# HIGH-PROBABILITY CONVERGENCE OF CLIPPED SGD UNDER HEAVY-TAILED NOISE AND $(L_0, L_1)$-SMOOTHNESS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Gradient clipping is widely used in language-model training to control heavy-tailed gradient noise and can improve convergence guarantees over stochastic gradient descent (SGD) under $(L_0, L_1)$-smoothness. Under these joint conditions, a central challenge is to obtain high-probability guarantees without exponential dependence on $L_1 R_0$, where $R_0$ bounds the initial distance to a minimizer. We resolve this challenge for convex objectives, establishing, to the best of our knowledge, the first such guarantees for standard Clip-SGD. We assume unbiased stochastic gradients with bounded central $\alpha$-th moment, $\alpha \in (1, 2]$. Our bounds have only polylogarithmic dependence on the inverse failure probability and recover known deterministic generalized-smoothness rates when the noise vanishes, as well as the standard high-probability rate under heavy-tailed noise in the classical $L$-smooth setting. The convex rate is attained by a computable output that averages iterates whose sampled stochastic gradients are not clipped, requiring neither function values nor extra oracle calls. We establish a lower bound for Clip-SGD with any fixed stepsize and clipping level, showing that, under the stated iteration conditions, our convex stochastic rate is optimal up to logarithmic factors in the iteration budget. Our upper-bound analysis uses a directional clipping-bias bound to absorb part of the bias into the progress generated by the clipped population gradient, avoiding an exponentially large local smoothness constant. We also obtain nonconvex high-probability guarantees that recover the noiseless rate, match the classical-smooth stochastic iteration rate when $L_1 = 0$, and have no explicit $L_1$ dependence in the asymptotically dominant stochastic term.

## 1 INTRODUCTION

Stochastic Gradient Descent (SGD) (Robbins & Monro, 1951) and its variants are central to large-scale machine learning, where computing full gradients can be prohibitively expensive (Shalev-Shwartz & Ben-David, 2014; Goodfellow et al., 2016). Gradient clipping stabilizes these methods by limiting the magnitude of stochastic gradients. It is used to control exploding gradients in recurrent neural networks (Pascanu et al., 2013) and is a standard component in training models such as BERT (Devlin et al., 2019), GPT-3 (Brown et al., 2020), Switch Transformers (Fedus et al., 2022), and LLaMA (Touvron et al., 2023). Two properties help explain its effectiveness: robustness to heavy-tailed gradient noise and adaptation to gradient-dependent curvature.

Heavy-tailed stochastic gradient noise has been observed in tasks such as BERT pre-training (Zhang et al., 2020c). Such noise need not have finite variance, placing it beyond the scope of classical bounded-variance analyses. Clipping limits the influence of extreme gradient samples and enables *high-probability* convergence guarantees under weaker moment assumptions. In particular, recent analyses obtain bounds with polylogarithmic dependence on the inverse failure probability $1/\delta$ when the noise has a bounded central $\alpha$-th moment, with $\alpha \in (1, 2]$ (Cutkosky & Mehta, 2021; Sadiev et al., 2023; Nguyen et al., 2023; Chezhegov et al., 2024). These guarantees balance clipping bias against stochastic fluctuations, often using a clipping threshold that grows with the iteration budget.

Clipping also helps when curvature depends on the gradient norm. Zhang et al. (2020b) observed this dependence along training trajectories, motivating the $(L_0, L_1)$-smoothness assumption. This



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

condition extends classical $L$-smoothness by allowing local curvature to increase with the gradient norm. Under this assumption, clipping can improve convergence guarantees over unclipped gradient methods (Zhang et al., 2020b;a; Koloskova et al., 2023; Gorbunov et al., 2025; Vankov et al., 2025). Several such analyses calibrate the clipping threshold to the smoothness parameters $L_0$ and $L_1$.

For convex $(L_0, L_1)$-smooth objectives with heavy-tailed gradient noise, we seek a high-probability guarantee for standard Clip-SGD, expressed in terms of an upper bound $R_0$ on the initial distance to a minimizer. A direct reduction to classical smoothness can introduce exponential dependence on $L_1 R_0$, since the effective smoothness constant over a localization region can grow exponentially with its radius. Our main question is therefore:

*Can standard Clip-SGD achieve high-probability convergence for convex $(L_0, L_1)$-smooth objectives under heavy-tailed noise without exponential dependence on $L_1 R_0$?*

We answer this question affirmatively. We first introduce the problem setup and assumptions, then present our main results.

### 1.1 Problem setup and assumptions

**Notation.** We denote the Euclidean norm on $\mathbb{R}^d$ by $\|\cdot\|$. For $\lambda > 0$, $\texttt{clip}(x, \lambda)$ denotes the Euclidean projection of $x$ onto the closed ball of radius $\lambda$ centered at the origin. When $f$ attains its minimum, we use $R_0$ to denote an upper bound on the distance from $x_0$ to some minimizer $x^\star$. We let $\mathcal{F}_k := \sigma(\xi_0, \dots, \xi_{k-1})$ denote the history generated by Clip-SGD before the $k$-th stochastic-gradient query. We also let $\nu \in (0, 1)$ denote the unique positive solution of $\nu e^\nu = 1$.

**Problem.** We focus on the classical stochastic optimization problem, which can be stated as

$$\min_{x \in \mathbb{R}^d} \{f(x) := \mathbb{E}_{\xi \sim \mathcal{D}}[f(x, \xi)]\}. \tag{1}$$

This formulation is foundational in machine learning (Shalev-Shwartz & Ben-David, 2014), where $f$ is the population loss, $x$ the model parameters, and $\xi \sim \mathcal{D}$ a data sample.

We write $f^\star := \inf_{x \in \mathbb{R}^d} f(x)$ and assume $f^\star > -\infty$. For our convex guarantees, we additionally assume that this infimum is attained. We consider Clip-SGD (Algorithm 1) applied to this problem.

**Algorithm 1** Clip-SGD

1: **Input:** Starting point $x_0$, level of clipping $\lambda$, learning rate $\gamma$
2: **for** $k = 0, \dots, K-1$ **do**
3: Sample $\nabla f(x_k, \xi_k)$
4: $x_{k+1} = x_k - \gamma \texttt{clip}(\nabla f(x_k, \xi_k), \lambda)$
5: **end for**

**Assumptions.** Our convex guarantees use the following three assumptions.

**Assumption 1** (Convexity). *$f$ is convex: for all $x, y \in \mathbb{R}^d$, $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle$.*

**Assumption 2** ($(L_0, L_1)$-smoothness). *There exist $L_0, L_1 \geq 0$ such that, for all $x, y \in \mathbb{R}^d$,*

$$\|\nabla f(x) - \nabla f(y)\| \leq \left( L_0 + L_1 \sup_{u \in [x,y]} \|\nabla f(u)\| \right) \|x - y\|.$$

This assumption is strictly more general than classical $L$-smoothness, i.e., the assumption that $\nabla f$ is globally $L$-Lipschitz, which is recovered by setting $(L_0, L_1) = (L, 0)$. It also covers functions that are not globally $L$-smooth, such as $f(x) = \exp(\|x\|) + \exp(-\|x\|)$ and $f(x) = \|x\|^n$ for $n > 2$ (Zhang et al., 2020b;a; Chen et al., 2023).

**Assumption 3** (Stochastic oracle). *The stochastic oracle is unbiased and has bounded central $\alpha$-th moment: $\exists\, \alpha \in (1, 2]$, $\sigma \geq 0$ such that, for all $x \in \mathbb{R}^d$,*

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

For $\alpha = 2$, this recovers the bounded-variance setting (Ghadimi & Lan, 2013), while $\alpha \in (1, 2)$ allows noise with infinite variance.

### 1.2 Our contributions

We establish, to the best of our knowledge, the first high-probability convergence guarantees for standard Clip-SGD on convex objectives under the joint assumptions of heavy-tailed noise and $(L_0, L_1)$-smoothness. Under Assumptions 1,2 and 3, with probability at least $1-\delta$, we obtain

$$f(\widehat{x}_K) - f^\star \leq \mathcal{O}\left(\max\left\{\frac{L_0 R_0^2}{K},\ \sigma R_0 \left(\frac{\max\{L_1 R_0, \log(12/\delta)\}}{K}\right)^{\frac{\alpha-1}{\alpha}}\right\}\right),$$

provided that, for $L_0 > 0$ and $R_0 > 0$,

$$K = \Omega\left(\log^2\left(\frac{12}{\delta}\right) + L_1 R_0 \log\left[\frac{12e}{\delta}\left(1 + \frac{K\Delta_0}{L_0 R_0^2}\right)\right]\right),$$

where $\Delta_0 := f(x_0) - f^\star$. An explicit sufficient iteration requirement is

$$K = \Omega\left(\log^2\left(\frac{12}{\delta}\right) + L_1 R_0 \log\left[\frac{12e}{\delta}\left(1 + \frac{L_1\Delta_0}{L_0 R_0}\right)\right]\right).$$

Here, $\widehat{x}_K$ averages the iterates whose sampled stochastic gradients are not clipped, requiring neither function values nor additional oracle calls. At fixed confidence, this explicit iteration requirement can be substantially smaller than a quadratic requirement in $L_1 R_0$.

Our convex guarantees recover deterministic rates under generalized smoothness when the noise vanishes (Gorbunov et al., 2025; Vankov et al., 2025), and the standard high-probability rate under heavy-tailed noise in the classical $L_0$-smooth setting ($L_1 = 0$) (Sadiev et al., 2023). The key ingredient is the directional clipping-bias bound in Lemma 9, which separates the bias into a heavy-tail remainder and a component that can be absorbed into the progress generated by the clipped population gradient. Together with structural inequalities for this gradient, this avoids introducing an exponentially large local smoothness constant. A comparison with prior convex guarantees appears in Table 1.

We complement the convex upper bound with a lower bound for Clip-SGD with arbitrary fixed stepsize and clipping level; see Theorem 3. For every $\delta \in (0, 1/4)$, under the iteration conditions of that theorem, and for every such parameter choice and every possibly randomized output rule based on the observed transcript, with output denoted by $x_K^{\mathrm{out}}$, we construct a one-dimensional convex $(1, L_1)$-smooth instance with a minimizer within distance $R_0$ of the initial point and an unbiased stochastic-gradient oracle satisfying the same central $\alpha$-th moment bound, such that

$$\mathbb{P}\left(f(x_K^{\mathrm{out}}) - f^\star \geq \frac{\sigma R_0}{128}\left[\frac{\max\{L_1 R_0/\log(eK), \log(12/\delta)\}}{K}\right]^{\frac{\alpha-1}{\alpha}}\right) > \delta.$$

Thus, our convex stochastic rate, including its confidence dependence, is optimal up to logarithmic factors in the iteration budget for fixed-parameter Clip-SGD.

We also establish a unified high-probability guarantee for the average gradient norm on nonconvex objectives; see Appendix D. It recovers the deterministic generalized-smoothness rate when $\sigma = 0$ and the classical-smooth heavy-tailed stochastic iteration rate when $L_1 = 0$. The noiseless recovery improves on the stated bounds of Nguyen et al. (2023) and An et al. (2025) under their prescribed tuning. For fixed problem parameters with $L_0\Delta_0 > 0$ and $\sigma > 0$, and fixed confidence, our asymptotically dominant stochastic term has the same polynomial iteration rate as the corresponding leading term of An et al. (2025), while containing no explicit $L_1$ dependence and no logarithmic factors in $K$ or $1/\delta$. In our bound, explicit $L_1$ dependence and logarithmic factors appear only in terms that decay faster with $K$.

**Comparison with closest prior work.** For classical $L$-smooth objectives, Sadiev et al. (2023) obtain the standard high-probability heavy-tailed rate through a localization argument. When such a guarantee is transferred to an $(L_0, L_1)$-smooth objective via local classical smoothness, the effective smoothness constant on a ball of radius $r$ can scale as $L_0 \exp(L_1 r)$. Since the relevant localization radius is of order $R_0$, this reduction can introduce exponential dependence on $L_1 R_0$.

For deterministic $(L_0, L_1)$-smooth optimization, Gorbunov et al. (2025) obtain the rate displayed in Table 1 under a quadratic sufficient transient in $L_1 R_0$. The refined analyses of Vankov et al. (2025)

Table 1: Convergence guarantees for convex objectives. Here $R_0 \geq \|x_0 - x^\star\|$ is an upper bound on the initial distance to a minimizer, and $M_\delta = \max\{L_1 R_0, \log(12/\delta)\}$. Stochastic bounds hold with high probability. The symbols $\widetilde{\mathcal{O}}$ and $\widetilde{\Theta}$ hide logarithmic factors in $K$ and $1/\delta$. Method details and sufficient iteration budgets are provided in the comparison with closest prior work.

| Reference | Stochasticity | Clipping scale | Rate |
|---|---|---|---|
| | | $L$-**smooth case** | |
| Sadiev et al. (2023) | HT | $\widetilde{\Theta}\left(\max\left\{LR_0,\ \sigma K^{\frac{1}{\alpha}}\right\}\right)$ | $\widetilde{\mathcal{O}}\left(\max\left\{\frac{LR_0^2}{K},\ \frac{R_0\sigma}{K^{\frac{\alpha-1}{\alpha}}}\right\}\right)$ |
| | | $(L_0, L_1)$-**smooth case** | |
| Gorbunov et al. (2025)<br>Vankov et al. (2025)<br>Lobanov et al. (2024) | Det. | $\Theta\left(\frac{L_0}{L_1}\right)$ | $\mathcal{O}\left(\frac{L_0 R_0^2}{K}\right)$ |
| Gaash et al. (2025) | LT | $\widetilde{\Theta}\left(\max\left\{\frac{L_0}{L_1},\ \frac{\sigma\sqrt{K}}{L_1 R_0}\right\}\right)$ | $\widetilde{\mathcal{O}}\left(\max\left\{\frac{L_0 R_0^2}{K},\ \frac{R_0\sigma}{\sqrt{K}}\right\}\right)$ |
| **This work** | HT | $\Theta\left(\max\left\{\frac{L_0 R_0}{M_\delta},\ \sigma\left(\frac{K}{M_\delta}\right)^{\frac{1}{\alpha}}\right\}\right)$ | $\mathcal{O}\left(\max\left\{\frac{L_0 R_0^2}{K},\ R_0\sigma\left(\frac{M_\delta}{K}\right)^{\frac{\alpha-1}{\alpha}}\right\}\right)$ |

*Stochasticity:* Det. = deterministic; LT = light-tailed noise; HT = heavy-tailed noise.

and Lobanov et al. (2024) obtain the same rate under a sufficient iteration condition that depends logarithmically on the initial objective gap and can be substantially weaker than the quadratic requirement on favorable instances. In the noiseless case, our guarantee recovers this same rate under the corresponding refined sufficient iteration condition, while our stochastic result extends this refined transient picture to heavy-tailed gradients. The accelerated method of Vankov et al. (2025) is not included in Table 1, as it relies on a line search using function values.

For comparison with the closest stochastic result, we consider the sub-Gaussian condition

$$\mathbb{E}\left[\exp\left(\|\nabla f(x,\xi) - \nabla f(x)\|^2/\sigma^2\right)\right] \leq \exp(1), \qquad \forall x \in \mathbb{R}^d, \tag{2}$$

for some $\sigma > 0$. Together with unbiasedness, this implies Assumption 3 for every $\alpha \in (1, 2]$.

The closest stochastic result is Gaash et al. (2025), who obtain high-probability guarantees for convex $(L_0, L_1)$-smooth objectives under the light-tailed noise condition in Eq. (2). The rate displayed in Table 1 holds under a sufficient iteration budget quadratic in $L_1 R_0$. Their method uses a double-sampling scheme in which one stochastic gradient determines the clipping test and an independent stochastic gradient determines the update. In contrast, we analyze standard Clip-SGD, where the same stochastic gradient is clipped and used for the update, under the weaker bounded central $\alpha$-th moment condition in Assumption 3, with $\alpha \in (1, 2]$. Our sufficient iteration condition retains the refined dependence on the initial objective gap discussed above rather than imposing a quadratic transient uniformly. Further details on all comparisons above are provided in Appendix C.3.

## 2 Related Work

**Convergence under $(L_0, L_1)$-smoothness - convex case.** Early studies of first-order methods under $(L_0, L_1)$-smoothness focused mainly on the non-convex setting (Zhang et al., 2020b;a; Zhao et al., 2021; Chen et al., 2023; Hübler et al., 2024; Khirirat et al., 2024; Crawshaw et al., 2022; Faw et al., 2023; Wang et al., 2022; 2023; Li et al., 2023b; Bilel, 2024; Liu & Zhou, 2025; Vankov et al., 2025).

In the convex setting, Koloskova et al. (2023) provided convergence guarantees for Clip-GD under convexity, $(L_0, L_1)$-smoothness and $L$-smoothness, showing that clipping can outperform standard GD for suitable parameter choices. Building on this, Takezawa et al. (2024) analyzed GD with Polyak stepsizes, while Li et al. (2023a) considered GD and Nesterov's accelerated gradient (Nesterov, 1983) under the broader class of $(r, \ell)$-smooth functions. However, the additional smoothness quantities in Koloskova et al. (2023); Takezawa et al. (2024); Li et al. (2023a) can lead to exponentially large factors in $L_1 R_0$. More recent results of Gorbunov et al. (2025); Vankov et al. (2025); Lobanov et al. (2024) avoid this drawback and provide tighter bounds for deterministic convex

$(L_0, L_1)$-smooth problems. Earlier complexity bounds are detailed in Appendix A, while recent rates and sufficient iteration conditions are compared in Table 1 and Appendix C.3. In addition, Tyurin (2024) gives a unified analysis of GD for convex and non-convex problems under the more general $\ell(\|\nabla f(x)\|)$-smoothness condition, while Yu et al. (2025b) studies Mirror Descent and its variants under a non-Euclidean version of $(r, \ell)$-smoothness (Li et al., 2023a).

Many of these works also give stochastic convergence results (Zhang et al., 2020b;c; Zhao et al., 2021; Chen et al., 2023; Crawshaw et al., 2022; Faw et al., 2023; Wang et al., 2022; 2023; Li et al., 2023b; Hübler et al., 2024; Gorbunov et al., 2025; Yu et al., 2025b). Yang et al. (2024) analyze a variant of Normalized SGD with independent normalization, while Yu et al. (2025a) study accelerated SGD with constant and adaptive stepsizes under $(L_0, L_1)$-smoothness and relaxed affine variance assumptions. Tovmasyan et al. (2025) introduce $\psi$-smoothness and derive convergence bounds for the Stochastic Proximal Point Method (Bertsekas, 2011). The bounds of Faw et al. (2023); Wang et al. (2023) have inverse-power dependence on $\delta$, whereas Li et al. (2023b); Yu et al. (2025a) assume sub-Gaussian noise related to equation 2. Further details on light-tailed high-probability guarantees are given in Appendix A.

For convex objectives, the closest stochastic result is Gaash et al. (2025), who derive high-probability rates with polylogarithmic dependence on $1/\delta$ under sub-Gaussian noise. Their method uses two independent stochastic gradients, one for the update direction and another for the clipping multiplier. While this avoids exponentially large factors in $L_1 R_0$, its behavior under heavy-tailed noise remains unclear.

**High-probability convergence under heavy-tailed noise.** Gradient clipping is a standard approach for obtaining high-probability guarantees under heavy-tailed noise. Early work includes the truncated Stochastic Mirror Descent method of Nazin et al. (2019), which established high-probability bounds for convex and strongly convex problems under bounded variance (Assumption 3 with $\alpha = 2$). Gorbunov et al. (2020) later gave the first comprehensive high-probability bounds for Clip-SGD (Algorithm 1) and introduced an accelerated variant based on the Stochastic Similar Triangles Method (SSTM) (Gasnikov & Nesterov, 2016). Subsequent work extended these results to Assumption 3 with $\alpha < 2$ (Cutkosky & Mehta, 2021; Sadiev et al., 2023; Nguyen et al., 2023; Gorbunov et al., 2024b; Chezhegov et al., 2026). Extensions to non-smooth optimization, variational inequalities, adaptive methods, zeroth-order methods, distributed optimization, and differential privacy are discussed in Appendix A.

Despite these advances, it remained open whether standard Clip-SGD admits high-probability guarantees for convex $(L_0, L_1)$-smooth objectives under heavy-tailed noise, with polylogarithmic dependence on $1/\delta$ and without exponential dependence on $L_1 R_0$.

## 3 High-Probability Guarantees for Convex Objectives

In this section, we present our main high-probability convergence guarantees for Clip-SGD under convexity, $(L_0, L_1)$-smoothness, and heavy-tailed stochastic gradients.

### 3.1 Best-Iterate Convergence

We establish a high-probability convergence guarantee for the best iterate generated by Clip-SGD.

The case $L_0 R_0 + \sigma = 0$ is immediate. Indeed, if $R_0 = 0$, then $x_0 = x^\star$. If $R_0 > 0$, then $L_0 = \sigma = 0$, and since $f$ admits a minimizer, Assumption 2 implies that $f$ is constant. Therefore, in either case, $\min_{0 \le k < K} (f(x_k) - f^\star) = 0$. We therefore restrict attention to $L_0 R_0 + \sigma > 0$. This condition ensures that the clipping level in the following theorem is strictly positive and, consequently, that the displayed stepsize is well-defined.

**Theorem 1.** *Suppose that Assumptions 1, 2, and 3 hold. Let $\delta \in (0, 1)$ and assume that $L_0 R_0 + \sigma > 0$. Suppose that $K \ge \frac{160}{\nu^2} \max\left\{L_1 R_0, \log \frac{4}{\delta}\right\}^2$. Choose the clipping level $\lambda$ and stepsize $\gamma$ as specified in Table 3. Then, with probability at least $1 - \delta$,*

$$\min_{0 \le k < K} (f(x_k) - f^\star) \le \frac{40}{\nu} \max\left\{ \frac{L_0 R_0^2}{K}, \frac{R_0 \sigma \max\left\{L_1 R_0, \log \frac{4}{\delta}\right\}^{\frac{\alpha-1}{\alpha}}}{K^{\frac{\alpha-1}{\alpha}}} \right\}.$$

**Proof overview.** **1) Deterministic backbone.** We first recall the standard argument for gradient descent on a convex $L$-smooth function. With $x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k)$, we have

$$\|x_{k+1} - x^\star\|^2 = \|x_k - x^\star\|^2 - \frac{2}{L}\langle \nabla f(x_k), x_k - x^\star\rangle + \frac{1}{L^2}\|\nabla f(x_k)\|^2 .$$

The argument relies on the two inequalities

$$\langle \nabla f(x_k), x_k - x^\star\rangle \geq f(x_k) - f^\star, \qquad \|\nabla f(x_k)\|^2 \leq L\langle \nabla f(x_k), x_k - x^\star\rangle ,$$

where the first follows from convexity and the second from convexity and $L$-smoothness. The second inequality absorbs the quadratic term into the progress term, while the first converts the remaining progress into an objective-gap bound:

$$\frac{1}{L}\big(f(x_k) - f^\star\big) \leq \|x_k - x^\star\|^2 - \|x_{k+1} - x^\star\|^2 .$$

Our analysis follows the same structure for the clipped population gradient $q_k := \texttt{clip}(\nabla f(x_k), \lambda)$. Under convexity and $(L_0, L_1)$-smoothness, our structural lemmas establish

$$\langle q_k, x_k - x^\star\rangle \geq \min\left\{f(x_k) - f^\star, \frac{\nu\lambda^2}{2(L_0 + L_1\lambda)}\right\}, \qquad \|q_k\|^2 \leq \frac{2(L_0 + L_1\lambda)}{\nu}\langle q_k, x_k - x^\star\rangle .$$

The second inequality again allows us to absorb the quadratic term generated by the update. The first converts the remaining progress into the objective gap, truncated at the level $\frac{\nu\lambda^2}{2(L_0+L_1\lambda)}$.

Crucially, these inequalities depend on the scale $L_0 + L_1\lambda$ without requiring a uniform bound on $\|\nabla f(x)\|$ over the localization region. Such a bound can grow exponentially in $L_1 R_0$ under generalized smoothness, introducing an exponentially large local smoothness constant. Working directly with the clipped population gradient avoids this reduction.

**2) Stochastic perturbations.** Let $G_k := \nabla f(x_k, \xi_k)$, $g_k := \texttt{clip}(G_k, \lambda)$, and $e_k := g_k - q_k$. Since $x_{k+1} = x_k - \gamma(q_k + e_k)$, the structural inequalities imply, for $\gamma \leq \frac{\nu}{8(L_0+L_1\lambda)}$,

$$\begin{aligned}\frac{3\gamma}{2}\min\left\{f(x_k) - f^\star, \frac{\nu\lambda^2}{2(L_0 + L_1\lambda)}\right\} &\leq \frac{3\gamma}{2}\langle q_k, x_k - x^\star\rangle \\ &\leq \|x_k - x^\star\|^2 - \|x_{k+1} - x^\star\|^2 \\ &\quad - 2\gamma\langle e_k - \mathbb{E}[e_k \mid \mathcal{F}_k], x_k - x^\star\rangle \\ &\quad - 2\gamma\langle \mathbb{E}[e_k \mid \mathcal{F}_k], x_k - x^\star\rangle + 2\gamma^2\|e_k\|^2 .\end{aligned} \tag{3}$$

Thus, three stochastic terms perturb the deterministic argument: a centered martingale term, the clipping bias, and a quadratic error term.

We first address the clipping bias $b_k := \mathbb{E}[e_k \mid \mathcal{F}_k]$. A direct norm bound on $b_k$ can be too crude when the population gradient is large. Instead, we control its component in the direction $x_k - x^\star$, which is the quantity appearing in the distance recursion. Our key ingredient, Lemma 9, gives

$$-\langle b_k, x_k - x^\star\rangle \leq \frac{\sigma}{\lambda}\langle q_k, x_k - x^\star\rangle + \|x_k - x^\star\|\frac{\sigma^\alpha}{\lambda^{\alpha-1}}.$$

Suppose temporarily that $\|x_k - x^\star\| \leq R$. For $\lambda \geq 8\sigma$, this yields

$$-2\gamma\langle b_k, x_k - x^\star\rangle \leq \frac{\gamma}{4}\langle q_k, x_k - x^\star\rangle + 2\gamma R\frac{\sigma^\alpha}{\lambda^{\alpha-1}}.$$

The first term is absorbed into the $\frac{3\gamma}{2}\langle q_k, x_k - x^\star\rangle$ progress term in Eq. (3), leaving $\frac{5\gamma}{4}\langle q_k, x_k - x^\star\rangle$. The first structural inequality then converts this remaining progress into the truncated objective gap. Thus, the bias contributes only the heavy-tail remainder $2\gamma R\sigma^\alpha/\lambda^{\alpha-1}$ after absorption.

We next control the quadratic error. Although Assumption 3 only assumes a bounded $\alpha$-th moment, clipping gives a conditional second-moment bound. Indeed, nonexpansiveness of clipping and the bound on the clipped vectors imply

$$\|e_k\| \leq \min\{\|G_k - \nabla f(x_k)\|, 2\lambda\}, \qquad \mathbb{E}[\|e_k\|^2 \mid \mathcal{F}_k] \leq 2^{2-\alpha}\sigma^\alpha\lambda^{2-\alpha}.$$

In particular, this bound does not require the population gradient to lie below the clipping threshold. Splitting the cumulative quadratic error into its conditional mean and a centered martingale sum, and applying Freedman's inequality, gives the high-probability bound in Lemma 7.

Finally, let $\zeta_k := e_k - b_k$. The term $-2\gamma \langle \zeta_k, x_k - x^\star \rangle$ is a martingale difference. Under the temporary radius bound, its magnitude is of order $\gamma R\lambda$, and its conditional variance is controlled by $\gamma^2 R^2 \mathbb{E}[\|e_k\|^2 \mid \mathcal{F}_k]$. The same second-moment estimate therefore allows another application of Freedman's inequality to control its cumulative contribution.

**3) Localization.** We now justify the temporary radius bound. Define $r_k := \|x_k - x^\star\|$ and $\tau := \inf\{k \geq 0 : r_k > 2R_0\}$, with $\inf \varnothing = +\infty$. Multiplying Eq. (3) by $\mathbf{1}_{\{k<\tau\}}$ and summing makes the distance terms telescope, since $\mathbf{1}_{\{k<\tau\}}\left(r_k^2 - r_{k+1}^2\right) = r_{k\wedge\tau}^2 - r_{(k+1)\wedge\tau}^2$. Combining the bias and concentration bounds for the stopped process, with $R = 2R_0$, gives, with probability at least $1-\delta$,

$$r_{K\wedge\tau}^2 + \frac{5\gamma}{4}\sum_{k=0}^{K-1} \mathbf{1}_{\{k<\tau\}} \min\left\{f(x_k) - f^\star, \frac{\nu\lambda^2}{2(L_0 + L_1\lambda)}\right\} \leq R_0^2 + A_K.$$

Here $A_K$ collects the accumulated bias, martingale fluctuations, and quadratic error. The parameter choices in Theorem 1 ensure $A_K \leq 2R_0^2$, as verified in the proof of that theorem.

On this event, $r_{K\wedge\tau}^2 \leq 3R_0^2$. If $\tau \leq K$, the definition of $\tau$ would instead give $r_{K\wedge\tau}^2 = r_\tau^2 > 4R_0^2$, a contradiction. Hence $\tau > K$. Thus, the temporary radius bound is valid for the entire trajectory up to time $K$, and the stopping indicators can be removed:

$$\frac{1}{K}\sum_{k=0}^{K-1} \min\left\{f(x_k) - f^\star, \frac{\nu\lambda^2}{2(L_0 + L_1\lambda)}\right\} \leq \frac{12R_0^2}{5\gamma K}.$$

The assumed lower bound on $K$ places the right-hand side strictly below the truncation level. Therefore, at least one iterate has an untruncated objective gap at most $\frac{12R_0^2}{5\gamma K}$.

To see how the rate follows, write $M := \max\{L_1R_0, \log(4/\delta)\}$. The theorem chooses

$$\lambda = \max\left\{\frac{L_0R_0}{M}, \sigma\left(\frac{K}{M}\right)^{1/\alpha}\right\}, \qquad \gamma = \frac{\nu R_0}{16\lambda M}.$$

The first term in $\lambda$ ensures $L_0 \leq \lambda M/R_0$; together with $L_1R_0 \leq M$, this gives $L_0 + L_1\lambda \leq 2\lambda M/R_0$, which ensures the required smoothness restriction on $\gamma$. The second term ensures $(\sigma/\lambda)^\alpha \leq M/K$, which keeps the heavy-tail remainder at the scale of the final optimization error. Substituting these choices into the preceding bound gives

$$\min_{0\leq k<K}\left(f(x_k) - f^\star\right) \leq \frac{40}{\nu}\max\left\{\frac{L_0R_0^2}{K}, \frac{R_0\sigma M^{(\alpha-1)/\alpha}}{K^{(\alpha-1)/\alpha}}\right\}.$$

### 3.2 A Computable Output

The guarantee in Theorem 1 concerns the best iterate, which is generally not observable. We now show that the same convergence rate can be achieved by an output constructed solely from the stochastic gradients already queried by Clip-SGD, without any additional oracle calls.

**Theorem 2.** *Suppose that Assumptions 1, 2, and 3 hold. Let $\delta \in (0,1)$ and assume that $L_0, R_0 > 0$. Suppose that $K \geq 1024\left[\log^2\left(\frac{12}{\delta}\right) + L_1R_0\log\left\{\frac{12e}{\delta}\left(1 + \frac{K\Delta_0}{L_0R_0^2}\right)\right\}\right]$. Choose the clipping level $\lambda$ and stepsize $\gamma$ as specified in Table 4. Let $\mathcal{U} = \{0 \leq k < K : \|G_k\| \leq \lambda\}$, and set $\widehat{x}_K = |\mathcal{U}|^{-1}\sum_{k\in\mathcal{U}} x_k$ when $\mathcal{U} \neq \varnothing$, and $\widehat{x}_K = x_0$ otherwise. Then, with probability at least $1-\delta$,*

$$f(\widehat{x}_K) - f^\star \leq \frac{128}{\nu}\max\left\{\frac{L_0R_0^2}{K}, \frac{R_0\sigma\max\left\{L_1R_0, \log\frac{12}{\delta}\right\}^{\frac{\alpha-1}{\alpha}}}{K^{\frac{\alpha-1}{\alpha}}}\right\}.$$

**Proof overview.** We show that the observed clipping test selects enough iterates, control a trimmed average used only in the analysis, and transfer its guarantee to the actual output.

**1) Ensuring enough selected iterates.** Let $G_k := \nabla f(x_k, \xi_k)$, $q_k := \texttt{clip}(\nabla f(x_k), \lambda)$, and $M := \max\{L_1 R_0, \log(12/\delta)\}$. Repeating the localization argument with $\delta$ replaced by $\delta/3$, while retaining the population progress term, gives $\tau > K$, $\|x_k - x^\star\| \le 2R_0$ for all $k \le K$, and $\sum_{k<K} \langle q_k, x_k - x^\star \rangle \le 8R_0^2/(5\gamma)$ with probability at least $1 - \delta/3$. We combine this event with the concentration events below by a union bound; their intersection has probability at least $1 - \delta$.

For our parameter choices, Lemma 12 gives $\mathbb{E}[f(x_{k+1}) - f(x_k) \mid \mathcal{F}_k] \le -\frac{3\gamma}{4} \langle \nabla f(x_k), q_k \rangle + \frac{R_0 \lambda}{K}$.

