AVOIDING BIAS IN CLIPSGD FOR OVERPARAMETERIZED MODELS UNDER GENERALIZED SMOOTHNESS

Anonymous authors
Paper under double-blind review

ABSTRACT

Modern machine learning is dominated by complex, overparameterized architectures capable of interpolating data and achieving zero training loss. For such models, we investigate the convergence properties of two popular modifications to standard SGD: clipped SGD and normalized SGD. We show that under overparameterization and an accuracy-independent sufficient batch size, both clipped and normalized SGD do not suffer from the bias typically introduced by clipping, converging effectively at the same rate as their deterministic counterparts. This provides a rigorous theoretical justification for the empirical success of gradient clipping methods. In our analysis, we employ the (L_0, L_1)-smoothness condition, under which we obtain convergence rates that improve upon the best known results in prior work. Furthermore, we extend our analysis to specific challenging regimes, including heavy-tailed noise, (H_0, H_1)-smoothness (which is strictly weaker than standard assumptions in optimization literature) and the deterministic regime.

1 INTRODUCTION

Modern machine learning (ML) models, particularly deep learning (DL) architectures, are typically heavily overparameterized, often containing more parameters than available training samples. This capacity allows models to interpolate the training data and achieve near-zero training loss (Ma et al., 2018). The training of these models is commonly performed using iterative first-order stochastic optimization methods, most notably variants of stochastic gradient descent (SGD) (Robbins & Monro, 1951; Nemirovski & Yudin, 1983) and adaptive counterparts, e.g., Adam (Kingma, 2014).

To ensure training stability, gradient clipping (Mikolov et al., 2012) has become a widely used algorithmic modification in practical optimization. Beyond stability, it serves as an essential component of privacy-preserving training algorithms, such as DP-SGD (Abadi et al., 2016). However, despite its ubiquity, the theoretical properties of gradient clipping are surprisingly poorly understood. Standard theoretical analyses typically predict that clipped SGD (ClipSGD) (Koloskova et al., 2023) introduces a systematic bias, preventing convergence to the true optimum and instead stagnating at an error floor determined by the stochastic noise of the gradients and clipping threshold.

This theoretical prediction contradicts the empirical reality where overparameterized models routinely converge to zero loss. Furthermore, classical analyses often rely on the simplifying assumption of global L-smoothness (Lan, 2020; Bach, 2024). In practice, modern optimization landscapes are often non-smooth or exhibit rapidly growing gradients that violate standard Lipschitz continuity conditions. Some methods designed under L-smoothness assumption and exhibiting strong theoretical guarantees (see, for example, variance-reduced methods for finite-sum problems in Johnson & Zhang, 2013; Defazio et al., 2014; Reddi et al., 2016; Schmidt et al., 2017) perform significantly worse in practice in DL (Defazio & Bottou, 2019). To bridge this gap between theory and practice, recent works proposed relaxed smoothness assumptions, including (L_0, L_1)-smoothness (Zhang et al., 2020b;a) and (H_0, H_1)-smoothness (Vaswani & Babanezhad, 2025; Liu et al., 2025; Alimisis et al., 2025).

In this work, we provide a rigorous theoretical justification for the success of gradient clipping in the interpolation regime. We analyze two clipping algorithms: ClipSGD (Mikolov et al., 2012) and Normalized SGD (NSGD) (Nesterov, 1984) under the Strong Growth Condition (SGC) (Vaswani et al., 2019a), viewing SGC as a structured interpolation condition that can arise in some overparameterized regimes, rather than as a consequence of overparameterization alone. We demonstrate that under this condition and an accuracy-independent condition on the batch size, the bias typically

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.

associated with clipping disappears, allowing these algorithms to converge with speed comparable to their deterministic counterparts.

More specifically, our contributions are summarized as follows:

- (**Section 4**). We provide the first convergence results for ClipSGD and NSGD under the SGC for both non-convex and convex, (L_0, L_1)-smooth objectives. Our analysis is the first to theoretically show that, under an accuracy-independent condition on the batch size, these algorithms do not suffer from the bias typically introduced by clipping and converge with nearly the same efficiency as their deterministic counterparts, and outperform standard SGD.
- (**Section 4**). We extend our analysis of NSGD to settings where stochastic gradients follow heavy-tailed distributions. We provide the first result showing that, even under heavy-tailed noise (combined with the strong growth condition), NSGD maintains its efficiency, converging similarly to its deterministic counterpart.
- (**Section 6**). We show across four datasets that ClipSGD and NSGD train successfully with moderate fixed batches far outside the sufficient low-noise regime, and use a direct CIFAR-10 ablation to guide joint batch-size and step-size selection.
- (**Appendix D**). In the stochastic setting, we extend our analysis of clipped SGD and normalized SGD to convex, (H_0, H_1)-smooth objectives (Vaswani & Babanezhad, 2025; Liu et al., 2025). In the deterministic setup, we analyze Gradient Descent (GD) with a warm-up step size strategy (viewed as a clipping-based algorithm). Under the (H_0, H_1)-smoothness condition, our analysis improves upon the best-known convergence guarantees for non-convex, convex, and strongly convex objectives. We further improve these rates under the Polyak-Łojasiewicz (PL) condition (Polyak, 1963; Karimi et al., 2016; Yue et al., 2023).

2 Problem Setup

In this paper we consider optimization problems of the form:

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

where f is a possibly non-convex, and possibly stochastic function. This problem formulation is general and encompasses a wide range of applications, including the optimization of deterministic functions when $\forall \xi : f\left(x, \xi\right) \equiv f\left(x\right)$. Moreover, by taking $\mathcal{D}$ to be the discrete uniform distribution over the interval $[1, n]$, problem equation 1 can be reformulated as a finite-sum formulation: $f(x) := \frac{1}{n} \sum_{\xi=1}^{n} \left[f_\xi(x) \right]$, which also covers many practical scenarios. For example, in the context of supervised machine learning, f_ξ represents the loss of the model x on the data point ξ. Next, we will narrow the class of problems under consideration by introducing assumptions on the objective function and the gradient oracle. We start with assumptions on function f.

2.1 Assumptions on the objective function

Throughout the paper, our analysis is conducted under the (L_0, L_1)-smoothness condition. This condition was introduced in (Zhang et al., 2020b;a) as a natural relaxation of the classical L-smoothness assumption (see, e.g., Nesterov, 2013). In contrast to standard L-smoothness, which imposes a uniform global Lipschitz constant on the gradient, (L_0, L_1)-smoothness allows the local smoothness parameter to depend on the norm of the gradient $\|\nabla f\left(x\right)\|$.

Assumption 2.1. Let f be (L_0, L_1)-smooth, then $\forall x, y \in \mathbb{R}^d$, as long as $\|y - x\| \leq \frac{1}{L_1}$, we have

$$\|\nabla f\left(y\right) - \nabla f\left(x\right)\| \leq \left(L_0 + L_1 \|\nabla f\left(x\right)\|\right) \|y - x\| .$$

We use this as the main assumption in our work. The motivation for Assumption 2.1 comes from the empirical findings of (Zhang et al., 2020b). The authors observed that, in a range of DL problems, the local smoothness constant decreases during training and is closely related to the gradient norm. Moreover, in their experiments, the curvature at the final stages of training could be dramatically smaller than the curvature at initialization; for instance, in LSTM training on the PTB dataset, it was reported to be up to 1000 times smaller. This indicates that the L-smoothness assumption may be too pessimistic for such problems, whereas Assumption 2.1 offers a more adaptive and realistic alternative. We note that Assumption 2.1 is strictly weaker than classical L-smoothness. Indeed, (L_0, L_1)-

smoothness reduces to L-smoothness when $L_1 = 0$. Moreover, the inclusion is strict: the exponential of an inner product is (L_0, L_1)-smooth but not L-smooth (see Example 1.5, Gorbunov et al., 2025).

2.2 Assumptions on the gradient oracle

Throughout this work, we adopt a standard assumption commonly found in the stochastic optimization literature (Nemirovski et al., 2009; Moulines & Bach, 2011; Juditsky et al., 2011; Roux et al., 2012; Lan, 2012; Ghadimi & Lan, 2012; 2013; Stich et al., 2018; Koloskova et al., 2023; Lobanov et al., 2024a): *the unbiasedness* of the gradient oracle. This oracle, when queried at any point $x \in \mathbb{R}^d$, provides access solely to a stochastic gradient $\nabla f(x, \xi)$, satisfying:

Assumption 2.2. The gradient oracle is unbiased, i.e.

$$\mathbb{E}_\xi \left[\nabla f(x, \xi)\right] = \nabla f(x). \tag{2}$$

In this work we also focus on the interpolation regime, which is typically observed in large ML models(particularly DL architectures),such as overparameterized neural networks, where the number of parameters exceeds the number of training examples, thereby enabling perfect data fitting (Ma et al., 2018; Allen-Zhu et al., 2019; Gower et al., 2019; 2021; Cooper, 2021; Nakkiran et al., 2021).

Definition 2.3. The interpolation condition holds if

$$x^* \in \arg\min f \Rightarrow x^* \in \arg\min f(\cdot, \xi) \quad \text{a.s.}$$

As a restriction on the stochastic noise of gradient oracle, we employ the generalized strong growth condition (Vaswani et al., 2019a; Koloskova et al., 2020; Lobanov et al., 2024c):

Assumption 2.4. The generalized strong growth condition holds if $\forall x \in \mathbb{R}^d$, $\rho \geq 1$ and $\sigma \geq 0$

$$\mathbb{E}\left[\|\nabla f(x, \xi)\|^2\right] \leq \rho \|\nabla f(x)\|^2 + \sigma^2. \tag{3}$$

We introduced Assumption 2.4 for generality, as with $\rho = 1$ it reduces to the standard bounded variance assumption $\mathbb{E}[\|\nabla f(x, \xi) - \nabla f(x)\|^2] \leq \sigma^2$, which is widely used in the literature (see, e.g., Karimireddy et al., 2020; Yang et al., 2022). However, it can be observed that the general form equation 3 does not satisfy the interpolation condition. Consequently, the main results of this work focus on perhaps the most common assumption that satisfies Definition 2.3: Assumption 2.4 with $\sigma = 0$ — known as *the strong growth condition* (SGC) (Cevher & Vũ, 2019; Vaswani et al., 2019b;a; Hermant et al., 2025). See Appendix B.1 for a discussion of the generality of form equation 3.

It is worth noting that in some works (e.g., see Vaswani et al., 2019b;a), SGC is classified as a *stationary-point interpolation*, which implies that holds: $\nabla f(x^*) = 0 \Rightarrow \nabla f(x^*, \xi) = 0$.

In part of this work, we provide a generalization to the heavy-tailed (HT). We assume a generalized p-th central moment (Liu & Zhou, 2025), which can be reduced to Assumption 2.4 for HT noises.

Assumption 2.5 (Optional)**.** We assume that gradient oracle has a generalized p-th central moment, i.e. there exists $\rho \geq 1$, $\sigma \geq 0$ such that, for all $x \in \mathbb{R}^d$ and $p \in (1, 2)$:

$$\mathbb{E}\left[\|\nabla f(x, \xi) - \nabla f(x)\|^p\right] \leq (\rho - 1)\|\nabla f(x)\|^p + \sigma^p.$$

When $\rho = 1$, Assumption 2.5 reduces to the bounded p-th central moment condition, which can be considered a standard in the heavy-tailed setting (Zhang et al., 2020c; Liu et al., 2023; Hübler et al., 2025). Moreover, Assumption 2.5 can be linked to a form of equation 3 (a p-th moment) by applying finite form of Jensen's inequality: $\mathbb{E}\left[\|\nabla f(x, \xi)\|^p\right] \leq 2^{p-1}\left(\rho\|\nabla f(x)\|^p + \sigma^p\right)$.

In this work, the case $\sigma = 0$ in Assumption 2.5 is of primary interest. It is straightforward to see that this case satisfies Definition 2.3 and can also be classified as stationary-point interpolation. And this specific form $\mathbb{E}\left[\|\nabla f(x, \xi)\|^p\right] \leq 2^{p-1}\rho\|\nabla f(x)\|^p$ we term *the SGC under heavy-tailed noise*.

Example 2.6. Consider the function $f(x) = \frac{1}{2}\|x\|^2$ with gradient oracle $\nabla f(x, \xi) = (Z + 1)x$ (where Z is a random variable with a symmetric Pareto distribution). Then, for $\alpha \in (p, 2]$ we obtain the strong growth condition under heavy-tailed noise $\mathbb{E}\left[\|\nabla f(x, \xi)\|^p\right] \leq 2^{p-1}\rho\|\nabla f(x)\|^p$ with $\rho = \frac{\alpha}{\alpha - p} + 1 < \infty$, and $\mathbb{E}\left[\|\nabla f(x, \xi)\|^2\right] = \infty$.

3 RELATED WORKS

This section discusses the most relevant related work.

Clipped SGD has been studied extensively across settings. In non-convex deterministic (L_0, L_1)-smooth optimization, Zhang et al. (2020b) first theoretically demonstrated the benefits of clipping. This was later refined by Koloskova et al. (2023), who improved bounds for ClipGD and analyzed ClipSGD under stochastic (L_0, L_1)-smoothness with bounded variance; however, their guarantees converge only to an error floor $\min\{\sigma, \sigma^2/c\}$, even with mini-batching. Another line of work mitigates clipping bias using error feedback and momentum (Islamov et al., 2025), but for modified clipping schemes rather than vanilla ClipSGD, which is the focus of our analysis. For convex objectives, deterministic analyses identified linear and sublinear regimes (Lobanov et al., 2024b; Vankov et al., 2025), while stochastic works established high-probability guarantees under light- and heavy-tailed noise by choosing sufficiently large clipping radius (Gaash et al., 2025; Chezhegov et al., 2025). Other recent works study convex interpolation under (L_0, L_1)- and (H_0, H_1)-smoothness, but rely on non-standard metrics (Gorbunov et al., 2025; Alimisis et al., 2025). *We analyze ClipSGD with an arbitrary clipping radius in standard metrics (the expected minimum gradient norm and the expected optimality gap over N iterations), proving convergence to arbitrary accuracy under SGC and an accuracy-independent batch-size $\mathcal{O}(\rho)$, with two regimes for both non-convex and convex objectives.*

Normalized SGD dates back to Nesterov (1984) and has since been studied under bounded variance (BV) and heavy-tailed noise. In convex (L_0, L_1)-smooth BV settings, Lobanov & Gasnikov (2025) showed that NSGD improves the iteration complexity of ClipSGD, but at the cost of a much worse oracle complexity, $\mathcal{O}(\varepsilon^{-5})$. For non-convex L-smooth BV problems, momentum NSGD achieves $\mathcal{O}(\varepsilon^{-4})$ complexity (Cutkosky & Mehta, 2020; Zhao et al., 2021); the same rate was obtained under non-convex (L_0, L_1)-smoothness in (Hübler et al., 2024), although their analysis does not separate L_0 and L_1. In the heavy-tailed regime, (Hübler et al., 2025; Liu & Zhou, 2025) analyzed momentum NSGD for non-convex L- and (L_0, L_1)-smooth objectives, respectively, obtaining $\mathcal{O}(\varepsilon^{-(3p-2)/(p-1)})$ sample complexity. *In this work, we identify a regime where NSGD improves the iteration complexity of Clipped SGD without degrading oracle complexity, by exploiting the strong growth condition under heavy-tailed noise. We show that with batch size $\mathcal{O}(\rho^{1/(p-1)})$, NSGD matches its deterministic counterpart. While (Liu & Zhou, 2025) uses a related Assumption 2.5, their analysis is restricted to $\rho = 1$ and thus does not capture our results for $\rho > 1$. Moreover, their method requires access to the full gradient and deterministic function value to set β and η, which is typically impractical for large-scale models. Our analysis removes these limitations for both $\rho > 1$ and $\sigma > 0$.*

4 STOCHASTIC SETUP VIA CONSTANT STEP SIZE

In this section, we analyze the convergence of two gradient clipping algorithms in a stochastic setup: Clipped Stochastic Gradient Descent (ClipSGD, which clips the gradient when its norm exceeds a given radius) and Normalized Stochastic Gradient Descent (NSGD, which performs normalization/clipping at every iteration). For brevity, both algorithms are presented in Algorithm 1. Note that in Step 2 of Algorithm 1, $\nabla f\left(x^k, \boldsymbol{\xi}^k\right)$ denotes the stochastic gradient computed on a mini-batch of size $B \geq 1$. In Step 3, depending on the chosen method, we define the update direction as follows: for ClipSGD, we use the clipping operator $\mathbf{g}(x^k, \boldsymbol{\xi}^k) = \alpha_{\boldsymbol{\xi}^k} \nabla f(x^k, \boldsymbol{\xi}^k)$, where $\alpha_{\boldsymbol{\xi}^k} = \min\left\{1, c/\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\|\right\}$; for NSGD, we use the normalized gradient $\mathbf{g}(x^k, \boldsymbol{\xi}^k) = \frac{\nabla f(x^k, \boldsymbol{\xi}^k)}{\|\nabla f(x^k, \boldsymbol{\xi}^k)\| + \lambda}$ with parameter $\lambda > 0$. For each methods in Algorithm 1, we consider a constant step size $\eta > 0$. In our theoretical analysis, we show that under the SGC (Assumption 2.4

Algorithm 1 Clipping Algorithms in a Stochastic Setup

Method: ClipSGD or NSGD
Input: initial point $x_0 \in \mathbb{R}^d$, iterations number N, batch size B, step size $\eta > 0$ and parameters $c > 0$, $\lambda > 0$
for $k = 0$ **to** $N - 1$ **do**
1. Draw fresh i.i.d. samples $\xi_1^k, ..., \xi_B^k$
2. $\nabla f(x^k, \boldsymbol{\xi}^k) = \frac{1}{B} \sum_{i=1}^{B} \nabla f(x^k, \xi_i^k)$
3. Compute $\mathbf{g}(x^k, \boldsymbol{\xi}^k)$ for the selected algorithm.
3.1. For ClipSGD: $\mathbf{g}(x^k, \boldsymbol{\xi}^k) = \alpha_{\boldsymbol{\xi}^k} \nabla f(x^k, \boldsymbol{\xi}^k)$
3.2. For NSGD: $\mathbf{g}(x^k, \boldsymbol{\xi}^k) = \frac{\nabla f(x^k, \boldsymbol{\xi}^k)}{\|\nabla f(x^k, \boldsymbol{\xi}^k)\| + \lambda}$
4. $x^{k+1} \leftarrow x^k - \eta \cdot \mathbf{g}(x^k, \boldsymbol{\xi}^k)$
end for
Return: x^N

with $\sigma = 0$), ClipSGD and NSGD converge to a stationary point for non-convex functions (or to a minimizer for convex functions) with nearly the same efficiency as their deterministic counterparts.

