HIGH-PROBABILITY BOUNDS FOR THE LAST ITERATE OF CLIPPED SGD

Anonymous authors
Paper under double-blind review

ABSTRACT

We study the problem of minimizing a convex objective when only noisy gradient estimates are available. Assuming that stochastic gradients have finite α-th moments for some $\alpha \in (1, 2]$, we show - for the first time - that the last iterate of clipped stochastic gradient descent (Clipped-SGD) converges with high probability at rate $1/K^{(2\alpha-2)/(3\alpha)}$ on smooth objectives, with only polylogarithmic dependence on the confidence parameter. We complement our theoretical results with empirical evidence that supports and illustrates these findings.

1 INTRODUCTION

Stochastic first-order optimization methods such as SGD (Robbins & Monro, 1951), Adam (Kingma & Ba, 2014), and their numerous variants are central to the training of modern machine learning models. In practice, these algorithms are almost always combined with additional techniques that enhance stability and performance. One such technique – *gradient clipping* (Pascanu et al., 2013) – has become a standard component in the training of large language models (LLMs) (Devlin et al., 2019; Brown et al., 2020; Fedus et al., 2022; Touvron et al., 2023).

Originally introduced to address the problem of exploding gradients in recurrent neural networks, gradient clipping has since proven valuable well beyond this initial motivation. It has been shown to improve convergence under generalized smoothness conditions (Zhang et al., 2020a), to provide robustness against heavy-tailed noise where gradient variance can be unbounded (Zhang et al., 2020b), and to enable *strong high-probability* convergence guarantees (Gorbunov et al., 2020; Cutkosky & Mehta, 2021; Parletta et al., 2024; Sadiev et al., 2023; Nguyen et al., 2023). Nevertheless, in the case of *convex problems*, most existing theoretical results for clipped methods (including Clipped-SGD) analyze the behavior of *averaged iterates*, while the practically more relevant *last iterate* remains largely unexplored.

Contributions. In this work, we close a long-standing gap in the theory of clipped stochastic gradient descent by analyzing the *last iterate* under heavy-tailed noise. Our main contributions are:

- **First high-probability last-iterate guarantees for Clipped-SGD.** We establish the first high-probability convergence rate for the *last iterate* of Clipped-SGD on convex smooth objectives. Assuming stochastic gradients have finite α-th moments with $\alpha \in (1, 2]$, we prove that after K iterations the method achieves an error of at most

$$\mathcal{O}\left(\frac{\text{polylog}(K/\delta)}{K^{2(\alpha-1)/3\alpha}}\right)$$

  with failure probability at most $\delta \in (0, 1)$. In the special case of $\alpha = 2$, this results in a polynomial gap compared to the best-known in-expectation rate of $1/\sqrt{K}$, where clipping is unnecessary. Crucially, our result covers the full spectrum of heavy-tailed noise and provides high-probability guarantees for a single run – significantly stronger than standard in-expectation bounds.
- **General analysis of step-size and clipping schedules.** We develop a unified analysis for polynomially decaying step-sizes and varying clipping levels, bounding the optimization error as a function of these schedules. This yields principled guidelines for tuning and

identifies optimal exponents, while avoiding restrictive assumptions such as a bounded optimization domain.

- **Any-time parameter choices.** Our parameter selection is *horizon-free*: it does not require prior knowledge of the number of iterations and remains valid in streaming or indefinite-training scenarios, where restarting schemes are impractical. Our results hold without reliance on large minibatches, making them applicable in resource-constrained settings.

We complement our result with empirical evidence supporting the advantage of the last iterate over the average.

2 Notation and Problem Setup

In this section, we introduce the main notation and discuss the assumptions used in the analysis.

Notation. The norm $\|x\| \coloneqq \sqrt{\langle x, x\rangle}$ denotes the Euclidean norm in $\mathbb{R}^d$. $\mathbb{E}_\xi[\,\cdot\,]$ denotes the expectation w.r.t. the random variable ξ. We also denote the clipping operator as $\mathrm{clip}(x, \lambda) \coloneqq \min\left\{1, \frac{\lambda}{\|x\|}\right\} x$. The initial distance, i.e., the distance between the starting point x_0 and a solution[1] x^*, we denote as $R_0 \coloneqq \|x_0 - x^*\|$.

Problem. We study the following problem:

$$\min_{x\in\mathbb{R}^d} f(x), \tag{1}$$

under the following standard hypothesis.

Assumption 1 (Convexity)**.** *The differentiable function f is convex, i.e.,*

$$f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle \quad \forall\, x, y \in \mathbb{R}^d.$$

In addition to convexity, we assume that f is L-smooth.

Assumption 2 (Smoothness)**.** *The differentiable function f is L-smooth, i.e.,*

$$f(y) \leq f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|y - x\|^2 \quad \forall\, x, y \in \mathbb{R}^d.$$

Finally, although the optimizer does not have direct access to f, we assume access to a stochastic gradient oracle $\nabla f(x, \xi)$ satisfying the following condition.

Assumption 3 (Stochastic oracle)**.** *The stochastic oracle $\nabla f(x, \xi)$ is unbiased and have bounded α-th central moment with $\alpha \in (1, 2]$, i.e.*

$$\mathbb{E}\left[\nabla f(x, \xi)\right] = \nabla f(x); \qquad \mathbb{E}\left[\|\nabla f(x, \xi) - \nabla f(x)\|^\alpha\right] \leq \sigma^\alpha.$$

This assumption was introduced by Nemirovskij & Yudin (1983) and later rediscovered by Zhang et al. (2020b), after which it has become standard in the analysis of stochastic methods under heavy-tailed noise. For problem (1), we study Clipped-SGD with time-varying stepsize γ_k and clipping level λ_k:

$$x_{k+1} = x_k - \gamma_k \cdot \mathrm{clip}(\nabla f_{\xi_k}(x_k), \lambda_k),$$

where $\nabla f_{\xi_k}(x_k) = \nabla f(x_k, \xi_k)$ is the stochastic gradient sampled independently of the past.

Stochastic Optimization. The above problem encompasses, as a special case, the classical stochastic optimization problem:

$$\min_{x\in\mathbb{R}^d} \Big[f(x) := \mathbb{E}_{\xi\sim\mathcal{D}}\left[\ell(x, \xi)\right] \Big], \tag{2}$$

where ℓ is the loss function, x are the model parameters and ξ represents the randomness due to data sampling from the unknown distribution $\mathcal{D}$. In this setting, the stochastic gradient can be computed

[1] Our results hold for any solution of the considered problem.

from the sample ξ as $\nabla f(x,\xi) = \nabla \ell(x,\xi)$. Note that, when $\xi = (Z,Y) \in \mathbb{R}^d \times \mathbb{R}$, Equation (2) also covers the statical supervise learning problem. Another important special case of Equation (2) is the finite-sum setting,

$$f(x) = \sum_{i=1}^{n} f_i(x).$$

which underlies many training procedures in machine learning.

High-probability convergence guarantees. For an iterative sequence $x_k{}_{k=0}^{K}$ (produced by some stochastic method) and a target criterion $C(\{x_k\}_{k=0}^{K})$, the standard goal is to ensure

$$\mathbb{E}\left[C\left(\{x_k\}_{k=0}^{K}\right)\right] \leq \varepsilon.$$

Such in-expectation bounds describe average behavior, but they do not capture the variability of the random process. *High-probability bounds*, by contrast, guarantee that the performance criterion is met with confidence at least $1-\delta$, i.e.,

$$\mathbb{P}\left\{C\left(\{x_k\}_{k=0}^{K}\right) \leq \varepsilon\right\} \geq 1-\delta.$$

thereby directly limiting the probability of unfavorable deviations.

Although one can obtain such bounds from expectation guarantees via Markov's inequality, this typically yields rates with an undesirable polynomial $1/\delta$ dependence. Modern approaches instead achieve bounds with only *polylogarithmic* dependence on $1/\delta$, which greatly improves reliability and reduces the number of iterations required to reach a target confidence level. In other words, the goal of high-probability convergence analysis is to establish convergence rates of the same order as the optimal in-expectation guarantees, with only a minimal dependence on the confidence parameter δ, ideally $\mathcal{O}(\sqrt{\log(1/\delta)})$.

3 Related Works

The literature on SGD and Clipped-SGD is vast and multifaceted, and a comprehensive survey is beyond the scope of this work. In what follows, we focus only on the closely related works.

In-expectation convergence bounds. Early studies on SGD for smooth and non-smooth (but Lipschitz) objectives investigated convergence in expectation under finite-variance noise (Nemirovskij & Yudin, 1983; Nemirovski et al., 2009; Ghadimi & Lan, 2013b;a). In this setting, the average iterate of SGD achieves a rate of $\widetilde{\mathcal{O}}(1/\sqrt{K})$[2], which is known to be optimal (Agarwal et al., 2012). Similar optimal rates for the last iterate in the non-smooth case were established in (Shamir & Zhang, 2013; Jain et al., 2021), while for smooth objectives the best known rate remained $\mathcal{O}(1/K^{1/3})$ (Moulines & Bach, 2011) for a long time, until (Liu & Zhou, 2024) — building on ideas from (Zamani & Glineur, 2023) — proved the optimal rate $\mathcal{O}(1/\sqrt{K})$, thereby unifying the analysis of both smooth and non-smooth cases.

Under the more general bounded $\alpha \in (1,2]$ moments model considered in this work, Zhang et al. (2020b) show that plain SGD does not converge (in terms[3] of $\mathbb{E}[\|x_k - x^*\|^2]$, and prove in-expectation convergence bounds for Clipped-SGD for non-convex smooth and strongly convex problems with bounded gradients. Vural et al. (2022) derive average-iterate $\mathcal{O}(1/K^{(\alpha-1)/\alpha})$ bound for Stochastic Mirror Descent (SMD) over convex Lipschitz objectives and show that it is optimal. In the same non-smooth setting, Parletta et al. (2025) show that the last iterate of Clipped-SGD enjoys the same optimal rate. Moreover, for strongly convex objectives, Jakovetić et al. (2023) investigate the last-iterate convergence of a general class of robust SGD variants assuming only $\alpha = 1$. However, these results require additional assumptions, such as symmetry of the noise distribution and constraints on its *effective dimension*. To the best of our knowledge, there exist no *last-iterate* in-expectation convergence bounds for Clipped-SGD in the case of convex L-smooth problems with the stochastic oracle satisfying Assumption 3.

[2]The $\widetilde{\mathcal{O}}$ hides poly-logarithmic factors.

[3]In the case of non-smooth problems, Fatkhullin et al. (2025) prove $\mathcal{O}(1/K^{(\alpha-1)/\alpha})$ average-iterate convergence rate of SGD.

Table 1: Comparison of the state-of-the-art *non-accelerated* in-expectation and high-probability convergence results for SGD/Clipped-SGD-like methods applied to *smooth convex* problems.

Reference	Convergence type	Iterate	Stochasticity	Rate
(Ghadimi & Lan, 2013b)	In-expectation	Average	As. 3, $\alpha = 2$	$\mathcal{O}\left(\frac{LR_0^2}{K} + \frac{R_0\sigma}{\sqrt{K}}\right)$
(Taylor & Bach, 2019)	In-expectation	Last	As. 3, $\alpha = 2$	$\mathcal{O}\left(\frac{LR_0^2+\sigma^2}{K^{1/3}}\right)$
(Liu & Zhou, 2024)	In-expectation	Last	As. 3, $\alpha \in (1, 2]$	$\mathcal{O}\left(\frac{LR_0^2}{K^{2(\alpha-1)/\alpha}} + \frac{R_0\sigma}{K^{(\alpha-1)/\alpha}}\right)$
(Ghadimi & Lan, 2013b)	High probability	Average	Sub-Gaussian	$\mathcal{O}\left(\frac{LR_0^2}{K} + \frac{R_0\sigma}{\sqrt{K}}\right)$
(Liu & Zhou, 2024)	High probability	Last	Sub-Weibull	$\mathcal{O}\left(\frac{LR_0^2}{K} + \frac{R_0\sigma}{\sqrt{K}}\right)$
(Sadiev et al., 2023)	High probability	Average	As. 3, $\alpha \in (1, 2]$[(1)]	$\widetilde{\mathcal{O}}\left(\frac{LR_0^2}{K} + \frac{R_0\sigma}{K^{(\alpha-1)/\alpha}}\right)$
(Nguyen et al., 2023)	High probability	Average	As. 3, $\alpha \in (1, 2]$	$\widetilde{\mathcal{O}}\left(\frac{LR_0^2}{K} + \frac{R_0\sigma}{K^{(\alpha-1)/\alpha}}\right)$[(2)]
This work	High probability	Last	As. 3, $\alpha \in (1, 2]$	$\widetilde{\mathcal{O}}\left(\frac{LR_0^2}{K} + \frac{B}{K^{2(\alpha-1)/3\alpha}}\right)$[(3)]

[(1)] Sadiev et al. (2023) make all assumptions only on a ball centered at x^* with radius $\sim R_0$.
[(2)] The rate from Nguyen et al. (2023) has better logarithmic factor than the one from Sadiev et al. (2023).
[(3)] $B \coloneqq \max\left\{R_0\sigma, L^{(\alpha-1)/(3\alpha-1)}R_0^{(4\alpha-2)/(3\alpha-1)}\sigma^{2\alpha/(3\alpha-1)}, L^{1/3}R_0^{4/3}\sigma^{2/3}\right\}$.

High-probability convergence bounds. The first high-probability results for SGD were established under sub-Gaussian noise assumptions (Nemirovski et al., 2009; Harvey et al., 2019), which are considerably stronger than those considered in this work. Moreover, these guarantees apply only to the average iterate. In this setting and for non-smooth objectives, Liu et al. (2023) showed that the average iterate of Stochastic Mirror Descent (and hence of SGD) converges at the optimal rate with only a modest confidence overhead of $\mathcal{O}(\sqrt{\log(1/\delta)})$. Similar last-iterate guarantees (including for smooth objectives) were later obtained by Eldowa & Paudice (2024) and Liu & Zhou (2024). Both works also relaxed the sub-Gaussian assumption to the broader class of sub-Weibull tails (Vladimirova et al., 2020), which were further explored in non-convex settings by Madden et al. (2024). While more general, these tail models still imply the existence of moments of all orders.

In contrast, under the heavy-tailed noise model studied here, several works (Nazin et al., 2019; Gorbunov et al., 2020; Nguyen et al., 2023; Sadiev et al., 2023; Liu & Zhou, 2024; Gorbunov et al., 2024; Parletta et al., 2024; 2025) have shown that clipping yields convergence of the average iterate at rate $\widetilde{\mathcal{O}}(1/K^{(\alpha-1)/\alpha})$ for convex L-smooth objectives. To the best of our knowledge, results for the last iterate in this regime require additional structural assumptions such as strong convexity or the PL condition (Sadiev et al., 2023). Our work closes this gap by establishing, for the first time, high-probability convergence rates for the last iterate of SGD on convex smooth objectives. Finally, refinements of the $\alpha \in (1, 2]$ model have been proposed in Puchkin et al. (2024), who demonstrate that in certain cases it is possible to surpass the $\widetilde{\mathcal{O}}(1/K^{(\alpha-1)/\alpha})$ barrier.

4 Main Results

In this section we state our high-probability last-iterate guarantee for Clipped-SGD and explain the ideas behind its proof.

Theorem 1. *Suppose that Assumptions 1, 2 and 3 hold. Then, if we choose*

$$\gamma_k = \min\left\{\frac{1}{1024L\ln^3\left(\frac{6(k+1)^2}{\delta}\right)}, \frac{p(\Phi_0, L, \sigma)}{256\cdot 4^{1/\alpha}(k+1)^\beta \ln^3\left(\frac{6(k+1)^2}{\delta}\right)}\right\},$$

$$\lambda_k = \frac{\sqrt{\Phi_0/C}}{256\sqrt{b_k}\gamma_k \ln^{5/2}\left(\frac{6(k+1)^2}{\delta}\right)},$$

where

$$b_k = \begin{cases} d_k, & t > 0; \\ 1, & t = 0; \end{cases} \qquad p(\Phi_0, L, \sigma) = \min\left\{ \frac{\sqrt{\Phi_0/C}}{\sigma}, \frac{\sqrt{\Phi_0/C}^{\frac{2\alpha}{3\alpha-1}}}{C^{\frac{\alpha-1}{3\alpha-1}}\sigma^{\frac{2\alpha}{3\alpha-1}}}, \frac{\sqrt{\Phi_0/C}^{\frac{2}{3}}}{C^{\frac{1}{3}}\sigma^{\frac{2}{3}}} \right\},$$

parameters β *satisfies*

$$\beta \geq \frac{2+\alpha}{3\alpha},$$

and

$$C = \max\left\{ L, \frac{1}{4^{1-1/\alpha} p(\Phi_0, L, \sigma)} \right\},$$

then, after K *iterations of* Clipped-SGD*, we have that*

$$f(x_K) - f^* = \mathcal{O}\left(\frac{LR_0^2 \log^4\left(\frac{6(K+1)^2}{\delta}\right)}{K} + \frac{\max\left\{ R_0\sigma, L^{\frac{\alpha-1}{3\alpha-1}} R_0^{\frac{4\alpha-2}{3\alpha-1}} \sigma^{\frac{2\alpha}{3\alpha-1}}, L^{\frac{1}{3}} R_0^{\frac{4}{3}} \sigma^{\frac{2}{3}} \right\} \log^4\left(\frac{6(K+1)^2}{\delta}\right)}{K^{1-\beta}} \right)$$

hold with probability at least $1 - \delta \sum_{t=1}^{K} \frac{1}{t^2}$.

Corollary 1. Let the conditions of Theorem 1 hold. Then, if we choose β in the optimal way, i.e. $\beta = \frac{2+\alpha}{3\alpha}$, we derive that

$$f(x_K) - f^* = \tilde{\mathcal{O}}\left(\frac{LR_0^2}{K} + \frac{\max\left\{ R_0\sigma, L^{\frac{\alpha-1}{3\alpha-1}} R_0^{\frac{4\alpha-2}{3\alpha-1}} \sigma^{\frac{2\alpha}{3\alpha-1}}, L^{\frac{1}{3}} R_0^{\frac{4}{3}} \sigma^{\frac{2}{3}} \right\}}{K^{\frac{2\alpha-2}{3\alpha}}} \right)$$

holds with probability at least $1 - 2\delta$. Here $\tilde{\mathcal{O}}(\cdot)$ denotes polylogarithmic dependency.