Let $\mathcal{C} := \{0 \le k < K : \|\nabla f(x_k)\| < 5\lambda/6\}$ and $D_k := f(x_k) - f^\star + R_0\lambda$. For $k < \tau$ and $k \notin \mathcal{C}$, convexity and clipping imply $\langle \nabla f(x_k), q_k \rangle \ge \lambda D_k/(4R_0)$. For the iterations such that $k < \tau$ and $k \in \mathcal{C}$, we drop the nonpositive descent term $-\frac{3\gamma}{4} \langle \nabla f(x_k), q_k \rangle$. Since $D_k \ge R_0\lambda$, both cases yield, for $k < \tau$,

$$\mathbb{E}[D_{k+1} \mid \mathcal{F}_k] \le \left(1 - \frac{3\gamma\lambda}{16R_0}\mathbf{1}_{\{k \notin \mathcal{C}\}} + \frac{1}{K}\right) D_k \le \exp\left(-\frac{3\gamma\lambda}{16R_0}\mathbf{1}_{\{k \notin \mathcal{C}\}} + \frac{1}{K}\right) D_k.$$

Every iteration with a large population gradient therefore contributes a fixed negative term to the exponent. Since the recurrence holds for $k < \tau$, we apply it up to the stopping time $\tau$, which gives an exponential-moment bound on the number of large-gradient iterations before $\tau$. The exponential form of Markov's inequality then gives

$$\sum_{k<K} \mathbf{1}_{\{k<\tau,\ k \notin \mathcal{C}\}} \le \frac{16R_0}{3\gamma\lambda} \log\left(\frac{6e}{\delta}\left(1 + \frac{\Delta_0}{R_0\lambda}\right)\right).$$

This explains the logarithmic dependence on the initial objective gap. The iteration requirement makes the right-hand side at most $K/4$, so localization gives $|\mathcal{C}| \ge 3K/4$; see Lemma 13.

Define $\mathcal{U} := \{0 \le k < K : \|G_k\| \le \lambda\}$. If $k \in \mathcal{C}$ is not selected, then $\|G_k - \nabla f(x_k)\| > \lambda/6$, so $\Pr(k \in \mathcal{C} \setminus \mathcal{U} \mid \mathcal{F}_k) \le (6\sigma/\lambda)^\alpha$. Martingale concentration bounds the number of these rejected indices by $K/12$, giving $|\mathcal{U}| \ge |\mathcal{C}| - |\mathcal{C} \setminus \mathcal{U}| \ge \frac{2K}{3}$. The details are given in Lemma 14.

**2) Controlling the trimmed average.** A selected iterate can have a large population gradient only if the stochastic error is large enough to reduce the observed gradient below the threshold. Introduce the exceptional set $\mathcal{O} := \left\{k \in \mathcal{U} : \|G_k - \nabla f(x_k)\| > \sigma\left(K/\log(12/\delta)\right)^{1/\alpha}\right\}$. This exclusion is used only in the analysis; the actual output retains all indices in $\mathcal{U}$. The moment assumption and martingale concentration give $|\mathcal{O}| < 4\log(12/\delta)$. We control the remaining selected indices by splitting according to their population gradient norms.

For $\|\nabla f(x_k)\| \le 2\lambda$, convexity gives $f(x_k) - f^\star \le 2\langle q_k, x_k - x^\star \rangle$. The population progress bound therefore controls their total contribution.

For $\|\nabla f(x_k)\| > 2\lambda$, selection $k \in \mathcal{U}$ requires $\|G_k - \nabla f(x_k)\| \ge \|\nabla f(x_k)\| - \lambda$. On this large-gradient event,

$$\Pr(k \in \mathcal{U} \mid \mathcal{F}_k) \le \frac{\sigma^\alpha}{\left(\|\nabla f(x_k)\| - \lambda\right)^\alpha}.$$

This upper bound decreases with the population gradient, while convexity and localization give $f(x_k) - f^\star \le 2R_0 \|\nabla f(x_k)\|$. Define $Y_k := \mathbf{1}_{\{k<\tau,\ k \in \mathcal{U}\setminus\mathcal{O},\ \|\nabla f(x_k)\| > 2\lambda\}}\left(f(x_k) - f^\star\right)$. Combining the preceding bounds gives $\mathbb{E}[Y_k \mid \mathcal{F}_k] \le 4R_0\sigma^\alpha/\lambda^{\alpha-1}$ and $\sum_{k<K} \mathbb{E}[Y_k \mid \mathcal{F}_k] \le 4R_0\lambda M$, where the second inequality uses $(\sigma/\lambda)^\alpha \le M/K$. Thus, rare selection compensates for the potentially large objective gap. On $\mathcal{U} \setminus \mathcal{O}$, we also have $\|\nabla f(x_k)\| \le \lambda + \sigma\left(K/\log(12/\delta)\right)^{1/\alpha}$. Together with the stopping indicator, this bounds $Y_k$ deterministically. Martingale concentration then gives $\sum_{k<K} Y_k = O(R_0\lambda M)$. Combining this with the small-gradient contribution and $\gamma = \nu R_0/(16\lambda M)$ gives

$$\sum_{k \in \mathcal{U}\setminus\mathcal{O}} \left(f(x_k) - f^\star\right) = O\left(\frac{R_0\lambda M}{\nu}\right).$$

Since $|\mathcal{U}| \ge 2K/3$ and $|\mathcal{O}| < 4\log(12/\delta)$, the iteration requirement ensures that $\mathcal{U} \setminus \mathcal{O}$ contains a constant fraction of all iterates. The trimmed average $\bar{x}_K := \frac{1}{|\mathcal{U}\setminus\mathcal{O}|} \sum_{k \in \mathcal{U}\setminus\mathcal{O}} x_k$ is therefore well-defined on the common event, and convexity gives $\Delta := f(\bar{x}_K) - f^\star = O\left(\frac{R_0\lambda M}{\nu K}\right)$.

**3) Transferring the guarantee to the actual output.** The actual output $\widehat{x}_K$ averages all of $\mathcal{U}$. Localization bounds the distance between any iterate and $\bar{x}_K$ by $4R_0$, so

$$D_K := \|\widehat{x}_K - \bar{x}_K\| \le \frac{4R_0|\mathcal{O}|}{|\mathcal{U}|} < \frac{24R_0\log(12/\delta)}{K}.$$

The iteration requirement ensures $L_1 D_K < 1/32$. Lemma 3 gives $\|\nabla f(\bar{x}_K)\| \le 4L_1\Delta + 2\sqrt{L_0\Delta}$. Combining this with local smoothness and Young's inequality yields $f(\widehat{x}_K) - f^\star \le \frac{5}{4}\Delta + 12L_0 D_K^2$. Moreover, the iteration requirement and $\lambda M \ge L_0 R_0$ give $L_0 D_K^2 = O(R_0\lambda M/K)$. Thus, restoring the excluded indices preserves the rate. Keeping the numerical constants, as verified in the full proof, gives

$$f(\widehat{x}_K) - f^\star \le \frac{128}{\nu}\frac{R_0\lambda M}{K} = \frac{128}{\nu}\max\left\{\frac{L_0R_0^2}{K},\, R_0\sigma\left(\frac{M}{K}\right)^{(\alpha-1)/\alpha}\right\}.$$

The common event has probability at least $1-\delta$, proving the claim.

### 3.3 Optimality for Fixed-Parameter Clip-SGD

We next establish a lower bound for Clip-SGD with any fixed stepsize and clipping level. Under the iteration conditions below, it matches the dependence of the stochastic term in Theorem 2 on $L_1R_0$ and $\log(1/\delta)$ up to logarithmic factors in the iteration budget, even when the output is an arbitrary randomized function of the observed transcript. We state the result for $L_0 = 1$.

**Theorem 3.** *Let $L_1, R_0, \sigma > 0$, $\alpha \in (1,2]$, and $\delta \in (0, 1/4)$. Suppose the integer $K$ satisfies*

$$K \ge \max\left\{1,\, L_1R_0,\, \sigma L_1,\, \log\frac{1}{2\delta}\max\left\{1,\, \left(\frac{\sigma}{R_0}\right)^{\alpha/(\alpha-1)}\right\}\right\}.$$

*Then, for every fixed $\gamma, \lambda > 0$ and every fixed possibly randomized output rule depending only on the observed transcript and independent auxiliary randomness, let $x_K^{\mathrm{out}}$ denote the point returned by this rule. There exist a convex $(1, L_1)$-smooth function $f : \mathbb{R} \to \mathbb{R}$, with a minimizer $x^\star$ satisfying $|x^\star| \le R_0$, and an unbiased stochastic-gradient oracle satisfying Assumption 3, such that Clip-SGD, started from $x_0 = 0$ and using fresh independent oracle samples, satisfies*

$$\mathbb{P}\left(f(x_K^{\mathrm{out}}) - f^\star \ge \frac{\sigma R_0}{128}\left[\frac{\max\{L_1R_0/\log(eK),\, \log(12/\delta)\}}{K}\right]^{(\alpha-1)/\alpha}\right) > \delta.$$

*The fixed parameters and output rule may depend on $K$ and $\delta$.*

## 4 Conclusion

We established high-probability convergence guarantees for standard Clip-SGD on convex $(L_0, L_1)$-smooth objectives under heavy-tailed noise, without exponential dependence on $L_1R_0$. The analysis combines a directional clipping-bias bound with structural inequalities for the clipped population gradient. The resulting guarantee applies to a computable average of iterates selected by the observed clipping test, requiring neither function values nor additional oracle calls. We also established a lower bound showing that, under the stated iteration conditions, our convex stochastic rate, including its confidence dependence, is optimal up to logarithmic factors in the iteration budget for Clip-SGD with any fixed stepsize and clipping level. We further obtained high-probability bounds for the average gradient norm on nonconvex $(L_0, L_1)$-smooth objectives under the same heavy-tailed noise assumption; see Appendix D. Our convex and nonconvex bounds recover the corresponding noiseless rates and the classical-smooth stochastic iteration rates when $L_1 = 0$. In the nonconvex case, the asymptotically dominant stochastic term has no explicit $L_1$ dependence and no logarithmic factors in $K$ or $1/\delta$.

Interesting directions include understanding whether adaptive parameter choices or different algorithms can improve the convex stochastic rate, and developing accelerated methods with comparable high-probability guarantees under heavy-tailed noise.

We hope that our results inspire further research in these directions and contribute to the broader understanding of stochastic optimization under realistic noise and smoothness assumptions.

### AI USE STATEMENT

Generative AI tools were used only for language editing and proofreading, including improving grammar, clarity, and presentation. All mathematical claims, proofs, derivations, and technical conclusions were developed and carried out by the authors. The authors take full responsibility for the content of the paper.

### REPRODUCIBILITY STATEMENT

This work is primarily theoretical. All assumptions, algorithmic details, and parameter choices required to reproduce the stated guarantees are provided in the main text and appendices. Complete proofs of the main results and supporting lemmas are given in the appendices.

## REFERENCES


Weixin An, Yuanyuan Liu, Fanhua Shang, Han Yu, Junkang Liu, and Hongying Liu. Tight high-probability bounds for nonconvex heavy-tailed scenario under weaker assumptions. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=vEFAR8KH1l.

Dimitri P Bertsekas. Incremental proximal methods for large scale convex optimization. *Mathematical programming*, 129(2):163–195, 2011.

Alina Beygelzimer, John Langford, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandit algorithms with supervised learning guarantees. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 19–26. JMLR Workshop and Conference Proceedings, 2011.

Bensaid Bilel. Complexities of armijo-like algorithms in deep learning context. *arXiv preprint arXiv:2412.14637*, 2024.

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.

Ziyi Chen, Yi Zhou, Yingbin Liang, and Zhaosong Lu. Generalized-smooth nonconvex optimization is as efficient as smooth nonconvex optimization. In *International Conference on Machine Learning*, pp. 5396–5427. PMLR, 2023.

Savelii Chezhegov, Yaroslav Klyukin, Andrei Semenov, Aleksandr Beznosikov, Alexander Gasnikov, Samuel Horváth, Martin Takáč, and Eduard Gorbunov. Clipping improves adam-norm and adagrad-norm when the noise is heavy-tailed. *arXiv preprint arXiv:2406.04443*, 2024.

Savelii Chezhegov, Daniela Angela Parletta, Andrea Paudice, and Eduard Gorbunov. High-probability bounds for the last iterate of clipped SGD. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=4sGEvpwyxN.

Michael Crawshaw and Mingrui Liu. Complexity lower bounds of adaptive gradient algorithms for non-convex stochastic optimization under relaxed smoothness. In *International Conference on Learning Representations*, 2025.

Michael Crawshaw, Mingrui Liu, Francesco Orabona, Wei Zhang, and Zhenxun Zhuang. Robustness to unbounded smoothness of generalized signsgd. *Advances in neural information processing systems*, 35:9955–9968, 2022.

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.

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.

Pavel Dvurechensky and Alexander Gasnikov. Stochastic intermediate gradient method for convex problems with stochastic inexact oracle. *Journal of Optimization Theory and Applications*, 171: 121–145, 2016.

Matthew Faw, Litu Rout, Constantine Caramanis, and Sanjay Shakkottai. Beyond uniform smoothness: A stopped analysis of adaptive sgd. In *The Thirty Sixth Annual Conference on Learning Theory*, pp. 89–160. PMLR, 2023.

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.

Ofir Gaash, Kfir Yehuda Levy, and Yair Carmon. Convergence of clipped SGD on convex $(l_0, l_1)$-smooth functions. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Alexander Gasnikov and Yurii Nesterov. Universal fast gradient method for stochastic composit optimization problems. *arXiv preprint arXiv:1604.05275*, 2016.

Saeed Ghadimi and Guanghui Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization i: A generic algorithmic framework. *SIAM Journal on Optimization*, 22(4):1469–1492, 2012.

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

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

Eduard Gorbunov, Marina Danilova, and Alexander Gasnikov. Stochastic optimization with heavy-tailed noise via accelerated gradient clipping. *Advances in Neural Information Processing Systems*, 33:15042–15053, 2020.

Eduard Gorbunov, Marina Danilova, David Dobre, Pavel Dvurechenskii, Alexander Gasnikov, and Gauthier Gidel. Clipped stochastic methods for variational inequalities with heavy-tailed noise. *Advances in Neural Information Processing Systems*, 35:31319–31332, 2022.

Eduard Gorbunov, Marina Danilova, Innokentiy Shibaev, Pavel Dvurechensky, and Alexander Gasnikov. High-probability complexity bounds for non-smooth stochastic convex optimization with heavy-tailed noise. *Journal of Optimization Theory and Applications*, pp. 1–60, 2024a.

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

Eduard Gorbunov, Nazarii Tupitsa, Sayantan Choudhury, Alen Aliev, Peter Richtárik, Samuel Horváth, and Martin Takáč. Methods for convex $(L_0, L_1)$-smooth optimization: Clipping, acceleration, and adaptivity. *International Conference on Learning Representations*, 2025.

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.

Florian Hübler, Junchi Yang, Xiang Li, and Niao He. Parameter-agnostic optimization under relaxed smoothness. In *International Conference on Artificial Intelligence and Statistics*, pp. 4861–4869. PMLR, 2024.

Saleh Vatan Khah, Savelii Chezhegov, Shahrokh Farahmand, Samuel Horváth, and Eduard Gorbunov. Differentially private clipped-sgd: High-probability convergence with arbitrary clipping level. *arXiv preprint arXiv:2507.23512*, 2025.

Sarit Khirirat, Abdurakhmon Sadiev, Artem Riabinin, Eduard Gorbunov, and Peter Richtárik. Error feedback under $(L_0, L_1)$-smoothness: Normalization and momentum. *arXiv preprint arXiv:2410.16871*, 2024.

Anastasia Koloskova, Hadrien Hendrikx, and Sebastian U Stich. Revisiting gradient clipping: Stochastic bias and tight convergence guarantees. In *International Conference on Machine Learning*, pp. 17343–17363. PMLR, 2023.

Nikita Kornilov, Ohad Shamir, Aleksandr Lobanov, Darina Dvinskikh, Alexander Gasnikov, Innokentiy Shibaev, Eduard Gorbunov, and Samuel Horváth. Accelerated zeroth-order method for non-smooth stochastic convex optimization problem with infinite variance. *Advances in Neural Information Processing Systems*, 36:64083–64102, 2023.

Su Hyeong Lee, Manzil Zaheer, and Tian Li. Efficient distributed optimization under heavy-tailed noise. *arXiv preprint arXiv:2502.04164*, 2025.

Haochuan Li, Jian Qian, Yi Tian, Alexander Rakhlin, and Ali Jadbabaie. Convex and non-convex optimization under generalized smoothness. *Advances in Neural Information Processing Systems*, 36:40238–40271, 2023a.

Haochuan Li, Alexander Rakhlin, and Ali Jadbabaie. Convergence of adam under relaxed assumptions. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023b. URL https://openreview.net/forum?id=yEewbkBNzi.

Shaojie Li and Yong Liu. High probability analysis for non-convex stochastic optimization with clipping. In *ECAI 2023*, pp. 1406–1413. IOS Press, 2023.

Xiaoyu Li and Francesco Orabona. A high probability analysis of adaptive sgd with momentum. *arXiv preprint arXiv:2007.14294*, 2020.

Zijian Liu and Zhengyuan Zhou. Nonconvex stochastic optimization under heavy-tailed noises: Optimal convergence without gradient clipping. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=NKotdPUc3L.

Zijian Liu, Ta Duy Nguyen, Thien Hang Nguyen, Alina Ene, and Huy Nguyen. High probability convergence of stochastic gradient methods. In *International Conference on Machine Learning*, pp. 21884–21914. PMLR, 2023.

Aleksandr Lobanov, Alexander Gasnikov, Eduard Gorbunov, and Martin Takáč. Linear convergence rate in convex setup is possible! gradient descent method variants under $(L_0, L_1)$-smoothness. *arXiv preprint arXiv:2412.17050*, 2024.

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.

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*, 80:1607–1627, 2019.

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

Yurii Evgenievich Nesterov. A method for solving the convex programming problem with convergence rate O$(1/k^2)$. In *Dokl. akad. nauk Sssr*, volume 269, pp. 543–547, 1983.

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.

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.

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.

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

Abdurakhmon Sadiev, Marina Danilova, Eduard Gorbunov, Samuel Horváth, Gauthier Gidel, Pavel Dvurechensky, Alexander Gasnikov, and Peter Richtárik. High-probability bounds for stochastic optimization and variational inequalities: the case of unbounded variance. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 29563–29648. PMLR, 23–29 Jul 2023.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Yuki Takezawa, Han Bao, Ryoma Sato, Kenta Niwa, and Makoto Yamada. Parameter-free clipped gradient descent meets polyak. *Advances in Neural Information Processing Systems*, 37:44575–44599, 2024.

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.

Zhirayr Tovmasyan, Grigory Malinovsky, Laurent Condat, and Peter Richtárik. Revisiting stochastic proximal point methods: Generalized smoothness and similarity. *arXiv preprint arXiv:2502.03401*, 2025.

Alexander Tyurin. Toward a unified theory of gradient descent under generalized smoothness. *arXiv preprint arXiv:2412.11773*, 2024.

Daniil Vankov, Anton Rodomanov, Angelia Nedich, Lalitha Sankar, and Sebastian U Stich. Optimizing $(L_0, L_1)$-smooth functions by gradient methods. *International Conference on Learning Representations*, 2025.

Bohan Wang, Yushun Zhang, Huishuai Zhang, Qi Meng, Zhi-Ming Ma, Tie-Yan Liu, and Wei Chen. Provable adaptivity in adam. *arXiv preprint arXiv:2208.09900*, 2022.

Bohan Wang, Huishuai Zhang, Zhiming Ma, and Wei Chen. Convergence of adagrad for non-convex objectives: Simple proofs and relaxed assumptions. In *The Thirty Sixth Annual Conference on Learning Theory*, pp. 161–190. PMLR, 2023.

Yufeng Yang, Erin Tripp, Yifan Sun, Shaofeng Zou, and Yi Zhou. Independently-normalized sgd for generalized-smooth nonconvex optimization. *arXiv preprint arXiv:2410.14054*, 2024.

Chenhao Yu, Yusu Hong, and Junhong Lin. Convergence analysis of stochastic accelerated gradient methods for generalized smooth optimizations. *arXiv preprint arXiv:2502.11125*, 2025a.

Dingzhi Yu, Wei Jiang, Yuanyu Wan, and Lijun Zhang. Mirror descent under generalized smoothness. *arXiv preprint arXiv:2502.00753*, 2025b.

Bohang Zhang, Jikai Jin, Cong Fang, and Liwei Wang. Improved analysis of clipping algorithms for non-convex optimization. In *Advances in Neural Information Processing Systems*, 2020a.

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*, 2020b.

Jingzhao Zhang, Sai Praneeth Karimireddy, Andreas Veit, Seungyeon Kim, Sashank Reddi, Sanjiv Kumar, and Suvrit Sra. Why are adaptive methods good for attention models? *Advances in Neural Information Processing Systems*, 33:15383–15393, 2020c.

Shen-Yi Zhao, Yin-Peng Xie, and Wu-Jun Li. On the convergence and improvement of stochastic normalized gradient descent. *Science China Information Sciences*, 64:1–13, 2021.

## CONTENTS

**1 Introduction** **1**

1.1 Problem setup and assumptions . . . . . 2

1.2 Our contributions . . . . . 3

**2 Related Work** **4**

**3 High-Probability Guarantees for Convex Objectives** **5**

3.1 Best-Iterate Convergence . . . . . 5

3.2 A Computable Output . . . . . 7

3.3 Optimality for Fixed-Parameter Clip-SGD . . . . . 9

**4 Conclusion** **9**

**A Additional Related Work** **15**

**B Preliminary Results and Notation Table** **15**

**C High-Probability Guarantees for Convex Objectives** **18**

C.1 Best-Iterate Convergence . . . . . 18

C.1.1 Proofs . . . . . 22

C.2 A Computable Output . . . . . 29

C.2.1 Proofs . . . . . 30

C.3 Detailed comparison underlying Table 1 . . . . . 37

**D High-Probability Guarantees for Nonconvex Objectives** **39**

D.1 Convergence Analysis . . . . . 39

D.2 Comparison with prior work . . . . . 43

D.3 Proofs . . . . . 45

**E A Lower Bound for Fixed-Parameter Clip-SGD in the Convex Setting** **52**

E.1 Dependence on the failure probability . . . . . 53

E.2 Dependence on $L_1 R_0$ . . . . . 54

E.3 Proofs . . . . . 54

# A Additional Related Work

This section complements the related-work discussion in the main text by providing additional details on deterministic convex guarantees, high-probability results under light-tailed noise, and further applications of gradient clipping.

**Detailed deterministic convex guarantees.** Koloskova et al. (2023) provided convergence guarantees for Clip-GD under convexity, $(L_0, L_1)$-smoothness and $L$-smoothness, deriving a complexity bound of

$$\mathcal{O}\left(\max\left\{(L_0+\lambda L_1)R_0^2/\varepsilon, \sqrt{R_0^4 L(L_0+\lambda L_1)^2/\lambda^2\varepsilon}\right\}\right).$$

The leading term in this complexity bound is independent of $L_1$ and $L$ if $\lambda \sim L_0/L_1$, and can significantly improve upon the corresponding bound for standard GD. Building on this, Takezawa et al. (2024) analyzed GD with Polyak stepsizes and derived the complexity bound

$$\mathcal{O}\left(\max\left\{L_0 R_0^2/\varepsilon, \sqrt{R_0^4 L L_1^2/\varepsilon}\right\}\right).$$

For convex $(L_0, L_1)$-smooth objectives, Li et al. (2023a) obtain $\mathcal{O}(1/\varepsilon)$ and $\mathcal{O}(1/\sqrt{\varepsilon})$ iteration complexities for GD and Nesterov's accelerated gradient method (Nesterov, 1983), respectively, with constants depending on the generalized-smoothness parameters and the initialization. Through the additional smoothness quantities appearing in these analyses, the bounds from Koloskova et al. (2023); Takezawa et al. (2024); Li et al. (2023a) can introduce exponential dependence on $L_1 R_0$, a drawback addressed by the more recent results of Gorbunov et al. (2025); Vankov et al. (2025); Lobanov et al. (2024).

**High-probability convergence under light-tailed noise.** High-probability convergence guarantees have long been a critical component in the analysis of stochastic first-order methods, particularly when the noise in the stochastic gradients is light-tailed. In these settings, methods like SGD and its variants can achieve convergence rates with polylogarithmic dependence on the inverse failure probability $1/\delta$. Under sub-Gaussian noise assumptions, this behavior has been rigorously established for SGD (Nemirovski et al., 2009; Harvey et al., 2019), its accelerated counterparts (Ghadimi & Lan, 2012; Dvurechensky & Gasnikov, 2016), and adaptive methods like AdaGrad (Li & Orabona, 2020; Liu et al., 2023). Recent extensions to broader classes of noise, such as sub-Weibull distributions, have further expanded this theoretical framework (Madden et al., 2024). The results of Li et al. (2023b); Yu et al. (2025a) rely on light-tailed noise assumptions related to Eq. (2). In particular, Yu et al. (2025a) use a more general version of Eq. (2), with $\sigma^2$ replaced by

$$A(f(x) - f^\star) + B\|\nabla f(x)\|^2 + C.$$

**Further results under heavy-tailed noise.** High-probability clipping guarantees have also been extended to broader problem classes, including nonsmooth optimization (Gorbunov et al., 2024a; Parletta et al., 2024) and unconstrained variational inequalities (Gorbunov et al., 2022). Adaptive variants have also been developed: Li & Liu (2023) analyzed Clip-AdaGrad with scalar stepsizes, while Chezhegov et al. (2024) obtained similar bounds for both scalar and coordinate-wise versions of Clip-AdaGrad and Clip-Adam. In the zeroth-order setting, Kornilov et al. (2023) proposed a clipped variant of SSTM. For distributed optimization, Lee et al. (2025) provided in-expectation convergence rates for the TailOPT method. Moreover, clipping was adapted to differential privacy with high-probability convergence guarantees (Khah et al., 2025).

# B Preliminary Results and Notation Table

**Proposition 1** ((Gorbunov et al., 2025, Lemma 2.2)). *Suppose Assumption 2 holds. Then, for all* $x \in \mathbb{R}^d$,

$$\nu\|\nabla f(x)\|^2 \le 2\left(L_0 + L_1\|\nabla f(x)\|\right)(f(x) - f^\star),$$

*where* $\nu$ *is the solution of* $x\exp(x) = 1$.

**Lemma 1** (Consequence of (Gorbunov et al., 2025, Lemma 2.1) )**.** *Assume that Assumption 2 holds. Let $x, y \in \mathbb{R}^d$, and set $s := y - x$. If $L_1 \|s\| \leq 1$, then*

$$f(y) \leq f(x) + \langle \nabla f(x), s \rangle + 2\left(L_0 + L_1 \|\nabla f(x)\|\right)\|s\|^2 .$$

**Lemma 2** ((Nguyen et al., 2023, Lemma A.1))**.** *Let $(X_t)_{t\geq 1}$ be a martingale difference sequence with respect to a filtration $(\mathcal{G}_t)_{t\geq 0}$. Assume that there exists a constant $c$ such that $|X_t| < c$ almost surely for all $t \geq 1$, and define*

$$\sigma_t^2 = \mathbb{E}\left[X_t^2 \mid \mathcal{G}_{t-1}\right].$$

*Then, for all $b > 0$, $F > 0$, and $T \geq 1$,*

$$\Pr\left[\left|\sum_{t=1}^{T} X_t\right| > b \text{ and } \sum_{t=1}^{T} \sigma_t^2 < F\right] \leq 2\exp\left(-\frac{b^2}{2F + 2cb/3}\right).$$

**Corollary 1.** *Let $(X_t)_{t\geq 1}$ be a martingale difference sequence with respect to a filtration $(\mathcal{G}_t)_{t\geq 0}$ such that*

$$|X_t| \leq c \quad \text{almost surely for all } t \geq 1.$$

*Define $V_T := \sum_{t=1}^{T} \mathbb{E}\left[X_t^2 \mid \mathcal{G}_{t-1}\right]$. Assume that, for some deterministic $F > 0$, $V_T \leq F$ almost surely. Then, for every $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\left|\sum_{t=1}^{T} X_t\right| \leq \sqrt{2F \log\frac{2}{\delta}} + \frac{2c}{3}\log\frac{2}{\delta}.$$

*Proof of Corollary 1.* Let $\varepsilon > 0$. Since $|X_t| \leq c < c + \varepsilon$ and $V_T \leq F < F + \varepsilon$, we may apply Lemma 2 with $c + \varepsilon$ and $F + \varepsilon$. Let $b_\varepsilon := \sqrt{2(F+\varepsilon)\log\frac{2}{\delta}} + \frac{2(c+\varepsilon)}{3}\log\frac{2}{\delta}$. Then $b_\varepsilon^2 \geq \log\frac{2}{\delta}\left(2(F+\varepsilon) + \frac{2(c+\varepsilon)b_\varepsilon}{3}\right)$. Therefore,

$$2\exp\left(-\frac{b_\varepsilon^2}{2(F+\varepsilon) + 2(c+\varepsilon)b_\varepsilon/3}\right) \leq 2e^{-\log\frac{2}{\delta}} = \delta.$$

Since $V_T < F + \varepsilon$ almost surely, Lemma 2 gives

$$\Pr\left[\left|\sum_{t=1}^{T} X_t\right| > b_\varepsilon\right] \leq \delta.$$

Letting $\varepsilon \downarrow 0$ gives the result. □

The following proposition is a direct consequence of the random-variance Freedman inequality of Beygelzimer et al. (2011, Theorem 1). It provides an absorbable bound when the predictable quadratic variation is controlled by the same nonnegative random quantity that appears in the descent inequality.