4.1 NON-CONVEX FUNCTIONS

In this part, we focus on non-convex functions that satisfy (L_0, L_1)-smoothness (see Assumption 2.1). We start by presenting our novel convergence result for Clipped SGD.

Theorem 4.1 (Non-convex, ClipSGD). *Let f is (L_0, L_1)-smooth (Assumption 2.1) and gradient oracle is unbiased (Assumption 2.2) and satisfies SGC (Assumption 2.4 with $\sigma = 0$), then ClipSGD (Algorithm 1) with step size $\eta \leq [9\,(L_0 + L_1 c)]^{-1}$, and batch size $B \geq 72(\rho - 1)$ guarantees:*

$$\min_{k \in [0, N-1]} \mathbb{E}\left[\left\|\nabla f\left(x^k\right)\right\|\right] \leq \sqrt{\frac{18 F_0}{\eta N}} + \frac{18 F_0}{\eta c N},$$

where N is the number of iterations, $F_0 = f\left(x^0\right) - f^*$.

The proof of Theorem 4.1 is provided in Appendix C.1.1 and corresponds to the case $\sigma = 0$. Note that by selecting the maximum step size $\eta = [9\,(L_0 + L_1 c)]^{-1}$, we obtain the following convergence result: $\min_{k \in [0, N-1]} \mathbb{E}\left[\left\|\nabla f\left(x^k\right)\right\|\right]$ is upper bounded by $\mathcal{O}\left(\sqrt{\frac{L_0 F_0}{N}} + \sqrt{\frac{L_1 c F_0}{N}} + \frac{L_0 F_0}{cN} + \frac{L_1 F_0}{N}\right)$.

Comparison to the prior works. First, the convergence result presented in Theorem 4.1 matches (when $B = \mathcal{O}(\rho)$) the convergence of ClipGD in the non-convex *deterministic setting* under (L_0, L_1)-smoothness (see Theorem 2.1, Koloskova et al., 2023). Second, the first term in Theorem 4.1 coincides with the convergence rate of SGD (*without clipping*) under the L-smoothness, SGC assumptions with step size $\eta \leq (\rho L)^{-1}$ (see Theorem 3, Vaswani et al., 2019a): $\mathcal{O}\left((\eta N)^{-1/2}\right)$. Therefore, it is straightforward to observe that Theorem 4.1 improves over SGD in both: Iteration complexity (due to a larger step size—an advantage from using mini-batches, and through the refined assumption of (L_0, L_1)-smoothness); Oracle complexity (in scenarios analogous to those in Zhang et al. (2020b) where $L_0 \lll L$ and $L_1 \lll L$, and potentially due to the addition of the term $\frac{F_0}{\eta c N}$ which depends on the clipping radius c; note, however, that the term $\mathcal{O}\left((\eta N)^{-1/2}\right)$ remains dominant).

Proposition 4.2 (Generalized SGC). *If Assumption 2.4 holds for $\sigma \geq 0$, then Theorem 4.1 takes the following form with $\eta \leq [9\,(L_0 + L_1 c)]^{-1}$ and $B \geq 72(\rho - 1)$ guarantees an error:*

$$\min_{k \in [0, N-1]} \mathbb{E}\left[\left\|\nabla f\left(x^k\right)\right\|\right] \lesssim \sqrt{\frac{F_0}{\eta N}} + \frac{F_0}{\eta c N} + \frac{\sigma}{\sqrt{B}} + \frac{\sigma^2}{cB}.$$

It is not hard to see that with a sufficiently large batch size, Proposition 4.2 guarantees convergence to the desired accuracy ε : $\min_{k \in [0, N-1]} \mathbb{E}\left[\left\|\nabla f\left(x^k\right)\right\|\right] \leq \varepsilon$, thereby improving upon the prior result (see Appendix C.4.1, Koloskova et al., 2023), which only ensured convergence to an error floor (asymptote) independent of the batch size B.

Next, we present the results for Normalized SGD.

Theorem 4.3 (Non-convex, NSGD). *Let f is (L_0, L_1)-smooth (Assumption 2.1) and gradient oracle is unbiased (Assumption 2.2) and satisfies SGC (Assumption 2.4 with $\sigma = 0$), then NSGD (Algorithm 1) with any $\lambda > 0$, step size $\eta \leq \frac{\lambda}{L_0 + L_1 \lambda}$, batch size $B \geq 64(\rho - 1)$ guarantees:*

$$\min_{k \in [0, N-1]} \mathbb{E}\left[\left\|\nabla f\left(x^k\right)\right\|\right] \leq \frac{4 F_0}{\eta N} + 6\lambda,$$

where N is the number of iterations, $F_0 = f\left(x^0\right) - f^*$.

The proof of Theorem 4.3 is provided in Appendix C.2.1 and corresponds to the case $\sigma = 0$.

Comparison to ClipSGD (Theorem 4.1). From Theorem 4.3, we observe that Normalized SGD converges to hyperparameter $\lambda > 0$; however, the smaller λ is, the more iterations are required, as λ directly influences the step size. Note that for NSGD to reach the desired accuracy ε : $\min_{k \in [0, N-1]} \mathbb{E}\left[\left\|\nabla f\left(x^k\right)\right\|\right] \leq \varepsilon$, it is necessary to take $\lambda \leq \frac{\varepsilon}{12}$. Then, by setting the maximum

allowable hyperparameter $\lambda = \frac{\varepsilon}{12}$ and the maximum step size $\eta = \frac{\varepsilon}{12L_0 + L_1\varepsilon}$, we obtain the iteration complexity: $N = \mathcal{O}\left(\frac{L_0 F_0}{\varepsilon^2} + \frac{L_1 F_0}{\varepsilon}\right)$, which coincides with ClipSGD (see the discussion following Theorem 4.1 with $c = \lambda \simeq \varepsilon$). Thus, Normalized SGD asymptotically behaves like Clipped SGD with a small clipping radius $c \simeq \varepsilon$, demonstrating improved coefficients. However, if the clipping radius is sufficiently large $c \geq \varepsilon$, then NSGD improves the iteration complexity compared to ClipSGD (specifically the term involving L_1) without increasing the oracle call numbers. In other words, in this regime, Normalized SGD outperforms ClipSGD in iteration and oracle complexity.

Comparison to the prior works. First, when $L_1 = 0$, Theorem 4.3 achieves oracle complexity matching *the deterministic counterpart with momentum* under L-smoothness (see Theorem 1, Cutkosky & Mehta, 2020, with $\sigma = 0$), up to an $\mathcal{O}(\rho)$ factor (since we employ mini-batches), i.e., $T = \mathcal{O}\left(\varepsilon^{-2}\right)$. We also note that due to the relaxed (L_0, L_1)-smoothness assumption (in particular, because $L_0 \lll L$ and $L_1 \lll L$, and includes an additive $\mathcal{O}\left(L_1\varepsilon^{-1}\right)$ term) the iteration complexity in Theorem 4.3 improves upon that of (Cutkosky & Mehta, 2020). Second, since our analysis relies on the SGC, the complexity $T = \mathcal{O}\left(\rho L_0\varepsilon^{-2} + \rho L_1\varepsilon^{-1}\right)$ established in Theorem 4.3 is strictly better than the oracle complexity $T = N \cdot B = \mathcal{O}\left(L\varepsilon^{-4}\right)$ (with the batch size $B = \mathcal{O}\left(\varepsilon^{-2}\right)$) of *NSGD under L-smoothness* (see Theorem 1, Zhao et al., 2021). Third, by employing mini-batches and leveraging the SGC, Theorem 4.3 improves both the iteration and oracle complexity of *NSGD with momentum under (L_0, L_1)-smoothness* (see Theorem 2, Hübler et al., 2024). Unlike (Hübler et al., 2024), our analysis separates the influence of L_0, L_1, which yields benefits in regimes $L_0 \lll L_1$.

Proposition 4.4 (Heavy-tailed). *Let f is (L_0, L_1)-smooth (Assumption 2.1) and gradient oracle is unbiased (Assumption 2.2) and satisfies SGC under heavy-tailed noise (Assumption 2.5 with $\sigma = 0$), then NSGD with any $\lambda > 0$, $\eta \leq \frac{\lambda}{L_0 + L_1\lambda}$ and $B \geq 2^{\frac{3p+1}{p-1}}(\rho - 1)^{\frac{1}{p-1}}$ guarantees an error:*

$$\min_{k \in [0, N-1]} \mathbb{E}\left[\left\|\nabla f\left(x^k\right)\right\|\right] \leq \frac{4F_0}{\eta N} + 6\lambda,$$

where N is the number of iterations, $F_0 = f\left(x^0\right) - f^*$.

Proposition 4.4 shows that the strong growth condition (SGC) produces a similar effect in the heavy-tailed setup. Specifically, Proposition 4.4 establishes an oracle complexity for $\lambda = \varepsilon/12$ and maximum step size: $\mathcal{O}\left(\rho^{\frac{1}{p-1}} L_0 F_0 \varepsilon^{-2} + \rho^{\frac{1}{p-1}} L_1 F_0 \varepsilon^{-1}\right)$. In comparison, (see Corollary 3, Hübler et al., 2025) derived a complexity for normalized stochastic gradient descent (NSGD) under L-smoothness $\mathcal{O}\left(\frac{LF_0}{\varepsilon^2}\left(\frac{\sigma}{\varepsilon}\right)^{\frac{p}{p-1}}\right)$, which requires a substantially larger number of oracle calls.

In Appendices C.2.1 and C.2.3, we provide proofs for the cases satisfying the generalized SGC (Assumption 2.4) and the generalized p-th central moment condition (Assumption 2.5), respectively.

4.2 Convex functions

In this subsection, we consider the class of convex and smoothness functions $f(\cdot, \xi)$.

Assumption 4.5. Let $f(\cdot, \xi)$ be $(\mathcal{L}_0, \mathcal{L}_1)$-smooth and convex for almost every $\xi \in \mathcal{D}$. That is, $\forall x, y, z \in \mathbb{R}^d$ such that $\|y - z\| \leq \frac{1}{\mathcal{L}_1}$, we have with $\mathcal{L}_\xi(z) := \mathcal{L}_0 + \mathcal{L}_1 \|\nabla f(z, \xi)\|$:

$$f(x, \xi) + \langle \nabla f(x, \xi), y - x \rangle \leq f(y, \xi) \leq f(z, \xi) + \langle \nabla f(z, \xi), y - z \rangle + \frac{\mathcal{L}_\xi(z)}{2}\|y - z\|^2.$$

We note that Assumption 4.5 is widely used in the literature under interpolation conditions (Woodworth & Srebro, 2021; Gorbunov et al., 2025; Alimisis et al., 2025). Assumption 4.5 can be seen to reduce to Assumption 2.1 under Assumption 2.4 (SGC) with $\sigma = 0$, $L_0 = \mathcal{L}_0$ and $L_1 = \sqrt{\rho}\mathcal{L}_1$ (see Appendix B.2 for details). Below, we present the convergence results for ClipSGD.

Theorem 4.6 (Convex, ClipSGD). *Let $f(x, \xi)$ be $(\mathcal{L}_0, \mathcal{L}_1)$-smooth and convex (Assumption 4.5), gradient oracle is unbiased (Assumption 2.2) and satisfies SGC (Assumption 2.4 with $\sigma = 0$). Then ClipSGD (Algorithm 1) with step size $\eta \leq [9(\mathcal{L}_0 + \sqrt{\rho}\mathcal{L}_1 c)]^{-1}$ and batch size $B \geq 36(\rho - 1)$ yields:*

$$\mathbb{E}\left[f\left(x^N\right)\right] - f^* \leq \max\left\{(1 - \bar{\alpha})^{N/2} F_0, \frac{18R_0^2}{N\eta}\right\},$$

where N is the number of iterations, $F_0 = f\left(x^0\right) - f^*$, and $\bar{\alpha} := \min\{\eta c/(36R_0), 1\}$.

The proof of Theorem 4.6 is provided in Appendix C.1.2. It is straightforward to see that if the step size is chosen as $\eta = [9(\mathcal{L}_0 + \sqrt{\rho}\mathcal{L}_1 c)]^{-1}$, the iteration complexity N takes the following form: $\max\left\{\mathcal{O}\left(\max\left\{1, \frac{[\mathcal{L}_0+\sqrt{\rho}\mathcal{L}_1 c]R_0}{c}\right\}\log\left(\frac{1}{\varepsilon}\right)\right), \mathcal{O}\left(\frac{(\mathcal{L}_0+\sqrt{\rho}\mathcal{L}_1 c)R_0^2}{\varepsilon}\right)\right\}$, where $R_0 = \left\|x^0 - x^*\right\|$.

Comparison to the prior works. First, Theorem 4.6 reveals two convergence regimes, closely matching the convergence rate of ClipGD in the convex *deterministic setting* (see Theorem 3.1, Lobanov et al., 2024b). The key distinction is that due to the monotonicity of the gradient norm in the deterministic case, ClipGD initially exhibits a linear rate, which slows down to a sublinear rate after the threshold c is reached. Second, Theorem 4.6 holds for an arbitrary clipping radius $c > 0$. In contrast, (Chezhegov et al., 2025; Gaash et al., 2025) require a specific, sufficiently large clipping radius c to establish convergence guarantees for *ClipSGD in the convex (L_0, L_1)-smooth setting*. This choice of clipping radius effectively ignores the term corresponding to the exponential decrease of the function (although we note that the sublinear term ultimately dominates). Third, the prior best known convergence rate of ClipSGD for convex (L_0, L_1)-smooth functions with *an arbitrary* c (see Theorem 3.1, Lobanov & Gasnikov, 2025) requires strong assumptions: a uniform bounded gradient norm. We successfully addressed these issues through an improved analysis and the application of SGC. Fourth, Theorem 4.6 provides guarantees in terms of the standard metric $\mathbb{E}\left[f(x^N)\right] - f^*$, whereas the analysis in (Gorbunov et al., 2025) relies on a non-standard metric.

We now present the convergence results for NSGD.

Theorem 4.7 (Convex, NSGD). *Let $f(x, \xi)$ be $(\mathcal{L}_0, \mathcal{L}_1)$-smooth and convex (Assumption 4.5), gradient oracle is unbiased (Assumption 2.2) and satisfies SGC (Assumption 2.4 with $\sigma = 0$). Then NSGD with $\lambda > 0$, step size $\eta \leq \min\left\{\frac{\lambda}{\mathcal{L}_0+\sqrt{\rho}\mathcal{L}_1\lambda}, 4R_0\right\}$ and batch size $B \geq 64(\rho - 1)$ yields:*

$$\mathbb{E}\left[f\left(x^N\right)\right] - f^* \leq \left(1 - \frac{\eta}{4R_0}\right)^N F_0 + 6\lambda R_0,$$

where N is the number of iterations, $F_0 = f\left(x^0\right) - f^*$ and $R_0 = \left\|x^0 - x^*\right\|$.

The proof of Theorem 4.7 is provided in Appendix C.2.2. Similar to the non-convex case, Theorem 4.7 shows that NSGD converges to an error floor of $6\lambda R_0$. To ensure convergence to a desired accuracy ε, i.e., $\mathbb{E}[f(x^N)] - f^* \leq \varepsilon$, the hyperparameter must satisfy $\lambda \leq \frac{\varepsilon}{12R_0}$. Setting λ to its maximum admissible value $\lambda = \frac{\varepsilon}{12R_0}$ and the step size to $\eta = \min\left\{\frac{\varepsilon}{12\mathcal{L}_0 R_0+\sqrt{\rho}\mathcal{L}_1\varepsilon}, 4R_0\right\}$, NSGD achieves the iteration complexity: $N = \mathcal{O}\left(\max\left\{1, \mathcal{L}_0 R_0^2/\varepsilon + \sqrt{\rho}\mathcal{L}_1 R_0\right\}\log\frac{F_0}{\varepsilon}\right)$ with batch size $B = \mathcal{O}(\rho)$.