4.1 Proof Sketch and Technical Novelties

We now outline the analysis and highlight the three key innovations.

Potential-based *high-probability* convergence proof. We analyze the method using the following potential:

$$\Phi_k = d_k(f(x_k) - f^*) + \frac{C}{2}\|x_k - x^*\|^2, \quad d_{k+1} = d_k + \gamma_k C, \; d_0 = 0,$$

where $C \geq L$ plays the role of an *effective smoothness* constant that absorbs the stochastic terms appearing in the high-probability bounds. In the special case of $C = L$, this potential reduces to the one proposed by Bansal & Gupta (2017) to study the convergence of Gradient Descent. Moreover, Taylor & Bach (2019) consider Φ_k with $C = L$ to derive the last-iterate *in-expectation* convergence rate (see Table 1). Since we consider high-probability convergence, our descent lemma differs from the one derived by Taylor & Bach (2019) and yields

$$\Phi_K \leq \Phi_0 - \sum_{k=0}^{K-1} \gamma_k C \langle x_k - x^*, \theta_k \rangle - \sum_{k=0}^{K-1} (\gamma_k d_{k+1} - 2e_k) \langle \nabla f(x_k), \theta_k \rangle + \sum_{k=0}^{K-1} e_k \|\theta_k\|^2, \quad (3)$$

with $e_k \coloneqq \frac{(Ld_{k+1}+C)\gamma_k^2}{2}$ and $\theta_k \coloneqq g_k - \nabla f(x_k)$, $g_k \coloneqq \mathrm{clip}(\nabla f_{\xi_k}(x_k), \lambda_k)$. We control the three martingale-type sums produced by the right-hand side via Bernstein/Freedman inequalities with a time-varying failure budget $\delta_t \sim 1/t^2$ and a union bound over t.

Clipping level $\lambda_k \sim 1/\sqrt{b_k}$. A central technical choice is

$$\boxed{\lambda_k = \frac{\sqrt{\Phi_0/C}}{256\sqrt{b_k}\,\gamma_k\,\ln^{5/2}\left(\frac{6(k+1)^2}{\delta}\right)}} \quad \text{with} \quad b_k = d_k\,(k>0),$$

rather than proportional to $1/\gamma_k$ like in the existing average-iterate convergence bounds, e.g., (Sadiev et al., 2023; Nguyen et al., 2023). The additional division by $\sqrt{b_k}$ comes from the following observation: since we prove by induction that $f(x_k) - f^*$ decreases as $\mathcal{O}(1/b_k)$ with high probability, we also get that $\|\nabla f(x_k)\| \leq \sqrt{2L(f(x_k)-f^*)} = \mathcal{O}(1/\sqrt{b_k})$. Therefore, we can better balance the bias/variance terms produced by clipping under Assumption 3 if we select $\lambda_k \sim 1/\sqrt{b_k}$ as well.

Horizon-agnostic schedules and $\log$-factors. Both γ_k and λ_k are any-time, i.e., no prior knowledge of K is assumed. This influences two aspects of the proof:

- Similarly to the prior high-probability analysis of Clipped-SGD (Gorbunov et al., 2020; Sadiev et al., 2023) for smooth convex objectives, we bound the sums from (3) for each $K > 0$ with high probability and then apply the union bound for estimating the probability of the "good" event E_K. However, since the horizon is unknown, we cannot select the failure probability at each step as δ/K for each $k = 0, \ldots, K-1$. Instead, the failure probability for step k is upper bounded by δ/k^2 in our proof. The choice $\delta_k \sim 1/k^2$ (and the resulting $\sum_k \delta_k \leq \delta$) introduces at most polylogarithmic dependence on $1/\delta$, while keeping the schedules horizon-free.
- Moreover, due to the horizon independence of the parameters, the derived upper bound for Φ_K contains a logarithmic factor $\sim \ln(6(K+1)^2/\delta)$. This leads to the $\ln^3$ and $\ln^{5/2}$ exponents in γ_k and λ_k, respectively. We refer to Appendix B.3 for further details.

Remark 1. Using C in place of L in the potential allows us to bound stochastic terms through (R_0, σ) as required by the high-probability analysis, while seamlessly recovering the deterministic case $C = L$ when $\sigma = 0$.

Remark 2. When $\alpha \to 1$, then our result shows convergence to some (finite) neighborhood, which is well-aligned with existing average-iterate results (see Table 1) and the lower bound for the Lipschitz convex case (Vural et al., 2022).

5 Experiments

In this section, we present the results of numerical simulations showing the practical advantages of the last iterate over the average. We consider the problem of minimizing a convex and smooth function $f\colon \mathbb{R}^d \to \mathbb{R}$ from noisy estimates $\widehat{\nabla} f(x)$ of its gradients. In all experiments, we run Clipped-SGD with the step-size and clipping level schedules suggested by the theory, optimizing the constants via a grid-search procedure. Finally we report the performance of both the average and the last iterate in terms of the 0.95-percentile of the function values across 1000 repetitions.

Corrupted gradients. We set $\widehat{\nabla} f(x) = \nabla f(x) + N$, where N is a random vector with components sampled i.i.d. from a Pareto distribution rescaled and reshaped so that it satisfies $\mathbb{E}[\widehat{\nabla} f(x)] = \nabla f(x)$, $\mathbb{E}[\|\widehat{\nabla} f(x) - \nabla f(x)\|^2] \leq 1$, and all moments of order greater than 2.001 are infinite, which closely matches our assumption with $\alpha = 2$. We set $d = 100$ and consider two cases: first, for a fixed unit vector a we let $f(x) = \ln(1 + \exp(\langle x, a\rangle))$; second, we also consider $f(x) = (1/2)\|x\|^2$. Notice that both objectives are smooth (and convex), but the second is also strongly convex. The results are shown in the left and central plots of Figure 1, where it is possible to see that the last iterate performs better than the average.

Statistical learning. We also consider the following statistical learning problem, in which we aim to minimize

$$f(x) = \mathbb{E}_{(Z,Y)\sim\mu}\left[\ln\left(1 + \exp(-Y\langle x, Z\rangle)\right)\right]$$

using only sampling access to $\mu = \mu_Z \cdot \mu_{Y|Z}$. We set $d = 10$ and take μ_Z to be an isotropic distribution with components sampled from a Student-t distribution with 2.001 degrees of freedom,

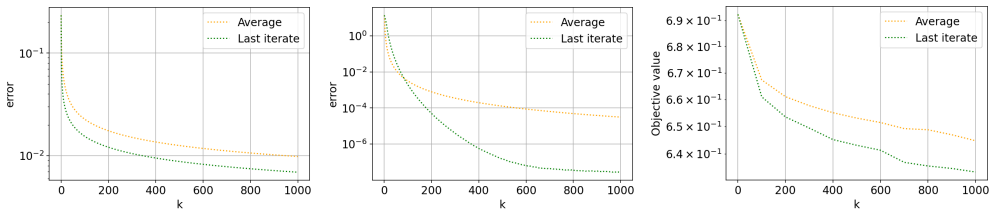

Figure 1: Experimental results: (left) $f(x) = \ln(1 + \exp(\langle x, a\rangle))$, (center) $f(x) = \frac{1}{2}\|x\|^2$, and (right) $f(x) = \mathbb{E}_{(Z,Y)\sim\mu}[\ln(1 + \exp(-Y\langle x, Z\rangle))]$. The x-axis shows the iteration counter k, while the y-axis (in logarithmic scale) reports the objective value $f(z_k)$ at the iterate z_k. Note that in the left and center plots, the y-axis corresponds to the optimization error since $\inf_{x\in\mathbb{R}^d} f(x) = 0$.

so that each component has variance $1/d$ but infinite moments of higher order. Moreover, we set $\mu_{Y|Z=z} = \text{Ber}(p(z))$ over $\{\pm 1\}$ with $p(z) = \text{sigmoid}(\langle w, z\rangle)$. We use $\hat{\nabla} f(x) = \nabla_x \ln(1 + \exp(-Y\langle x, Z\rangle))$ for $(Z, Y) \sim \mu$ and note that $\mathbb{E}[\hat{\nabla} f(x)] = \nabla f(x)$. We estimate $f(z)$ at a given point z via the Median-of-Means[4] estimator with 10^4 samples from μ. The results are shown in the right plot of Figure 1, where it is possible to see that the last iterate performs better than the average.

Discussion. For fairness, both the average and the last iterate are evaluated under the same step-size and clipping level schedules: namely, those for which our theory guarantees convergence of the last iterate. We emphasize that the optimal schedules for the average iterate differ from these ones. In additional experiments (see appendix), we observed that the last iterate actually performs even better when the average iterate is run under its own optimal schedule. This suggests that our current $\widetilde{\mathcal{O}}(1/K^{1/3})$ bound for the last iterate may be an artifact of the analysis, and that there might exist a schedule (possibly the one already known to be optimal for the average iterate) that makes the last iterate achieve the optimal rate as well. A rigorous proof of this conjecture is left to future work.

6 Conclusion

We presented the first high-probability last-iterate guarantees for Clipped-SGD on convex L-smooth objectives under heavy-tailed noise with finite α-th moments, $\alpha \in (1, 2]$. Our analysis is based on a potential function tailored to high-probability control, a new clipping schedule that scales as $1/(\sqrt{b_k}\gamma_k)$, and horizon-agnostic parameter choices. These ingredients yield a rate of $\widetilde{\mathcal{O}}(1/K^{2(\alpha-1)/3\alpha})$ for the last iterate with only polylogarithmic dependence on $1/\delta$. Empirically, we observe a clear advantage of the last iterate over the average under heavy-tailed perturbations.

Limitations and future work. The rate at $\alpha = 2$ leaves a polynomial gap from the $\widetilde{\mathcal{O}}(1/\sqrt{K})$ expectation benchmark (Liu & Zhou, 2024); tightening the last-iterate high-probability rate is a compelling direction. Extending the theory to (L_0, L_1)-smooth objectives, which is done for the *average iterate* by Gaash et al. (2025) under sub-Gaussian noise assumption and by Chezhegov et al. (2025) under Assumption 3, without losing horizon-freeness, is another natural next step.

References

Alekh Agarwal, Peter L. Bartlett, Pradeep Ravikumar, and Martin J. Wainwright. Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization. *IEEE Transactions on Information Theory*, 58(5):3235–3249, 2012. (Cited on page 3)

Nikhil Bansal and Anupam Gupta. Potential-function proofs for first-order methods. *arXiv preprint arXiv:1712.04581*, 2017. (Cited on page 5)

[4]The use of Median-of-Means ensures an estimation error of the order $\sqrt{\frac{\log(1/\beta)}{m}}$ with confidence $1 - \beta$ over m samples, which is exponentially better than the $\sqrt{\frac{1}{\beta \cdot m}}$ exhibited by the standard Monte-Carlo estimate (Lugosi & Mendelson, 2019).

George Bennett. Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association*, 57(297):33–45, 1962. (Cited on page 11)

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. (Cited on page 1)

Savelii Chezhegov, Aleksandr Beznosikov, Samuel Horváth, and Eduard Gorbunov. Convergence of clipped-sgd for convex (L_0, L_1)-smooth optimization with heavy-tailed noise. *arXiv preprint arXiv:2505.20817*, 2025. (Cited on page 7)

Ashok Cutkosky and Harsh Mehta. High-probability bounds for non-convex stochastic optimization with heavy tails. *Advances in Neural Information Processing Systems*, 34:4883–4895, 2021. (Cited on page 1)

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, 2019. (Cited on page 1)

Kacha Dzhaparidze and JH Van Zanten. On bernstein-type inequalities for martingales. *Stochastic processes and their applications*, 93(1):109–117, 2001. (Cited on page 11)

Khaled Eldowa and Andrea Paudice. General tail bounds for non-smooth stochastic mirror descent. In *Proceedings of the 27-th International Conference on Artificial Intelligence and Statistics*, pp. 3205–3213, 2024. (Cited on page 4)

Ilyas Fatkhullin, Florian Hübler, and Guanghui Lan. Can sgd handle heavy-tailed noise? *arXiv preprint arXiv:2508.04860*, 2025. (Cited on page 3)

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022. (Cited on page 1)

David A Freedman et al. On tail probabilities for martingales. *the Annals of Probability*, 3(1): 100–118, 1975. (Cited on page 11)

Ofir Gaash, Kfir Yehuda Levy, and Yair Carmon. Convergence of clipped sgd on convex (L_0, L_1)-smooth functions. *arXiv preprint arXiv:2502.16492*, 2025. (Cited on page 7)

Saeed Ghadimi and Guanghui Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization, ii: shrinking procedures and optimal algorithms. *SIAM Journal on Optimization*, 23(4):2061–2089, 2013a. (Cited on page 3)

Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013b. (Cited on pages 3 and 4)

Eduard Gorbunov, Marina Danilova, and Alexander Gasnikov. Stochastic optimization with heavy-tailed noise via accelerated gradient clipping. In *Proceedings of the 34th Annual Conference on Neural Information Processing Systems*, pp. 15042–15053, 2020. (Cited on pages 1, 4, and 6)

Eduard Gorbunov, Marina Danilova, Innokentiy Shibaev, Pavel E. Dvurechensky, and Alexander V. Gasnikov. High-probability complexity bounds for non-smooth stochastic convex optimization with heavy-tailed noise. *J. Optim. Theory Appl.*, 203(3):2679–2738, 2024. (Cited on page 4)

Nicholas JA Harvey, Christopher Liaw, and Sikander Randhawa. Simple and optimal high-probability bounds for strongly-convex stochastic gradient descent. *arXiv preprint arXiv:1909.00843*, 2019. (Cited on page 4)

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. (Cited on page 3)

Dusan Jakovetić, Dragana Bajović, Anit Kumar Sahu, Soummya Kar, Nemanja Milosević, and Dusan Stamenković. Nonlinear gradient mappings and stochastic optimization: A general framework with applications to heavy-tail noise. *SIAM Journal on Optimization*, 33(2):394–423, 2023. (Cited on page 3)

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. (Cited on page 1)

Zijian Liu and Zhengyuan Zhou. Revisiting the last-iterate convergence of stochastic gradient methods. In *Proceedings of the 12-th International Conference on Learning Representations (to appear)*, 2024. (Cited on pages 3, 4, and 7)

Zijian Liu, Ta Duy Nguyen, Thien Hang Nguyen, Alina Ene, and Huy Nguyen. High probability convergence of stochastic gradient methods. In *Proceedings of the 40-th International Conference on Machine Learning*, pp. 21884–21914, 2023. (Cited on page 4)

Gábor Lugosi and Shahar Mendelson. Mean estimation and regression under heavy-tailed distributions: A survey. *Foundations of Computational Mathematics*, 19(5):1145–1190, 2019. (Cited on page 7)

Liam Madden, Emiliano Dall'Anese, and Stephen Becker. High probability convergence bounds for non-convex stochastic gradient descent with sub-weibull noise. *Journal of Machine Learning Research*, 25(241):1–36, 2024. (Cited on page 4)

Eric Moulines and Francis Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Proceedings of the 7th Annual Conference on Advances in Neural Information Processing Systems*, 2011. (Cited on page 3)

Alexander V. Nazin, Arkadi S. Nemirovsky, Alexandre B. Tsybakov, and Anatoli B. Juditsky. Algorithms of robust stochastic optimization based on mirror descent method. *Automation and Remote Control*, 80(9):1607–1627, 2019. (Cited on page 4)

A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009. (Cited on pages 3 and 4)

Arkadij Semenovič Nemirovskij and David Borisovich Yudin. *Problem complexity and method efficiency in optimization*. Wiley-Interscience, 1983. (Cited on pages 2 and 3)

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 (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 24191–24222. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/4c454d34f3a4c8d6b4ca85a918e5d7ba-Paper-Conference.pdf. (Cited on pages 1, 4, and 6)

Daniela Angela Parletta, Andrea Paudice, Massimiliano Pontil, and Saverio Salzo. High probability bounds for stochastic subgradient schemes with heavy tailed noise. *SIAM Journal on Mathematics of Data Science*, 6(4):953–977, 2024. (Cited on pages 1 and 4)

Daniela Angela Parletta, Andrea Paudice, and Saverio Salzo. An improved analysis of the clipped stochastic subgradient method under heavy-tailed noise, 2025. URL https://arxiv.org/abs/2410.00573. (Cited on pages 3 and 4)

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In *International conference on machine learning*, pp. 1310–1318. Pmlr, 2013. (Cited on page 1)

Nikita Puchkin, Eduard Gorbunov, Nickolay Kutuzov, and Alexander Gasnikov. Breaking the heavy-tailed noise barrier in stochastic optimization problems. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, 2024. (Cited on page 4)

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pp. 400–407, 1951. (Cited on page 1)

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 *Proceedings of the 40-th International Conference on Machine Learning*, 2023. (Cited on pages 1, 4, 6, 11, and 15)

Ohad Shamir and Tong Zhang. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In *Proceedings of the 30th International Conference on Machine Learning*, 2013. (Cited on page 3)

Adrien Taylor and Francis Bach. Stochastic first-order methods: non-asymptotic and computer-aided analyses via potential functions. In *Conference on Learning Theory*, pp. 2934–2992. PMLR, 2019. (Cited on pages 4, 5, and 12)

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. (Cited on page 1)

Mariia Vladimirova, Stéphane Girard, Hien Nguyen, and Julyan Arbel. Sub-weibull distributions: Generalizing sub-gaussian and sub-exponential properties to heavier tailed distributions. *Stat*, 9 (1):e318, 2020. (Cited on page 4)

Nuri Mert Vural, Lu Yu, Krishnakumar Balasubramanian, Stanislav Volgushev, and Murat A. Erdogdu. Mirror descent strikes again: Optimal stochastic convex optimization under infinite noise variance. In *Proceedings of the 35-th Conference on Learning Theory*, 2022. (Cited on pages 3 and 6)

Moslem Zamani and Francois Glineur. Exact convergence rate of the last iterate in subgradient methods, 2023. URL https://arxiv.org/abs/2307.11134. (Cited on page 3)

Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity. In *International Conference on Learning Representations*, 2020a. (Cited on page 1)

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 *Proceedings of the 34th Annual Conference on Neural Information Processing Systems*, 2020b. (Cited on pages 1, 2, and 3)

A Appendix

In this section, we provide additional known details intended to support our analysis.

