt{p}}\eta_t^2\tau^{2-\mathtt{p}} + \left(\sum_{t=1}^T \frac{\sigma_{\mathtt{s}}\sigma_1^{\mathtt{p}-1}\eta_t}{\tau^{\mathtt{p}-1}} + \min_{\beta\in\left[0,\frac{1}{2}\right]} \frac{\sigma_1^{\mathtt{p}}\bar{G}^{1-\beta\mathtt{p}}\eta_t}{\tau^{(1-\beta)\mathtt{p}}}\right)^2\right] \leq \bar{D}^2, \tag{20}$$

there is

$$\Pr\left[\|\boldsymbol{x}_\star - \boldsymbol{x}_t\| \leq \bar{D}, \forall t \in [T+1]\right] \geq 1-\delta.$$

Now, we claim that

$$\bar{D} \triangleq \sqrt{5+32\mathtt{C}}D = \Theta(D) \tag{21}$$

satisfies (20) under the current choices of $\{\eta_t\}_{t=1}^T$ and $\{\tau_t\}_{t=1}^T$. To prove the claim, we first upper bound

$$\begin{aligned}
&\max_{t\in[T]} \eta_t^2\rho^2 \ln^2\frac{3}{\delta} + \sum_{t=1}^T \sigma_1^{\mathtt{p}}\eta_t^2\tau^{2-\mathtt{p}} + \left(\sum_{t=1}^T \frac{\sigma_{\mathtt{s}}\sigma_1^{\mathtt{p}-1}\eta_t}{\tau^{\mathtt{p}-1}} + \min_{\beta\in\left[0,\frac{1}{2}\right]} \frac{\sigma_1^{\mathtt{p}}\bar{G}^{1-\beta\mathtt{p}}\eta_t}{\tau^{(1-\beta)\mathtt{p}}}\right)^2 \\
\overset{(a)}{\leq}& \eta_\star^2 T^2\left(\rho^2\ln^2\frac{3}{\delta} + \sigma_1^{\mathtt{p}}\tau^{2-\mathtt{p}}T + 2\left(\frac{\sigma_{\mathtt{s}}^2\sigma_1^{2\mathtt{p}-2}}{\tau^{2\mathtt{p}-2}}T^2 + \min_{\beta\in\left[0,\frac{1}{2}\right]} \frac{\sigma_1^{2\mathtt{p}}T^2}{\bar{G}^{2\beta\mathtt{p}}\tau^{2(1-\beta)\mathtt{p}}}\bar{G}^2\right)\right) \\
\overset{(b)}{\leq}& \eta_\star^2 T^2\left(\left(\min_{\beta\in\left[0,\frac{1}{2}\right]} \frac{2\sigma_1^{2\mathtt{p}}T^2}{\bar{G}^{2\beta\mathtt{p}}\tau^{2(1-\beta)\mathtt{p}}} + 4\ln^2\frac{3}{\delta}\right)\bar{G}^2 + \tau^2\ln^2\frac{3}{\delta} + \sigma_1^{\mathtt{p}}\tau^{2-\mathtt{p}}T + 2\frac{\sigma_{\mathtt{s}}^2\sigma_1^{2\mathtt{p}-2}}{\tau^{2\mathtt{p}-2}}T^2\right) \\
\overset{(c)}{=}& \eta_\star^2 T^2\left(\left(\min_{\beta\in\left[0,\frac{1}{2}\right]} \frac{2\sigma_1^{2\mathtt{p}}T^{2\beta}}{\bar{G}^{2\beta\mathtt{p}}\tau_\star^{2(1-\beta)\mathtt{p}}} + 4\ln^2\frac{3}{\delta}\right)\bar{G}^2 + \left(\tau_\star^2\ln^2\frac{3}{\delta} + \sigma_1^{\mathtt{p}}\tau_\star^{2-\mathtt{p}} + 2\frac{\sigma_{\mathtt{s}}^2\sigma_1^{2\mathtt{p}-2}}{\tau_\star^{2\mathtt{p}-2}}\right)T^{\frac{2}{\mathtt{p}}}\right) \\
\overset{(d)}{\leq}& \eta_\star^2 T^2\left(\left(2\varphi_{\mathrm{smo}}^2 + 4\ln^2\frac{3}{\delta}\right)\left(L(\bar{D}+D)+G_1\right)^2 + \left(\tau_\star^2\ln^2\frac{3}{\delta} + \sigma_1^{\mathtt{p}}\tau_\star^{2-\mathtt{p}} + 2\frac{\sigma_{\mathtt{s}}^2\sigma_1^{2\mathtt{p}-2}}{\tau_\star^{2\mathtt{p}-2}}\right)T^{\frac{2}{\mathtt{p}}}\right) \\
\overset{(e)}{\leq}& \frac{\bar{D}^2+D^2}{8\mathtt{C}} + 4D^2,
\end{aligned} \tag{22}$$

where (a) is by $\eta_t = \eta_\star(T-t+1) \leq \eta_\star T, \forall t \in [T]$ and $(x+y)^2 \leq 2x^2+2y^2$, (b) is due to $\rho = \max\left\{2\bar{G}, \tau\right\} \Rightarrow \rho^2 \leq 4\bar{G}^2+\tau^2$, (c) holds after plugging in $\tau = \tau_\star T^{\frac{1}{\mathtt{p}}}$, (d) is due to $\bar{G} = L(\bar{D}+D)+G_1 \overset{(21)}{=} \Theta(LD+G_1)$ and the definition of φ_{smo} (see (12)), and (e) is true by noticing that

$$\begin{aligned}
\eta_\star^2 T^2 \cdot \left(2\varphi_{\mathrm{smo}}^2 + 4\ln^2\frac{3}{\delta}\right)\left(L(\bar{D}+D)+G_1\right)^2 &\overset{(19)}{\leq} \frac{\bar{D}^2+D^2}{8\mathtt{C}}, \\
\eta_\star^2 T^2 \cdot \left(\tau_\star^2\ln^2\frac{3}{\delta} + \sigma_1^{\mathtt{p}}\tau_\star^{2-\mathtt{p}} + 2\frac{\sigma_{\mathtt{s}}^2\sigma_1^{2\mathtt{p}-2}}{\tau_\star^{2\mathtt{p}-2}}\right)T^{\frac{2}{\mathtt{p}}} &\overset{(8),(19)}{\leq} 4D^2.
\end{aligned}$$

Combining (20) and (22), we only need to show that $\bar{D} = \sqrt{5+32\mathtt{C}}D$ satisfies the inequality $2D^2 + 4\mathtt{C}\left(\frac{\bar{D}^2+D^2}{8\mathtt{C}} + 4D^2\right) \leq \bar{D}^2$, which holds after a direct verification. Hence, we obtain that

$$\Pr\left[\|\boldsymbol{x}_\star - \boldsymbol{x}_t\| \leq \bar{D}, \forall t \in [T+1]\right] \geq 1-\delta. \tag{23}$$

Next, Lemma 4 ($\omega = 0$ now) yields that

$$F(\boldsymbol{x}_{T+1}) - F_\star \leq \frac{v_0^2D^2 + 2P_T^{\mathrm{cvx}}}{\eta_T}, \tag{24}$$

where

$$P_T^{\mathrm{cvx}} = 8\max_{t\in[T]}\left(\sum_{s=1}^t M_s\right)^2 + 2\sum_{t=1}^T v_t^2\eta_t^2\left\|\boldsymbol{d}_t^{\mathrm{u}}\right\|^2 + 4\left(\sum_{t=1}^T v_t\eta_t\left\|\boldsymbol{d}_t^{\mathrm{b}}\right\|\right)^2,$$

and $\left\{v_t^2\right\}_{t=0}^T$ is a sequence defined in (43) as

$$v_t^2 \triangleq \frac{\eta_T}{\sum_{s=t}^T \eta_s}, \forall t \in [T] \quad \text{and} \quad v_0^2 \triangleq v_1^2.$$

Observe that

$$\Pr\left[P_T^{\text{cvx}} = \bar{P}_T^{\text{cvx}}\right] \geq \Pr\left[\zeta_t = 1, \forall t \in [T+1]\right] \overset{(23)}{\geq} 1 - \delta,$$

in which

$$\bar{P}_T^{\text{cvx}} = 8 \max_{t \in [T]} \left(\sum_{s=1}^t M_s \zeta_s\right)^2 + 2\sum_{t=1}^T v_t^2 \eta_t^2 \left\|\boldsymbol{d}_t^{\text{u}}\right\|^2 \zeta_t^2 + 4\left(\sum_{t=1}^T v_t \eta_t \left\|\boldsymbol{d}_t^{\text{b}}\right\| \zeta_t\right)^2,$$
$$\zeta_t = \mathbb{1}\left[\left\|\boldsymbol{x}_\star - \boldsymbol{x}_t\right\| \leq \bar{D}\right], \forall t \in [T],$$

are introduced in Lemma 6. Thus, with probability at least $1 - \delta$,

$$F(\boldsymbol{x}_{T+1}) - F_\star \overset{(24)}{\leq} \frac{v_0^2 D^2 + 2P_T^{\text{cvx}}}{\eta_T} = \frac{v_0^2 D^2 + 2\bar{P}_T^{\text{cvx}}}{\eta_T}. \tag{25}$$

Note that $\{\eta_t\}_{t=1}^T$ and $\{\tau_t\}_{t=1}^T$ also satisfy Conditions 1 and 2 in Lemma 6, which implies that $\Pr\left[\bar{P}_T^{\text{cvx}} \leq \bar{U}_T^{\text{cvx}}\right] \geq 1 - \delta$, where $\bar{U}_T^{\text{cvx}}$ is a constant satisfying

$$\bar{U}_T^{\text{cvx}} = \mathcal{O}\left(\max_{t \in [T]} v_t^2 \eta_t^2 \rho^2 \ln^2 \frac{3}{\delta} + \sum_{t=1}^T \sigma_1^{\mathtt{p}} v_t^2 \eta_t^2 \tau^{2-\mathtt{p}} + \left(\sum_{t=1}^T \frac{\sigma_{\mathtt{s}} \sigma_1^{\mathtt{p}-1} v_t \eta_t}{\tau^{\mathtt{p}-1}} + \min_{\beta \in \left[0, \frac{1}{2}\right]} \frac{\sigma_1^{\mathtt{p}} \bar{G}^{1-\beta_{\mathtt{p}}} v_t \eta_t}{\tau^{(1-\beta)\mathtt{p}}}\right)^2\right), \tag{26}$$

in which $\tau = \tau_\star T^{\frac{1}{\mathtt{p}}}$, $\bar{G} = L(\bar{D} + D) + G_1 \overset{(21)}{=} \Theta(LD + G_1)$, and $\rho = \max\left\{2\bar{G}, \tau\right\}$. As such, we find that, with probability at least $1 - 2\delta$,

$$F(\boldsymbol{x}_{T+1}) - F_\star \leq \frac{v_0^2 D^2 + 2\bar{U}_T^{\text{cvx}}}{\eta_T}. \tag{27}$$

Finally, by following essentially the same steps in proving Theorem 5, we conclude that with probability at least $1 - 2\delta$,

$$F(\boldsymbol{x}_{T+1}) - F_\star = \mathcal{O}\left(\frac{\left(\varphi_{\text{smo}} + \ln \frac{3}{\delta}\right)(LD + G_1) D}{T} + \frac{\left(\sigma_{\mathtt{s}}^{\frac{2}{\mathtt{p}}-1} \sigma_1^{2-\frac{2}{\mathtt{p}}} + \sigma_{\mathtt{s}}^{\frac{1}{\mathtt{p}}} \sigma_1^{1-\frac{1}{\mathtt{p}}} \ln^{1-\frac{1}{\mathtt{p}}} \frac{3}{\delta}\right) D}{T^{1-\frac{1}{\mathtt{p}}}}\right).$$

□

B.2 Strongly Star-Convex Optimization

We turn our attention to strongly star-convex optimization, i.e., $\mu > 0$ in Assumption 2.

B.2.1 Lipschitz Objectives

Theorem 7 (Restatement of Theorem 3)**.** *Under Assumptions 1, 2 (with $\mu > 0$), 3, and 5, for any $T \geq 2$ and $\delta \in \left(0, \frac{1}{2}\right]$, setting $\eta_t = \begin{cases} \frac{6}{\mu t} & t \leq S \\ \frac{T-t+1}{\mu T^2} & t > S \end{cases}$, $\tau_t = \begin{cases} \max\left\{2G, \tau_\star t^{\frac{1}{\mathtt{p}}}\right\} & t \leq S \\ \max\left\{2G, \tau_\star \left\lfloor \frac{T}{2} \right\rfloor^{\frac{1}{\mathtt{p}}}\right\} & t > S \end{cases}$, $\forall t \in [T]$, where $S \triangleq \left\lceil \frac{T}{2} \right\rceil$, then* `Clipped-SGD` *(Algorithm 1) guarantees that, with probability at least $1 - 2\delta$,*

$$F(\boldsymbol{x}_{T+1}) - F_\star = \mathcal{O}\left(\frac{\mu D^2}{T^3} + \frac{\left(\varphi_{\text{lip}}^2 + \ln^2 \frac{3}{\delta}\right) G^2}{\mu T^2} + \frac{G^2}{\mu T} + \frac{\sigma_{\mathtt{s}}^{\frac{4}{\mathtt{p}}-2} \sigma_1^{4-\frac{4}{\mathtt{p}}} + \sigma_{\mathtt{s}}^{\frac{2}{\mathtt{p}}} \sigma_1^{2-\frac{2}{\mathtt{p}}} \ln^{2-\frac{2}{\mathtt{p}}} \frac{3}{\delta}}{\mu T^{2-\frac{2}{\mathtt{p}}}}\right),$$

in which $\varphi_{\text{lip}} \leq \varphi_\star$ is a constant defined in (11).

Proof. Given $T \in \mathbb{N}$ and $\delta \in \left(0, \frac{1}{2}\right]$, note that $\{\eta_t\}_{t=1}^S$ satisfies Lemma 8 with $\eta = 6$ under the time horizon S, and $\{\eta_t\}_{t=1}^S$ and $\{\tau_t\}_{t=1}^S$ satisfy Conditions 1 and 2 in Lemma 9. Therefore, Lemmas 8 ($\eta = 6$ and $\omega = 1$ now) and 9 ($\eta = 6$ now) together yield that, with probability at least $1 - \delta$,

$$\|\boldsymbol{x}_\star - \boldsymbol{x}_{S+1}\|^2 \leq \frac{8D^2 + 4A_S^{\mathrm{str}} + 16\sum_{t=1}^S \gamma_t \eta_t^2 G^2}{\gamma_S}, \tag{28}$$

where A_S^{str} is a constant satisfying

$$\begin{aligned} A_S^{\mathrm{str}} =& \mathcal{O}\left(\max_{t\in[S]} \gamma_t \eta_t^2 \tau_t^2 \ln^2 \frac{3}{\delta} + \sum_{t=1}^S \sigma_1^{\mathtt{p}} \gamma_t \eta_t^2 \tau_t^{2-\mathtt{p}} + \sum_{t=1}^S \left(\frac{\sigma_{\mathtt{s}}^{\mathtt{p}} \gamma_t \eta_t^2}{\tau_t^{\mathtt{p}-2}} + \frac{\sigma_1^{\mathtt{p}} G^2 \gamma_t \eta_t^2}{\tau_t^{\mathtt{p}}}\right) \ln \frac{3}{\delta} \right. \\ &\left. + \sum_{t=1}^S \left(\frac{\sigma_{\mathtt{s}}^2 \sigma_1^{2\mathtt{p}-2} \gamma_t \eta_t}{\tau_t^{2\mathtt{p}-2}} + \frac{\sigma_1^{2\mathtt{p}} G^2 \gamma_t \eta_t}{\tau_t^{2\mathtt{p}}}\right) \frac{1}{\mu}\right), \end{aligned} \tag{29}$$

and $\{\gamma_t\}_{t=1}^S$ is a sequence defined in (75).