Comparison to ClipSGD (Theorem 4.6). Similar to the non-convex setting, the convergence behavior of NSGD mirrors that of ClipSGD with $c \simeq \varepsilon/R_0$. The distinction lies in the fact that for ClipSGD, the term involving $\mathcal{L}_1$ differ between its two convergence regimes, yielding a bound proportional to $\max\left\{\mathcal{O}\left(\sqrt{\rho}\mathcal{L}_1 R_0 \log(1/\varepsilon)\right), \mathcal{O}\left(\sqrt{\rho}\mathcal{L}_1 R_0\right)\right\}$. Specifically, for small ε, the $\mathcal{O}\left(\sqrt{\rho}\mathcal{L}_1 R_0 \log(1/\varepsilon)\right)$ term dominates in ClipSGD, which aligns with the rate of NSGD. However, for sufficiently large ε, the $\mathcal{O}\left(\sqrt{\rho}\mathcal{L}_1 R_0\right)$ term becomes dominant for ClipSGD, a rate that can potentially be worse than that of NSGD. We also note that for a large clipping radius $c \gtrsim \varepsilon/R_0$, NSGD is more efficient than ClipSGD in terms of both iteration N and oracle $T = N \cdot B$ complexity.

Comparison to the prior works. First, unlike ClipSGD, Theorem 4.7 shows that the NSGD achieves the same efficiency as its *deterministic counterpart* (see Theorem 3.3, Lobanov et al., 2024b). However, the analysis in (Lobanov et al., 2024b) introduces the hyperparameter under the condition $\lambda > 0 : \|\nabla f(x^k)\| \geq \lambda$ (for all k). In this work, we explicitly characterize the convergence behavior for any $\lambda > 0$. Second, (Theorem 4.1, Lobanov & Gasnikov, 2025) relies on a strong assumption of bounded gradients and requires a sufficiently large batch size of order $B = \mathcal{O}(\varepsilon^{-3})$. In contrast, the results presented in Theorem 4.7 eliminate the need for the bounded gradient. Furthermore, by leveraging the SGC, the required batch size is $B = \mathcal{O}(\rho)$.

Proposition 4.8 (Heavy-tailed). *Let $f(\cdot, \xi)$ be $(\mathcal{L}_0, \mathcal{L}_1)$-smooth and convex (Assumption 4.5), gradient oracle is unbiased (Assumption 2.2) and satisfies SGC under heavy-tailed noise (Assumption 2.5 with $\sigma = 0$). Then NSGD (Algorithm 1) with hyperparameter $\lambda > 0$, step size $\eta \leq \min\left\{\frac{\lambda}{\mathcal{L}_0+\sqrt{\rho}\mathcal{L}_1\lambda}, 4R_0\right\}$ and batch size $B \geq 2^{\frac{3p+1}{p-1}}(\rho - 1)^{\frac{1}{p-1}}$ guarantees convergence rate:*

$$\mathbb{E}\left[f\left(x^N\right)\right] - f^* \leq \left(1 - \frac{\eta}{4R_0}\right)^N F_0 + 6\lambda R_0.$$

Here $R_0 = \left\|x^0 - x^*\right\|$ *and* $F_0 = f\left(x^0\right) - f^*$.

Proposition 4.8 shows that in the convex setting, the SGC under heavy-tailed noise yields the effect: NSGD achieves efficiency matching that of its deterministic counterpart. The proof of Proposition 4.8 is provided in Appendix C.2.3. We also present generalizations of Theorem 4.7 to the case $\sigma \geq 0$ (generalized strong growth condition, Assumption 2.4) in Appendix C.2.2.

5 Discussion and Limitation

First, due to space constraints in the main text, we refer the reader to Appendix D. There, we extend the analysis of Section 4 to convex $(\mathcal{H}_0, \mathcal{H}_1)$-smooth (which is strictly weaker than standard assumptions in optimization literature) functions and provide an improved analysis of GD with a warm-up step-size strategy under (H_0, H_1)-smoothness. The latter covers several regimes, including non-convex, convex, and strongly convex objectives, as well as under the PL condition.

Second, the results of Theorems 4.6 and 4.7, together with Proposition 4.8, show that the iteration complexity scales as $N \sim \sqrt{\rho}$. This dependence stems from assuming convexity and smoothness at the level of individual realizations, i.e., for almost every $\xi \in \mathcal{D}$ (see Assumption 4.5). It can be avoided if Assumption 4.5 is imposed directly on the mini-batch oracle associated with $\boldsymbol{\xi}$. Third, the GD step size in Theorems D.16-D.19 requires knowledge of f^*; however, this information is often readily available in modern machine learning applications (e.g., for the squared loss in regression and the logistic loss in classification, $f^* = 0$ Bartlett et al., 2006).

Finally, as directions for future work, we identify the following avenues: extending the results of Subsection 4.1 to (H_0, H_1)-smooth objectives; generalizing Theorem 4.6 to the case $\sigma > 0$; extending our ClipSGD guarantees to heavy-tailed noise (Assumption 2.5). Another promising direction is to understand how the results of Subsection 4.2 change when accelerated methods are considered.

6 Numerical Experiments

Under SGC with $\sigma = 0$, theory requires a fixed $B = \mathcal{O}(\rho)$ that is independent of the desired accuracy ε. Our experiments address three questions: whether a moderate fixed batch works throughout training, how conservative the sufficient low-noise regime is, and which diagnostics identify a useful operational batch. Full protocols, additional datasets, and complete tables are in Appendix A.

Protocol and metric. With $B = 128$ and seed 86, we train CIFAR-10/ResNet-18 for 100 epochs using ClipSGD ($\eta = 10, c = 0.1$), strict NSGD ($\eta = 1, \lambda = 0$), or SGD ($\eta = 0.01$). These are the best-performing configurations within a prespecified 100-epoch grid search at $B = 128$, rather than theory-selected defaults; Appendix A gives the complete search spaces and selection protocol. For a single-example gradient g and batch mean g_B, define $\rho = \frac{\mathbb{E}\|g\|^2}{\|\mathbb{E}g\|^2}, q_B = \rho_B - 1 = \frac{\mathbb{E}\|g_B\|^2}{\|\mathbb{E}g\|^2} - 1$.

For i.i.d. sampling, $q_B = (\rho - 1)/B$; thus a measured batch ratio is ρ_B, not ρ. The appendix gives direct validation and a finite-population correction.

Finding 1: the sufficient regime is conservative. Clipped SGD and Normalized SGD reach 84.65% test accuracy versus 77.67% for SGD (Figure 1(a)). Their trajectories coincide because Clipped SGD remains in the clipped regime and $\eta c = 1$, giving the same update as strict normalization. Nevertheless, neither sufficient threshold is met: the final checkpoint has $q_{128} = 12.00$ and direct $\rho = 1543.49$, implying $B \geq 111{,}060$ for Clipped SGD and $B \geq 98{,}720$ for Normalized SGD. These values expose conservative sufficient constants; they are neither practical recommendations nor certified uniform SGC bounds. We use strict Normalized SGD, the undamped boundary case $\lambda = 0$, to isolate the effect of direction normalization from denominator regularization.

The diagnostic is accurate but oracle-dependent. The conversion from ρ_B to ρ is independently validated using per-example gradients. At the final normalized checkpoint, direct $\rho = 1543.49$ predicts $q_{128} = 12.051$, versus the measured 12.002; for SGD, direct $\rho = 1294.02$ predicts 10.102, versus 9.952. The respective relative gaps are 0.4% and 1.5%. Including random crop and flip in the stochastic oracle raises q_{128} to 15.67 ± 0.71 for ClipSGD/NSGD and 17.21 ± 0.60 for SGD over five

passes, but leaves both ratios orders of magnitude above the sufficient thresholds. Thus the qualitative conclusion is robust, although the numerical value of ρ depends on the oracle definition.

Finding 2: the gap persists across tasks. The same $B = 128$ diagnostic on UCI HAR, FEMNIST Digits, and CIFAR-100 yields direct ρ values from 21.13 to 1359.55 for the clipping-based methods. Substitution into the i.i.d. theorem conditions gives sufficient batches from 1289 to 97,816, yet the $B = 128$ runs reach 92.82%, 98.23%, and 50.79% test accuracy, respectively. These screening runs are not controlled optimizer comparisons, but they show that the theory–practice gap is not specific to CIFAR-10 and that ρ is task-, trajectory-, and method-dependent.

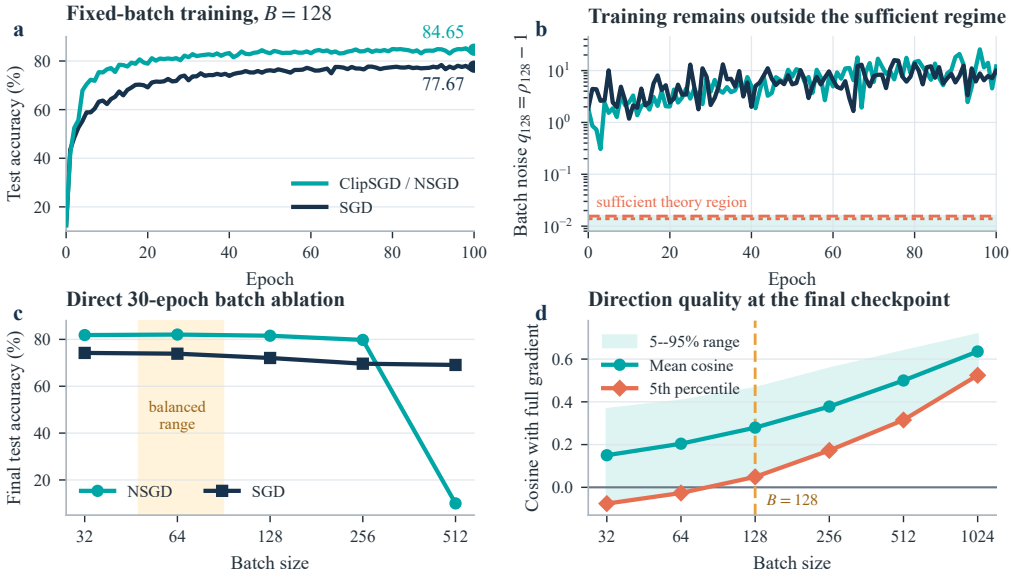


Figure 1: Practical batching on CIFAR-10/ResNet-18. **(a)** Fixed-B training. **(b)** Batch noise versus sufficient thresholds. **(c)** 30-epoch batch ablation with learning rates scaled from the tuned $B = 128$ baseline. **(d)** Gradient alignment versus B.

Finding 3: direction quality alone does not select the batch. In the 30-epoch NSGD ablation, $B = 32, 64, 128$ attain 81.84%, 82.05%, and 81.56%, while $B = 256$ falls to 79.77% and $B = 512$ collapses to 10.00% under linear learning-rate scaling. Mean alignment with the full gradient improves monotonically, and its fifth percentile becomes positive at $B = 128$ (Figure 1(c–d)). Direction quality is thus informative but not sufficient: update budget, throughput, and step-size tuning also matter. **Practical takeaway.** Here $B = 64$ is a balanced pilot default and $B \in [32, 128]$ is the useful range; in a new task, we recommend screening $B \in \{32, 64, 128, 256\}$ with retuned step sizes rather than using a theorem bound as a batch prescription. Full cross-task results are in Appendix A.3.

7 Conclusion

We establish arbitrary-accuracy convergence guarantees for ClipSGD and NSGD under SGC and relaxed (L_0, L_1)-smoothness; for NSGD, the normalization parameter is selected according to the desired accuracy. With $\sigma = 0$, the sufficient batch size scales with ρ but is independent of the desired optimization accuracy ε. Higher accuracy is achieved through the iteration budget and, for NSGD, a smaller λ, rather than through a growing mini-batch. The guarantees cover convex and non-convex objectives, arbitrary clipping radii for ClipSGD, and heavy-tailed gradients for NSGD.

The experiments establish three complementary facts. First, direct per-example computation confirms the conversion between the observable mini-batch ratio ρ_B and the single-example constant ρ. Second, ClipSGD and NSGD optimize successfully even when the measured batch noise lies far outside the sufficient regime, consistently across four datasets. Third, larger batches improve direction quality but need not improve finite-budget accuracy, making direct ablation essential. These findings do not contradict the theory: they distinguish a clean sufficient low-noise regime from the much broader regime in which the methods work empirically. The central message is therefore twofold: clipping bias can be removed without increasing the batch as the desired accuracy becomes more stringent, and theorem-derived sufficient batches should not be interpreted as practical batch-size prescriptions.

AI USE STATEMENT

Generative AI tools were used for language editing and to improve the clarity of the manuscript. We reviewed all resulting changes and take full responsibility for the final content.

REFERENCES

Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, CCS'16, pp. 308–318. ACM, October 2016. doi: 10.1145/2976749.2978318. URL http://dx.doi.org/10.1145/2976749.2978318.

Foivos Alimisis, Rustem Islamov, and Aurelien Lucchi. Why do we need warm-up? a theoretical perspective. *arXiv preprint arXiv:2510.03164*, 2025.

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International conference on machine learning*, pp. 242–252. PMLR, 2019.

Francis Bach. *Learning theory from first principles*. MIT press, 2024.

Peter L Bartlett, Michael I Jordan, and Jon D McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

Volkan Cevher and Bang Công Vũ. On the linear convergence of the stochastic gradient method with constant step-size. *Optimization Letters*, 13(5):1177–1187, 2019.

Yeshwanth Cherapanamjeri, Nilesh Tripuraneni, Peter Bartlett, and Michael Jordan. Optimal mean estimation without a variance. In *Conference on Learning Theory*, pp. 356–357. PMLR, 2022.

Savelii Chezhegov, Aleksandr Beznosikov, Samuel Horváth, and Eduard Gorbunov. Convergence of clipped-sgd for convex (l_0, l_1)-smooth optimization with heavy-tailed noise. *arXiv preprint arXiv:2505.20817*, 2025.

Yaim Cooper. Global minima of overparameterized neural networks. *SIAM Journal on Mathematics of Data Science*, 3(2):676–691, 2021.

David C Cox. Note on a martingale inequality of pisier. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 92, pp. 163–165. Cambridge University Press, 1982.

Ashok Cutkosky and Harsh Mehta. Momentum improves normalized sgd. In *International conference on machine learning*, pp. 2260–2268. PMLR, 2020.

Aaron Defazio and Léon Bottou. On the ineffectiveness of variance reduced optimization for deep learning. *Advances in Neural Information Processing Systems*, 32, 2019.

Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. *Advances in neural information processing systems*, 27, 2014.

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.

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.

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. In *The Thirteenth International Conference on Learning Representations*, 2025.

Robert Gower, Othmane Sebbouh, and Nicolas Loizou. Sgd for structured nonconvex functions: Learning rates, minibatching and interpolation. In *International Conference on Artificial Intelligence and Statistics*, pp. 1315–1323. PMLR, 2021.

Robert Mansel Gower, Nicolas Loizou, Xun Qian, Alibek Sailanbayev, Egor Shulgin, and Peter Richtárik. Sgd: General analysis and improved rates. In *International conference on machine learning*, pp. 5200–5209. PMLR, 2019.

Julien Hermant, Marien Renaud, Jean-François Aujol, Charles Dossal, and Aude Rondepierre. Gradient correlation is a key ingredient to accelerate sgd with momentum. In *International Conference on Learning Representations, 2025*, 2025.

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.

Florian Hübler, Ilyas Fatkhullin, and Niao He. From gradient clipping to normalization for heavy tailed sgd. In *International Conference on Artificial Intelligence and Statistics*, pp. 2413–2421. PMLR, 2025.

Rustem Islamov, Samuel Horvath, Aurelien Lucchi, Peter Richtarik, and Eduard Gorbunov. Double momentum and error feedback for clipping with fast rates and differential privacy. *arXiv preprint arXiv:2502.11682*, 2025.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. *Advances in neural information processing systems*, 26, 2013.

Anatoli Juditsky, Arkadi Nemirovski, et al. First order methods for nonsmooth convex large-scale optimization, ii: utilizing problems structure. *Optimization for Machine Learning*, 30(9):149–183, 2011.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Joint European conference on machine learning and knowledge discovery in databases*, pp. 795–811. Springer, 2016.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International conference on machine learning*, pp. 5132–5143. PMLR, 2020.

Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Anastasia Koloskova, Nicolas Loizou, Sadra Boreiri, Martin Jaggi, and Sebastian Stich. A unified theory of decentralized sgd with changing topology and local updates. In *International conference on machine learning*, pp. 5381–5393. PMLR, 2020.

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.

Guanghui Lan. An optimal method for stochastic composite optimization. *Mathematical Programming*, 133(1):365–397, 2012.

Guanghui Lan. *First-order and stochastic optimization methods for machine learning*, volume 1. Springer, 2020.

Yuxing Liu, Yuze Ge, Rui Pan, An Kang, and Tong Zhang. Theoretical analysis on how learning rate warmup accelerates convergence. *arXiv preprint arXiv:2509.07972*, 2025.

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, Jiawei Zhang, and Zhengyuan Zhou. Breaking the lower bound with (little) structure: Acceleration in non-convex stochastic optimization with heavy-tailed noise. In *The Thirty Sixth Annual Conference on Learning Theory*, pp. 2266–2290. PMLR, 2023.

Aleksandr Lobanov and Alexander Gasnikov. Power of generalized smoothness in stochastic convex optimization: First-and zero-order algorithms. *arXiv preprint arXiv:2501.18198*, 2025.