A.1 Technical Details

Additional notation. We introduce the following notation:

$$
\begin{aligned}
g_t &= \operatorname{clip}\left(\nabla f_{\xi_t}(x_t), \lambda_t\right), \\
\theta_t &= g_t - \nabla f(x_t), \\
\theta_t^u &= g_t - \mathbb{E}_{\xi_t}[g_t], \\
\theta_t^b &= \mathbb{E}_{\xi_t}[g_t] - \nabla f(x_t), \\
R_t &= \|x_t - x^*\|, \\
\Delta_t &= f(x_t) - f_*.
\end{aligned}
$$

We also use the following standard results.

Lemma 1 (Lemma 5.1 from (Sadiev et al., 2023))**.** Let X be a random vector from $\mathbb{R}^d$ and $\widehat{X} = \operatorname{clip}(X, \lambda)$. Then, $\left\|\widehat{X} - \mathbb{E}\left[\widehat{X}\right]\right\| \leq 2\lambda$. Moreover, if for some $\sigma \geq 0$ and $\alpha \in (1, 2]$ we have $\mathbb{E}[X] = x \in \mathbb{R}^d$, $\mathbb{E}\left[\|X - x\|^\alpha\right] \leq \sigma^\alpha$, and $\|x\| \leq \frac{\lambda}{2}$, then

$$
\begin{aligned}
\left\|\mathbb{E}\left[\widehat{X}\right] - x\right\| &\leq \frac{2^\alpha \sigma^\alpha}{\lambda^{\alpha-1}}, \\
\mathbb{E}\left[\left\|\widehat{X} - x\right\|^2\right] &\leq 18\lambda^{2-\alpha}\sigma^\alpha, \\
\mathbb{E}\left[\left\|\widehat{X} - \mathbb{E}\left[\widehat{X}\right]\right\|^2\right] &\leq 18\lambda^{2-\alpha}\sigma^\alpha.
\end{aligned}
$$

This lemma provides sufficient bounds for quantities such as the bias and variance of the clipped stochastic gradient, which satisfies Assumption 3.

Next, we use one of the most popular concentration inequalities: Bernstein's inequality (Bennett, 1962; Dzhaparidze & Van Zanten, 2001; Freedman et al., 1975).

Lemma 2 (Bernstein's inequality)**.** Let the sequence of random variables $\{X_i\}_{i \geq 1}$ form a martingale difference sequence, i.e., $\mathbb{E}\left[X_i \mid X_{i-1}, \ldots, X_1\right] = 0$ for all $i \geq 1$. Assume that conditional variances $\sigma_i^2 = \mathbb{E}\left[X_i^2 \mid X_{i-1}, \ldots, X_1\right]$ exist and are bounded and also assume that there exists deterministic constant $c > 0$ such that $|X_i| \leq c$ almost surely for all $i \geq 1$. Then for all $b > 0$, $G > 0$ and $n \geq 1$

$$
\mathbb{P}\left\{\left|\sum_{i=1}^n X_i\right| > b \text{ and } \sum_{i=1}^n \sigma_i^2 \leq G\right\} \leq 2\exp\left(-\frac{b^2}{2G + \frac{2cb}{3}}\right).
$$

Additionally, we formulate Young's inequality.

Proposition 1 (Young's inequality.)**.** For any $x, y \in \mathbb{R}^d$ and $p > 0$ the following inequality holds:

$$
\|x + y\|^2 \leq (1 + p)\|x\|^2 + \left(1 + \frac{1}{p}\right)\|y\|^2.
$$

In particular, for $p = 1$

$$
\|x + y\|^2 \leq 2\|x\|^2 + 2\|y\|^2.
$$

B MISSING PROOFS

This section is organized as follows. First, we introduce an auxiliary numerical lemma required for the main proof. Next, we state the descent lemma, which serves as the foundation for deriving the main result. Finally, we present the convergence rate theorem for Clipped-SGD based on the last iterate.

B.1 AUXILIARY NUMERICAL LEMMA

Lemma 3 (Numerical lemma)**.** Suppose that $t \geq 0$, $\beta \in (0,1)$, $m \geq 0$ and $n > 0$. Then, we have

$$\left(\frac{(t+1)^{(1-\beta)}-1}{1-\beta}\right)^n (t+1)^{-m} \leq \frac{\ln^n(t+1)}{(t+1)^{m-(1-\beta)n}}.$$

Proof. Let us consider

$$\left(\frac{(t+1)^{1-\beta}-1}{1-\beta}\right)^n (t+1)^{-m}.$$

For the case $t = 0$ it is obvious: $0 \leq 0$. For $t \geq 1$, it can be rewritten as

$$\begin{aligned}\left(\frac{(t+1)^{1-\beta}-1}{1-\beta}\right)^n (t+1)^{-m} &= \left(\ln(t+1)\frac{e^{(1-\beta)\ln(t+1)}-1}{(1-\beta)\ln(t+1)}\right)^n (t+1)^{-m} \\ &= \ln^n(t+1)^{-m}\left(\frac{e^{(1-\beta)\ln(t+1)}-1}{(1-\beta)\ln(t+1)}\right)^n.\end{aligned}$$

What is more, it is known that for all $x > 0$ we have

$$\frac{e^x-1}{x} \leq e^x.$$

It is enough to apply Taylor series or compare the growth of both parts. Therefore, with $\beta < 1$ and $\ln(t+1) > 0$ since $t \geq 1$, we have

$$\begin{aligned}\left(\frac{(t+1)^{1-\beta}-1}{1-\beta}\right)^n (t+1)^{-m} &= \ln^n(t+1)^{-m}\left(\frac{e^{(1-\beta)\ln(t+1)}-1}{(1-\beta)\ln(t+1)}\right)^n \\ &\leq \ln^n(t+1)^{-m}\left(e^{(1-\beta)\ln(t+1)}\right)^n \\ &= \ln^n(t+1)^{-m}(t+1)^{(1-\beta)n} \\ &= \frac{\ln^n(t+1)}{(t+1)^{m-(1-\beta)n}}.\end{aligned}$$

This finishes the proof. □

B.2 DESCENT LEMMA

Lemma 4 (Descent lemma)**.** Suppose that Assumptions 1 and 2 hold. Then, after K iterations of Clipped-SGD, we have

$$\Phi_K \leq \Phi_0 - \sum_{k=0}^{K-1} \gamma_k C \langle x_k - x^*, \theta_k\rangle - \sum_{k=0}^{K-1} (\gamma_k d_{k+1} - 2e_k)\langle \nabla f(x_k), \theta_k\rangle + \sum_{k=0}^{K-1} e_k \|\theta_k\|^2,$$

where $d_{k+1} := d_k + \gamma_k C$, $d_0 := 0$, $e_k := \frac{(Ld_{k+1}+C)\gamma_k^2}{2}$ and $\Phi_k := d_k(f(x_k) - f^*) + \frac{C}{2}\|x_k - x^*\|^2$ with $\gamma_k d_{k+1} \geq e_k$.

Proof. Our proof closely follows the proof of Theorem 5 from Taylor & Bach (2019). Main idea lies in constructing the potential which reflects the convergence of the algorithm. According to

Assumptions 1 and 2, we have

$$
\begin{aligned}
0 \geq\ & (d_{k+1} - d_k)(f(x_k) - f^* + \langle \nabla f(x_k), x^* - x_k \rangle) \\
& - d_{k+1} \left(f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2} \|x_{k+1} - x_k\|^2 - f(x_{k+1}) \right) \\
=\ & d_{k+1}(f(x_{k+1}) - f^*) - d_k(f(x_k) - f^*) + (d_{k+1} - d_k) \langle \nabla f(x_k), x^* - x_k \rangle \\
& - d_{k+1} \left(\langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2} \|x_{k+1} - x_k\|^2 \right) \\
=\ & d_{k+1}(f(x_{k+1}) - f^*) + \frac{C}{2} \|x_{k+1} - x^*\|^2 - \frac{C}{2} \|x_{k+1} - x^*\|^2 \\
& - d_k(f(x_k) - f^*) - \frac{C}{2} \|x_k - x^*\|^2 + \frac{C}{2} \|x_k - x^*\|^2 \\
& + (d_{k+1} - d_k) \langle \nabla f(x_k), x^* - x_k \rangle \\
& - d_{k+1} \left(\langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2} \|x_{k+1} - x_k\|^2 \right),
\end{aligned}
$$

where in the inequality we apply Assumptions 1 and 2, and in the second equation we add and subtract the same terms twice. Decomposing $\|x_{k+1} - x^*\|^2$, we obtain

$$
\begin{aligned}
0 \geq\ & d_{k+1}(f(x_{k+1}) - f^*) + \frac{C}{2} \|x_{k+1} - x^*\|^2 - \frac{C}{2} \|x_{k+1} - x^*\|^2 \\
& - d_k(f(x_k) - f^*) - \frac{C}{2} \|x_k - x^*\|^2 + \frac{C}{2} \|x_k - x^*\|^2 \\
& + (d_{k+1} - d_k) \langle \nabla f(x_k), x^* - x_k \rangle \\
& - d_{k+1} \left(\langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2} \|x_{k+1} - x_k\|^2 \right) \\
=\ & d_{k+1}(f(x_{k+1}) - f^*) + \frac{C}{2} \|x_{k+1} - x^*\|^2 - d_k(f(x_k) - f^*) - \frac{C}{2} \|x_k - x^*\|^2 \\
& - \frac{C}{2} \left(\|x_k - x^*\|^2 + 2 \langle x_{k+1} - x_k, x_k - x^* \rangle + \|x_{k+1} - x_k\|^2 \right) + \frac{C}{2} \|x_k - x^*\|^2 \\
& + (d_{k+1} - d_k) \langle \nabla f(x_k), x^* - x_k \rangle - d_{k+1} \left(\langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2} \|x_{k+1} - x_k\|^2 \right) \\
=\ & d_{k+1}(f(x_{k+1}) - f^*) + \frac{C}{2} \|x_{k+1} - x^*\|^2 - d_k(f(x_k) - f^*) - \frac{C}{2} \|x_k - x^*\|^2 \\
& - \frac{C}{2} \left(2 \langle x_{k+1} - x_k, x_k - x^* \rangle + \|x_{k+1} - x_k\|^2 \right) \\
& + (d_{k+1} - d_k) \langle \nabla f(x_k), x^* - x_k \rangle - d_{k+1} \left(\langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2} \|x_{k+1} - x_k\|^2 \right).
\end{aligned}
$$

Using the notation for the considered potential, applying the update rule and noting that $d_{k+1} - d_k = \gamma_k C$, we have

$$
\begin{aligned}
0 \geq\ & d_{k+1}(f(x_{k+1}) - f^*) + \frac{C}{2} \|x_{k+1} - x^*\|^2 - d_k(f(x_k) - f^*) - \frac{C}{2} \|x_k - x^*\|^2 \\
& - \frac{C}{2} \left(2 \langle x_{k+1} - x_k, x_k - x^* \rangle + \|x_{k+1} - x_k\|^2 \right) \\
& + (d_{k+1} - d_k) \langle \nabla f(x_k), x^* - x_k \rangle - d_{k+1} \left(\langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2} \|x_{k+1} - x_k\|^2 \right) \\
=\ & \Phi_{k+1} - \Phi_k - C \langle x_{k+1} - x_k, x_k - x^* \rangle - \frac{L d_{k+1} + C}{2} \|x_{k+1} - x_k\|^2 \\
& + \gamma_k C \langle \nabla f(x_k), x^* - x_k \rangle - d_{k+1} \langle \nabla f(x_k), x_{k+1} - x_k \rangle \\
=\ & \Phi_{k+1} - \Phi_k + \gamma_k C \langle x_k - x^*, g_k \rangle - \frac{(L d_{k+1} + C)\gamma_k^2}{2} \|g_k\|^2 + \gamma_k C \langle \nabla f(x_k), x^* - x_k \rangle \\
& + \gamma_k d_{k+1} \langle \nabla f(x_k), g_k \rangle .
\end{aligned}
$$

Using the notation of g_k and θ_k, we get

$$
\begin{aligned}
0 &\geq \Phi_{k+1} - \Phi_k + \gamma_k C \langle x_k - x^*, g_k \rangle - \frac{(Ld_{k+1} + C)\gamma_k^2}{2} \|g_k\|^2 + \gamma_k C \langle \nabla f(x_k), x^* - x_k \rangle \\
&+ \gamma_k d_{k+1} \langle \nabla f(x_k), g_k \rangle \\
&= \Phi_{k+1} - \Phi_k + \gamma_k C \langle x_k - x^*, \theta_k \rangle + \gamma_k d_{k+1} \langle \nabla f(x_k), \theta_k \rangle + \gamma_k d_{k+1} \|\nabla f(x_k)\|^2 \\
&- \frac{(Ld_{k+1} + C)\gamma_k^2}{2} \left(\|\nabla f(x_k)\|^2 + 2 \langle \nabla f(x_k), \theta_k \rangle + \|\theta_k\|^2 \right) \\
&= \Phi_{k+1} - \Phi_k + \gamma_k C \langle x_k - x^*, \theta_k \rangle - \frac{(Ld_{k+1} + C)\gamma_k^2}{2} \|\theta_k\|^2 \\
&+ (\gamma_k d_{k+1} - (Ld_{k+1} + C)\gamma_k^2) \langle \nabla f(x_k), \theta_k \rangle + \left(\gamma_k d_{k+1} - \frac{(Ld_{k+1} + C)\gamma_k^2}{2} \right) \|\nabla f(x_k)\|^2.
\end{aligned}
$$

Due to the notation of $e_k := \frac{(Ld_{k+1}+C)\gamma_k^2}{2}$, rearranging terms gives the following inequality:

$$
\Phi_{k+1} \leq \Phi_k - \gamma_C L \langle x_k - x^*, \theta_k \rangle - (\gamma_k d_{k+1} - 2e_k) \langle \nabla f(x_k), \theta_k \rangle + e_k \|\theta_k\|^2,
$$

where we apply $\gamma_k d_{k+1} \geq e_k$ to eliminate the term related to $\|\nabla f(x_k)\|^2$. Summing up finishes the proof. □

B.3 Proof of Theorem 1

Theorem 2. *Suppose that Assumptions 1, 2 and 3 hold. Then, if we choose*

$$
\gamma_k = \min \left\{ \frac{1}{1024L \ln^3 \left(\frac{6(k+1)^2}{\delta} \right)}, \frac{p(\Phi_0, L, \sigma)}{256 \cdot 4^{1/\alpha} (k+1)^\beta \ln^3 \left(\frac{6(k+1)^2}{\delta} \right)} \right\} \tag{4}
$$

and

$$
\lambda_k = \frac{\sqrt{\Phi_0 / C}}{256 \sqrt{b_k} \gamma_k \ln^{5/2} \left(\frac{6(k+1)^2}{\delta} \right)}, \tag{5}
$$

where

$$
b_k = \begin{cases} d_k, & t > 0; \\ 1, & t = 0, \end{cases} \tag{6}
$$

$$
p(\Phi_0, L, \sigma) = \min \left\{ \frac{\sqrt{\Phi_0 / C}}{\sigma}, \frac{\sqrt{\Phi_0 / C}^{\frac{2\alpha}{3\alpha - 1}}}{C^{\frac{\alpha - 1}{3\alpha - 1}} \sigma^{\frac{2\alpha}{3\alpha - 1}}}, \frac{\sqrt{\Phi_0 / C}^{\frac{2}{3}}}{C^{\frac{1}{3}} \sigma^{\frac{2}{3}}} \right\}, \tag{7}
$$

parameter β is clarified later, and

$$
C = \max \left\{ L, \frac{1}{4^{1 - 1/\alpha} p(\Phi_0, L, \sigma)} \right\}, \tag{8}
$$

then, after K iterations of Clipped-SGD*, we have that*

$$
\begin{aligned}
&f(x_K) - f^* \\
&= \mathcal{O} \left(\frac{L R_0^2 \ln^4 \left(\frac{6(K+1)^2}{\delta} \right)}{K} + \frac{\max \left\{ R_0 \sigma, L^{\frac{\alpha - 1}{3\alpha - 1}} R_0^{\frac{4\alpha - 2}{3\alpha - 1}} \sigma^{\frac{2\alpha}{3\alpha - 1}}, L^{\frac{1}{3}} R_0^{\frac{4}{3}} \sigma^{\frac{2}{3}} \right\} \ln^4 \left(\frac{6(K+1)^2}{\delta} \right)}{K^{1 - \beta}} \right)
\end{aligned}
$$

hold with probability at least $1 - \delta \sum_{t=1}^{K} \frac{1}{t^2}$.

Remark 3. During the proof, some constrains over β appear. This parameter obviously has a strong impact on a final convergence bound. This is why, in the course of the proof, parameter β must be chosen with greater care. These constraints are formulated as **Conditions**.

Remark 4. In the definition of $p(\Phi_0, L, \sigma)$ we use the *recurrent* equation. Indeed, after substitution of (8) into (7), function $p(\Phi_0, L, \sigma)$ can be expressed as a function of itself. However, this form is more intuitive for the proof, since C plays the role of *effective* constant of smoothness for the potential Φ_t. Of course, in the end of the proof of Theorem 2, we provide an explicit form of $p(\Phi_0, L, \sigma)$ in terms of problem parameters R_0, L and σ for formulating the final convergence bound.