Next, we compute

$$\gamma_t \overset{(75)}{=} \prod_{s=2}^t \frac{1 + \mu\eta_{s-1}}{1 + \mu\eta_s/2} \overset{\eta_s = \frac{6}{\mu s}}{=} \prod_{s=2}^t \frac{s(s+5)}{(s-1)(s+3)} = \frac{t(t+4)(t+5)}{30}, \forall t \in [S].$$

By the above equation and (28), we have

$$\|\boldsymbol{x}_\star - \boldsymbol{x}_{S+1}\|^2 = \mathcal{O}\left(\frac{D^2 + A_S^{\mathrm{str}} + \frac{G^2}{\mu^2}\sum_{t=1}^S t}{S^3}\right) \overset{S = \lceil \frac{T}{2} \rceil}{=} \mathcal{O}\left(\frac{D^2 + A_S^{\mathrm{str}}}{T^3} + \frac{G^2}{\mu^2 T}\right). \tag{30}$$

By similar steps in proving Theorem 14 of [23], we can show

$$\begin{aligned} A_S^{\mathrm{str}} &= \mathcal{O}\left(\frac{\left(\min_{\beta\in\left[0,\frac{1}{2}\right]} \frac{\sigma_1^{2\mathtt{p}} S^{2\beta}}{G^{2\beta\mathtt{p}} \tau_\star^{2(1-\beta)\mathtt{p}}} + \ln^2 \frac{3}{\delta}\right) G^2}{\mu^2} S + \frac{\tau_\star^2 \ln^2 \frac{3}{\delta} + \sigma_1^{\mathtt{p}} \tau_\star^{2-\mathtt{p}} + \frac{\sigma_{\mathtt{s}}^2 \sigma_1^{2\mathtt{p}-2}}{\tau_\star^{2\mathtt{p}-2}}}{\mu^2} S^{1+\frac{2}{\mathtt{p}}}\right) \\ &\overset{S=\lceil \frac{T}{2} \rceil, (8), (11)}{=} \mathcal{O}\left(\frac{\left(\varphi_{\mathrm{lip}}^2 + \ln^2 \frac{3}{\delta}\right) G^2}{\mu^2} T + \frac{\sigma_{\mathtt{s}}^{\frac{4}{\mathtt{p}}-2} \sigma_1^{4-\frac{4}{\mathtt{p}}} + \sigma_{\mathtt{s}}^{\frac{2}{\mathtt{p}}} \sigma_1^{2-\frac{2}{\mathtt{p}}} \ln^{2-\frac{2}{\mathtt{p}}} \frac{3}{\delta}}{\mu^2} T^{1+\frac{2}{\mathtt{p}}}\right). \end{aligned} \tag{31}$$

Hence, we obtain that, with probability at least $1 - \delta$,

$$\|\boldsymbol{x}_\star - \boldsymbol{x}_{S+1}\|^2 \overset{(30),(31)}{=} \mathcal{O}\left(\frac{D^2}{T^3} + \frac{\left(\varphi_{\mathrm{lip}}^2 + \ln^2 \frac{3}{\delta}\right) G^2}{\mu^2 T^2} + \frac{G^2}{\mu^2 T} + \frac{\sigma_{\mathtt{s}}^{\frac{4}{\mathtt{p}}-2} \sigma_1^{4-\frac{4}{\mathtt{p}}} + \sigma_{\mathtt{s}}^{\frac{2}{\mathtt{p}}} \sigma_1^{2-\frac{2}{\mathtt{p}}} \ln^{2-\frac{2}{\mathtt{p}}} \frac{3}{\delta}}{\mu^2 T^{2-\frac{2}{\mathtt{p}}}}\right). \tag{32}$$

Finally, when $T \geq 2$, we observe that $\boldsymbol{x}_{T+1}$ can be viewed as the output of Algorithm 1 after $T - S = \left\lfloor \frac{T}{2} \right\rfloor$ iterations with the initial point $\boldsymbol{x}_{S+1}$. Therefore, let $\eta_\star \triangleq \frac{1}{\mu T^2}$, by our choices of $\{\eta_t\}_{t=S+1}^T$ and $\{\tau_t\}_{t=S+1}^T$ and the proof of Theorem 5, we have, with probability at least $1 - \delta$,

$$\begin{aligned} F(\boldsymbol{x}_{T+1}) - F_\star \overset{(18)}{=} \mathcal{O}\left(\right. & \frac{\|\boldsymbol{x}_\star - \boldsymbol{x}_{S+1}\|^2}{\eta_\star \left\lfloor \frac{T}{2} \right\rfloor^2} + \eta_\star \left(\varphi_{\mathrm{lip}}^2 + \ln^2 \frac{3}{\delta}\right) G^2 + \eta_\star G^2 \left\lfloor \frac{T}{2} \right\rfloor \\ & + \eta_\star \left(\sigma_{\mathtt{s}}^{\frac{4}{\mathtt{p}}-2} \sigma_1^{4-\frac{4}{\mathtt{p}}} + \sigma_{\mathtt{s}}^{\frac{2}{\mathtt{p}}} \sigma_1^{2-\frac{2}{\mathtt{p}}} \ln^{2-\frac{2}{\mathtt{p}}} \frac{3}{\delta}\right) \left\lfloor \frac{T}{2} \right\rfloor^{\frac{2}{\mathtt{p}}} \left.\right) \\ \overset{\eta_\star = \frac{1}{\mu T^2}}{=} \mathcal{O}\left(\right. & \mu \|\boldsymbol{x}_\star - \boldsymbol{x}_{S+1}\|^2 + \frac{\left(\varphi_{\mathrm{lip}}^2 + \ln^2 \frac{3}{\delta}\right) G^2}{\mu T^2} + \frac{G^2}{\mu T} \\ & + \frac{\sigma_{\mathtt{s}}^{\frac{4}{\mathtt{p}}-2} \sigma_1^{4-\frac{4}{\mathtt{p}}} + \sigma_{\mathtt{s}}^{\frac{2}{\mathtt{p}}} \sigma_1^{2-\frac{2}{\mathtt{p}}} \ln^{2-\frac{2}{\mathtt{p}}} \frac{3}{\delta}}{\mu T^{2-\frac{2}{\mathtt{p}}}} \left.\right). \end{aligned} \tag{33}$$

By combining (32) and (33), we obtain the desired result. □

B.2.2 Smooth Objectives

Theorem 8 (Restatement of Theorem 4)**.** *Under Assumptions 1, 2 (with $\mu > 0$), 4, 5, and 7, for any $T \geq 2$ and $\delta \in \left(0, \frac{1}{2}\right]$, setting $\eta_t = \begin{cases} \frac{6}{\mu(t+12\kappa)} & t \leq S \\ \frac{T-t+1}{\mu T(T+\kappa)} & t > S \end{cases}$, $\tau_t = \begin{cases} \max\left\{2L\left\|\boldsymbol{x}_t - \boldsymbol{x}_1\right\| + 2G_1, \tau_\star \left[T(T+\kappa)\right]^{\frac{1}{2\mathtt{p}}}\right\} & t \leq S \\ \max\left\{2(1+2\kappa)G_1, \tau_\star \left\lfloor \frac{T}{2} \right\rfloor^{\frac{1}{\mathtt{p}}}\right\} & t > S \end{cases}$, $\forall t \in [T]$, where $S \triangleq \left\lceil \frac{T}{2} \right\rceil$ and $\kappa \triangleq \frac{L}{\mu}$, then* `Clipped-SGD` *(Algorithm 1) guarantees that, with probability at least $1 - 2\delta$,*

$$F(\boldsymbol{x}_{T+1}) - F_\star = \mathcal{O}\left(\frac{\left(\varphi_{\mathrm{smo}}^2 + \ln^2\frac{3}{\delta}\right)(1+\kappa)^2 G_1^2}{\mu T^2} + \frac{\sigma_{\mathtt{s}}^{\frac{4}{\mathtt{p}}-2}\sigma_1^{4-\frac{4}{\mathtt{p}}} + \sigma_{\mathtt{s}}^{\frac{2}{\mathtt{p}}}\sigma_1^{2-\frac{2}{\mathtt{p}}}\ln^{2-\frac{2}{\mathtt{p}}}\frac{3}{\delta}}{\mu\left[T(T+\kappa)\right]^{1-\frac{1}{\mathtt{p}}}}\right),$$

in which $\varphi_{\mathrm{smo}} \leq \varphi_\star$ is a constant defined in (12).

Proof. Given $T \in \mathbb{N}$ and $\delta \in \left(0, \frac{1}{2}\right]$, note that $\{\eta_t\}_{t=1}^S$ satisfies Lemma 8 with $\eta = 6$ under the time horizon S, $\{\eta_t\}_{t=1}^S$ and $\{\tau_t\}_{t=1}^S$ satisfy Conditions 1 and 2 in Lemma 10, and I_S^{str} in Lemma 8 equals $\bar{I}_S^{\mathrm{str}}$ in Lemma 10 under $\bar{D} \triangleq \frac{2G_1}{\mu}$ due to Assumption 7. Therefore, Lemmas 8 ($\eta = 6$ and $\omega = 0$ now) and 10 ($\eta = 6$ and $\bar{D} = \frac{2G_1}{\mu}$ now) together yield that, with probability at least $1 - \delta$,

$$\left\|\boldsymbol{x}_\star - \boldsymbol{x}_{S+1}\right\|^2 \leq \frac{8D^2 + 4\bar{A}_S^{\mathrm{str}}}{\gamma_S}, \tag{34}$$

where $\bar{A}_S^{\mathrm{str}}$ is a constant satisfying

$$\begin{aligned}\bar{A}_S^{\mathrm{str}} =& \mathcal{O}\left(\max_{t\in[S]}\gamma_t\eta_t^2\rho^2\ln^2\frac{3}{\delta} + \sum_{t=1}^S \sigma_1^{\mathtt{p}}\gamma_t\eta_t^2\tau^{2-\mathtt{p}} + \sum_{t=1}^S\left(\frac{\sigma_{\mathtt{s}}^{\mathtt{p}}\gamma_t\eta_t^2\rho}{\tau^{\mathtt{p}-1}} + \min_{\beta\in\left[0,\frac{1}{2}\right]}\frac{\sigma_1^{\mathtt{p}}\bar{G}^{2-\beta\mathtt{p}}\gamma_t\eta_t^2}{\tau^{(1-\beta)\mathtt{p}}}\right)\ln\frac{3}{\delta}\right.\\ &\left. + \left(\sum_{t=1}^S\frac{\sigma_{\mathtt{s}}^2\sigma_1^{2\mathtt{p}-2}\gamma_t\eta_t}{\tau^{2\mathtt{p}-2}} + \min_{\beta\in\left[0,\frac{1}{2}\right]}\frac{\sigma_1^{2\mathtt{p}}\bar{G}^{2-2\beta\mathtt{p}}\gamma_t\eta_t}{\tau^{2(1-\beta)\mathtt{p}}}\right)\frac{1}{\mu}\right),\end{aligned} \tag{35}$$

$\tau = \tau_\star\left[T(T+\kappa)\right]^{\frac{1}{2\mathtt{p}}}$, $\bar{G} = L(\bar{D}+D) + G_1 \overset{0\leq D\leq\bar{D}=2G_1/\mu}{=} \Theta((1+\kappa)G_1)$, $\rho = \max\left\{2\bar{G}, \tau\right\}$, and $\{\gamma_t\}_{t=1}^S$ is a sequence defined in (75).

Next, we compute

$$\begin{aligned}\gamma_t &\overset{(75)}{=} \prod_{s=2}^t \frac{1+\mu\eta_{s-1}}{1+\mu\eta_s/2} \overset{\eta_s=\frac{6}{\mu(s+12\kappa)}}{=} \prod_{s=2}^t \frac{(s+12\kappa)(s+5+12\kappa)}{(s-1+12\kappa)(s+3+12\kappa)}\\ &= \frac{(t+12\kappa)(t+4+12\kappa)(t+5+12\kappa)}{(1+12\kappa)(5+12\kappa)(6+12\kappa)}, \forall t \in [S].\end{aligned}$$

By the above equation and (34), we have

$$\left\|\boldsymbol{x}_\star - \boldsymbol{x}_{S+1}\right\|^2 = \mathcal{O}\left(\frac{(1+\kappa)^3\left(D^2 + \bar{A}_S^{\mathrm{str}}\right)}{(S+\kappa)^3}\right) \overset{S=\left\lceil\frac{T}{2}\right\rceil}{=} \mathcal{O}\left(\frac{(1+\kappa)^3\left(D^2 + \bar{A}_S^{\mathrm{str}}\right)}{(T+\kappa)^3}\right). \tag{36}$$

Noticing that $\gamma_t\eta_t = \Theta\left(\frac{(t+\kappa)^2}{\mu(1+\kappa)^3}\right)$, $\gamma_t\eta_t^2 = \Theta\left(\frac{t+\kappa}{\mu^2(1+\kappa)^3}\right)$, and $S = \Theta(T)$, we can write (35) into

$$\begin{aligned}
\bar{A}_S^{\mathrm{str}} =&\mathcal{O}\left(\frac{\rho^2(T+\kappa)\ln^2\frac{3}{\delta}}{\mu^2(1+\kappa)^3} + \frac{\sigma_1^{\mathtt{p}}\tau^{2-\mathtt{p}}(T+\kappa)T}{\mu^2(1+\kappa)^3}\right.\\
&+\left(\frac{\sigma_{\mathtt{s}}^{\mathtt{p}}\rho(T+\kappa)T}{\tau^{\mathtt{p}-1}\mu^2(1+\kappa)^3} + \min_{\beta\in\left[0,\frac{1}{2}\right]}\frac{\sigma_1^{\mathtt{p}}\bar{G}^{2-\beta\mathtt{p}}(T+\kappa)T}{\tau^{(1-\beta)\mathtt{p}}\mu^2(1+\kappa)^3}\right)\ln\frac{3}{\delta}\\
&\left.+\frac{\sigma_{\mathtt{s}}^2\sigma_1^{2\mathtt{p}-2}(T+\kappa)^2T}{\tau^{2\mathtt{p}-2}\mu^2(1+\kappa)^3} + \min_{\beta\in\left[0,\frac{1}{2}\right]}\frac{\sigma_1^{2\mathtt{p}}\bar{G}^{2-2\beta\mathtt{p}}(T+\kappa)^2T}{\tau^{2(1-\beta)\mathtt{p}}\mu^2(1+\kappa)^3}\right),\\
=&\mathcal{O}\left(\frac{\rho^2(T+\kappa)\ln^2\frac{3}{\delta}}{\mu^2(1+\kappa)^3} + \frac{\sigma_1^{\mathtt{p}}\tau^{2-\mathtt{p}}(T+\kappa)T}{\mu^2(1+\kappa)^3}\right.\\
&\left.+\frac{\sigma_{\mathtt{s}}^2\sigma_1^{2\mathtt{p}-2}(T+\kappa)^2T}{\tau^{2\mathtt{p}-2}\mu^2(1+\kappa)^3} + \min_{\beta\in\left[0,\frac{1}{2}\right]}\frac{\sigma_1^{2\mathtt{p}}\bar{G}^{2-2\beta\mathtt{p}}(T+\kappa)^2T}{\tau^{2(1-\beta)\mathtt{p}}\mu^2(1+\kappa)^3}\right),
\end{aligned}\tag{37}$$

where the second equation is due to

$$\begin{aligned}
&\frac{\rho^2(T+\kappa)\ln^2\frac{3}{\delta}}{\mu^2(1+\kappa)^3} + \frac{\sigma_{\mathtt{s}}^2\sigma_1^{2\mathtt{p}-2}(T+\kappa)^2T}{\tau^{2\mathtt{p}-2}\mu^2(1+\kappa)^3} + \min_{\beta\in\left[0,\frac{1}{2}\right]}\frac{\sigma_1^{2\mathtt{p}}\bar{G}^{2-2\beta\mathtt{p}}(T+\kappa)^2T}{\tau^{2(1-\beta)\mathtt{p}}\mu^2(1+\kappa)^3}\\
\overset{(a)}{=}&\Omega\left(\frac{\sigma_{\mathtt{s}}\sigma_1^{\mathtt{p}-1}\rho(T+\kappa)^{\frac{3}{2}}T^{\frac{1}{2}}\ln\frac{3}{\delta}}{\tau^{\mathtt{p}-1}\mu^2(1+\kappa)^3} + \min_{\beta\in\left[0,\frac{1}{2}\right]}\frac{\sigma_1^{\mathtt{p}}\bar{G}^{1-\beta\mathtt{p}}\rho(T+\kappa)^{\frac{3}{2}}T^{\frac{1}{2}}\ln\frac{3}{\delta}}{\tau^{(1-\beta)\mathtt{p}}\mu^2(1+\kappa)^3}\right)\\
\overset{(b)}{=}&\Omega\left(\frac{\sigma_{\mathtt{s}}^{\mathtt{p}}\rho(T+\kappa)T\ln\frac{3}{\delta}}{\tau^{\mathtt{p}-1}\mu^2(1+\kappa)^3} + \min_{\beta\in\left[0,\frac{1}{2}\right]}\frac{\sigma_1^{\mathtt{p}}\bar{G}^{2-\beta\mathtt{p}}(T+\kappa)T\ln\frac{3}{\delta}}{\tau^{(1-\beta)\mathtt{p}}\mu^2(1+\kappa)^3}\right),
\end{aligned}$$

in which (a) is by AM-GM inequality and (b) is due to $\sigma_1 \geq \sigma_{\mathtt{s}}$, $\mathtt{p} \geq 1$, $\rho = \max\left\{2\bar{G}, \tau\right\} \geq \bar{G}$, and $T + \kappa \geq T$. Now, we use $\rho^2 \leq 4\bar{G}^2 + \tau^2$ and $\tau = \tau_\star\left[T(T+\kappa)\right]^{\frac{1}{2\mathtt{p}}}$ to simplify (37) into

$$\begin{aligned}
\bar{A}_S^{\mathrm{str}} =& \mathcal{O}\left(\frac{\left(\min_{\beta\in\left[0,\frac{1}{2}\right]}\frac{\sigma_1^{2\mathtt{p}}[T(T+\kappa)]^{\beta}}{\bar{G}^{2\beta\mathtt{p}}\tau_\star^{2(1-\beta)\mathtt{p}}} + \ln^2\frac{3}{\delta}\right)\bar{G}^2}{\mu^2(1+\kappa)^3}(T+\kappa)\right.\\
&\left.+\frac{\tau_\star^2\ln^2\frac{3}{\delta} + \sigma_1^{\mathtt{p}}\tau_\star^{2-\mathtt{p}} + \frac{\sigma_{\mathtt{s}}^2\sigma_1^{2\mathtt{p}-2}}{\tau_\star^{2\mathtt{p}-2}}}{\mu^2(1+\kappa)^3}T^{\frac{1}{\mathtt{p}}}(T+\kappa)^{1+\frac{1}{\mathtt{p}}}\right)\\
\overset{(8),(12)}{=}&\mathcal{O}\left(\frac{\left(\varphi_{\mathrm{smo}}^2\sqrt{1+\frac{\kappa}{T}} + \ln^2\frac{3}{\delta}\right)\bar{G}^2}{\mu^2(1+\kappa)^3}(T+\kappa)\right.\\
&\left.+\frac{\sigma_{\mathtt{s}}^{\frac{4}{\mathtt{p}}-2}\sigma_1^{4-\frac{4}{\mathtt{p}}} + \sigma_{\mathtt{s}}^{\frac{2}{\mathtt{p}}}\sigma_1^{2-\frac{2}{\mathtt{p}}}\ln^{2-\frac{2}{\mathtt{p}}}\frac{3}{\delta}}{\mu^2(1+\kappa)^3}T^{\frac{1}{\mathtt{p}}}(T+\kappa)^{1+\frac{1}{\mathtt{p}}}\right).
\end{aligned}\tag{38}$$

Hence, we obtain that, with probability at least $1-\delta$,

$$\begin{aligned}
&\left\|\boldsymbol{x}_\star - \boldsymbol{x}_{S+1}\right\|^2\\
\overset{(36),(38)}{=}&\mathcal{O}\left(\frac{(1+\kappa)^3D^2}{(T+\kappa)^3} + \frac{\left(\varphi_{\mathrm{smo}}^2\sqrt{1+\frac{\kappa}{T}} + \ln^2\frac{3}{\delta}\right)\bar{G}^2}{\mu^2(T+\kappa)^2} + \frac{\sigma_{\mathtt{s}}^{\frac{4}{\mathtt{p}}-2}\sigma_1^{4-\frac{4}{\mathtt{p}}} + \sigma_{\mathtt{s}}^{\frac{2}{\mathtt{p}}}\sigma_1^{2-\frac{2}{\mathtt{p}}}\ln^{2-\frac{2}{\mathtt{p}}}\frac{3}{\delta}}{\mu^2(T+\kappa)^{2-\frac{1}{\mathtt{p}}}T^{-\frac{1}{\mathtt{p}}}}\right)\\
=&\mathcal{O}\left(\frac{\left(\varphi_{\mathrm{smo}}^2\sqrt{1+\frac{\kappa}{T}} + \ln^2\frac{3}{\delta}\right)\bar{G}^2}{\mu^2(T+\kappa)^2} + \frac{\sigma_{\mathtt{s}}^{\frac{4}{\mathtt{p}}-2}\sigma_1^{4-\frac{4}{\mathtt{p}}} + \sigma_{\mathtt{s}}^{\frac{2}{\mathtt{p}}}\sigma_1^{2-\frac{2}{\mathtt{p}}}\ln^{2-\frac{2}{\mathtt{p}}}\frac{3}{\delta}}{\mu^2(T+\kappa)^{2-\frac{1}{\mathtt{p}}}T^{-\frac{1}{\mathtt{p}}}}\right),
\end{aligned}\tag{39}$$

where the second step is due to $\frac{(1+\kappa)^3D^2}{(T+\kappa)^3} \leq \frac{(1+\kappa)^2D^2}{(T+\kappa)^2} \overset{\text{Assumption 7}}{=} \mathcal{O}\left(\frac{(1+\kappa)^2G_1^2}{\mu^2(T+\kappa)^2}\right) = \mathcal{O}\left(\frac{\bar{G}^2}{\mu^2(T+\kappa)^2}\right)$.