Aleksandr Lobanov, Nail Bashirov, and Alexander Gasnikov. The "black-box" optimization problem: Zero-order accelerated stochastic method via kernel approximation. *Journal of Optimization Theory and Applications*, 203(3):2451–2486, 2024a.

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*, 2024b.

Aleksandr Lobanov, Alexander Gasnikov, and Fedor Stonyakin. Highly smooth zeroth-order methods for solving optimization problems under the pl condition. *Computational Mathematics and Mathematical Physics*, 64(4):739–770, 2024c.

Siyuan Ma, Raef Bassily, and Mikhail Belkin. The power of interpolation: Understanding the effectiveness of sgd in modern over-parametrized learning. In *International Conference on Machine Learning*, pp. 3325–3334. PMLR, 2018.

Tomáš Mikolov et al. Statistical language models based on neural networks. *Presentation at Google, Mountain View, 2nd April*, 80(26), 2012.

Eric Moulines and Francis Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. *Advances in neural information processing systems*, 24, 2011.

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124003, 2021.

Arkadi Nemirovski and D Yudin. *Problem complexity and method efficiency in optimization*. Wiley, 1983.

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

Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

Yurii E Nesterov. Minimization methods for nonsmooth convex and quasiconvex functions. *Matekon*, 29(3):519–531, 1984.

Gilles Pisier. Martingales with values in uniformly convex spaces. *Israel Journal of Mathematics*, 20 (3):326–350, 1975.

Boris T Polyak. Gradient methods for the minimisation of functionals. *USSR Computational Mathematics and Mathematical Physics*, 3(4):864–878, 1963.

Sashank J Reddi, Ahmed Hefny, Suvrit Sra, Barnabas Poczos, and Alex Smola. Stochastic variance reduction for nonconvex optimization. In *International conference on machine learning*, pp. 314–323. PMLR, 2016.

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

Nicolas Roux, Mark Schmidt, and Francis Bach. A stochastic gradient method with an exponential convergence rate for finite training sets. *Advances in neural information processing systems*, 25, 2012.

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

Sebastian U Stich, Jean-Baptiste Cordonnier, and Martin Jaggi. Sparsified sgd with memory. *Advances in neural information processing systems*, 31, 2018.

Daniil Vankov, Anton Rodomanov, Angelia Nedich, Lalitha Sankar, and Sebastian U Stich. Optimizing (l_0, l_1)-smooth functions by gradient methods. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=GQ1Tc3vHbt.

Sharan Vaswani and Reza Babanezhad. Armijo line-search can make (stochastic) gradient descent provably faster. *arXiv preprint arXiv:2503.00229*, 2025.

Sharan Vaswani, Francis Bach, and Mark Schmidt. Fast and faster convergence of sgd for over-parameterized models and an accelerated perceptron. In *The 22nd international conference on artificial intelligence and statistics*, pp. 1195–1204. PMLR, 2019a.

Sharan Vaswani, Aaron Mishkin, Issam Laradji, Mark Schmidt, Gauthier Gidel, and Simon Lacoste-Julien. Painless stochastic gradient: Interpolation, line-search, and convergence rates. *Advances in neural information processing systems*, 32, 2019b.

Bengt von Bahr and Carl-Gustav Esseen. Inequalities for the rth absolute moment of a sum of random variables, $1 \leq r \leq 2$. *The Annals of Mathematical Statistics*, pp. 299–303, 1965.

Blake E Woodworth and Nathan Srebro. An even more optimal stochastic optimization algorithm: minibatching and interpolation learning. *Advances in neural information processing systems*, 34: 7333–7345, 2021.

Junchi Yang, Antonio Orvieto, Aurelien Lucchi, and Niao He. Faster single-loop algorithms for minimax optimization without strong concavity. In *International conference on artificial intelligence and statistics*, pp. 5485–5517. PMLR, 2022.

Pengyun Yue, Cong Fang, and Zhouchen Lin. On the lower bound of minimizing polyak-łojasiewicz functions. In *The Thirty Sixth Annual Conference on Learning Theory*, pp. 2948–2968. PMLR, 2023.

Bohang Zhang, Jikai Jin, Cong Fang, and Liwei Wang. Improved analysis of clipping algorithms for non-convex optimization. *Advances in Neural Information Processing Systems*, 33:15511–15521, 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(3):132103, 2021.

APPENDIX

Contents

1 Introduction **1**

2 Problem Setup **2**
2.1 Assumptions on the objective function 2
2.2 Assumptions on the gradient oracle 3

3 Related works **4**

4 Stochastic Setup via Constant Step Size **4**
4.1 Non-convex functions 5
4.2 Convex functions 6

5 Discussion and Limitation **8**

6 Numerical Experiments **8**

7 Conclusion **9**

A Detailed Numerical Experiments **15**
A.1 Protocol and estimand 15
A.2 CIFAR-10 numerical results 16
A.3 Cross-dataset screening 17

B Auxiliary Results **18**
B.1 On the generalized strong growth condition 19
B.2 On the generalized smoothness conditions 19
B.3 Auxiliary lemma for $(\mathcal{L}_0, \mathcal{L}_1)$-smoothness 20
B.4 Monotone behavior of the distance to optimum for ClipSGD 22
B.5 Monotone behavior of the distance to optimum for NSGD 23
B.6 Variance for mini-batch stochastic gradient 25
B.7 Motivation for the generalized strong growth condition under heavy-tailed noise . . 25

C Missing Proofs for Stochastic Part (ClipSGD & NSGD) **26**
C.1 Missing proofs for clipped stochastic gradient descent method 26
C.1.1 Non-convex setup under (L_0, L_1)-smooth and generalized strong growth conditions 26
C.1.2 Convex setup under $(\mathcal{L}_0, \mathcal{L}_1)$-smooth and strong growth conditions 31
C.2 Missing proofs for normalized stochastic gradient descent method 36
C.2.1 Non-convex setup under (L_0, L_1)-smooth and generalized strong growth conditions 36
C.2.2 Convex setup under $(\mathcal{L}_0, \mathcal{L}_1)$-smooth and strong growth conditions 38
C.2.3 Modification to the proof to include heavy-tailed noise 41

D Extending the Analysis to $(\mathcal{H}_0, \mathcal{H}_1)$-Smoothness **42**
D.1 Assumptions on the objective function 42
D.2 Implication of convexity and monotone behavior of the distance to optimum 44
D.3 Stochastic setting 48
D.3.1 Extension to heavy-tailed noise 56
D.4 Deterministic setting 57
D.4.1 Non-convex setting 58

D.4.2 Convex setting 60
D.4.3 Strongly convex setting 61

A Detailed Numerical Experiments

This section reports the numerical protocol and diagnostics underlying Section 6. The experiments are designed to separate two claims that should not be conflated. The theoretical claim is that, under SGC with $\sigma = 0$, the sufficient batch size depends on ρ but not on the desired optimization accuracy ε. The empirical claim is that ClipSGD and NSGD can train successfully with batches far below this sufficient bound. A finite experiment can illustrate successful fixed-batch training; it cannot certify a global SGC constant or prove accuracy independence.

A.1 Protocol and estimand

CIFAR-10 protocol. We use the 50,000-example CIFAR-10 training set and 10,000-example test set. The model is a torchvision ResNet-18 initialized from scratch with a ten-class output head. Training uses random 32×32 crops with four pixels of padding, random horizontal flips, and cross-entropy loss. Updates contain neither momentum nor weight decay. All runs use seed 86. The three 100-epoch configurations used in the main comparison are listed in Table 1.

Hyperparameter selection. The main configurations are selected by a prespecified exhaustive sweep at $B = 128$, comprising 49 configurations in total, with every candidate trained for 100 epochs under the same protocol. For ClipSGD we search $\eta \in \{0.01, 0.1, 0.5, 1, 5, 7, 10\}$ and $c \in \{10^{-4}, 10^{-3}, 10^{-2}, 0.1, 0.25\}$; for NSGD we search the same seven values of η with strict normalization ($\lambda = 0$); and for SGD we search $\eta \in \{6.25{\times}10^{-5}, 1.25{\times}10^{-4}, 6.25{\times}10^{-4}, 10^{-3}, 1.25{\times}10^{-3}, 10^{-2}, 10^{-1}\}$. Candidates are ranked by their best accuracy over the 100-epoch run on the fixed evaluation split. The selected grid optima are $(\eta, c) = (10, 0.1)$ for ClipSGD, $\eta = 1$ for NSGD, and $\eta = 0.01$ for SGD. Thus the step sizes used below are empirically selected within stated search spaces, not theorem-prescribed values or claims of global optimality. Because the evaluation split is also used for screening, the reported accuracies are within-protocol optimization outcomes rather than unbiased estimates from an untouched test set.

Table 1: Selected CIFAR-10 grid optima and accuracy. ClipSGD and NSGD have identical checkpoints for these hyperparameters.

Method	η	c / λ	Batch	Final acc.	Best acc.
ClipSGD	10	$c = 0.1$	128	84.65	85.24
NSGD	1	$\lambda = 0$	128	84.65	85.24
SGD	0.01	–	128	77.67	78.37

The NSGD run uses strict normalization, $g/\|g\|$, corresponding to the undamped boundary case $\lambda = 0$. This choice isolates direction normalization from the denominator regularization used in the theoretical formulation, which is stated for $\lambda > 0$. The SGC diagnostics and the conversion between ρ_B and ρ are unaffected by this choice once the optimizer trajectory is fixed. For the chosen ClipSGD parameters, stochastic gradients remain in the clipped regime and its update equals the strict-NSGD update: $10 \min\{1, 0.1/\|g\|\} g = g/\|g\|$. All state parameters and model buffers in the two final checkpoints agree exactly.

What is measured. For the fixed empirical objective, let $g_i = \nabla f_i(x)$, $\bar{g} = n^{-1} \sum_{i=1}^n g_i$, and let g_B be the mean of a uniformly sampled batch. We report

$$\rho = \frac{n^{-1} \sum_{i=1}^n \|g_i\|^2}{\|\bar{g}\|^2}, \qquad \rho_B = \frac{\mathbb{E}\|g_B\|^2}{\|\bar{g}\|^2}, \qquad q_B = \rho_B - 1. \tag{4}$$

For i.i.d. sampling and uniform sampling without replacement, respectively,

$$\rho_B = 1 + \frac{\rho - 1}{B}, \qquad \rho_B = 1 + \frac{n - B}{B(n-1)}(\rho - 1). \tag{5}$$

Thus dividing a measured batch ratio by B is incorrect; the i.i.d. inversion is $\rho = 1 + B(\rho_B - 1)$. To keep the diagnostic objective fixed, we disable augmentation, switch the model to evaluation

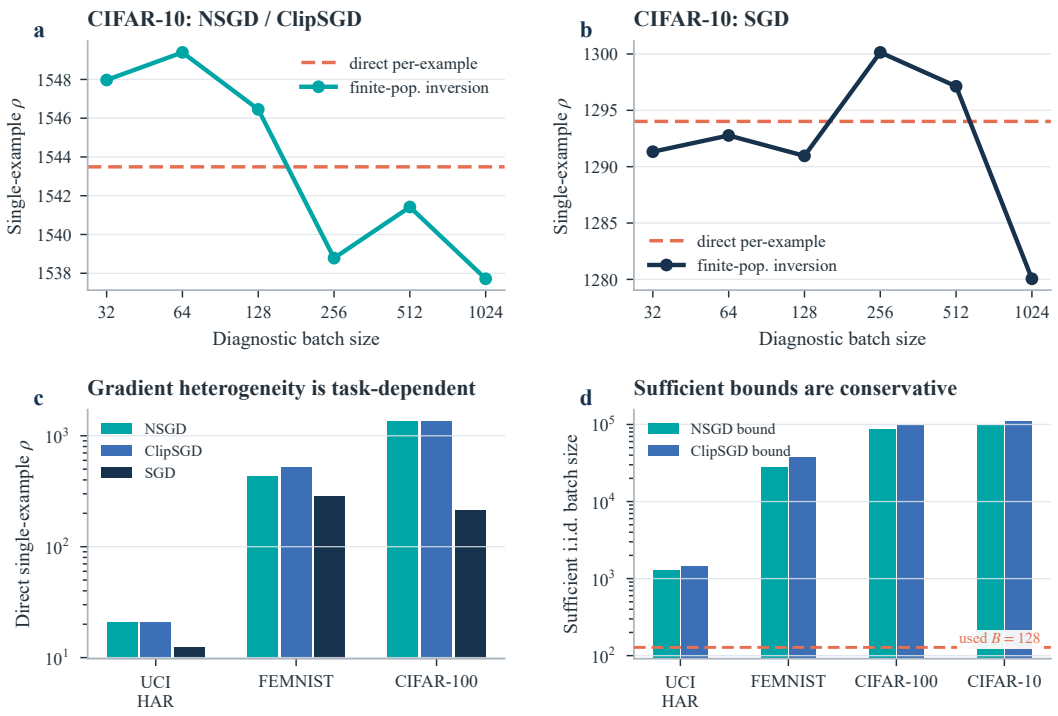


Figure 2: SGC diagnostics. **(a–b)** Finite-population inversions over a diagnostic batch grid agree with direct per-example computation at the final CIFAR-10 checkpoints. **(c)** Direct ρ varies by roughly two orders of magnitude across tasks and also depends on the optimizer trajectory. **(d)** Plugging these local values into the sufficient i.i.d. conditions produces batches far above the actually used $B = 128$. The bars describe proof conditions, not recommended training batches.

mode, use all training examples, and restore the training mode after evaluation. We estimate ρ_B using equal-size batches and independently compute single-example ρ with vectorized per-example gradients.

A.2 CIFAR-10 NUMERICAL RESULTS

Direct validation and augmentation. Table 2 compares the epoch-100 batch diagnostic with a direct pass over all 50,000 individual gradients. The batch inversion agrees within 0.4% for NSGD/ClipSGD and within 1.5% for SGD. Including random crop and flip as part of the stochastic oracle increases the measured batch noise: over five independent diagnostic passes, $q_{128} = 15.67 \pm 0.71$ for NSGD/ClipSGD and $q_{128} = 17.21 \pm 0.60$ for SGD. This changes the numerical scale but not the conclusion that the sufficient thresholds are not met.

Table 2: Validation of the CIFAR-10 SGC diagnostic at epoch 100. The prediction uses the directly measured single-example ρ and the i.i.d. relation in equation 5.

Method	Measured q_{128}	Direct ρ	Predicted q_{128}	Rel. gap
NSGD / ClipSGD	12.002	1543.49	12.051	0.4%
SGD	9.952	1294.02	10.102	1.5%

The inferred single-example ρ is not monotone along training. For the NSGD/ClipSGD trajectory it reaches an observed maximum of 3291.25 at epoch 96 and ends at 1537.27 by batch inversion. SGD reaches an observed maximum of 2039.86 and ends at 1274.88. Peaks often occur when the full-gradient norm is small, making the ratio in equation 4 sensitive to cancellation. Consequently, neither a final value nor a maximum over one finite trajectory is a global SGC constant.

Batch-size ablation. Table 3 gives the complete 30-epoch ablation from Figure 1(c). Learning rates are scaled linearly around the tuned $B = 128$ values; they are not retuned separately for each batch. Reducing the batch to 32 or 64 does not hurt final accuracy in this fixed-epoch comparison. Increasing it to 256 reduces accuracy, while NSGD with the naively scaled $B = 512$ learning rate collapses.

The experiment uses one seed and a single scaling rule, so it supports a local recommendation, not a universal optimal batch.

Table 3: CIFAR-10 batch-size ablation, 30 epochs, seed 86.

Method	B	Learning rate	Updates	Final acc.	Best acc.
NSGD	32	0.25	46,890	81.84	82.50
NSGD	64	0.50	23,460	82.05	82.47
NSGD	128	1.00	11,730	81.56	82.22
NSGD	256	2.00	5,880	79.77	80.16
NSGD	512	4.00	2,940	10.00	15.63
SGD	32	0.0025	46,890	74.22	74.22
SGD	64	0.0050	23,460	73.89	73.90
SGD	128	0.0100	11,730	72.06	74.00
SGD	256	0.0200	5,880	69.62	72.64
SGD	512	0.0400	2,940	69.09	72.10

Direction-quality diagnostic. For independently partitioned batches at the final NSGD/ClipSGD checkpoint, Table 4 reports cosine alignment with the full gradient. The fifth percentile changes sign between $B = 64$ and $B = 128$, and no negative inner products are observed at $B \geq 256$. Larger batches therefore give a cleaner stochastic direction, but this benefit does not translate monotonically into better fixed-epoch accuracy. In particular, $B = 512$ has better alignment than $B = 128$ but fails under naive learning-rate scaling.

Table 4: Direction quality at the final CIFAR-10 NSGD/ClipSGD checkpoint.

B	q_B	Mean cosine	5th pct. cosine	Negative fraction
32	48.313	0.150	-0.076	14.79%
64	24.163	0.204	-0.026	6.66%
128	12.043	0.279	0.049	2.82%
256	5.976	0.378	0.173	0.00%
512	2.978	0.500	0.315	0.00%
1024	1.470	0.636	0.524	0.00%

A.3 Cross-dataset screening

We apply the same diagnostic at $B = 128$ to three additional datasets that permit relatively inexpensive full-gradient measurements. UCI HAR uses all 7,352 training examples and a 561–128–6 MLP; FEMNIST Digits uses a fixed random subset of 50,000 examples and a small CNN; CIFAR-100 uses all 50,000 training examples and a 100-class ResNet-18. These are screening runs with dataset-specific learning rates and training lengths, not controlled optimizer comparisons. In particular, final accuracies are not matched.