**Proposition 2.** *Let $(X_t)_{t\geq 1}$ be a martingale-difference sequence with respect to $(\mathcal{G}_t)_{t\geq 0}$ such that $X_t \leq c$ almost surely. Suppose that, for some nonnegative random variable $S$ and deterministic numbers $a, b \geq 0$,*

$$\sum_{t=1}^{T} \mathbb{E}[X_t^2 \mid \mathcal{G}_{t-1}] \leq aS + b \quad \text{almost surely}.$$

*Then, for every $\eta > 0$ and $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\sum_{t=1}^{T} X_t \leq \eta S + \frac{3a}{4\eta}\log\frac{1}{\delta} + \sqrt{3b\log\frac{1}{\delta}} + c\log\frac{1}{\delta}.$$

*Proof of Proposition 2.* Define $V_T := \sum_{t=1}^{T} \mathbb{E}[X_t^2 \mid \mathcal{G}_{t-1}]$. By Beygelzimer et al. (2011, Theorem 1), for every deterministic $V'$ satisfying $V' \geq \frac{c^2}{e-2}\log\frac{1}{\delta}$, with probability at least $1-\delta$,

$$\sum_{t=1}^{T} X_t \leq \sqrt{(e-2)\log\frac{1}{\delta}}\left(\frac{V_T}{\sqrt{V'}} + \sqrt{V'}\right).$$

We choose $V'$ such that

$$\sqrt{V'} := \frac{a}{\eta}\sqrt{(e-2)\log\frac{1}{\delta}} + \sqrt{b} + \frac{c}{\sqrt{e-2}}\sqrt{\log\frac{1}{\delta}}.$$

This choice satisfies the required lower bound on $V'$. Using $V_T \le aS + b$, we obtain

$$\begin{aligned}\sum_{t=1}^{T} X_t &\le \sqrt{(e-2)\log\frac{1}{\delta}}\left(\frac{aS+b}{\sqrt{V'}} + \sqrt{V'}\right)\\ &\le \eta S + \frac{(e-2)a}{\eta}\log\frac{1}{\delta} + 2\sqrt{(e-2)b\log\frac{1}{\delta}} + c\log\frac{1}{\delta}.\end{aligned}$$

Finally, since $e-2 \le \frac{3}{4}$ and $2\sqrt{e-2} \le \sqrt{3}$, we conclude that

$$\sum_{t=1}^{T} X_t \le \eta S + \frac{3a}{4\eta}\log\frac{1}{\delta} + \sqrt{3b\log\frac{1}{\delta}} + c\log\frac{1}{\delta}.$$

□

**Lemma 3.** *Assume that Assumption 2 holds. Then, for every $x \in \mathbb{R}^d$,*

$$\|\nabla f(x)\| \le \frac{1}{\nu}\left(L_1\big(f(x)-f^\star\big) + \sqrt{L_1^2\big(f(x)-f^\star\big)^2 + 2\nu L_0\big(f(x)-f^\star\big)}\right).$$

*In particular, if $f(x) - f^\star \le 2\Delta_0$, then*

$$\|\nabla f(x)\| \le \frac{2}{\nu}\left(L_1\Delta_0 + \sqrt{L_1^2\Delta_0^2 + \nu L_0\Delta_0}\right).$$

*Proof of Lemma 3.* By Proposition 1,

$$\nu\,\|\nabla f(x)\|^2 - 2L_1\big(f(x)-f^\star\big)\,\|\nabla f(x)\| - 2L_0\big(f(x)-f^\star\big) \le 0.$$

Since $\nu > 0$, this quadratic inequality implies that $\|\nabla f(x)\|$ is bounded above by its larger root. Therefore,

$$\|\nabla f(x)\| \le \frac{1}{\nu}\left(L_1\big(f(x)-f^\star\big) + \sqrt{L_1^2\big(f(x)-f^\star\big)^2 + 2\nu L_0\big(f(x)-f^\star\big)}\right).$$

If $f(x) - f^\star \le 2\Delta_0$, we obtain

$$\begin{aligned}\|\nabla f(x)\| &\le \frac{1}{\nu}\left(2L_1\Delta_0 + \sqrt{4L_1^2\Delta_0^2 + 4\nu L_0\Delta_0}\right)\\ &= \frac{2}{\nu}\left(L_1\Delta_0 + \sqrt{L_1^2\Delta_0^2 + \nu L_0\Delta_0}\right).\end{aligned}$$

□

Table 2: Auxiliary notation used in the proofs.

| Symbol | Definition | Description |
|---|---|---|
| $\mathcal{F}_k$ | $\sigma(\xi_0, \dots, \xi_{k-1})$ | History before the $k$-th oracle call |
| $G_k$ | $\nabla f(x_k, \xi_k)$ | Stochastic gradient |
| $q_k$ | $\texttt{clip}(\nabla f(x_k), \lambda)$ | Clipped true gradient |
| $g_k$ | $\texttt{clip}(G_k, \lambda)$ | Clipped stochastic gradient |
| $e_k$ | $g_k - q_k$ | Clipping error |

Throughout the analysis, $x_0$ is deterministic and $\xi_k$ is sampled independently of the past. Consequently, $x_k$ and $q_k$ are $\mathcal{F}_k$-measurable. Moreover, Assumption 3, applied conditionally on $\mathcal{F}_k$ at $x_k$, gives

$$\mathbb{E}[G_k \mid \mathcal{F}_k] = \nabla f(x_k), \qquad \mathbb{E}\left[\|G_k - \nabla f(x_k)\|^\alpha \mid \mathcal{F}_k\right] \le \sigma^\alpha.$$

# C High-Probability Guarantees for Convex Objectives

## C.1 Best-Iterate Convergence

Table 3: Clipping level and stepsize for Theorem 1.

| Parameter | Choice |
|---|---|
| $\lambda$ | $\max\left\{\frac{L_0 R_0}{\max\{L_1 R_0, \log(4/\delta)\}},\ \sigma\left(\frac{K}{\max\{L_1 R_0, \log(4/\delta)\}}\right)^{1/\alpha}\right\}$ |
| $\gamma$ | $\frac{\nu R_0}{16\lambda \max\{L_1 R_0, \log(4/\delta)\}}$ |

**Lemma 4.** *Suppose that Assumptions 1 and 2 hold. For every $x \in \mathbb{R}^d$ and every $\lambda > 0$, define $q(x,\lambda) := \texttt{clip}(\nabla f(x), \lambda)$. Then*

$$\|q(x,\lambda)\|^2 \le \frac{2\,(L_0 + L_1\lambda)}{\nu} \langle q(x,\lambda), x - x^\star\rangle .$$

*Moreover, if $\|\nabla f(x)\| \ge \lambda$, then*

$$\langle q(x,\lambda), x - x^\star\rangle \ge \frac{\nu\lambda^2}{2\,(L_0 + L_1\lambda)}.$$

**Lemma 5.** *Suppose that Assumptions 1 and 2 hold. Let $q_k := \texttt{clip}(\nabla f(x_k), \lambda)$, where $\lambda > 0$. Then, for every $k$,*

$$\min\left\{f(x_k) - f^\star, \frac{\nu\lambda^2}{2\,(L_0 + L_1\lambda)}\right\} \le \langle q_k, x_k - x^\star\rangle .$$

**Lemma 6.** *Suppose that Assumptions 1 and 2 hold. Assume that $\gamma \le \frac{\nu}{8(L_0 + L_1\lambda)}$. Then, for every $k$,*

$$\begin{aligned}\frac{3\gamma}{2}\min\left\{f(x_k) - f^\star, \frac{\nu\lambda^2}{2(L_0 + L_1\lambda)}\right\} &\le \frac{3\gamma}{2}\langle q_k, x_k - x^\star\rangle \\ &\le \|x_k - x^\star\|^2 - \|x_{k+1} - x^\star\|^2 \\ &\quad - 2\gamma\langle e_k - \mathbb{E}[e_k \mid \mathcal{F}_k], x_k - x^\star\rangle \\ &\quad - 2\gamma\langle \mathbb{E}[e_k \mid \mathcal{F}_k], x_k - x^\star\rangle + 2\gamma^2\|e_k\|^2 .\end{aligned}$$

For $k \ge 0$, let $r_k := \|x_k - x^\star\|$, and define the stopping time

$$\tau := \inf\{k \ge 0 :\ r_k > 2R_0\},$$

with the convention $\inf\emptyset = +\infty$. On the event $\{k < \tau\}$, we have $r_k \le 2R_0$.

**Corollary 2.** *Suppose that Assumptions 1 and 2 hold. Assume that $\gamma \le \frac{\nu}{8(L_0 + L_1\lambda)}$. Then, for every $k$,*

$$\begin{aligned}\frac{3\gamma}{2}\mathbf{1}_{\{k<\tau\}}\min\left\{f(x_k) - f^\star, \frac{\nu\lambda^2}{2(L_0 + L_1\lambda)}\right\} &\le \frac{3\gamma}{2}\mathbf{1}_{\{k<\tau\}}\langle q_k, x_k - x^\star\rangle \\ &\le r_{k\wedge\tau}^2 - r_{(k+1)\wedge\tau}^2 \\ &\quad - 2\gamma\mathbf{1}_{\{k<\tau\}}\langle e_k - \mathbb{E}[e_k \mid \mathcal{F}_k], x_k - x^\star\rangle \\ &\quad - 2\gamma\mathbf{1}_{\{k<\tau\}}\langle \mathbb{E}[e_k \mid \mathcal{F}_k], x_k - x^\star\rangle \\ &\quad + 2\gamma^2\|e_k\|^2 .\end{aligned}$$

*Proof.* Multiplying Lemma 6 by $\mathbf{1}_{\{k<\tau\}}$, we use $\mathbf{1}_{\{k<\tau\}}\left(r_k^2 - r_{k+1}^2\right) = r_{k\wedge\tau}^2 - r_{(k+1)\wedge\tau}^2$ and $2\gamma^2\mathbf{1}_{\{k<\tau\}}\|e_k\|^2 \le 2\gamma^2\|e_k\|^2$. □

**Lemma 7.** *Under Assumption 3, for every $K \geq 1$ and every $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\sum_{k=0}^{K-1} \|e_k\|^2 \leq 2^{3-\alpha} K \sigma^\alpha \lambda^{2-\alpha} + 5\lambda^2 \log \frac{2}{\delta}.$$

*Consequently, with probability at least $1-\delta$,*

$$2\gamma^2 \sum_{k=0}^{K-1} \|e_k\|^2 \leq 2^{4-\alpha} K \gamma^2 \sigma^\alpha \lambda^{2-\alpha} + 10\gamma^2 \lambda^2 \log(2/\delta).$$

**Lemma 8.** *Under Assumption 3, for every $K \geq 1$ and every $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$-2\gamma \sum_{k=0}^{K-1} \mathbf{1}_{\{k<\tau\}} \langle e_k - \mathbb{E}[e_k \mid \mathcal{F}_k], x_k - x^\star \rangle \leq 8\gamma R_0 \sqrt{K \sigma^\alpha \lambda^{2-\alpha} \log \frac{2}{\delta}} + \frac{32}{3} \gamma R_0 \lambda \log \frac{2}{\delta}.$$

**Lemma 9.** *Let $x \in \mathbb{R}^d$ and define $G := \nabla f(x,\xi)$, $q := \texttt{clip}(\nabla f(x), \lambda)$, $g := \texttt{clip}(G, \lambda)$, and $b := \mathbb{E}[g] - q$. Suppose that Assumption 3 holds. Then, for every $d \in \mathbb{R}^d$,*

$$|\langle b, d \rangle| \leq \frac{\sigma}{\lambda} |\langle q, d \rangle| + \|d\| \frac{\sigma^\alpha}{\lambda^{\alpha-1}}.$$

**Remark 1.** *If we assume that both Assumptions 1 and 2 hold, then $\langle \nabla f(x), x - x^\star \rangle \geq f(x) - f^\star \geq 0$. Hence, Lemma 9 applies with $d = x - x^\star$, and yields*

$$-\langle b, x - x^\star \rangle \leq \frac{\sigma}{\lambda} \langle q, x - x^\star \rangle + \|x - x^\star\| \frac{\sigma^\alpha}{\lambda^{\alpha-1}}.$$

**Lemma 10.** *Suppose that Assumptions 1 and 3 hold. Assume $\lambda \geq 8\sigma$. Then, for every $k$,*

$$-2\gamma \mathbf{1}_{\{k<\tau\}} \langle \mathbb{E}[e_k \mid \mathcal{F}_k], x_k - x^\star \rangle \leq 4\gamma R_0 \frac{\sigma^\alpha}{\lambda^{\alpha-1}} + \frac{\gamma}{4} \mathbf{1}_{\{k<\tau\}} \langle q_k, x_k - x^\star \rangle.$$

**Proposition 3.** *Suppose that Assumptions 1, 2, and 3 hold. Assume that $\lambda \geq 8\sigma$ and $\gamma \leq \frac{\nu}{8(L_0 + L_1\lambda)}$. Then, for every $K \geq 1$ and every $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\begin{aligned} r^2_{\min(K,\tau)} + \frac{5\gamma}{4} \sum_{k=0}^{K-1} \mathbf{1}_{\{k<\tau\}} \min\left\{ f(x_k) - f^\star, \frac{\nu\lambda^2}{2(L_0 + L_1\lambda)} \right\} &\leq R_0^2 + 4\gamma K R_0 \frac{\sigma^\alpha}{\lambda^{\alpha-1}} \\ &\quad + 8\gamma R_0 \sqrt{K \sigma^\alpha \lambda^{2-\alpha} \log \frac{4}{\delta}} \\ &\quad + \frac{32}{3} \gamma R_0 \lambda \log \frac{4}{\delta} \\ &\quad + 2^{4-\alpha} K \gamma^2 \sigma^\alpha \lambda^{2-\alpha} \\ &\quad + 10\gamma^2 \lambda^2 \log \frac{4}{\delta}. \end{aligned}$$

*Proof.* Summing Corollary 2 from $k = 0$ to $K-1$, using $r_0 = \|x_0 - x^\star\| \leq R_0$, and applying Lemma 10, the term $\frac{\gamma}{4} \mathbf{1}_{\{k<\tau\}} \langle q_k, x_k - x^\star \rangle$ is absorbed into the $\frac{3\gamma}{2} \mathbf{1}_{\{k<\tau\}} \langle q_k, x_k - x^\star \rangle$ progress term, leaving the coefficient $\frac{5\gamma}{4}$. Applying Lemmas 7 and 8 with failure probability $\delta/2$ each, the claimed bound follows by a union bound. □

**Proposition 4.** *Suppose that Assumptions 1, 2, and 3 hold. Let $\delta \in (0,1)$. Assume that $\lambda \geq 8\sigma$, that $\gamma \leq \frac{\nu}{8(L_0 + L_1\lambda)}$, and that $\frac{\nu\lambda^2}{2(L_0 + L_1\lambda)} \geq \epsilon$. Let*

$$\begin{aligned} \mathcal{A}_K := 4\gamma K R_0 \frac{\sigma^\alpha}{\lambda^{\alpha-1}} &+ 8\gamma R_0 \sqrt{K \sigma^\alpha \lambda^{2-\alpha} \log \frac{4}{\delta}} \\ &+ \frac{32}{3} \gamma R_0 \lambda \log \frac{4}{\delta} + 2^{4-\alpha} K \gamma^2 \sigma^\alpha \lambda^{2-\alpha} + 10\gamma^2 \lambda^2 \log \frac{4}{\delta}. \end{aligned}$$

*Suppose moreover that $\mathcal{A}_K \leq 2R_0^2$ and $K > \frac{12 R_0^2}{5\gamma\epsilon}$. Then, with probability at least $1-\delta$, we have $\tau > K$ and*

$$\min_{0 \leq k < K} \big( f(x_k) - f^\star \big) \leq \epsilon.$$

**Derivation of the tuning scales.** For a fixed iteration budget $K$, we seek to choose $\lambda$ and $\gamma$ so that the accuracy bound $\epsilon$ certified by Proposition 4 is as small as possible. The three conditions used to derive the tuning envelope are $\gamma \le \frac{\nu}{8(L_0+L_1\lambda)}$, $A_K \le 2R_0^2$, and $K > \frac{12R_0^2}{5\gamma\epsilon}$. The last condition is equivalent to

$$\epsilon > \frac{12R_0^2}{5K\gamma}. \tag{4}$$

For a given $\lambda$, this lower bound decreases as $\gamma$ increases and is therefore minimized by taking $\gamma$ as large as the first two conditions allow. We consequently derive upper bounds on $\gamma$ from the stepsize and localization conditions, substitute them into the preceding inequality, and then minimize the resulting bound over $\lambda$.

The proposition also imposes the feasibility conditions $\lambda \ge 8\sigma$ and $\epsilon \le \frac{\nu\lambda^2}{2(L_0+L_1\lambda)}$. These conditions do not provide upper bounds on $\gamma$ and therefore do not enter the preceding elimination of $\gamma$. We first obtain candidate values of $\lambda$, $\gamma$, and $\epsilon$ from the three tuning conditions above. In the subsequent theorem, we verify that these candidates satisfy both feasibility conditions under the stated lower bound on $K$.

We now derive the resulting lower bounds on $\epsilon$.

If $R_0 = 0$, the right-hand side of every lower bound derived below is zero, so each bound follows from $\epsilon > 0$. We may therefore assume that $R_0 > 0$.

Combining Eq. (4) with $\gamma \le \frac{\nu}{8(L_0+L_1\lambda)}$ yields

$$\epsilon > \frac{R_0}{K}\frac{96}{5\nu}\left(L_0R_0 + \lambda L_1R_0\right).$$

Because all the terms defining $A_K$ are nonnegative, its first term is bounded by $A_K$. Therefore, $4\gamma K R_0 \frac{\sigma^\alpha}{\lambda^{\alpha-1}} \le A_K \le 2R_0^2$. For $\sigma > 0$, this gives $\gamma \le \frac{R_0\lambda^{\alpha-1}}{2K\sigma^\alpha}$, and hence

$$\epsilon > \frac{R_0}{K}\frac{24K\sigma^\alpha}{5\lambda^{\alpha-1}}.$$

When $\sigma = 0$, the right-hand side of the preceding inequality vanishes, so the same conclusion follows directly from $\epsilon > 0$. Thus, the inequality holds in both cases.

Similarly, the third term in $A_K$ satisfies $\frac{32}{3}\gamma R_0\lambda\log\frac{4}{\delta} \le A_K \le 2R_0^2$. Therefore, $\gamma \le \frac{3R_0}{16\lambda\log\frac{4}{\delta}}$, and Eq. (4) yields

$$\epsilon > \frac{R_0}{K}\frac{64}{5}\lambda\log\frac{4}{\delta}.$$

Combining the preceding three necessary inequalities, we obtain

$$\epsilon > \frac{R_0}{K}\max\left\{\frac{96}{5\nu}\left(L_0R_0 + \lambda L_1R_0\right), \frac{64}{5}\lambda\log\frac{4}{\delta}, \frac{24K\sigma^\alpha}{5\lambda^{\alpha-1}}\right\}.$$

Similarly, requiring the second, fourth, and fifth terms of $A_K$ to be bounded by $2R_0^2$ gives additional upper bounds on $\gamma$. After substituting these bounds into Eq. (4), the resulting lower bounds on $\epsilon$ are, up to numerical factors, dominated by

$$\frac{R_0}{K}\max\left\{\frac{K\sigma^\alpha}{\lambda^{\alpha-1}}, \lambda\log\frac{4}{\delta}\right\}.$$

Indeed, $\sqrt{K\sigma^\alpha\lambda^{2-\alpha}\log\frac{4}{\delta}} = \sqrt{\frac{K\sigma^\alpha}{\lambda^{\alpha-1}}\cdot\lambda\log\frac{4}{\delta}} \le \max\left\{\frac{K\sigma^\alpha}{\lambda^{\alpha-1}}, \lambda\log\frac{4}{\delta}\right\}$. Moreover, since $\alpha \in (1,2]$ and $\log(4/\delta) > 1$, $\sqrt{2^{3-\alpha}K\sigma^\alpha\lambda^{2-\alpha}} \le 2\max\left\{\frac{K\sigma^\alpha}{\lambda^{\alpha-1}}, \lambda\log\frac{4}{\delta}\right\}$, and $\lambda\sqrt{\log\frac{4}{\delta}} \le \lambda\log\frac{4}{\delta}$. Thus, these terms can change only the universal numerical constants and introduce no additional tuning scale.

**Lemma 11.** *Let $\delta \in (0,1)$ and $K \ge 1$. Suppose that $\lambda, \gamma, \epsilon > 0$ satisfy $\gamma \le \frac{\nu}{8(L_0+L_1\lambda)}$, $A_K \le 2R_0^2$ and $K > \frac{12R_0^2}{5\gamma\epsilon}$, where $A_K$ is defined in Proposition 4. Then*

$$\epsilon > \frac{R_0}{K}\max\left\{\frac{96}{5\nu}\left(L_0R_0 + \lambda L_1R_0\right), \frac{64}{5}\lambda\log\frac{4}{\delta}, \frac{24K\sigma^\alpha}{5\lambda^{\alpha-1}}\right\}.$$

*In particular, since $\nu \in (0, 1]$, the preceding inequality implies the simpler tuning bound*

$$\epsilon > \frac{R_0}{K} \max\left\{ L_0 R_0,\ \lambda \max\left\{ L_1 R_0, \log\frac{4}{\delta} \right\},\ \frac{K\sigma^\alpha}{\lambda^{\alpha-1}} \right\}.$$

*Moreover,*

$$\begin{aligned}\inf_{\eta>0} \max\left\{ L_0 R_0,\ \eta \max\left\{ L_1 R_0, \log\frac{4}{\delta} \right\},\ \frac{K\sigma^\alpha}{\eta^{\alpha-1}} \right\} \\ = \max\left\{ L_0 R_0,\ \sigma K^{\frac{1}{\alpha}} \max\left\{ L_1 R_0, \log\frac{4}{\delta} \right\}^{\frac{\alpha-1}{\alpha}} \right\}.\end{aligned}$$

*Provided that $L_0 R_0 + \sigma > 0$, one minimizer is*

$$\lambda^\star = \max\left\{ \frac{L_0 R_0}{\max\left\{ L_1 R_0, \log\frac{4}{\delta} \right\}},\ \sigma \left( \frac{K}{\max\left\{ L_1 R_0, \log\frac{4}{\delta} \right\}} \right)^{\frac{1}{\alpha}} \right\}.$$

**Choice of parameters.** For brevity, let $M := \max\left\{ L_1 R_0, \log\frac{4}{\delta} \right\}$. In the nontrivial case $L_0 R_0 + \sigma > 0$, the minimizer identified by Lemma 11 gives

$$\lambda = \max\left\{ \frac{L_0 R_0}{M}, \sigma \left( \frac{K}{M} \right)^{\frac{1}{\alpha}} \right\}.$$

This choice of $\lambda$ also determines the relevant stepsize scale. Indeed, consider the two possible regimes in its definition.

If $\lambda = \frac{L_0 R_0}{M}$, then the generalized-smoothness restriction gives

$$\gamma \le \frac{\nu}{8(L_0 + L_1\lambda)} \le \frac{\nu}{8L_0} = \frac{\nu R_0}{8\lambda M}.$$

If instead $\lambda = \sigma\left(\frac{K}{M}\right)^{\frac{1}{\alpha}}$, then the first term of the localization condition $A_K \le 2R_0^2$ gives $4\gamma K R_0 \frac{\sigma^\alpha}{\lambda^{\alpha-1}} \le 2R_0^2$. Consequently,

$$\gamma \le \frac{R_0 \lambda^{\alpha-1}}{2K\sigma^\alpha} = \frac{R_0}{2\lambda M},$$

where the equality follows from $\lambda^\alpha = K\frac{\sigma^\alpha}{M}$. Thus, in either regime, the conditions of Proposition 4 restrict the stepsize to the scale $\frac{R_0}{\lambda M}$.

Since Eq. (4) favors taking $\gamma$ as large as the remaining conditions permit, we choose a uniform stepsize at this scale. More precisely, we write

$$\gamma = \frac{c\nu R_0}{\lambda M},$$

where $c > 0$ is a constant. By the definitions of $\lambda$ and $M$, $L_0 \le \frac{\lambda M}{R_0}$ and $L_1\lambda \le \frac{\lambda M}{R_0}$. Therefore, $L_0 + L_1\lambda \le \frac{2\lambda M}{R_0}$. It follows that

$$\gamma \le \frac{\nu}{8(L_0 + L_1\lambda)}$$

is guaranteed whenever $c \le \frac{1}{16}$.

We next verify that the same choice also controls the localization budget. By construction, $\frac{K\sigma^\alpha}{\lambda^\alpha} \le M$, and the definition of $M$ gives $\log\frac{4}{\delta} \le M$. Using these inequalities, together with $M > 1$, in the five terms defining $A_K$, respectively, yields

$$A_K \le \left( 4c\nu + 8c\nu + \frac{32}{3}c\nu + 8c^2\nu^2 + 10c^2\nu^2 \right) R_0^2 = \left( \frac{68}{3}c\nu + 18c^2\nu^2 \right) R_0^2.$$

We therefore choose $c = \frac{1}{16}$, that is, $\gamma = \frac{\nu R_0}{16\lambda M}$. This choice satisfies the generalized-smoothness restriction. Moreover, since $\nu < \frac{3}{5}$,

$$\begin{aligned} A_K &\le \left(\frac{17\nu}{12} + \frac{9\nu^2}{128}\right) R_0^2 \\ &< \left(\frac{17}{20} + \frac{81}{3200}\right) R_0^2 \\ &= \frac{2801}{3200} R_0^2 \\ &< R_0^2. \end{aligned}$$

It remains to verify the two feasibility conditions of Proposition 4 that were not used to derive the tuning scale, namely $\lambda \ge 8\sigma$ and $\epsilon \le \frac{\nu\lambda^2}{2(L_0+L_1\lambda)}$. For the chosen stepsize, Eq. (4) requires $\epsilon > \frac{192 R_0 \lambda M}{5\nu K}$. We therefore set

$$\epsilon := \frac{40 R_0 \lambda M}{\nu K}.$$

Since $L_0 + L_1\lambda \le \frac{2\lambda M}{R_0}$, the second feasibility condition is satisfied whenever $K \ge \frac{160}{\nu^2} M^2$.

The same lower bound also implies $\frac{K}{M} \ge 160 > 8^\alpha$, and hence $\lambda \ge \sigma \left(\frac{K}{M}\right)^{\frac{1}{\alpha}} \ge 8\sigma$. Thus, all the conditions of Proposition 4 are satisfied.

### C.1.1 Proofs

*Proof of Lemma 4.* **1)** We distinguish two cases.

- If $\|\nabla f(x)\| \le \lambda$, then $q(x,\lambda) = \nabla f(x)$. By Proposition 1 and convexity,

$$\begin{aligned} \|q(x,\lambda)\|^2 &= \|\nabla f(x)\|^2 \\ &\le \frac{2}{\nu} \left(L_0 + L_1 \|\nabla f(x)\|\right) (f(x) - f^\star) \\ &\le \frac{2(L_0 + L_1\lambda)}{\nu} (f(x) - f^\star) \\ &\le \frac{2(L_0 + L_1\lambda)}{\nu} \langle q(x,\lambda), x - x^\star \rangle . \end{aligned}$$

- If $\|\nabla f(x)\| \ge \lambda$, then

$$q(x,\lambda) = \lambda \frac{\nabla f(x)}{\|\nabla f(x)\|}.$$

Hence, by convexity and Proposition 1,

$$\begin{aligned} \langle q(x,\lambda), x - x^\star \rangle &= \lambda \frac{\langle \nabla f(x), x - x^\star \rangle}{\|\nabla f(x)\|} \\ &\ge \lambda \frac{f(x) - f^\star}{\|\nabla f(x)\|} \\ &\ge \frac{\nu\lambda \|\nabla f(x)\|}{2\left(L_0 + L_1 \|\nabla f(x)\|\right)}. \end{aligned}$$

Since $t \mapsto \frac{t}{L_0 + L_1 t}$ is nondecreasing on $(0,\infty)$ and $\|\nabla f(x)\| \ge \lambda$, we obtain

$$\langle q(x,\lambda), x - x^\star \rangle \ge \frac{\nu\lambda^2}{2(L_0 + L_1\lambda)}.$$

Since $\|q(x,\lambda)\| = \lambda$, it follows that

$$\|q(x,\lambda)\|^2 \le \frac{2(L_0 + L_1\lambda)}{\nu} \langle q(x,\lambda), x - x^\star \rangle .$$

**2)** It is already proven in the first part. □

*Proof of Lemma 5.* We distinguish two cases.

- If $\|\nabla f(x_k)\| \leq \lambda$, then $q_k = \nabla f(x_k)$. Hence, by convexity,

$$\begin{aligned}\min\left\{f(x_k) - f^\star, \frac{\nu\lambda^2}{2(L_0 + L_1\lambda)}\right\} &\leq f(x_k) - f^\star \\ &\leq \langle \nabla f(x_k), x_k - x^\star\rangle \\ &= \langle q_k, x_k - x^\star\rangle .\end{aligned}$$

- If $\|\nabla f(x_k)\| > \lambda$, then, by Lemma 4,

$$\begin{aligned}\min\left\{f(x_k) - f^\star, \frac{\nu\lambda^2}{2(L_0 + L_1\lambda)}\right\} &\leq \frac{\nu\lambda^2}{2(L_0 + L_1\lambda)} \\ &\leq \langle q_k, x_k - x^\star\rangle .\end{aligned}$$

□

*Proof of Lemma 6.* We have $x_{k+1} = x_k - \gamma g_k = x_k - \gamma(q_k + e_k)$. Therefore,

$$\begin{aligned}\|x_{k+1} - x^\star\|^2 &= \|x_k - x^\star - \gamma(q_k + e_k)\|^2 \\ &= \|x_k - x^\star\|^2 - 2\gamma\langle q_k, x_k - x^\star\rangle - 2\gamma\langle e_k, x_k - x^\star\rangle + \gamma^2\|q_k + e_k\|^2 .\end{aligned}$$

Using $\|q_k + e_k\|^2 \leq 2\|q_k\|^2 + 2\|e_k\|^2$, we obtain

$$\begin{aligned}\|x_{k+1} - x^\star\|^2 \leq \|x_k - x^\star\|^2 &- 2\gamma\langle q_k, x_k - x^\star\rangle - 2\gamma\langle e_k, x_k - x^\star\rangle \\ &+ 2\gamma^2\|q_k\|^2 + 2\gamma^2\|e_k\|^2 .\end{aligned}$$

By Lemma 4, $\|q_k\|^2 \leq \frac{2(L_0+L_1\lambda)}{\nu}\langle q_k, x_k - x^\star\rangle$. Hence,

$$2\gamma^2\|q_k\|^2 \leq \frac{4\gamma^2(L_0 + L_1\lambda)}{\nu}\langle q_k, x_k - x^\star\rangle .$$

Since $\gamma \leq \frac{\nu}{8(L_0+L_1\lambda)}$, we have $\frac{4\gamma^2(L_0+L_1\lambda)}{\nu} \leq \frac{\gamma}{2}$. Therefore,

$$\begin{aligned}\|x_{k+1} - x^\star\|^2 \leq \|x_k - x^\star\|^2 &- \frac{3\gamma}{2}\langle q_k, x_k - x^\star\rangle \\ &- 2\gamma\langle e_k, x_k - x^\star\rangle + 2\gamma^2\|e_k\|^2 .\end{aligned}$$

Rearranging gives

$$\begin{aligned}\frac{3\gamma}{2}\langle q_k, x_k - x^\star\rangle \leq \|x_k - x^\star\|^2 &- \|x_{k+1} - x^\star\|^2 \\ &- 2\gamma\langle e_k, x_k - x^\star\rangle + 2\gamma^2\|e_k\|^2 .\end{aligned}$$

Decomposing $e_k = e_k - \mathbb{E}[e_k \mid \mathcal{F}_k] + \mathbb{E}[e_k \mid \mathcal{F}_k]$, we obtain

$$\begin{aligned}\frac{3\gamma}{2}\langle q_k, x_k - x^\star\rangle \leq \|x_k - x^\star\|^2 &- \|x_{k+1} - x^\star\|^2 \\ &- 2\gamma\langle e_k - \mathbb{E}[e_k \mid \mathcal{F}_k], x_k - x^\star\rangle \\ &- 2\gamma\langle \mathbb{E}[e_k \mid \mathcal{F}_k], x_k - x^\star\rangle + 2\gamma^2\|e_k\|^2 .\end{aligned}$$

Finally, by Lemma 5,

$$\min\left\{f(x_k) - f^\star, \frac{\nu\lambda^2}{2(L_0 + L_1\lambda)}\right\} \leq \langle q_k, x_k - x^\star\rangle .$$

Therefore,

$$\begin{aligned}\frac{3\gamma}{2}\min\left\{f(x_k)-f^\star,\ \frac{\nu\lambda^2}{2(L_0+L_1\lambda)}\right\} &\le \frac{3\gamma}{2}\langle q_k, x_k-x^\star\rangle \\ &\le \|x_k-x^\star\|^2-\|x_{k+1}-x^\star\|^2 \\ &\quad -2\gamma\langle e_k-\mathbb{E}[e_k\mid\mathcal{F}_k], x_k-x^\star\rangle \\ &\quad -2\gamma\langle\mathbb{E}[e_k\mid\mathcal{F}_k], x_k-x^\star\rangle+2\gamma^2\|e_k\|^2,\end{aligned}$$

□

*Proof of Lemma 7.* Set $X_k := \|e_k\|^2$. Since the clipping map is nonexpansive,

$$\|e_k\| \le \|G_k-\nabla f(x_k)\|.$$

Moreover, both clipped vectors have norm at most $\lambda$, hence $\|e_k\|\le 2\lambda$. Therefore,

$$\begin{aligned}X_k=\|e_k\|^2=\|e_k\|^{2-\alpha}\|e_k\|^\alpha &\le (2\lambda)^{2-\alpha}\|G_k-\nabla f(x_k)\|^\alpha.\end{aligned}$$

Using Assumption 3, we obtain

$$\begin{aligned}\mathbb{E}[X_k\mid\mathcal{F}_k] &\le (2\lambda)^{2-\alpha}\mathbb{E}[\|G_k-\nabla f(x_k)\|^\alpha\mid\mathcal{F}_k] \\ &\le 2^{2-\alpha}\sigma^\alpha\lambda^{2-\alpha}.\end{aligned}$$

Thus

$$\sum_{k=0}^{K-1}\mathbb{E}[X_k\mid\mathcal{F}_k]\le 2^{2-\alpha}\sigma^\alpha\lambda^{2-\alpha}K.$$

We now use the decomposition

$$\sum_{k=0}^{K-1}X_k=\sum_{k=0}^{K-1}\mathbb{E}[X_k\mid\mathcal{F}_k]+\sum_{k=0}^{K-1}\Big(X_k-\mathbb{E}[X_k\mid\mathcal{F}_k]\Big).$$

Let $Z_k := X_k-\mathbb{E}[X_k\mid\mathcal{F}_k]$. Then $Z_k$ is $\mathcal{F}_{k+1}$-measurable and $\mathbb{E}[Z_k\mid\mathcal{F}_k]=0$. Thus, after reindexing $Y_{k+1}:=Z_k$, $(Y_k)_{k\ge1}$ is a martingale difference sequence with respect to $(\mathcal{F}_k)_{k\ge0}$. Also,

$$|Z_k|\le 4\lambda^2.$$

Indeed, $0\le X_k\le 4\lambda^2$ and $0\le\mathbb{E}[X_k\mid\mathcal{F}_k]\le 4\lambda^2$. We also have

$$\begin{aligned}\mathbb{E}[Z_k^2\mid\mathcal{F}_k] &\le \mathbb{E}[X_k^2\mid\mathcal{F}_k] \\ &\le 4\lambda^2\,\mathbb{E}[X_k\mid\mathcal{F}_k] \\ &\le 2^{4-\alpha}\sigma^\alpha\lambda^{4-\alpha}.\end{aligned}$$

Hence

$$\sum_{k=0}^{K-1}\mathbb{E}[Z_k^2\mid\mathcal{F}_k]\le 2^{4-\alpha}\sigma^\alpha\lambda^{4-\alpha}K.$$

By Corollary 1, with probability at least $1-\delta$,

$$\sum_{k=0}^{K-1}Z_k\le\sqrt{2\cdot 2^{4-\alpha}\sigma^\alpha\lambda^{4-\alpha}K\log\frac{2}{\delta}}+\frac{8\lambda^2}{3}\log\frac{2}{\delta}.$$

Therefore, on the same event,

$$\sum_{k=0}^{K-1}X_k\le 2^{2-\alpha}\sigma^\alpha\lambda^{2-\alpha}K+\sqrt{2\cdot 2^{4-\alpha}\sigma^\alpha\lambda^{4-\alpha}K\log\frac{2}{\delta}}+\frac{8\lambda^2}{3}\log\frac{2}{\delta}.$$

Using $2\sqrt{ab}\le a+b$, with $a:=2^{2-\alpha}\sigma^\alpha\lambda^{2-\alpha}K$ and $b:=2\lambda^2\log\frac{2}{\delta}$, we get

$$\sqrt{2\cdot 2^{4-\alpha}\sigma^\alpha\lambda^{4-\alpha}K\log\frac{2}{\delta}}\le 2^{2-\alpha}\sigma^\alpha\lambda^{2-\alpha}K+2\lambda^2\log\frac{2}{\delta}.$$