Finally, when $T \geq 2$, we observe that $\boldsymbol{x}_{T+1}$ can be viewed as the output of Algorithm 1 after $T - S = \left\lfloor \frac{T}{2} \right\rfloor$ iterations with the initial point $\boldsymbol{x}_{S+1}$. Moreover, note that for any $t > S$,

$$\begin{aligned}\|\nabla f(\boldsymbol{x}_t)\| &\leq \|\nabla f(\boldsymbol{x}_t) - \nabla f(\boldsymbol{x}_1)\| + \|\nabla f(\boldsymbol{x}_1)\| \leq L \|\boldsymbol{x}_t - \boldsymbol{x}_1\| + \|\nabla f(\boldsymbol{x}_1)\| \\ &\overset{(c)}{\leq} 2\kappa G_1 + G_1 = (1 + 2\kappa) G_1 \leq \frac{\tau_t}{2},\end{aligned}$$

where (c) follows from Assumption 7. Therefore, let $\eta_\star \triangleq \frac{1}{\mu T(T+\kappa)}$, by our choices of $\{\eta_t\}_{t=S+1}^T$ and $\{\tau_t\}_{t=S+1}^T$ and the proof of Theorem 6 (in which we can directly set $\bar{D} = \frac{2G_1}{\mu}$ due to Assumption 7, resulting in $\zeta_t = 1$ almost surely), we have, for $\bar{G} = (1 + 2\kappa)G_1$, with probability at least $1 - \delta$,

$$\begin{aligned}F(\boldsymbol{x}_{T+1}) - F_\star &\overset{(26),(27)}{=} \mathcal{O}\Bigg(\frac{\|\boldsymbol{x}_\star - \boldsymbol{x}_{S+1}\|^2}{\eta_\star \left\lfloor \frac{T}{2} \right\rfloor^2} + \eta_\star \left(\varphi_{\mathrm{smo}}^2 + \ln^2 \frac{3}{\delta}\right) \bar{G}^2 \\ &\qquad + \eta_\star \left(\sigma_{\mathtt{s}}^{\frac{4}{\mathtt{p}}-2} \sigma_{\mathtt{l}}^{4-\frac{4}{\mathtt{p}}} + \sigma_{\mathtt{s}}^{\frac{2}{\mathtt{p}}} \sigma_{\mathtt{l}}^{2-\frac{2}{\mathtt{p}}} \ln^{2-\frac{2}{\mathtt{p}}} \frac{3}{\delta}\right) \left\lfloor \frac{T}{2} \right\rfloor^{\frac{2}{\mathtt{p}}}\Bigg) \\ &\overset{\eta_\star = \frac{1}{\mu T(T+\kappa)}}{=} \mathcal{O}\Bigg(\left(1 + \frac{\kappa}{T}\right) \mu \|\boldsymbol{x}_\star - \boldsymbol{x}_{S+1}\|^2 + \frac{\left(\varphi_{\mathrm{smo}}^2 + \ln^2 \frac{3}{\delta}\right) \bar{G}^2}{\mu T(T+\kappa)} \\ &\qquad + \frac{\sigma_{\mathtt{s}}^{\frac{4}{\mathtt{p}}-2} \sigma_{\mathtt{l}}^{4-\frac{4}{\mathtt{p}}} + \sigma_{\mathtt{s}}^{\frac{2}{\mathtt{p}}} \sigma_{\mathtt{l}}^{2-\frac{2}{\mathtt{p}}} \ln^{2-\frac{2}{\mathtt{p}}} \frac{3}{\delta}}{\mu T^{1-\frac{2}{\mathtt{p}}}(T+\kappa)}\Bigg). \end{aligned} \tag{40}$$

By combining (39) and (40), we obtain the desired result. □

C Theoretical Analysis

This section contains the missing analysis of each intermediate lemma used in Section B. Before proceeding, we recall that $\mathcal{F}_t = \sigma(Z_1, \ldots, Z_t)$ denotes the natural filtration. By the structure of Algorithm 1, we naturally have $\tau_t \in \mathcal{F}_{t-1}$ and $\eta_t \in \mathcal{F}_t$, which implies that $\boldsymbol{x}_t \in \mathcal{F}_{t-1}$ and $\nabla f(\boldsymbol{x}_t) \in \mathcal{F}_{t-1}$. For each $t \in [T]$, we define the following standard notation in the literature on `Clipped-SGD`,

$$\boldsymbol{d}_t^{\mathrm{c}} \triangleq \boldsymbol{g}_t^{\mathrm{c}} - \nabla f(\boldsymbol{x}_t), \quad \boldsymbol{d}_t^{\mathrm{u}} \triangleq \boldsymbol{g}_t^{\mathrm{c}} - \mathbb{E}\left[\boldsymbol{g}_t^{\mathrm{c}} \mid \mathcal{F}_{t-1}\right], \quad \boldsymbol{d}_t^{\mathrm{b}} \triangleq \mathbb{E}\left[\boldsymbol{g}_t^{\mathrm{c}} \mid \mathcal{F}_{t-1}\right] - \nabla f(\boldsymbol{x}_t), \tag{41}$$

where $\boldsymbol{d}_t^{\mathrm{c}} \in \mathcal{F}_t$ denotes the difference between the clipped stochastic gradient and the true gradient, $\boldsymbol{d}_t^{\mathrm{u}}$ is the unbiased part in $\boldsymbol{d}_t^{\mathrm{c}}$, and $\boldsymbol{d}_t^{\mathrm{b}}$ is the biased part in $\boldsymbol{d}_t^{\mathrm{c}}$.

C.1 General Lemmas

This subsection provides two general lemmas used in the proof.

First, Lemma 2 below gives the clipping error bounds under Assumption 5 recently proved by [23].

Lemma 2 (Lemma 2 of [23])**.** *Under Assumption 5, let $\chi_t \triangleq \mathbb{1}\left[\tau_t \geq 2\|\nabla f(\boldsymbol{x}_t)\|\right]$, then there are:*

1. $\|\boldsymbol{d}_t^{\mathrm{u}}\| \leq 2\tau_t$.
2. $\mathbb{E}\left[\|\boldsymbol{d}_t^{\mathrm{u}}\|^2 \mid \mathcal{F}_{t-1}\right] \leq 4\sigma_{\mathtt{l}}^{\mathtt{p}} \tau_t^{2-\mathtt{p}}$.
3. $\left\|\mathbb{E}\left[\boldsymbol{d}_t^{\mathrm{u}} (\boldsymbol{d}_t^{\mathrm{u}})^\top \mid \mathcal{F}_{t-1}\right]\right\|_{\mathtt{op}} \chi_t \leq 4\sigma_{\mathtt{s}}^{\mathtt{p}} \tau_t^{2-\mathtt{p}} + 2^{\mathtt{p}+1} \sigma_{\mathtt{l}}^{\mathtt{p}} \|\nabla f(\boldsymbol{x}_t)\|^{\mathtt{p}} \tau_t^{-\mathtt{p}}$.
4. $\left\|\boldsymbol{d}_t^{\mathrm{b}}\right\| \chi_t \leq \sigma_{\mathtt{s}} \sigma_{\mathtt{l}}^{\mathtt{p}-1} \tau_t^{1-\mathtt{p}} + 2^{\mathtt{p}-1} \sigma_{\mathtt{l}}^{\mathtt{p}} \|\nabla f(\boldsymbol{x}_t)\| \tau_t^{-\mathtt{p}}$.

Remark 6. The original Lemma 2 of [23] is stated more generally and includes two additional inequalities. Here, we present a simplified version containing only the ones used in this paper.

Next, we give a one-step descent inequality. Because it is standard in the literature, we omit the proof to save space.

Lemma 3. *Under Assumptions 2 and 3 (resp. 4), for any $\boldsymbol{y} \in \mathcal{X}$, $t \in \mathbb{N}$, and $\eta_t > 0$ (resp. $0 < \eta_t \leq \frac{1}{2L}$), `Clipped-SGD` (Algorithm 1) guarantees:*

$$F(\boldsymbol{x}_{t+1}) - F(\boldsymbol{y}) \leq \frac{\|\boldsymbol{y} - \boldsymbol{x}_t\|^2 - \|\boldsymbol{y} - \boldsymbol{x}_{t+1}\|^2}{2\eta_t} + \langle \boldsymbol{d}_t^{\mathrm{c}}, \boldsymbol{y} - \boldsymbol{x}_t \rangle + \eta_t \|\boldsymbol{d}_t^{\mathrm{c}}\|^2 + 4\eta_t G^2 \omega,$$

where $\omega \triangleq \begin{cases} 1 & \text{under Assumption 3} \\ 0 & \text{under Assumption 4} \end{cases}$.

C.2 Lemmas for General Convex Optimization

In this subsection, we will prove the lemmas used in general convex optimization, i.e., $\mu = 0$ in Assumption 2.

Following [26] and [42], we first introduce an auxiliary sequence $\{\boldsymbol{z}_t\}_{t=0}^T$ defined as follows

$$\boldsymbol{z}_0 \triangleq \boldsymbol{x}_\star \quad \text{and} \quad \boldsymbol{z}_t \triangleq \frac{v_{t-1}^2}{v_t^2}\boldsymbol{z}_{t-1} + \left(1 - \frac{v_{t-1}^2}{v_t^2}\right)\boldsymbol{x}_t, \forall t \in [T], \tag{42}$$

where $\boldsymbol{x}_\star$ is the optimal solution given in Assumption 1, and the sequence $\left\{v_t^2\right\}_{t=0}^T$ is defined as

$$v_t^2 \triangleq \frac{\eta_T}{\sum_{s=t}^T \eta_s}, \forall t \in [T] \quad \text{and} \quad v_0^2 \triangleq v_1^2. \tag{43}$$

Equivalently, there is

$$\boldsymbol{z}_t = \frac{v_0^2}{v_t^2}\boldsymbol{x}_\star + \sum_{s=1}^t \frac{v_s^2 - v_{s-1}^2}{v_t^2}\boldsymbol{x}_s, \forall t \in [T], \tag{44}$$

which implies that $\boldsymbol{z}_t$ is a convex combination of $\boldsymbol{x}_\star$ and $\{\boldsymbol{x}_s\}_{s=1}^t$ (since $\left\{v_t^2\right\}_{t=0}^T$ is positive and nondecreasing), and hence falls into $\mathcal{X}$.

Next, inspired by [23], we introduce another two sequences $\{R_t\}_{t=1}^T$ and $\{M_t\}_{t=1}^T$ as follows

$$R_t \triangleq \max_{s \in [t]} v_{s-1} \|\boldsymbol{z}_{s-1} - \boldsymbol{x}_s\| \quad \text{and} \quad M_t \triangleq \left\langle v_{t-1}\eta_t \boldsymbol{d}_t^{\mathrm{u}}, \frac{v_{t-1}(\boldsymbol{z}_{t-1} - \boldsymbol{x}_t)}{R_t} \right\rangle. \tag{45}$$

One more step, like [25], we define two more sequences $\{D_t\}_{t=1}^T$ and $\{N_t\}_{t=1}^T$ in the following

$$D_t \triangleq \max_{s \in [t]} \|\boldsymbol{x}_\star - \boldsymbol{x}_s\| \quad \text{and} \quad N_t \triangleq \left\langle \eta_t \boldsymbol{d}_t^{\mathrm{u}}, \frac{\boldsymbol{x}_\star - \boldsymbol{x}_t}{D_t} \right\rangle. \tag{46}$$

As one may observe, $\{D_t\}_{t=1}^T$ (resp. $\{N_t\}_{t=1}^T$) corresponds to $\{R_t\}_{t=1}^T$ (resp. $\{M_t\}_{t=1}^T$) when $v_t = 1, \forall t \in \{0\} \cup [T]$.

Now, we are ready to prove the following Lemma 4, which not only characterizes the last-iterate progress made by `Clipped-SGD` after T iterations, but also provides a bound on $\|\boldsymbol{x}_\star - \boldsymbol{x}_{t+1}\|^2$ for any $t \in [T]$.

Lemma 4. *Under Assumptions 1, 2 (with $\mu = 0$), and 3 (resp. 4), for any $T \in \mathbb{N}$ and $\{\eta_t\}_{t=1}^T$ satisfying $\eta_t > 0, \forall t \in [T]$ (resp. $0 < \eta_t \leq \frac{1}{2L}, \forall t \in [T]$), `Clipped-SGD` (Algorithm 1) guarantees*

$$F(\boldsymbol{x}_{T+1}) - F_\star \leq \frac{v_0^2 D^2 + 2P_T^{\mathrm{cvx}} + 8\sum_{t=1}^T v_t^2 \eta_t^2 G^2 \omega}{\eta_T},$$

$$\|\boldsymbol{x}_\star - \boldsymbol{x}_{t+1}\|^2 \leq 2D^2 + 4I_t^{\mathrm{cvx}} + 16\sum_{s=1}^t \eta_s^2 G^2 \omega, \forall t \in [T],$$

where

$$P_T^{\mathrm{cvx}} = 8\max_{t \in [T]} \left(\sum_{s=1}^t M_s\right)^2 + 2\sum_{t=1}^T v_t^2 \eta_t^2 \|\boldsymbol{d}_t^{\mathrm{u}}\|^2 + 4\left(\sum_{t=1}^T v_t \eta_t \left\|\boldsymbol{d}_t^{\mathrm{b}}\right\|\right)^2,$$

$$I_t^{\mathrm{cvx}} = 8\max_{t' \in [t]} \left(\sum_{s=1}^{t'} N_s\right)^2 + 2\sum_{s=1}^t \eta_s^2 \|\boldsymbol{d}_s^{\mathrm{u}}\|^2 + 4\left(\sum_{s=1}^t \eta_s \left\|\boldsymbol{d}_s^{\mathrm{b}}\right\|\right)^2, \forall t \in [T],$$

and $\omega = \begin{cases} 1 & \textit{under Assumption 3} \\ 0 & \textit{under Assumption 4} \end{cases}$ *is introduced in Lemma 3.*

Proof. We invoke Lemma 3 with $\boldsymbol{y} = \boldsymbol{z}_t$ (defined in (42)) and replace the subscript t with s to have, for any $s \in [T]$,

$$F(\boldsymbol{x}_{s+1}) - F(\boldsymbol{z}_s) \leq \frac{\|\boldsymbol{z}_s - \boldsymbol{x}_s\|^2 - \|\boldsymbol{z}_s - \boldsymbol{x}_{s+1}\|^2}{2\eta_s} + \langle \boldsymbol{d}_s^{\mathrm{c}}, \boldsymbol{z}_s - \boldsymbol{x}_s \rangle + \eta_s \|\boldsymbol{d}_s^{\mathrm{c}}\|^2 + 4\eta_s G^2 \omega,$$

where $\omega = \begin{cases} 1 & \text{under Assumption 3} \\ 0 & \text{under Assumption 4} \end{cases}$. Multiply both sides of the above inequality by $v_s^2 \eta_s$ and use $\boldsymbol{z}_s - \boldsymbol{x}_s \overset{(42)}{=} \frac{v_{s-1}^2}{v_s^2}(\boldsymbol{z}_{s-1} - \boldsymbol{x}_s)$ and $v_{s-1}^2 \leq v_s^2$ to obtain

$$\begin{aligned} v_s^2 \eta_s \left(F(\boldsymbol{x}_{s+1}) - F(\boldsymbol{z}_s)\right) \leq & \frac{v_{s-1}^2 \|\boldsymbol{z}_{s-1} - \boldsymbol{x}_s\|^2 - v_s^2 \|\boldsymbol{z}_s - \boldsymbol{x}_{s+1}\|^2}{2} \\ & + \langle v_{s-1}\eta_s \boldsymbol{d}_s^{\mathrm{c}}, v_{s-1}(\boldsymbol{z}_{s-1} - \boldsymbol{x}_s) \rangle + v_s^2 \eta_s^2 \|\boldsymbol{d}_s^{\mathrm{c}}\|^2 + 4 v_s^2 \eta_s^2 G^2 \omega, \end{aligned}$$

sum up which from $s = 1$ to t where $t \in [T]$, use $\boldsymbol{z}_0 \overset{(42)}{=} \boldsymbol{x}_\star$ and $D = \|\boldsymbol{x}_\star - \boldsymbol{x}_1\|$, and rearrange terms to have

$$\begin{aligned} & \frac{v_t^2 \|\boldsymbol{z}_t - \boldsymbol{x}_{t+1}\|^2}{2} + \sum_{s=1}^t v_s^2 \eta_s \left(F(\boldsymbol{x}_{s+1}) - F(\boldsymbol{z}_s)\right) \\ \leq & \frac{v_0^2 D^2}{2} + \sum_{s=1}^t \langle v_{s-1}\eta_s \boldsymbol{d}_s^{\mathrm{c}}, v_{s-1}(\boldsymbol{z}_{s-1} - \boldsymbol{x}_s) \rangle + \sum_{s=1}^t v_s^2 \eta_s^2 \|\boldsymbol{d}_s^{\mathrm{c}}\|^2 + 4 \sum_{s=1}^t v_s^2 \eta_s^2 G^2 \omega. \end{aligned} \tag{47}$$