Table 5: Cross-dataset screening summary. Direct ρ is computed from all individual gradients at the final checkpoint.

Dataset	Method	Epochs	Final acc.	Best acc.	Direct ρ
UCI HAR	NSGD	50	92.82	97.40	21.13
UCI HAR	ClipSGD	50	92.82	97.40	21.13
UCI HAR	SGD	50	85.76	97.53	12.37
FEMNIST Digits-50k	NSGD	10	98.23	98.23	437.70
FEMNIST Digits-50k	ClipSGD	10	98.23	98.23	527.02
FEMNIST Digits-50k	SGD	10	93.74	93.74	286.57
CIFAR-100	NSGD	30	50.79	50.84	1359.55
CIFAR-100	ClipSGD	30	50.79	50.84	1359.55
CIFAR-100	SGD	30	40.28	43.01	213.58

The direct values range from about 12 on UCI HAR/SGD to about 1360 on CIFAR-100/NSGD, demonstrating that the numerical scale is strongly dependent on the task, model, optimizer, and iterate. Independent batch grids $B \in \{32, 64, 128, 256, 512\}$ agree with direct computation: for example, the five finite-population inversions lie in $[431.9, 440.6]$ for FEMNIST/NSGD and $[1348.8, 1370.2]$ for CIFAR-100/NSGD, versus direct values 437.7 and 1359.5. At CIFAR-100 and $B = 128$, the i.i.d. inversion differs from direct measurement by about 0.002%.

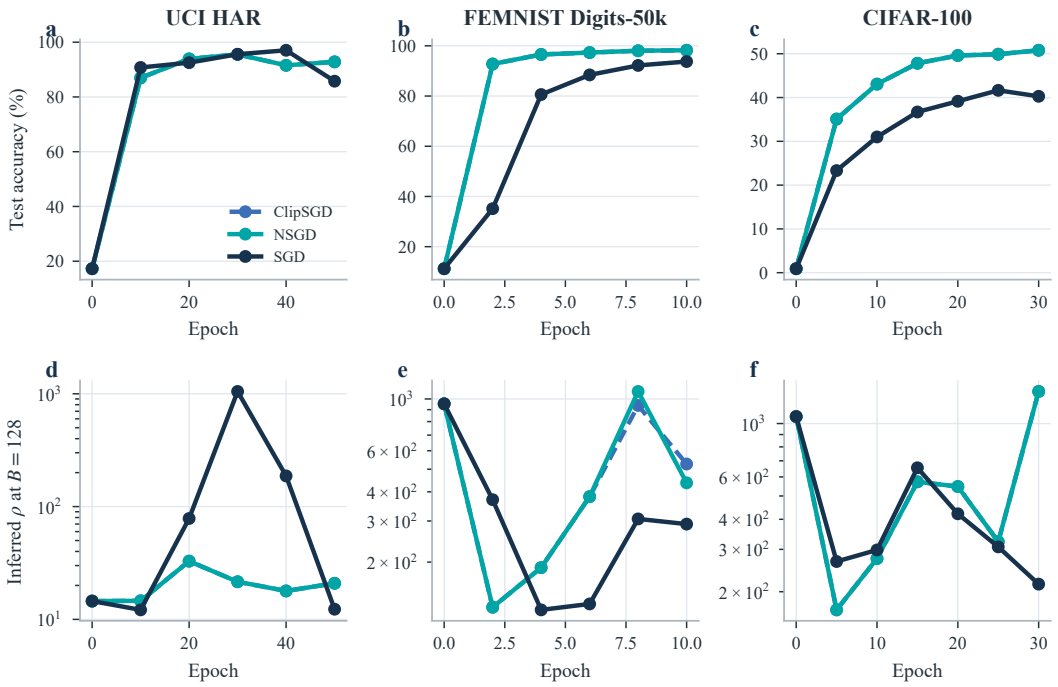


Figure 3: Cross-dataset screening at $B = 128$. The top row shows test accuracy; the bottom row shows the single-example ρ inferred from batches using the i.i.d. relation. ClipSGD and NSGD overlap when their updates coincide. The pronounced non-monotonicity of ρ confirms that one endpoint should not be interpreted as a global SGC constant.

Table 6 translates each final direct value into the corresponding sufficient batch condition. The finite-population column solves the without-replacement inequality in equation 5; it is smaller than the i.i.d. value but remains far above 128. All experiments make substantial optimization progress despite operating outside these sufficient regimes. This supports using ρ to quantify gradient heterogeneity and proof conservatism, while choosing an operational batch by direct ablation and direction-quality diagnostics.

Table 6: Sufficient batches obtained by substituting final direct ρ values into the non-convex theorem conditions. These are not practical batch recommendations.

Dataset	Method	i.i.d. sufficient B	Finite-pop. sufficient B
UCI HAR	NSGD	1,289	1,097
UCI HAR	ClipSGD	1,450	1,211
FEMNIST Digits-50k	NSGD	27,949	17,928
FEMNIST Digits-50k	ClipSGD	37,874	21,551
CIFAR-100	NSGD	86,948	31,746
CIFAR-100	ClipSGD	97,816	33,088
CIFAR-10	NSGD	98,720	33,191
CIFAR-10	ClipSGD	111,060	34,478

Scope and reproducibility limitations. All reported runs use a single seed. FEMNIST uses a fixed 50,000-example subset of the cached 341,873 training images. BatchNorm is active during training but frozen during the fixed-function diagnostic; augmentation noise is measured separately. The cross-dataset runs use unmatched checkpoints and must not be read as a causal ranking of optimizers by ρ. Finally, a maximum observed along finitely many iterates is only a trajectory diagnostic, not a certified uniform SGC constant. These limitations motivate the narrow practical recommendation in the main text: screen a small grid such as $B \in \{32, 64, 128, 256\}$ with retuned step sizes, rather than transferring a theorem-derived batch or a single optimum across tasks.

B Auxiliary Results

In this section, we provide auxiliary technical results that are used in our analysis.

Basic inequalities. For all $a, b \in \mathbb{R}^d$ ($d \geq 1$) and $\lambda > 0$, the following inequalities hold:

$$\langle a, b \rangle = \frac{\|a\|^2}{2} + \frac{\|b\|^2}{2} - \frac{\|b-a\|^2}{2}, \tag{6}$$

$$\langle a, b \rangle \leq \|a\| \cdot \|b\|, \tag{7}$$

$$\langle a, b \rangle \leq \frac{\|a\|^2}{2\lambda} + \frac{\lambda \|b\|^2}{2}. \tag{8}$$

(L_0, L_1)-smoothness. Assumption 2.1 has the equivalent form $\forall x, y \in \mathbb{R}^d$ with $\|y - x\| \leq \frac{1}{L_1}$:

$$f(y) - f(x) \leq \langle \nabla f(x), y - x \rangle + \frac{L_0 + L_1 \|\nabla f(x)\|}{2} \|y - x\|^2. \tag{9}$$

Assumption B.1 (Convexity)**.** The function f is convex, i.e., for all $x, y \in \mathbb{R}^d$,

$$f(x) - f(y) \leq \langle \nabla f(x), x - y \rangle, \qquad \forall x, y \in \mathbb{R}^d. \tag{10}$$

B.1 On the generalized strong growth condition

Generalized strong growth condition. In this work, we assume that the generalized strong growth condition holds, i.e.

$$\mathbb{E}\left[\|\nabla f(x, \xi)\|^2\right] \leq \rho \|\nabla f(x)\|^2 + \sigma^2.$$

Subtracting the norm of the full gradient from both sides, we have that

$$\mathbb{E}\left[\|\nabla f(x, \xi)\|^2\right] - \|\nabla f(x)\|^2 \leq (\rho - 1) \|\nabla f(x)\|^2 + \sigma^2.$$

We know that

$$\begin{aligned}\mathbb{E}\left[\|\nabla f(x, \xi) - \nabla f(x)\|^2\right] &= \mathbb{E}\left[\|\nabla f(x, \xi)\|^2 - 2\langle \nabla f(x, \xi), \nabla f(x) \rangle + \|\nabla f(x)\|^2\right] \\ &= \mathbb{E}\left[\|\nabla f(x, \xi)\|^2\right] - \|\nabla f(x)\|^2.\end{aligned}$$

Therefore, on the left-hand side we now have exactly the variance, and hence it is equivalent to

$$\mathbb{E}\left[\|\nabla f(x, \xi) - \nabla f(x)\|^2\right] \leq (\rho - 1) \|\nabla f(x)\|^2 + \sigma^2. \tag{11}$$

Strong growth condition. We note that interpolation regime is achieved when $\sigma = 0$, i.e.

$$\mathbb{E}\left[\|\nabla f(x, \xi)\|^2\right] \leq \rho \|\nabla f(x)\|^2.$$

From equation 11, we see that in this case the variance has the following upper bound:

$$\mathbb{E}\left[\|\nabla f(x, \xi) - \nabla f(x)\|^2\right] \leq (\rho - 1) \|\nabla f(x)\|^2.$$

This means that ρ cannot be smaller than 1! If ρ is smaller than 1, it means that expectation of the positive number has to be negative and no function satisfies this.

B.2 On the generalized smoothness conditions

(L_0, L_1)-Smoothness. In this work, we rely on two smoothness assumptions: Assumption 2.1 for the non-convex setting (see Section 4.1) and Assumption 4.5 for the convex setting (see Section 4.2). It is straightforward to verify that, under the strong growth condition (Assumption 2.4 with $\sigma = 0$), Assumption 4.5 implies Assumption 2.1 with $L_0 = \mathcal{L}_0$ and $L_1 = \sqrt{\rho}\mathcal{L}_1$:

$$\begin{aligned}\|\nabla f(y) - \nabla f(x)\| &\overset{①}{=} \|\mathbb{E}\left[\nabla f(y, \xi) - \nabla f(x, \xi)\right]\| \\ &\overset{②}{\leq} \mathbb{E}\left[\|\nabla f(y, \xi) - \nabla f(x, \xi)\|\right]\end{aligned}$$

$$\overset{③}{\leq} \left(\mathcal{L}_0 + \mathcal{L}_1 \mathbb{E}\left[\|\nabla f(x,\xi)\|\right]\right)\|y-x\|$$
$$= \left(\mathcal{L}_0 + \mathcal{L}_1 \mathbb{E}\left[\left[\|\nabla f(x,\xi)\|^2\right]^{1/2}\right]\right)\|y-x\|$$
$$\overset{④}{\leq} \left(\mathcal{L}_0 + \mathcal{L}_1 \left[\mathbb{E}\left[\|\nabla f(x,\xi)\|^2\right]\right]^{1/2}\right)\|y-x\|$$
$$\overset{⑤}{\leq} \left(\mathcal{L}_0 + \sqrt{\rho}\mathcal{L}_1 \|\nabla f(x)\|\right)\|y-x\|, \tag{12}$$

where in ① we used the unbiasedness relation $\nabla f(x) = \mathbb{E}\left[\nabla f(x,\xi)\right]$ and $\nabla f(y) = \mathbb{E}\left[\nabla f(y,\xi)\right]$ (Assumption 2.2), in ②, we applied Jensen's inequality to the norm, in ③, we used the $(\mathcal{L}_0, \mathcal{L}_1)$-smoothness assumption for each realization ξ (Assumption 4.5), in ④, we used Jensen's inequality, namely $\mathbb{E}\left[\|Z\|\right] \leq \left(\mathbb{E}\left[\|Z\|^2\right]\right)^{1/2}$, in ⑤, we applied the strong growth condition (Assumption 2.4).

Similarly, Assumption 4.5 also implies the deterministic descent inequality equation 9 with $L_0 = \mathcal{L}_0$ and $L_1 = \sqrt{\rho}\mathcal{L}_1$,

$$f(y) - f(x) = \mathbb{E}\left[f(y,\xi) - f(x,\xi)\right] \overset{①}{\leq} \mathbb{E}\left[\left\langle \nabla f(x,\xi), y-x\right\rangle + \frac{\mathcal{L}_0 + \mathcal{L}_1\|\nabla f(x,\xi)\|}{2}\|y-x\|^2\right]$$
$$= \left\langle \mathbb{E}\left[\nabla f(x,\xi)\right], y-x\right\rangle + \frac{\mathcal{L}_0 + \mathcal{L}_1\mathbb{E}\left[\|\nabla f(x,\xi)\|\right]}{2}\|y-x\|^2$$
$$\overset{②}{=} \left\langle \nabla f(x), y-x\right\rangle + \frac{\mathcal{L}_0 + \mathcal{L}_1\mathbb{E}\left[\|\nabla f(x,\xi)\|\right]}{2}\|y-x\|^2$$
$$\overset{③}{\leq} \left\langle \nabla f(x), y-x\right\rangle + \frac{\mathcal{L}_0 + \mathcal{L}_1\left[\mathbb{E}\left[\|\nabla f(x,\xi)\|^2\right]\right]^{1/2}}{2}\|y-x\|^2$$
$$\overset{④}{\leq} \left\langle \nabla f(x), y-x\right\rangle + \frac{\mathcal{L}_0 + \sqrt{\rho}\mathcal{L}_1\|\nabla f(x)\|}{2}\|y-x\|^2,$$

where in ① we used $(\mathcal{L}_0, \mathcal{L}_1)$-smoothness inequality from Assumption 4.5; in ② we used unbiasedness relation $\nabla f(x) = \mathbb{E}\left[\nabla f(x,\xi)\right]$ (Assumption 2.2); in ③ we used Jensen's inequality, $\mathbb{E}\left[\|Z\|\right] \leq (\mathbb{E}\left[\|Z\|^2\right])^{1/2}$; and in ④ we applied the strong growth condition (Assumption 2.4 with $\sigma = 0$).

(H_0, H_1)-Smoothness. In this work, we extend our analysis to (H_0, H_1)-smooth functions (see Assumption D.2). Unlike (L_0, L_1)-smoothness, the (H_0, H_1)-smoothness condition depends on the function gap rather than on the gradient norm. As a result, taking expectation directly yields the corresponding deterministic condition. Consequently, if $(\mathcal{H}_0, \mathcal{H}_1)$-smoothness (see Assumption D.1) holds for each realization, then the deterministic (H_0, H_1)-smoothness condition in Assumption D.2 is also satisfied with the same constants $H_0 = \mathcal{H}_0$ and $H_1 = \mathcal{H}_1$.

B.3 Auxiliary lemma for $(\mathcal{L}_0, \mathcal{L}_1)$-smoothness

Lemma B.2. *Let $f(\cdot,\xi)$ be $(\mathcal{L}_0, \mathcal{L}_1)$-smooth and convex for each realization $\xi \in \mathcal{D}$ (Assumption 4.5), and let the interpolation condition (Definition 2.3) hold, then $\forall x \in \mathbb{R}^d$ we have upper bound for norm of mini-batch gradient $\|\nabla f(x,\boldsymbol{\xi})\| = \left\|\frac{1}{B}\sum_{i=1}^B \nabla f(x,\xi_i)\right\|$:*

$$\|\nabla f(x,\boldsymbol{\xi})\|^2 \leq 2\left(\mathcal{L}_0 + \mathcal{L}_1\|\nabla f(x,\boldsymbol{\xi})\|\right)\left\langle \nabla f(x,\boldsymbol{\xi}), x - x^*\right\rangle. \tag{13}$$

Proof. By the $(\mathcal{L}_0, \mathcal{L}_1)$-smoothness of $f(\cdot,\xi)$, for every realization $\xi \in \mathcal{D}$ (see Assumption 4.5) we have the following descent inequality:

$$f(y,\xi) \leq f(x,\xi) + \left\langle \nabla f(x,\xi), y-x\right\rangle + \frac{\mathcal{L}_0 + \mathcal{L}_1\|\nabla f(x,\xi)\|}{2}\|y-x\|^2,$$

whenever the locality condition required by $(\mathcal{L}_0, \mathcal{L}_1)$-smoothness is satisfied. Let

$$y = x - \frac{\nabla f(x,\xi)}{\mathcal{L}_0 + \mathcal{L}_1\|\nabla f(x,\xi)\|}.$$

Then, for $\mathcal{L}_1 > 0$,

$$\|y - x\| = \frac{\|\nabla f(x,\xi)\|}{\mathcal{L}_0 + \mathcal{L}_1 \|\nabla f(x,\xi)\|} \leq \frac{1}{\mathcal{L}_1},$$

and hence the locality condition is satisfied. If $\mathcal{L}_1 = 0$, the corresponding descent inequality is global and the argument below remains valid.