Hence

$$\sum_{k=0}^{K-1} X_k \le 2^{3-\alpha}\sigma^\alpha\lambda^{2-\alpha}K + \left(2+\frac{8}{3}\right)\lambda^2\log\frac{2}{\delta}$$
$$\le 2^{3-\alpha}\sigma^\alpha\lambda^{2-\alpha}K + 5\lambda^2\log\frac{2}{\delta}.$$

We deduce that

$$2\gamma^2\sum_{k=0}^{K-1}\|e_k\|^2 \le 2^{4-\alpha}K\gamma^2\sigma^\alpha\lambda^{2-\alpha} + 10\gamma^2\lambda^2\log\frac{2}{\delta}.$$

□

*Proof of Lemma 8.* Define

$$Z_k := -2\gamma\mathbf{1}_{\{k<\tau\}}\langle e_k - \mathbb{E}[e_k \mid \mathcal{F}_k], x_k - x^\star\rangle.$$

Then $Z_k$ is $\mathcal{F}_{k+1}$-measurable and

$$\mathbb{E}[Z_k \mid \mathcal{F}_k] = 0.$$

Thus, after reindexing $Y_{k+1} := Z_k$, $(Y_k)_{k\ge1}$ is a martingale difference sequence with respect to $(\mathcal{F}_k)_{k\ge0}$. We have

$$\| e_k - \mathbb{E}[e_k \mid \mathcal{F}_k] \|_2 \le \| e_k \|_2 + \| \mathbb{E}[e_k \mid \mathcal{F}_k] \|_2 \le 4\lambda.$$

On the event $\{k<\tau\}$, we have $\| x_k - x^\star \|_2 \le 2R_0$. Therefore,

$$|Z_k| \le 16\gamma R_0\lambda.$$

Moreover,

$$\mathbb{E}[\| e_k - \mathbb{E}[e_k \mid \mathcal{F}_k] \|_2^2 \mid \mathcal{F}_k] \le \mathbb{E}[\| e_k \|_2^2 \mid \mathcal{F}_k].$$

As in Lemma 7,

$$\mathbb{E}[\| e_k \|_2^2 \mid \mathcal{F}_k] \le 2^{2-\alpha}\sigma^\alpha\lambda^{2-\alpha}.$$

Thus

$$\mathbb{E}[Z_k^2 \mid \mathcal{F}_k] \le 4\gamma^2\mathbf{1}_{\{k<\tau\}}\| x_k - x^\star \|_2^2\,\mathbb{E}[\| e_k - \mathbb{E}[e_k \mid \mathcal{F}_k] \|_2^2 \mid \mathcal{F}_k]$$
$$\le 16\gamma^2R_0^2\mathbb{E}[\| e_k \|_2^2 \mid \mathcal{F}_k]$$
$$\le 2^{6-\alpha}\gamma^2R_0^2\sigma^\alpha\lambda^{2-\alpha}.$$

Consequently,

$$\sum_{k=0}^{K-1}\mathbb{E}[Z_k^2 \mid \mathcal{F}_k] \le F,$$

where $F := 2^{6-\alpha}\gamma^2R_0^2K\sigma^\alpha\lambda^{2-\alpha}$. By Corollary 1, with probability at least $1-\delta$,

$$\sum_{k=0}^{K-1} Z_k \le \sqrt{2F\log\frac{2}{\delta}} + \frac{2}{3}(16\gamma R_0\lambda)\log\frac{2}{\delta}.$$

Since $\alpha \in (1,2]$,

$$\sqrt{2F\log\frac{2}{\delta}} \le 8\gamma R_0\sqrt{K\sigma^\alpha\lambda^{2-\alpha}\log\frac{2}{\delta}}.$$

Therefore, with probability at least $1-\delta$,

$$-2\gamma\sum_{k=0}^{K-1}\mathbf{1}_{\{k<\tau\}}\langle e_k - \mathbb{E}[e_k \mid \mathcal{F}_k], x_k - x^\star\rangle \le 8\gamma R_0\sqrt{K\sigma^\alpha\lambda^{2-\alpha}\log\frac{2}{\delta}} + \frac{32}{3}\gamma R_0\lambda\log\frac{2}{\delta}.$$

□

*Proof of Lemma 9.* We have $\langle b, d\rangle = \mathbb{E}\left\langle \left(\min\left\{1, \frac{\lambda}{\|G\|}\right\} - \min\left\{1, \frac{\lambda}{\|\nabla f(x)\|}\right\}\right) G, d\right\rangle$. Hence,

$$\langle b, d\rangle = \mathbb{E}\left[\left(\min\left\{1, \frac{\lambda}{\|G\|}\right\} - \min\left\{1, \frac{\lambda}{\|\nabla f(x)\|}\right\}\right)\langle G, d\rangle\right].$$

This implies that

$$|\langle b, d\rangle| \leq \underbrace{\left|\mathbb{E}\left[\min\left\{1, \frac{\lambda}{\|G\|}\right\} - \min\left\{1, \frac{\lambda}{\|\nabla f(x)\|}\right\}\right]\right| |\langle \nabla f(x), d\rangle|}_{\text{alignment term}}$$
$$+ \underbrace{\mathbb{E}\left[\left|\min\left\{1, \frac{\lambda}{\|G\|}\right\} - \min\left\{1, \frac{\lambda}{\|\nabla f(x)\|}\right\}\right| |\langle G - \nabla f(x), d\rangle|\right]}_{\text{noise term}}.$$

**1) Bounding the noise term.** Let $\phi(r) := \min\{1, \frac{\lambda}{r}\}$, with $\phi(0) = 1$. On $[0, \lambda]$, $\phi$ is 0-Lipschitz. On $[\lambda, \infty)$, $\phi(r) = \frac{\lambda}{r}$ and $\forall r \geq \lambda$, $\left|\left(\phi|_{[\lambda,\infty)}\right)'(r)\right| = \frac{\lambda}{r^2} \leq \frac{1}{\lambda}$, so $\phi$ is $\frac{1}{\lambda}$-Lipschitz on $[\lambda, \infty)$. If $x \leq \lambda \leq y$, then

$$|\phi(x) - \phi(y)| \leq |\phi(x) - \phi(\lambda)| + |\phi(\lambda) - \phi(y)| \leq \frac{|x - y|}{\lambda}.$$

Hence, $\phi$ is $\frac{1}{\lambda}$-Lipschitz on $[0, \infty)$. Moreover, since $\phi$ takes values in $[0, 1]$

$$\begin{aligned}\text{noise term} &\leq \|d\|\ \mathbb{E}\left[\|G - \nabla f(x)\| \min\left\{1, \frac{\|G - \nabla f(x)\|}{\lambda}\right\}\right] \\ &\leq \frac{\|d\|}{\lambda^{\alpha-1}}\mathbb{E}\left[\|G - \nabla f(x)\|^{\alpha}\right] \\ &\leq \|d\|\,\frac{\sigma^{\alpha}}{\lambda^{\alpha-1}},\end{aligned}$$

where in the second inequality we used $r \min\{1, \frac{r}{\lambda}\} \leq \frac{r^{\alpha}}{\lambda^{\alpha-1}}$ for all $r \geq 0$.

**2) Bounding the alignment term.** We now show that $\phi$ satisfies a stronger property than $\frac{1}{\lambda}$-Lipschitzness:

$$|\phi(x) - \phi(y)| \leq \phi(x)\frac{|x - y|}{\lambda}.$$

Indeed, since $\phi(r) = \frac{\lambda}{\max\{\lambda, r\}}$,

$$\begin{aligned}|\phi(x) - \phi(y)| &= \frac{\lambda|\max\{\lambda, x\} - \max\{\lambda, y\}|}{\max\{\lambda, x\}\max\{\lambda, y\}} \\ &\leq \frac{|\max\{\lambda, x\} - \max\{\lambda, y\}|}{\max\{\lambda, x\}} \\ &\leq \frac{|x - y|}{\max\{\lambda, x\}} \\ &= \phi(x)\frac{|x - y|}{\lambda},\end{aligned}$$

where we used that $r \mapsto \max\{\lambda, r\}$ is 1-Lipschitz. Therefore,

$$\begin{aligned}\text{alignment term} &\leq \frac{\mathbb{E}[\|G - \nabla f(x)\|]}{\lambda} \min\left\{1, \frac{\lambda}{\|\nabla f(x)\|}\right\} |\langle \nabla f(x), d\rangle| \\ &= \frac{\mathbb{E}[\|G - \nabla f(x)\|]}{\lambda} |\langle q, d\rangle| \\ &\leq \frac{\sigma}{\lambda} |\langle q, d\rangle|.\end{aligned}$$

□

*Proof of Lemma 10.* Let $b_k := \mathbb{E}[e_k \mid \mathcal{F}_k]$. By Lemma 9 with $d = x_k - x^\star$, and using convexity, $\langle \nabla f(x_k), x_k - x^\star \rangle \geq 0$, we obtain

$$-\langle b_k, x_k - x^\star \rangle \leq \frac{\sigma}{\lambda} \langle q_k, x_k - x^\star \rangle + \|x_k - x^\star\| \frac{\sigma^\alpha}{\lambda^{\alpha-1}}.$$

Multiplying by $2\gamma \mathbf{1}_{\{k<\tau\}}$ gives

$$-2\gamma \mathbf{1}_{\{k<\tau\}} \langle b_k, x_k - x^\star \rangle \leq \frac{2\gamma\sigma}{\lambda} \mathbf{1}_{\{k<\tau\}} \langle q_k, x_k - x^\star \rangle + 2\gamma \mathbf{1}_{\{k<\tau\}} \|x_k - x^\star\| \frac{\sigma^\alpha}{\lambda^{\alpha-1}}.$$

By the definition of $\tau$, on $\{k < \tau\}$, $\|x_k - x^\star\| \leq 2R_0$. Therefore,

$$-2\gamma \mathbf{1}_{\{k<\tau\}} \langle b_k, x_k - x^\star \rangle \leq \frac{2\gamma\sigma}{\lambda} \mathbf{1}_{\{k<\tau\}} \langle q_k, x_k - x^\star \rangle + 4\gamma R_0 \frac{\sigma^\alpha}{\lambda^{\alpha-1}}.$$

Finally, since $\lambda \geq 8\sigma$, we have $\frac{2\sigma}{\lambda} \leq \frac{1}{4}$. Hence,

$$-2\gamma \mathbf{1}_{\{k<\tau\}} \langle \mathbb{E}[e_k \mid \mathcal{F}_k], x_k - x^\star \rangle \leq \frac{\gamma}{4} \mathbf{1}_{\{k<\tau\}} \langle q_k, x_k - x^\star \rangle + 4\gamma R_0 \frac{\sigma^\alpha}{\lambda^{\alpha-1}},$$

which proves the claim. □

*Proof of Proposition 4.* We work on the event from Proposition 3, which holds with probability at least $1 - \delta$. On this event,

$$r^2_{\min(K,\tau)} + \frac{5\gamma}{4} \sum_{k=0}^{K-1} \mathbf{1}_{\{k<\tau\}} \min\left\{ f(x_k) - f^\star, \frac{\nu\lambda^2}{2(L_0 + L_1\lambda)} \right\} \leq R_0^2 + \mathcal{A}_K.$$

Since $\mathcal{A}_K \leq 2R_0^2$, we obtain

$$r^2_{\min(K,\tau)} + \frac{5\gamma}{4} \sum_{k=0}^{K-1} \mathbf{1}_{\{k<\tau\}} \min\left\{ f(x_k) - f^\star, \frac{\nu\lambda^2}{2(L_0 + L_1\lambda)} \right\} \leq 3R_0^2.$$

In particular, $r^2_{\min(K,\tau)} \leq 3R_0^2$. If $\tau \leq K$, then by the definition of $\tau$,

$$r^2_{\min(K,\tau)} = r^2_\tau > 4R_0^2,$$

which is a contradiction. Hence, $\tau > K$. Therefore,

$$\frac{5\gamma}{4} \sum_{k=0}^{K-1} \min\left\{ f(x_k) - f^\star, \frac{\nu\lambda^2}{2(L_0 + L_1\lambda)} \right\} \leq 3R_0^2,$$

and consequently,

$$\frac{1}{K} \sum_{k=0}^{K-1} \min\left\{ f(x_k) - f^\star, \frac{\nu\lambda^2}{2(L_0 + L_1\lambda)} \right\} \leq \frac{12R_0^2}{5\gamma K}.$$

Since $K > \frac{12R_0^2}{5\gamma\epsilon}$, we have

$$\frac{1}{K} \sum_{k=0}^{K-1} \min\left\{ f(x_k) - f^\star, \frac{\nu\lambda^2}{2(L_0 + L_1\lambda)} \right\} < \epsilon.$$

Hence, there exists some $k \in \{0, \ldots, K-1\}$ such that

$$\min\left\{ f(x_k) - f^\star, \frac{\nu\lambda^2}{2(L_0 + L_1\lambda)} \right\} < \epsilon.$$

By assumption, $\frac{\nu\lambda^2}{2(L_0+L_1\lambda)} \geq \epsilon$. Therefore, necessarily $f(x_k) - f^\star < \epsilon$. Thus,

$$\min_{0 \leq k < K} \big( f(x_k) - f^\star \big) \leq \epsilon.$$

□

*Proof of Lemma 11.* The first inequality was established in the derivation preceding the lemma. Since $\nu \in (0,1]$, $\frac{96}{5\nu} > 1$, $\frac{64}{5} > 1$, $\frac{24}{5} > 1$, and $L_0R_0 + \lambda L_1 R_0 \geq \max\{L_0R_0, \lambda L_1 R_0\}$. Therefore,

$$\epsilon > \frac{R_0}{K} \max\left\{L_0R_0, \lambda \max\left\{L_1R_0, \log\frac{4}{\delta}\right\}, \frac{K\sigma^\alpha}{\lambda^{\alpha-1}}\right\}.$$

It remains to minimize the expression inside the maximum. Let

$$M := \max\left\{L_1R_0, \log\frac{4}{\delta}\right\}.$$

Since $\delta \in (0,1)$, we have $M > 0$. Define $h : (0,\infty) \to \mathbb{R}$ by

$$h(\eta) := \max\left\{L_0R_0, \eta M, \frac{K\sigma^\alpha}{\eta^{\alpha-1}}\right\}.$$

Suppose first that $\sigma > 0$, and define the balancing point $\eta_{\text{bal}} := \sigma\left(\frac{K}{M}\right)^{\frac{1}{\alpha}}$. At this point,

$$\eta_{\text{bal}} M = \frac{K\sigma^\alpha}{\eta_{\text{bal}}^{\alpha-1}} = \sigma K^{\frac{1}{\alpha}} M^{\frac{\alpha-1}{\alpha}}.$$

If $\eta \geq \eta_{\text{bal}}$, then $\max\left\{\eta M, \frac{K\sigma^\alpha}{\eta^{\alpha-1}}\right\} \geq \eta M \geq \eta_{\text{bal}} M$. If $\eta \leq \eta_{\text{bal}}$, then $\max\left\{\eta M, \frac{K\sigma^\alpha}{\eta^{\alpha-1}}\right\} \geq \frac{K\sigma^\alpha}{\eta^{\alpha-1}} \geq \frac{K\sigma^\alpha}{\eta_{\text{bal}}^{\alpha-1}}$. Consequently, for every $\eta > 0$,

$$h(\eta) \geq \max\left\{L_0R_0, \sigma K^{\frac{1}{\alpha}} M^{\frac{\alpha-1}{\alpha}}\right\}.$$

Equality holds at $\eta = \eta_{\text{bal}}$, because the two $\eta$-dependent terms are equal there. Although $\eta_{\text{bal}}$ is already a minimizer when $\sigma > 0$, we introduce the following alternative minimizer because the same formula also covers the deterministic case:

$$\eta^\star := \max\left\{\frac{L_0R_0}{M}, \sigma\left(\frac{K}{M}\right)^{\frac{1}{\alpha}}\right\}.$$

When $\sigma > 0$, we have $\eta^\star M = \max\left\{L_0R_0, \sigma K^{\frac{1}{\alpha}} M^{\frac{\alpha-1}{\alpha}}\right\}$ and $\frac{K\sigma^\alpha}{(\eta^\star)^{\alpha-1}} \leq \frac{K\sigma^\alpha}{\eta_{\text{bal}}^{\alpha-1}} = \sigma K^{\frac{1}{\alpha}} M^{\frac{\alpha-1}{\alpha}}$. Therefore,

$$h(\eta^\star) = \max\left\{L_0R_0, \sigma K^{\frac{1}{\alpha}} M^{\frac{\alpha-1}{\alpha}}\right\},$$

so $\eta^\star$ attains the preceding lower bound.

If $\sigma = 0$, then

$$h(\eta) = \max\{L_0R_0, \eta M\}.$$

If $L_0R_0 > 0$, its infimum is attained at

$$\eta^\star = \frac{L_0R_0}{M}.$$

If $L_0R_0 = 0$, the infimum is zero but is not attained over $\eta > 0$. Hence, whenever $L_0R_0 + \sigma > 0$, $\eta^\star$ attains the stated infimum. Substituting the definition of $M$ and setting $\lambda^\star = \eta^\star$ gives the claimed minimizer. □

*Proof of Theorem 1.* If $R_0 = 0$, then $x_0 = x^\star$, and the conclusion is immediate. Hence, assume that $R_0 > 0$.

Let

$$M := \max\left\{L_1R_0, \log\frac{4}{\delta}\right\}, \qquad \epsilon := \frac{40R_0\lambda M}{\nu K}.$$

The conditions of Proposition 4 have already been verified for these choices of $\lambda$, $\gamma$, and $\epsilon$. Therefore, with probability at least $1-\delta$,

$$\min_{0\leq k<K}(f(x_k) - f^\star) \leq \epsilon.$$

By the definition of $\lambda$, $\lambda M = \max\left\{L_0 R_0, \sigma K^{\frac{1}{\alpha}} M^{\frac{\alpha-1}{\alpha}}\right\}$. Consequently,

$$\epsilon = \frac{40}{\nu} \max\left\{\frac{L_0 R_0^2}{K}, \frac{R_0 \sigma M^{\frac{\alpha-1}{\alpha}}}{K^{\frac{\alpha-1}{\alpha}}}\right\}.$$

Substituting the definition of $M$ gives the claimed result. □

### C.2 A Computable Output

Table 4: Clipping level and stepsize for Theorem 2.

| Parameter | Choice |
|---|---|
| $\lambda$ | $\max\left\{\frac{L_0 R_0}{\max\{L_1 R_0, \log\left(\frac{12}{\delta}\right)\}}, \sigma\left(\frac{K}{\max\{L_1 R_0, \log\left(\frac{12}{\delta}\right)\}}\right)^{1/\alpha}\right\}$ |
| $\gamma$ | $\frac{\nu R_0}{16\lambda \max\{L_1 R_0, \log\left(\frac{12}{\delta}\right)\}}$ |

The cases $R_0 = 0$ and constant $f$ are immediate. We therefore consider $R_0 > 0$ and $L_0 > 0$. We assume throughout this subsection that

$$K \geq 1024\left[\log^2\left(\frac{12}{\delta}\right) + L_1 R_0 \log\left\{\frac{12e}{\delta}\left(1 + \frac{K\Delta_0}{L_0 R_0^2}\right)\right\}\right].$$

Choose

$$\begin{cases} \lambda = \max\left\{\frac{L_0 R_0}{\max\{L_1 R_0, \log\left(\frac{12}{\delta}\right)\}}, \sigma\left(\frac{K}{\max\{L_1 R_0, \log\left(\frac{12}{\delta}\right)\}}\right)^{1/\alpha}\right\}, \\ \gamma = \frac{\nu R_0}{16\lambda \max\{L_1 R_0, \log\left(\frac{12}{\delta}\right)\}}. \end{cases}$$

Writing $M := \max\{L_1 R_0, \log\left(\frac{12}{\delta}\right)\}$, the above choice implies

$$\begin{cases} \frac{\sigma}{\lambda} \leq \frac{1}{16}, \\ 2\gamma(L_0 + L_1\lambda) \leq \frac{\nu}{4}. \end{cases}$$

Indeed, $\lambda \geq \frac{L_0 R_0}{M}$ and $M \geq L_1 R_0$ imply $L_0 + L_1\lambda \leq \frac{2\lambda M}{R_0}$, and hence $2\gamma(L_0 + L_1\lambda) \leq \frac{\nu}{4}$. Moreover, the assumed lower bound on $K$ gives $K \geq 1024 M \log\left(\frac{12}{\delta}\right) \geq 16^\alpha M$. Consequently, $\left(\frac{\sigma}{\lambda}\right)^\alpha \leq \frac{M}{K}$ and $\frac{\sigma}{\lambda} \leq \left(\frac{M}{K}\right)^{1/\alpha} \leq \frac{1}{16}$.

The localization argument from the proof of Theorem 1, with failure probability $\frac{\delta}{6}$ for each of its two concentration bounds, gives an event $\mathcal{E}_1$ of probability at least $1 - \frac{\delta}{3}$ on which

$$\begin{cases} \tau > K, \\ \forall k \in \{0, \dots, K\},\ \|x_k - x^\star\| \leq 2R_0, \\ \sum_{k<K} \langle q_k, x_k - x^\star\rangle \leq \frac{8R_0^2}{5\gamma}. \end{cases}$$

**Lemma 12.** *Suppose that Assumptions 2 and 3 hold. Let $c_1, c_2 > 0$ satisfy $c_1^{-1} + 2c_2 < 1$. If $\lambda > 0$ and $\gamma > 0$ satisfy $\lambda \geq c_1\sigma$ and $\gamma(L_0 + L_1\lambda) \leq c_2$, then, for every iteration,*

$$\mathbb{E}[f(x_{k+1}) - f(x_k) \mid \mathcal{F}_k] \leq -\left(1 - \frac{1}{c_1} - 2c_2\right)\gamma\langle\nabla f(x_k), q_k\rangle + \left(4c_2 + \frac{c_1}{2}(1 - 4c_2)^2\right)\gamma\sigma^\alpha\lambda^{2-\alpha}.$$

**Lemma 13.** *Let $\mathcal{C} = \{0 \le k < K : \|\nabla f(x_k)\| < 5\lambda/6\}$. There exists an event $\mathcal{E}_{\mathcal{C}} \subseteq \mathcal{E}_1$ with $\Pr(\mathcal{E}_{\mathcal{C}}) \ge 1 - \delta/2$ such that, on $\mathcal{E}_{\mathcal{C}}$,*

$$|\mathcal{C}| \ge \frac{3K}{4}.$$

**Lemma 14.** *Let $\mathcal{U} = \{0 \le k < K : \|G_k\| \le \lambda\}$. There exists an event $\mathcal{E}_2 \subseteq \mathcal{E}_1$ with $\Pr(\mathcal{E}_2) \ge 1 - \frac{2\delta}{3}$ such that, on $\mathcal{E}_2$,*

$$|\mathcal{U}| \ge \frac{2K}{3}.$$

**Lemma 15.** *Let $\mathcal{O} := \left\{k \in \mathcal{U} : \|G_k - \nabla f(x_k)\| > \sigma \left(\frac{K}{\log(12/\delta)}\right)^{1/\alpha}\right\}$. There exists an event $\mathcal{E}_3 \subseteq \mathcal{E}_2$ with $\Pr(\mathcal{E}_3) \ge 1 - \delta$ such that, on $\mathcal{E}_3$, $|\mathcal{O}| < 4\log(12/\delta)$ and*

$$\sum_{k \in \mathcal{U}\setminus\mathcal{O}} \big(f(x_k) - f^\star\big) \le \frac{16R_0^2}{5\gamma} + 13R_0\lambda \max\{L_1 R_0, \log(12/\delta)\}.$$

**Lemma 16.** *Let $\bar{x}_K := \frac{1}{|\mathcal{U}\setminus\mathcal{O}|}\sum_{k\in\mathcal{U}\setminus\mathcal{O}} x_k$. Then, on $\mathcal{E}_3$,*

$$\begin{cases} f(\bar{x}_K) - f^\star < \dfrac{97}{\nu}\dfrac{R_0\lambda M}{K}, \\ \|\widehat{x}_K - \bar{x}_K\| < \dfrac{24R_0\log(12/\delta)}{K}. \end{cases}$$

### C.2.1 Proofs

*Proof of Lemma 12.* Recall that $e_k = g_k - q_k$. Nonexpansiveness of clipping and the oracle moment bound give $\mathbb{E}[\|e_k\| \mid \mathcal{F}_k] \le \sigma$ and $\mathbb{E}[\|e_k\|^2 \mid \mathcal{F}_k] \le 2\sigma^\alpha\lambda^{2-\alpha}$. We have

$$L_1\|x_{k+1} - x_k\| \le \gamma L_1\lambda \le \gamma(L_0 + L_1\lambda) \le c_2 < 1.$$

Thus Lemma 1 gives

$$f(x_{k+1}) - f(x_k) \le -\gamma\langle\nabla f(x_k), g_k\rangle + 2\gamma^2\big(L_0 + L_1\|\nabla f(x_k)\|\big)\|g_k\|^2.$$

- Assume that $\|\nabla f(x_k)\| \le \lambda$. Then $q_k = \nabla f(x_k)$. Using $2\gamma(L_0 + L_1\lambda) \le 2c_2$, we obtain

$$\begin{aligned} f(x_{k+1}) - f(x_k) &\le -\gamma\langle\nabla f(x_k), g_k\rangle + 2\gamma c_2\|g_k\|^2 \\ &= -\gamma\Big(\langle q_k, g_k\rangle - 2c_2\|g_k\|^2\Big). \end{aligned}$$

  Using $g_k = q_k + e_k$, we obtain

$$\langle q_k, g_k\rangle - 2c_2\|g_k\|^2 = (1-2c_2)\|q_k\|^2 + (1-4c_2)\langle q_k, e_k\rangle - 2c_2\|e_k\|^2.$$

  By Young's inequality, $(1-4c_2)\langle q_k, e_k\rangle \ge -\frac{1}{c_1}\|q_k\|^2 - \frac{c_1}{4}(1-4c_2)^2\|e_k\|^2$. Then

$$f(x_{k+1}) - f(x_k) \le -\gamma\left(1 - \frac{1}{c_1} - 2c_2\right)\|q_k\|^2 + \gamma\left(2c_2 + \frac{c_1}{4}(1-4c_2)^2\right)\|e_k\|^2.$$

  Taking conditional expectations and using $\|q_k\|^2 = \langle\nabla f(x_k), q_k\rangle$, we obtain

$$\begin{aligned} \mathbb{E}[f(x_{k+1}) - f(x_k) \mid \mathcal{F}_k] \le &-\gamma\left(1 - \frac{1}{c_1} - 2c_2\right)\langle\nabla f(x_k), q_k\rangle \\ &+ \gamma\left(4c_2 + \frac{c_1}{2}(1-4c_2)^2\right)\sigma^\alpha\lambda^{2-\alpha}. \end{aligned}$$

- Assume that $\|\nabla f(x_k)\| > \lambda$. Since $\|g_k\| \le \lambda$,

$$\begin{aligned} \big(L_0 + L_1\|\nabla f(x_k)\|\big)\|g_k\|^2 &\le \big(L_0 + L_1\|\nabla f(x_k)\|\big)\lambda^2 \\ &\le (L_0 + L_1\lambda)\lambda\|\nabla f(x_k)\| \\ &= (L_0 + L_1\lambda)\langle\nabla f(x_k), q_k\rangle. \end{aligned}$$

Moreover, $g_k = q_k + e_k$ and $\lambda \geq c_1\sigma$ imply

$$\begin{aligned}\mathbb{E}[\langle \nabla f(x_k), g_k\rangle \mid \mathcal{F}_k] &\geq \langle \nabla f(x_k), q_k\rangle - \sigma \left\|\nabla f(x_k)\right\| \\ &= \left(1 - \frac{\sigma}{\lambda}\right)\langle \nabla f(x_k), q_k\rangle \\ &\geq \left(1 - \frac{1}{c_1}\right)\langle \nabla f(x_k), q_k\rangle .\end{aligned}$$

Substituting these two bounds into the descent inequality yields

$$\begin{aligned}\mathbb{E}[f(x_{k+1}) - f(x_k) \mid \mathcal{F}_k] &\leq -\gamma\left(1 - \frac{1}{c_1} - 2\gamma(L_0 + L_1\lambda)\right)\langle \nabla f(x_k), q_k\rangle \\ &\leq -\gamma\left(1 - \frac{1}{c_1} - 2c_2\right)\langle \nabla f(x_k), q_k\rangle .\end{aligned}$$

The second case also implies the claimed inequality, since its noise term is nonnegative. □

*Proof of Lemma 13.* We apply Lemma 12 with $c_1 = 16$ and $c_2 = 3/32$. Indeed, the parameter choices give $\lambda \geq 16\sigma$ and $\gamma(L_0 + L_1\lambda) \leq \nu/8 < 3/32$. For these constants, the descent coefficient is $3/4$ and the noise coefficient is $7/2 \leq 4$. Moreover,

$$4\gamma\sigma^\alpha\lambda^{2-\alpha} = \frac{\nu R_0\lambda}{4M}\left(\frac{\sigma}{\lambda}\right)^\alpha \leq \frac{\nu R_0\lambda}{4K} \leq \frac{R_0\lambda}{K}.$$

Hence

$$\mathbb{E}[f(x_{k+1}) - f(x_k) \mid \mathcal{F}_k] \leq -\frac{3\gamma}{4}\langle \nabla f(x_k), q_k\rangle + \frac{R_0\lambda}{K}.$$

For $k < \tau$ with $k \notin \mathcal{C}$, convexity and $\|\nabla f(x_k)\| \geq \frac{5\lambda}{6}$ give

$$f(x_k) - f^\star + R_0\lambda \leq \frac{16}{5}R_0\left\|\nabla f(x_k)\right\| \leq \frac{4R_0}{\lambda}\langle \nabla f(x_k), q_k\rangle ,$$

where the last inequality uses $\|q_k\| \geq \frac{5\lambda}{6}$. Hence

$$\langle \nabla f(x_k), q_k\rangle \geq \frac{\lambda}{4R_0}\big(f(x_k) - f^\star + R_0\lambda\big).$$

Therefore, if $k < \tau$ and $k \notin \mathcal{C}$, then

$$\begin{aligned}&\mathbb{E}[f(x_{k+1}) - f^\star + R_0\lambda \mid \mathcal{F}_k] \\ &\leq f(x_k) - f^\star + R_0\lambda - \frac{3\gamma\lambda}{16R_0}\big(f(x_k) - f^\star + R_0\lambda\big) + \frac{R_0\lambda}{K} \\ &\leq \left(1 - \frac{3\gamma\lambda}{16R_0} + \frac{1}{K}\right)\big(f(x_k) - f^\star + R_0\lambda\big),\end{aligned}$$

For $k < \tau$ with $k \in \mathcal{C}$, using the same drift bound and $f(x_k) - f^\star + R_0\lambda \geq R_0\lambda$, we obtain

$$\mathbb{E}[f(x_{k+1}) - f^\star + R_0\lambda \mid \mathcal{F}_k] \leq \left(1 + \frac{1}{K}\right)\big(f(x_k) - f^\star + R_0\lambda\big).$$

From the two previous cases, for every $k < \tau$,

$$\begin{aligned}\mathbb{E}[f(x_{k+1}) - f^\star + R_0\lambda \mid \mathcal{F}_k] &\leq \left(1 - \frac{3\gamma\lambda}{16R_0}\,\mathbf{1}_{\{k\notin\mathcal{C}\}} + \frac{1}{K}\right)\big(f(x_k) - f^\star + R_0\lambda\big) \\ &\leq \exp\left(-\frac{3\gamma\lambda}{16R_0}\,\mathbf{1}_{\{k\notin\mathcal{C}\}} + \frac{1}{K}\right)\big(f(x_k) - f^\star + R_0\lambda\big),\end{aligned}$$

where we used $1 + t \leq e^t$. On $\{k < \tau\}$, we have $k \wedge \tau = k$ and $(k+1) \wedge \tau = k + 1$. On $\{k \geq \tau\}$, both stopped indices equal $\tau$, and the exponential multiplier below is $e^{-1/K} \leq 1$. Thus, for every $0 \leq k < K$,

$$\mathbb{E}\left[\big(f(x_{(k+1)\wedge\tau}) - f^\star + R_0\lambda\big)\exp\left\{\frac{3\gamma\lambda}{16R_0}\,\mathbf{1}_{\{k<\tau,\ k\notin\mathcal{C}\}} - \frac{1}{K}\right\}\,\middle|\,\mathcal{F}_k\right] \leq f(x_{k\wedge\tau}) - f^\star + R_0\lambda.$$

Multiplying by the exponential factor accumulated over $j < k$, which is $\mathcal{F}_k$-measurable, and taking expectations gives

$$\mathbb{E}\left[\left(f(x_{(k+1)\wedge\tau}) - f^\star + R_0\lambda\right)\exp\left\{\frac{3\gamma\lambda}{16R_0}\sum_{j\le k}\mathbf{1}_{\{j<\tau,\, j\notin\mathcal{C}\}} - \frac{k+1}{K}\right\}\right]$$
$$\le \mathbb{E}\left[\left(f(x_{k\wedge\tau}) - f^\star + R_0\lambda\right)\exp\left\{\frac{3\gamma\lambda}{16R_0}\sum_{j<k}\mathbf{1}_{\{j<\tau,\, j\notin\mathcal{C}\}} - \frac{k}{K}\right\}\right].$$

Iterating for $k = 0, \dots, K-1$ yields

$$\mathbb{E}\left[\left(f(x_{K\wedge\tau}) - f^\star + R_0\lambda\right)\exp\left\{\frac{3\gamma\lambda}{16R_0}\sum_{k<K}\mathbf{1}_{\{k<\tau,\, k\notin\mathcal{C}\}} - 1\right\}\right] \le \Delta_0 + R_0\lambda.$$

Since $f(x_{K\wedge\tau}) \ge f^\star$, we obtain

$$\mathbb{E}\left[\exp\left\{\frac{3\gamma\lambda}{16R_0}\sum_{k<K}\mathbf{1}_{\{k<\tau,\, k\notin\mathcal{C}\}}\right\}\right] \le e\left(1 + \frac{\Delta_0}{R_0\lambda}\right).$$

Therefore, by Markov's inequality, there exists an event $\widetilde{\mathcal{E}}_{\mathcal{C}}$ with $\Pr(\widetilde{\mathcal{E}}_{\mathcal{C}}) \ge 1 - \delta/6$ such that

$$\sum_{k<K}\mathbf{1}_{\{k<\tau,\, k\notin\mathcal{C}\}} \le \frac{16R_0}{3\gamma\lambda}\log\left\{\frac{6e}{\delta}\left(1 + \frac{\Delta_0}{R_0\lambda}\right)\right\}.$$

Define $\mathcal{E}_{\mathcal{C}} := \mathcal{E}_1 \cap \widetilde{\mathcal{E}}_{\mathcal{C}}$. Since $\Pr(\mathcal{E}_1) \ge 1 - \delta/3$, the union bound gives $\Pr(\mathcal{E}_{\mathcal{C}}) \ge 1 - \delta/2$. On $\mathcal{E}_{\mathcal{C}}$, we have $\tau > K$. Substituting $\gamma$ gives

$$K - |\mathcal{C}| \le \frac{256M}{3\nu}\log\left\{\frac{6e}{\delta}\left(1 + \frac{\Delta_0}{R_0\lambda}\right)\right\}.$$

We now show that this is at most $\frac{K}{4}$ under the assumed budget.