Next, we apply the convexity of F to upper bound

$$F(\boldsymbol{z}_s) - F_\star \overset{(44)}{\leq} \sum_{\ell=1}^s \frac{v_\ell^2 - v_{\ell-1}^2}{v_s^2} \left(F(\boldsymbol{x}_\ell) - F_\star\right) = \sum_{\ell=2}^s \frac{v_\ell^2 - v_{\ell-1}^2}{v_s^2} \left(F(\boldsymbol{x}_\ell) - F_\star\right),$$

where the last equation is due to $v_1 \overset{(43)}{=} v_0$. Hence, we know

$$\begin{aligned} & \sum_{s=1}^t v_s^2 \eta_s \left(F(\boldsymbol{x}_{s+1}) - F(\boldsymbol{z}_s)\right) \\ \geq & \sum_{s=1}^t v_s^2 \eta_s \left(F(\boldsymbol{x}_{s+1}) - F_\star - \sum_{\ell=2}^s \frac{v_\ell^2 - v_{\ell-1}^2}{v_s^2} \left(F(\boldsymbol{x}_\ell) - F_\star\right)\right) \\ = & \sum_{s=1}^t v_s^2 \eta_s \left(F(\boldsymbol{x}_{s+1}) - F_\star\right) - \sum_{s=2}^t \sum_{\ell=2}^s \eta_s (v_\ell^2 - v_{\ell-1}^2) \left(F(\boldsymbol{x}_\ell) - F_\star\right) \\ = & v_t^2 \eta_t \left(F(\boldsymbol{x}_{t+1}) - F_\star\right) + \sum_{s=2}^t \left[v_{s-1}^2 \eta_{s-1} - \left(\sum_{\ell=s}^t \eta_\ell\right)(v_s^2 - v_{s-1}^2)\right] \left(F(\boldsymbol{x}_s) - F_\star\right) \\ \overset{(a)}{=} & v_t^2 \eta_t \left(F(\boldsymbol{x}_{t+1}) - F_\star\right) + \sum_{s=2}^t \left(v_s^2 - v_{s-1}^2\right) \left(\sum_{\ell=t+1}^T \eta_\ell\right) \left(F(\boldsymbol{x}_s) - F_\star\right) \\ \overset{(b)}{\geq} & v_t^2 \eta_t \left(F(\boldsymbol{x}_{t+1}) - F_\star\right), \end{aligned} \tag{48}$$

where (a) is by

$$v_{s-1}^2 \eta_{s-1} - \left(\sum_{\ell=s}^t \eta_\ell\right)(v_s^2 - v_{s-1}^2) = v_{s-1}^2 \left(\sum_{\ell=s-1}^t \eta_\ell\right) - v_s^2 \left(\sum_{\ell=s}^t \eta_\ell\right) \overset{(43)}{=} \left(v_s^2 - v_{s-1}^2\right) \left(\sum_{\ell=t+1}^T \eta_\ell\right),$$

and (b) is due to $v_s^2 \geq v_{s-1}^2 \geq 0$ and $F(\boldsymbol{x}_s) - F_\star \geq 0$. We plug (48) back into (47) to obtain

$$\begin{aligned}
&\frac{v_t^2 \left\|\boldsymbol{z}_t - \boldsymbol{x}_{t+1}\right\|^2}{2} + v_t^2 \eta_t \left(F(\boldsymbol{x}_{t+1}) - F_\star\right) \\
\leq &\frac{v_0^2 D^2}{2} + \sum_{s=1}^{t} \left\langle v_{s-1}\eta_s \boldsymbol{d}_s^{\mathrm{c}}, v_{s-1}\left(\boldsymbol{z}_{s-1} - \boldsymbol{x}_s\right)\right\rangle + \sum_{s=1}^{t} v_s^2 \eta_s^2 \left\|\boldsymbol{d}_s^{\mathrm{c}}\right\|^2 + 4\sum_{s=1}^{t} v_s^2 \eta_s^2 G^2 \omega. \qquad (49)
\end{aligned}$$

Moreover, by $\boldsymbol{d}_s^{\mathrm{c}} \overset{(41)}{=} \boldsymbol{d}_s^{\mathrm{u}} + \boldsymbol{d}_s^{\mathrm{b}}$, we have the decomposition

$$\begin{aligned}
&\left\langle v_{s-1}\eta_s \boldsymbol{d}_s^{\mathrm{c}}, v_{s-1}\left(\boldsymbol{z}_{s-1} - \boldsymbol{x}_s\right)\right\rangle \overset{(45)}{=} R_s M_s + \left\langle v_{s-1}\eta_s \boldsymbol{d}_s^{\mathrm{b}}, v_{s-1}\left(\boldsymbol{z}_{s-1} - \boldsymbol{x}_s\right)\right\rangle \\
\leq &R_s M_s + v_{s-1}\eta_s \left\|\boldsymbol{d}_s^{\mathrm{b}}\right\| \cdot v_{s-1}\left\|\boldsymbol{z}_{s-1} - \boldsymbol{x}_s\right\| \overset{(45), v_{s-1}\leq v_s}{\leq} R_s M_s + R_s v_s \eta_s \left\|\boldsymbol{d}_s^{\mathrm{b}}\right\|,
\end{aligned}$$

which implies that

$$\begin{aligned}
&\sum_{s=1}^{t} \left\langle v_{s-1}\eta_s \boldsymbol{d}_s^{\mathrm{c}}, v_{s-1}\left(\boldsymbol{z}_{s-1} - \boldsymbol{x}_s\right)\right\rangle \leq \sum_{s=1}^{t} R_s M_s + \sum_{s=1}^{t} R_s v_s \eta_s \left\|\boldsymbol{d}_s^{\mathrm{b}}\right\| \\
\overset{(c)}{\leq} &2R_t \max_{t'\in[t]} \left|\sum_{s=1}^{t'} M_s\right| + R_t \sum_{s=1}^{t} v_s \eta_s \left\|\boldsymbol{d}_s^{\mathrm{b}}\right\| \overset{(d)}{\leq} \frac{R_t^2}{4} + 8\max_{t'\in[t]} \left(\sum_{s=1}^{t'} M_s\right)^2 + 2\left(\sum_{s=1}^{t} v_s \eta_s \left\|\boldsymbol{d}_s^{\mathrm{b}}\right\|\right)^2,
\end{aligned}$$

where (c) is due to Lemma 13, and (d) is by AM-GM inequality. Combine the above inequality and (49) to know

$$\begin{aligned}
&\frac{v_t^2 \left\|\boldsymbol{z}_t - \boldsymbol{x}_{t+1}\right\|^2}{2} + v_t^2 \eta_t \left(F(\boldsymbol{x}_{t+1}) - F_\star\right) \\
\leq &\frac{v_0^2 D^2}{2} + \frac{R_t^2}{4} + 8\max_{t'\in[t]} \left(\sum_{s=1}^{t'} M_s\right)^2 + 2\left(\sum_{s=1}^{t} v_s \eta_s \left\|\boldsymbol{d}_s^{\mathrm{b}}\right\|\right)^2 + \sum_{s=1}^{t} v_s^2 \eta_s^2 \left\|\boldsymbol{d}_s^{\mathrm{c}}\right\|^2 + 4\sum_{s=1}^{t} v_s^2 \eta_s^2 G^2 \omega \\
\leq &\frac{R_t^2}{4} + \frac{v_0^2 D^2}{2} + \underbrace{8\max_{t'\in[t]} \left(\sum_{s=1}^{t'} M_s\right)^2 + 2\sum_{s=1}^{t} v_s^2 \eta_s^2 \left\|\boldsymbol{d}_s^{\mathrm{u}}\right\|^2 + 4\left(\sum_{s=1}^{t} v_s \eta_s \left\|\boldsymbol{d}_s^{\mathrm{b}}\right\|\right)^2}_{\triangleq P_t^{\mathrm{cvx}}} + 4\sum_{s=1}^{t} v_s^2 \eta_s^2 G^2 \omega,
\end{aligned} \qquad (50)$$

where the last step is by

$$\begin{aligned}
&\left\|\boldsymbol{d}_s^{\mathrm{c}}\right\|^2 \overset{(41)}{=} \left\|\boldsymbol{d}_s^{\mathrm{u}} + \boldsymbol{d}_s^{\mathrm{b}}\right\|^2 \leq 2\left\|\boldsymbol{d}_s^{\mathrm{u}}\right\|^2 + 2\left\|\boldsymbol{d}_s^{\mathrm{b}}\right\|^2 \\
\Rightarrow &\sum_{s=1}^{t} v_s^2 \eta_s^2 \left\|\boldsymbol{d}_s^{\mathrm{c}}\right\|^2 \leq 2\sum_{s=1}^{t} v_s^2 \eta_s^2 \left\|\boldsymbol{d}_s^{\mathrm{u}}\right\|^2 + 2\sum_{s=1}^{t} v_s^2 \eta_s^2 \left\|\boldsymbol{d}_s^{\mathrm{b}}\right\|^2 \leq 2\sum_{s=1}^{t} v_s^2 \eta_s^2 \left\|\boldsymbol{d}_s^{\mathrm{u}}\right\|^2 + 2\left(\sum_{s=1}^{t} v_s \eta_s \left\|\boldsymbol{d}_s^{\mathrm{b}}\right\|\right)^2.
\end{aligned}$$

Now, let $a_t \triangleq \frac{v_{t-1}^2 \left\|\boldsymbol{z}_{t-1} - \boldsymbol{x}_t\right\|^2}{2}$, $b_t \triangleq v_t^2 \eta_t \left(F(\boldsymbol{x}_{t+1}) - F_\star\right)$, and $c_t \triangleq \frac{v_0^2 D^2}{2} + P_{t-1}^{\mathrm{cvx}} + 4\sum_{s=1}^{t-1} v_s^2 \eta_s^2 G^2 \omega$. Then, (50) is saying that

$$a_{t+1} + b_t \leq \frac{\max_{s\in[t]} a_s}{2} + c_{t+1}, \forall t \in [T].$$

Note that $b_t \geq 0$, c_t is nondecreasing, and $a_1 = \frac{v_0^2 D^2}{2} \leq v_0^2 D^2 = 2c_1$. Thus, by Lemma 14, we have

$$a_{t+1} + b_t \leq 2c_{t+1}, \forall t \in [T],$$

implying that

$$v_T^2 \eta_T \left(F(\boldsymbol{x}_{T+1}) - F_\star\right) = b_T \leq a_{T+1} + b_T \leq 2c_{T+1} = v_0^2 D^2 + 2P_T^{\mathrm{cvx}} + 8\sum_{t=1}^{T} v_t^2 \eta_t^2 G^2 \omega.$$

Finally, we divide both sides by $v_T^2 \eta_T$ and use $v_T^2 \overset{(43)}{=} 1$ to complete the proof for the first inequality.

To prove the inequality for $\|\boldsymbol{x}_\star - \boldsymbol{x}_{t+1}\|^2$, we can simply repeat the above analysis, up to replacing $\boldsymbol{y} = \boldsymbol{z}_s$ with $\boldsymbol{y} = \boldsymbol{x}_\star$. One new quantity defined in the proof is

$$I_t^{\mathrm{cvx}} \triangleq 8 \max_{t' \in [t]} \left(\sum_{s=1}^{t'} N_s \right)^2 + 2 \sum_{s=1}^{t} \eta_s^2 \left\| \boldsymbol{d}_s^{\mathrm{u}} \right\|^2 + 4 \left(\sum_{s=1}^{t} \eta_s \left\| \boldsymbol{d}_s^{\mathrm{b}} \right\| \right)^2, \forall t \in [T] .$$

For simplicity, we omit the details. □

C.2.1 Lipschitz Objectives

In this part, we consider Lipschitz objectives (i.e., Assumption 3). To prove the last-iterate rate, it is sufficient to bound the residual term P_T^{cvx} in Lemma 4. To do so, we rely on the following Lemma 5.

Lemma 5. *Under Assumptions 3 and 5, and the following two conditions:*

1. *η_t and τ_t are deterministic for all $t \in [T]$.*
2. *$\tau_t \geq 2G$ for all $t \in [T]$.*

We have for any $\delta \in (0, 1]$, with probability at least $1 - \delta$, $P_T^{\mathrm{cvx}} \leq U_T^{\mathrm{cvx}}$, where P_T^{cvx} is introduced in Lemma 4 and U_T^{cvx} is a constant in the order of

$$\mathcal{O}\left(\max_{t \in [T]} v_t^2 \eta_t^2 \tau_t^2 \ln^2 \frac{3}{\delta} + \sum_{t=1}^{T} \frac{\sigma_1^{\mathtt{p}} v_t^2 \eta_t^2}{\tau_t^{\mathtt{p}-2}} + \left(\sum_{t=1}^{T} \frac{\sigma_{\mathtt{s}} \sigma_1^{\mathtt{p}-1} v_t \eta_t}{\tau_t^{\mathtt{p}-1}} + \frac{\sigma_1^{\mathtt{p}} G v_t \eta_t}{\tau_t^{\mathtt{p}}} \right)^2 \right).$$

Proof. Note that $\tau_t \overset{\text{Condition 2}}{\geq} 2G \overset{\text{Assumption 3}}{\geq} 2 \left\| \nabla f(\boldsymbol{x}_t) \right\|, \forall t \in [T]$ holds, implying that χ_t in Lemma 2 equals 1 for all $t \in [T]$. Then Lemma 2 and Assumption 3 together yield that the following collection of inequalities holds for any $t \in [T]$:

$$\left\| \boldsymbol{d}_t^{\mathrm{u}} \right\| \leq 2\tau_t, \tag{51}$$

$$\mathbb{E}\left[\left\| \boldsymbol{d}_t^{\mathrm{u}} \right\|^2 \mid \mathcal{F}_{t-1} \right] \leq \frac{4\sigma_1^{\mathtt{p}}}{\tau_t^{\mathtt{p}-2}}, \tag{52}$$

$$\left\| \mathbb{E}\left[\boldsymbol{d}_t^{\mathrm{u}} \left(\boldsymbol{d}_t^{\mathrm{u}} \right)^\top \mid \mathcal{F}_{t-1} \right] \right\|_{\mathtt{op}} \leq \frac{4\sigma_{\mathtt{s}}^{\mathtt{p}}}{\tau_t^{\mathtt{p}-2}} + \frac{2^{\mathtt{p}+1} \sigma_1^{\mathtt{p}} G^2}{\tau_t^{\mathtt{p}}}, \tag{53}$$

$$\left\| \boldsymbol{d}_t^{\mathrm{b}} \right\| \leq \frac{\sigma_{\mathtt{s}} \sigma_1^{\mathtt{p}-1}}{\tau_t^{\mathtt{p}-1}} + \frac{2^{\mathtt{p}-1} \sigma_1^{\mathtt{p}} G}{\tau_t^{\mathtt{p}}}. \tag{54}$$

Now, we are ready to bound P_T^{cvx}.