By interpolation (see Definition 2.3), x^* is a minimizer of each realization $f(\cdot,\xi)$. Therefore, $f(x^*,\xi) \leq f(y,\xi)$. Applying the descent inequality with the above choice of y, we obtain

$$\begin{aligned} f(x^*,\xi) &\leq f(x,\xi) - \frac{\|\nabla f(x,\xi)\|^2}{\mathcal{L}_0 + \mathcal{L}_1 \|\nabla f(x,\xi)\|} + \frac{\mathcal{L}_0 + \mathcal{L}_1 \|\nabla f(x,\xi)\|}{2} \left(\frac{\|\nabla f(x,\xi)\|}{\mathcal{L}_0 + \mathcal{L}_1 \|\nabla f(x,\xi)\|}\right)^2 \\ &= f(x,\xi) - \frac{\|\nabla f(x,\xi)\|^2}{2(\mathcal{L}_0 + \mathcal{L}_1 \|\nabla f(x,\xi)\|)}. \end{aligned}$$

Hence,

$$f(x,\xi) - f(x^*,\xi) \geq \frac{\|\nabla f(x,\xi)\|^2}{2(\mathcal{L}_0 + \mathcal{L}_1 \|\nabla f(x,\xi)\|)}.$$

On the other hand, by convexity of $f(\cdot,\xi)$ (see Assumption 4.5),

$$f(x,\xi) - f(x^*,\xi) \leq \langle \nabla f(x,\xi), x - x^* \rangle.$$

Combining the last two inequalities gives, for each realization ξ,

$$\langle \nabla f(x,\xi), x - x^* \rangle \geq \frac{\|\nabla f(x,\xi)\|^2}{2(\mathcal{L}_0 + \mathcal{L}_1 \|\nabla f(x,\xi)\|)}.$$

Averaging this inequality over the mini-batch, we get

$$\langle \nabla f(x,\boldsymbol{\xi}), x - x^* \rangle = \frac{1}{B}\sum_{i=1}^{B} \langle \nabla f(x,\xi_i), x - x^* \rangle \geq \frac{1}{2B}\sum_{i=1}^{B} \frac{\|\nabla f(x,\xi_i)\|^2}{\mathcal{L}_0 + \mathcal{L}_1 \|\nabla f(x,\xi_i)\|}.$$

Now define, for $t \geq 0$,

$$\phi(t) = \frac{t^2}{\mathcal{L}_0 + \mathcal{L}_1 t}.$$

This function is nondecreasing and convex, since

$$\begin{aligned} \phi'(t) &= \frac{t(2\mathcal{L}_0 + \mathcal{L}_1 t)}{(\mathcal{L}_0 + \mathcal{L}_1 t)^2} \geq 0, \\ \phi''(t) &= \frac{2\mathcal{L}_0^2}{(\mathcal{L}_0 + \mathcal{L}_1 t)^3} \geq 0. \end{aligned}$$

Therefore, by Jensen's inequality and then by the triangle inequality together with the monotonicity of ϕ, we have

$$\begin{aligned} \frac{1}{B}\sum_{i=1}^{B} \phi(\|\nabla f(x,\xi_i)\|) &\geq \phi\left(\frac{1}{B}\sum_{i=1}^{B} \|\nabla f(x,\xi_i)\|\right) \\ &\geq \phi\left(\left\|\frac{1}{B}\sum_{i=1}^{B} \nabla f(x,\xi_i)\right\|\right) \\ &= \phi(\|\nabla f(x,\boldsymbol{\xi})\|) \\ &= \frac{\|\nabla f(x,\boldsymbol{\xi})\|^2}{\mathcal{L}_0 + \mathcal{L}_1 \|\nabla f(x,\boldsymbol{\xi})\|}. \end{aligned}$$

Consequently,

$$\langle \nabla f(x,\boldsymbol{\xi}), x - x^* \rangle \geq \frac{\|\nabla f(x,\boldsymbol{\xi})\|^2}{2(\mathcal{L}_0 + \mathcal{L}_1 \|\nabla f(x,\boldsymbol{\xi})\|)}.$$

Rearranging the last inequality yields

$$\|\nabla f(x,\boldsymbol{\xi})\|^2 \leq 2(\mathcal{L}_0 + \mathcal{L}_1 \|\nabla f(x,\boldsymbol{\xi})\|) \langle \nabla f(x,\boldsymbol{\xi}), x - x^* \rangle.$$

□

B.4 MONOTONE BEHAVIOR OF THE DISTANCE TO OPTIMUM FOR CLIPSGD

Lemma B.3. *Let $f(\cdot,\xi)$ be $(\mathcal{L}_0,\mathcal{L}_1)$-smooth and convex for each realization ξ (Assumption 4.5), and let the interpolation condition from Definition 2.3 hold. Let $R_0 := \left\|x^0 - x^*\right\|$ and let $\rho \geq 1$ be the strong growth parameter. Then ClipSGD with step size $\eta \leq \left[9\left(\mathcal{L}_0 + \sqrt{\rho}\mathcal{L}_1 c\right)\right]^{-1}$ guarantees*

$$\left\|x^{k+1} - x^*\right\|^2 \leq \left\|x^k - x^*\right\|^2.$$

Consequently,

$$\left\|x^k - x^*\right\| \leq \left\|x^0 - x^*\right\| = R_0 \quad \textit{for all } k \geq 0.$$

Proof. We prove the statement by induction. The claim is trivial for $k = 0$. Assume that

$$\left\|x^k - x^*\right\| \leq R_0.$$

We show that

$$\left\|x^{k+1} - x^*\right\| \leq \left\|x^k - x^*\right\|.$$

For a mini-batch $\boldsymbol{\xi}^k$, define

$$F_k(\boldsymbol{\xi}^k) := f\left(x^k, \boldsymbol{\xi}^k\right) - f\left(x^*, \boldsymbol{\xi}^k\right).$$

By the interpolation condition, x^* minimizes each stochastic realization almost surely, and hence it also minimizes the mini-batch objective. Therefore,

$$F_k(\boldsymbol{\xi}^k) \geq 0. \tag{14}$$

Moreover, by convexity equation 10,

$$F_k(\boldsymbol{\xi}^k) \leq \left\langle \nabla f\left(x^k, \boldsymbol{\xi}^k\right), x^k - x^*\right\rangle. \tag{15}$$

The ClipSGD update has the form

$$x^{k+1} = x^k - \eta \alpha_{\boldsymbol{\xi}^k} \nabla f\left(x^k, \boldsymbol{\xi}^k\right),$$

where

$$\alpha_{\boldsymbol{\xi}^k} = \min\left\{1, \frac{c}{\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\|}\right\}.$$

Then

$$\begin{aligned}\left\|x^{k+1} - x^*\right\|^2 &= \left\|x^k - x^* - \eta \alpha_{\boldsymbol{\xi}^k} \nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\|^2 \\ &= \left\|x^k - x^*\right\|^2 - 2\eta \alpha_{\boldsymbol{\xi}^k} \left\langle \nabla f\left(x^k, \boldsymbol{\xi}^k\right), x^k - x^*\right\rangle + \eta^2 \alpha_{\boldsymbol{\xi}^k}^2 \left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\|^2. \end{aligned} \tag{16}$$

By the self-bounding property of convex $(\mathcal{L}_0,\mathcal{L}_1)$-smooth functions applied to the mini-batch objective, and using $\rho \geq 1$, we have

$$\begin{aligned}\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\|^2 &\overset{equation\ 13}{\leq} 2\left(\mathcal{L}_0 + \mathcal{L}_1 \left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\|\right) \left\langle \nabla f\left(x^k, \boldsymbol{\xi}^k\right), x^k - x^*\right\rangle \\ &\leq 2\left(\mathcal{L}_0 + \sqrt{\rho}\mathcal{L}_1 \left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\|\right) \left\langle \nabla f\left(x^k, \boldsymbol{\xi}^k\right), x^k - x^*\right\rangle. \end{aligned} \tag{17}$$

Substituting equation 17 into equation 16, we obtain

$$\begin{aligned}\left\|x^{k+1} - x^*\right\|^2 \leq \left\|x^k - x^*\right\|^2 &- 2\eta \alpha_{\boldsymbol{\xi}^k} \left\langle \nabla f\left(x^k, \boldsymbol{\xi}^k\right), x^k - x^*\right\rangle \\ &+ 2\eta^2 \alpha_{\boldsymbol{\xi}^k}^2 \left(\mathcal{L}_0 + \sqrt{\rho}\mathcal{L}_1 \left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\|\right) \left\langle \nabla f\left(x^k, \boldsymbol{\xi}^k\right), x^k - x^*\right\rangle. \end{aligned} \tag{18}$$

We now consider two cases.

Case 1: $\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\| \leq c$**.** In this case, $\alpha_{\boldsymbol{\xi}^k} = 1$. Using equation 18, we get

$$\begin{aligned}\left\|x^{k+1} - x^*\right\|^2 \leq \left\|x^k - x^*\right\|^2 &- 2\eta \left\langle \nabla f\left(x^k, \boldsymbol{\xi}^k\right), x^k - x^*\right\rangle \\ &+ 2\eta^2 \left(\mathcal{L}_0 + \sqrt{\rho}\mathcal{L}_1 c\right) \left\langle \nabla f\left(x^k, \boldsymbol{\xi}^k\right), x^k - x^*\right\rangle \end{aligned}$$

$$
\begin{aligned}
&\leq \left\|x^k - x^*\right\|^2 - 2\eta \left\langle \nabla f\left(x^k, \boldsymbol{\xi}^k\right), x^k - x^* \right\rangle + \frac{2}{9}\eta \left\langle \nabla f\left(x^k, \boldsymbol{\xi}^k\right), x^k - x^* \right\rangle \\
&= \left\|x^k - x^*\right\|^2 - \frac{16}{9}\eta \left\langle \nabla f\left(x^k, \boldsymbol{\xi}^k\right), x^k - x^* \right\rangle \\
&\overset{equation\ 15}{\leq} \left\|x^k - x^*\right\|^2 - \frac{16}{9}\eta F_k(\boldsymbol{\xi}^k) \\
&\overset{equation\ 14}{\leq} \left\|x^k - x^*\right\|^2.
\end{aligned}
$$

Case 2: $\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\| > c$**.** In this case,

$$
\alpha_{\boldsymbol{\xi}^k} = \frac{c}{\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\|}.
$$

Since $\alpha_{\boldsymbol{\xi}^k} \leq 1$, we have

$$
\begin{aligned}
\alpha_{\boldsymbol{\xi}^k}\left(\mathcal{L}_0 + \sqrt{\rho}\mathcal{L}_1 \left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\|\right) &= \alpha_{\boldsymbol{\xi}^k}\mathcal{L}_0 + \sqrt{\rho}\mathcal{L}_1 c \\
&\leq \mathcal{L}_0 + \sqrt{\rho}\mathcal{L}_1 c.
\end{aligned}
$$

Using this bound in equation 18, we obtain

$$
\begin{aligned}
\left\|x^{k+1} - x^*\right\|^2 &\leq \left\|x^k - x^*\right\|^2 - 2\eta\alpha_{\boldsymbol{\xi}^k} \left\langle \nabla f\left(x^k, \boldsymbol{\xi}^k\right), x^k - x^* \right\rangle \\
&\quad + 2\eta^2\alpha_{\boldsymbol{\xi}^k}\left(\mathcal{L}_0 + \sqrt{\rho}\mathcal{L}_1 c\right) \left\langle \nabla f\left(x^k, \boldsymbol{\xi}^k\right), x^k - x^* \right\rangle \\
&\leq \left\|x^k - x^*\right\|^2 - 2\eta\alpha_{\boldsymbol{\xi}^k} \left\langle \nabla f\left(x^k, \boldsymbol{\xi}^k\right), x^k - x^* \right\rangle + \frac{2}{9}\eta\alpha_{\boldsymbol{\xi}^k} \left\langle \nabla f\left(x^k, \boldsymbol{\xi}^k\right), x^k - x^* \right\rangle \\
&= \left\|x^k - x^*\right\|^2 - \frac{16}{9}\eta\alpha_{\boldsymbol{\xi}^k} \left\langle \nabla f\left(x^k, \boldsymbol{\xi}^k\right), x^k - x^* \right\rangle \\
&\overset{equation\ 15}{\leq} \left\|x^k - x^*\right\|^2 - \frac{16}{9}\eta\alpha_{\boldsymbol{\xi}^k} F_k(\boldsymbol{\xi}^k) \\
&\overset{equation\ 14}{\leq} \left\|x^k - x^*\right\|^2.
\end{aligned}
$$

Combining both cases, we conclude that

$$
\left\|x^{k+1} - x^*\right\|^2 \leq \left\|x^k - x^*\right\|^2.
$$

Thus, by induction,

$$
\left\|x^k - x^*\right\| \leq \left\|x^0 - x^*\right\| = R_0 \quad \text{for all } k \geq 0.
$$

Finally, the same step-size condition implies

$$
\left\|x^{k+1} - x^k\right\| = \eta \left\|\text{clip}_c\left(\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right)\right\| \leq \eta c \leq \frac{1}{\sqrt{\rho}\mathcal{L}_1},
$$

which ensures that the ClipSGD update stays within the locality radius required by the corresponding (L_0, L_1)-smoothness condition with $L_1 = \sqrt{\rho}\mathcal{L}_1$. This completes the proof. □

B.5 Monotone behavior of the distance to optimum for NSGD

Lemma B.4. *Let $f(\cdot, \xi)$ be $(\mathcal{L}_0, \mathcal{L}_1)$-smooth and convex for each realization ξ (Assumption 4.5), and let the interpolation condition from Definition 2.3 hold. Let $R_0 := \left\|x^0 - x^*\right\|$ and let $\rho \geq 1$ be the strong growth parameter. Then NSGD with any $\lambda > 0$ and step size $\eta \leq \frac{\lambda}{\mathcal{L}_0 + \sqrt{\rho}\mathcal{L}_1\lambda}$ guarantees*

$$
\left\|x^{k+1} - x^*\right\|^2 \leq \left\|x^k - x^*\right\|^2.
$$

Consequently,

$$
\left\|x^k - x^*\right\| \leq \left\|x^0 - x^*\right\| = R_0 \quad \textit{for all } k \geq 0.
$$

Proof. We prove the statement by induction. The claim is trivial for $k = 0$. Assume that

$$\left\|x^k - x^*\right\| \leq R_0.$$

We show that

$$\left\|x^{k+1} - x^*\right\| \leq \left\|x^k - x^*\right\|.$$

For a mini-batch $\boldsymbol{\xi}^k$, define

$$F_k(\boldsymbol{\xi}^k) := f\left(x^k, \boldsymbol{\xi}^k\right) - f\left(x^*, \boldsymbol{\xi}^k\right).$$

By the interpolation condition, x^* minimizes each stochastic realization almost surely, and hence it also minimizes the mini-batch objective. Therefore,

$$F_k(\boldsymbol{\xi}^k) \geq 0. \tag{19}$$

Moreover, by convexity,

$$F_k(\boldsymbol{\xi}^k) \leq \left\langle \nabla f\left(x^k, \boldsymbol{\xi}^k\right), x^k - x^* \right\rangle. \tag{20}$$

Introduce the NSGD direction

$$G_\lambda\left(x^k, \boldsymbol{\xi}^k\right) := \frac{\nabla f\left(x^k, \boldsymbol{\xi}^k\right)}{\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\| + \lambda}.$$

Then the update rule of NSGD can be written as

$$x^{k+1} = x^k - \eta G_\lambda\left(x^k, \boldsymbol{\xi}^k\right).$$

Hence,

$$\begin{aligned}
\left\|x^{k+1} - x^*\right\|^2 &= \left\|x^k - x^* - \eta G_\lambda\left(x^k, \boldsymbol{\xi}^k\right)\right\|^2 \\
&= \left\|x^k - x^*\right\|^2 - 2\eta \left\langle G_\lambda\left(x^k, \boldsymbol{\xi}^k\right), x^k - x^* \right\rangle + \eta^2 \left\|G_\lambda\left(x^k, \boldsymbol{\xi}^k\right)\right\|^2 \\
&= \left\|x^k - x^*\right\|^2 - \frac{2\eta}{\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\| + \lambda} \left\langle \nabla f\left(x^k, \boldsymbol{\xi}^k\right), x^k - x^* \right\rangle \\
&\quad + \eta^2 \left(\frac{\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\|}{\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\| + \lambda}\right)^2.
\end{aligned} \tag{21}$$

By the self-bounding property of convex $(\mathcal{L}_0, \mathcal{L}_1)$-smooth functions applied to the mini-batch objective, and using $\rho \geq 1$, we have

$$\begin{aligned}
\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\|^2 &\overset{equation\ 13}{\leq} 2\left(\mathcal{L}_0 + \mathcal{L}_1 \left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\|\right) \left\langle \nabla f\left(x^k, \boldsymbol{\xi}^k\right), x^k - x^* \right\rangle \\
&\leq 2\left(\mathcal{L}_0 + \sqrt{\rho}\mathcal{L}_1 \left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\|\right) \left\langle \nabla f\left(x^k, \boldsymbol{\xi}^k\right), x^k - x^* \right\rangle ..
\end{aligned} \tag{22}$$

Substituting equation 22 into equation 21, we obtain

$$\begin{aligned}
\left\|x^{k+1} - x^*\right\|^2 &\leq \left\|x^k - x^*\right\|^2 - \frac{2\eta}{\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\| + \lambda} \left\langle \nabla f\left(x^k, \boldsymbol{\xi}^k\right), x^k - x^* \right\rangle \\
&\quad + \frac{2\eta^2\left(\mathcal{L}_0 + \sqrt{\rho}\mathcal{L}_1 \left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\|\right)}{\left(\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\| + \lambda\right)^2} \left\langle \nabla f\left(x^k, \boldsymbol{\xi}^k\right), x^k - x^* \right\rangle \\
&= \left\|x^k - x^*\right\|^2 \\
&\quad - \frac{2\eta \left\langle \nabla f\left(x^k, \boldsymbol{\xi}^k\right), x^k - x^* \right\rangle}{\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\| + \lambda} \left(1 - \eta \frac{\mathcal{L}_0 + \sqrt{\rho}\mathcal{L}_1 \left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\|}{\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\| + \lambda}\right).
\end{aligned} \tag{23}$$