If $L_1R_0 > \log(\frac{12}{\delta})$, then $M = L_1R_0$ and $K \ge M$. Since $R_0\lambda \ge \frac{L_0R_0^2}{M} \ge \frac{L_0R_0^2}{K}$,

$$\begin{aligned} K - |\mathcal{C}| &\le \frac{256L_1R_0}{3\nu}\log\left\{\frac{12e}{\delta}\left(1 + \frac{K\Delta_0}{L_0R_0^2}\right)\right\} \\ &\le \frac{K}{12\nu} < \frac{K}{4}. \end{aligned}$$

If $L_1R_0 \le \log(\frac{12}{\delta})$, then $M = \log(\frac{12}{\delta}) > 2$. Generalized smoothness and $\nabla f(x^\star) = 0$ give $\Delta_0 \le \frac{L_0R_0^2}{2}e^{L_1R_0}$ so $1 + \frac{\Delta_0}{R_0\lambda} \le 1 + \frac{M}{2}e^{L_1R_0} \le Me^{L_1R_0}$. Consequently, using $\log t \le t - 1$ with $t = \frac{M}{2}$,

$$\begin{aligned} \log\left\{\frac{6e}{\delta}\left(1 + \frac{\Delta_0}{R_0\lambda}\right)\right\} &\le \log\left(\frac{6e}{\delta}\right) + \log M + L_1R_0 \\ &= M - \log 2 + 1 + \log M + L_1R_0 \\ &= L_1R_0 + M + 1 + \log\left(\frac{M}{2}\right) \\ &\le L_1R_0 + M + 1 + \left(\frac{M}{2} - 1\right) \\ &= L_1R_0 + \frac{3}{2}M, \end{aligned}$$

Thus, on $\mathcal{E}_{\mathcal{C}}$, $|\mathcal{C}| \ge \frac{3K}{4}$. □

*Proof of Lemma 14.* If $k \in \mathcal{C} \setminus \mathcal{U}$, then $\|\nabla f(x_k)\| < \frac{5\lambda}{6}$ and $\|G_k\| > \lambda$. Then $\|G_k - \nabla f(x_k)\| > \frac{\lambda}{6}$. By Markov's inequality and the moment bound in Assumption 3,

$$\Pr(k \in \mathcal{C} \setminus \mathcal{U} \mid \mathcal{F}_k) \le 6^\alpha \left(\frac{\sigma}{\lambda}\right)^\alpha \le \frac{36M}{K},$$

where we used $\alpha \le 2$. Define

$$Z_k := \mathbf{1}_{\{k \in \mathcal{C} \setminus \mathcal{U}\}} - \mathbb{E}\left[\mathbf{1}_{\{k \in \mathcal{C} \setminus \mathcal{U}\}} \mid \mathcal{F}_k\right].$$

Then $\mathbb{E}[Z_k \mid \mathcal{F}_k] = 0$, $|Z_k| \le 1$, and $\mathbb{E}[Z_k^2 \mid \mathcal{F}_k] \le \mathbb{E}\left[\mathbf{1}_{\{k \in \mathcal{C} \setminus \mathcal{U}\}} \mid \mathcal{F}_k\right] \le \frac{36M}{K}$. Thus,

$$\sum_{k<K} \mathbb{E}[Z_k^2 \mid \mathcal{F}_k] \le 36M.$$

By Corollary 1, with probability at least $1 - \delta/6$,

$$\sum_{k<K} Z_k \le \sqrt{72M \log\left(\frac{12}{\delta}\right)} + \frac{2}{3}\log\left(\frac{12}{\delta}\right).$$

On this event,

$$\begin{aligned}
|\mathcal{C} \setminus \mathcal{U}| &= \sum_{k<K} \mathbb{E}\left[\mathbf{1}_{\{k \in \mathcal{C} \setminus \mathcal{U}\}} \mid \mathcal{F}_k\right] + \sum_{k<K} Z_k \\
&\le 36M + \sqrt{72M \log\left(\frac{12}{\delta}\right)} + \frac{2}{3}\log\left(\frac{12}{\delta}\right) \\
&\le \left(36 + \sqrt{72} + \frac{2}{3}\right) M < 46M,
\end{aligned}$$

where we used $M \ge \log(12/\delta)$. Using $K \ge 1024M$, we deduce

$$|\mathcal{C} \setminus \mathcal{U}| < 46M \le \frac{K}{12}.$$

Let $\tilde{\mathcal{E}}$ denote the event established above and define $\mathcal{E}_2 := \mathcal{E}_{\mathcal{C}} \cap \tilde{\mathcal{E}}$. Since $\Pr(\mathcal{E}_{\mathcal{C}}) \ge 1 - \frac{\delta}{2}$ and $\Pr(\tilde{\mathcal{E}}) \ge 1 - \frac{\delta}{6}$, the union bound gives $\Pr(\mathcal{E}_2) \ge 1 - \frac{2\delta}{3}$. On $\mathcal{E}_2$, we have

$$|\mathcal{U}| \ge |\mathcal{C} \cap \mathcal{U}| = |\mathcal{C}| - |\mathcal{C} \setminus \mathcal{U}| \ge \frac{3K}{4} - \frac{K}{12} = \frac{2K}{3}.$$

□

*Proof of Lemma 15.* Recall that $M := \max\{L_1 R_0, \log(12/\delta)\}$.

If $\sigma = 0$, then $G_k = \nabla f(x_k)$ almost surely, so $\mathcal{O} = \varnothing$. Moreover, for $k \in \mathcal{U}$ we have $\|\nabla f(x_k)\| = \|G_k\| \le \lambda$. Thus, on $\mathcal{E}_2 \subseteq \mathcal{E}_1$, we have $f(x_k) - f^\star \le \langle \nabla f(x_k), x_k - x^\star \rangle = \langle q_k, x_k - x^\star \rangle$, and hence

$$\sum_{k \in \mathcal{U}} \big(f(x_k) - f^\star\big) \le \sum_{k<K} \langle q_k, x_k - x^\star \rangle \le \frac{8R_0^2}{5\gamma}.$$

Therefore the claim holds with $\mathcal{E}_3 = \mathcal{E}_2$. Hence, assume $\sigma > 0$.

We first bound $|\mathcal{O}|$. By Markov's inequality and the moment bound in Assumption 3,

$$\Pr(k \in \mathcal{O} \mid \mathcal{F}_k) \le \Pr\left(\|G_k - \nabla f(x_k)\| > \sigma \left(\frac{K}{\log(12/\delta)}\right)^{1/\alpha} \,\middle|\, \mathcal{F}_k\right) \le \frac{\log(12/\delta)}{K}.$$

Define $Z_k := \mathbf{1}_{\{k \in \mathcal{O}\}} - \mathbb{E}\left[\mathbf{1}_{\{k \in \mathcal{O}\}} \mid \mathcal{F}_k\right]$. Then $\mathbb{E}[Z_k \mid \mathcal{F}_k] = 0$, $|Z_k| \le 1$, and

$$\sum_{k<K} \mathbb{E}[Z_k^2 \mid \mathcal{F}_k] \le \sum_{k<K} \Pr(k \in \mathcal{O} \mid \mathcal{F}_k) \le \log(12/\delta).$$

By Corollary 1, with probability at least $1 - \delta/6$,

$$\sum_{k<K} Z_k \le \sqrt{2}\log(12/\delta) + \frac{2}{3}\log(12/\delta).$$

Therefore, on this event,

$$|\mathcal{O}| \le \left(1 + \sqrt{2} + \frac{2}{3}\right)\log(12/\delta) < 4\log(12/\delta).$$

We next control the selected iterates with $\|\nabla f(x_k)\| > 2\lambda$. Define

$$Y_k := \mathbf{1}_{\{k<\tau\}}\mathbf{1}_{\{k\in\mathcal{U}\setminus\mathcal{O},\,\|\nabla f(x_k)\|>2\lambda\}}\big(f(x_k) - f^\star\big).$$

For $k < \tau$, by convexity

$$f(x_k) - f^\star \le \langle \nabla f(x_k), x_k - x^\star\rangle \le 2R_0\,\|\nabla f(x_k)\|\,.$$

If $\|\nabla f(x_k)\| > 2\lambda$ and $k \in \mathcal{U}$, then $\|G_k - \nabla f(x_k)\| \ge \|\nabla f(x_k)\| - \lambda$. Hence, by Markov's inequality and Assumption 3,

$$\Pr(k \in \mathcal{U} \mid \mathcal{F}_k) \le \frac{\sigma^\alpha}{\big(\|\nabla f(x_k)\| - \lambda\big)^\alpha}.$$

Therefore, on $\{k < \tau, \|\nabla f(x_k)\| > 2\lambda\}$,

$$\mathbb{E}[Y_k \mid \mathcal{F}_k] \le \frac{2R_0\,\|\nabla f(x_k)\|\,\sigma^\alpha}{\big(\|\nabla f(x_k)\| - \lambda\big)^\alpha} \le \frac{4R_0\sigma^\alpha}{\big(\|\nabla f(x_k)\| - \lambda\big)^{\alpha-1}} \le \frac{4R_0\sigma^\alpha}{\lambda^{\alpha-1}},$$

where we used $\|\nabla f(x_k)\| \le 2(\|\nabla f(x_k)\| - \lambda)$. Outside this event, $Y_k = 0$, so the same bound holds for every $k$. Since $\left(\frac{\sigma}{\lambda}\right)^\alpha \le \frac{M}{K}$,

$$\sum_{k<K} \mathbb{E}[Y_k \mid \mathcal{F}_k] \le \frac{4KR_0\sigma^\alpha}{\lambda^{\alpha-1}} \le 4R_0\lambda M.$$

On $\{k \in \mathcal{U}\setminus\mathcal{O}\}$, $\|\nabla f(x_k)\| \le \|G_k\| + \|G_k - \nabla f(x_k)\| \le \lambda + \sigma\left(\frac{K}{\log(12/\delta)}\right)^{1/\alpha}$. Thus

$$0 \le Y_k \le 2R_0\left[\lambda + \sigma\left(\frac{K}{\log(12/\delta)}\right)^{1/\alpha}\right].$$

Since $\lambda \ge \sigma\left(\frac{K}{M}\right)^{1/\alpha}$ and $M \ge \log\left(\frac{12}{\delta}\right)$,

$$\begin{aligned}2R_0\left[\lambda + \sigma\left(\frac{K}{\log(12/\delta)}\right)^{1/\alpha}\right]\log(12/\delta) &\le 2R_0\lambda\left[\log(12/\delta) + M^{1/\alpha}\log^{1-1/\alpha}(12/\delta)\right]\\ &\le 4R_0\lambda M.\end{aligned}$$

Moreover,

$$\begin{aligned}\sum_{k<K}\mathbb{E}\Big[(Y_k - \mathbb{E}[Y_k\mid\mathcal{F}_k])^2 \mid \mathcal{F}_k\Big] &\le \sum_{k<K}\mathbb{E}[Y_k^2\mid\mathcal{F}_k]\\ &\le 2R_0\left[\lambda + \sigma\left(\frac{K}{\log(12/\delta)}\right)^{1/\alpha}\right]\sum_{k<K}\mathbb{E}[Y_k\mid\mathcal{F}_k]\\ &\le 8R_0^2\lambda M\left[\lambda + \sigma\left(\frac{K}{\log(12/\delta)}\right)^{1/\alpha}\right]\\ &\le \frac{16(R_0\lambda M)^2}{\log(12/\delta)}.\end{aligned}$$

Applying Corollary 1 and using the preceding bounds, with probability at least $1-\delta/6$,

$$\sum_{k<K}\left(Y_k-\mathbb{E}[Y_k\mid\mathcal{F}_k]\right)\le\sqrt{32}R_0\lambda M+\frac{8}{3}R_0\lambda M.$$

Consequently, on the same event,

$$\sum_{k<K}Y_k\le\left(4+\sqrt{32}+\frac{8}{3}\right)R_0\lambda M<13R_0\lambda M.$$

Let $\mathcal{E}_3$ be the intersection of $\mathcal{E}_2$ with the two concentration events established above. Since $\Pr(\mathcal{E}_2)\ge 1-\frac{2\delta}{3}$, the union bound gives

$$\Pr(\mathcal{E}_3)\ge 1-\frac{2\delta}{3}-\frac{\delta}{6}-\frac{\delta}{6}=1-\delta.$$

We now work on $\mathcal{E}_3$. Since $\mathcal{E}_3\subseteq\mathcal{E}_2\subseteq\mathcal{E}_1$, we have $\tau>K$ and

$$\sum_{k<K}\langle q_k,x_k-x^\star\rangle\le\frac{8R_0^2}{5\gamma}.$$

For $\|\nabla f(x_k)\|\le 2\lambda$, the definition of $q_k$ and convexity give

$$f(x_k)-f^\star\le\langle\nabla f(x_k),x_k-x^\star\rangle\le 2\langle q_k,x_k-x^\star\rangle.$$

Then

$$\sum_{\substack{k\in\mathcal{U}\setminus\mathcal{O}\\ \|\nabla f(x_k)\|\le 2\lambda}}\left(f(x_k)-f^\star\right)\le 2\sum_{k<K}\langle q_k,x_k-x^\star\rangle\le\frac{16R_0^2}{5\gamma}.$$

Since $\tau>K$, the definition of $Y_k$ gives

$$\sum_{\substack{k\in\mathcal{U}\setminus\mathcal{O}\\ \|\nabla f(x_k)\|>2\lambda}}\left(f(x_k)-f^\star\right)=\sum_{k<K}Y_k<13R_0\lambda M.$$

We conclude that

$$\sum_{k\in\mathcal{U}\setminus\mathcal{O}}\left(f(x_k)-f^\star\right)\le\frac{16R_0^2}{5\gamma}+13R_0\lambda M.$$

□

*Proof of Lemma 16.* We first show that $\bar{x}_K$ is well-defined on $\mathcal{E}_3$. By Lemma 14 and Lemma 15, on $\mathcal{E}_3$, $|\mathcal{U}|\ge\frac{2K}{3}$ and $|\mathcal{O}|<4\log\left(\frac{12}{\delta}\right)$. Moreover, the assumed iteration requirement gives $K\ge 1024\log^2\left(\frac{12}{\delta}\right)$. Since $\log\left(\frac{12}{\delta}\right)>2$, it holds that $4\log\left(\frac{12}{\delta}\right)<\frac{K}{512}$. Therefore,

$$|\mathcal{U}\setminus\mathcal{O}|=|\mathcal{U}|-|\mathcal{O}|>\frac{2K}{3}-\frac{K}{512}>\frac{85K}{128}.$$

By convexity and Lemma 15,

$$f(\bar{x}_K)-f^\star\le\frac{1}{|\mathcal{U}\setminus\mathcal{O}|}\sum_{k\in\mathcal{U}\setminus\mathcal{O}}\left(f(x_k)-f^\star\right)<\frac{128}{85K}\left(\frac{16R_0^2}{5\gamma}+13R_0\lambda M\right).$$

Using $\gamma=\frac{\nu R_0}{16\lambda M}$, we have $\frac{16R_0^2}{5\gamma}=\frac{256}{5\nu}R_0\lambda M$. Hence

$$\begin{aligned}f(\bar{x}_K)-f^\star&<\frac{128}{85}\left(\frac{256}{5\nu}+13\right)\frac{R_0\lambda M}{K}\\&<\frac{128}{85\nu}\left(\frac{256}{5}+13\right)\frac{R_0\lambda M}{K}<\frac{97}{\nu}\frac{R_0\lambda M}{K},\end{aligned}$$

where we used $\nu < 1$.

It remains to bound $\|\widehat{x}_K - \bar{x}_K\|$. Since $\mathcal{O} \subseteq \mathcal{U}$,

$$\widehat{x}_K = \frac{|\mathcal{U} \setminus \mathcal{O}|}{|\mathcal{U}|}\bar{x}_K + \frac{1}{|\mathcal{U}|}\sum_{k\in\mathcal{O}} x_k,$$

and therefore

$$\widehat{x}_K - \bar{x}_K = \frac{1}{|\mathcal{U}|}\sum_{k\in\mathcal{O}}(x_k - \bar{x}_K).$$

Since $\mathcal{E}_3 \subseteq \mathcal{E}_1$, $\|x_k - x^\star\| \le 2R_0$ for every $k \le K$. Also, since $\bar{x}_K$ is an average of such iterates,

$$\|\bar{x}_K - x^\star\| \le \frac{1}{|\mathcal{U} \setminus \mathcal{O}|}\sum_{k\in\mathcal{U}\setminus\mathcal{O}} \|x_k - x^\star\| \le 2R_0.$$

Thus, for every $k \in \mathcal{O}$,

$$\|x_k - \bar{x}_K\| \le \|x_k - x^\star\| + \|\bar{x}_K - x^\star\| \le 4R_0.$$

Consequently,

$$\|\widehat{x}_K - \bar{x}_K\| \le \frac{4R_0|\mathcal{O}|}{|\mathcal{U}|} < \frac{16R_0\log\left(\frac{12}{\delta}\right)}{2K/3} = \frac{24R_0\log\left(\frac{12}{\delta}\right)}{K}.$$

□

*Proof of Theorem 2.* Let $M := \max\{L_1R_0, \log(12/\delta)\}$. By Lemma 16, on $\mathcal{E}_3$,

$$\begin{cases} f(\bar{x}_K) - f^\star < \dfrac{97}{\nu}\dfrac{R_0\lambda M}{K}, \\ D_K < \dfrac{24R_0\log(12/\delta)}{K}, \end{cases}$$

where $D_K := \|\widehat{x}_K - \bar{x}_K\|$. Since $\Pr(\mathcal{E}_3) \ge 1 - \delta$, it remains to prove the claimed bound on this event.

We have $K \ge 1024M\log\left(\frac{12}{\delta}\right)$. Since $M \ge L_1R_0$, we obtain

$$L_1D_K < \frac{24L_1R_0\log(12/\delta)}{K} \le \frac{24M\log(12/\delta)}{K} \le \frac{3}{128} < \frac{1}{32}.$$

Therefore, applying Lemma 1 with $x = \bar{x}_K$ and $y = \widehat{x}_K$, we obtain

$$\begin{aligned} f(\widehat{x}_K) - f^\star &\le f(\bar{x}_K) - f^\star + \|\nabla f(\bar{x}_K)\| D_K + 2\left(L_0 + L_1\|\nabla f(\bar{x}_K)\|\right)D_K^2 \\ &= f(\bar{x}_K) - f^\star + (1 + 2L_1D_K)\|\nabla f(\bar{x}_K)\| D_K + 2L_0D_K^2 \\ &\le f(\bar{x}_K) - f^\star + \frac{17}{16}\|\nabla f(\bar{x}_K)\| D_K + 2L_0D_K^2. \end{aligned}$$

Let $\Delta := f(\bar{x}_K) - f^\star$. By Lemma 3, $\|\nabla f(\bar{x}_K)\| \le \frac{1}{\nu}\left(L_1\Delta + \sqrt{L_1^2\Delta^2 + 2\nu L_0\Delta}\right)$. Using $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ and $\nu > 1/2$, follows that

$$\|\nabla f(\bar{x}_K)\| \le \frac{2L_1}{\nu}\Delta + \sqrt{\frac{2}{\nu}}\sqrt{L_0\Delta} \le 4L_1\Delta + 2\sqrt{L_0\Delta}.$$

Therefore

$$f(\widehat{x}_K) - f^\star \le \Delta + \frac{17}{4}L_1D_K\,\Delta + \frac{17}{8}D_K\sqrt{L_0\Delta} + 2L_0D_K^2.$$

Since $L_1D_K < 1/32$, it holds that $\frac{17}{4}L_1D_K\,\Delta \le \frac{17}{128}\Delta$. By Young's inequality, we obtain

$$\frac{17}{8}D_K\sqrt{L_0\Delta} \le \frac{15}{128}\Delta + \frac{289}{30}L_0D_K^2.$$

Hence

$$f(\widehat{x}_K) - f^\star \le \left(1 + \frac{17}{128} + \frac{15}{128}\right)\Delta + \left(2 + \frac{289}{30}\right) L_0 D_K^2$$
$$< \frac{5}{4}\Delta + 12 L_0 D_K^2.$$

By Lemma 16,

$$\begin{cases} \Delta < \dfrac{97}{\nu}\dfrac{R_0 \lambda M}{K}, \\ 12 L_0 D_K^2 < \dfrac{6912 L_0 R_0^2 \log^2(12/\delta)}{K^2}. \end{cases}$$

Since $K \ge 1024 \log^2(12/\delta)$, it follows that $12 L_0 D_K^2 \le \frac{27}{4}\frac{L_0 R_0^2}{K}$. By the definition of $\lambda$, we have $\lambda \ge \frac{L_0 R_0}{M}$, and hence $\frac{L_0 R_0^2}{K} \le \frac{R_0 \lambda M}{K}$. Therefore

$$f(\widehat{x}_K) - f^\star < \frac{5}{4}\frac{97}{\nu}\frac{R_0 \lambda M}{K} + \frac{27}{4}\frac{R_0 \lambda M}{K}$$
$$\le \frac{1}{4\nu}(485 + 27)\frac{R_0 \lambda M}{K}$$
$$= \frac{128}{\nu}\frac{R_0 \lambda M}{K},$$

where we used $\nu < 1$.

By the definition of $\lambda$,

$$\frac{R_0 \lambda M}{K} = \max\left\{\frac{L_0 R_0^2}{K},\ R_0 \sigma \left(\frac{M}{K}\right)^{\frac{\alpha-1}{\alpha}}\right\}.$$

Substituting $M = \max\{L_1 R_0, \log(12/\delta)\}$, we obtain

$$f(\widehat{x}_K) - f^\star \le \frac{128}{\nu} \max\left\{\frac{L_0 R_0^2}{K},\ \frac{R_0 \sigma \max\left\{L_1 R_0, \log\frac{12}{\delta}\right\}^{\frac{\alpha-1}{\alpha}}}{K^{\frac{\alpha-1}{\alpha}}}\right\}.$$

Since $\Pr(\mathcal{E}_3) \ge 1 - \delta$, the result follows. □

### C.3 Detailed comparison underlying Table 1

We provide here the technical details supporting the comparisons summarized in Table 1. Throughout this subsection, $\Delta_0 = f(x_0) - f^\star$.

**Reduction to classical smoothness.** For classical $L$-smooth objectives, Sadiev et al. (2023) obtain the standard heavy-tailed rate using a localization argument. Their assumptions are imposed on a ball containing the iterates with high probability. When this guarantee is applied to an $(L_0, L_1)$-smooth objective, the gradient is classically Lipschitz on $B_r(x^\star)$ with constant of order $L_0 \exp(L_1 r)$. Since the localization radius is of order $R_0$, such a reduction can therefore introduce exponential dependence on $L_1 R_0$.

**Deterministic benchmarks and refined transients.** For deterministic $(L_0, L_1)$-smooth optimization, the rate displayed for Gorbunov et al. (2025) in Table 1 holds under the sufficient requirement $K = \Omega((L_1 R_0)^2)$. Subsequent refined analyses relax this quadratic transient. In particular, for $L_0, R_0 > 0$, Vankov et al. (2025, Theorem 3.2) and Lobanov et al. (2024, Theorem 3.5) imply the same rate under $K = \Omega\left(L_1 R_0 \log\left(\frac{K\Delta_0}{L_0 R_0^2}\right)\right)$. If $\Delta_0 \le \frac{2 L_0 R_0^2}{K}$, the initial point is already at the displayed $\mathcal{O}\left(\frac{L_0 R_0^2}{K}\right)$ accuracy scale, so the logarithmic transient is relevant only in the complementary regime. The implicit condition above admits the convenient explicit sufficient form $K = \Omega\left(L_1 R_0 \log\left(e + \frac{L_1 \Delta_0}{L_0 R_0}\right)\right)$. Thus, the refined initial-gap-dependent condition can be substantially smaller than the quadratic requirement.

**A favorable instance.** To illustrate this improvement, consider the convex function $f(x) = e^x - x - 1$, which is globally $(1, 1)$-smooth since $e^x \le 1 + |e^x - 1|$ for all $x \in \mathbb{R}$. Moreover, $L_1 = 1$ is sharp globally: for any $\widetilde{L}_1 < 1$ and finite $\widetilde{L}_0$, the inequality $f''(x) \le \widetilde{L}_0 + \widetilde{L}_1 |f'(x)|$ fails for sufficiently large $x$. Taking $x_0 = -R_0$ gives $\Delta_0 = e^{-R_0} + R_0 - 1 = \Theta(R_0)$, so the explicit refined sufficient budget scales as $\Theta(L_1 R_0)$ for large $R_0$, compared with the quadratic sufficient budget in the earlier analysis. On the other hand, generalized smoothness and $\nabla f(x^\star) = 0$ imply $\Delta_0 \le \frac{L_0 R_0^2}{2} \exp(L_1 R_0)$, and hence $\frac{L_1 \Delta_0}{L_0 R_0} \le \frac{L_1 R_0}{2} \exp(L_1 R_0)$. Therefore, the explicit refined condition has at most $\mathcal{O}((L_1 R_0)^2)$ transient scaling in the worst case, while allowing substantially smaller transients on favorable instances.

**Accelerated deterministic methods.** Our comparison with Vankov et al. (2025) concerns their gradient method in Theorem 3.2. Their accelerated method in Theorem 6.2 achieves the $\mathcal{O}\left(\frac{L_0 R_0^2}{K^2}\right)$ rate under a weaker transient requirement. Here $K$ counts outer iterations, and each iteration requires a line search using function values, whose oracle cost is additional. These function values are not provided by our stochastic-gradient oracle, so this accelerated guarantee does not directly apply in our setting.

**Comparison with the light-tailed setting.** The sub-Gaussian result of Gaash et al. (2025) also treats convex $(L_0, L_1)$-smooth objectives and obtains the rate displayed in Table 1 under a quadratic sufficient transient. Their analysis considers a double-sampling variant in which one stochastic gradient determines the clipping test and an independent stochastic gradient determines the update; the output averages iterates selected by the clipping test. In contrast, our result applies directly to standard Clip-SGD, which clips the same stochastic gradient used for the update, and uses a computable average of its iterates. Moreover, our oracle assumption only requires a bounded central $\alpha$-th moment, with $\alpha \in (1, 2]$, rather than sub-Gaussian noise.

**Stochastic transient.** Our result extends the refined transient picture above to stochastic gradients with bounded central $\alpha$-th moment. For $L_0, R_0 > 0$ and any $\sigma \ge 0$, a sufficient iteration budget is $K = \Omega\Big(\log^2\left(\frac{12}{\delta}\right) + L_1 R_0 \log\left\{\frac{12e}{\delta}\left(1 + \frac{K\Delta_0}{L_0 R_0^2}\right)\right\}\Big)$, while the convergence guarantee is given in Table 1. Thus, rather than requiring a quadratic transient uniformly, our stochastic condition retains the refined dependence on the initial gap while allowing heavy-tailed noise.

**Noiseless reduction.** To make the deterministic comparison precise, fix the confidence parameter used in the tuning to $1/2$. Since $\sigma = 0$, the oracle, the iterates, and the output are deterministic, so any positive-probability guarantee holds with probability one. Consider first the nontrivial regime $\frac{K\Delta_0}{L_0 R_0^2} > 2$. Indeed, if $\frac{K\Delta_0}{L_0 R_0^2} \le 2$, then objective monotonicity and convexity give $f(\widehat{x}_K) - f^\star \le \Delta_0 \le \frac{2 L_0 R_0^2}{K}$, so the desired deterministic rate already holds.

In the nontrivial regime, $\log\left(1 + \frac{K\Delta_0}{L_0 R_0^2}\right) \le 2 \log\left(\frac{K\Delta_0}{L_0 R_0^2}\right)$. After fixing the confidence parameter, the general iteration requirement becomes $K = \Omega\Big(1 + L_1 R_0 + L_1 R_0 \log\left(1 + \frac{K\Delta_0}{L_0 R_0^2}\right)\Big)$. If $L_1 R_0 \ge 1$, the nontrivial regime also gives $\log\left(\frac{K\Delta_0}{L_0 R_0^2}\right) > \log 2$, and hence $1 + L_1 R_0 = \mathcal{O}\Big(L_1 R_0 \log\left(\frac{K\Delta_0}{L_0 R_0^2}\right)\Big)$. Together with the preceding logarithmic bound, this reduces the sufficient iteration requirement to $K = \Omega\Big(L_1 R_0 \log\left(\frac{K\Delta_0}{L_0 R_0^2}\right)\Big)$.

If $L_1 R_0 < 1$, generalized smoothness gives $\Delta_0 \le \frac{L_0 R_0^2}{2} \exp(L_1 R_0) \le \frac{e}{2} L_0 R_0^2$. After fixing the confidence parameter, the right-hand side of the original iteration requirement is therefore bounded by $C(1 + \log(1 + K))$ for a universal constant $C$. Hence, there exists a universal constant $K_0$ such that the original requirement holds for every $K \ge K_0$. For $1 \le K < K_0$, objective monotonicity and convexity give $f(\widehat{x}_K) - f^\star \le \frac{e}{2} L_0 R_0^2 \le \frac{e K_0}{2} \frac{L_0 R_0^2}{K}$. Thus, when $\sigma = 0$, our method achieves the $\mathcal{O}\left(\frac{L_0 R_0^2}{K}\right)$ rate under the same refined sufficient iteration condition as the gradient methods analyzed in Vankov et al. (2025, Theorem 3.2) and Lobanov et al. (2024, Theorem 3.5). When $L_1 R_0 < 1$, the

argument additionally shows that this rate holds for every $K \geq 1$, with a possibly larger universal numerical constant.

# D High-Probability Guarantees for Nonconvex Objectives

## D.1 Convergence Analysis

**Lemma 17.** *Assume that Assumption 3 holds and that $\lambda \geq 8\sigma$. Then, for every $k \geq 0$, almost surely,*

$$-\langle \nabla f(x_k), \mathbb{E}[e_k \mid \mathcal{F}_k] \rangle \leq \frac{1}{4} \|\nabla f(x_k)\| \min\{\|\nabla f(x_k)\|, \lambda\} + 2\frac{\sigma^{2\alpha}}{\lambda^{2\alpha-2}}.$$

**Lemma 18.** *Assume that Assumptions 2 and 3 hold. Suppose that $\lambda \geq 8\sigma$, $\gamma L_0 \leq \frac{1}{64}$ and $\gamma L_1 \lambda \leq \frac{1}{64}$. For every $k \geq 0$, define*

$$\Phi_k := -\gamma \langle \nabla f(x_k), g_k \rangle + 2\gamma^2 (L_0 + L_1 \|\nabla f(x_k)\|) \|g_k\|^2.$$

*Then, almost surely,*

$$f(x_{k+1}) - f(x_k) \leq \Phi_k,$$

*and*

$$\mathbb{E}[\Phi_k \mid \mathcal{F}_k] \leq -\frac{\gamma}{2} \|\nabla f(x_k)\| \min\{\|\nabla f(x_k)\|, \lambda\} + 3\gamma \frac{\sigma^{2\alpha}}{\lambda^{2\alpha-2}} + 8\gamma^2 L_0 \sigma^\alpha \lambda^{2-\alpha}.$$

*Moreover,*

$$\begin{cases} |\Phi_k - \mathbb{E}[\Phi_k \mid \mathcal{F}_k]| \leq 4\gamma\lambda \|\nabla f(x_k)\| + 8\gamma^2 L_0 \lambda^2, \\ \quad \mathrm{Var}(\Phi_k \mid \mathcal{F}_k) \leq 64\sigma^\alpha \lambda^{2-\alpha} \left(\gamma^2 \|\nabla f(x_k)\|^2 + \gamma^4 L_0^2 \lambda^2\right). \end{cases}$$

Let $\Delta_k := f(x_k) - f^\star$ and define the stopping time

$$\widetilde{\tau} := \inf\{k \geq 0 : \ \Delta_k > 2\Delta_0\},$$

with the convention $\inf \emptyset = +\infty$. On the event $\{k < \widetilde{\tau}\}$, we have $\Delta_k \leq 2\Delta_0$.