- We first observe that the sequence $\{v_t\}_{t=1}^T$ is now deterministic by Condition 1 and its definition in (43), which further implies that $\boldsymbol{z}_t \in \mathcal{F}_{t-1}$ and $R_t \in \mathcal{F}_{t-1}$ (see (44) and (45) for their definitions). Therefore, we know

$$\mathbb{E}\left[M_t \mid \mathcal{F}_{t-1} \right] \overset{(45)}{=} \left\langle v_{t-1} \eta_t \mathbb{E}\left[\boldsymbol{d}_t^{\mathrm{u}} \mid \mathcal{F}_{t-1} \right], \frac{v_{t-1} \left(\boldsymbol{z}_{t-1} - \boldsymbol{x}_t \right)}{R_t} \right\rangle = 0, \forall t \in [T] ,$$

  meaning that $\{M_t\}_{t=1}^T$ is a real-valued MDS. By Cauchy-Schwarz inequality, we can bound

$$\begin{aligned} |M_t| &\overset{(45)}{=} \left| \left\langle v_{t-1} \eta_t \boldsymbol{d}_t^{\mathrm{u}}, \frac{v_{t-1} \left(\boldsymbol{z}_{t-1} - \boldsymbol{x}_t \right)}{R_t} \right\rangle \right| \leq v_{t-1} \eta_t \left\| \boldsymbol{d}_t^{\mathrm{u}} \right\| \cdot \frac{v_{t-1} \left\| \boldsymbol{z}_{t-1} - \boldsymbol{x}_t \right\|}{R_t} \\ &\overset{(45), v_{t-1} \leq v_t}{\leq} v_t \eta_t \left\| \boldsymbol{d}_t^{\mathrm{u}} \right\| \overset{(51)}{\leq} 2 v_t \eta_t \tau_t \leq 2 \max_{t \in [T]} v_t \eta_t \tau_t. \end{aligned}$$

In addition, the conditional variance of M_t can be controlled by

$$\mathbb{E}\left[M_t^2 \mid \mathcal{F}_{t-1}\right] = \left(\frac{v_{t-1}\left(\boldsymbol{z}_{t-1}-\boldsymbol{x}_t\right)}{R_t}\right)^\top v_{t-1}^2\eta_t^2\mathbb{E}\left[\boldsymbol{d}_t^{\mathrm{u}}\left(\boldsymbol{d}_t^{\mathrm{u}}\right)^\top \mid \mathcal{F}_{t-1}\right]\frac{v_{t-1}\left(\boldsymbol{z}_{t-1}-\boldsymbol{x}_t\right)}{R_t}$$
$$\overset{(45),v_{t-1}\le v_t}{\le} v_{t-1}^2\eta_t^2\left\|\mathbb{E}\left[\boldsymbol{d}_t^{\mathrm{u}}\left(\boldsymbol{d}_t^{\mathrm{u}}\right)^\top \mid \mathcal{F}_{t-1}\right]\right\|_{\mathtt{op}} \overset{(53)}{\le} \frac{4\sigma_{\mathtt{s}}^{\mathtt{p}}v_t^2\eta_t^2}{\tau_t^{\mathtt{p}-2}} + \frac{2^{\mathtt{p}+1}\sigma_1^{\mathtt{p}}G^2v_t^2\eta_t^2}{\tau_t^{\mathtt{p}}}.$$

Therefore, by Freedman's inequality (Lemma 11), with probability at least $1-2\delta/3$,

$$\left|\sum_{s=1}^t M_s\right| \le \frac{2}{3}\max_{s\in[T]} v_s\eta_s\tau_s\ln\frac{3}{\delta} + \sqrt{8\sum_{s=1}^T\left(\frac{\sigma_{\mathtt{s}}^{\mathtt{p}}v_s^2\eta_s^2}{\tau_s^{\mathtt{p}-2}} + \frac{2^{\mathtt{p}-1}\sigma_1^{\mathtt{p}}G^2v_s^2\eta_s^2}{\tau_s^{\mathtt{p}}}\right)\ln\frac{3}{\delta}}, \forall t\in[T],$$

which implies that

$$\max_{t\in[T]}\left(\sum_{s=1}^t M_s\right)^2 \le \frac{8}{9}\max_{t\in[T]} v_t^2\eta_t^2\tau_t^2\ln^2\frac{3}{\delta} + 16\sum_{t=1}^T\left(\frac{\sigma_{\mathtt{s}}^{\mathtt{p}}v_t^2\eta_t^2}{\tau_t^{\mathtt{p}-2}} + \frac{2^{\mathtt{p}-1}\sigma_1^{\mathtt{p}}G^2v_t^2\eta_t^2}{\tau_t^{\mathtt{p}}}\right)\ln\frac{3}{\delta}. \quad (55)$$

- Next, note that $\{v_t\eta_t\left\|\boldsymbol{d}_t^{\mathrm{u}}\right\|\}_{t=1}^T$ is a sequence of random variables satisfying

$$v_t\eta_t\left\|\boldsymbol{d}_t^{\mathrm{u}}\right\| \overset{(51)}{\le} 2v_t\eta_t\tau_t \le 2\max_{t\in[T]} v_t\eta_t\tau_t \quad \text{and} \quad \mathbb{E}\left[v_t^2\eta_t^2\left\|\boldsymbol{d}_t^{\mathrm{u}}\right\|^2 \mid \mathcal{F}_{t-1}\right] \overset{(52)}{\le} \frac{4\sigma_1^{\mathtt{p}}v_t^2\eta_t^2}{\tau_t^{\mathtt{p}-2}}.$$

Therefore, by Lemma 12, with probability at least $1-\delta/3$,

$$\sum_{t=1}^T v_t^2\eta_t^2\left\|\boldsymbol{d}_t^{\mathrm{u}}\right\|^2 \le \frac{10}{3}\max_{t\in[T]} v_t^2\eta_t^2\tau_t^2\ln\frac{3}{\delta} + 8\sum_{t=1}^T\frac{\sigma_1^{\mathtt{p}}v_t^2\eta_t^2}{\tau_t^{\mathtt{p}-2}}$$
$$\overset{\ln\frac{3}{\delta}\ge 1}{\le} \frac{10}{3}\max_{t\in[T]} v_t^2\eta_t^2\tau_t^2\ln^2\frac{3}{\delta} + 8\sum_{t=1}^T\frac{\sigma_1^{\mathtt{p}}v_t^2\eta_t^2}{\tau_t^{\mathtt{p}-2}}. \quad (56)$$

- Lastly, there is

$$\sum_{t=1}^T v_t\eta_t\left\|\boldsymbol{d}_t^{\mathrm{b}}\right\| \overset{(54)}{\le} \sum_{t=1}^T\frac{\sigma_{\mathtt{s}}\sigma_1^{\mathtt{p}-1}v_t\eta_t}{\tau_t^{\mathtt{p}-1}} + \frac{2^{\mathtt{p}-1}\sigma_1^{\mathtt{p}}Gv_t\eta_t}{\tau_t^{\mathtt{p}}}. \quad (57)$$

Combine (55), (56), and (57) to have with probability at least $1-\delta$,

$$P_T^{\mathrm{cvx}} = 8\max_{t\in[T]}\left(\sum_{s=1}^t M_s\right)^2 + 2\sum_{t=1}^T v_t^2\eta_t^2\left\|\boldsymbol{d}_t^{\mathrm{u}}\right\|^2 + 4\left(\sum_{t=1}^T v_t\eta_t\left\|\boldsymbol{d}_t^{\mathrm{b}}\right\|\right)^2 \le U_T^{\mathrm{cvx}},$$

where

$$U_T^{\mathrm{cvx}} \triangleq \frac{124}{9}\max_{t\in[T]} v_t^2\eta_t^2\tau_t^2\ln^2\frac{3}{\delta} + 128\sum_{t=1}^T\left(\frac{\sigma_{\mathtt{s}}^{\mathtt{p}}v_t^2\eta_t^2}{\tau_t^{\mathtt{p}-2}} + \frac{2^{\mathtt{p}-1}\sigma_1^{\mathtt{p}}G^2v_t^2\eta_t^2}{\tau_t^{\mathtt{p}}}\right)\ln\frac{3}{\delta}$$
$$+ 16\sum_{t=1}^T\frac{\sigma_1^{\mathtt{p}}v_t^2\eta_t^2}{\tau_t^{\mathtt{p}-2}} + 4\left(\sum_{t=1}^T\frac{\sigma_{\mathtt{s}}\sigma_1^{\mathtt{p}-1}v_t\eta_t}{\tau_t^{\mathtt{p}-1}} + \frac{2^{\mathtt{p}-1}\sigma_1^{\mathtt{p}}Gv_t\eta_t}{\tau_t^{\mathtt{p}}}\right)^2$$
$$= \mathcal{O}\left(\max_{t\in[T]} v_t^2\eta_t^2\tau_t^2\ln^2\frac{3}{\delta} + \sum_{t=1}^T\frac{\sigma_1^{\mathtt{p}}v_t^2\eta_t^2}{\tau_t^{\mathtt{p}-2}} + \left(\sum_{t=1}^T\frac{\sigma_{\mathtt{s}}\sigma_1^{\mathtt{p}-1}v_t\eta_t}{\tau_t^{\mathtt{p}-1}} + \frac{\sigma_1^{\mathtt{p}}Gv_t\eta_t}{\tau_t^{\mathtt{p}}}\right)^2\right),$$

in which the second equation is by noticing that

$$\begin{aligned}
&\max_{t\in[T]} v_t^2\eta_t^2\tau_t^2\ln^2\frac{3}{\delta}+\left(\sum_{t=1}^T\frac{\sigma_{\mathtt{s}}\sigma_{\mathtt{1}}^{\mathtt{p}-1}v_t\eta_t}{\tau_t^{\mathtt{p}-1}}+\frac{\sigma_{\mathtt{1}}^{\mathtt{p}}Gv_t\eta_t}{\tau_t^{\mathtt{p}}}\right)^2\\
\overset{(a)}{\geq}&2\left(\max_{t\in[T]} v_t\eta_t\tau_t\right)\sum_{t=1}^T\left(\frac{\sigma_{\mathtt{s}}\sigma_{\mathtt{1}}^{\mathtt{p}-1}v_t\eta_t}{\tau_t^{\mathtt{p}-1}}+\frac{\sigma_{\mathtt{1}}^{\mathtt{p}}Gv_t\eta_t}{\tau_t^{\mathtt{p}}}\right)\ln\frac{3}{\delta}\\
\geq&2\sum_{t=1}^T\left(\frac{\sigma_{\mathtt{s}}\sigma_{\mathtt{1}}^{\mathtt{p}-1}v_t^2\eta_t^2}{\tau_t^{\mathtt{p}-2}}+\frac{\sigma_{\mathtt{1}}^{\mathtt{p}}Gv_t^2\eta_t^2}{\tau_t^{\mathtt{p}-1}}\right)\ln\frac{3}{\delta}\\
\overset{(b)}{=}&\Omega\left(\sum_{t=1}^T\left(\frac{\sigma_{\mathtt{s}}^{\mathtt{p}}v_t^2\eta_t^2}{\tau_t^{\mathtt{p}-2}}+\frac{\sigma_{\mathtt{1}}^{\mathtt{p}}G^2v_t^2\eta_t^2}{\tau_t^{\mathtt{p}}}\right)\ln\frac{3}{\delta}\right),
\end{aligned}$$

where (a) is due to AM-GM inequality and (b) is by $\sigma_{\mathtt{1}}\geq\sigma_{\mathtt{s}}$, $\mathtt{p}\geq 1$, and $\tau_t \overset{\text{Condition 2}}{\geq} 2G$. □

C.2.2 Smooth Objectives

In this part, we consider smooth objectives (i.e., Assumption 4). One may still think that, like the Lipschitz case, we only need to bound the residual term P_T^{cvx} in Lemma 4. This is not a wrong thought. However, an obstacle will arise when considering smooth objectives, that is, how to bound $\|\nabla f(\boldsymbol{x}_t)\|$ on the R.H.S. of inequalities in Lemma 2. Why is this important? Because, to apply Freedman's inequality (Lemma 11), the conditional variance of the martingale should be upper bounded by a constant.

To handle this challenge, our solution is to first bound $\bar{P}_T^{\text{cvx}}$ in the following Lemma 6, a variant of P_T^{cvx} under the condition that $\|\boldsymbol{x}_\star-\boldsymbol{x}_t\|$ is uniformly bounded for all $t\in[T]$. Then, Lemma 7 provides a sufficient condition to ensure that $\|\boldsymbol{x}_\star-\boldsymbol{x}_t\|$ is uniformly bounded.

Lemma 6. *Under Assumptions 1, 4, 5, 6, and the following two conditions:*

1. *η_t is deterministic for all $t\in[T]$.*
2. *$\tau_t=\max\{2G_t,\tau\}$ for all $t\in[T]$, where $G_t=L\|\boldsymbol{x}_t-\boldsymbol{x}_1\|+G_1$ and $\tau\geq 0$ is a constant.*

We have for any $\delta\in(0,1]$, with probability at least $1-\delta$, $\bar{P}_T^{\text{cvx}}\leq\bar{U}_T^{\text{cvx}}$, where

$$\begin{aligned}
\bar{P}_T^{\text{cvx}}\triangleq&\,8\max_{t\in[T]}\left(\sum_{s=1}^t M_s\zeta_s\right)^2+2\sum_{t=1}^T v_t^2\eta_t^2\left\|\boldsymbol{d}_t^{\text{u}}\right\|^2\zeta_t^2+4\left(\sum_{t=1}^T v_t\eta_t\left\|\boldsymbol{d}_t^{\text{b}}\right\|\zeta_t\right)^2,\\
\zeta_t\triangleq&\,\mathbb{1}\left[\|\boldsymbol{x}_\star-\boldsymbol{x}_t\|\leq\bar{D}\right],\forall t\in[T],
\end{aligned}$$

$\bar{D}\geq 0$ can be any constant, and $\bar{U}_T^{\text{cvx}}$ is a constant in the order of

$$\mathcal{O}\left(\max_{t\in[T]} v_t^2\eta_t^2\rho^2\ln^2\frac{3}{\delta}+\sum_{t=1}^T\frac{\sigma_{\mathtt{1}}^{\mathtt{p}}v_t^2\eta_t^2}{\tau^{\mathtt{p}-2}}+\left(\sum_{t=1}^T\frac{\sigma_{\mathtt{s}}\sigma_{\mathtt{1}}^{\mathtt{p}-1}v_t\eta_t}{\tau^{\mathtt{p}-1}}+\min_{\beta\in\left[0,\frac{1}{2}\right]}\frac{\sigma_{\mathtt{1}}^{\mathtt{p}}\bar{G}^{1-\beta\mathtt{p}}v_t\eta_t}{\tau^{(1-\beta)\mathtt{p}}}\right)^2\right),$$

in which $\bar{G}=L(\bar{D}+D)+G_1$ and $\rho=\max\left\{2\bar{G},\tau\right\}$.

Proof. By L-smoothness (Assumption 4) and $\|\nabla f(\boldsymbol{x}_1)\|\leq G_1$ (Assumption 6), we have

$$\|\nabla f(\boldsymbol{x}_t)\|\leq\|\nabla f(\boldsymbol{x}_t)-\nabla f(\boldsymbol{x}_1)\|+\|\nabla f(\boldsymbol{x}_1)\|\leq L\|\boldsymbol{x}_t-\boldsymbol{x}_1\|+G_1=G_t,\forall t\in[T]. \tag{58}$$

Moreover, under Assumption 1, we note that $G_t\leq L\|\boldsymbol{x}_\star-\boldsymbol{x}_t\|+L\|\boldsymbol{x}_\star-\boldsymbol{x}_1\|+G_1$, implying that

$$\|\nabla f(\boldsymbol{x}_t)\|\zeta_t\overset{(58)}{\leq}G_t\zeta_t\leq L(\bar{D}+D)+G_1\triangleq\bar{G},\forall t\in[T], \tag{59}$$

and

$$\tau_t\zeta_t=\max\{2G_t,\tau\}\zeta_t\overset{\zeta_t\leq 1}{\leq}\max\{2G_t\zeta_t,\tau\}\overset{(59)}{\leq}\max\left\{2\bar{G},\tau\right\}\triangleq\rho,\forall t\in[T]. \tag{60}$$

We observe that

$$2\left\|\nabla f(\boldsymbol{x}_t)\right\| \overset{(58)}{\leq} 2G_t \leq \max\left\{2G_t, \tau\right\} = \tau_t, \forall t \in [T], \tag{61}$$

which implies that χ_t in Lemma 2 equals 1 for all $t \in [T]$. Then Lemma 2 and (58) together yield that the following collection of inequalities holds for any $t \in [T]$:

$$\left\|\boldsymbol{d}_t^{\mathrm{u}}\right\| \leq 2\tau_t, \tag{62}$$

$$\mathbb{E}\left[\left\|\boldsymbol{d}_t^{\mathrm{u}}\right\|^2 \mid \mathcal{F}_{t-1}\right] \leq \frac{4\sigma_1^{\mathtt{p}}}{\tau_t^{\mathtt{p}-2}}, \tag{63}$$

$$\left\|\mathbb{E}\left[\boldsymbol{d}_t^{\mathrm{u}}\left(\boldsymbol{d}_t^{\mathrm{u}}\right)^\top \mid \mathcal{F}_{t-1}\right]\right\|_{\mathtt{op}} \leq \frac{4\sigma_{\mathtt{s}}^{\mathtt{p}}}{\tau_t^{\mathtt{p}-2}} + \frac{2^{\mathtt{p}+1}\sigma_1^{\mathtt{p}}G_t^2}{\tau_t^{\mathtt{p}}}, \tag{64}$$

$$\left\|\boldsymbol{d}_t^{\mathrm{b}}\right\| \leq \frac{\sigma_{\mathtt{s}}\sigma_1^{\mathtt{p}-1}}{\tau_t^{\mathtt{p}-1}} + \frac{2^{\mathtt{p}-1}\sigma_1^{\mathtt{p}}G_t}{\tau_t^{\mathtt{p}}}. \tag{65}$$

Now, we are ready to bound $\bar{P}_T^{\mathrm{cvx}}$.