It remains to show that the term in parentheses is non-negative. For any $G \geq 0$,

$$\frac{\mathcal{L}_0 + \sqrt{\rho}\mathcal{L}_1\lambda}{\lambda} - \frac{\mathcal{L}_0 + \sqrt{\rho}\mathcal{L}_1 G}{G + \lambda} = \frac{\mathcal{L}_0 G + \sqrt{\rho}\mathcal{L}_1\lambda^2}{\lambda(G + \lambda)} \geq 0.$$

Applying this inequality with $G = \left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\|$, and using the step-size condition, gives

$$1 - \eta \frac{\mathcal{L}_0 + \sqrt{\rho}\mathcal{L}_1 \left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\|}{\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\| + \lambda} \geq 0.$$

Substituting this into equation 23, as well as using $F_k(\boldsymbol{\xi}^k) \leq \left\langle \nabla f\left(x^k, \boldsymbol{\xi}^k\right), x^k - x^* \right\rangle$ from equation 20 $F_k(\boldsymbol{\xi}^k) \geq 0$ from equation 19, we conclude that

$$\left\|x^{k+1} - x^*\right\|^2 \leq \left\|x^k - x^*\right\|^2.$$

Thus, by induction,

$$\left\|x^k - x^*\right\| \leq \left\|x^0 - x^*\right\| = R_0 \quad \text{for all } k \geq 0.$$

Finally, the same step-size condition implies

$$\left\|x^{k+1} - x^k\right\| = \eta \left\| \frac{\nabla f\left(x^k, \boldsymbol{\xi}^k\right)}{\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\| + \lambda} \right\| \leq \eta \leq \frac{1}{\sqrt{\rho}\mathcal{L}_1},$$

which ensures that the NSGD update stays within the locality radius required by the corresponding (L_0, L_1)-smoothness condition with $L_1 = \sqrt{\rho}\mathcal{L}_1$. This completes the proof. □

B.6 Variance for mini-batch stochastic gradient

If we use the mini-batch stochastic gradient $\nabla f(x, \boldsymbol{\xi}) = \frac{1}{B}\sum_{i=1}^{B} \nabla f(x, \xi_i)$ instead of the stochastic gradient $\nabla f(x, \xi)$, then the variance equation 11 takes the following form:

$$\begin{aligned}
\mathbb{E}\left[\left\|\nabla f(x, \boldsymbol{\xi}) - \nabla f(x)\right\|^2\right] &\overset{①}{=} \mathbb{E}\left[\left\|\frac{1}{B}\sum_{i=1}^{B} \nabla f(x, \xi_i) - \nabla f(x)\right\|^2\right] \\
&\overset{②}{=} \mathbb{E}\left[\left\|\frac{1}{B}\sum_{i=1}^{B} \left(\nabla f(x, \xi_i) - \nabla f(x)\right)\right\|^2\right] \\
&= \frac{1}{B^2}\mathbb{E}\left[\left\|\sum_{i=1}^{B} \left(\nabla f(x, \xi_i) - \nabla f(x)\right)\right\|^2\right] \\
&\overset{③}{=} \frac{1}{B^2}\sum_{i=1}^{B}\sum_{j=1}^{B} \mathbb{E}\left[\left\langle \nabla f(x, \xi_i) - \nabla f(x), \nabla f(x, \xi_j) - \nabla f(x)\right\rangle\right] \\
&= \frac{1}{B^2}\sum_{i=1}^{B} \mathbb{E}\left[\left\|\nabla f(x, \xi_i) - \nabla f(x)\right\|^2\right] \\
&\qquad + \frac{1}{B^2}\sum_{i\neq j}^{B} \left\langle \mathbb{E}\left[\nabla f(x, \xi_i)\right] - \nabla f(x), \mathbb{E}\left[\nabla f(x, \xi_j)\right] - \nabla f(x)\right\rangle \\
&\overset{④}{=} \frac{1}{B^2}\sum_{i=1}^{B} \mathbb{E}\left[\left\|\nabla f(x, \xi_i) - \nabla f(x)\right\|^2\right] \\
&\overset{\text{equation 11}}{\leq} \frac{1}{B^2}\sum_{i=1}^{B} \left[(\rho - 1)\left\|\nabla f(x)\right\|^2 + \sigma^2\right] \\
&= \frac{\rho - 1}{B}\left\|\nabla f(x)\right\|^2 + \frac{\sigma^2}{B},
\end{aligned} \tag{24}$$

where in ① we used $\nabla f(x, \boldsymbol{\xi}) = \frac{1}{B}\sum_{i=1}^{B} \nabla f(x, \xi_i)$, in ② we used $\nabla f(x) = \frac{1}{B}\sum_{i=1}^{B} \nabla f(x)$, in ③ we utilized the definition of the Euclidean norm, in ④ we used that $\mathbb{E}\left[\nabla f(x, \xi_i)\right] = \nabla f(x)$.

B.7 Motivation for the generalized strong growth condition under heavy-tailed noise

Example B.5. Consider the function $f(x) = \frac{1}{2}\|x\|^2$ with gradient oracle $\nabla f(x,\xi) = (Z+1)x$ (where Z is a random variable with a symmetric Pareto distribution). Then, for $\alpha \in (p, 2]$ we obtain the strong growth condition under heavy-tailed noise $\mathbb{E}\left[\|\nabla f(x,\xi)\|^p\right] \leq 2^{p-1}\rho\|\nabla f(x)\|^p$ with $\rho = \frac{\alpha}{\alpha-p} + 1 < \infty$, and $\mathbb{E}\left[\|\nabla f(x,\xi)\|^2\right] = \infty$.

Proof. We start by estimating the p-th moment $\mathbb{E}\left[\|\nabla f(x,\xi)\|^p\right] \overset{①}{=} \mathbb{E}\left[\|Zx + x\|^p\right]$:

$$\mathbb{E}\left[\|Zx+x\|^p\right] \overset{②}{\leq} 2^{p-1}\left(\mathbb{E}\left[|Z|^p\right]+1\right)\|x\|^p \overset{③}{=} 2^{p-1}\left(\alpha\int_1^\infty |z|^{p-\alpha-1}dz + 1\right)\|x\|^p$$
$$\overset{④}{=} 2^{p-1}\underbrace{\left(\frac{\alpha}{\alpha-p}+1\right)}_{=:\rho<\infty}\|\nabla f(x)\|^p,$$

where in ① we used $\nabla f(x,\xi) = (Z+1)x$, in ② we applied Jensen's inequality for finite form, in ③ we applied the definition of expectation with probability density function of a symmetric Pareto random variable $f_Z(z) = \frac{\alpha}{2}|z|^{-(\alpha+1)}$ with $|z| \geq 1$ and $|z+1| \leq 2|z|$, in ④ we noted that $\int_1^\infty |z|^{p-\alpha-1}dz = \frac{1}{\alpha-p} < \infty$ since $p < \alpha$, and used $\nabla f(x) = x$.

To analyze the second moment, we follow the same steps, substituting p with 2. In this case, we obtain $\int_1^\infty |z|^{2-\alpha-1}dz = \infty$, since $\alpha \leq (p, 2]$. Consequently, the second moment is unbounded: $\mathbb{E}\left[\|\nabla f(x,\xi)\|^2\right] = \infty$.

□

C Missing Proofs for Stochastic Part (ClipSGD & NSGD)

C.1 Missing proofs for clipped stochastic gradient descent method

This section provides the proofs for ClipSGD in the non-convex and convex settings. The analysis follows the same general line as Koloskova et al. (2023), but the stochastic noise is controlled through the generalized strong growth condition and a mini-batch gradient. The main proof ingredient is an alignment argument: once the batch size is proportional to the growth parameter ρ, the clipped stochastic direction remains sufficiently aligned with the full gradient.

C.1.1 Non-convex setup under (L_0, L_1)-smooth and generalized strong growth conditions

The proof is divided into two regimes: $c \leq \frac{6\sqrt{2}\sigma}{\sqrt{B}}$ and $c > \frac{6\sqrt{2}\sigma}{\sqrt{B}}$. In each regime, we further split according to the deterministic gradient norm $\left\|\nabla f\left(x^k\right)\right\|$.

Proof. Throughout the proof, we use the shorthand

$$F_k := f\left(x^k\right) - f^*.$$

In the one-step estimates below, expectations are taken with respect to the fresh mini-batch $\boldsymbol{\xi}^k$ and are written explicitly as conditional expectations given the current iterate x^k. After taking full expectation, we return to the notation $\mathbb{E}[\cdot]$.

We start from the descent inequality for (L_0, L_1)-smooth functions:

$$f\left(x^{k+1}\right) - f\left(x^k\right) \overset{①}{\leq} \left\langle \nabla f\left(x^k\right), x^{k+1} - x^k\right\rangle + \frac{L_0 + L_1\left\|\nabla f\left(x^k\right)\right\|}{2}\left\|x^{k+1} - x^k\right\|^2$$
$$\overset{②}{=} -\eta\left\langle \nabla f\left(x^k\right), \mathrm{clip}_c\left(\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right)\right\rangle$$
$$+ \frac{\eta^2\left(L_0 + L_1\left\|\nabla f\left(x^k\right)\right\|\right)}{2}\left\|\mathrm{clip}_c\left(\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right)\right\|^2, \tag{25}$$

Here, in ① we used the descent inequality equation 9; its locality condition $\left\|x^{k+1}-x^k\right\| \leq \frac{1}{L_1}$ is verified in Remark C.1. In ② we used the ClipSGD update $x^{k+1}=x^k-\eta \cdot \operatorname{clip}_c\left(\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right)$.

We now split the proof into two regimes.

First regime: $c \leq \frac{6\sqrt{2}\sigma}{\sqrt{B}}$**.** We consider two cases: $\left\|\nabla f\left(x^k\right)\right\| \geq \frac{6\sqrt{2}\sigma}{\sqrt{B}}$ and $\left\|\nabla f\left(x^k\right)\right\| < \frac{6\sqrt{2}\sigma}{\sqrt{B}}$.

1.1) The case $\left\|\nabla f\left(x^k\right)\right\| \geq \frac{6\sqrt{2}\sigma}{\sqrt{B}}$.

Define the bad event

$$\delta=\mathbb{1}\left\{\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)-\nabla f\left(x^k\right)\right\|>3 \sqrt{\frac{\rho-1}{B}\left\|\nabla f\left(x^k\right)\right\|^2+\frac{\sigma^2}{B}}\right\}.$$

We decompose the first term in equation 25 according to this event:

$$\begin{aligned}
\mathbb{E}\left[-\eta\left\langle\nabla f\left(x^k\right), \operatorname{clip}_c\left(\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right)\right\rangle \mid x^k\right] &\overset{\circledast}{=} \mathbb{E}\left[-\eta \alpha_{\boldsymbol{\xi}^k}\left\langle\nabla f\left(x^k\right), \nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\rangle \mid x^k\right] \\
&\leq \mathbb{P}\left(\delta=0 \mid x^k\right) \underbrace{\mathbb{E}\left[-\eta \alpha_{\boldsymbol{\xi}^k}\left\langle\nabla f\left(x^k\right), \nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\rangle \mid x^k, \delta=0\right]}_{=: T_1} \\
&\quad+\mathbb{P}\left(\delta=1 \mid x^k\right) \underbrace{\mathbb{E}\left[-\eta \alpha_{\boldsymbol{\xi}^k}\left\langle\nabla f\left(x^k\right), \nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\rangle \mid x^k, \delta=1\right]}_{=: T_2},
\end{aligned} \tag{26}$$

where in ⊛ we used $\operatorname{clip}_c\left(\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right)=\alpha_{\boldsymbol{\xi}^k} \cdot \nabla f\left(x^k, \boldsymbol{\xi}^k\right)$. We first estimate the probability of the bad event. By Markov's inequality,

$$\begin{aligned}
\mathbb{P}\left(\delta=1 \mid x^k\right) &= \mathbb{P}\left(\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)-\nabla f\left(x^k\right)\right\|^2>9\left(\frac{\rho-1}{B}\left\|\nabla f\left(x^k\right)\right\|^2+\frac{\sigma^2}{B}\right) \mid x^k\right) \\
&\leq \frac{\mathbb{E}\left[\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)-\nabla f\left(x^k\right)\right\|^2 \mid x^k\right]}{9(\rho-1)\left\|\nabla f\left(x^k\right)\right\|^2+\sigma^2} \cdot B \\
&\overset{equation\ 24}{\leq} \frac{1}{9}.
\end{aligned} \tag{27}$$

Consequently, $\mathbb{P}\left(\delta=0 \mid x^k\right)=1-\mathbb{P}\left(\delta=1 \mid x^k\right) \geq \frac{8}{9}$. We next bound the contribution of the good event:

$$\begin{aligned}
&\mathbb{E}\left[-\eta \alpha_{\boldsymbol{\xi}^k}\left\langle\nabla f\left(x^k\right), \nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\rangle \mid x^k, \delta=0\right] \\
&=\mathbb{E}\left[-\eta \alpha_{\boldsymbol{\xi}^k}\left\langle\nabla f\left(x^k\right), \nabla f\left(x^k, \boldsymbol{\xi}^k\right) \pm \nabla f\left(x^k\right)\right\rangle \mid x^k, \delta=0\right] \\
&=\mathbb{E}\left[-\eta \alpha_{\boldsymbol{\xi}^k}\left\|\nabla f\left(x^k\right)\right\|^2-\eta \alpha_{\boldsymbol{\xi}^k}\left\langle\nabla f\left(x^k\right), \nabla f\left(x^k, \boldsymbol{\xi}^k\right)-\nabla f\left(x^k\right)\right\rangle \mid x^k, \delta=0\right] \\
&\overset{equation\ 7}{\leq} \mathbb{E}\left[-\eta \alpha_{\boldsymbol{\xi}^k}\left\|\nabla f\left(x^k\right)\right\|^2+\eta \alpha_{\boldsymbol{\xi}^k}\left\|\nabla f\left(x^k\right)\right\|\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)-\nabla f\left(x^k\right)\right\| \mid x^k, \delta=0\right] \\
&\overset{①}{\leq} \mathbb{E}\left[-\eta \alpha_{\boldsymbol{\xi}^k}\left\|\nabla f\left(x^k\right)\right\|^2+3 \eta \alpha_{\boldsymbol{\xi}^k}\left\|\nabla f\left(x^k\right)\right\| \sqrt{\frac{\rho-1}{B}\left\|\nabla f\left(x^k\right)\right\|^2+\frac{\sigma^2}{B}} \mid x^k, \delta=0\right] \\
&\overset{②}{\leq} \mathbb{E}\left[-\eta \alpha_{\boldsymbol{\xi}^k}\left\|\nabla f\left(x^k\right)\right\|^2+3 \eta \alpha_{\boldsymbol{\xi}^k} \sqrt{\frac{\rho-1}{B}+\frac{1}{72}} \cdot\left\|\nabla f\left(x^k\right)\right\|^2 \mid x^k, \delta=0\right] \\
&\overset{③}{\leq} \mathbb{E}\left[-\eta \alpha_{\boldsymbol{\xi}^k}\left\|\nabla f\left(x^k\right)\right\|^2+3 \eta \alpha_{\boldsymbol{\xi}^k} \sqrt{\frac{1}{72}+\frac{1}{72}} \cdot\left\|\nabla f\left(x^k\right)\right\|^2 \mid x^k, \delta=0\right] \\
&=\mathbb{E}\left[-\eta \alpha_{\boldsymbol{\xi}^k}\left\|\nabla f\left(x^k\right)\right\|^2+\frac{\eta \alpha_{\boldsymbol{\xi}^k}}{2}\left\|\nabla f\left(x^k\right)\right\|^2 \mid x^k, \delta=0\right] \\
&=\mathbb{E}\left[-\frac{\eta \alpha_{\boldsymbol{\xi}^k}}{2}\left\|\nabla f\left(x^k\right)\right\|^2 \mid x^k, \delta=0\right]
\end{aligned}$$

$$\overset{④}{\leq} -\frac{\eta c}{4}\left\|\nabla f\left(x^k\right)\right\|, \tag{28}$$

Here, in ① we used the definition of the event $\{\delta = 0\}$, in ② we used $\frac{\sigma}{\sqrt{B}} \leq \frac{\left\|\nabla f\left(x^k\right)\right\|}{6\sqrt{2}}$, and in ③ we used $B \geq 72(\rho - 1)$. In ④ we used the bound $\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\| \leq 2\left\|\nabla f\left(x^k\right)\right\|$, which implies $\alpha_{\boldsymbol{\xi}^k} \geq c/(2\left\|\nabla f\left(x^k\right)\right\|)$. On the complementary event, we only use the boundedness induced by clipping:

$$\begin{aligned}\mathbb{E}\left[-\eta\alpha_{\boldsymbol{\xi}^k}\left\langle\nabla f\left(x^k\right), \nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\rangle \mid x^k, \delta = 1\right] &\overset{equation\ 7}{\leq} \eta\left\|\nabla f\left(x^k\right)\right\| \mathbb{E}\left[\alpha_{\boldsymbol{\xi}^k}\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\| \mid x^k, \delta = 1\right] \\ &\overset{①}{\leq} \eta c\left\|\nabla f\left(x^k\right)\right\|, \end{aligned} \tag{29}$$

where in ① we used $\left\|\text{clip}_c\left(\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right)\right\| \leq c$.