Define

$$\begin{cases} H_0 := \frac{2}{\nu}\left(L_1\Delta_0 + \sqrt{L_1^2\Delta_0^2 + \nu L_0 \Delta_0}\right), \\ \rho_0 := \max\left\{1, \frac{H_0}{\lambda}\right\}. \end{cases}$$

**Lemma 19.** *Let $K \geq 1$. Assume that Assumptions 2 and 3 hold. Suppose that $\lambda \geq 8\sigma$, $\gamma L_0 \leq \frac{1}{64}$ and $\gamma L_1 \lambda \leq \frac{1}{64}$. Define $S_K := \sum_{k=0}^{K-1} \mathbf{1}_{\{k<\widetilde{\tau}\}} \|\nabla f(x_k)\| \min\{\|\nabla f(x_k)\|, \lambda\}$ and*

$$Z_k := \mathbf{1}_{\{k<\widetilde{\tau}\}} (\Phi_k - \mathbb{E}[\Phi_k \mid \mathcal{F}_k]).$$

*Then $|Z_k| \leq \frac{5}{2}\gamma\lambda H_0 + 4\gamma^2 L_0 \lambda^2$ almost surely and,*

$$\sum_{k=0}^{K-1} \mathbb{E}[Z_k^2 \mid \mathcal{F}_k] \leq 64\sigma^\alpha \lambda^{2-\alpha} \left(\gamma^2 \rho_0 S_K + K\gamma^4 L_0^2 \lambda^2\right).$$

**Proposition 5.** *Let $K \geq 1$ and $\delta \in (0,1)$. Assume that Assumptions 2 and 3 hold. Suppose that $\lambda \geq 8\sigma$, $\gamma L_0 \leq \frac{1}{64}$ and $\gamma L_1 \lambda \leq \frac{1}{64}$. Then, with probability at least $1-\delta$,*

$$\begin{aligned} \Delta_{K\wedge\widetilde{\tau}} + \frac{\gamma}{4} S_K \leq & \Delta_0 + 3\gamma K \frac{\sigma^{2\alpha}}{\lambda^{2\alpha-2}} + \left(8 + 4\sqrt{3}\right)\gamma^2 L_0 K \sigma^\alpha \lambda^{2-\alpha} \\ & + \left[\left(4 + 4\sqrt{3}\right)\gamma^2 L_0 \lambda^2 + 192\gamma\sigma^\alpha \lambda^{2-\alpha}\rho_0 + \frac{5}{2}\gamma\lambda H_0\right] \log\frac{1}{\delta}. \end{aligned}$$

**Proposition 6.** *Let $K \geq 1$, $\delta \in (0,1)$, and $\varepsilon > 0$. Assume that Assumptions 2 and 3 hold. Suppose that $\lambda \geq \max\{8\sigma, \varepsilon\}$, $\gamma L_0 \leq \frac{1}{64}$ and $\gamma L_1 \lambda \leq \frac{1}{64}$. Let*

$$\mathcal{B}_K := 3\gamma K \frac{\sigma^{2\alpha}}{\lambda^{2\alpha-2}} + \left(8 + 4\sqrt{3}\right) \gamma^2 L_0 K \sigma^\alpha \lambda^{2-\alpha} + \left[ \left(4 + 4\sqrt{3}\right) \gamma^2 L_0 \lambda^2 + 192\gamma\sigma^\alpha \lambda^{2-\alpha} \rho_0 + \frac{5}{2}\gamma\lambda H_0 \right] \log \frac{1}{\delta}.$$

*Suppose moreover that $\mathcal{B}_K \leq \Delta_0$ and $K > \frac{32\Delta_0}{\gamma\varepsilon^2}$. Then, with probability at least $1-\delta$, we have $\widetilde{\tau} > K$ and*

$$\frac{1}{K} \sum_{k=0}^{K-1} \|\nabla f(x_k)\| \leq \varepsilon.$$

**Derivation of the tuning scales.** For a fixed iteration budget $K$, we seek to choose $\lambda$ and $\gamma$ so that the accuracy $\varepsilon$ certified by Proposition 6 is as small as possible. The stepsize is constrained by the smoothness and localization conditions $\gamma L_0 \leq \frac{1}{64}$, $\gamma L_1 \lambda \leq \frac{1}{64}$ and $\mathcal{B}_K \leq \Delta_0$. The proposition then certifies any accuracy $\varepsilon$ satisfying $K > \frac{32\Delta_0}{\gamma\varepsilon^2}$, or equivalently

$$\varepsilon > \sqrt{\frac{32\Delta_0}{K\gamma}}. \tag{5}$$

For a fixed $\lambda$, this lower bound decreases as $\gamma$ increases. We therefore derive upper bounds on $\gamma$ from the three preceding conditions, substitute them into Eq. (5), and then minimize the resulting accuracy bound over $\lambda$.

The smoothness conditions give $\gamma \leq \frac{1}{64L_0}$, $\gamma \leq \frac{1}{64L_1\lambda}$, and hence

$$\varepsilon > 32\sqrt{2} \max\left\{ \sqrt{\frac{\Delta_0 L_0}{K}}, \sqrt{\frac{\Delta_0 L_1 \lambda}{K}} \right\}.$$

The condition $\mathcal{B}_K \leq \Delta_0$ implies

$$\begin{cases} \gamma \leq \dfrac{\Delta_0 \lambda^{2\alpha-2}}{3K\sigma^{2\alpha}}, \\ \gamma \leq \sqrt{\dfrac{\Delta_0}{\left(8+4\sqrt{3}\right) L_0 K \sigma^\alpha \lambda^{2-\alpha}}}, \\ \gamma \leq \sqrt{\dfrac{\Delta_0}{\left(4+4\sqrt{3}\right) L_0 \lambda^2 \log\frac{1}{\delta}}}, \\ \gamma \leq \dfrac{\Delta_0}{192\sigma^\alpha \lambda^{2-\alpha} \rho_0 \log\frac{1}{\delta}}, \\ \gamma \leq \dfrac{2\Delta_0}{5\lambda H_0 \log\frac{1}{\delta}}. \end{cases}$$

Substituting these bounds into Eq. (5) shows that, up to universal numerical factors,

$$\varepsilon \gtrsim \max\left\{ \frac{\sigma^\alpha}{\lambda^{\alpha-1}}, \left(\frac{\Delta_0 L_0 \sigma^\alpha \lambda^{2-\alpha}}{K}\right)^{\frac{1}{4}}, \left(\frac{\Delta_0 L_0 \lambda^2 \log\frac{1}{\delta}}{K^2}\right)^{\frac{1}{4}}, \sqrt{\frac{\sigma^\alpha \lambda^{2-\alpha} \rho_0 \log\frac{1}{\delta}}{K}}, \sqrt{\frac{\lambda H_0 \log\frac{1}{\delta}}{K}} \right\}.$$

We next remove the redundant scales. By the definition of $H_0$, $L_1\Delta_0 \leq \frac{\nu}{4} H_0$ and $L_0\Delta_0 \leq \frac{\nu}{4} H_0^2$. Then

$$\begin{cases} \sqrt{\dfrac{\Delta_0 L_1 \lambda}{K}} \lesssim \sqrt{\dfrac{\lambda H_0 \log\frac{e}{\delta}}{K}}, \\ \left(\dfrac{\Delta_0 L_0 \lambda^2 \log\frac{1}{\delta}}{K^2}\right)^{\frac{1}{4}} \lesssim \sqrt{\dfrac{\lambda H_0 \log\frac{e}{\delta}}{K}}. \end{cases}$$

It remains to simplify the term involving $\rho_0$. Since $\rho_0 = \max\left\{1, \frac{H_0}{\lambda}\right\} \le 1 + \frac{H_0}{\lambda}$ and $\lambda \ge 8\sigma$, we have

$$\begin{aligned}\sigma^\alpha \lambda^{2-\alpha} \rho_0 &\le \sigma^\alpha \lambda^{2-\alpha} + \sigma^\alpha H_0 \lambda^{1-\alpha} \\ &= \sigma^\alpha \lambda^{2-\alpha} + \lambda H_0 \left(\frac{\sigma}{\lambda}\right)^\alpha \\ &\le \sigma^\alpha \lambda^{2-\alpha} + \frac{\lambda H_0}{8^\alpha}.\end{aligned}$$

Consequently,

$$\sqrt{\frac{\sigma^\alpha \lambda^{2-\alpha} \rho_0 \log\frac{1}{\delta}}{K}} \le \sqrt{\frac{\sigma^\alpha \lambda^{2-\alpha} \log\frac{e}{\delta}}{K}} + \frac{1}{8^{\alpha/2}} \sqrt{\frac{\lambda H_0 \log\frac{e}{\delta}}{K}}.$$

The second term is already present in the accuracy bound. For the first term, observe that $\frac{\sigma^\alpha \lambda^{2-\alpha} \log\frac{e}{\delta}}{K} = \frac{\sigma^\alpha}{\lambda^{\alpha-1}} \frac{\lambda \log\frac{e}{\delta}}{K}$. If $\frac{\sigma^\alpha}{\lambda^{\alpha-1}} \le H_0$, then $\sqrt{\frac{\sigma^\alpha \lambda^{2-\alpha} \log\frac{e}{\delta}}{K}} \le \sqrt{\frac{\lambda H_0 \log\frac{e}{\delta}}{K}}$. If instead $\frac{\sigma^\alpha}{\lambda^{\alpha-1}} > H_0$, then, on the event of Proposition 5, the condition $\mathcal{B}_K \le \Delta_0$ implies $\widetilde{\tau} > K$ by the stopping-time argument. Consequently, $\frac{1}{K}\sum_{k=0}^{K-1} \|\nabla f(x_k)\| \le H_0 < \frac{\sigma^\alpha}{\lambda^{\alpha-1}}$. Thus, in either case, the term involving $\rho_0$ introduces no additional tuning scale.

Combining the preceding estimates, it suffices, up to universal numerical factors, to consider

$$\varepsilon \gtrsim \max\left\{\sqrt{\frac{\Delta_0 L_0}{K}}, \frac{\sigma^\alpha}{\lambda^{\alpha-1}}, \sqrt{\frac{\lambda H_0 \log\frac{e}{\delta}}{K}}, \left(\frac{\Delta_0 L_0 \sigma^\alpha \lambda^{2-\alpha}}{K}\right)^{\frac{1}{4}}\right\}. \tag{6}$$

Before optimizing over $\lambda$, we incorporate the condition $\lambda \ge \max\{8\sigma, \varepsilon\}$ required by Proposition 6. In view of Eq. (6), the condition $\varepsilon \le \lambda$ is ensured if every term in the maximum is at most $\lambda$.

The first term requires $\lambda \ge \sqrt{\frac{\Delta_0 L_0}{K}}$. The second is automatically controlled by $\lambda \ge 8\sigma$, since $\frac{\sigma^\alpha}{\lambda^{\alpha-1}} = \lambda\left(\frac{\sigma}{\lambda}\right)^\alpha \le \frac{\lambda}{8^\alpha}$. The third term is at most $\lambda$ whenever $\lambda \ge \frac{H_0 \log\frac{e}{\delta}}{K}$. Finally, the fourth term introduces no additional condition. Indeed, if $\lambda \ge \sqrt{\frac{\Delta_0 L_0}{K}}$ and $\lambda \ge 8\sigma$, then $\frac{\Delta_0 L_0 \sigma^\alpha \lambda^{2-\alpha}}{K} \le \lambda^2 \left(\frac{\lambda}{8}\right)^\alpha \lambda^{2-\alpha} = \frac{\lambda^4}{8^\alpha}$, and hence $\left(\frac{\Delta_0 L_0 \sigma^\alpha \lambda^{2-\alpha}}{K}\right)^{\frac{1}{4}} \le \frac{\lambda}{8^{\alpha/4}}$.

We are therefore led to the feasible range

$$\lambda \ge \underline{\lambda}_K, \qquad \underline{\lambda}_K := \max\left\{8\sigma, \sqrt{\frac{\Delta_0 L_0}{K}}, \frac{H_0 \log\frac{e}{\delta}}{K}\right\}.$$

**Lemma 20.** *Let $K \ge 1$, $\delta \in (0,1)$, and $\alpha \in (1,2]$, and suppose that $\Delta_0 > 0$. Define*

$$\begin{cases} \lambda_K := \max\left\{\underline{\lambda}_K, \min\left\{\left(\frac{K\sigma^{2\alpha}}{H_0 \log\frac{e}{\delta}}\right)^{\frac{1}{2\alpha-1}}, \left(\frac{K\sigma^{3\alpha}}{\Delta_0 L_0}\right)^{\frac{1}{3\alpha-2}}\right\}\right\}, \\ \varepsilon_K := \max\left\{\sqrt{\frac{\Delta_0 L_0}{K}}, \frac{H_0 \log\frac{e}{\delta}}{K}, \left(\frac{H_0^{\alpha-1}\sigma^\alpha \log^{\alpha-1}\frac{e}{\delta}}{K^{\alpha-1}}\right)^{\frac{1}{2\alpha-1}}, \left(\frac{(\Delta_0 L_0)^{\alpha-1}\sigma^\alpha}{K^{\alpha-1}}\right)^{\frac{1}{3\alpha-2}}\right\}. \end{cases}$$

*When $L_0 = 0$, the quantity $\left(\frac{K\sigma^{3\alpha}}{\Delta_0 L_0}\right)^{\frac{1}{3\alpha-2}}$ is interpreted as $+\infty$.*

*Then*

$$\begin{aligned}\varepsilon_K &\le \inf_{\lambda \ge \underline{\lambda}_K} \max\left\{\sqrt{\frac{\Delta_0 L_0}{K}}, \frac{\sigma^\alpha}{\lambda^{\alpha-1}}, \sqrt{\frac{\lambda H_0 \log\frac{e}{\delta}}{K}}, \left(\frac{\Delta_0 L_0 \sigma^\alpha \lambda^{2-\alpha}}{K}\right)^{\frac{1}{4}}\right\} \\ &= \max\left\{\sqrt{\frac{\Delta_0 L_0}{K}}, \frac{\sigma^\alpha}{\lambda_K^{\alpha-1}}, \sqrt{\frac{\lambda_K H_0 \log\frac{e}{\delta}}{K}}, \left(\frac{\Delta_0 L_0 \sigma^\alpha \lambda_K^{2-\alpha}}{K}\right)^{\frac{1}{4}}\right\} \\ &\le 2\sqrt{2}\,\varepsilon_K.\end{aligned}$$

*In particular, $\lambda_K$ is an exact minimizer.*

**Choice of parameters.** By Lemma 20, we choose $\lambda := \lambda_K$ and denote

$$E_K := \max\left\{\sqrt{\frac{\Delta_0 L_0}{K}}, \frac{\sigma^\alpha}{\lambda_K^{\alpha-1}}, \sqrt{\frac{\lambda_K H_0 \log\frac{e}{\delta}}{K}}, \left(\frac{\Delta_0 L_0 \sigma^\alpha \lambda_K^{2-\alpha}}{K}\right)^{\frac{1}{4}}\right\}.$$

Then

$$\varepsilon_K \le E_K \le 2\sqrt{2}\,\varepsilon_K.$$

By Eq. (5), the relevant stepsize scale is $\gamma \asymp \frac{\Delta_0}{K E_K^2}$. To simultaneously control the term involving $\rho_0$ in the localization budget, we set

$$\gamma := c \min\left\{\frac{\Delta_0}{K E_K^2}, \frac{\Delta_0}{\sigma^\alpha \lambda_K^{2-\alpha} \rho_0 \log\frac{e}{\delta}}\right\},$$

where the second quantity is interpreted as $+\infty$ when $\sigma = 0$, and $c > 0$ is a universal constant to be fixed below.

Since $\gamma \le c\frac{\Delta_0}{K E_K^2}$ and $E_K^2 \ge \frac{\Delta_0 L_0}{K}$, we have

$$\gamma L_0 \le c\frac{\Delta_0 L_0}{K E_K^2} \le c.$$

Similarly, since $E_K^2 \ge \frac{\lambda_K H_0 \log\frac{e}{\delta}}{K}$, and $L_1 \Delta_0 \le \frac{\nu}{4} H_0$, we obtain

$$\gamma L_1 \lambda_K \le c\frac{\Delta_0 L_1 \lambda_K}{K E_K^2} \le c\frac{L_1 \Delta_0}{H_0 \log\frac{e}{\delta}} \le c\frac{L_1 \Delta_0}{H_0} \le \frac{c\nu}{4}.$$

We next verify the localization condition. Since $E_K^2 \ge \frac{\sigma^{2\alpha}}{\lambda_K^{2\alpha-2}}$, we have

$$3\gamma K \frac{\sigma^{2\alpha}}{\lambda_K^{2\alpha-2}} \le 3c\Delta_0.$$

Since $E_K^4 \ge \frac{\Delta_0 L_0 \sigma^\alpha \lambda_K^{2-\alpha}}{K}$, we have

$$(8+4\sqrt{3})\gamma^2 L_0 K \sigma^\alpha \lambda_K^{2-\alpha} \le (8+4\sqrt{3})c^2 \frac{\Delta_0^2 L_0 \sigma^\alpha \lambda_K^{2-\alpha}}{K E_K^4} \le (8+4\sqrt{3})c^2\Delta_0.$$

Since $E_K^4 \ge \frac{\lambda_K^2 H_0^2 \log^2\frac{e}{\delta}}{K^2}$, and $L_0\Delta_0 \le \frac{\nu}{4} H_0^2$, we have

$$(4+4\sqrt{3})\gamma^2 L_0 \lambda_K^2 \log\frac{1}{\delta} \le (4+4\sqrt{3})c^2 \frac{\Delta_0^2 L_0 \lambda_K^2 \log\frac{1}{\delta}}{K^2 E_K^4} \le \frac{\nu}{4}(4+4\sqrt{3})c^2\Delta_0.$$

By the definition of $\gamma$,

$$192\gamma\sigma^\alpha \lambda_K^{2-\alpha}\rho_0 \log\frac{1}{\delta} \le 192c\Delta_0.$$

Since $E_K^2 \ge \frac{\lambda_K H_0 \log\frac{e}{\delta}}{K}$, we have

$$\frac{5}{2}\gamma\lambda_K H_0 \log\frac{1}{\delta} \le \frac{5}{2}c\Delta_0 \frac{\lambda_K H_0 \log\frac{1}{\delta}}{K E_K^2} \le \frac{5}{2}c\Delta_0.$$

Consequently,

$$\mathcal{B}_K \le \left[\frac{395}{2}c + \left(8+4\sqrt{3}+\nu(1+\sqrt{3})\right)c^2\right]\Delta_0.$$

Since $\nu \in (0,1)$, choosing $c := \frac{1}{256}$ gives

$$\begin{cases}\gamma L_0 < \frac{1}{64},\\ \gamma L_1 \lambda_K < \frac{1}{64},\\ \mathcal{B}_K < \Delta_0.\end{cases}$$

We now give the parameter choices used in Theorem 4. Let $\ell_\delta := \log \frac{e}{\delta}$. Define $H_0 := \frac{2}{\nu}\left(L_1\Delta_0 + \sqrt{L_1^2\Delta_0^2 + \nu L_0\Delta_0}\right)$ and $\underline{\lambda}_K := \max\left\{8\sigma, \sqrt{\frac{\Delta_0 L_0}{K}}, \frac{H_0\ell_\delta}{K}\right\}$. The clipping level is

$$\lambda_K := \max\left\{\underline{\lambda}_K, \min\left\{\left(\frac{K\sigma^{2\alpha}}{H_0\ell_\delta}\right)^{\frac{1}{2\alpha-1}}, \left(\frac{K\sigma^{3\alpha}}{\Delta_0 L_0}\right)^{\frac{1}{3\alpha-2}}\right\}\right\}.$$

For this choice of $\lambda_K$, set

$$\begin{cases} E_K := \max\left\{\sqrt{\frac{\Delta_0 L_0}{K}}, \frac{\sigma^\alpha}{\lambda_K^{\alpha-1}}, \sqrt{\frac{\lambda_K H_0\ell_\delta}{K}}, \left(\frac{\Delta_0 L_0\sigma^\alpha\lambda_K^{2-\alpha}}{K}\right)^{\frac{1}{4}}\right\}, \\ \rho_0 := \max\left\{1, \frac{H_0}{\lambda_K}\right\}. \end{cases}$$

The stepsize is then chosen as

$$\gamma_K := \frac{1}{256}\min\left\{\frac{\Delta_0}{KE_K^2}, \frac{\Delta_0}{\sigma^\alpha\lambda_K^{2-\alpha}\rho_0\ell_\delta}\right\}.$$

The second balancing term defining $\lambda_K$ is understood as $+\infty$ when $L_0 = 0$, and the second quantity in the minimum defining $\gamma_K$ is understood as $+\infty$ when $\sigma = 0$.

**Theorem 4.** *Assume that Assumptions 2 and 3 hold. Let $\delta \in (0,1)$, $K \geq 1$, and suppose that $\Delta_0 > 0$. Run* Clip-SGD *with the clipping level and stepsize specified above. Then, with probability at least $1-\delta$,*

$$\frac{1}{K}\sum_{k=0}^{K-1}\|\nabla f(x_k)\| \leq 4704\sqrt{2}\max\left\{\sqrt{\frac{L_0\Delta_0}{K}}, \frac{\frac{4}{\nu}\left(L_1\Delta_0 + \sqrt{L_0\Delta_0}\right)\log\frac{e}{\delta}}{K}, \left(\frac{\left(\frac{4}{\nu}\right)^{\alpha-1}\left(L_1\Delta_0 + \sqrt{L_0\Delta_0}\right)^{\alpha-1}\sigma^\alpha\log^{\alpha-1}\frac{e}{\delta}}{K^{\alpha-1}}\right)^{\frac{1}{2\alpha-1}}, \left(\frac{(L_0\Delta_0)^{\alpha-1}\sigma^\alpha}{K^{\alpha-1}}\right)^{\frac{1}{3\alpha-2}}\right\}.$$

### D.2 Comparison with prior work

We compare the nonconvex guarantees using the averaged gradient norm

$$\frac{1}{K}\sum_{k=0}^{K-1}\|\nabla f(x_k)\|.$$

When a prior result controls the averaged squared-gradient norm, we apply Cauchy–Schwarz:

$$\frac{1}{K}\sum_{k=0}^{K-1}\|\nabla f(x_k)\| \leq \left(\frac{1}{K}\sum_{k=0}^{K-1}\|\nabla f(x_k)\|^2\right)^{1/2}.$$

Thus, the comparisons below concern guarantees for the averaged gradient norm.

Let $\Delta_0 := f(x_0) - f^\star$, $G_0 := L_1\Delta_0 + \sqrt{L_0\Delta_0}$ and $\ell_\delta := \log\frac{e}{\delta}$. For the prescribed parameter choices, Theorem 4 gives, with probability at least $1-\delta$,

$$\frac{1}{K}\sum_{k=0}^{K-1}\|\nabla f(x_k)\| \leq \mathcal{O}\left(\max\left\{\sqrt{\frac{L_0\Delta_0}{K}}, \frac{G_0\ell_\delta}{K}, \left(\frac{G_0^{\alpha-1}\sigma^\alpha\ell_\delta^{\alpha-1}}{K^{\alpha-1}}\right)^{\frac{1}{2\alpha-1}}, \left(\frac{(L_0\Delta_0)^{\alpha-1}\sigma^\alpha}{K^{\alpha-1}}\right)^{\frac{1}{3\alpha-2}}\right\}\right).$$

where $\mathcal{O}$ hides only universal numerical constants. For fixed problem parameters satisfying $L_0\Delta_0 > 0$ and $\sigma > 0$, and fixed $\delta$, the asymptotically dominant term is

$$\left(\frac{(L_0\Delta_0)^{\alpha-1}\sigma^\alpha}{K^{\alpha-1}}\right)^{\frac{1}{3\alpha-2}}.$$

This term contains no explicit dependence on $L_1$ or $\delta$. The full bound and the parameter choices retain such dependencies through the other terms.

**Comparison with Nguyen et al. (2023) in the nonconvex case ($L_1 = 0$).** In the classical $L$-smooth setting, Nguyen et al. (2023) establish a time-optimal high-probability guarantee for Clip-SGD under bounded central $\alpha$-th moment noise. Applying Cauchy–Schwarz to their averaged squared-gradient bound gives the optimal stochastic dependence $K^{-\frac{\alpha-1}{3\alpha-2}}$ for the averaged gradient norm. To compare the problem-parameter dependence, suppose that $L_0\Delta_0 > 0$ and $\sigma > 0$, and set $z := \sigma/\sqrt{L_0\Delta_0}$. Ignoring logarithmic factors in $\delta$, their stochastic terms with this $K$-dependence scale as $\sqrt{L_0\Delta_0}\,K^{-\frac{\alpha-1}{3\alpha-2}} \max\left\{z^{\frac{1}{2}}, z^{\frac{\alpha}{2(\alpha-1)}}\right\}$, whereas our leading stochastic term is $\sqrt{L_0\Delta_0}\,K^{-\frac{\alpha-1}{3\alpha-2}} z^{\frac{\alpha}{3\alpha-2}}$. Since, for every $\alpha \in (1,2]$, $\frac{1}{2} \le \frac{\alpha}{3\alpha-2} < \frac{\alpha}{2(\alpha-1)}$, our leading term has uniformly no worse joint polynomial dependence on $L_0\Delta_0$ and $\sigma$. For $\alpha \in (1,2)$, the improvement is strict whenever $z \neq 1$, with equality at $z = 1$. For $\alpha = 2$, the dependence coincides when $0 < z \le 1$ and is strictly better when $z > 1$. Our leading term also contains no logarithmic dependence on $\delta$. These comparisons concern the leading stochastic terms; the full bounds also contain faster-decaying terms.

**Comparison with An et al. (2025) in the nonconvex case.** An et al. (2025) also study Clip-SGD under heavy-tailed noise and $(L_0, L_1)$-smoothness. After applying Cauchy–Schwarz to their averaged squared-gradient bound, their stated rate becomes $\mathcal{O}\left(K^{-\frac{\alpha-1}{3\alpha-2}} \log^{\frac{\alpha-1}{2\alpha-1}} \frac{K}{\delta}\right)$, with the problem-parameter dependence suppressed. To make this dependence explicit, we use the quantities $R$ and $\mathcal{R}$ introduced in Appendix B.1 of An et al. (2025), just below their Eq. (15) and just after their Eq. (19), respectively. Writing their initial objective gap as $\Delta_0$, these are $R := \frac{L_0+(4L_1+1)\Delta_0}{2}$ and $\mathcal{R} := L_0 + 2(L_1+1)R$. We also set $\rho_K := \log(4K/\delta)$. Substituting their clipping and stepsize choices in Eq. (32) into their descent bound in Eq. (33), and applying Cauchy–Schwarz, gives $\frac{1}{K}\sum_{k=0}^{K-1} \|\nabla f(x_k)\| \le 16\sqrt{\mathcal{R}\rho_K\lambda}\,K^{-\frac{2\alpha-1}{2(3\alpha-2)}}$. Writing $a := (\alpha-1)/(3\alpha-2)$ and $b := (2\alpha-1)/(2(3\alpha-2))$, expanding their displayed clipping threshold gives the full parameter-dependent expression

$$\mathcal{O}\left(\max\left\{\frac{\sqrt{\mathcal{R}R\rho_K}}{K^b}, \frac{(\mathcal{R}\rho_K)^{\frac{\alpha-1}{2\alpha-1}}\sigma^{\frac{\alpha}{2\alpha-1}}}{K^a}, \frac{\mathcal{R}^{\frac{\alpha-1}{2\alpha}}\Delta_0^{\frac{1}{2\alpha}}\sigma^{1/2}\rho_K^{\frac{1}{2}-\frac{1}{\alpha}}}{K^a}, \frac{\sqrt{\mathcal{R}\sigma}\left(1+\frac{L_1\Delta_0}{32\mathcal{R}}\right)^{1/\alpha}\rho_K^{\frac{1}{2}-\frac{1}{\alpha}}}{K^a}\right\}\right).$$

Here only universal numerical constants are suppressed. For fixed problem parameters with $\sigma > 0$ and fixed $\delta$, the first stochastic term eventually dominates, giving $\mathcal{O}\left(\frac{(\mathcal{R}\rho_K)^{\frac{\alpha-1}{2\alpha-1}}\sigma^{\frac{\alpha}{2\alpha-1}}}{K^{\frac{\alpha-1}{3\alpha-2}}}\right)$. Thus, their leading stochastic prefactor retains an $L_1$-dependence through $\mathcal{R}$. Our corresponding term, $\left(\frac{(L_0\Delta_0)^{\alpha-1}\sigma^\alpha}{K^{\alpha-1}}\right)^{\frac{1}{3\alpha-2}}$, has the same polynomial dependence on $K$, contains no explicit $L_1$ dependence, and has no logarithmic factors in $K$ or $1/\delta$. This comparison concerns the asymptotically dominant stochastic terms for fixed confidence and fixed problem parameters with $L_0\Delta_0 > 0$ and $\sigma > 0$. The guarantees also differ in the noiseless regime. Formally specializing their displayed parameter choices and bound to $\sigma = 0$ gives $\lambda = 4R$ and the expression $\sqrt{\mathcal{R}R\rho_K}\,K^{-\frac{2\alpha-1}{2(3\alpha-2)}}$. For every $\alpha > 1$, this decays more slowly than $K^{-1/2}$, so their prescribed tuning does not recover the deterministic generalized-smoothness rate. Our bound reduces to $\mathcal{O}\left(\sqrt{\frac{L_0\Delta_0}{K}} + \frac{(L_1\Delta_0+\sqrt{L_0\Delta_0})\log(e/\delta)}{K}\right)$, recovering the deterministic rate up to a confidence logarithm in the faster-decaying transient term.

**Comparison with Liu & Zhou (2025).** Liu & Zhou (2025, Theorem 3.2) establish an in-expectation guarantee for normalized SGD with momentum under $(L_0, L_1)$-smoothness and heavy-tailed noise. With batch size one, their leading stochastic term matches ours in its dependence on $K, L_0, \Delta_0, \sigma$ and contains no explicit dependence on $L_1$. Their stated bound additionally contains the faster-decaying term $\sigma/K^{(\alpha-1)/(2\alpha-1)}$. Our result establishes the same leading stochastic dependence with high probability for standard Clip-SGD.

**Dependence on the initial objective gap.** The nonconvex comparisons above are made at fixed $\Delta_0 = f(x_0) - f^\star$. Under $(L_0, L_1)$-smoothness, this gap can grow exponentially with $L_1 R_0$. Consequently, the absence of explicit $L_1$-dependence in our leading stochastic term does not imply

Table 5: Comparison of deterministic and high-probability convergence guarantees for first-order nonconvex optimization under $L$-smoothness and $(L_0, L_1)$-smoothness (Assumption 2). All rates are stated for the averaged gradient norm $\frac{1}{K}\sum_{k=0}^{K-1}\|\nabla f(x_k)\|$. We use $G_0 := L_1\Delta_0 + \sqrt{L_0\Delta_0}$ and $\ell_\delta := \log(e/\delta)$.

| Reference | Smoothness | Heavy-tailed noise | Rate |
|---|---|---|---|
| Vankov et al. (2025)[1] | $(L_0, L_1)$ | ✗ | $\mathcal{O}\left(\sqrt{\frac{L_0\Delta_0}{K}} + \frac{L_1\Delta_0}{K}\right)$ |
| Nguyen et al. (2023)[2] | $L$-smooth | ✓ | $\mathcal{O}\left((L_0\Delta_0)^{\frac{1}{4}}\sqrt{\ell_\delta}\max\left\{\left(\frac{\ell_\delta}{\sqrt{L_0\Delta_0}}\right)^{\frac{1}{2(\alpha-1)}}\frac{\sigma^{\frac{\alpha}{2(\alpha-1)}}}{K^{\frac{\alpha-1}{3\alpha-2}}}, \frac{(L_0\Delta_0)^{\frac{1}{4}}}{K^{\frac{2\alpha-1}{2(3\alpha-2)}}}, \frac{\sqrt{\sigma}}{K^{\frac{\alpha-1}{3\alpha-2}}}\right\}\right)$ |
| An et al. (2025) | $(L_0, L_1)$ | ✓ | $\mathcal{O}\left(\frac{\log^{\frac{\alpha-1}{2\alpha-1}}(K/\delta)}{K^{\frac{\alpha-1}{3\alpha-2}}}\right)$ [3] |
| **This work** | $(L_0, L_1)$ | ✓ | $\mathcal{O}\left(\max\left\{\sqrt{\frac{L_0\Delta_0}{K}}, \frac{G_0\ell_\delta}{K}, \left(\frac{G_0^{\alpha-1}\sigma^\alpha\ell_\delta^{\alpha-1}}{K^{\alpha-1}}\right)^{\frac{1}{2\alpha-1}}, \left(\frac{(L_0\Delta_0)^{\alpha-1}\sigma^\alpha}{K^{\alpha-1}}\right)^{\frac{1}{3\alpha-2}}\right\}\right)$ |

[1] Deterministic setting. The result of Vankov et al. (2025) is written for the averaged gradient norm to match our stationarity criterion.
[2] Nguyen et al. (2023) consider the $L$-smooth case, corresponding to $L_1 = 0$ and $L = L_0$. Their averaged squared-gradient bound is converted using Cauchy–Schwarz. We replace $\max\{\log(1/\delta), 1\}$ by the equivalent quantity $\log(e/\delta)$ up to universal constants. For a parameter-level comparison with our result, see the comparison with Nguyen et al. (2023).
[3] The displayed rate of An et al. (2025) is their stated stochastic iteration rate, converted from the averaged squared-gradient criterion using Cauchy–Schwarz. The dependence on $L_0$, $L_1$, $\Delta_0$, and $\sigma$ is suppressed. The full parameter-dependent expression and the formal noiseless specialization of their prescribed tuning are given in the comparison with An et al. (2026).

polynomial dependence on $R_0$. The distance-based guarantee avoiding exponential dependence on $L_1 R_0$ is specific to our convex results.