- We first observe that the sequence $\{v_t\}_{t=1}^T$ is now deterministic by Condition 1 and its definition in (43), which further implies that $\boldsymbol{z}_t \in \mathcal{F}_{t-1}$ and $R_t \in \mathcal{F}_{t-1}$ (see (44) and (45) for their definitions). Therefore, using $\zeta_t \in \mathcal{F}_{t-1}$, we know

$$\mathbb{E}\left[M_t\zeta_t \mid \mathcal{F}_{t-1}\right] \overset{(45)}{=} \left\langle v_{t-1}\eta_t\mathbb{E}\left[\boldsymbol{d}_t^{\mathrm{u}} \mid \mathcal{F}_{t-1}\right], \frac{v_{t-1}\left(\boldsymbol{z}_{t-1}-\boldsymbol{x}_t\right)}{R_t}\right\rangle \zeta_t = 0, \forall t \in [T],$$

  meaning that $\{M_t\zeta_t\}_{t=1}^T$ is a real-valued MDS. By Cauchy-Schwarz inequality, we can bound

$$\begin{aligned}\left|M_t\zeta_t\right| &\overset{(45)}{=} \left|\left\langle v_{t-1}\eta_t\boldsymbol{d}_t^{\mathrm{u}}, \frac{v_{t-1}\left(\boldsymbol{z}_{t-1}-\boldsymbol{x}_t\right)}{R_t}\right\rangle\right|\zeta_t \leq v_{t-1}\eta_t\left\|\boldsymbol{d}_t^{\mathrm{u}}\right\| \cdot \frac{v_{t-1}\left\|\boldsymbol{z}_{t-1}-\boldsymbol{x}_t\right\|}{R_t}\cdot\zeta_t \\ &\overset{(45),v_{t-1}\leq v_t}{\leq} v_t\eta_t\left\|\boldsymbol{d}_t^{\mathrm{u}}\right\|\zeta_t \overset{(62)}{\leq} 2v_t\eta_t\tau_t\zeta_t \overset{(60)}{\leq} 2v_t\eta_t\rho \leq 2\max_{t\in[T]} v_t\eta_t\rho.\end{aligned}$$

  In addition, the conditional variance of $M_t\zeta_t$ can be controlled by

$$\begin{aligned}\mathbb{E}\left[\left(M_t\zeta_t\right)^2 \mid \mathcal{F}_{t-1}\right] &= \left(\frac{v_{t-1}\left(\boldsymbol{z}_{t-1}-\boldsymbol{x}_t\right)}{R_t}\right)^\top v_{t-1}^2\eta_t^2\mathbb{E}\left[\boldsymbol{d}_t^{\mathrm{u}}\left(\boldsymbol{d}_t^{\mathrm{u}}\right)^\top \mid \mathcal{F}_{t-1}\right]\frac{v_{t-1}\left(\boldsymbol{z}_{t-1}-\boldsymbol{x}_t\right)}{R_t}\zeta_t^2 \\ &\overset{(45),v_{t-1}\leq v_t}{\leq} v_t^2\eta_t^2\left\|\mathbb{E}\left[\boldsymbol{d}_t^{\mathrm{u}}\left(\boldsymbol{d}_t^{\mathrm{u}}\right)^\top \mid \mathcal{F}_{t-1}\right]\right\|_{\mathtt{op}}\zeta_t^2 \overset{(64)}{\leq} \frac{4\sigma_{\mathtt{s}}^{\mathtt{p}}v_t^2\eta_t^2\zeta_t^2}{\tau_t^{\mathtt{p}-2}} + \frac{2^{\mathtt{p}+1}\sigma_1^{\mathtt{p}}G_t^2v_t^2\eta_t^2\zeta_t^2}{\tau_t^{\mathtt{p}}} \\ &\leq \frac{4\sigma_{\mathtt{s}}^{\mathtt{p}}v_t^2\eta_t^2\rho}{\tau^{\mathtt{p}-1}} + \min_{\beta\in\left[0,\frac{1}{2}\right]}\frac{2^{\mathtt{p}+1}\sigma_1^{\mathtt{p}}\bar{G}^{2-\beta_{\mathtt{p}}}v_t^2\eta_t^2}{2^{\beta_{\mathtt{p}}}\tau^{(1-\beta)\mathtt{p}}},\end{aligned}$$

  where the last step is due to

$$\frac{\zeta_t^2}{\tau_t^{\mathtt{p}-2}} = \frac{\tau_t\zeta_t^2}{\tau_t^{\mathtt{p}-1}} \overset{\tau_t\geq\tau,\mathtt{p}\geq 1}{\leq} \frac{\tau_t\zeta_t^2}{\tau^{\mathtt{p}-1}} \overset{\zeta_t\leq 1,(60)}{\leq} \frac{\rho}{\tau^{\mathtt{p}-1}},$$

  and, for any $\beta \in \left[0, \frac{1}{2}\right]$,

$$\frac{G_t^2\zeta_t^2}{\tau_t^{\mathtt{p}}} \overset{\tau_t\geq\tau,(61)}{\leq} \frac{G_t^{2-\beta_{\mathtt{p}}}\zeta_t^2}{2^{\beta_{\mathtt{p}}}\tau^{(1-\beta)\mathtt{p}}} \overset{\zeta_t\leq 1,\beta_{\mathtt{p}}\leq 2,(59)}{\leq} \frac{\bar{G}^{2-\beta_{\mathtt{p}}}}{2^{\beta_{\mathtt{p}}}\tau^{(1-\beta)\mathtt{p}}}.$$

  Therefore, by Freedman's inequality (Lemma 11), with probability at least $1-2\delta/3$,

$$\left|\sum_{s=1}^t M_s\zeta_s\right| \leq \frac{2}{3}\max_{s\in[T]} v_s\eta_s\rho\ln\frac{3}{\delta} + \sqrt{8\sum_{s=1}^T\left(\frac{\sigma_{\mathtt{s}}^{\mathtt{p}}v_s^2\eta_s^2\rho}{\tau^{\mathtt{p}-1}} + \min_{\beta\in\left[0,\frac{1}{2}\right]}\frac{2^{\mathtt{p}-1}\sigma_1^{\mathtt{p}}\bar{G}^{2-\beta_{\mathtt{p}}}v_s^2\eta_s^2}{2^{\beta_{\mathtt{p}}}\tau^{(1-\beta)\mathtt{p}}}\right)\ln\frac{3}{\delta}},$$

  which implies that

$$\max_{t\in[T]}\left(\sum_{s=1}^t M_s\zeta_s\right)^2 \leq \frac{8}{9}\max_{t\in[T]} v_t^2\eta_t^2\rho^2\ln^2\frac{3}{\delta} + 16\sum_{t=1}^T\left(\frac{\sigma_{\mathtt{s}}^{\mathtt{p}}v_t^2\eta_t^2\rho}{\tau^{\mathtt{p}-1}} + \min_{\beta\in\left[0,\frac{1}{2}\right]}\frac{2^{\mathtt{p}-1}\sigma_1^{\mathtt{p}}\bar{G}^{2-\beta_{\mathtt{p}}}v_t^2\eta_t^2}{2^{\beta_{\mathtt{p}}}\tau^{(1-\beta)\mathtt{p}}}\right)\ln\frac{3}{\delta}. \tag{66}$$

- Next, note that $\{v_t\eta_t \left\|\boldsymbol{d}_t^{\mathrm{u}}\right\| \zeta_t\}_{t=1}^T$ is a sequence of random variables satisfying

$$v_t\eta_t \left\|\boldsymbol{d}_t^{\mathrm{u}}\right\| \zeta_t \overset{(62)}{\leq} 2v_t\eta_t\tau_t\zeta_t \overset{(60)}{\leq} 2v_t\eta_t\rho \leq 2\max_{t\in[T]} v_t\eta_t\rho,$$

and

$$\mathbb{E}\left[v_t^2\eta_t^2 \left\|\boldsymbol{d}_t^{\mathrm{u}}\right\|^2 \zeta_t^2 \mid \mathcal{F}_{t-1}\right] \overset{(63)}{\leq} \frac{4\sigma_1^{\mathtt{p}} v_t^2\eta_t^2\zeta_t^2}{\tau_t^{\mathtt{p}-2}} \leq \frac{4\sigma_1^{\mathtt{p}} v_t^2\eta_t^2}{\tau^{\mathtt{p}-2}} + \min_{\beta\in\left[0,\frac{1}{2}\right]} \frac{4\sigma_1^{\mathtt{p}}(2\bar{G})^{2-\beta_{\mathtt{p}}} v_t^2\eta_t^2}{\tau^{(1-\beta)\mathtt{p}}},$$

where the second step is due to, for any $\beta \in \left[0, \frac{1}{2}\right]$,

$$\begin{aligned}\frac{\zeta_t^2}{\tau_t^{\mathtt{p}-2}} = \frac{\tau_t^2\zeta_t^2}{\tau_t^{\mathtt{p}}} &\overset{\text{Condition 2}}{\leq} \frac{\left(\tau^2 + 4G_t^2\right)\zeta_t^2}{\tau_t^{\mathtt{p}}} \overset{\tau_t\geq\tau,\zeta_t\leq 1}{\leq} \frac{1}{\tau^{\mathtt{p}-2}} + \frac{4G_t^2\zeta_t^2}{\tau_t^{\mathtt{p}}} \\ &\overset{\tau_t\geq\tau,(61)}{\leq} \frac{1}{\tau^{\mathtt{p}-2}} + \frac{(2G_t)^{2-\beta_{\mathtt{p}}}\zeta_t^2}{\tau^{(1-\beta)\mathtt{p}}} \overset{\zeta_t\leq 1,\beta_{\mathtt{p}}\leq 2,(59)}{\leq} \frac{1}{\tau^{\mathtt{p}-2}} + \frac{(2\bar{G})^{2-\beta_{\mathtt{p}}}}{\tau^{(1-\beta)\mathtt{p}}}.\end{aligned}$$

Therefore, by Lemma 12, with probability at least $1-\delta/3$,

$$\begin{aligned}\sum_{t=1}^T v_t^2\eta_t^2 \left\|\boldsymbol{d}_t^{\mathrm{u}}\right\|^2 \zeta_t^2 &\leq \frac{10}{3}\max_{t\in[T]} v_t^2\eta_t^2\rho^2 \ln\frac{3}{\delta} + 8\sum_{t=1}^T\left(\frac{\sigma_1^{\mathtt{p}} v_t^2\eta_t^2}{\tau^{\mathtt{p}-2}} + \min_{\beta\in\left[0,\frac{1}{2}\right]} \frac{\sigma_1^{\mathtt{p}}(2\bar{G})^{2-\beta_{\mathtt{p}}} v_t^2\eta_t^2}{\tau^{(1-\beta)\mathtt{p}}}\right) \\ &\overset{\ln\frac{3}{\delta}\geq 1}{\leq} \frac{10}{3}\max_{t\in[T]} v_t^2\eta_t^2\rho^2 \ln^2\frac{3}{\delta} + 8\sum_{t=1}^T\left(\frac{\sigma_1^{\mathtt{p}} v_t^2\eta_t^2}{\tau^{\mathtt{p}-2}} + \min_{\beta\in\left[0,\frac{1}{2}\right]} \frac{\sigma_1^{\mathtt{p}}(2\bar{G})^{2-\beta_{\mathtt{p}}} v_t^2\eta_t^2}{\tau^{(1-\beta)\mathtt{p}}}\right).\end{aligned} \tag{67}$$

- Lastly, there is

$$v_t\eta_t \left\|\boldsymbol{d}_t^{\mathrm{b}}\right\| \zeta_t \overset{(65)}{\leq} \frac{\sigma_{\mathtt{s}}\sigma_1^{\mathtt{p}-1} v_t\eta_t\zeta_t}{\tau_t^{\mathtt{p}-1}} + \frac{2^{\mathtt{p}-1}\sigma_1^{\mathtt{p}} G_t v_t\eta_t\zeta_t}{\tau_t^{\mathtt{p}}} \leq \frac{\sigma_{\mathtt{s}}\sigma_1^{\mathtt{p}-1} v_t\eta_t}{\tau^{\mathtt{p}-1}} + \min_{\beta\in\left[0,\frac{1}{2}\right]} \frac{2^{\mathtt{p}-1}\sigma_1^{\mathtt{p}}\bar{G}^{1-\beta_{\mathtt{p}}} v_t\eta_t}{2^{\beta_{\mathtt{p}}}\tau^{(1-\beta)\mathtt{p}}},$$

where the second step is due to

$$\frac{\zeta_t}{\tau_t^{\mathtt{p}-1}} \overset{\tau_t\geq\tau,\mathtt{p}\geq 1}{\leq} \frac{\zeta_t}{\tau^{\mathtt{p}-1}} \overset{\zeta_t\leq 1}{\leq} \frac{1}{\tau^{\mathtt{p}-1}},$$

and, for any $\beta \in \left[0, \frac{1}{2}\right]$,

$$\frac{G_t\zeta_t}{\tau_t^{\mathtt{p}}} \overset{\tau_t\geq\tau,(61)}{\leq} \frac{G_t^{1-\beta_{\mathtt{p}}}\zeta_t}{2^{\beta_{\mathtt{p}}}\tau^{(1-\beta)\mathtt{p}}} \overset{\zeta_t\leq 1,\beta_{\mathtt{p}}\leq 1,(59)}{\leq} \frac{\bar{G}^{1-\beta_{\mathtt{p}}}}{2^{\beta_{\mathtt{p}}}\tau^{(1-\beta)\mathtt{p}}}.$$

Hence, we obtain

$$\sum_{t=1}^T v_t\eta_t \left\|\boldsymbol{d}_t^{\mathrm{b}}\right\| \zeta_t \leq \sum_{t=1}^T \frac{\sigma_{\mathtt{s}}\sigma_1^{\mathtt{p}-1} v_t\eta_t}{\tau^{\mathtt{p}-1}} + \min_{\beta\in\left[0,\frac{1}{2}\right]} \frac{2^{\mathtt{p}-1}\sigma_1^{\mathtt{p}}\bar{G}^{1-\beta_{\mathtt{p}}} v_t\eta_t}{2^{\beta_{\mathtt{p}}}\tau^{(1-\beta)\mathtt{p}}}. \tag{68}$$

Combine (66), (67), and (68) to have with probability at least $1-\delta$,

$$\bar{P}_T^{\mathrm{cvx}} = 8\max_{t\in[T]}\left(\sum_{s=1}^t M_s\zeta_s\right)^2 + 2\sum_{t=1}^T v_t^2\eta_t^2 \left\|\boldsymbol{d}_t^{\mathrm{u}}\right\|^2 \zeta_t^2 + 4\left(\sum_{t=1}^T v_t\eta_t \left\|\boldsymbol{d}_t^{\mathrm{b}}\right\| \zeta_t\right)^2 \leq \bar{U}_T^{\mathrm{cvx}},$$

where

$$\begin{aligned}\bar{U}_T^{\mathrm{cvx}} \triangleq{}& \frac{124}{9}\max_{t\in[T]} v_t^2\eta_t^2\rho^2 \ln^2\frac{3}{\delta} + 128\sum_{t=1}^T\left(\frac{\sigma_{\mathtt{s}}^{\mathtt{p}} v_t^2\eta_t^2\rho}{\tau^{\mathtt{p}-1}} + \min_{\beta\in\left[0,\frac{1}{2}\right]} \frac{2^{\mathtt{p}-1}\sigma_1^{\mathtt{p}}\bar{G}^{2-\beta_{\mathtt{p}}} v_t^2\eta_t^2}{2^{\beta_{\mathtt{p}}}\tau^{(1-\beta)\mathtt{p}}}\right)\ln\frac{3}{\delta} \\ &+ 16\sum_{t=1}^T\left(\frac{\sigma_1^{\mathtt{p}} v_t^2\eta_t^2}{\tau^{\mathtt{p}-2}} + \min_{\beta\in\left[0,\frac{1}{2}\right]} \frac{\sigma_1^{\mathtt{p}}(2\bar{G})^{2-\beta_{\mathtt{p}}} v_t^2\eta_t^2}{\tau^{(1-\beta)\mathtt{p}}}\right) \\ &+ 4\left(\sum_{t=1}^T \frac{\sigma_{\mathtt{s}}\sigma_1^{\mathtt{p}-1} v_t\eta_t}{\tau^{\mathtt{p}-1}} + \min_{\beta\in\left[0,\frac{1}{2}\right]} \frac{2^{\mathtt{p}-1}\sigma_1^{\mathtt{p}}\bar{G}^{1-\beta_{\mathtt{p}}} v_t\eta_t}{2^{\beta_{\mathtt{p}}}\tau^{(1-\beta)\mathtt{p}}}\right)^2 \\ ={}& \mathcal{O}\left(\max_{t\in[T]} v_t^2\eta_t^2\rho^2 \ln^2\frac{3}{\delta} + \sum_{t=1}^T \frac{\sigma_1^{\mathtt{p}} v_t^2\eta_t^2}{\tau^{\mathtt{p}-2}} + \left(\sum_{t=1}^T \frac{\sigma_{\mathtt{s}}\sigma_1^{\mathtt{p}-1} v_t\eta_t}{\tau^{\mathtt{p}-1}} + \min_{\beta\in\left[0,\frac{1}{2}\right]} \frac{\sigma_1^{\mathtt{p}}\bar{G}^{1-\beta_{\mathtt{p}}} v_t\eta_t}{\tau^{(1-\beta)\mathtt{p}}}\right)^2\right),\end{aligned} \tag{69}$$

in which the second equation is by noticing that

$$\begin{aligned}
&\max_{t\in[T]} v_t^2\eta_t^2\rho^2\ln^2\frac{3}{\delta}+\left(\sum_{t=1}^{T}\frac{\sigma_{\mathtt{s}}\sigma_1^{\mathtt{p}-1}v_t\eta_t}{\tau^{\mathtt{p}-1}}+\min_{\beta\in\left[0,\frac{1}{2}\right]}\frac{2^{\mathtt{p}-1}\sigma_1^{\mathtt{p}}\bar{G}^{1-\beta_{\mathtt{p}}}v_t\eta_t}{2^{\beta_{\mathtt{p}}}\tau^{(1-\beta)\mathtt{p}}}\right)^2\\
\overset{(a)}{\geq}&2\left(\max_{t\in[T]} v_t\eta_t\rho\right)\sum_{t=1}^{T}\left(\frac{\sigma_{\mathtt{s}}\sigma_1^{\mathtt{p}-1}v_t\eta_t}{\tau^{\mathtt{p}-1}}+\min_{\beta\in\left[0,\frac{1}{2}\right]}\frac{2^{\mathtt{p}-1}\sigma_1^{\mathtt{p}}\bar{G}^{1-\beta_{\mathtt{p}}}v_t\eta_t}{2^{\beta_{\mathtt{p}}}\tau^{(1-\beta)\mathtt{p}}}\right)\ln\frac{3}{\delta}\\
\geq&2\sum_{t=1}^{T}\left(\frac{\sigma_{\mathtt{s}}\sigma_1^{\mathtt{p}-1}v_t^2\eta_t^2\rho}{\tau^{\mathtt{p}-1}}+\min_{\beta\in\left[0,\frac{1}{2}\right]}\frac{2^{\mathtt{p}-1}\sigma_1^{\mathtt{p}}\bar{G}^{1-\beta_{\mathtt{p}}}v_t^2\eta_t^2\rho}{2^{\beta_{\mathtt{p}}}\tau^{(1-\beta)\mathtt{p}}}\right)\ln\frac{3}{\delta}\\
\overset{(b)}{=}&\Omega\left(\sum_{t=1}^{T}\left(\frac{\sigma_{\mathtt{s}}^{\mathtt{p}}v_t^2\eta_t^2\rho}{\tau^{\mathtt{p}-1}}+\min_{\beta\in\left[0,\frac{1}{2}\right]}\frac{2^{\mathtt{p}-1}\sigma_1^{\mathtt{p}}\bar{G}^{2-\beta_{\mathtt{p}}}v_t^2\eta_t^2}{2^{\beta_{\mathtt{p}}}\tau^{(1-\beta)\mathtt{p}}}\right)\ln\frac{3}{\delta}+\sum_{t=1}^{T}\min_{\beta\in\left[0,\frac{1}{2}\right]}\frac{\sigma_1^{\mathtt{p}}(2\bar{G})^{2-\beta_{\mathtt{p}}}v_t^2\eta_t^2}{\tau^{(1-\beta)\mathtt{p}}}\right),
\end{aligned}$$

where (a) is due to AM-GM inequality and (b) is by $\sigma_1\geq\sigma_{\mathtt{s}}$, $\mathtt{p}\geq 1$, $\rho\overset{(60)}{\geq}2\bar{G}$, and $\ln\frac{3}{\delta}\geq 1$. □

To find a sufficient condition for uniformly upper bounded $\|\boldsymbol{x}_\star-\boldsymbol{x}_t\|$, we employ the idea introduced before in Section 4.