Proof. The proof in constructed in the similar manner as in Sadiev et al. (2023). For each $k = 0, 1, \ldots$ let us consider probabilistic events E_k: inequalities

$$-\sum_{l=0}^{t-1}\gamma_l C\left\langle x_l - x^*, \theta_l\right\rangle - \sum_{l=0}^{t-1}(\gamma_l d_{l+1} - 2e_l)\left\langle \nabla f(x_l), \theta_l\right\rangle + \sum_{l=0}^{t-1} e_l\|\theta_l\|^2 \leq \Phi_0 \ln\left(\frac{6(t+1)^2}{\delta}\right);$$

$$\Phi_t \leq 2\Phi_0 \ln\left(\frac{6(t+1)^2}{\delta}\right)$$

hold for all $t = 0, \ldots, k$ simultaneously. We want to show via induction that $\mathbb{P}\{E_k\} \geq 1 - \delta\sum\limits_{t=1}^{k}\frac{1}{t^2}$. For $k = 0$ it is obvious. Now, let us assume that $\mathbb{P}\{E_{T-1}\} \geq 1 - \delta\sum\limits_{t=1}^{T-1}\frac{1}{t^2}$ for some $T \geq 1$. Then, applying Lemma 4, we have

$$\Phi_t \leq \Phi_0 - \sum_{l=0}^{t-1}\gamma_l C\left\langle x_l - x^*, \theta_l\right\rangle - \sum_{l=0}^{t-1}(\gamma_l d_{l+1} - 2e_l)\left\langle \nabla f(x_l), \theta_l\right\rangle + \sum_{l=0}^{t-1} e_l\|\theta_l\|^2$$

for all $t = 0, \ldots, T$. The only thing we should check is to guarantee $\gamma_t d_{t+1} \geq e_t$. In fact, we have

$$\gamma_t d_{t+1} \geq e_t = \frac{(Ld_{t+1} + C)\gamma_t^2}{2} \Leftrightarrow 2d_{t+1} \geq (Ld_{t+1} + C)\gamma_t.$$

Then, it is enough to guarantee

$$\gamma_t \leq \frac{2d_{t+1}}{C(d_{t+1}+1)}, \tag{9}$$

since $L \leq C$. Now, it can be obviously shown that the RHS of (9) is increasing due to the growth of d_{t+1}, and the LHS is decreasing according to the (4). Therefore, it is *enough* to guarantee

$$\gamma_0 \leq \frac{2d_1}{C(d_1+1)}.$$

Thus, we have

$$(\gamma_0 C + 1)\gamma_0 C \leq 2\gamma_0 C \Leftrightarrow (\gamma_0 C)^2 \leq \gamma_0 C \Leftrightarrow \gamma_0 \leq \frac{1}{C}.$$

Moreover, one can have

$$\begin{aligned}\gamma_t &= \frac{1}{1024\ln^3\left(\frac{6(t+1)^2}{\delta}\right)}\min\left\{\frac{1}{L}, \frac{4^{1-1/\alpha}p(\Phi_0, L, \sigma)}{(t+1)^\beta}\right\}\\ &\leq \frac{1}{1024\ln^3\left(\frac{6(t+1)^2}{\delta}\right)}\min\left\{\frac{1}{L}, 4^{1-1/\alpha}p(\Phi_0, L, \sigma)\right\} = \frac{1}{1024C\ln^3\left(\frac{6(t+1)^2}{\delta}\right)},\end{aligned} \tag{10}$$

where the first inequality holds since the logarithmic factor appears in the denominator. Consequently, we can apply Lemma 4 with (4). Continuing, for all $t = 0, \ldots, T-1$ the event E_{T-1} implies

$$\begin{aligned}\Phi_t &\leq \Phi_0 - \sum_{l=0}^{t-1}\gamma_l C\left\langle x_l - x^*, \theta_l\right\rangle - \sum_{l=0}^{t-1}(\gamma_l d_{l+1} - 2e_l)\left\langle \nabla f(x_l), \theta_l\right\rangle + \sum_{l=0}^{t-1} e_l\|\theta_l\|^2\\ &\leq 2\Phi_0 \ln\left(\frac{6(t+1)^2}{\delta}\right).\end{aligned}$$

What is more, the event E_{T-1} implies

$$\Phi_T \leq \Phi_0 - \sum_{t=0}^{T-1} \gamma_t C \langle x_t - x^*, \theta_t \rangle - \sum_{t=0}^{T-1} (\gamma_t d_{t+1} - 2e_t) \langle \nabla f(x_t), \theta_t \rangle + \sum_{t=0}^{T-1} e_t \|\theta_t\|^2. \tag{11}$$

At the same time, the event E_{T-1} implies

$$\begin{aligned}\|\nabla f(x_t)\| &\leq \begin{cases} \sqrt{2L(f(x_t) - f^*)}, & t > 0; \\ L\|x_0 - x^*\|, & t = 0; \end{cases} \\ &\leq \begin{cases} \sqrt{\frac{4L\Phi_0 \ln\left(\frac{6(t+1)^2}{\delta}\right)}{d_t}}, & t > 0; \\ \sqrt{2L\Phi_0}, & t = 0; \end{cases},\end{aligned}$$

where b_t is defined in (6) and in the second inequality we use that $R_0 = \sqrt{\frac{2\Phi_0}{C}} \leq \sqrt{\frac{2\Phi_0}{L}}$. Hence, with (10) we get

$$\|\nabla f(x_t)\| \leq 2\sqrt{\frac{L\Phi_0 \ln\left(\frac{6(t+1)^2}{\delta}\right)}{b_t}} \overset{(4),(5)}{\leq} \frac{\lambda_t}{2}, \tag{12}$$

since

$$\frac{\lambda_t}{2} \overset{(5)}{=} \frac{\sqrt{\Phi_0/C}}{512\sqrt{b_t}\gamma_t \ln^{5/2}\left(\frac{6(t+1)^2}{\delta}\right)} \overset{(10)}{\geq} \frac{2C\sqrt{\Phi_0/C}\sqrt{\ln\left(\frac{6(t+1)^2}{\delta}\right)}}{\sqrt{b_t}} \geq 2\sqrt{\frac{L\Phi_0 \ln\left(\frac{6(t+1)^2}{\delta}\right)}{b_t}},$$

where we also apply $C \geq L$. Simultaneously, we have

$$\|x_t - x^*\| \leq \sqrt{\frac{2\Phi_t}{C}} \leq 2\sqrt{\frac{\Phi_0 \ln\left(\frac{6(t+1)^2}{\delta}\right)}{C}}$$

Now, let us decompose (11) using the notation of bias and unbiased part:

$$\begin{aligned}\Phi_T &\leq \Phi_0 - \sum_{t=0}^{T-1} \gamma_t C \langle x_t - x^*, \theta_t \rangle - \sum_{t=0}^{T-1} (\gamma_t d_{t+1} - 2e_t) \langle \nabla f(x_t), \theta_t \rangle + \sum_{t=0}^{T-1} e_t \|\theta_t\|^2 \\ &\leq \Phi_0 \underbrace{- \sum_{t=0}^{T-1} \gamma_t C \langle x_t - x^*, \theta_t^u \rangle}_{①} \underbrace{- \sum_{t=0}^{T-1} \gamma_t C \langle x_t - x^*, \theta_t^b \rangle}_{②} \\ &\underbrace{- \sum_{t=0}^{T-1} (\gamma_t d_{t+1} - 2e_t) \langle \nabla f(x_t), \theta_t^u \rangle}_{③} \underbrace{- \sum_{t=0}^{T-1} (\gamma_t d_{t+1} - 2e_t) \langle \nabla f(x_t), \theta_t^b \rangle}_{④} \\ &\underbrace{+ \sum_{t=0}^{T-1} 2e_t \left(\|\theta_t^u\|^2 - \mathbb{E}_{\xi_t}\left[\|\theta_t^u\|^2\right]\right)}_{⑤} + \underbrace{\sum_{t=0}^{T-1} 2e_t \mathbb{E}_{\xi_t}\left[\|\theta_t^u\|^2\right]}_{⑥} + \underbrace{\sum_{t=0}^{T-1} 2e_t \|\theta_t^b\|^2}_{⑦},\end{aligned}$$

where we apply Young's inequality for $\left\|\theta_t^u + \theta_t^b\right\|^2$.

Next, we are going to give sufficient bound on each term: ①, …, ⑦. However, before we start, let us give extra boundaries for the sequence $\{b_t\}$. For the case $t = 0$ we have $b_t = 1$. On the other

hand, for $t > 0$ we get

$$\frac{1}{1024\ln^3\left(\frac{6}{\delta}\right)} = \gamma_0 C \leq b_t = d_t = \sum_{k=0}^{t-1} \gamma_k C \overset{y\geq 0,(4)}{\leq} \sum_{k=0}^{t-1} \frac{p(\Phi_0, L, \sigma)C}{256 \cdot 4^{1/\alpha}(k+1)^\beta}$$
$$\leq \frac{p(\Phi_0, L, \sigma)C}{256 \cdot 4^{1/\alpha}} \left(1 + \int_1^t \frac{1}{x^\beta} dx\right) \leq \frac{p(\Phi_0, L, \sigma)C}{256 \cdot 4^{1/\alpha}} \left(1 + \frac{t^{1-\beta} - 1}{1-\beta}\right). \tag{13}$$

To reflect the correct behavior of convergence with $\alpha \to 1$, we need to bound b_t as above. Nevertheless, to obtain the correct bound for fixed $\alpha \in (1, 2]$, it is *enough* to use

$$\frac{p(\Phi_0, L, \sigma)C}{256 \cdot 4^{1/\alpha}} \frac{t^{1-\beta}}{1-\beta}. \tag{14}$$

The problem arises when we consider the limit $\alpha \to 1$. In the next part of the proof it will be shown that

$$\alpha \to 1 \Rightarrow \beta \to 1$$

due to the constraints over β. Consequently, we could get

$$\lim_{\beta \to 1} \frac{t^{1-\beta}}{1-\beta} = \infty$$

instead of

$$\lim_{\beta \to 1} \frac{t^{1-\beta} - 1}{1-\beta} = \ln(t+1).$$

As a result, for more reasonable analysis, we *must* use (13) instead of (14).
Next, since boundaries over the true gradients hold, we can apply Lemma 1 to obtain

$$\|\theta_t^u\| \leq 2\lambda_t,$$

$$\left\|\theta_t^b\right\| \leq \frac{2^\alpha \sigma^\alpha}{\lambda_t^{\alpha-1}},$$

$$\mathbb{E}_{\xi_t}\left[\|\theta_t^u\|^2\right] \leq 18\lambda_t^{2-\alpha}\sigma^\alpha.$$

Now we are ready to bound terms ① − ⑦.
Bound for ①. By definition θ_t^u, we get

$$\mathbb{E}_{\xi_t}[-\gamma_t C \langle x_t - x^*, \theta_t^u\rangle] = 0.$$

Moreover, we have that event E_{T-1} implies

$$|-\gamma_t C \langle x_t - x^*, \theta_t^u\rangle| \leq \gamma_t C \|x_t - x^*\| \|\theta_t^u\| \leq 4\gamma_t \lambda_t C \sqrt{\frac{\Phi_0 \ln\left(\frac{6(t+1)^2}{\delta}\right)}{C}} \overset{(5)}{=} \frac{\Phi_0}{64\sqrt{b_t}\ln^2\left(\frac{6(t+1)^2}{\delta}\right)}$$
$$\leq \frac{\Phi_0}{2\ln^{1/2}\left(\frac{6(t+1)^2}{\delta}\right)} \leq \frac{\Phi_0}{2} := c.$$

Next, we define σ_t^2:

$$\sigma_t^2 = \mathbb{E}_{\xi_t}[\gamma_t^2 C^2 \langle x_t - x^*, \theta_t^u\rangle^2] \leq 4\gamma_t^2 \Phi_0 C \ln\left(\frac{6(t+1)^2}{\delta}\right) \mathbb{E}_{\xi_t}\left[\|\theta_t^u\|^2\right]$$

Therefore, we can apply Bernstein's inequality with $b = {}^{\Phi_0 \ln\left(\frac{6(T)^2}{\delta}\right)}/_2$ and $G = {}^{\Phi_0^2 \ln\left(\frac{6(T)^2}{\delta}\right)}/_{24}$:

$$\mathbb{P}\left\{|①| > b \text{ and } \sum_{t=0}^{T-1} \sigma_t^2 \leq G\right\} \leq 2\exp\left(-\frac{b^2}{2G + \frac{2cb}{3}}\right) = \frac{\delta}{3T^2}.$$

Consequently, we have

$$\mathbb{P}\left\{|①| \leq b \text{ either } \sum_{t=0}^{T-1} \sigma_t^2 > G\right\} \geq 1 - \frac{\delta}{3T^2}.$$

What is more, we have

$$\sum_{t=0}^{T-1} \sigma_t^2 \leq \sum_{t=0}^{T-1} 4\gamma_t^2 \Phi_0 C \ln\left(\frac{6(t+1)^2}{\delta}\right) \mathbb{E}_{\xi_t}\left[\|\theta_t^u\|^2\right] \leq \sum_{t=0}^{T-1} 72\gamma_t^2 \Phi_0 C \ln\left(\frac{6(t+1)^2}{\delta}\right) \lambda_t^{2-\alpha}\sigma^\alpha$$
$$\overset{(5)}{\leq} \sum_{t=0}^{T-1} \frac{72\sqrt{\Phi_0/C}^{2-\alpha}\Phi_0 C\gamma_t^\alpha\sigma^\alpha \ln\left(\frac{6(t+1)^2}{\delta}\right)}{256^{2-\alpha} b_t^{1-\alpha/2} \ln^{5-5\alpha/2}\left(\frac{6(t+1)^2}{\delta}\right)}.$$

Using the notation of b_t and emphasizing that $b_t \geq \frac{1}{1024\ln^3\left(\frac{6}{\delta}\right)}$, we obtain

$$\sum_{t=0}^{T-1} \sigma_t^2 \leq \sum_{t=0}^{T-1} \frac{72\sqrt{\Phi_0/C}^{2-\alpha}\Phi_0 C\gamma_t^\alpha\sigma^\alpha \ln\left(\frac{6(t+1)^2}{\delta}\right)}{256^{2-\alpha} b_t^{1-\alpha/2} \ln^{5-5\alpha/2}\left(\frac{6(t+1)^2}{\delta}\right)}$$
$$\leq \sum_{t=0}^{T-1} \frac{72\cdot 1024^{1-\alpha/2}\ln^{3-3\alpha/2}\left(\frac{6}{\delta}\right)\sqrt{\Phi_0/C}^{2-\alpha}\Phi_0 C\gamma_t^\alpha\sigma^\alpha \ln\left(\frac{6(t+1)^2}{\delta}\right)}{256^{2-\alpha}\ln^{5-5\alpha/2}\left(\frac{6(t+1)^2}{\delta}\right)}$$
$$\leq \sum_{t=0}^{T-1} \frac{72\cdot 1024^{1-\alpha/2}\ln^{3-3\alpha/2}\left(\frac{6}{\delta}\right)\Phi_0^2}{256^2\cdot 4(t+1)^{\beta\alpha}\ln^{4-\alpha}\left(\frac{6(t+1)^2}{\delta}\right)}.$$

To get sufficient bound depending on $\ln\left(\frac{6(t+1)^2}{\delta}\right)$, we must have this term in power of *one*. Thus, we obtain

Condition 1:

$$\left\{\beta\alpha \geq 1 \Rightarrow \beta \geq \tfrac{1}{\alpha}. \right. \tag{15}$$

Moreover, we get

$$\sum_{t=0}^{T-1} \frac{1}{(t+1)^{\beta\alpha}} \leq \sum_{t=0}^{T-1} \frac{1}{(t+1)} \leq 1 + \int_1^T \frac{1}{x}dx = 1 + \ln(T) \leq 3\ln\left(\frac{\sqrt{6}T}{\sqrt{\delta}}\right) = 3/2\ln\left(\frac{6T^2}{\delta}\right) \tag{16}$$

since $T \geq 1$. Therefore, with (17) one can have

$$\sum_{t=0}^{T-1} \sigma_t^2 \leq \sum_{t=0}^{T-1} \frac{72\cdot 1024^{1-\alpha/2}\ln^{3-3\alpha/2}\left(\frac{6}{\delta}\right)\Phi_0^2}{256^2\cdot 4(t+1)^{\beta\alpha}\ln^{4-\alpha}\left(\frac{6(t+1)^2}{\delta}\right)} \overset{(15),(16)}{\leq} \frac{108\cdot 32\Phi_0^2}{256^2\cdot 4}\ln\left(\frac{6T^2}{\delta}\right)$$
$$\leq \frac{\Phi_0^2\ln\left(\frac{6T^2}{\delta}\right)}{24},$$

where we used that $3 - 3\alpha/2 \leq 4 - \alpha$.

Bound for ②. The event E_{T-1} implies

$$-\sum_{t=0}^{T-1}\gamma_t C\left\langle x_t - x^*, \theta_t^b\right\rangle \leq \sum_{t=0}^{T-1}\gamma_t C\|x_t - x^*\|\left\|\theta_t^b\right\| \leq \sum_{t=0}^{T-1} 2\sqrt{\Phi_0 C\ln\left(\frac{6(t+1)^2}{\delta}\right)}\frac{2^\alpha\sigma^\alpha\gamma_t}{\lambda_t^{\alpha-1}}$$
$$= \sum_{t=0}^{T-1} 2\sqrt{\Phi_0 C}\ln^{5\alpha/2-2}\left(\frac{6(t+1)^2}{\delta}\right)\frac{2^\alpha\cdot 256^{\alpha-1}\sigma^\alpha\gamma_t^\alpha b_t^{\frac{\alpha-1}{2}}}{\sqrt{\Phi_0/C}^{\alpha-1}}$$
$$= \sum_{t=0}^{T-1} 2C\sqrt{\Phi_0/C}^{2-\alpha}\ln^{5\alpha/2-2}\left(\frac{6(t+1)^2}{\delta}\right)2^\alpha\cdot 256^{\alpha-1}\sigma^\alpha\gamma_t^\alpha b_t^{\frac{\alpha-1}{2}}.$$