### D.3 Proofs

*Proof of Lemma 17.* By Lemma 9, we distinguish two cases.

- If $\|\nabla f(x_k)\| \le \lambda$, then

$$\begin{aligned} -\langle \nabla f(x_k), \mathbb{E}[e_k \mid \mathcal{F}_k]\rangle &\le \frac{1}{8}\|\nabla f(x_k)\| \min\{\|\nabla f(x_k)\|, \lambda\} + \|\nabla f(x_k)\| \frac{\sigma^\alpha}{\lambda^{\alpha-1}} \\ &\le \frac{1}{4}\|\nabla f(x_k)\| \min\{\|\nabla f(x_k)\|, \lambda\} + 2\frac{\sigma^{2\alpha}}{\lambda^{2\alpha-2}}. \end{aligned}$$

  The first inequality uses $\frac{\sigma}{\lambda} \le \frac{1}{8}$. For the second inequality, since $\|\nabla f(x_k)\| \le \lambda$, we have $\|\nabla f(x_k)\| \min\{\|\nabla f(x_k)\|, \lambda\} = \|\nabla f(x_k)\|^2$, and Young's inequality gives

$$\|\nabla f(x_k)\| \frac{\sigma^\alpha}{\lambda^{\alpha-1}} \le \frac{1}{8}\|\nabla f(x_k)\|^2 + 2\frac{\sigma^{2\alpha}}{\lambda^{2\alpha-2}}.$$

- If $\|\nabla f(x_k)\| > \lambda$, then

$$\begin{aligned} -\langle \nabla f(x_k), \mathbb{E}[e_k \mid \mathcal{F}_k]\rangle &\le \frac{1}{8}\|\nabla f(x_k)\| \min\{\|\nabla f(x_k)\|, \lambda\} + \lambda\|\nabla f(x_k)\| \frac{\sigma^\alpha}{\lambda^\alpha} \\ &\le \frac{1}{4}\|\nabla f(x_k)\| \min\{\|\nabla f(x_k)\|, \lambda\}, \end{aligned}$$

  where we used $\lambda\|\nabla f(x_k)\| = \|\nabla f(x_k)\| \min\{\|\nabla f(x_k)\|, \lambda\}$ and $(\sigma/\lambda)^\alpha \le 1/8$.

□

*Proof of Lemma 18.* Fix $k \ge 0$. We have

$$L_1\|x_{k+1} - x_k\| = \gamma L_1 \|g_k\| \le \gamma L_1 \lambda \le 1.$$

By Lemma 1,

$$\begin{aligned} f(x_{k+1}) - f(x_k) &\le -\gamma\langle \nabla f(x_k), g_k\rangle + 2\gamma^2 (L_0 + L_1\|\nabla f(x_k)\|)\|g_k\|^2 \\ &= \Phi_k. \end{aligned}$$

We remark that $\langle \nabla f(x_k), q_k \rangle = \|\nabla f(x_k)\| \min\{\|\nabla f(x_k)\|, \lambda\}$. By the proof of Lemma 7,

$$\mathbb{E}[\|e_k\|^2 \mid \mathcal{F}_k] \leq 2^{2-\alpha}\sigma^\alpha \lambda^{2-\alpha} \leq 2\sigma^\alpha \lambda^{2-\alpha}.$$

Consequently,

$$\mathbb{E}[\|g_k\|^2 \mid \mathcal{F}_k] \leq 2\|q_k\|^2 + 2\mathbb{E}[\|e_k\|^2 \mid \mathcal{F}_k] \leq 2\min\{\|\nabla f(x_k)\|, \lambda\}^2 + 4\sigma^\alpha \lambda^{2-\alpha}.$$

Moreover, by Lemma 17,

$$-\langle \nabla f(x_k), \mathbb{E}[e_k \mid \mathcal{F}_k] \rangle \leq \frac{1}{4}\|\nabla f(x_k)\| \min\{\|\nabla f(x_k)\|, \lambda\} + 2\frac{\sigma^{2\alpha}}{\lambda^{2\alpha-2}}.$$

It follows that

$$\begin{aligned}\mathbb{E}[\Phi_k \mid \mathcal{F}_k] \leq &-\frac{3\gamma}{4}\|\nabla f(x_k)\| \min\{\|\nabla f(x_k)\|, \lambda\} + 2\gamma\frac{\sigma^{2\alpha}}{\lambda^{2\alpha-2}} \\ &+ 4\gamma^2 (L_0 + L_1\|\nabla f(x_k)\|) \min\{\|\nabla f(x_k)\|, \lambda\}^2 \\ &+ 8\gamma^2 (L_0 + L_1\|\nabla f(x_k)\|)\sigma^\alpha \lambda^{2-\alpha}.\end{aligned} \tag{7}$$

Since $(L_0 + L_1\|\nabla f(x_k)\|)\min\{\|\nabla f(x_k)\|, \lambda\}^2 \leq (L_0 + L_1\lambda)\|\nabla f(x_k)\| \min\{\|\nabla f(x_k)\|, \lambda\}$, the stepsize conditions imply

$$4\gamma^2 (L_0 + L_1\|\nabla f(x_k)\|)\min\{\|\nabla f(x_k)\|, \lambda\}^2 \leq \frac{\gamma}{8}\|\nabla f(x_k)\| \min\{\|\nabla f(x_k)\|, \lambda\}.$$

We now control the last term of Eq. (7) . If $\|\nabla f(x_k)\| \leq \lambda$, Young's inequality gives

$$\begin{aligned}8\gamma^2 L_1 \|\nabla f(x_k)\| \sigma^\alpha \lambda^{2-\alpha} &\leq \frac{\gamma}{8}\|\nabla f(x_k)\|^2 + 128\gamma^3 L_1^2 \sigma^{2\alpha}\lambda^{4-2\alpha} \\ &\leq \frac{\gamma}{8}\|\nabla f(x_k)\| \min\{\|\nabla f(x_k)\|, \lambda\} + \frac{\gamma}{32}\frac{\sigma^{2\alpha}}{\lambda^{2\alpha-2}},\end{aligned}$$

where the last inequality uses $\gamma L_1 \lambda \leq \frac{1}{64}$.

If $\|\nabla f(x_k)\| > \lambda$, then

$$\begin{aligned}8\gamma^2 L_1 \|\nabla f(x_k)\| \sigma^\alpha \lambda^{2-\alpha} &= 8\gamma\|\nabla f(x_k)\| \lambda (\gamma L_1 \lambda)\left(\frac{\sigma}{\lambda}\right)^\alpha \\ &\leq \frac{\gamma}{64}\|\nabla f(x_k)\| \min\{\|\nabla f(x_k)\|, \lambda\}.\end{aligned}$$

Thus, in both cases,

$$8\gamma^2 L_1 \|\nabla f(x_k)\| \sigma^\alpha \lambda^{2-\alpha} \leq \frac{\gamma}{8}\|\nabla f(x_k)\| \min\{\|\nabla f(x_k)\|, \lambda\} + \frac{\gamma}{32}\frac{\sigma^{2\alpha}}{\lambda^{2\alpha-2}}.$$

We deduce that

$$\mathbb{E}[\Phi_k \mid \mathcal{F}_k] \leq -\frac{\gamma}{2}\|\nabla f(x_k)\| \min\{\|\nabla f(x_k)\|, \lambda\} + 3\gamma\frac{\sigma^{2\alpha}}{\lambda^{2\alpha-2}} + 8\gamma^2 L_0 \sigma^\alpha \lambda^{2-\alpha}.$$

It remains to prove the last two inequalities of the lemma. Conditionally on $\mathcal{F}_k$, define

$$\varphi_k(z) := -\gamma\langle \nabla f(x_k), z\rangle + 2\gamma^2 (L_0 + L_1\|\nabla f(x_k)\|)\|z\|^2.$$

For any $z, z'$ satisfying $\|z\|, \|z'\| \leq \lambda$,

$$\begin{aligned}|\varphi_k(z) - \varphi_k(z')| &\leq \left(\gamma\|\nabla f(x_k)\| + 4\gamma^2 (L_0 + L_1\|\nabla f(x_k)\|)\lambda\right)\|z - z'\| \\ &\leq \left(2\gamma\|\nabla f(x_k)\| + 4\gamma^2 L_0 \lambda\right)\|z - z'\|,\end{aligned}$$

where the last inequality uses $\gamma L_1 \lambda \leq \frac{1}{64}$.

Since $\Phi_k = \varphi_k(g_k)$ and $\|g_k\| \leq \lambda$,

$$|\Phi_k - \mathbb{E}[\Phi_k \mid \mathcal{F}_k]| \le 2\lambda \left(2\gamma \|\nabla f(x_k)\| + 4\gamma^2 L_0 \lambda\right)$$
$$= 4\gamma\lambda \|\nabla f(x_k)\| + 8\gamma^2 L_0 \lambda^2.$$

Since $\Phi_k = \varphi_k(g_k)$,

$$\begin{aligned}\mathrm{Var}(\Phi_k \mid \mathcal{F}_k) &\le \mathbb{E}\left[\left(\varphi_k(g_k) - \varphi_k(q_k)\right)^2 \middle| \mathcal{F}_k\right] \\ &\le \left(2\gamma \|\nabla f(x_k)\| + 4\gamma^2 L_0 \lambda\right)^2 \mathbb{E}[\|e_k\|^2 \mid \mathcal{F}_k] \\ &\le 2^{2-\alpha}\sigma^\alpha \lambda^{2-\alpha} \left(2\gamma \|\nabla f(x_k)\| + 4\gamma^2 L_0 \lambda\right)^2 \\ &\le 64\sigma^\alpha \lambda^{2-\alpha} \left(\gamma^2 \|\nabla f(x_k)\|^2 + \gamma^4 L_0^2 \lambda^2\right),\end{aligned}$$

where the third inequality uses

$$\mathbb{E}[\|e_k\|^2 \mid \mathcal{F}_k] \le 2^{2-\alpha}\sigma^\alpha \lambda^{2-\alpha},$$

established in the proof of Lemma 7, and the last inequality uses $2^{2-\alpha} \le 2$ and $(a+b)^2 \le 2a^2 + 2b^2$.

□

*Proof of Lemma 19.* We have $\Phi_k := -\gamma \langle \nabla f(x_k), g_k \rangle + 2\gamma^2 \left(L_0 + L_1 \|\nabla f(x_k)\|\right) \|g_k\|^2$. By Lemma 3,

$$|\Phi_k| 1_{\{k<\tilde{\tau}\}} \le \gamma\lambda H_0 + 2\gamma^2 L_0 \lambda^2 + 2\gamma\lambda \frac{1}{64} H_0 = \frac{33}{32}\gamma\lambda H_0 + 2\gamma^2 L_0 \lambda^2.$$

Similarly, we have

$$|\mathbb{E}[\Phi_k \mid \mathcal{F}_k]| 1_{\{k<\tilde{\tau}\}} \le \frac{33}{32}\gamma\lambda H_0 + 2\gamma^2 L_0 \lambda^2.$$

Therefore $|Z_k| \le \frac{5}{2}\gamma\lambda H_0 + 4\gamma^2 L_0 \lambda^2$ almost surely. We note that $\mathbf{1}_{\{k<\tilde{\tau}\}}$ is $\mathcal{F}_k$-measurable. By Lemma 18,

$$\begin{aligned}\sum_{k=0}^{K-1} \mathbb{E}[Z_k^2 \mid \mathcal{F}_k] &\le \sum_{k=0}^{K-1} \mathbf{1}_{\{k<\tilde{\tau}\}} \mathrm{Var}(\Phi_k \mid \mathcal{F}_k) \\ &\le 64\sigma^\alpha \lambda^{2-\alpha} \left(\gamma^2 \sum_{k=0}^{K-1} \mathbf{1}_{\{k<\tilde{\tau}\}} \|\nabla f(x_k)\|^2 + K\gamma^4 L_0^2 \lambda^2\right) \\ &\le 64\sigma^\alpha \lambda^{2-\alpha} \left(\gamma^2 \rho_0 S_K + K\gamma^4 L_0^2 \lambda^2\right).\end{aligned}$$

□

*Proof of Proposition 5.* By Lemma 19, $(Z_k)_{k\ge 0}$ is a martingale-difference sequence satisfying $|Z_k| \le \frac{5}{2}\gamma\lambda H_0 + 4\gamma^2 L_0 \lambda^2$ almost surely and

$$\sum_{k=0}^{K-1} \mathbb{E}[Z_k^2 \mid \mathcal{F}_k] \le 64\sigma^\alpha \lambda^{2-\alpha} \left(\gamma^2 \rho_0 S_K + K\gamma^4 L_0^2 \lambda^2\right).$$

Applying Proposition 2 with $\eta = \frac{\gamma}{4}$ gives, with probability at least $1-\delta$,

$$\begin{aligned}\sum_{k=0}^{K-1} Z_k \le& \frac{\gamma}{4} S_K + 192\gamma\sigma^\alpha \lambda^{2-\alpha} \rho_0 \log\frac{1}{\delta} + \sqrt{192 K\gamma^4 L_0^2 \sigma^\alpha \lambda^{4-\alpha} \log\frac{1}{\delta}} \\ &+ \left(\frac{5}{2}\gamma\lambda H_0 + 4\gamma^2 L_0 \lambda^2\right) \log\frac{1}{\delta}.\end{aligned}$$

Young's inequality gives

$$\sqrt{192K\gamma^4 L_0^2\sigma^\alpha\lambda^{4-\alpha}\log\frac{1}{\delta}} = 8\sqrt{3}\gamma^2 L_0\lambda\sqrt{K\sigma^\alpha\lambda^{2-\alpha}\log\frac{1}{\delta}} \\ \leq 4\sqrt{3}\gamma^2 L_0\left(K\sigma^\alpha\lambda^{2-\alpha}+\lambda^2\log\frac{1}{\delta}\right).$$

Therefore,

$$\sum_{k=0}^{K-1} Z_k \leq \frac{\gamma}{4}S_K + 4\sqrt{3}\gamma^2 L_0 K\sigma^\alpha\lambda^{2-\alpha} + \left(4+4\sqrt{3}\right)\gamma^2 L_0\lambda^2\log\frac{1}{\delta} \\ + 192\gamma\sigma^\alpha\lambda^{2-\alpha}\rho_0\log\frac{1}{\delta} + \frac{5}{2}\gamma\lambda H_0\log\frac{1}{\delta}.$$

On the other hand, by Lemma 18,

$$\begin{aligned}\Delta_{K\wedge\widetilde{\tau}} - \Delta_0 &\leq \sum_{k=0}^{K-1}\mathbf{1}_{\{k<\widetilde{\tau}\}}\Phi_k \\ &= \sum_{k=0}^{K-1}\mathbf{1}_{\{k<\widetilde{\tau}\}}\mathbb{E}[\Phi_k \mid \mathcal{F}_k] + \sum_{k=0}^{K-1} Z_k \\ &\leq -\frac{\gamma}{2}S_K + 3\gamma K\frac{\sigma^{2\alpha}}{\lambda^{2\alpha-2}} + 8\gamma^2 L_0 K\sigma^\alpha\lambda^{2-\alpha} + \sum_{k=0}^{K-1} Z_k.\end{aligned}$$

We deduce that

$$\Delta_{K\wedge\widetilde{\tau}} + \frac{\gamma}{4}S_K \leq \Delta_0 + 3\gamma K\frac{\sigma^{2\alpha}}{\lambda^{2\alpha-2}} + \left(8+4\sqrt{3}\right)\gamma^2 L_0 K\sigma^\alpha\lambda^{2-\alpha} \\ + \left(4+4\sqrt{3}\right)\gamma^2 L_0\lambda^2\log\frac{1}{\delta} + 192\gamma\sigma^\alpha\lambda^{2-\alpha}\rho_0\log\frac{1}{\delta} + \frac{5}{2}\gamma\lambda H_0\log\frac{1}{\delta}.$$

□

*Proof of Proposition 6.* We work on the event of Proposition 5, which has probability at least $1-\delta$. Since $\mathcal{B}_K \leq \Delta_0$, on this event,

$$\Delta_{K\wedge\widetilde{\tau}} + \frac{\gamma}{4}S_K \leq \Delta_0 + \mathcal{B}_K \leq 2\Delta_0.$$

If $\widetilde{\tau} \leq K$, then $\Delta_{K\wedge\widetilde{\tau}} = \Delta_{\widetilde{\tau}} > 2\Delta_0$, which is a contradiction. Hence $\widetilde{\tau} > K$, and therefore

$$S_K = \sum_{k=0}^{K-1}\|\nabla f(x_k)\|\min\{\|\nabla f(x_k)\|, \lambda\}.$$

For every $u \geq 0$,

$$u \leq \sqrt{u\min\{u,\lambda\}} + \frac{u\min\{u,\lambda\}}{\lambda}.$$

Indeed, the first term equals $u$ when $u \leq \lambda$, while the second term equals $u$ when $u > \lambda$. Applying this inequality with $u = \|\nabla f(x_k)\|$ and using Jensen's inequality gives

$$\begin{aligned}\frac{1}{K}\sum_{k=0}^{K-1}\|\nabla f(x_k)\| &\leq \frac{1}{K}\sum_{k=0}^{K-1}\sqrt{\|\nabla f(x_k)\|\min\{\|\nabla f(x_k)\|,\lambda\}} + \frac{S_K}{K\lambda} \\ &\leq \sqrt{\frac{S_K}{K}} + \frac{S_K}{K\lambda} \\ &\leq \sqrt{\frac{8\Delta_0}{\gamma K}} + \frac{8\Delta_0}{\gamma K\lambda}.\end{aligned}$$

Finally, since $K > \frac{32\Delta_0}{\gamma\varepsilon^2}$ and $\lambda \geq \varepsilon$, we obtain

$$\frac{1}{K}\sum_{k=0}^{K-1}\|\nabla f(x_k)\| < \frac{\varepsilon}{2} + \frac{\varepsilon^2}{4\lambda} \leq \frac{3\varepsilon}{4} \leq \varepsilon.$$

□

*Proof of Lemma 20.* For brevity, set $A := \sqrt{\frac{\Delta_0 L_0}{K}}$ and $B := \frac{H_0 \log\frac{e}{\delta}}{K}$. Then

$$\underline{\lambda}_K = \max\{8\sigma, A, B\}.$$

If $\sigma = 0$, then, by the definition of $\lambda_K$, $\lambda_K = \underline{\lambda}_K = \max\{A, B\}$. Hence

$$\inf_{\lambda \geq \underline{\lambda}_K} \max\{A, \sqrt{\lambda B}\} = \max\left\{A, \sqrt{\underline{\lambda}_K B}\right\} = \max\{A, B\} = \varepsilon_K.$$

Thus, the claim holds with equality. Suppose now that $\sigma > 0$, and define

$$F(\lambda) := \max\left\{A, \frac{\sigma^\alpha}{\lambda^{\alpha-1}}, \sqrt{\lambda B}, \left(A^2\sigma^\alpha\lambda^{2-\alpha}\right)^{\frac{1}{4}}\right\}.$$

Among the three $\lambda$-dependent terms, $\frac{\sigma^\alpha}{\lambda^{\alpha-1}}$, $\sqrt{\lambda B}$ and $\left(A^2\sigma^\alpha\lambda^{2-\alpha}\right)^{\frac{1}{4}}$, the first is decreasing in $\lambda$, whereas the other two are nondecreasing.

The first term balances the second when $\lambda = \left(\frac{\sigma^{2\alpha}}{B}\right)^{\frac{1}{2\alpha-1}}$, and balances the third when $\lambda = \left(\frac{\sigma^{3\alpha}}{A^2}\right)^{\frac{1}{3\alpha-2}}$, where the latter is interpreted as $+\infty$ when $A = 0$. Therefore, the maximum of the three terms is nonincreasing until the smaller of these two balancing points and nondecreasing afterward. Its unconstrained minimum is thus attained at

$$\bar{\lambda} := \min\left\{\left(\frac{\sigma^{2\alpha}}{B}\right)^{\frac{1}{2\alpha-1}}, \left(\frac{\sigma^{3\alpha}}{A^2}\right)^{\frac{1}{3\alpha-2}}\right\}.$$

The term $A$ is independent of $\lambda$, so the above point remains a minimizer after adding $A$. Since $F$ is nondecreasing on $[\bar{\lambda}, \infty)$, imposing $\lambda \geq \underline{\lambda}_K$ yields a constrained minimizer at

$$\max\left\{\underline{\lambda}_K, \bar{\lambda}\right\} = \lambda_K.$$

It remains to compare the constrained minimum with $\varepsilon_K$. We first note that the unconstrained minimum has a simple form.

Since the first term is decreasing, the smaller balancing point corresponds to the larger balancing value. Hence,

$$F(\bar{\lambda}) = \max\left\{A, \left(B^{\alpha-1}\sigma^\alpha\right)^{\frac{1}{2\alpha-1}}, \left(A^{2\alpha-2}\sigma^\alpha\right)^{\frac{1}{3\alpha-2}}\right\}.$$

Moreover, by the definitions of $A$, $B$, and $\varepsilon_K$,

$$\varepsilon_K = \max\left\{B, F(\bar{\lambda})\right\}.$$

Since $\bar{\lambda}$ is an unconstrained minimizer, $F(\lambda_K) \geq F(\bar{\lambda})$. Moreover, $\lambda_K \geq \underline{\lambda}_K \geq B$, so $F(\lambda_K) \geq \sqrt{\lambda_K B} \geq B$. Therefore,

$$F(\lambda_K) \geq \max\left\{B, F(\bar{\lambda})\right\} = \varepsilon_K.$$

It remains to bound $F(\lambda_K)$ from above. Since $\lambda_K \geq \bar{\lambda}$,

$$\frac{\sigma^\alpha}{\lambda_K^{\alpha-1}} \leq \frac{\sigma^\alpha}{\bar{\lambda}^{\alpha-1}} \leq F(\bar{\lambda}) \leq \varepsilon_K.$$

Next, since $\lambda_K = \max\left\{8\sigma, A, B, \bar{\lambda}\right\}$, we have $\sqrt{\lambda_K B} = \max\left\{\sqrt{8\sigma B}, \sqrt{AB}, B, \sqrt{\bar{\lambda} B}\right\}$. We remark that

$$\begin{cases} \sqrt{AB} \le \max\{A, B\} \le \varepsilon_K, \\ \sqrt{\bar{\lambda} B} \le F(\bar{\lambda}) \le \varepsilon_K, \\ \sqrt{\sigma B} = B^{\frac{1}{2\alpha}} \left[\left(B^{\alpha-1}\sigma^{\alpha}\right)^{\frac{1}{2\alpha-1}}\right]^{\frac{2\alpha-1}{2\alpha}} \le \varepsilon_K. \end{cases}$$

Hence $\sqrt{\lambda_K B} \le 2\sqrt{2}\,\varepsilon_K$.

Finally, if $A = 0$, the last term vanishes. Otherwise, since $2 - \alpha \ge 0$,

$$\left(A^2\sigma^{\alpha}\lambda_K^{2-\alpha}\right)^{\frac{1}{4}} = \max\left\{ \left(A^2\sigma^{\alpha}(8\sigma)^{2-\alpha}\right)^{\frac{1}{4}}, \left(A^2\sigma^{\alpha}A^{2-\alpha}\right)^{\frac{1}{4}}, \left(A^2\sigma^{\alpha}B^{2-\alpha}\right)^{\frac{1}{4}}, \left(A^2\sigma^{\alpha}\bar{\lambda}^{2-\alpha}\right)^{\frac{1}{4}} \right\}.$$

The last term is bounded by $F(\bar{\lambda}) \le \varepsilon_K$, while the first three satisfy

$$\left(A^2\sigma^{\alpha}(8\sigma)^{2-\alpha}\right)^{\frac{1}{4}} = 8^{\frac{2-\alpha}{4}} A^{\frac{2-\alpha}{2\alpha}} \left[\left(A^{2\alpha-2}\sigma^{\alpha}\right)^{\frac{1}{3\alpha-2}}\right]^{\frac{3\alpha-2}{2\alpha}},$$

$$\left(A^2\sigma^{\alpha}A^{2-\alpha}\right)^{\frac{1}{4}} = A^{\frac{3(2-\alpha)}{4}} \left[\left(A^{2\alpha-2}\sigma^{\alpha}\right)^{\frac{1}{3\alpha-2}}\right]^{\frac{3\alpha-2}{4}},$$

$$\left(A^2\sigma^{\alpha}B^{2-\alpha}\right)^{\frac{1}{4}} = A^{\frac{2-\alpha}{2}} B^{\frac{2-\alpha}{4}} \left[\left(A^{2\alpha-2}\sigma^{\alpha}\right)^{\frac{1}{3\alpha-2}}\right]^{\frac{3\alpha-2}{4}}.$$

Therefore $\left(A^2\sigma^{\alpha}\lambda_K^{2-\alpha}\right)^{\frac{1}{4}} \le 8^{\frac{2-\alpha}{4}}\varepsilon_K \le 2\sqrt{2}\varepsilon_K$. Combining the above bounds gives

$$F(\lambda_K) \le 2\sqrt{2}\,\varepsilon_K.$$

We deduce that

$$\varepsilon_K \le \inf_{\lambda \ge \lambda_K} F(\lambda) = F(\lambda_K) \le 2\sqrt{2}\,\varepsilon_K,$$

□

*Proof of Theorem 4.* Set $\lambda := \lambda_K$ and $\gamma := \gamma_K$. By the preceding parameter verification, with $c = \frac{1}{256}$,

$$\begin{cases} \lambda \ge 8\sigma, \\ \gamma L_0 < \dfrac{1}{64}, \\ \gamma L_1 \lambda < \dfrac{1}{64}, \\ \mathcal{B}_K < \Delta_0. \end{cases}$$

We now apply Proposition 5. On an event of probability at least $1 - \delta$,

$$\Delta_{K \wedge \widetilde{\tau}} + \frac{\gamma}{4} S_K \le \Delta_0 + \mathcal{B}_K < 2\Delta_0.$$

If $\widetilde{\tau} \le K$, then, by the definition of $\widetilde{\tau}$, $\Delta_{\widetilde{\tau}} > 2\Delta_0$, which contradicts the preceding inequality. Thus $\widetilde{\tau} > K$. Since $\Delta_K \ge 0$, it follows that

$$S_K = \sum_{k=0}^{K-1} \|\nabla f(x_k)\| \min\left\{\|\nabla f(x_k)\|, \lambda\right\} \le \frac{8\Delta_0}{\gamma}.$$

For every $u \ge 0$, $u \le \sqrt{u \min\{u, \lambda\}} + \frac{u \min\{u,\lambda\}}{\lambda}$. Therefore, by Cauchy–Schwarz,

$$\begin{aligned} \frac{1}{K} \sum_{k=0}^{K-1} \|\nabla f(x_k)\| &\le \sqrt{\frac{S_K}{K}} + \frac{S_K}{K\lambda} \\ &\le \sqrt{\frac{8\Delta_0}{\gamma K}} + \frac{8\Delta_0}{\gamma K \lambda}. \end{aligned}$$

The only remaining difficulty is that $\gamma$ is defined as the minimum of two quantities. To obtain a bound of order $E_K$ from the preceding estimate, we need $\gamma$ to remain, up to a universal constant, of the scale $\Delta_0/(KE_K^2)$. The additional term introduced to control the $\rho_0$-dependent localization error may make $\gamma$ smaller. We show that either this term is of the same order as the first one, or the desired stationarity bound follows directly from localization. This leads to two cases.

*Case 1.* Suppose that $\frac{\sigma^\alpha}{\lambda^{\alpha-1}} \le H_0$. In this regime, the additional $\rho_0$-dependent term does not reduce the stepsize beyond the scale $\Delta_0/(KE_K^2)$. Indeed, since $\rho_0 \le 1 + \frac{H_0}{\lambda}$ and $\lambda \ge 8\sigma$,

$$\sigma^\alpha \lambda^{2-\alpha}\rho_0 \le \sigma^\alpha \lambda^{2-\alpha} + \sigma^\alpha H_0 \lambda^{1-\alpha} = \lambda \frac{\sigma^\alpha}{\lambda^{\alpha-1}} + \lambda H_0 \left(\frac{\sigma}{\lambda}\right)^\alpha \le \left(1 + \frac{1}{8^\alpha}\right)\lambda H_0 \le \frac{9}{8}\lambda H_0.$$

Since $E_K^2 \ge \frac{\lambda H_0 \log\frac{e}{\delta}}{K}$, we obtain $\sigma^\alpha \lambda^{2-\alpha}\rho_0 \log\frac{e}{\delta} \le \frac{9}{8} K E_K^2$. Hence, by the definition of $\gamma$,

$$\gamma \ge \frac{8c}{9}\frac{\Delta_0}{KE_K^2}.$$

We also need $E_K \le \lambda$. By the definition of $\lambda_K$, $\lambda \ge \max\left\{8\sigma, \sqrt{\frac{\Delta_0 L_0}{K}}, \frac{H_0 \log\frac{e}{\delta}}{K}\right\}$. Therefore,

$$\begin{cases} \sqrt{\dfrac{\Delta_0 L_0}{K}} \le \lambda, \\ \dfrac{\sigma^\alpha}{\lambda^{\alpha-1}} \le \dfrac{\lambda}{8^\alpha}, \\ \sqrt{\dfrac{\lambda H_0 \log\frac{e}{\delta}}{K}} \le \lambda, \\ \left(\dfrac{\Delta_0 L_0 \sigma^\alpha \lambda^{2-\alpha}}{K}\right)^{\frac{1}{4}} \le \dfrac{\lambda}{8^{\alpha/4}}, \end{cases}$$

and thus $E_K \le \lambda$. Substituting the lower bound on $\gamma$ into the preceding stationarity estimate gives $\sqrt{\frac{8\Delta_0}{\gamma K}} \le \frac{3}{\sqrt{c}} E_K$ and $\frac{8\Delta_0}{\gamma K \lambda} \le \frac{9}{c}\frac{E_K^2}{\lambda} \le \frac{9}{c} E_K$. Hence

$$\frac{1}{K}\sum_{k=0}^{K-1} \|\nabla f(x_k)\| \le \left(\frac{3}{\sqrt{c}} + \frac{9}{c}\right) E_K.$$

Since $c = \frac{1}{256}$,

$$\frac{1}{K}\sum_{k=0}^{K-1} \|\nabla f(x_k)\| \le 2352 E_K.$$

*Case 2.* Suppose instead that $\frac{\sigma^\alpha}{\lambda^{\alpha-1}} > H_0$. In this regime, we do not need to control the stepsize from below: localization already gives the desired stationarity bound. Indeed, since $\widetilde{\tau} > K$, we have $\Delta_k \le 2\Delta_0$ for every $k = 0, \dots, K-1$. By the localized gradient bound,

$$\|\nabla f(x_k)\| \le H_0 < \frac{\sigma^\alpha}{\lambda^{\alpha-1}} \le E_K.$$

Therefore,

$$\frac{1}{K}\sum_{k=0}^{K-1} \|\nabla f(x_k)\| \le E_K.$$

Combining the two cases,

$$\frac{1}{K}\sum_{k=0}^{K-1} \|\nabla f(x_k)\| \le 2352 E_K.$$

By Lemma 20,

$$E_K \le 2\sqrt{2}\,\max\left\{\sqrt{\frac{L_0\Delta_0}{K}}, \frac{H_0 \log\frac{e}{\delta}}{K}, \left(\frac{H_0^{\alpha-1}\sigma^\alpha \log^{\alpha-1}\frac{e}{\delta}}{K^{\alpha-1}}\right)^{\frac{1}{2\alpha-1}}, \left(\frac{(L_0\Delta_0)^{\alpha-1}\sigma^\alpha}{K^{\alpha-1}}\right)^{\frac{1}{3\alpha-2}}\right\}.$$

Therefore,

$$\frac{1}{K}\sum_{k=0}^{K-1}\|\nabla f(x_k)\| \le 4704\sqrt{2}\max\left\{\sqrt{\frac{L_0\Delta_0}{K}}, \frac{H_0\log\frac{e}{\delta}}{K}, \left(\frac{H_0^{\alpha-1}\sigma^\alpha\log^{\alpha-1}\frac{e}{\delta}}{K^{\alpha-1}}\right)^{\frac{1}{2\alpha-1}}, \left(\frac{(L_0\Delta_0)^{\alpha-1}\sigma^\alpha}{K^{\alpha-1}}\right)^{\frac{1}{3\alpha-2}}\right\}.$$

It remains only to eliminate $H_0$. By its definition, $H_0 = \frac{2}{\nu}\left(L_1\Delta_0 + \sqrt{L_1^2\Delta_0^2 + \nu L_0\Delta_0}\right)$. Using $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ and $\nu \le 1$, we obtain

$$H_0 \le \frac{2}{\nu}\left(2L_1\Delta_0 + \sqrt{\nu L_0\Delta_0}\right) \le \frac{4}{\nu}\left(L_1\Delta_0 + \sqrt{L_0\Delta_0}\right).$$

We deduce that

$$\frac{1}{K}\sum_{k=0}^{K-1}\|\nabla f(x_k)\| \le 4704\sqrt{2}\max\left\{\sqrt{\frac{L_0\Delta_0}{K}}, \frac{\frac{4}{\nu}\left(L_1\Delta_0 + \sqrt{L_0\Delta_0}\right)\log\frac{e}{\delta}}{K}, \left(\frac{\left(\frac{4}{\nu}\right)^{\alpha-1}\left(L_1\Delta_0 + \sqrt{L_0\Delta_0}\right)^{\alpha-1}\sigma^\alpha\log^{\alpha-1}\frac{e}{\delta}}{K^{\alpha-1}}\right)^{\frac{1}{2\alpha-1}}, \left(\frac{(L_0\Delta_0)^{\alpha-1}\sigma^\alpha}{K^{\alpha-1}}\right)^{\frac{1}{3\alpha-2}}\right\},$$

□

# E A Lower Bound for Fixed-Parameter Clip-SGD in the Convex Setting

We establish complementary lower bounds for the dependence on $L_1R_0$ and on the failure probability. Together, under the iteration conditions of Theorem 3, they match the convex stochastic upper bound up to logarithmic factors in the iteration budget for fixed-parameter Clip-SGD. Here, $R_0$ denotes a common upper bound on the initial distance to a minimizer. The lower bound applies to any possibly randomized output rule measurable with respect to the observed iterates and stochastic gradients. The hard-function family below is the one used in Crawshaw & Liu (2025, proof of Lemma 8). For the construction below, we fix $L_0 = 1$; the resulting objectives are therefore $(1, L_1)$-smooth.