Lemma 7. *Under Assumptions 1, 2 (with $\mu=0$), 4, 5, 6, and the following two conditions:*

1. *η_t is deterministic and satisfies $0<\eta_t\leq\frac{1}{2L}$ for all $t\in[T]$.*
2. *$\tau_t=\max\{2G_t,\tau\}$ for all $t\in[T]$, where $G_t=L\|\boldsymbol{x}_t-\boldsymbol{x}_1\|+G_1$ and $\tau\geq 0$ is a constant.*

`Clipped-SGD` *(Algorithm 1) guarantees that, for any $\delta\in(0,1]$, with probability at least $1-\delta$,*

$$\|\boldsymbol{x}_\star-\boldsymbol{x}_t\|\leq\bar{D},\forall t\in[T+1],$$

where $\bar{D}>0$ is any constant satisfying (72).

Proof. Let $\bar{D}>0$ be any constant. We first upper bound the following term:

$$\bar{I}_T^{\mathrm{cvx}}\triangleq 8\max_{t\in[T]}\left(\sum_{s=1}^{t}N_s\zeta_s\right)^2+2\sum_{t=1}^{T}\eta_t^2\left\|\boldsymbol{d}_t^{\mathrm{u}}\right\|^2\zeta_t^2+4\left(\sum_{t=1}^{T}\eta_t\left\|\boldsymbol{d}_t^{\mathrm{b}}\right\|\zeta_t\right)^2,\tag{70}$$

where $\zeta_t\triangleq\mathbb{1}\left[\|\boldsymbol{x}_\star-\boldsymbol{x}_t\|\leq\bar{D}\right],\forall t\in[T+1]$. Following similar steps of proving (69), we can show that, for any $\delta\in(0,1]$,

$$\begin{aligned}
&\Pr[E]\geq 1-\delta,\text{ where }E\triangleq\left\{\bar{I}_T^{\mathrm{cvx}}\leq\bar{A}_T^{\mathrm{cvx}}\right\}\text{ and}\\
&\bar{A}_T^{\mathrm{cvx}}\triangleq\mathtt{C}\left[\max_{t\in[T]}\eta_t^2\rho^2\ln^2\frac{3}{\delta}+\sum_{t=1}^{T}\sigma_1^{\mathtt{p}}\eta_t^2\tau^{2-\mathtt{p}}+\left(\sum_{t=1}^{T}\frac{\sigma_{\mathtt{s}}\sigma_1^{\mathtt{p}-1}\eta_t}{\tau^{\mathtt{p}-1}}+\min_{\beta\in\left[0,\frac{1}{2}\right]}\frac{\sigma_1^{\mathtt{p}}\bar{G}^{1-\beta_{\mathtt{p}}}\eta_t}{\tau^{(1-\beta)\mathtt{p}}}\right)^2\right],
\end{aligned}\tag{71}$$

in which $\bar{G}=L(\bar{D}+D)+G_1$, $\rho=\max\left\{2\bar{G},\tau\right\}$, and $\mathtt{C}\geq 1$ is a numerical constant.

Next, suppose $\bar{D}$ satisfies the following inequality

$$2D^2+4\bar{A}_T^{\mathrm{cvx}}\leq\bar{D}^2.\tag{72}$$

We will show

$$\Pr\left[\|\boldsymbol{x}_\star-\boldsymbol{x}_t\|\leq\bar{D},\forall t\in[T+1]\right]\geq 1-\delta.$$

Let us define a group of events $E_t\triangleq\left\{\zeta_s=1,\forall s\in[t]\right\},\forall t\in[T+1]$. We claim that there is always

$$E\cap E_t\subseteq\left\{\zeta_{t+1}=1\right\},\forall t\in[T].\tag{73}$$

Note that, by Condition 1 and Lemma 4 ($\omega=0$ now), every sample path satisfies

$$\|\boldsymbol{x}_\star-\boldsymbol{x}_{t+1}\|^2\leq 2D^2+4\left[8\max_{t'\in[t]}\left(\sum_{s=1}^{t'}N_s\right)^2+2\sum_{s=1}^{t}\eta_s^2\left\|\boldsymbol{d}_s^{\mathrm{u}}\right\|^2+4\left(\sum_{s=1}^{t}\eta_s\left\|\boldsymbol{d}_s^{\mathrm{b}}\right\|\right)^2\right],\forall t\in[T].$$

Given a fixed $t \in [T]$ and a sample path in $E \cap E_t$, we know $\zeta_s = 1, \forall s \in [t]$ and $\bar{I}_T^{\mathrm{cvx}} \leq \bar{A}_T^{\mathrm{cvx}}$, implying that

$$\begin{aligned}
\left\|\boldsymbol{x}_\star - \boldsymbol{x}_{t+1}\right\|^2 &\leq 2D^2 + 4\left[8\max_{t'\in[t]}\left(\sum_{s=1}^{t'} N_s\right)^2 + 2\sum_{s=1}^{t}\eta_s^2\left\|\boldsymbol{d}_s^{\mathrm{u}}\right\|^2 + 4\left(\sum_{s=1}^{t}\eta_s\left\|\boldsymbol{d}_s^{\mathrm{b}}\right\|\right)^2\right] \\
&= 2D^2 + 4\left[8\max_{t'\in[t]}\left(\sum_{s=1}^{t'} N_s\zeta_s\right)^2 + 2\sum_{s=1}^{t}\eta_s^2\left\|\boldsymbol{d}_s^{\mathrm{u}}\right\|^2\zeta_s^2 + 4\left(\sum_{s=1}^{t}\eta_s\left\|\boldsymbol{d}_s^{\mathrm{b}}\right\|\zeta_s\right)^2\right] \\
&\leq 2D^2 + 4\left[8\max_{t\in[T]}\left(\sum_{s=1}^{t} N_s\zeta_s\right)^2 + 2\sum_{t=1}^{T}\eta_t^2\left\|\boldsymbol{d}_t^{\mathrm{u}}\right\|^2\zeta_t^2 + 4\left(\sum_{t=1}^{T}\eta_t\left\|\boldsymbol{d}_t^{\mathrm{b}}\right\|\zeta_t\right)^2\right] \\
&\overset{(70)}{=} 2D^2 + 4\bar{I}_T^{\mathrm{cvx}} \leq 2D^2 + 4\bar{A}_T^{\mathrm{cvx}} \overset{(72)}{\leq} \bar{D}^2 \Rightarrow \zeta_{t+1} = 1.
\end{aligned}$$

Therefore, (73) holds for all $t \in [T]$, which implies that

$$E \cap E_{t+1} = E \cap E_t \cap \{\zeta_{t+1} = 1\} = E \cap E_t, \forall t \in [T].$$

Subsequently, we obtain

$$E_{T+1} \supseteq E \cap E_{T+1} = E \cap E_1 = E, \tag{74}$$

where the last step holds because $E_1 = \{\zeta_1 = 1\} = \left\{\|\boldsymbol{x}_\star - \boldsymbol{x}_1\| \leq \bar{D}\right\}$ is the set that contains all possible sample paths, since $\|\boldsymbol{x}_\star - \boldsymbol{x}_1\| = D \leq \sqrt{2}D \overset{(72)}{\leq} \bar{D}$. Finally, we have

$$\Pr\left[\|\boldsymbol{x}_\star - \boldsymbol{x}_t\| \leq \bar{D}, \forall t \in [T+1]\right] = \Pr\left[E_{T+1}\right] \overset{(74)}{\geq} \Pr\left[E\right] \overset{(71)}{\geq} 1 - \delta.$$

□

C.3 Lemmas for Strongly Star-Convex Optimization

In this subsection, we will prove the lemmas used in strongly star-convex optimization, i.e., $\mu > 0$ in Assumption 2.

Before proceeding with the analysis, we need some new notation. Inspired by [23], we introduce an auxiliary sequence $\{\gamma_t\}_{t=1}^T$ defined as follows

$$\gamma_t \triangleq \prod_{s=2}^{t}\frac{1+\mu\eta_{s-1}}{1+\mu\eta_s/2}, \forall t \in [T]. \tag{75}$$

With $\{\gamma_t\}_{t=1}^T$ in hand, we redefine $\{D_t\}_{t=1}^T$ and $\{N_t\}_{t=1}^T$ in the following

$$D_t \triangleq \max_{s\in[t]}\sqrt{\gamma_s(1+\mu\eta_s/2)}\left\|\boldsymbol{x}_\star - \boldsymbol{x}_s\right\|, \tag{76}$$

$$N_t \triangleq \left\langle\sqrt{\frac{\gamma_t}{1+\mu\eta_t/2}}\eta_t\boldsymbol{d}_t^{\mathrm{u}}, \frac{\sqrt{\gamma_t(1+\mu\eta_t/2)}\left(\boldsymbol{x}_\star - \boldsymbol{x}_t\right)}{D_t}\right\rangle. \tag{77}$$

As one can see, these two new definitions degenerate to the old ones in (46) when $\mu = 0$.

Although our goal is to prove last-iterate convergence in terms of function value gap, we still need the following Lemma 8, which provides a bound on $\left\|\boldsymbol{x}_\star - \boldsymbol{x}_{T+1}\right\|^2$.

Lemma 8. *Under Assumptions 1, 2 (with $\mu > 0$), and 3 (resp. 4), for any $T \in \mathbb{N}$ and $\{\eta_t\}_{t=1}^T$ satisfying $0 < \eta_t \leq \frac{\eta}{\mu}, \forall t \in [T]$ (resp. $0 < \eta_t \leq \min\left\{\frac{1}{2L}, \frac{\eta}{\mu}\right\}, \forall t \in [T]$) where $\eta > 0$ is a constant,* `Clipped-SGD` *(Algorithm 1) guarantees*

$$\left\|\boldsymbol{x}_\star - \boldsymbol{x}_{T+1}\right\|^2 \leq \frac{(2+\eta)D^2 + 4I_T^{\mathrm{str}} + 16\sum_{t=1}^{T}\gamma_t\eta_t^2G^2\omega}{\gamma_T},$$

where

$$I_T^{\mathrm{str}} = 4 \max_{t\in[T]} \left(\sum_{s=1}^{t} N_s \right)^2 + 2 \sum_{t=1}^{T} \gamma_t \eta_t^2 \left\| \boldsymbol{d}_t^{\mathrm{u}} \right\|^2 + \frac{1+2\eta}{\mu} \sum_{t=1}^{T} \gamma_t \eta_t \left\| \boldsymbol{d}_t^{\mathrm{b}} \right\|^2,$$

and $\omega = \begin{cases} 1 & \textit{under Assumption 3} \\ 0 & \textit{under Assumption 4} \end{cases}$ *is introduced in Lemma 3.*

Proof. We invoke Lemma 3 with $\boldsymbol{y} = \boldsymbol{x}_\star$ and replace the subscript t with s to have, for any $s \in [T]$,

$$F(\boldsymbol{x}_{s+1}) - F_\star \leq \frac{\left\| \boldsymbol{x}_\star - \boldsymbol{x}_s \right\|^2 - \left\| \boldsymbol{x}_\star - \boldsymbol{x}_{s+1} \right\|^2}{2\eta_s} + \left\langle \boldsymbol{d}_s^{\mathrm{c}}, \boldsymbol{x}_\star - \boldsymbol{x}_s \right\rangle + \eta_s \left\| \boldsymbol{d}_s^{\mathrm{c}} \right\|^2 + 4\eta_s G^2 \omega,$$

where $\omega = \begin{cases} 1 & \text{under Assumption 3} \\ 0 & \text{under Assumption 4} \end{cases}$. Note that $F_\star = F(\boldsymbol{x}_\star) = \inf_{\boldsymbol{x}\in\mathcal{X}} F(\boldsymbol{x})$, which implies that there exists $\nabla F(\boldsymbol{x}_\star) \in \partial F(\boldsymbol{x}_\star)$ such that $\langle \nabla F(\boldsymbol{x}_\star), \boldsymbol{x} - \boldsymbol{x}_\star \rangle \geq 0, \forall \boldsymbol{x} \in \mathcal{X}$. Therefore, we know

$$F(\boldsymbol{x}_{s+1}) - F_\star \overset{\text{Lemma 1}}{\geq} \frac{\mu}{2} \left\| \boldsymbol{x}_{s+1} - \boldsymbol{x}_\star \right\|^2.$$

As such, there is

$$0 \leq \frac{\left\| \boldsymbol{x}_\star - \boldsymbol{x}_s \right\|^2 - (1+\mu\eta_s) \left\| \boldsymbol{x}_\star - \boldsymbol{x}_{s+1} \right\|^2}{2\eta_s} + \left\langle \boldsymbol{d}_s^{\mathrm{c}}, \boldsymbol{x}_\star - \boldsymbol{x}_s \right\rangle + \eta_s \left\| \boldsymbol{d}_s^{\mathrm{c}} \right\|^2 + 4\eta_s G^2 \omega.$$

With the above inequality in hand, the following proof is identical to the analysis for Lemma 7 of [23]. Therefore, we omit all details to save space, but only remind the reader of the following newly defined quantity:

$$I_t^{\mathrm{str}} \triangleq 4 \max_{t'\in[t]} \left(\sum_{s=1}^{t'} N_s \right)^2 + 2 \sum_{s=1}^{t} \gamma_s \eta_s^2 \left\| \boldsymbol{d}_s^{\mathrm{u}} \right\|^2 + \frac{1+2\eta}{\mu} \sum_{s=1}^{t} \gamma_s \eta_s \left\| \boldsymbol{d}_s^{\mathrm{b}} \right\|^2, \forall t \in [T].$$

□

C.3.1 Lipschitz Objectives

In this part, we consider Lipschitz objectives (i.e., Assumption 3). Our goal is to bound the residual term I_T^{str} in Lemma 8. Fortunately, this has been done previously by [23].

Lemma 9 (Lemma 9 of [23])**.** *Under Assumptions 3 and 5, and the following two conditions:*

1. *η_t and τ_t are deterministic for all $t \in [T]$.*
2. *$\tau_t \geq 2G$ for all $t \in [T]$.*

We have for any $\delta \in (0, 1]$, with probability at least $1-\delta$, $I_T^{\mathrm{str}} \leq A_T^{\mathrm{str}}$, where I_T^{str} is introduced in Lemma 8 and A_T^{str} is a constant in the order of

$$\mathcal{O}\left(\max_{t\in[T]} \gamma_t \eta_t^2 \tau_t^2 \ln^2 \frac{3}{\delta} + \sum_{t=1}^{T} \frac{\sigma_1^{\mathtt{p}} \gamma_t \eta_t^2}{\tau_t^{\mathtt{p}-2}} + \sum_{t=1}^{T} \left(\frac{\sigma_{\mathtt{s}}^{\mathtt{p}} \gamma_t \eta_t^2}{\tau_t^{\mathtt{p}-2}} + \frac{\sigma_1^{\mathtt{p}} G^2 \gamma_t \eta_t^2}{\tau_t^{\mathtt{p}}} \right) \ln \frac{3}{\delta} + \sum_{t=1}^{T} \left(\frac{\sigma_{\mathtt{s}}^2 \sigma_1^{2\mathtt{p}-2} \gamma_t \eta_t}{\tau_t^{2\mathtt{p}-2}} + \frac{\sigma_1^{2\mathtt{p}} G^2 \gamma_t \eta_t}{\tau_t^{2\mathtt{p}}} \right) \frac{1+\eta}{\mu} \right).$$

Remark 7. Condition 2 in the original Lemma 9 of [23] is stated in a more general form, i.e., $\tau_t \geq \frac{G}{1-\alpha}, \forall t \in [T]$ for a constant $\alpha \in (0, 1)$. Here, we set $\alpha = \frac{1}{2}$ to accommodate the present paper.

C.3.2 Smooth Objectives

In this part, we consider smooth objectives (i.e., Assumption 4). As discussed before, the difficulty in the smooth case is to bound $\|\nabla f(\boldsymbol{x}_t)\|$ properly. Luckily, on a compact domain (i.e., Assumption 7), this is relatively easy to achieve. However, to maintain generality, we prove the following lemma in a manner analogous to Lemma 6 under the weaker Assumption 6.

Lemma 10. *Under Assumptions 1, 4, 5, 6, and the following two conditions:*

1. *η_t is deterministic for all $t \in [T]$.*
2. *$\tau_t = \max\{2G_t, \tau\}$ holds for all $t \in [T]$, where $G_t = L\|\boldsymbol{x}_t - \boldsymbol{x}_1\| + G_1$ and $\tau \geq 0$ is deterministic.*

We have for any $\delta \in (0,1]$, with probability at least $1-\delta$, $\bar{I}_T^{\mathrm{str}} \leq \bar{A}_T^{\mathrm{str}}$, where

$$\bar{I}_T^{\mathrm{str}} \triangleq 4\max_{t\in[T]}\left(\sum_{s=1}^{t} N_s\zeta_s\right)^2 + 2\sum_{t=1}^{T}\gamma_t\eta_t^2\left\|\boldsymbol{d}_t^{\mathrm{u}}\right\|^2\zeta_t^2 + \frac{1+2\eta}{\mu}\sum_{t=1}^{T}\gamma_t\eta_t\left\|\boldsymbol{d}_t^{\mathrm{b}}\right\|^2\zeta_t^2,$$
$$\zeta_t \triangleq \mathbb{1}\left[\|\boldsymbol{x}_\star - \boldsymbol{x}_t\| \leq \bar{D}\right], \forall t \in [T],$$

$\bar{D} \geq 0$ can be any constant, and $\bar{A}_T^{\mathrm{str}}$ is a constant in the order of

$$\mathcal{O}\left(\max_{t\in[T]}\gamma_t\eta_t^2\rho^2\ln^2\frac{3}{\delta} + \sum_{t=1}^{T}\frac{\sigma_1^{\mathtt{p}}\gamma_t\eta_t^2}{\tau^{\mathtt{p}-2}} + \sum_{t=1}^{T}\left(\frac{\sigma_{\mathtt{s}}^{\mathtt{p}}\gamma_t\eta_t^2\rho}{\tau^{\mathtt{p}-1}} + \min_{\beta\in\left[0,\frac{1}{2}\right]}\frac{\sigma_1^{\mathtt{p}}\bar{G}^{2-\beta_{\mathtt{p}}}\gamma_t\eta_t^2}{\tau^{(1-\beta)\mathtt{p}}}\right)\ln\frac{3}{\delta}\right.$$
$$\left. + \left(\sum_{t=1}^{T}\frac{\sigma_{\mathtt{s}}^2\sigma_1^{2\mathtt{p}-2}\gamma_t\eta_t}{\tau^{2\mathtt{p}-2}} + \min_{\beta\in\left[0,\frac{1}{2}\right]}\frac{\sigma_1^{2\mathtt{p}}\bar{G}^{2-2\beta_{\mathtt{p}}}\gamma_t\eta_t}{\tau^{2(1-\beta)\mathtt{p}}}\right)\frac{1+\eta}{\mu}\right),$$

in which $\bar{G} = L(\bar{D}+D)+G_1$ and $\rho = \max\left\{2\bar{G}, \tau\right\}$.