Therefore, applying (13), one can obtain

$$
\begin{aligned}
② &\le \sum_{t=0}^{T-1} 2C\sqrt{\Phi_0/C}^{2-\alpha} \ln^{5\alpha/2-2}\left(\frac{6(t+1)^2}{\delta}\right) 2^\alpha \cdot 256^{\alpha-1}\sigma^\alpha \gamma_t^\alpha b_t^{\frac{\alpha-1}{2}} \\
&\overset{(13),t\le t+1}{\le} 2C\sqrt{\Phi_0/C}^{2-\alpha} \ln^{5\alpha/2-2}\left(\frac{6}{\delta}\right) 2^\alpha \cdot 256^{\alpha-1}\sigma^\alpha \gamma_0^\alpha \\
&+ \sum_{t=1}^{T-1} 2C\sqrt{\Phi_0/C}^{2-\alpha} \ln^{5\alpha/2-2}\left(\frac{6(t+1)^2}{\delta}\right) 2^\alpha \cdot 256^{\alpha-1}\sigma^\alpha \gamma_t^\alpha b_t^{\frac{\alpha-1}{2}} \\
&\times \left(\frac{3p(\Phi_0,L,\sigma)C}{256\cdot 4^{1/\alpha}} \frac{(t+1)^{1-\beta}-1}{1-\beta}\right)^{\frac{\alpha-1}{2}},
\end{aligned}
$$

where in the last inequality we apply $b_t \le b_{t+1}$ and $1 + \frac{t^{1-\beta}-1}{1-\beta} \le 1 + \frac{(t+1)^{1-\beta}-1}{1-\beta} \le 3\frac{(t+1)^{1-\beta}-1}{1-\beta}$, where $t \ge 1$. Consequently, we get

$$
\begin{aligned}
② &\le 2C\sqrt{\Phi_0/C}^{2-\alpha} \ln^{5\alpha/2-2}\left(\frac{6}{\delta}\right) 2^\alpha \cdot 256^{\alpha-1}\sigma^\alpha \gamma_0^\alpha \\
&+ \sum_{t=1}^{T-1} \frac{6\cdot 256^{\alpha-1} C\sqrt{\Phi_0/C}^{2-\alpha}(p(\Phi_0,L,\sigma)C)^{\frac{\alpha-1}{2}}}{(256\cdot 4^{1/\alpha})^{\frac{\alpha-1}{2}}} \ln^{5\alpha/2-2}\left(\frac{6(t+1)^2}{\delta}\right) 2^\alpha \sigma^\alpha \gamma_t^\alpha \times \left(\frac{(t+1)^{1-\beta}-1}{1-\beta}\right)^{\frac{\alpha-1}{2}} \\
&\overset{(4)}{\le} \frac{2C\sqrt{\Phi_0/C}^{2-\alpha} \ln^{5\alpha/2-2}\left(\frac{6}{\delta}\right) 2^\alpha \cdot 256^{\alpha-1}\sigma^\alpha p^\alpha(\Phi_0,L,\sigma)}{(256\cdot 4^{1/\alpha})^\alpha \ln^{3\alpha}\left(\frac{6}{\delta}\right)} \\
&+ \sum_{t=1}^{T-1} \frac{6\cdot 256^{\alpha-1} C\sqrt{\Phi_0/C}^{2-\alpha} p^{\frac{3\alpha-1}{2}}(\Phi_0,L,\sigma)C^{\frac{\alpha-1}{2}}}{(256\cdot 4^{1/\alpha})^{\frac{3\alpha-1}{2}}(t+1)^{\beta\alpha}\ln^{3\alpha}\left(\frac{6(t+1)^2}{\delta}\right)} \ln^{5\alpha/2-2}\left(\frac{6(t+1)^2}{\delta}\right) 2^\alpha \sigma^\alpha \times \left(\frac{(t+1)^{1-\beta}-1}{1-\beta}\right)^{\frac{\alpha-1}{2}}.
\end{aligned}
$$

Applying (7) and using that $3\alpha \ge \frac{5\alpha}{2} - 2$, one can derive

$$
② \le \frac{8\cdot 256^{\alpha-1}\Phi_0}{(256\cdot 4^{1/\alpha})^\alpha} + \sum_{t=1}^{T-1} \frac{24\cdot 256^{\alpha-1}\Phi_0}{(256\cdot 4^{1/\alpha})^{\frac{3\alpha-1}{2}}(t+1)^{\beta\alpha}\ln^{2+\alpha/2}\left(\frac{6(t+1)^2}{\delta}\right)} \times \left(\frac{(t+1)^{1-\beta}-1}{1-\beta}\right)^{\frac{\alpha-1}{2}}.
$$

Consequently, applying Lemma 3, one can get

$$
\begin{aligned}
② &\le \frac{8\cdot 256^{\alpha-1}\Phi_0}{(256\cdot 4^{1/\alpha})^\alpha} + \sum_{t=1}^{T-1} \frac{24\cdot 256^{\alpha-1}\Phi_0}{(256\cdot 4^{1/\alpha})^{\frac{3\alpha-1}{2}}(t+1)^{\beta\alpha}\ln^{2+\alpha/2}\left(\frac{6(t+1)^2}{\delta}\right)} \times \left(\frac{(t+1)^{1-\beta}-1}{1-\beta}\right)^{\frac{\alpha-1}{2}} \\
&\overset{(Lemma\ 3)}{\le} \frac{8\cdot 256^{\alpha-1}\Phi_0}{(256\cdot 4^{1/\alpha})^\alpha} + \sum_{t=1}^{T-1} \frac{24\cdot 256^{\alpha-1}\Phi_0}{(256\cdot 4^{1/\alpha})^{\frac{3\alpha-1}{2}}\ln^{2+\alpha/2-\alpha-1/2}\left(\frac{6(t+1)^2}{\delta}\right)} \times \frac{1}{(t+1)^{\beta\alpha-(1-\beta)(\alpha-1)/2}} \\
&\le \frac{8\cdot 256^{\alpha-1}\Phi_0}{(256\cdot 4^{1/\alpha})^\alpha} + \sum_{t=1}^{T-1} \frac{24\cdot 256^{\alpha-1}\Phi_0}{(256\cdot 4^{1/\alpha})^{\frac{3\alpha-1}{2}}} \frac{1}{(t+1)^{\beta\alpha-(1-\beta)(\alpha-1)/2}} \\
&\le \sum_{t=0}^{T-1} 2\Phi_0 \max\left\{\frac{8\cdot 256^{\alpha-1}}{(256\cdot 4^{1/\alpha})^\alpha}, \frac{24\cdot 256^{\alpha-1}}{(256\cdot 4^{1/\alpha})^{\frac{3\alpha-1}{2}}}\right\} \frac{1}{(t+1)^{\beta\alpha-(1-\beta)(\alpha-1)/2}}.
\end{aligned}
$$

Similar to the bound of ①, next inequality should hold:

Condition 2:

$$
\left\{\beta\alpha - \frac{(1-\beta)(\alpha-1)}{2} \ge 1 \Rightarrow \beta \ge \frac{\alpha+1}{3\alpha-1}. \right. \tag{17}
$$

Due to these constraints, we derive

$$② \leq \sum_{t=0}^{T-1} 2\Phi_0 \max\left\{\frac{8\cdot 256^{\alpha-1}}{(256\cdot 4^{1/\alpha})^{\alpha}}, \frac{24\cdot 256^{\alpha-1}}{(256\cdot 4^{1/\alpha})^{\frac{3\alpha-1}{2}}}\right\} \frac{1}{(t+1)^{\beta\alpha-(1-\beta)(\alpha-1)/2}}$$
$$\overset{(17),(16)}{\leq} 3\Phi_0 \max\left\{\frac{8\cdot 256^{\alpha-1}}{(256\cdot 4^{1/\alpha})^{\alpha}}, \frac{24\cdot 256^{\alpha-1}}{(256\cdot 4^{1/\alpha})^{\frac{3\alpha-1}{2}}}\right\} \ln\left(\frac{6T^2}{\delta}\right) \leq \frac{\Phi_0 \ln\left(\frac{6T^2}{\delta}\right)}{12},$$

where we bound $\sum_{t=0}^{T-1} \frac{1}{t+1}$ in the similar way as for ①.

Bound for ③. Due to the definition θ_t^u, we get

$$\mathbb{E}_{\xi_t}\left[-(\gamma_t d_{t+1} - 2e_t)\left\langle \nabla f(x_t), \theta_t^u\right\rangle\right] = 0.$$

Moreover, it is uniformly bounded:

$$\left|-(\gamma_t d_{t+1} - 2e_t)\left\langle \nabla f(x_t), \theta_t^u\right\rangle\right| \leq 3\gamma_t d_{t+1}\left\|\nabla f(x_t)\right\|\left\|\theta_t^u\right\| \leq 12\gamma_t\lambda_t d_{t+1}\sqrt{\frac{\Phi_0 C \ln\left(\frac{6(t+1)^2}{\delta}\right)}{b_t}}$$
$$\overset{(5)}{\leq} \frac{3\Phi_0 d_{t+1}}{64 b_t} = \frac{3\Phi_0(d_t+\gamma_t C)}{64 b_t}.$$

For the case $t=0$ we have $\frac{d_0+\gamma_0 C}{b_0} = \frac{C_1}{\ln\left(\frac{6}{\delta}\right)} \leq 1$; at the same time, if $t>0$, one can obtain $\frac{d_t+\gamma_t C}{b_t} = \frac{d_t+\gamma_t C}{d_t} \leq 2$ since the sequence of stepsizes $\{\gamma_t\}$ is non-increasing: $\frac{\gamma_t C}{d_t} \leq \frac{\gamma_0 C}{d_t} \leq 1$. Therefore,

$$\left|-(\gamma_t d_{t+1} - 2e_t)\left\langle \nabla f(x_t), \theta_t^u\right\rangle\right| \leq \frac{3\Phi_0(d_t+\gamma_t L)}{64 b_t} \leq \frac{3\Phi_0}{32} := c.$$

Let us define $\sigma_t^2 = \mathbb{E}_{\xi_t}\left[\left((\gamma_t d_{t+1} - 2e_t)\left\langle \nabla f(x_t), \theta_t^u\right\rangle\right)^2\right]$. Hence, we get

$$\sigma_t^2 \leq \mathbb{E}_{\xi_t}\left[9\gamma_t^2 d_{t+1}^2\left\|\nabla f(x_t)\right\|^2\left\|\theta_t^u\right\|^2\right] = 9\gamma_t^2 d_{t+1}^2\left\|\nabla f(x_t)\right\|^2\mathbb{E}_{\xi_t}\left[\left\|\theta_t^u\right\|^2\right].$$

Consequently, we can apply Bernstein's inequality with $b = \Phi_0\ln\left(\frac{6(T)^2}{\delta}\right)/12$ and $G = \Phi_0^2\ln\left(\frac{6(T)^2}{\delta}\right)/1152$:

$$\mathbb{P}\left\{|③| > b \text{ and } \sum_{t=0}^{T-1}\sigma_t^2 \leq G\right\} \leq 2\exp\left(-\frac{b^2}{2G+\frac{2cb}{3}}\right) = \frac{\delta}{3T^2}.$$

It automatically leads to

$$\mathbb{P}\left\{|③| \leq b \text{ either } \sum_{t=0}^{T-1}\sigma_t^2 > G\right\} \geq 1 - \frac{\delta}{3T^2}.$$

What is more, we get

$$\sum_{t=0}^{T-1}\sigma_t^2 \leq \sum_{t=0}^{T-1} 9\gamma_t^2 d_{t+1}^2\left\|\nabla f(x_t)\right\|^2\mathbb{E}_{\xi_t}\left[\left\|\theta_t^u\right\|^2\right] \leq \sum_{t=0}^{T-1} 648\gamma_t^2 d_{t+1}^2\frac{L\Phi_0\ln\left(\frac{6(t+1)^2}{\delta}\right)}{b_t}\lambda_t^{2-\alpha}\sigma^{\alpha}$$
$$\overset{(5)}{\leq} \sum_{t=0}^{T-1}\frac{648\gamma_t^{\alpha} d_{t+1}^2 L\Phi_0\sqrt{\Phi_0/C}^{2-\alpha}\sigma^{\alpha}}{256^{2-\alpha} b_t^{2-\alpha/2}\ln^{4-5\alpha/2}\left(\frac{6(t+1)^2}{\delta}\right)}$$
$$\overset{(4)}{\leq} \sum_{t=0}^{T-1}\frac{1296 p^{\alpha}(\Phi_0, L, \sigma) d_{t+1}^{\alpha/2} L\Phi_0\sqrt{\Phi_0/C}^{2-\alpha}\sigma^{\alpha}}{256^2\cdot 4(t+1)^{\beta\alpha}\ln^{4+\alpha/2}\left(\frac{6(t+1)^2}{\delta}\right)},$$

where in the last inequality we substitute the second choice of γ_t and bound ${}^{d_{t+1}}/_{b_t} \leq 2$. What is more, $d_{t+1} = b_{t+1}$ for all $t \geq 0$. Hence, one can obtain

$$\sum_{t=0}^{T-1} \sigma_t^2 \leq \sum_{t=0}^{T-1} \frac{1296 p^{\alpha}(\Phi_0, L, \sigma) d_{t+1}^{\alpha/2} L \Phi_0 \sqrt{{}^{\Phi_0}/_{C}}^{2-\alpha} \sigma^{\alpha}}{256^2 \cdot 4 (t+1)^{\beta\alpha} \ln^{4+\alpha/2}\left(\frac{6(t+1)^2}{\delta}\right)}$$
$$\overset{(13)}{\leq} \sum_{t=0}^{T-1} \frac{1296 p^{\alpha}(\Phi_0, L, \sigma) L \Phi_0 \sqrt{{}^{\Phi_0}/_{C}}^{2-\alpha} \sigma^{\alpha}}{256^2 \cdot 4 (t+1)^{\beta\alpha} \ln^{4+\alpha/2}\left(\frac{6(t+1)^2}{\delta}\right)} \times \left(\frac{p(\Phi_0, L, \sigma) C}{256 \cdot 4^{1/\alpha}} \left(1 + \frac{(t+1)^{1-\beta} - 1}{1-\beta}\right)\right)^{\alpha/2}.$$

Noting that $(a+b)^{\alpha/2} \leq a^{\alpha/2} + b^{\alpha/2}$ due to the $\alpha/2 \leq 1$, we derive

$$\sum_{t=0}^{T-1} \sigma_t^2 \overset{L \leq C}{\leq} \sum_{t=0}^{T-1} \frac{1296 p^{3\alpha/2}(\Phi_0, L, \sigma) C^{1+\alpha/2} \Phi_0 \sqrt{{}^{\Phi_0}/_{C}}^{2-\alpha} \sigma^{\alpha}}{256^{2+\alpha/2} \cdot 4^{3/2} (t+1)^{\beta\alpha} \ln^{4+\alpha/2}\left(\frac{6(t+1)^2}{\delta}\right)}$$
$$+ \sum_{t=0}^{T-1} \frac{1296 p^{3\alpha/2}(\Phi_0, L, \sigma) C^{1+\alpha/2} \Phi_0 \sqrt{{}^{\Phi_0}/_{C}}^{2-\alpha} \sigma^{\alpha}}{256^{2+\alpha/2} \cdot 4^{3/2} (t+1)^{\beta\alpha} \ln^{4+\alpha/2}\left(\frac{6(t+1)^2}{\delta}\right)} \times \left(\frac{(t+1)^{1-\beta} - 1}{1-\beta}\right)^{\alpha/2}$$
$$\overset{(7)}{\leq} \sum_{t=0}^{T-1} \frac{1296 \Phi_0^2}{256^{2+\alpha/2} \cdot 4^{3/2} (t+1)^{\beta\alpha} \ln^{4+\alpha/2}\left(\frac{6(t+1)^2}{\delta}\right)}$$
$$+ \sum_{t=0}^{T-1} \frac{1296 \Phi_0^2}{256^{2+\alpha/2} \cdot 4^{3/2} (t+1)^{\beta\alpha} \ln^{4+\alpha/2}\left(\frac{6(t+1)^2}{\delta}\right)} \times \left(\frac{(t+1)^{1-\beta} - 1}{1-\beta}\right)^{\alpha/2},$$

where in the last inequality we substitute the third choice of $p(\Phi_0, L, \sigma)$. Applying Lemma 3, we obtain

$$\sum_{t=0}^{T-1} \sigma_t^2 \leq \sum_{t=0}^{T-1} \frac{1296 \Phi_0^2}{256^{2+\alpha/2} \cdot 4^{3/2} (t+1)^{\beta\alpha} \ln^{4+\alpha/2}\left(\frac{6(t+1)^2}{\delta}\right)}$$
$$+ \sum_{t=0}^{T-1} \frac{1296 \Phi_0^2}{256^{2+\alpha/2} \cdot 4^{3/2} (t+1)^{\beta\alpha - (1-\beta)\alpha/2} \ln^{4}\left(\frac{6(t+1)^2}{\delta}\right)}.$$

Therefore, conditions over β can be formulated as follows:

Condition 3:

$$\begin{cases} \beta\alpha \geq 1 \Rightarrow \beta \geq \frac{1}{\alpha}; \\ \beta\alpha - (1-\beta)\alpha/2 \geq 1 \Rightarrow \beta \geq \frac{2+\alpha}{3\alpha}. \end{cases} \tag{18}$$

Consequently, with (18), we have

$$\sum_{t=0}^{T-1} \sigma_t^2 \overset{(18)}{\leq} \sum_{t=0}^{T-1} \frac{2592 \Phi_0^2}{256^{2+\alpha/2} \cdot 4^{3/2} (t+1)} \overset{(16)}{\leq} \frac{3888 \Phi_0^2 \ln\left(\frac{6T^2}{\delta}\right)}{256^{2+\alpha/2} \cdot 4^{3/2}} \leq \frac{\Phi_0^2 \ln\left(\frac{6T^2}{\delta}\right)}{1152}.$$