**Lemma 21.** *Let $L_1 > 0$ and $a > 0$. Define*

$$\phi_a(x) := \int_0^{|x|} \min\left\{a, \frac{e^{L_1 t} - 1}{L_1}\right\} dt, \qquad r_a := \frac{\log(1 + L_1 a)}{L_1}.$$

*Then $\phi_a$ is convex and continuously differentiable, with minimum $\phi_a(0) = 0$, and, for every $x < y$, $|\phi_a'(y) - \phi_a'(x)| \le \left(1 + L_1 \sup_{u\in[x,y]}|\phi_a'(u)|\right)(y - x)$. Moreover,*

$$\phi_a'(x) = \begin{cases} \operatorname{sign}(x)\dfrac{e^{L_1|x|} - 1}{L_1}, & |x| < r_a, \\ a\,\operatorname{sign}(x), & |x| \ge r_a, \end{cases}$$

*so $|\phi_a'(x)| \le a$. In addition, $r_a \le a$ and $\phi_a(x) \ge a(|x| - r_a)_+$, where $(v)_+ := \max\{v, 0\}$.*

For $i \in \{1, 2\}$, let

$$\mathcal{T}_i := \left(x_0^{(i)}, G_0^{(i)}, x_1^{(i)}, G_1^{(i)}, \ldots, x_{K-1}^{(i)}, G_{K-1}^{(i)}, x_K^{(i)}\right)$$

denote the transcript observed by the algorithm on instance $i$, where $G_k^{(i)}$ is the stochastic gradient observed at $x_k^{(i)}$.

The hard function consists of a curved region around its minimizer and linear tails with slopes $\pm a$. The linear tails are useful because shifted copies of the function have exactly the same gradient on regions lying on the same side of both minimizers. Thus, while the two objectives have different minimizers, an algorithm whose trajectory remains in such a region may receive identical stochastic-gradient observations on both instances. The next lemma converts this indistinguishability into an objective lower bound.

**Lemma 22.** *Let $f_i(x) := \phi_a(x - u_i)$ for $i \in \{1, 2\}$, where $\Delta_u := |u_1 - u_2| > 2r_a$. Then $u_i$ is a minimizer of $f_i$ and $f_i^\star := f_i(u_i) = 0$.*

*Suppose the randomness of the two runs can be defined jointly so that each run has its original distribution and*

$$\mathbb{P}(\mathcal{T}_1 = \mathcal{T}_2) \geq q$$

*for some $q \in [0, 1]$.*

*Then, for any fixed possibly randomized output rule used on both instances, depending only on the observed transcript and on independent auxiliary randomness, let $x_K^{\mathrm{out},(i)}$ denote the point returned by this rule on instance $i$. There exists $i \in \{1, 2\}$ such that*

$$\mathbb{P}_i\left(f_i(x_K^{\mathrm{out},(i)}) - f_i^\star \geq a\left(\frac{\Delta_u}{2} - r_a\right)\right) \geq \frac{q}{2}.$$

*Here, $\mathbb{P}_i$ denotes probability under instance $i$, including all randomness used by the algorithm and the output rule.*

**Lemma 23.** *Let $f : \mathbb{R} \to \mathbb{R}$ be differentiable, with $|f'(x)| \leq a$ for all $x$. For $p \in (0, 1]$ and $Z \sim \mathrm{Bernoulli}(p)$, define*

$$G(x, Z) := \frac{f'(x)}{p} Z.$$

*Then $\mathbb{E}[G(x, Z)] = f'(x)$ and, for every $\alpha \in (1, 2]$,*

$$\mathbb{E}\left[|G(x, Z) - f'(x)|^\alpha\right] \leq 2a^\alpha p^{1-\alpha}.$$

*With fresh independent draws at successive iterations, these statements also hold conditionally on the past.*

### E.1 Dependence on the failure probability

To obtain the dependence on $\delta$, we consider two hard instances whose minimizers are located at $-R_0$ and $R_0$. We use the stochastic oracle

$$G(x, Z) = \frac{f'(x)}{p} Z, \qquad Z \sim \mathrm{Bernoulli}(p),$$

so that a nonzero stochastic gradient is observed only with probability $p$. On the event that all $K$ Bernoulli draws are zero, Clip-SGD receives the same sequence of zero gradients on both instances. Consequently, both trajectories remain at $x_0 = 0$ and the two observed transcripts are identical, regardless of the choices of $\gamma$ and $\lambda$.

We choose $p$ so that the all-zero event has probability greater than $2\delta$. Hence, with probability of order $\delta$, the algorithm cannot distinguish between two objectives whose minimizers are separated by $2R_0$. The indistinguishability argument then yields the desired high-probability lower bound.

**Lemma 24.** *Let $L_1, R_0, \sigma > 0$, $\alpha \in (1, 2]$, and $\delta \in (0, 1/4)$. Suppose the integer $K$ satisfies*

$$K \geq \log\left(\frac{1}{2\delta}\right) \max\left\{1, \left(\frac{\sigma}{R_0}\right)^{\alpha/(\alpha-1)}\right\}.$$

*For every fixed $\gamma, \lambda > 0$ and every fixed, possibly randomized, output rule depending only on the observed transcript and independent auxiliary randomness, let $x_K^{\mathrm{out}}$ denote the point returned by this rule. Then there exist a convex $(1, L_1)$-smooth function $f : \mathbb{R} \to \mathbb{R}$, with a minimizer $|x^\star| \leq R_0$, and an unbiased stochastic-gradient oracle satisfying Assumption 3, such that Clip-SGD, started from $x_0 = 0$ and using fresh independent oracle samples, satisfies*

$$\mathbb{P}\left(f(x_K^{\mathrm{out}}) - f^\star \geq \frac{\sigma R_0}{4}\left(\frac{\log(1/(2\delta))}{K}\right)^{(\alpha-1)/\alpha}\right) > \delta.$$

### E.2 Dependence on $L_1 R_0$

The proof separates the fixed choices of $\gamma$ and $\lambda$ according to whether the clipped step generated by a common tail gradient is large or small. If this update is large, we use two noiseless shifted copies of $\phi_H$ whose minimizers lie at different positive locations to the right of $x_0 = 0$. The shifts are chosen so that $x_0$ lies in the common left linear tail of both objectives, while the first update lands in their common right linear tail. Thus both runs observe the same negative gradient at $x_0$, make the same first update, then observe the same positive gradient and return to $x_0$. This two-step trajectory repeats, so the two transcripts are identical.

If instead the update is small, we consider two shifted copies of $\phi_a$ with minimizers at $R_0/2$ and $R_0$, equipped with Bernoulli oracles. We choose the sampling probability so that, with constant probability, the total displacement over $K$ iterations is less than $R_0/4$. We then choose the slope $a$ to satisfy the noise constraint and verify that the corresponding transition radius satisfies $r_a < R_0/4$. On this event, both trajectories remain in the common left linear tail of the two objectives, where their population gradients coincide. Using the same Bernoulli draws for both runs therefore makes their stochastic gradients, iterates, and hence their entire transcripts identical. In both constructions, the indistinguishability lemma converts this transcript equality into the desired objective lower bound.

**Lemma 25.** *Let $L_1, R_0, \sigma > 0$ and $\alpha \in (1, 2]$. Suppose the integer $K$ satisfies*

$$\begin{cases} K \geq \max\{1, L_1 R_0, \sigma L_1\}, \\ \dfrac{L_1 R_0}{\log(eK)} > 64. \end{cases}$$

*For every fixed $\gamma, \lambda > 0$ and every fixed possibly randomized output rule depending only on the observed transcript and independent auxiliary randomness, let $x_K^{\mathrm{out}}$ denote the point returned by this rule. There exist a convex $(1, L_1)$-smooth function $f : \mathbb{R} \to \mathbb{R}$, with a minimizer $x^\star \in [0, R_0]$, and an unbiased stochastic-gradient oracle satisfying Assumption 3, such that Clip-SGD, started from $x_0 = 0$ and using fresh independent oracle samples, satisfies*

$$\mathbb{P}\left( f(x_K^{\mathrm{out}}) - f^\star \geq \frac{\sigma R_0}{128} \left( \frac{L_1 R_0}{K \log(eK)} \right)^{(\alpha-1)/\alpha} \right) \geq \frac{1}{4}.$$

### E.3 Proofs

*Proof of Lemma 21.* Define $g$ on $[0, \infty)$ by $g(t) := \min\left\{a, \frac{e^{L_1 t} - 1}{L_1}\right\}$. Since $g$ is continuous, nonnegative, and nondecreasing, $\phi_a(x) = \int_0^{|x|} g(t)\, dt$ is continuously differentiable and convex. Since $g \geq 0$, $\phi_a(x) \geq 0 = \phi_a(0)$ for all $x$, so 0 is a global minimizer.

We have, $\phi_a'(x) = \mathrm{sign}(x)\, g(|x|)$ and by the definition of $r_a$, we also have $\frac{e^{L_1 r_a} - 1}{L_1} = a$. Therefore

$$\phi_a'(x) = \begin{cases} \mathrm{sign}(x) \dfrac{e^{L_1 |x|} - 1}{L_1}, & |x| < r_a, \\ a\ \mathrm{sign}(x), & |x| \geq r_a. \end{cases}$$

In particular, $|\phi_a'(x)| \leq a$.

To verify generalized smoothness, fix $x < y$ and split $[x, y]$, if necessary, at $-r_a$, 0, and $r_a$. On each resulting subinterval $[s, t]$, either $\phi_a'$ is constant, or

$$\frac{d}{du} \phi_a'(u) = e^{L_1 |u|} = 1 + L_1 |\phi_a'(u)|.$$

Hence, by the mean-value theorem,

$$|\phi_a'(t) - \phi_a'(s)| \leq \left( 1 + L_1 \sup_{u \in [s,t]} |\phi_a'(u)| \right) (t - s).$$

Summing these bounds over the subintervals and using $\sup_{u \in [s,t]} |\phi_a'(u)| \leq \sup_{u \in [x,y]} |\phi_a'(u)|$ gives

$$|\phi_a'(y) - \phi_a'(x)| \leq \left( 1 + L_1 \sup_{u \in [x,y]} |\phi_a'(u)| \right) (y - x).$$

Finally, using $\log(1+z) \le z$ for $z \ge 0$, we obtain $r_a = \frac{\log(1+L_1 a)}{L_1} \le a$.

It remains to prove the lower bound on $\phi_a$. If $|x| \le r_a$, then $a(|x| - r_a)_+ = 0 \le \phi_a(x)$. If $|x| > r_a$, then $g(t) = a$ for every $t \in [r_a, |x|]$, and therefore

$$\phi_a(x) = \int_0^{r_a} g(t)\, dt + \int_{r_a}^{|x|} a\, dt \ge a(|x| - r_a).$$

Hence, for every $x \in \mathbb{R}$,

$$\phi_a(x) \ge a(|x| - r_a)_+.$$

□

*Proof of Lemma 22.* Let $E := \{\mathcal{T}_1 = \mathcal{T}_2\}$. By assumption, $\mathbb{P}(E) \ge q$.

Let $\mu$ denote the distribution of the auxiliary randomness used by the randomized output rule. Since the same output rule is used on both instances, there exists a measurable function $\Psi$ such that, given a transcript $t$, the output can be represented as

$$\Psi(t, U), \qquad U \sim \mu.$$

Enlarge the probability space supporting the coupling $(\mathcal{T}_1, \mathcal{T}_2)$ by an independent random variable $U \sim \mu$, and define

$$x_K^{\mathrm{out},(i)} := \Psi(\mathcal{T}_i, U), \qquad i \in \{1, 2\}.$$

On the event $E$, we have $\mathcal{T}_1 = \mathcal{T}_2$, and therefore

$$x_K^{\mathrm{out},(1)} = \Psi(\mathcal{T}_1, U) = \Psi(\mathcal{T}_2, U) = x_K^{\mathrm{out},(2)}.$$

Hence, on $E$,

$$\left|x_K^{\mathrm{out},(1)} - u_1\right| + \left|x_K^{\mathrm{out},(1)} - u_2\right| \ge |u_1 - u_2| = \Delta_u.$$

Consequently, on $E$,

$$\max\left\{\left|x_K^{\mathrm{out},(1)} - u_1\right|, \left|x_K^{\mathrm{out},(1)} - u_2\right|\right\} \ge \frac{\Delta_u}{2}.$$

Thus, on $E$, for at least one $i \in \{1, 2\}$, $\left|x_K^{\mathrm{out},(i)} - u_i\right| \ge \frac{\Delta_u}{2}$. For such an $i$, Lemma 21 gives

$$\begin{aligned} f_i(x_K^{\mathrm{out},(i)}) - f_i^\star &= \phi_a(x_K^{\mathrm{out},(i)} - u_i) \\ &\ge a\left(\left|x_K^{\mathrm{out},(i)} - u_i\right| - r_a\right)_+ \\ &\ge a\left(\frac{\Delta_u}{2} - r_a\right), \end{aligned}$$

where the last quantity is positive because $\Delta_u > 2r_a$.

For $i \in \{1, 2\}$, define

$$A_i := \left\{ f_i(x_K^{\mathrm{out},(i)}) - f_i^\star \ge a\left(\frac{\Delta_u}{2} - r_a\right)\right\}.$$

The preceding argument shows that $E \subseteq A_1 \cup A_2$. Hence, $\mathbf{1}_{A_1} + \mathbf{1}_{A_2} \ge \mathbf{1}_E$. Therefore

$$\mathbb{P}(A_1) + \mathbb{P}(A_2) \ge \mathbb{P}(E) \ge q.$$

Since each coupled run has its original marginal distribution,

$$\mathbb{P}(A_i) = \mathbb{P}_i\left(f_i(x_K^{\mathrm{out},(i)}) - f_i^\star \ge a\left(\frac{\Delta_u}{2} - r_a\right)\right).$$

We deduce that at least one $i \in \{1, 2\}$,

$$\mathbb{P}_i\left(f_i(x_K^{\mathrm{out},(i)}) - f_i^\star \ge a\left(\frac{\Delta_u}{2} - r_a\right)\right) \ge \frac{q}{2}.$$

□

*Proof of Lemma 23.* We have $\mathbb{E}[G(x)] = \frac{f'(x)}{p}\mathbb{E}[Z] = f'(x)$. Moreover,

$$\begin{aligned}\mathbb{E}\left[|G(x) - f'(x)|^\alpha\right] &= |f'(x)|^\alpha \left[p\left|\frac{1}{p} - 1\right|^\alpha + (1-p)\right] \\ &= |f'(x)|^\alpha \left[p^{1-\alpha}(1-p)^\alpha + (1-p)\right].\end{aligned}$$

Since $|f'(x)| \le a$ and $p \in (0,1]$, we obtain

$$\begin{aligned}\mathbb{E}\left[|G(x) - f'(x)|^\alpha\right] &\le a^\alpha \left(p^{1-\alpha} + 1\right) \\ &\le 2a^\alpha p^{1-\alpha},\end{aligned}$$

where the last inequality follows from $p^{1-\alpha} \ge 1$ since $\alpha > 1$. □

*Proof of Lemma 24.* We consider the two hard instances $f_1(x) = \phi_a(x + R_0)$ and $f_2(x) = \phi_a(x - R_0)$, whose minimizers are located at $-R_0$ and $R_0$, respectively. We couple their stochastic oracles so that, with probability greater than $2\delta$, both runs observe only zero gradients.

Consider the oracles $G_i(x, Z) = f_i'(x)Z/p$ with $Z \sim \text{Bernoulli}(p)$, for $i \in \{1, 2\}$. If the same Bernoulli draws are used for the two instances, then the two transcripts are identical on the event that all $K$ draws are zero. This event has probability $(1-p)^K$. For $p \in (0, 1/2]$, $\log(1-p) > -2p$, and hence $(1-p)^K > e^{-2Kp}$. To make this lower bound equal to $2\delta$, we choose $p$ so that $e^{-2Kp} = 2\delta$, which gives

$$p = \frac{\log(1/(2\delta))}{2K}.$$

Since $K \ge \log(1/(2\delta))$, we indeed have $p \le 1/2$.

We now determine the largest tail slope allowed by the stochastic-oracle assumption. Since $|f_i'(x)| \le a$ for $i \in \{1, 2\}$, Lemma 23 gives

$$\mathbb{E}\left[|G_i(x, Z) - f_i'(x)|^\alpha\right] \le 2a^\alpha p^{1-\alpha}.$$

Thus, to ensure the stochastic-oracle assumption, it suffices to impose $2a^\alpha p^{1-\alpha} \le \sigma^\alpha$.

On the other hand, if $r_a \le R_0/2$, then the hard-function construction in Lemma 22 gives an objective gap of at least $a(R_0 - r_a) \ge aR_0/2$. Hence, in this regime, the guaranteed lower bound increases with $a$. We therefore take the boundary value permitted by the moment sufficient condition:

$$a = \frac{\sigma}{2}\left(\frac{\log(1/(2\delta))}{K}\right)^{(\alpha-1)/\alpha}.$$

It remains to verify that $r_a \le R_0/2$. We have

$$r_a \le \frac{R_0}{2} \iff a \le \frac{e^{L_1 R_0/2} - 1}{L_1} \iff K \ge \log\left(\frac{1}{2\delta}\right)\left(\frac{\sigma L_1}{2(e^{L_1 R_0/2} - 1)}\right)^{\alpha/(\alpha-1)}.$$

Since $e^x - 1 \ge x$ for $x \ge 0$,

$$\frac{\sigma L_1}{2(e^{L_1 R_0/2} - 1)} \le \frac{\sigma}{R_0},$$

so the required condition $r_a \le R_0/2$ follows from the assumed lower bound on $K$.

Therefore, since $\mathbb{P}(\mathcal{T}_1 = \mathcal{T}_2) \ge \mathbb{P}(E) > 2\delta$, Lemma 22 implies that there exists $i \in \{1, 2\}$ such that

$$\mathbb{P}_i\left(f_i(x_K^{\text{out},(i)}) - f_i^\star \ge a(R_0 - r_a)\right) > \delta.$$

Since $r_a \le R_0/2$, the definition of $a$ gives

$$a(R_0 - r_a) \ge \frac{aR_0}{2} = \frac{\sigma R_0}{4}\left(\frac{\log(1/(2\delta))}{K}\right)^{(\alpha-1)/\alpha}.$$

□

*Proof of Lemma 25.* Fix $\gamma, \lambda$ and the output rule. Since $K \ge L_1 R_0$ and $\log(eK) \ge 1$, it follows that $\frac{L_1 R_0}{K \log(eK)} \le 1$. Thus an objective gap of $\sigma R_0$ is stronger than the claimed lower bound.

**First, consider large updates.** Let $H, b > 0$ satisfy $r_H < b/2$ and $2b \le R_0$, and consider

$$f_1(x) := \phi_H(x - b), \qquad f_2(x) := \phi_H(x - 2b).$$

Their minimizers are located at $b$ and $2b$, respectively. For this construction, we use the exact-gradient oracle. Since the minimizers are separated by $b > 2r_H$, Lemma 22 gives the objective threshold $H\left(\frac{b}{2} - r_H\right)$.

Let

$$S_H := \gamma \min\{\lambda, H\}.$$

Because $r_H < b/2$, the point $x_0 = 0$ lies in the common left linear tail of the two objectives, where both gradients are equal to $-H$. If

$$S_H \ge 2b + r_H,$$

then the first update moves both runs from 0 into their common right linear tail, where both gradients are equal to $H$. The next update therefore returns both runs to 0. The same two-step trajectory then repeats, and the two transcripts are identical. Hence, for one of the two instances,

$$\mathbb{P}_i\left(f_i(x_K^{\mathrm{out},(i)}) - f_i^\star \ge H\left(\frac{b}{2} - r_H\right)\right) \ge \frac{1}{2}.$$

It remains to consider the complementary case $S_H < 2b + r_H$.

**Next, consider small updates.** Consider

$$h_1(x) := \phi_a\left(x - \frac{R_0}{2}\right), \qquad h_2(x) := \phi_a(x - R_0),$$

where $a > 0$ will be chosen later. Both runs start from $x_0 = 0$. For each instance, use the Bernoulli oracle

$$G_i(x, Z) := \frac{h_i'(x)}{p} Z, \qquad Z \sim \mathrm{Bernoulli}(p),$$

where $p \in (0, 1)$ will be chosen later. Successive oracle calls use independent Bernoulli draws.

We require $a/p \le H$. Since $|h_i'(x)| \le a$, every update corresponding to a draw $Z$ has length at most

$$\gamma \min\left\{\lambda, \frac{a}{p}\right\} Z \le S_H Z,$$

where $S_H < 2b + r_H$ in the present case.

For the comparison of the two instances, couple the runs using the same Bernoulli draws $Z_0, \ldots, Z_{K-1}$, and let

$$N := \sum_{k=0}^{K-1} Z_k$$

be the number of nonzero Bernoulli draws. Since $\mathbb{E}[N] = Kp$, Markov's inequality gives $\mathbb{P}(N \le 2Kp) \ge \frac{1}{2}$. To control the total distance traveled, we impose

$$2Kp(2b + r_H) \le \frac{R_0}{4}.$$

On the event $\{N \le 2Kp\}$, for every $i \in \{1, 2\}$ and $0 \le k \le K$,

$$|x_k^{(i)}| \le S_H \sum_{j=0}^{k-1} Z_j \le S_H N \le 2KpS_H.$$

Since $S_H < 2b + r_H$ and $2Kp(2b + r_H) \le R_0/4$, it follows that $|x_k^{(i)}| < \frac{R_0}{4}$.

Since $p < 1$ and $a/p \le H$, we have $a \le pH < H$. Together with $r_H < b/2$ and $2b \le R_0$, this gives $r_a \le r_H < b/2 \le R_0/4$. Then every $x \le R_0/4$ lies in the common left linear tail, where $h_1'(x) = h_2'(x) = -a$. Thus, on $\{N \le 2Kp\}$, the shared Bernoulli draws give

$$G_k^{(1)} = G_k^{(2)} = -\frac{a}{p} Z_k.$$

Since both runs start from zero and receive the same stochastic gradient at every iteration, their iterates also coincide. Consequently, $\mathbb{P}(\mathcal{T}_1 = \mathcal{T}_2) \geq \frac{1}{2}$.

The minimizers are separated by $R_0/2 > 2r_a$. By Lemma 22, there exists $i \in \{1, 2\}$ such that

$$\mathbb{P}_i\left(h_i(x_K^{\mathrm{out},(i)}) - h_i^\star \geq a\left(\frac{R_0}{4} - r_a\right)\right) \geq \frac{1}{4}.$$

It remains to ensure that the Bernoulli oracle satisfies the stochastic oracle assumption. By Lemma 23, it is enough to have $2a^\alpha p^{1-\alpha} \leq \sigma^\alpha$.

We have therefore reduced the proof to choosing $a, p, H$, and $b$, with $p \in (0, 1)$, so that

$$\frac{a}{p} \leq H, \qquad 2Kp(2b + r_H) \leq \frac{R_0}{4}, \qquad 2a^\alpha p^{1-\alpha} \leq \sigma^\alpha.$$

For the large-update construction, we also need $r_H < \frac{b}{2}$ and $2b \leq R_0$. The objective thresholds in the two constructions are

$$H\left(\frac{b}{2} - r_H\right) \qquad \text{and} \qquad a\left(\frac{R_0}{4} - r_a\right).$$

We now choose the parameters so that these conditions hold and both thresholds are at least the claimed lower bound.

Assuming $r_H < b/2$, we have $2b + r_H < \frac{5b}{2}$. Hence $2Kp(2b + r_H) < 5Kpb$, so the displacement condition is guaranteed by

$$5Kpb \leq \frac{R_0}{4}.$$

For fixed $b$, increasing $p$ increases the largest admissible slope $a$ under the noise constraint, while decreasing the corresponding value of $a/p$. We therefore take the largest value allowed by the sufficient displacement condition above,

$$p := \frac{R_0}{20Kb}.$$

We will verify that $p < 1$ after choosing $b$. For this value of $p$, the stochastic oracle condition gives $a \leq 2^{-1/\alpha}\sigma p^{(\alpha-1)/\alpha}$. We take

$$a := 2^{-1/\alpha}\sigma p^{(\alpha-1)/\alpha}.$$

We next choose $H$. To obtain an objective threshold of at least $\sigma R_0$ in the large-update construction, we need $H\left(\frac{b}{2} - r_H\right) \geq \sigma R_0$, or equivalently

$$\frac{b}{2} \geq r_H + \frac{\sigma R_0}{H}.$$

Increasing $H$ makes the first term larger and the second term smaller. We choose the slope that makes their sum as small as possible, subject to $H \geq a/p$.

For a fixed $b$, the values of $a$ and $p$ are already determined. We therefore set

$$H := \operatorname*{arg\,min}_{u \geq a/p}\left\{r_u + \frac{\sigma R_0}{u}\right\}.$$

We have

$$\frac{d}{du}\left(r_u + \frac{\sigma R_0}{u}\right) = \frac{u^2 - \sigma L_1 R_0 u - \sigma R_0}{u^2(1 + L_1 u)}.$$

The numerator has exactly one positive root, $u_\star := \frac{\sigma L_1 R_0 + \sqrt{(\sigma L_1 R_0)^2 + 4\sigma R_0}}{2}$. The sum decreases up to $u_\star$ and increases after it. Taking into account the constraint $u \geq a/p$, we obtain

$$H = \max\left\{\frac{a}{p}, u_\star\right\}.$$

We first derive a bound on $H$ under the condition $2b \le R_0$, and then verify that our choice of $b$ indeed satisfies this condition. Under $2b \le R_0$, the definitions of $a$ and $p$ give

$$L_1 \frac{a}{p} = \sigma L_1 \left(\frac{10Kb}{R_0}\right)^{1/\alpha} \le \sigma L_1 (5K)^{1/\alpha} \le 5K^2.$$

We also have, $L_1 u_\star \le (\sigma L_1)(L_1 R_0) + \sqrt{(\sigma L_1)(L_1 R_0)} \le K^2 + K \le 2K^2$. Hence

$$L_1 H \le 5K^2.$$

Since $H \ge u_\star > \sigma L_1 R_0$, it holds that $\frac{\sigma R_0}{H} < \frac{1}{L_1}$, and therefore $r_H + \frac{\sigma R_0}{H} < \frac{\log(1+5K^2)+1}{L_1}$. We choose

$$b := \frac{2\big(\log(1+5K^2)+1\big)}{L_1}.$$

Since $K \ge 1$, we have $\frac{2}{L_1} \le b \le \frac{6\log(eK)}{L_1} < \frac{3R_0}{32}$. In particular,

$$2b < \frac{3R_0}{16} < R_0,$$

so the condition $2b \le R_0$ used above is indeed satisfied.

These bounds also give

$$0 < \frac{L_1 R_0}{120K\log(eK)} \le p \le \frac{L_1 R_0}{40K} \le \frac{1}{40} < 1.$$

Finally, by the choice of $b$, $r_H + \frac{\sigma R_0}{H} < \frac{b}{2}$. Hence $r_H < b/2$ and

$$H\left(\frac{b}{2} - r_H\right) > \sigma R_0,$$

which gives the required threshold in the large-update case.

For the small-update case, $a/p \le H$ holds by construction, and

$$2Kp(2b + r_H) < 5Kpb = \frac{R_0}{4}, \qquad 2a^\alpha p^{1-\alpha} = \sigma^\alpha.$$

Since $a \le pH < H$, we also have $r_a \le r_H < b/2 < 3R_0/64 < R_0/8$. Therefore,

$$\begin{aligned} a\left(\frac{R_0}{4} - r_a\right) &\ge \frac{aR_0}{8} = \frac{\sigma R_0}{16}(2p)^{(\alpha-1)/\alpha} \\ &\ge \frac{\sigma R_0}{16}\left(\frac{L_1 R_0}{60K\log(eK)}\right)^{(\alpha-1)/\alpha} \\ &\ge \frac{\sigma R_0}{128}\left(\frac{L_1 R_0}{K\log(eK)}\right)^{(\alpha-1)/\alpha}, \end{aligned}$$

where the last inequality uses $60^{(\alpha-1)/\alpha} \le \sqrt{60} < 8$.

The large-update case gives the claimed lower bound with probability at least $1/2$, and the small-update case with probability at least $1/4$. □

*Proof of Theorem 3.* Fix $\gamma, \lambda$ and the output rule. First, suppose that

$$\frac{L_1 R_0}{\log(eK)} > 64 \qquad \text{and} \qquad \frac{L_1 R_0}{\log(eK)} \ge \log(12/\delta).$$

Then

$$\max\left\{\frac{L_1 R_0}{\log(eK)}, \log(12/\delta)\right\} = \frac{L_1 R_0}{\log(eK)}.$$

By Lemma 25, there exists an instance such that the objective gap is at least

$$\frac{\sigma R_0}{128}\left(\frac{L_1 R_0}{K\log(eK)}\right)^{(\alpha-1)/\alpha}$$

with probability at least $1/4 > \delta$. This proves the claimed bound in this case.

It remains to consider the case where $L_1 R_0/\log(eK) \le 64$ or $L_1 R_0/\log(eK) < \log(12/\delta)$. Since $\log(12/\delta) > 1$, we have

$$\max\left\{\frac{L_1 R_0}{\log(eK)}, \log(12/\delta)\right\} \le 64\log(12/\delta).$$

Set

$$p := \frac{\log(1/(2\delta))}{2K}, \qquad a := 2^{-1/\alpha}\sigma p^{(\alpha-1)/\alpha} = \frac{\sigma}{2}\left(\frac{\log(1/(2\delta))}{K}\right)^{(\alpha-1)/\alpha},$$

and consider

$$f_1(x) := \phi_a(x + R_0), \qquad f_2(x) := \phi_a(x - R_0).$$

Their minimizers are located at $-R_0$ and $R_0$, respectively. Both objectives are convex and $(1, L_1)$-smooth by Lemma 21. The iteration assumptions give $0 < p \le 1/2$ and $r_a \le a \le R_0/2$.

For each instance, use the Bernoulli oracle $G_i(x, Z) := f_i'(x)Z/p$, with fresh independent draws. By Lemma 23, both oracles are unbiased and have central $\alpha$-th moments at most

$$2a^\alpha p^{1-\alpha} = \sigma^\alpha.$$

Couple the runs using the same Bernoulli draws. On the event that all $K$ draws are zero, both runs remain at zero and their transcripts coincide. Since $\log(1-p) > -2p$ for $0 < p \le 1/2$, this event has probability

$$(1-p)^K > e^{-2Kp} = 2\delta.$$

The minimizers are separated by $2R_0 > 2r_a$. By Lemma 22, there exists $i \in \{1, 2\}$ such that

$$\mathbb{P}_i\left(f_i(x_K^{\mathrm{out},(i)}) - f_i^\star \ge a(R_0 - r_a)\right) > \delta.$$

Since $r_a \le R_0/2$, we have $a(R_0 - r_a) \ge \frac{aR_0}{2} = \frac{\sigma R_0}{4}\left(\frac{\log(1/(2\delta))}{K}\right)^{(\alpha-1)/\alpha}$.

Finally, since $\delta < 1/4$, it holds that

$$\log(12/\delta) = \log(1/(2\delta)) + \log 24 < 6\log(1/(2\delta)).$$

Using $4 \cdot 6^{(\alpha-1)/\alpha} < 16$ and $64^{(\alpha-1)/\alpha} \le 8$, we obtain

$$\begin{aligned} a(R_0 - r_a) &\ge \frac{\sigma R_0}{16}\left(\frac{\log(12/\delta)}{K}\right)^{(\alpha-1)/\alpha} \\ &\ge \frac{\sigma R_0}{128}\left[\frac{\max\{L_1 R_0/\log(eK), \log(12/\delta)\}}{K}\right]^{(\alpha-1)/\alpha}. \end{aligned}$$

□