Proof. Similar to the proof of Lemma 6, let $\bar{G} \triangleq L(\bar{D}+D)+G_1$ and $\rho \triangleq \max\left\{2\bar{G}, \tau\right\}$, then the following inequalities hold for any $t \in [T]$:

$$2\|\nabla f(\boldsymbol{x}_t)\| \leq 2G_t \leq \tau_t, \quad \|\nabla f(\boldsymbol{x}_t)\|\zeta_t \leq \bar{G}, \quad \tau_t\zeta_t \leq \rho. \tag{78}$$

In addition, χ_t in Lemma 2 equals 1 for all $t \in [T]$. Then Lemma 2 and (78) together yield that the following collection of inequalities holds for any $t \in [T]$:

$$\|\boldsymbol{d}_t^{\mathrm{u}}\| \leq 2\tau_t, \tag{79}$$
$$\mathbb{E}\left[\left\|\boldsymbol{d}_t^{\mathrm{u}}\right\|^2 \mid \mathcal{F}_{t-1}\right] \leq \frac{4\sigma_1^{\mathtt{p}}}{\tau_t^{\mathtt{p}-2}}, \tag{80}$$
$$\left\|\mathbb{E}\left[\boldsymbol{d}_t^{\mathrm{u}}\left(\boldsymbol{d}_t^{\mathrm{u}}\right)^\top \mid \mathcal{F}_{t-1}\right]\right\|_{\mathtt{op}} \leq \frac{4\sigma_{\mathtt{s}}^{\mathtt{p}}}{\tau_t^{\mathtt{p}-2}} + \frac{2^{\mathtt{p}+1}\sigma_1^{\mathtt{p}}G_t^2}{\tau_t^{\mathtt{p}}}, \tag{81}$$
$$\left\|\boldsymbol{d}_t^{\mathrm{b}}\right\| \leq \frac{\sigma_{\mathtt{s}}\sigma_1^{\mathtt{p}-1}}{\tau_t^{\mathtt{p}-1}} + \frac{2^{\mathtt{p}-1}\sigma_1^{\mathtt{p}}G_t}{\tau_t^{\mathtt{p}}}. \tag{82}$$

Now, we are ready to bound $\bar{I}_T^{\mathrm{str}}$.

- Similar to (66), we can show with probability at least $1-2\delta/3$,

$$\max_{t\in[T]}\left(\sum_{s=1}^{t} N_s\zeta_s\right)^2 \leq \frac{8}{9}\max_{t\in[T]}\gamma_t\eta_t^2\rho^2\ln^2\frac{3}{\delta} + 16\sum_{t=1}^{T}\left(\frac{\sigma_{\mathtt{s}}^{\mathtt{p}}\gamma_t\eta_t^2\rho}{\tau^{\mathtt{p}-1}} + \min_{\beta\in\left[0,\frac{1}{2}\right]}\frac{2^{\mathtt{p}-1}\sigma_1^{\mathtt{p}}\bar{G}^{2-\beta_{\mathtt{p}}}\gamma_t\eta_t^2}{2^{\beta_{\mathtt{p}}}\tau^{(1-\beta)\mathtt{p}}}\right)\ln\frac{3}{\delta}. \tag{83}$$

- Similar to (67), we can show with probability at least $1-\delta/3$,

$$\sum_{t=1}^{T}\gamma_t\eta_t^2\left\|\boldsymbol{d}_t^{\mathrm{u}}\right\|^2\zeta_t^2\le\frac{10}{3}\max_{t\in[T]}\gamma_t\eta_t^2\rho^2\ln\frac{3}{\delta}+8\sum_{t=1}^{T}\left(\frac{\sigma_1^{\mathtt{p}}\gamma_t\eta_t^2}{\tau^{\mathtt{p}-2}}+\min_{\beta\in\left[0,\frac{1}{2}\right]}\frac{\sigma_1^{\mathtt{p}}(2\bar{G})^{2-\beta_{\mathtt{p}}}\gamma_t\eta_t^2}{\tau^{(1-\beta)\mathtt{p}}}\right)$$
$$\overset{\ln\frac{3}{\delta}\ge1}{\le}\frac{10}{3}\max_{t\in[T]}\gamma_t\eta_t^2\rho^2\ln^2\frac{3}{\delta}+8\sum_{t=1}^{T}\left(\frac{\sigma_1^{\mathtt{p}}\gamma_t\eta_t^2}{\tau^{\mathtt{p}-2}}+\min_{\beta\in\left[0,\frac{1}{2}\right]}\frac{\sigma_1^{\mathtt{p}}(2\bar{G})^{2-\beta_{\mathtt{p}}}\gamma_t\eta_t^2}{\tau^{(1-\beta)\mathtt{p}}}\right). \tag{84}$$

- Similar to (68), we can show

$$\sum_{t=1}^{T}\gamma_t\eta_t\left\|\boldsymbol{d}_t^{\mathrm{b}}\right\|^2\zeta_t^2\le2\sum_{t=1}^{T}\left(\frac{\sigma_{\mathtt{s}}^2\sigma_1^{2\mathtt{p}-2}\gamma_t\eta_t}{\tau^{2\mathtt{p}-2}}+\min_{\beta\in\left[0,\frac{1}{2}\right]}\frac{2^{2\mathtt{p}-2}\sigma_1^{2\mathtt{p}}\bar{G}^{2-2\beta_{\mathtt{p}}}\gamma_t\eta_t}{2^{2\beta_{\mathtt{p}}}\tau^{2(1-\beta)\mathtt{p}}}\right). \tag{85}$$

Combine (83), (84), and (85) to have with probability at least $1-\delta$,

$$\bar{I}_T^{\mathrm{str}}=4\max_{t\in[T]}\left(\sum_{s=1}^{t}N_s\zeta_s\right)^2+2\sum_{t=1}^{T}\gamma_t\eta_t^2\left\|\boldsymbol{d}_t^{\mathrm{u}}\right\|^2\zeta_t^2+\frac{1+2\eta}{\mu}\sum_{t=1}^{T}\gamma_t\eta_t\left\|\boldsymbol{d}_t^{\mathrm{b}}\right\|^2\zeta_t^2\le\bar{A}_T^{\mathrm{str}},$$

where

$$\begin{aligned}\bar{A}_T^{\mathrm{str}}\triangleq&\frac{92}{9}\max_{t\in[T]}\gamma_t\eta_t^2\rho^2\ln^2\frac{3}{\delta}+64\sum_{t=1}^{T}\left(\frac{\sigma_{\mathtt{s}}^{\mathtt{p}}\gamma_t\eta_t^2\rho}{\tau^{\mathtt{p}-1}}+\min_{\beta\in\left[0,\frac{1}{2}\right]}\frac{2^{\mathtt{p}-1}\sigma_1^{\mathtt{p}}\bar{G}^{2-\beta_{\mathtt{p}}}\gamma_t\eta_t^2}{2^{\beta_{\mathtt{p}}}\tau^{(1-\beta)\mathtt{p}}}\right)\ln\frac{3}{\delta}\\&+16\sum_{t=1}^{T}\left(\frac{\sigma_1^{\mathtt{p}}\gamma_t\eta_t^2}{\tau^{\mathtt{p}-2}}+\min_{\beta\in\left[0,\frac{1}{2}\right]}\frac{\sigma_1^{\mathtt{p}}(2\bar{G})^{2-\beta_{\mathtt{p}}}\gamma_t\eta_t^2}{\tau^{(1-\beta)\mathtt{p}}}\right)\\&+2\left(\sum_{t=1}^{T}\frac{\sigma_{\mathtt{s}}^2\sigma_1^{2\mathtt{p}-2}\gamma_t\eta_t}{\tau^{2\mathtt{p}-2}}+\min_{\beta\in\left[0,\frac{1}{2}\right]}\frac{2^{2\mathtt{p}-2}\sigma_1^{2\mathtt{p}}\bar{G}^{2-2\beta_{\mathtt{p}}}\gamma_t\eta_t}{2^{2\beta_{\mathtt{p}}}\tau^{2(1-\beta)\mathtt{p}}}\right)\frac{1+2\eta}{\mu}\\=&\mathcal{O}\left(\max_{t\in[T]}\gamma_t\eta_t^2\rho^2\ln^2\frac{3}{\delta}+\sum_{t=1}^{T}\frac{\sigma_1^{\mathtt{p}}\gamma_t\eta_t^2}{\tau^{\mathtt{p}-2}}+\sum_{t=1}^{T}\left(\frac{\sigma_{\mathtt{s}}^{\mathtt{p}}\gamma_t\eta_t^2\rho}{\tau^{\mathtt{p}-1}}+\min_{\beta\in\left[0,\frac{1}{2}\right]}\frac{\sigma_1^{\mathtt{p}}\bar{G}^{2-\beta_{\mathtt{p}}}\gamma_t\eta_t^2}{\tau^{(1-\beta)\mathtt{p}}}\right)\ln\frac{3}{\delta}\right.\\&\left.+\left(\sum_{t=1}^{T}\frac{\sigma_{\mathtt{s}}^2\sigma_1^{2\mathtt{p}-2}\gamma_t\eta_t}{\tau^{2\mathtt{p}-2}}+\min_{\beta\in\left[0,\frac{1}{2}\right]}\frac{\sigma_1^{2\mathtt{p}}\bar{G}^{2-2\beta_{\mathtt{p}}}\gamma_t\eta_t}{\tau^{2(1-\beta)\mathtt{p}}}\right)\frac{1+\eta}{\mu}\right),\end{aligned}$$

in which the second equation is by $\ln\frac{3}{\delta}\ge1$. □

D Technical Lemmas

In this section, we list some technical lemmas used in the analysis. For results already existing in the literature, we omit their proofs to save space.

First, Lemma 11 is known as Freedman's inequality, a useful tool to obtain a high-probability concentration bound for a real-valued MDS.

Lemma 11 (Freedman's inequality [12])**.** *Suppose $\{X_t\}_{t=1}^T$ is a real-valued MDS adapted to the filtration $\{\mathcal{F}_t\}_{t=0}^T$ satisfying for any $t\in[T]$, $X_t\le b$ and $\mathbb{E}\left[X_t^2\mid\mathcal{F}_{t-1}\right]\le\sigma_t^2$ almost surely, where $b\ge0$ and σ_t^2 are both constant, then for any $\delta\in(0,1]$, there is*

$$\Pr\left[\sum_{s=1}^{t}X_s\le\frac{b}{3}\ln\frac{1}{\delta}+\sqrt{2\sum_{s=1}^{T}\sigma_s^2\ln\frac{1}{\delta}},\forall t\in[T]\right]\ge1-\delta.$$

Next, we introduce another concentration inequality in Lemma 12.

Lemma 12. *Suppose $\{X_t\}_{t=1}^T$ is a sequence of random variables adapted to the filtration $\{\mathcal{F}_t\}_{t=0}^T$ satisfying for any $t \in [T]$, $|X_t| \le b$ and $\mathbb{E}\left[X_t^2 \mid \mathcal{F}_{t-1}\right] \le \sigma_t^2$ almost surely, where $b \ge 0$ is a constant and $\sigma_t^2 \in \mathcal{F}_{t-1}$ is predictable, then for any $\delta \in (0,1]$, there is*

$$\Pr\left[\sum_{t=1}^T X_t^2 \le \frac{5b^2}{6}\ln\frac{1}{\delta} + 2\sum_{t=1}^T \sigma_t^2\right] \ge 1-\delta.$$

Remark 8. At first glance, this inequality seems to exist in the prior literature (e.g., [8, 23, 25, 44]). However, we emphasize that our lemma is in fact stronger than existing ones, as ours allows σ_t^2 to be a random variable rather than a constant assumed in prior work. Although its proof is not difficult, we believe it can provide some fresh insights to the community.

Proof. We first have

$$\sum_{t=1}^T X_t^2 = \sum_{t=1}^T \underbrace{X_t^2 - \mathbb{E}\left[X_t^2 \mid \mathcal{F}_{t-1}\right]}_{\triangleq Y_t} + \sum_{t=1}^T \mathbb{E}\left[X_t^2 \mid \mathcal{F}_{t-1}\right] \le \sum_{t=1}^T Y_t + \sum_{t=1}^T \sigma_t^2. \tag{86}$$

Note that $\{Y_t\}_{t=1}^T$ is a real-valued MDS adapted to the filtration $\{\mathcal{F}_t\}_{t=0}^T$ satisfying

$$Y_t \le X_t^2 \le b^2 \quad \text{and} \quad \mathbb{E}\left[Y_t^2 \mid \mathcal{F}_{t-1}\right] \le \mathbb{E}\left[X_t^4 \mid \mathcal{F}_{t-1}\right] \le b^2\sigma_t^2. \tag{87}$$

Next, let $h(x) \triangleq \begin{cases} \frac{\exp(x)-1-x}{x^2} & x \ne 0 \\ \frac{1}{2} & x = 0 \end{cases}$. It is easy to verify that h is continuous and increasing. Therefore, given $\lambda > 0$, we have for any $t \in [T]$,

$$\begin{aligned} \exp(\lambda Y_t) &= 1 + \lambda Y_t + (\lambda Y_t)^2 h(\lambda Y_t) \overset{(87)}{\le} 1 + \lambda Y_t + (\lambda Y_t)^2 h(\lambda b^2) \\ \Rightarrow \mathbb{E}\left[\exp(\lambda Y_t) \mid \mathcal{F}_{t-1}\right] &\le 1 + \lambda\mathbb{E}\left[Y_t \mid \mathcal{F}_{t-1}\right] + \lambda^2\mathbb{E}\left[Y_t^2 \mid \mathcal{F}_{t-1}\right] h(\lambda b^2) \\ &\overset{(87)}{\le} 1 + \lambda^2 b^2\sigma_t^2 h(\lambda b^2) \le \exp\left(\lambda^2 b^2\sigma_t^2 h(\lambda b^2)\right). \end{aligned} \tag{88}$$

As such, we know

$$\begin{aligned} \Pr\left[\sum_{t=1}^T Y_t - \lambda b^2 h(\lambda b^2)\sum_{t=1}^T \sigma_t^2 \ge \frac{\ln\frac{1}{\delta}}{\lambda}\right] &= \Pr\left[\exp\left(\lambda\sum_{t=1}^T Y_t - \lambda^2 b^2 h(\lambda b^2)\sum_{t=1}^T \sigma_t^2\right) \ge \frac{1}{\delta}\right] \\ &\overset{(a)}{\le} \delta\mathbb{E}\left[\exp\left(\lambda\sum_{t=1}^T Y_t - \lambda^2 b^2 h(\lambda b^2)\sum_{t=1}^T \sigma_t^2\right)\right] \overset{(b)}{\le} \delta, \end{aligned} \tag{89}$$

where (a) is due to Markov's inequality and (b) is by applying (88) recursively.

Finally, by (86) and (89), for any given $\lambda > 0$,

$$\Pr\left[\sum_{t=1}^T X_t^2 \le \frac{\ln\frac{1}{\delta}}{\lambda} + \left(\lambda b^2 h(\lambda b^2) + 1\right)\sum_{t=1}^T \sigma_t^2\right] \ge 1-\delta.$$

We pick $\lambda = \frac{6}{5b^2}$ and use $\frac{6}{5}h(\frac{6}{5}) \le 1$ to conclude. □

Moreover, our analysis relies on the following inequality proposed in [19].

Lemma 13 (Lemma C.2 of [19])**.** *Let $\{a_t\}_{t=1}^T$ and $\{b_t\}_{t=1}^T$ be two sequences in $\mathbb{R}$ such that a_t is nonnegative and nondecreasing, then there is*

$$\left|\sum_{s=1}^t a_s b_s\right| \le 2a_t \max_{t' \in [t]}\left|\sum_{s=1}^{t'} b_s\right|, \forall t \in [T].$$

Finally, we need an algebraic inequality proved in [23]. As stated in [23], Lemma 14 (or the same idea) has been used in previous works such as [19] and [25].

Lemma 14 (Lemma 14 of [23]). *Let $\{a_t\}_{t=1}^{T+1}$, $\{b_t\}_{t=1}^{T}$, and $\{c_t\}_{t=1}^{T+1}$ be three sequences in $\mathbb{R}$ such that b_t is nonnegative and c_t is nondecreasing. If $a_1 \leq 2c_1$ and*

$$a_{t+1} + b_t \leq \frac{\max_{s\in[t]} a_s}{2} + c_{t+1}, \forall t \in [T],$$

then there is

$$a_{t+1} + b_t \leq 2c_{t+1}, \forall t \in [T].$$

Remark 9. Although the above lemma differs slightly from the original one presented in [23], it follows essentially the same proof given in [23].

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims made in the abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: In Section 5, we discuss the limitations of the work.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: For each theoretical result, the paper provides the full set of assumptions and a complete (and correct) proof.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [N/A]

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [N/A]

Justification: The paper does not include experiments requiring code.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [N/A]

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [N/A]

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [N/A]

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper conforms, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: There is no societal impact of the work performed.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: The paper does not use existing assets.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.