Bound for ④. The event E_{T-1} implies

$$
\begin{aligned}
-\sum_{t=0}^{T-1}(\gamma_t d_{t+1}-2e_t)\left\langle\nabla f(x_t),\theta_t^b\right\rangle &\le \sum_{t=0}^{T-1}3\gamma_t d_{t+1}\left\|\nabla f(x_t)\right\|\left\|\theta_t^b\right\| \\
&\le \sum_{t=0}^{T-1}6\gamma_t d_{t+1}\sqrt{\frac{L\Phi_0\ln\left(\frac{6(t+1)^2}{\delta}\right)}{b_t}}\frac{2^\alpha\sigma^\alpha}{\lambda_t^{\alpha-1}} \\
&\overset{(5)}{\le} \sum_{t=0}^{T-1}\frac{24\cdot 256^{\alpha-1}\gamma_t^\alpha d_{t+1}\sqrt{L\Phi_0}\sigma^\alpha}{\sqrt{\Phi_0/C}^{\alpha-1}b_t^{1-\alpha/2}}\ln^{5\alpha/2-2}\left(\frac{6(t+1)^2}{\delta}\right) \\
&\le \sum_{t=0}^{T-1}\frac{24\cdot 256^{\alpha-1}\gamma_t^\alpha d_{t+1}C\sqrt{\Phi_0/C}^{2-\alpha}\sigma^\alpha}{b_t^{1-\alpha/2}}\ln^{5\alpha/2-2}\left(\frac{6(t+1)^2}{\delta}\right) \\
&\le \sum_{t=0}^{T-1}48\cdot 256^{\alpha-1}\gamma_t^\alpha d_{t+1}^{\alpha/2}C\sqrt{\Phi_0/C}^{2-\alpha}\sigma^\alpha\ln^{5\alpha/2-2}\left(\frac{6(t+1)^2}{\delta}\right),
\end{aligned}
$$

where in the last inequality we apply $d_{t+1}/b_t \le 2$. Thus, we have

$$
\begin{aligned}
④ &\le \sum_{t=0}^{T-1}48\cdot 256^{\alpha-1}\gamma_t^\alpha d_{t+1}^{\alpha/2}C\sqrt{\Phi_0/C}^{2-\alpha}\sigma^\alpha\ln^{5\alpha/2-2}\left(\frac{6(t+1)^2}{\delta}\right) \\
&\overset{(4)}{\le} \sum_{t=0}^{T-1}\frac{48\cdot 256^{\alpha-1}p^\alpha(\Phi_0,L,\sigma)d_{t+1}^{\alpha/2}C\sqrt{\Phi_0/C}^{2-\alpha}\sigma^\alpha}{(256\cdot 4^{1/\alpha})^\alpha(t+1)^{\beta\alpha}\ln^{2+\alpha/2}\left(\frac{6(t+1)^2}{\delta}\right)} \\
&\overset{d_{t+1}=b_{t+1},(13)}{\le} \sum_{t=0}^{T-1}\frac{48\cdot 256^{\alpha-1}p^\alpha(\Phi_0,L,\sigma)C\sqrt{\Phi_0/C}^{2-\alpha}\sigma^\alpha}{(256\cdot 4^{1/\alpha})^\alpha(t+1)^{\beta\alpha}\ln^{2+\alpha/2}\left(\frac{6(t+1)^2}{\delta}\right)} \\
&\times\left(\frac{p(\Phi_0,L,\sigma)C}{256\cdot 4^{1/\alpha}}\left(1+\frac{(t+1)^{1-\beta}-1}{1-\beta}\right)\right)^{\alpha/2} \\
&\le \sum_{t=0}^{T-1}\frac{48\cdot 256^{\alpha-1}p^{3\alpha/2}(\Phi_0,L,\sigma)C^{1+\alpha/2}\sqrt{\Phi_0/C}^{2-\alpha}\sigma^\alpha}{(256\cdot 4^{1/\alpha})^{3\alpha/2}(t+1)^{\beta\alpha}\ln^{2+\alpha/2}\left(\frac{6(t+1)^2}{\delta}\right)} \\
&+\sum_{t=0}^{T-1}\frac{48\cdot 256^{\alpha-1}p^{3\alpha/2}(\Phi_0,L,\sigma)C^{1+\alpha/2}\sqrt{\Phi_0/C}^{2-\alpha}\sigma^\alpha}{(256\cdot 4^{1/\alpha})^{3\alpha/2}(t+1)^{\beta\alpha}\ln^{2+\alpha/2}\left(\frac{6(t+1)^2}{\delta}\right)}\times\left(\frac{(t+1)^{1-\beta}-1}{1-\beta}\right)^{\alpha/2},
\end{aligned}
$$

where in the last inequality we use $(a+b)^{\alpha/2}\le a^{\alpha/2}+b^{\alpha/2}$. Therefore, substituting the third choice of $p(\Phi_0,L,\sigma)$ (7), one can obtain

$$
\begin{aligned}
④ &\le \sum_{t=0}^{T-1}\frac{48\cdot 256^{\alpha-1}\Phi_0}{(256\cdot 4^{1/\alpha})^{3\alpha/2}(t+1)^{\beta\alpha}\ln^{2+\alpha/2}\left(\frac{6(t+1)^2}{\delta}\right)} \\
&+\sum_{t=0}^{T-1}\frac{48\cdot 256^{\alpha-1}\Phi_0}{(256\cdot 4^{1/\alpha})^{3\alpha/2}(t+1)^{\beta\alpha}\ln^{2+\alpha/2}\left(\frac{6(t+1)^2}{\delta}\right)}\times\left(\frac{(t+1)^{1-\beta}-1}{1-\beta}\right)^{\alpha/2}.
\end{aligned}
$$

Applying Lemma 3 to the second term, we get

$$
\begin{aligned}
④ &\le \sum_{t=0}^{T-1}\frac{48\cdot 256^{\alpha-1}\Phi_0}{(256\cdot 4^{1/\alpha})^{3\alpha/2}(t+1)^{\beta\alpha}\ln^{2+\alpha/2}\left(\frac{6(t+1)^2}{\delta}\right)} \\
&+\sum_{t=0}^{T-1}\frac{48\cdot 256^{\alpha-1}\Phi_0}{(256\cdot 4^{1/\alpha})^{3\alpha/2}(t+1)^{\beta\alpha-(1-\beta)\alpha/2}\ln^{2}\left(\frac{6(t+1)^2}{\delta}\right)}.
\end{aligned}
$$

To derive a sufficient boundary, it is enough to apply next conditions over β:

Condition 4:

$$\begin{cases}\beta\alpha \geq 1 \Rightarrow \beta \geq \frac{1}{\alpha};\\ \beta\alpha - (1-\beta)\alpha/2 \geq 1 \Rightarrow \beta \geq \frac{2+\alpha}{3\alpha}.\end{cases} \tag{19}$$

Hence, we have

$$④ \overset{(19)}{\leq} \sum_{t=0}^{T-1} \frac{96 \cdot 256^{\alpha-1}\Phi_0}{(256 \cdot 4^{1/\alpha})^{3\alpha/2}(t+1)} \overset{(16)}{\leq} \frac{144 \cdot 256^{\alpha-1}\Phi_0 \ln\left(\frac{6T^2}{\delta}\right)}{(256 \cdot 4^{1/\alpha})^{3\alpha/2}} \leq \frac{\Phi_0 \ln\left(\frac{6T^2}{\delta}\right)}{12}.$$

Bound for ⑤. First of all, we have

$$\mathbb{E}_{\xi_t}\left[2e_t\left(\|\theta_t^u\|^2 - \mathbb{E}_{\xi_t}\left[\|\theta_t^u\|^2\right]\right)\right] = 0$$

Moreover, we have

$$\left|2e_t\left(\|\theta_t^u\|^2 - \mathbb{E}_{\xi_t}\left[\|\theta_t^u\|^2\right]\right)\right| \leq 8e_t\lambda_t^2 = 4(Ld_{t+1}+C)\gamma_t^2\lambda_t^2 \overset{(5)}{=} \frac{4(Ld_{t+1}+C)\Phi_0/C}{256^2 b_t \ln^5\left(\frac{6(t+1)^2}{\delta}\right)}$$

$$\leq \frac{4(d_{t+1}+1)\Phi_0}{256^2 b_t \ln^5\left(\frac{6(t+1)^2}{\delta}\right)} \overset{d_{t+1}/b_t \leq 2,(13)}{\leq} \frac{\Phi_0}{8} := c.$$

Let us define $\sigma_t^2 = \mathbb{E}_{\xi_t}\left[4e_t^2\left(\|\theta_t^u\|^2 - \mathbb{E}_{\xi_t}\left[\|\theta_t^u\|^2\right]\right)^2\right]$. Consequently, it can be bounded as follows:

$$\sigma_t^2 = \mathbb{E}_{\xi_t}\left[4e_t^2\left(\|\theta_t^u\|^2 - \mathbb{E}_{\xi_t}\left[\|\theta_t^u\|^2\right]\right)^2\right] \leq c\mathbb{E}_{\xi_t}\left[2e_t\left|\left(\|\theta_t^u\|^2 - \mathbb{E}_{\xi_t}\left[\|\theta_t^u\|^2\right]\right)\right|\right]$$

$$\leq 4ce_t\mathbb{E}_{\xi_t}\left[\|\theta_t^u\|^2\right].$$

Therefore, we can apply Bernstein's inequality with $b = \Phi_0 \ln\left(\frac{6T^2}{\delta}\right)/8$ and $G = \Phi_0^2 \ln\left(\frac{6T^2}{\delta}\right)/384$:

$$\mathbb{P}\left\{|⑤| > b \text{ and } \sum_{t=0}^{T-1}\sigma_t^2 \leq G\right\} \leq 2\exp\left(-\frac{b^2}{2G + \frac{2cb}{3}}\right) = \frac{\delta}{3T^2}.$$

Hence, we derive

$$\mathbb{P}\left\{|⑤| \leq b \text{ either } \sum_{t=0}^{T-1}\sigma_t^2 > G\right\} \geq 1 - \frac{\delta}{3T^2}.$$

What is more, we get

$$\sum_{t=0}^{T-1}\sigma_t^2 \leq \sum_{t=0}^{T-1} 4ce_t\mathbb{E}_{\xi_t}\left[\|\theta_t^u\|^2\right] \leq \sum_{t=0}^{T-1} 36c(Ld_{t+1}+C)\gamma_t^2\lambda_t^{2-\alpha}\sigma^\alpha$$

$$\overset{(5)}{=} \sum_{t=0}^{T-1} \frac{36c(Ld_{t+1}+C)\gamma_t^\alpha\sigma^\alpha\sqrt{\Phi_0/C}^{2-\alpha}}{256^{2-\alpha}b_t^{1-\alpha/2}\ln^{5-5\alpha/2}\left(\frac{6(t+1)^2}{\delta}\right)}$$

$$\leq \sum_{t=0}^{T-1} \frac{36cC(d_{t+1}+1)\gamma_t^\alpha\sigma^\alpha\sqrt{\Phi_0/C}^{2-\alpha}}{256^{2-\alpha}b_t^{1-\alpha/2}\ln^{5-5\alpha/2}\left(\frac{6(t+1)^2}{\delta}\right)}$$

$$= \sum_{t=0}^{T-1} \frac{36cCd_{t+1}\gamma_t^\alpha\sigma^\alpha\sqrt{\Phi_0/C}^{2-\alpha}}{256^{2-\alpha}b_t^{1-\alpha/2}\ln^{5-5\alpha/2}\left(\frac{6(t+1)^2}{\delta}\right)} + \sum_{t=0}^{T-1} \frac{36cC\gamma_t^\alpha\sigma^\alpha\sqrt{\Phi_0/C}^{2-\alpha}}{256^{2-\alpha}b_t^{1-\alpha/2}\ln^{5-5\alpha/2}\left(\frac{6(t+1)^2}{\delta}\right)}$$

$$\overset{d_{t+1}/b_t \leq 2,(13)}{\leq} \sum_{t=0}^{T-1} \frac{72cCd_{t+1}^{\alpha/2}\gamma_t^\alpha\sigma^\alpha\sqrt{\Phi_0/C}^{2-\alpha}}{256^{2-\alpha}\ln^{5-5\alpha/2}\left(\frac{6(t+1)^2}{\delta}\right)} + \sum_{t=0}^{T-1} \frac{36 \cdot 1024^{1-\alpha/2}cC\gamma_t^\alpha\sigma^\alpha\sqrt{\Phi_0/C}^{2-\alpha}\ln^{3-3\alpha/2}\left(\frac{6}{\delta}\right)}{256^{2-\alpha}\ln^{5-5\alpha/2}\left(\frac{6(t+1)^2}{\delta}\right)},$$

where in the last inequality we apply $d_{t+1} \leq 2b_t$ for the first term, and substitute the lower bound for b_t (see (13)) in the second term. Consequently, substituting the stepsize (4), we get

$$\begin{aligned}
\sum_{t=0}^{T-1}\sigma_t^2 &\overset{(4)}{\leq} \sum_{t=0}^{T-1}\frac{72cCp^{\alpha}(\Phi_0,L,\sigma)d_{t+1}^{\alpha/2}\sigma^{\alpha}\sqrt{\Phi_0/C}^{2-\alpha}}{256^2\cdot 4(t+1)^{\beta\alpha}\ln^{5+\alpha/2}\left(\frac{6(t+1)^2}{\delta}\right)} \\
&\quad+\sum_{t=0}^{T-1}\frac{36\cdot 1024^{1-\alpha/2}cCp^{\alpha}(\Phi_0,L,\sigma)\sigma^{\alpha}\sqrt{\Phi_0/C}^{2-\alpha}\ln^{3-3\alpha/2}\left(\frac{6}{\delta}\right)}{256^2\cdot 4(t+1)^{\beta\alpha}\ln^{5+\alpha/2}\left(\frac{6(t+1)^2}{\delta}\right)} \\
&\overset{(13)}{\leq} \sum_{t=0}^{T-1}\frac{72cCp^{\alpha}(\Phi_0,L,\sigma)\sigma^{\alpha}\sqrt{\Phi_0/C}^{2-\alpha}}{256^2\cdot 4(t+1)^{\beta\alpha}\ln^{5+\alpha/2}\left(\frac{6(t+1)^2}{\delta}\right)}\times\left(\frac{p(\Phi_0,L,\sigma)C}{256\cdot 4^{1/\alpha}}\left(1+\frac{(t+1)^{1-\beta}-1}{1-\beta}\right)\right)^{\alpha/2} \\
&\quad+\sum_{t=0}^{T-1}\frac{36\cdot 1024^{1-\alpha/2}cCp^{\alpha}(\Phi_0,L,\sigma)\sigma^{\alpha}\sqrt{\Phi_0/C}^{2-\alpha}\ln^{3-3\alpha/2}\left(\frac{6}{\delta}\right)}{256^2\cdot 4(t+1)^{\beta\alpha}\ln^{5+\alpha/2}\left(\frac{6(t+1)^2}{\delta}\right)},
\end{aligned}$$

where we used the upper bound for d_{t+1} (see also (13)). Using the fact that $(a+b)^{\alpha/2}\leq a^{\alpha/2}+b^{\alpha/2}$ and applying Lemma 3, we get

$$\begin{aligned}
\sum_{t=0}^{T-1}\sigma_t^2 &\leq \sum_{t=0}^{T-1}\frac{72cC^{1+\alpha/2}p^{3\alpha/2}(\Phi_0,L,\sigma)\sigma^{\alpha}\sqrt{\Phi_0/C}^{2-\alpha}}{256^{2+\alpha/2}\cdot 4^{3/2}(t+1)^{\beta\alpha}\ln^{5+\alpha/2}\left(\frac{6(t+1)^2}{\delta}\right)} \\
&\quad+\sum_{t=0}^{T-1}\frac{72cC^{1+\alpha/2}p^{3\alpha/2}(\Phi_0,L,\sigma)\sigma^{\alpha}\sqrt{\Phi_0/C}^{2-\alpha}}{256^{2+\alpha/2}\cdot 4^{3/2}(t+1)^{\beta\alpha}\ln^{5+\alpha/2}\left(\frac{6(t+1)^2}{\delta}\right)}\times\left(\frac{(t+1)^{1-\beta}-1}{1-\beta}\right)^{\alpha/2} \\
&\quad+\sum_{t=0}^{T-1}\frac{36\cdot 1024^{1-\alpha/2}cCp^{\alpha}(\Phi_0,L,\sigma)\sigma^{\alpha}\sqrt{\Phi_0/C}^{2-\alpha}\ln^{3-3\alpha/2}\left(\frac{6}{\delta}\right)}{256^2\cdot 4(t+1)^{\beta\alpha}\ln^{5+\alpha/2}\left(\frac{6(t+1)^2}{\delta}\right)} \\
&\leq \sum_{t=0}^{T-1}\frac{72cC^{1+\alpha/2}p^{3\alpha/2}(\Phi_0,L,\sigma)\sigma^{\alpha}\sqrt{\Phi_0/C}^{2-\alpha}}{256^{2+\alpha/2}\cdot 4^{3/2}(t+1)^{\beta\alpha}\ln^{5+\alpha/2}\left(\frac{6(t+1)^2}{\delta}\right)} \\
&\quad+\sum_{t=0}^{T-1}\frac{72cCp^{\alpha}(\Phi_0,L,\sigma)\sigma^{\alpha}\sqrt{\Phi_0/C}^{2-\alpha}}{256^{2+\alpha/2}\cdot 4^{3/2}(t+1)^{\beta\alpha-(1-\beta)\alpha/2}\ln^{5}\left(\frac{6(t+1)^2}{\delta}\right)} \\
&\quad+\sum_{t=0}^{T-1}\frac{36\cdot 1024^{1-\alpha/2}cC^{1+\alpha/2}p^{3\alpha/2}(\Phi_0,L,\sigma)\sigma^{\alpha}\sqrt{\Phi_0/C}^{2-\alpha}\ln^{3-3\alpha/2}\left(\frac{6}{\delta}\right)}{256^2\cdot 4(t+1)^{\beta\alpha}\ln^{5+\alpha/2}\left(\frac{6(t+1)^2}{\delta}\right)}.
\end{aligned}$$

Substitution of $p(\Phi_0,L,\sigma)$ gives

$$\begin{aligned}
\sum_{t=0}^{T-1}\sigma_t^2 &\leq \sum_{t=0}^{T-1}\frac{72c\Phi_0}{256^{2+\alpha/2}\cdot 4^{3/2}(t+1)^{\beta\alpha}\ln^{5+\alpha/2}\left(\frac{6(t+1)^2}{\delta}\right)} \\
&\quad+\sum_{t=0}^{T-1}\frac{72c\Phi_0}{256^{2+\alpha/2}\cdot 4^{3/2}(t+1)^{\beta\alpha-(1-\beta)\alpha/2}\ln^{5}\left(\frac{6(t+1)^2}{\delta}\right)} \\
&\quad+\sum_{t=0}^{T-1}\frac{36\cdot 1024^{1-\alpha/2}c\Phi_0\ln^{3-3\alpha/2}\left(\frac{6}{\delta}\right)}{256^2\cdot 4(t+1)^{\beta\alpha}\ln^{5+\alpha/2}\left(\frac{6(t+1)^2}{\delta}\right)}.
\end{aligned}$$

Thus, if we choose β as follows:

Condition 5:

$$\begin{cases}\beta\alpha \geq 1 \Rightarrow \beta \geq \frac{1}{\alpha};\\ \beta\alpha - (1-\beta)\alpha/2 \geq 1 \Rightarrow \beta \geq \frac{2+\alpha}{3\alpha},\end{cases} \tag{20}$$

we derive

$$\begin{aligned}\sum_{t=0}^{T-1}\sigma_t^2 &\overset{(20)}{\leq} \sum_{t=0}^{T-1}\frac{72c\Phi_0}{256^{2+\alpha/2}\cdot 4^{3/2}(t+1)} + \sum_{t=0}^{T-1}\frac{72c\Phi_0}{256^{2+\alpha/2}\cdot 4^{3/2}(t+1)}\\ &\quad + \sum_{t=0}^{T-1}\frac{36\cdot 1024^{1-\alpha/2}c\Phi_0}{256^2\cdot 4(t+1)}\\ &\overset{(16)}{\leq} 3c\Phi_0\max\left\{\frac{144}{256^{2+\alpha/2}\cdot 4^{3/2}}, \frac{36\cdot 1024^{1-\alpha/2}}{256^2\cdot 4}\right\}\ln\left(\frac{6T^2}{\delta}\right) \leq \frac{\Phi_0^2\ln\left(\frac{6T^2}{\delta}\right)}{384}.\end{aligned}$$

Bound for ⑥. The event E_{T-1} implies

$$\begin{aligned}\sum_{t=0}^{T-1}2e_t\mathbb{E}_{\xi_t}\left[\|\theta_t^u\|^2\right] &\leq \sum_{t=0}^{T-1}C(d_{t+1}+1)\gamma_t^2\mathbb{E}_{\xi_t}\left[\|\theta_t^u\|^2\right] \leq \sum_{t=0}^{T-1}18C(d_{t+1}+1)\gamma_t^2\lambda_t^{2-\alpha}\sigma^\alpha\\ &\overset{(5)}{=} \sum_{t=0}^{T-1}\frac{18C\sqrt{\Phi_0/C}^{2-\alpha}(d_{t+1}+1)\gamma_t^\alpha\sigma^\alpha}{256^{2-\alpha}b_t^{1-\alpha/2}\ln^{5-5\alpha/2}\left(\frac{6(t+1)^2}{\delta}\right)}\\ &\overset{d_{t+1}/b_t\leq 2,(13)}{\leq} \sum_{t=0}^{T-1}\frac{36C\sqrt{\Phi_0/C}^{2-\alpha}d_{t+1}^{\alpha/2}\gamma_t^\alpha\sigma^\alpha}{256^{2-\alpha}\ln^{5-5\alpha/2}\left(\frac{6(t+1)^2}{\delta}\right)}\\ &\quad + \sum_{t=0}^{T-1}\frac{18\cdot 1024^{1-\alpha/2}C\sqrt{\Phi_0/C}^{2-\alpha}\gamma_t^\alpha\sigma^\alpha\ln^{3-3\alpha/2}\left(\frac{6}{\delta}\right)}{256^{2-\alpha}\ln^{5-5\alpha/2}\left(\frac{6(t+1)^2}{\delta}\right)}.\end{aligned}$$

Substitution of (4) gives

$$\begin{aligned}\sum_{t=0}^{T-1}2e_t\mathbb{E}_{\xi_t}\left[\|\theta_t^u\|^2\right] &\leq \sum_{t=0}^{T-1}\frac{36C\sqrt{\Phi_0/C}^{2-\alpha}d_{t+1}^{\alpha/2}\gamma_t^\alpha\sigma^\alpha}{256^{2-\alpha}\ln^{5-5\alpha/2}\left(\frac{6(t+1)^2}{\delta}\right)}\\ &\quad + \sum_{t=0}^{T-1}\frac{18\cdot 1024^{1-\alpha/2}C\sqrt{\Phi_0/C}^{2-\alpha}\gamma_t^\alpha\sigma^\alpha\ln^{3-3\alpha/2}\left(\frac{6}{\delta}\right)}{256^{2-\alpha}\ln^{5-5\alpha/2}\left(\frac{6(t+1)^2}{\delta}\right)}\\ &\overset{(4)}{\leq} \sum_{t=0}^{T-1}\frac{36Cp^\alpha(\Phi_0,L,\sigma)\sqrt{\Phi_0/C}^{2-\alpha}d_{t+1}^{\alpha/2}\sigma^\alpha}{256^2\cdot 4(t+1)^{\beta\alpha}\ln^{5+\alpha/2}\left(\frac{6(t+1)^2}{\delta}\right)}\\ &\quad + \sum_{t=0}^{T-1}\frac{18\cdot 1024^{1-\alpha/2}Cp^\alpha(\Phi_0,L,\sigma)\sqrt{\Phi_0/C}^{2-\alpha}\sigma^\alpha\ln^{3-3\alpha/2}\left(\frac{6}{\delta}\right)}{256^2\cdot 4(t+1)^{\beta\alpha}\ln^{5+\alpha/2}\left(\frac{6(t+1)^2}{\delta}\right)}.\end{aligned}$$

Using (13) as the upper bound for d_{t+1}, we obtain

$$\begin{aligned}\sum_{t=0}^{T-1}2e_t\mathbb{E}_{\xi_t}\left[\|\theta_t^u\|^2\right] &\leq \sum_{t=0}^{T-1}\frac{36Cp^\alpha(\Phi_0,L,\sigma)\sqrt{\Phi_0/C}^{2-\alpha}\sigma^\alpha}{256^2\cdot 4(t+1)^{\beta\alpha}\ln^{5+\alpha/2}\left(\frac{6(t+1)^2}{\delta}\right)}\\ &\quad\times\left(\frac{p(\Phi_0,L,\sigma)C}{256\cdot 4^{1/\alpha}}\left(1+\frac{(t+1)^{1-\beta}-1}{1-\beta}\right)\right)^{\alpha/2}\\ &\quad + \sum_{t=0}^{T-1}\frac{18\cdot 1024^{1-\alpha/2}Cp^\alpha(\Phi_0,L,\sigma)\sqrt{\Phi_0/C}^{2-\alpha}\sigma^\alpha\ln^{3-3\alpha/2}\left(\frac{6}{\delta}\right)}{256^2\cdot 4(t+1)^{\beta\alpha}\ln^{5+\alpha/2}\left(\frac{6(t+1)^2}{\delta}\right)}.\end{aligned}$$

Using that $(a+b)^{\alpha/2} \leq a^{\alpha/2} + b^{\alpha/2}$ and applying Lemma 3, we get

$$\begin{aligned}\sum_{t=0}^{T-1} `2e_t \mathbb{E}_{\xi_t}\left[\|\theta_t^u\|^2\right] \leq &\sum_{t=0}^{T-1} \frac{36C^{1+\alpha/2}p^{3\alpha/2}(\Phi_0, L, \sigma)\sqrt{\Phi_0/C}^{2-\alpha}\sigma^\alpha}{256^{2+\alpha/2}\cdot 4^{3/2}(t+1)^{\beta\alpha}\ln^{5+\alpha/2}\left(\frac{6(t+1)^2}{\delta}\right)} \\ &+ \sum_{t=0}^{T-1} \frac{36C^{1+\alpha/2}p^{3\alpha/2}(\Phi_0, L, \sigma)\sqrt{\Phi_0/C}^{2-\alpha}\sigma^\alpha}{256^{2+\alpha/2}\cdot 4^{3/2}(t+1)^{\beta\alpha-(1-\beta)\alpha/2}\ln^{5}\left(\frac{6(t+1)^2}{\delta}\right)} \\ &+ \sum_{t=0}^{T-1} \frac{18\cdot 1024^{1-\alpha/2}Cp^\alpha(\Phi_0, L, \sigma)\sqrt{\Phi_0/C}^{2-\alpha}\sigma^\alpha \ln^{3-3\alpha/2}\left(\frac{6}{\delta}\right)}{256^2\cdot 4(t+1)^{\beta\alpha}\ln^{5+\alpha/2}\left(\frac{6(t+1)^2}{\delta}\right)}.\end{aligned}$$

Then, if we substitute the choice of $p(\Phi_0, L, \sigma)$ (7), we have

$$\begin{aligned}\sum_{t=0}^{T-1} `2e_t \mathbb{E}_{\xi_t}\left[\|\theta_t^u\|^2\right] \leq &\sum_{t=0}^{T-1} \frac{36\Phi_0}{256^{2+\alpha/2}\cdot 4^{3/2}(t+1)^{\beta\alpha}\ln^{5+\alpha/2}\left(\frac{6(t+1)^2}{\delta}\right)} \\ &+ \sum_{t=0}^{T-1} \frac{36\Phi_0}{256^{2+\alpha/2}\cdot 4^{3/2}(t+1)^{\beta\alpha-(1-\beta)\alpha/2}\ln^{5}\left(\frac{6(t+1)^2}{\delta}\right)} \\ &+ \sum_{t=0}^{T-1} \frac{18\cdot 1024^{1-\alpha/2}\Phi_0}{256^2\cdot 4(t+1)^{\beta\alpha}\ln^{2+2\alpha}\left(\frac{6(t+1)^2}{\delta}\right)}.\end{aligned}$$

Therefore, if we choose β in the following way:

Condition 6:

$$\begin{cases}\beta\alpha \geq 1 \Rightarrow \beta \geq \frac{1}{\alpha}; \\ \beta\alpha - (1-\beta)\alpha/2 \geq 1 \Rightarrow \beta \geq \frac{2+\alpha}{3\alpha},\end{cases} \tag{21}$$

one can obtain

$$\begin{aligned}\sum_{t=0}^{T-1} `2e_t \mathbb{E}_{\xi_t}\left[\|\theta_t^u\|^2\right] &\overset{(21)}{\leq} \sum_{t=0}^{T-1} \frac{36\Phi_0}{256^{2+\alpha/2}\cdot 4^{3/2}(t+1)} \\ &+ \sum_{t=0}^{T-1} \frac{36\Phi_0}{256^{2+\alpha/2}\cdot 4^{3/2}(t+1)} \\ &+ \sum_{t=0}^{T-1} \frac{18\cdot 1024^{1-\alpha/2}\Phi_0}{256^2\cdot 4(t+1)} \\ &\overset{(16)}{\leq} 3\Phi_0 \max\left\{\frac{72}{256^{2+\alpha/2}\cdot 4^{3/2}}, \frac{18\cdot 1024^{1-\alpha/2}}{256^2\cdot 4}\right\}\ln\left(\frac{6T^2}{\delta}\right) \\ &\leq \frac{\Phi_0 \ln\left(\frac{6T^2}{\delta}\right)}{16}.\end{aligned}$$

Bound for ⑦. According to the event E_{T-1}, we have

$$\begin{aligned}
\sum_{t=0}^{T-1} 2e_t \left\|\theta_t^b\right\|^2 &\leq \sum_{t=0}^{T-1} C(d_{t+1}+1)\gamma_t^2 \frac{4^\alpha \sigma^{2\alpha}}{\lambda_t^{2\alpha-2}} \\
&\overset{(5)}{=} \sum_{t=0}^{T-1} \frac{4^\alpha \cdot 256^{2\alpha-2} C(d_{t+1}+1)\gamma_t^{2\alpha}\sigma^{2\alpha} b_t^{\alpha-1} \ln^{5\alpha-5}\left(\frac{6(t+1)^2}{\delta}\right)}{\sqrt{\Phi_0/C}^{2\alpha-2}} \\
&= \frac{4^\alpha \cdot 256^{2\alpha-2} C(d_1+1)\gamma_0^{2\alpha}\sigma^{2\alpha} \ln^{5\alpha-5}\left(\frac{6}{\delta}\right)}{\sqrt{\Phi_0/C}^{2\alpha-2}} \\
&\quad + \sum_{t=1}^{T-1} \frac{4^\alpha \cdot C_3^{2\alpha-2} C(d_{t+1}+1)\gamma_t^{2\alpha}\sigma^{2\alpha} b_t^{\alpha-1} \ln^{5\alpha-5}\left(\frac{6(t+1)^2}{\delta}\right)}{\sqrt{\Phi_0/C}^{2\alpha-2}} \\
&\leq \frac{4^\alpha \cdot 256^{2\alpha-2} C(d_1+1)\gamma_0^{2\alpha}\sigma^{2\alpha} \ln^{5\alpha-5}\left(\frac{6}{\delta}\right)}{\sqrt{\Phi_0/C}^{2\alpha-2}} \\
&\quad + \sum_{t=1}^{T-1} \frac{4^\alpha \cdot 256^{2\alpha-2} C(d_{t+1}+1)\gamma_t^{2\alpha}\sigma^{2\alpha} d_{t+1}^{\alpha-1} \ln^{5\alpha-5}\left(\frac{6(t+1)^2}{\delta}\right)}{\sqrt{\Phi_0/C}^{2\alpha-2}},
\end{aligned}$$

where in the last inequality we use that $b_t = d_t \leq d_{t+1}$ for all $t \geq 1$. Hence, one can obtain

$$\begin{aligned}
\sum_{t=0}^{T-1} 2e_t \left\|\theta_t^b\right\|^2 &\leq \frac{4^\alpha \cdot 256^{2\alpha-2} C(d_1+1)\gamma_0^{2\alpha}\sigma^{2\alpha} \ln^{5\alpha-5}\left(\frac{6}{\delta}\right)}{\sqrt{\Phi_0/C}^{2\alpha-2}} \\
&\quad + \sum_{t=1}^{T-1} \frac{4^\alpha \cdot 256^{2\alpha-2} C d_{t+1}^{\alpha}\gamma_t^{2\alpha}\sigma^{2\alpha} \ln^{5\alpha-5}\left(\frac{6(t+1)^2}{\delta}\right)}{\sqrt{\Phi_0/C}^{2\alpha-2}} \\
&\quad + \sum_{t=1}^{T-1} \frac{4^\alpha \cdot 256^{2\alpha-2} C d_{t+1}^{\alpha-1}\gamma_t^{2\alpha}\sigma^{2\alpha} \ln^{5\alpha-5}\left(\frac{6(t+1)^2}{\delta}\right)}{\sqrt{\Phi_0/C}^{2\alpha-2}},
\end{aligned}$$

where we divide $(d_{t+1}+1)$ into two terms. As a result, substituting (13) to bound d_{t+1}, we get

$$\begin{aligned}
\sum_{t=0}^{T-1} 2e_t \left\|\theta_t^b\right\|^2 &\leq \frac{4^\alpha \cdot 256^{2\alpha-2} C(d_1+1)\gamma_0^{2\alpha}\sigma^{2\alpha} \ln^{5\alpha-5}\left(\frac{6}{\delta}\right)}{\sqrt{\Phi_0/C}^{2\alpha-2}} \\
&\quad + \sum_{t=1}^{T-1} \frac{4^\alpha \cdot 256^{2\alpha-2} C \gamma_t^{2\alpha}\sigma^{2\alpha} \ln^{5\alpha-5}\left(\frac{6(t+1)^2}{\delta}\right)}{\sqrt{\Phi_0/C}^{2\alpha-2}} \\
&\quad \times \left(\frac{p(\Phi_0, L, \sigma)C}{256 \cdot 4^{1/\alpha}}\left(1 + \frac{(t+1)^{1-\beta}-1}{1-\beta}\right)\right)^{\alpha} \\
&\quad + \sum_{t=1}^{T-1} \frac{4^\alpha \cdot 256^{2\alpha-2} C \gamma_t^{2\alpha}\sigma^{2\alpha} \ln^{5\alpha-5}\left(\frac{6(t+1)^2}{\delta}\right)}{\sqrt{\Phi_0/C}^{2\alpha-2}} \\
&\quad \times \left(\frac{p(\Phi_0, L, \sigma)C}{256 \cdot 4^{1/\alpha}}\left(1 + \frac{(t+1)^{1-\beta}-1}{1-\beta}\right)\right)^{\alpha-1}.
\end{aligned}$$

Substitution of (4) gives

$$\sum_{t=0}^{T-1} 2e_t\left\|\theta_t^b\right\|^2 \overset{(4)}{\leq} \frac{4^\alpha \cdot 256^{2\alpha-2} C(d_1+1)p^{2\alpha}(\Phi_0,L,\sigma)\sigma^{2\alpha}}{(256)^{2\alpha}\cdot 16\sqrt{\Phi_0/C}^{2\alpha-2}\ln^{5+\alpha}\left(\frac{6}{\delta}\right)}$$
$$+\sum_{t=1}^{T-1}\frac{4^\alpha \cdot 256^{2\alpha-2} Cp^{2\alpha}(\Phi_0,L,\sigma)\sigma^{2\alpha}}{(256)^{2\alpha}\cdot 16(t+1)^{2\beta\alpha}\sqrt{\Phi_0/C}^{2\alpha-2}\ln^{5+\alpha}\left(\frac{6(t+1)^2}{\delta}\right)}$$
$$\times\left(\frac{p(\Phi_0,L,\sigma)C}{256\cdot 4^{1/\alpha}}\left(1+\frac{(t+1)^{1-\beta}-1}{1-\beta}\right)\right)^\alpha$$
$$+\sum_{t=1}^{T-1}\frac{4^\alpha \cdot 256^{2\alpha-2} Cp^{2\alpha}(\Phi_0,L,\sigma)\sigma^{2\alpha}}{(256)^{2\alpha}\cdot 16(t+1)^{2\beta\alpha}\sqrt{\Phi_0/C}^{2\alpha-2}\ln^{5+\alpha}\left(\frac{6(t+1)^2}{\delta}\right)}$$
$$\times\left(\frac{p(\Phi_0,L,\sigma)C}{256\cdot 4^{1/\alpha}}\left(1+\frac{(t+1)^{1-\beta}-1}{1-\beta}\right)\right)^{\alpha-1}.$$

Abusing facts that $(a+b)^{\alpha-1}\leq a^{\alpha-1}+b^{\alpha-1}$ since $\alpha-1\leq 1$ and $(a+b)^\alpha\leq 2a^\alpha+2b^\alpha$ since $\alpha\leq 2$ allows to derive

$$\sum_{t=0}^{T-1} 2e_t\left\|\theta_t^b\right\|^2 \leq \frac{4^\alpha \cdot 256^{2\alpha-2} C(d_1+1)p^{2\alpha}(\Phi_0,L,\sigma)\sigma^{2\alpha}}{(256)^{2\alpha}\cdot 16\sqrt{\Phi_0/C}^{2\alpha-2}\ln^{5+\alpha}\left(\frac{6}{\delta}\right)}$$
$$+\sum_{t=1}^{T-1}\frac{2\cdot 4^\alpha \cdot 256^{2\alpha-2} C^{1+\alpha}p^{3\alpha}(\Phi_0,L,\sigma)\sigma^{2\alpha}}{(256)^{3\alpha}\cdot 64(t+1)^{2\beta\alpha}\sqrt{\Phi_0/C}^{2\alpha-2}\ln^{5+\alpha}\left(\frac{6(t+1)^2}{\delta}\right)}$$
$$+\sum_{t=1}^{T-1}\frac{2\cdot 4^\alpha \cdot 256^{2\alpha-2} C^{1+\alpha}p^{3\alpha}(\Phi_0,L,\sigma)\sigma^{2\alpha}}{(256)^{3\alpha}\cdot 64(t+1)^{2\beta\alpha}\sqrt{\Phi_0/C}^{2\alpha-2}\ln^{5+\alpha}\left(\frac{6(t+1)^2}{\delta}\right)}\times\left(\frac{(t+1)^{1-\beta}-1}{1-\beta}\right)^\alpha$$
$$+\sum_{t=1}^{T-1}\frac{4^\alpha \cdot 256^{2\alpha-2} C^{\alpha}p^{3\alpha-1}(\Phi_0,L,\sigma)\sigma^{2\alpha}}{(256)^{3\alpha-1}\cdot 4^{(3\alpha-1)/\alpha}(t+1)^{2\beta\alpha}\sqrt{\Phi_0/C}^{2\alpha-2}\ln^{5+\alpha}\left(\frac{6(t+1)^2}{\delta}\right)}$$
$$+\sum_{t=1}^{T-1}\frac{4^\alpha \cdot 256^{2\alpha-2} C^{\alpha}p^{3\alpha-1}(\Phi_0,L,\sigma)\sigma^{2\alpha}}{(256)^{3\alpha-1}\cdot 4^{(3\alpha-1)/\alpha}(t+1)^{2\beta\alpha}\sqrt{\Phi_0/C}^{2\alpha-2}\ln^{5+\alpha}\left(\frac{6(t+1)^2}{\delta}\right)}\times\left(\frac{(t+1)^{1-\beta}-1}{1-\beta}\right)^{\alpha-1}.$$

Applying Lemma 3, we get

$$\sum_{t=0}^{T-1} 2e_t\left\|\theta_t^b\right\|^2 \leq \frac{2\cdot 4^\alpha \cdot 256^{2\alpha-2} C(d_1+1)p^{2\alpha}(\Phi_0,L,\sigma)\sigma^{2\alpha}}{(256)^{2\alpha}\cdot 16\sqrt{\Phi_0/C}^{2\alpha-2}\ln^{5+\alpha}\left(\frac{6}{\delta}\right)}$$
$$+\sum_{t=1}^{T-1}\frac{2\cdot 4^\alpha \cdot 256^{2\alpha-2} C^{1+\alpha}p^{3\alpha}(\Phi_0,L,\sigma)\sigma^{2\alpha}}{(256)^{3\alpha}\cdot 64(t+1)^{2\beta\alpha}\sqrt{\Phi_0/C}^{2\alpha-2}\ln^{5+\alpha}\left(\frac{6(t+1)^2}{\delta}\right)}$$
$$+\sum_{t=1}^{T-1}\frac{2\cdot 4^\alpha \cdot 256^{2\alpha-2} C^{1+\alpha}p^{3\alpha}(\Phi_0,L,\sigma)\sigma^{2\alpha}}{(256)^{3\alpha}\cdot 64(t+1)^{2\beta\alpha-(1-\beta)\alpha}\sqrt{\Phi_0/C}^{2\alpha-2}\ln^{5}\left(\frac{6(t+1)^2}{\delta}\right)}$$
$$+\sum_{t=1}^{T-1}\frac{4^\alpha \cdot 256^{2\alpha-2} C^{\alpha}p^{3\alpha-1}(\Phi_0,L,\sigma)\sigma^{2\alpha}}{(256)^{3\alpha-1}\cdot 4^{(3\alpha-1)/\alpha}(t+1)^{2\beta\alpha}\sqrt{\Phi_0/C}^{2\alpha-2}\ln^{5+\alpha}\left(\frac{6(t+1)^2}{\delta}\right)}$$
$$+\sum_{t=1}^{T-1}\frac{4^\alpha \cdot 256^{2\alpha-2} C^{\alpha}p^{3\alpha-1}(\Phi_0,L,\sigma)\sigma^{2\alpha}}{(256)^{3\alpha-1}\cdot 4^{(3\alpha-1)/\alpha}(t+1)^{2\beta\alpha-(1-\beta)(\alpha-1)}\sqrt{\Phi_0/C}^{2\alpha-2}\ln^{6}\left(\frac{6(t+1)^2}{\delta}\right)}.$$

Then, substituting (7) for each term (to be precise, for the first term we use option 1, for second and third terms – option 3, and for the last two terms – option 2), with $d_1+1=\gamma_0 C+1\leq 2$, one can

obtain

$$\begin{aligned}\sum_{t=0}^{T-1} 2e_t\left\|\theta_t^b\right\|^2 \leq & \frac{2\cdot 4^\alpha\cdot 256^{2\alpha-2}\Phi_0}{(256)^{2\alpha}\cdot 16\ln^{5+\alpha}\left(\frac{6}{\delta}\right)} \\ & + \sum_{t=1}^{T-1}\frac{2\cdot 4^\alpha\cdot 256^{2\alpha-2}\Phi_0}{(256)^{3\alpha}\cdot 64(t+1)^{2\beta\alpha}\ln^{5+\alpha}\left(\frac{6(t+1)^2}{\delta}\right)} \\ & + \sum_{t=1}^{T-1}\frac{2\cdot 4^\alpha\cdot 256^{2\alpha-2}\Phi_0}{(256)^{3\alpha}\cdot 64(t+1)^{2\beta\alpha-(1-\beta)\alpha}\ln^{5}\left(\frac{6(t+1)^2}{\delta}\right)} \\ & + \sum_{t=1}^{T-1}\frac{4^\alpha\cdot 256^{2\alpha-2}\Phi_0}{(256)^{3\alpha-1}\cdot 4^{(3\alpha-1)/\alpha}(t+1)^{2\beta\alpha}\ln^{5+\alpha}\left(\frac{6(t+1)^2}{\delta}\right)} \\ & + \sum_{t=1}^{T-1}\frac{4^\alpha\cdot 256^{2\alpha-2}\Phi_0}{(256)^{3\alpha-1}\cdot 4^{(3\alpha-1)/\alpha}(t+1)^{2\beta\alpha-(1-\beta)(\alpha-1)}\ln^{6}\left(\frac{6(t+1)^2}{\delta}\right)}.\end{aligned}$$

Hence, to obtain the sufficient bound, it is enough to choose β as follows:

Condition 7:

$$\begin{cases}2\beta\alpha\geq 1\Rightarrow\beta\geq\frac{1}{2\alpha};\\ 2\beta\alpha-(1-\beta)(\alpha-1)\geq 1\Rightarrow\beta\geq\frac{\alpha}{3\alpha-1};\\ 2\beta\alpha-(1-\beta)\alpha\geq 1\Rightarrow\beta\geq\frac{1+\alpha}{3\alpha};\end{cases}\tag{22}$$

Consequently, we have

$$\begin{aligned}\sum_{t=0}^{T-1} 2e_t\left\|\theta_t^b\right\|^2 &\overset{(22)}{\leq} \frac{2\cdot 4^\alpha\cdot 256^{2\alpha-2}\Phi_0}{(256)^{2\alpha}\cdot 16}+\sum_{t=1}^{T-1}\frac{4\cdot 4^\alpha\cdot 256^{2\alpha-2}\Phi_0}{(256)^{3\alpha}\cdot 64(t+1)}+\sum_{t=1}^{T-1}\frac{2\cdot 4^\alpha\cdot 256^{2\alpha-2}\Phi_0}{(256)^{3\alpha-1}\cdot 4^{(3\alpha-1)/\alpha}(t+1)} \\ &\leq \sum_{t=0}^{T-1}\frac{4\cdot 4^\alpha\cdot 256^{2\alpha-2}\Phi_0}{(256)^{2\alpha}\cdot 16(t+1)} \overset{(16)}{\leq} \frac{\Phi_0\left(\frac{6T^2}{\delta}\right)}{16}.\end{aligned}$$

Final bound. If we formulate events $E_{①}$, $E_{③}$, $E_{⑤}$ as follows: $b = {}^{\Phi_0\ln\left(\frac{6(T)^2}{\delta}\right)}/_{2}$ and $G = \frac{\Phi_0^2\ln\left(\frac{6(T)^2}{\delta}\right)}{24}$

$$\begin{aligned}E_{①} &= \left\{|①|\leq\frac{\Phi_0\ln\left(\frac{6(T)^2}{\delta}\right)}{2}\text{ either }\sum_{t=0}^{T-1}\sigma_t^2>\frac{\Phi_0^2\ln\left(\frac{6(T)^2}{\delta}\right)}{24}\right\}\\ E_{③} &= \left\{|③|\leq\frac{\Phi_0\ln\left(\frac{6(T)^2}{\delta}\right)}{12}\text{ either }\sum_{t=0}^{T-1}\sigma_t^2>\frac{\Phi_0^2\ln\left(\frac{6(T)^2}{\delta}\right)}{1152}\right\}\\ E_{⑤} &= \left\{|⑤|\leq\frac{\Phi_0\ln\left(\frac{6(T)^2}{\delta}\right)}{8}\text{ either }\sum_{t=0}^{T-1}\sigma_t^2>\frac{\Phi_0^2\ln\left(\frac{6(T)^2}{\delta}\right)}{384}\right\},\end{aligned}$$

then the event $E_{T-1}\cap E_{①}\cap E_{③}\cap E_{⑤}$ implies

$$\Phi_T\leq\Phi_0+\frac{\Phi_0\ln\left(\frac{6T^2}{\delta}\right)}{2}+3\cdot\frac{\Phi_0\ln\left(\frac{6T^2}{\delta}\right)}{12}+\frac{\Phi_0\ln\left(\frac{6T^2}{\delta}\right)}{8}+2\cdot\frac{\Phi_0\ln\left(\frac{6T^2}{\delta}\right)}{16}\leq 2\Phi_0\ln\left(\frac{6T^2}{\delta}\right).$$

Consequently, we have

$$\begin{aligned}\mathbb{P}\{E_T\} &\geq \mathbb{P}\{E_{T-1} \cap E_{①} \cap E_{③} \cap E_{⑤}\} = 1 - \mathbb{P}\{\overline{E}_{T-1} \cap \overline{E}_{①} \cap \overline{E}_{③} \cap \overline{E}_{⑤}\} \\ &\geq 1 - \mathbb{P}\{\overline{E}_{T-1}\} - \mathbb{P}\{\overline{E}_{①}\} - \mathbb{P}\{\overline{E}_{③}\} - \mathbb{P}\{\overline{E}_{⑤}\} \geq 1 - \delta \sum_{t=1}^{T-1} \frac{1}{t^2} - \frac{3\delta}{3T^2} \\ &= 1 - \delta \sum_{t=1}^{T} \frac{1}{t^2}.\end{aligned}$$

This finishes the inductive step of the proof. Therefore, the event E_K implies

$$\Phi_K \leq 2\Phi_0 \ln\left(\frac{6(K+1)^2}{\delta}\right)$$

with probability at least $1 - \delta \sum\limits_{t=1}^{K} \frac{1}{t^2}$. Abusing the notation, with $d_0 = 0$ we have

$$d_K(f(x_K) - f^*) \leq 2CR_0^2 \ln\left(\frac{6(K+1)^2}{\delta}\right).$$

Thus, the event E_K implies

$$f(x_K) - f^* \leq \frac{2CR_0^2 \ln\left(\frac{6(K+1)^2}{\delta}\right)}{\sum\limits_{t=0}^{K-1} \gamma_t C} = \frac{2R_0^2 \ln\left(\frac{6(K+1)^2}{\delta}\right)}{\sum\limits_{t=0}^{K-1} \gamma_t}. \tag{23}$$

Moreover, it is known that

$$\frac{1}{\sum\limits_{i=1}^{n} \frac{1}{a_i}} \leq \frac{\sum\limits_{i=1}^{n} a_i}{n^2}$$

for $a_i > 0$, since it is AM-HM inequality. Thus, applying it with $a_i = \frac{1}{\gamma_i}$ to (23) leads to

$$f(x_K) - f^* \leq \frac{2R_0^2 \ln\left(\frac{6(K+1)^2}{\delta}\right)}{\sum\limits_{t=0}^{K-1} \gamma_t} \leq \frac{2R_0^2 \ln\left(\frac{6(K+1)^2}{\delta}\right)}{K^2} \sum_{t=0}^{K-1} \frac{1}{\gamma_t}.$$

Substituting (4), we have

$$\begin{aligned}f(x_K) - f^* &\leq \frac{2R_0^2 \ln\left(\frac{6(K+1)^2}{\delta}\right)}{K^2} \sum_{k=0}^{K-1} \frac{1}{\gamma_t} \\ &\leq \frac{2R_0^2 \ln\left(\frac{6(K+1)^2}{\delta}\right)}{K^2} \sum_{k=0}^{K-1} \max\left\{1024L \ln^3\left(\frac{6(k+1)^2}{\delta}\right), \frac{(k+1)^\beta \ln^3\left(\frac{6(k+1)^2}{\delta}\right)}{256 \cdot 4^{1/\alpha} p(\Phi, L, \sigma)}\right\}.\end{aligned} \tag{24}$$

Before we provide a final bound, let us emphasize two important parts. First of all, combining **Conditions 1-7**, we should choose $\beta \geq \max\left\{\frac{1}{\alpha}, \frac{\alpha+1}{3\alpha-1}, \frac{2+\alpha}{3\alpha}, \frac{\alpha}{3\alpha-1}, \frac{\alpha+1}{3\alpha}\right\}$. It is easy to verify that β should be greater than $\frac{2+\alpha}{3\alpha}$.

What is more, let us look at the choice of $p(\Phi_0, L, \sigma)$. Applying the definition of C (8), we have

$$
\begin{aligned}
p(\Phi_0, L, \sigma) &= \min\left\{\frac{\sqrt{\Phi_0/C}}{\sigma}, \frac{\sqrt{\Phi_0/C}^{\frac{2\alpha}{3\alpha-1}}}{C^{\frac{\alpha-1}{3\alpha-1}}\sigma^{\frac{2\alpha}{3\alpha-1}}}, \frac{\sqrt{\Phi_0/C}^{\frac{2}{3}}}{C^{\frac{1}{3}}\sigma^{\frac{2}{3}}}\right\} \\
&= \min\left\{\frac{R_0}{\sqrt{2}\sigma}, \frac{R_0^{\frac{2\alpha}{3\alpha-1}}}{2^{\frac{\alpha}{3\alpha-1}}C^{\frac{\alpha-1}{3\alpha-1}}\sigma^{\frac{2\alpha}{3\alpha-1}}}, \frac{R_0^{\frac{2}{3}}}{2^{\frac{1}{3}}C^{\frac{1}{3}}\sigma^{\frac{2}{3}}}\right\} \\
&= \min\left\{\frac{R_0}{\sqrt{2}\sigma}, \frac{R_0^{\frac{2\alpha}{3\alpha-1}}}{2^{\frac{\alpha}{3\alpha-1}}L^{\frac{\alpha-1}{3\alpha-1}}\sigma^{\frac{2\alpha}{3\alpha-1}}}, \frac{R_0^{\frac{2}{3}}}{2^{\frac{1}{3}}L^{\frac{1}{3}}\sigma^{\frac{2}{3}}},\right. \\
&\left.\frac{R_0^{\frac{2\alpha}{3\alpha-1}}(4^{1-1/\alpha}p(\Phi_0, L, \sigma))^{\frac{\alpha-1}{3\alpha-1}}}{2^{\frac{\alpha}{3\alpha-1}}\sigma^{\frac{2\alpha}{3\alpha-1}}}, \frac{R_0^{\frac{2}{3}}(4^{1-1/\alpha}p(\Phi_0, L, \sigma))^{\frac{1}{3}}}{2^{\frac{1}{3}}\sigma^{\frac{2}{3}}}\right\}. \qquad (25)
\end{aligned}
$$

If $p(\Phi_0, L, \sigma)$ equals fourth of fifth choice from (25), we can show that

$$
p(\Phi_0, L, \sigma) = \frac{const \cdot R_0^r p^{1-r}(\Phi_0, L, \sigma)}{\sigma^r},
$$

where r can be equal to $\frac{2\alpha}{3\alpha-1}$ and $\frac{2}{3}$. Therefore, we get

$$
p^r(\Phi_0, L, \sigma) = \frac{const \cdot R_0^r}{\sigma^r} \Rightarrow p(\Phi_0, L, \sigma) = \frac{\sqrt[r]{const} \cdot R_0}{\sigma}.
$$

As a result, we have

$$
p(\Phi_0, L, \sigma) = \Theta\left(\min\left\{\frac{R_0}{\sigma}, \frac{R_0^{\frac{2\alpha}{3\alpha-1}}}{L^{\frac{\alpha-1}{3\alpha-1}}\sigma^{\frac{2\alpha}{3\alpha-1}}}, \frac{R_0^{\frac{2}{3}}}{L^{\frac{1}{3}}\sigma^{\frac{2}{3}}}\right\}\right). \qquad (26)
$$

Finally, applying (26) to (24) and using that

$$
\ln\left(\frac{6(k+1)^2}{\delta}\right) \leq \ln\left(\frac{6(K+1)^2}{\delta}\right)
$$

and

$$
\sum_{k=0}^{K-1}(k+1)^\beta = \sum_{k=1}^{K}k^\beta \leq K^\beta + \int_1^K x^\beta dx = K^\beta + \frac{K^{1+\beta}-1}{1+\beta} \leq 2K^{1+\beta},
$$

we derive

$$
\begin{aligned}
f(x_K) - f^* &\leq \frac{2R_0^2\ln\left(\frac{6(K+1)^2}{\delta}\right)}{K^2}\sum_{k=0}^{K-1}\max\left\{1024L\ln^3\left(\frac{6(k+1)^2}{\delta}\right), \frac{(k+1)^\beta\ln^3\left(\frac{6(k+1)^2}{\delta}\right)}{256\cdot 4^{1/\alpha}p(\Phi, L, \sigma)}\right\} \\
&= \mathcal{O}\left(\frac{LR_0^2\ln^4\left(\frac{6(K+1)^2}{\delta}\right)}{K} + \frac{\max\left\{R_0\sigma, L^{\frac{\alpha-1}{3\alpha-1}}R_0^{\frac{4\alpha-2}{3\alpha-1}}\sigma^{\frac{2\alpha}{3\alpha-1}}, L^{\frac{1}{3}}R_0^{\frac{4}{3}}\sigma^{\frac{2}{3}}\right\}\ln^4\left(\frac{6(K+1)^2}{\delta}\right)}{K^{1-\beta}}\right),
\end{aligned}
$$

Noting that we choose β as the best possible one, i.e. $\beta = \frac{2+\alpha}{3\alpha}$, we obtain

$$
f(x_K) - f^* = \tilde{\mathcal{O}}\left(\frac{LR_0^2}{K} + \frac{\max\left\{R_0\sigma, L^{\frac{\alpha-1}{3\alpha-1}}R_0^{\frac{4\alpha-2}{3\alpha-1}}\sigma^{\frac{2\alpha}{3\alpha-1}}, L^{\frac{1}{3}}R_0^{\frac{4}{3}}\sigma^{\frac{2}{3}}\right\}}{K^{\frac{2\alpha-2}{3\alpha}}}\right),
$$

where $\tilde{\mathcal{O}}(\cdot)$ denotes polylogarithmic dependency. This concludes the proof. □