Combining the good and bad events yields

$$\begin{aligned}\mathbb{E}\left[-\eta\left\langle\nabla f\left(x^k\right), \text{clip}_c\left(\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right)\right\rangle \mid x^k\right] &\overset{equation\ 26}{\leq} \mathbb{P}\left(\delta = 0 \mid x^k\right) \underbrace{\mathbb{E}\left[-\eta\alpha_{\boldsymbol{\xi}^k}\left\langle\nabla f\left(x^k\right), \nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\rangle \mid x^k, \delta = 0\right]}_{=:T_1} \\ &\quad + \mathbb{P}\left(\delta = 1 \mid x^k\right) \underbrace{\mathbb{E}\left[-\eta\alpha_{\boldsymbol{\xi}^k}\left\langle\nabla f\left(x^k\right), \nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\rangle \mid x^k, \delta = 1\right]}_{=:T_2} \\ &\overset{equation\ 28, equation\ 29}{\leq} -\frac{\eta c}{4}\left\|\nabla f\left(x^k\right)\right\| \cdot \mathbb{P}\left(\delta = 0 \mid x^k\right) + \eta c\left\|\nabla f\left(x^k\right)\right\| \cdot \mathbb{P}\left(\delta = 1 \mid x^k\right) \\ &= -\frac{\eta c}{4}\left\|\nabla f\left(x^k\right)\right\| \cdot \left(1 - \mathbb{P}\left(\delta = 1 \mid x^k\right)\right) + \eta c\left\|\nabla f\left(x^k\right)\right\| \cdot \mathbb{P}\left(\delta = 1 \mid x^k\right) \\ &\overset{equation\ 27}{\leq} -\eta c\left(\frac{8}{9} \cdot \frac{1}{4} - \frac{1}{9}\right)\left\|\nabla f\left(x^k\right)\right\| \\ &= -\frac{\eta c}{9}\left\|\nabla f\left(x^k\right)\right\|. \end{aligned} \tag{30}$$

Substituting this estimate into equation 25, we obtain:

$$\begin{aligned}\mathbb{E}\left[F_{k+1} \mid x^k\right] - F_k &\overset{equation\ 25}{\leq} -\mathbb{E}\left[\eta\left\langle\nabla f\left(x^k\right), \text{clip}_c\left(\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right)\right\rangle \mid x^k\right] \\ &\quad + \frac{\eta^2\left(L_0 + L_1\left\|\nabla f\left(x^k\right)\right\|\right)}{2}\mathbb{E}\left[\left\|\text{clip}_c\left(\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right)\right\|^2 \mid x^k\right] \\ &\overset{equation\ 30}{\leq} -\frac{\eta c}{9}\left\|\nabla f\left(x^k\right)\right\| + \frac{\eta^2\left(L_0 + L_1\left\|\nabla f\left(x^k\right)\right\|\right)}{2}\mathbb{E}\left[\left\|\text{clip}_c\left(\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right)\right\|^2 \mid x^k\right] \\ &\overset{①}{\leq} -\frac{\eta c}{9}\left\|\nabla f\left(x^k\right)\right\| + \frac{\eta^2\left(L_0 + L_1\left\|\nabla f\left(x^k\right)\right\|\right)}{2} \cdot c^2 \\ &\overset{②}{\leq} -\frac{\eta c}{9}\left\|\nabla f\left(x^k\right)\right\| + \frac{\eta^2\left(L_0 + L_1 c\right)}{2} \cdot c \cdot \left\|\nabla f\left(x^k\right)\right\| \\ &\overset{③}{\leq} -\frac{\eta c}{9}\left\|\nabla f\left(x^k\right)\right\| + \frac{\eta c}{18}\left\|\nabla f\left(x^k\right)\right\| \\ &= -\frac{\eta c}{18}\left\|\nabla f\left(x^k\right)\right\|, \end{aligned}$$

where in ① we used $\left\|\text{clip}_c\left(\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right)\right\| \leq c$, in ② we used $\left\|\nabla f\left(x^k\right)\right\| \geq c$, in ③ we chose $\eta \leq \left[9\left(L_0 + L_1 c\right)\right]^{-1}$.

Hence, in the first case,

$$\left\|\nabla f\left(x^k\right)\right\| \leq \frac{18\left(F_k - \mathbb{E}\left[F_{k+1} \mid x^k\right]\right)}{\eta c}. \tag{31}$$

1.2) The case $\left\|\nabla f\left(x^k\right)\right\| < \frac{6\sqrt{2}\sigma}{\sqrt{B}}$.

This case is immediate from the case condition:

$$\left\|\nabla f\left(x^k\right)\right\| \leq \frac{6\sqrt{2}\sigma}{\sqrt{B}}.$$

Combining the first regime. Since both terms on the right-hand side are non-negative, the bounds from the two cases imply, for all $k \in [0, N-1]$,

$$\left\|\nabla f\left(x^k\right)\right\| \leq \frac{18\left(F_k - \mathbb{E}\left[F_{k+1} \mid x^k\right]\right)}{\eta c} + \frac{6\sqrt{2}\sigma}{\sqrt{B}}.$$

Taking full expectation, summing over all indices $k \in [0, N-1]$, and using telescoping of $\mathbb{E}[F_k]$ gives

$$\min_{k\in[0,N-1]} \mathbb{E}\left[\left\|\nabla f\left(x^k\right)\right\|\right] \leq \frac{1}{N}\sum_{k=0}^{N-1} \mathbb{E}\left[\left\|\nabla f\left(x^k\right)\right\|\right] \leq \frac{18F_0}{\eta c N} + \frac{6\sqrt{2}\sigma}{\sqrt{B}}. \tag{32}$$

Second regime: $c > \frac{6\sqrt{2}\sigma}{\sqrt{B}}$**.** We now consider two cases: $\left\|\nabla f\left(x^k\right)\right\| \geq c$ and $\left\|\nabla f\left(x^k\right)\right\| < c$.

2.1) The case $\left\|\nabla f\left(x^k\right)\right\| \geq c$.

This case is covered by the alignment argument used in 1.1). Indeed, the condition of 1.1) is

$$\left\|\nabla f\left(x^k\right)\right\| \geq \frac{6\sqrt{2}\sigma}{\sqrt{B}} \geq c,$$

while the present case satisfies

$$\left\|\nabla f\left(x^k\right)\right\| \geq c > \frac{6\sqrt{2}\sigma}{\sqrt{B}}.$$

Thus, the same proof applies without repetition.

Therefore,

$$c \cdot \left\|\nabla f\left(x^k\right)\right\| \overset{equation\ 31}{\leq} \frac{18\left(F_k - \mathbb{E}\left[F_{k+1} \mid x^k\right]\right)}{\eta}. \tag{33}$$

2.2) The case $\left\|\nabla f\left(x^k\right)\right\| < c$.

In this case, $\nabla f\left(x^k\right) = \text{clip}_c\left(\nabla f\left(x^k\right)\right)$. From equation 25, we have:

$$\begin{aligned}
\mathbb{E}\left[f\left(x^{k+1}\right) \mid x^k\right] - f\left(x^k\right) &\overset{equation\ 25}{\leq} -\eta\mathbb{E}\left[\left\langle \nabla f\left(x^k\right), \text{clip}_c\left(\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right)\right\rangle \mid x^k\right] \\
&\quad + \frac{\eta^2\left(L_0 + L_1\left\|\nabla f\left(x^k\right)\right\|\right)}{2}\mathbb{E}\left[\left\|\text{clip}_c\left(\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right)\right\|^2 \mid x^k\right] \\
&\overset{①}{\leq} -\eta\mathbb{E}\left[\left\langle \nabla f\left(x^k\right), \text{clip}_c\left(\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right)\right\rangle \mid x^k\right] \\
&\quad + \frac{\eta^2\left(L_0 + L_1 c\right)}{2}\mathbb{E}\left[\left\|\text{clip}_c\left(\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right)\right\|^2 \mid x^k\right] \\
&\overset{②}{=} -\frac{\eta}{2}\left\|\nabla f\left(x^k\right)\right\|^2 - \frac{\eta}{2}\mathbb{E}\left[\left\|\text{clip}_c\left(\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right)\right\|^2 \mid x^k\right] \\
&\quad + \frac{\eta}{2}\mathbb{E}\left[\left\|\text{clip}_c\left(\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right) - \nabla f\left(x^k\right)\right\|^2 \mid x^k\right] \\
&\quad + \frac{\eta^2\left(L_0 + L_1 c\right)}{2}\mathbb{E}\left[\left\|\text{clip}_c\left(\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right)\right\|^2 \mid x^k\right] \\
&\overset{③}{=} -\frac{\eta}{2}\left\|\nabla f\left(x^k\right)\right\|^2 + \frac{\eta}{2}\mathbb{E}\left[\left\|\text{clip}_c\left(\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right) - \text{clip}_c\left(\nabla f\left(x^k\right)\right)\right\|^2 \mid x^k\right] \\
&\quad - \frac{\eta}{2}\left(1 - \eta\left(L_0 + L_1 c\right)\right)\mathbb{E}\left[\left\|\text{clip}_c\left(\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right)\right\|^2 \mid x^k\right] \\
&\overset{④}{\leq} -\frac{\eta}{2}\left\|\nabla f\left(x^k\right)\right\|^2 + \frac{\eta}{2}\mathbb{E}\left[\left\|\text{clip}_c\left(\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right) - \text{clip}_c\left(\nabla f\left(x^k\right)\right)\right\|^2 \mid x^k\right],
\end{aligned}$$

where in ① we used $\left\|\nabla f\left(x^k\right)\right\| < c$, in ② we used inequality equation 6 with $a = \nabla f\left(x^k\right)$ and $b = \text{clip}_c\left(\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right)$, in ③ we used $\nabla f\left(x^k\right) = \text{clip}_c\left(\nabla f\left(x^k\right)\right)$, in ④ we chose $\eta \leq [9\left(L_0 + L_1 c\right)]^{-1}$.

Since clipping is the projection onto the Euclidean ball of radius c, it is a Lipschitz operator with Lipschitz constant 1. Therefore,

$$\begin{aligned}
\mathbb{E}\left[f\left(x^{k+1}\right) \mid x^k\right] - f\left(x^k\right) &\leq -\frac{\eta}{2}\left\|\nabla f\left(x^k\right)\right\|^2 + \frac{\eta}{2}\mathbb{E}\left[\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right) - \nabla f\left(x^k\right)\right\|^2 \mid x^k\right] \\
&\overset{equation\ 24}{\leq} -\frac{\eta}{2}\left\|\nabla f\left(x^k\right)\right\|^2 + \frac{\eta}{2}\left(\frac{(\rho-1)}{B}\left\|\nabla f\left(x^k\right)\right\|^2 + \frac{\sigma^2}{B}\right) \\
&\overset{①}{\leq} -\frac{\eta}{2}\left\|\nabla f\left(x^k\right)\right\|^2 + \frac{\eta}{144}\left\|\nabla f\left(x^k\right)\right\|^2 + \frac{\eta\sigma^2}{2B} \\
&\leq -\frac{\eta}{2}\left\|\nabla f\left(x^k\right)\right\|^2 + \frac{\eta}{9}\left\|\nabla f\left(x^k\right)\right\|^2 + \frac{\eta\sigma^2}{2B} \\
&= -\frac{7\eta}{18}\left\|\nabla f\left(x^k\right)\right\|^2 + \frac{\eta\sigma^2}{2B} \\
&\leq -\frac{\eta}{18}\left\|\nabla f\left(x^k\right)\right\|^2 + \frac{\eta\sigma^2}{2B},
\end{aligned}$$

where in ① we chose $B \geq 72(\rho - 1)$.

Hence, in the second case,

$$\left\|\nabla f\left(x^k\right)\right\|^2 \leq \frac{18\left(F_k - \mathbb{E}\left[F_{k+1} \mid x^k\right]\right)}{\eta} + \frac{9\sigma^2}{B}, \tag{34}$$

Combining the second regime. Let $\mathcal{F}_k$ denote the history before drawing the k-th mini-batch and define the $\mathcal{F}_k$-measurable indicator

$$I_k := \mathbf{1}\left\{\left\|\nabla f\left(x^k\right)\right\| \geq c\right\}, \qquad g_k := \left\|\nabla f\left(x^k\right)\right\|.$$

Multiplying equation 33 by I_k, multiplying equation 34 by $1 - I_k$, and adding the resulting inequalities gives, conditionally on $\mathcal{F}_k$,

$$I_k c g_k + (1 - I_k) g_k^2 \leq \frac{18}{\eta}\left(F_k - \mathbb{E}\left[F_{k+1} \mid \mathcal{F}_k\right]\right) + (1 - I_k)\frac{9\sigma^2}{B}. \tag{35}$$

Taking full expectation, summing over $k = 0, \ldots, N-1$, and using $F_N \geq 0$, we obtain

$$\frac{1}{N}\sum_{k=0}^{N-1}\mathbb{E}\left[I_k c g_k + (1 - I_k) g_k^2\right] \leq D_N, \qquad D_N := \frac{18F_0}{\eta N} + \frac{9\sigma^2}{B}. \tag{36}$$

Since both terms on the left-hand side are non-negative,

$$\frac{1}{N}\sum_{k=0}^{N-1}\mathbb{E}\left[I_k g_k\right] \leq \frac{D_N}{c}, \qquad \frac{1}{N}\sum_{k=0}^{N-1}\mathbb{E}\left[(1 - I_k) g_k^2\right] \leq D_N.$$

Moreover, Cauchy–Schwarz on the product probability space of a uniformly sampled index and the algorithmic randomness yields

$$\frac{1}{N}\sum_{k=0}^{N-1}\mathbb{E}\left[(1 - I_k) g_k\right] \leq \sqrt{\frac{1}{N}\sum_{k=0}^{N-1}\mathbb{E}\left[(1 - I_k) g_k^2\right]}\sqrt{\frac{1}{N}\sum_{k=0}^{N-1}\mathbb{E}\left[1 - I_k\right]} \leq \sqrt{D_N}.$$

Combining the two contributions and using $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ gives, in the second regime,

$$\begin{aligned}
\min_{k \in [0, N-1]}\mathbb{E}\left[g_k\right] &\leq \frac{1}{N}\sum_{k=0}^{N-1}\mathbb{E}\left[g_k\right] \leq \sqrt{D_N} + \frac{D_N}{c} \\
&\leq \sqrt{\frac{18F_0}{\eta N}} + \frac{18F_0}{\eta c N} + \frac{3\sigma}{\sqrt{B}} + \frac{9\sigma^2}{cB}.
\end{aligned} \tag{37}$$

Together with the first-regime estimate equation 32, this implies the uniform bound

$$\min_{k\in[0,N-1]} \mathbb{E}\left[\left\|\nabla f\left(x^k\right)\right\|\right] \leq \sqrt{\frac{18F_0}{\eta N}} + \frac{18F_0}{\eta c N} + \frac{6\sqrt{2}\sigma}{\sqrt{B}} + \frac{9\sigma^2}{cB}.$$

This proves the stated convergence rate for step size $\eta \leq \left[9\left(L_0 + L_1 c\right)\right]^{-1}$ and batch size $B \geq 72(\rho - 1)$. □

Remark C.1. *In the first step of equation 25, we use the descent lemma for (L_0, L_1)-smooth functions equation 9 (see Appendix A.1 in Zhang et al., 2020a); the condition $\left\|x^{k+1} - x^k\right\| \leq 1/L_1$ required for its application is guaranteed by the chosen step size $\eta \leq \left[9\left(L_0 + L_1 c\right)\right]^{-1}$:*

$$\left\|x^{k+1} - x^k\right\| \overset{①}{=} \eta \left\|clip_c\left(\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right)\right\| \overset{②}{\leq} \eta c \overset{③}{\leq} \frac{c}{9\left(L_0 + L_1 c\right)} \overset{④}{\leq} \frac{c}{9L_1 c} \overset{④}{\leq} \frac{1}{L_1},$$

where in ① we used $x^{k+1} = x^k - \eta \cdot clip_c\left(\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right)$, in ② we used $\left\|clip_c\left(\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right)\right\| = \left\|\min\left\{1, \frac{c}{\|\nabla f(x^k, \boldsymbol{\xi}^k)\|}\right\} \nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\| \leq c$, in ③ we used $\eta \leq \left[9\left(L_0 + L_1 c\right)\right]^{-1}$, where in ④ we used the fact that, for $0 < b \leq a$, one has $\frac{1}{a} \leq \frac{1}{b}$.

C.1.2 Convex setup under $(\mathcal{L}_0, \mathcal{L}_1)$-smooth and strong growth conditions

We now prove the convex ClipSGD result. The proof follows the same alignment argument as in the non-convex case, but convexity allows us to convert gradient-norm bounds into a recursion for the optimality gap.

Proof. Throughout the proof, we use the shorthand

$$F_k := f\left(x^k\right) - f^*.$$

In the one-step estimates below, expectations are taken with respect to the fresh mini-batch $\boldsymbol{\xi}^k$ and are written explicitly as conditional expectations given the current iterate x^k. After taking full expectation, we return to the notation $\mathbb{E}[\cdot]$. By convexity, the Cauchy–Schwarz inequality, and the monotonicity of the distance to the optimum from Lemma B.3, we have

$$F_k \leq \left\langle \nabla f\left(x^k\right), x^k - x^* \right\rangle \leq \left\|\nabla f\left(x^k\right)\right\| \left\|x^k - x^*\right\| \leq R_0 \left\|\nabla f\left(x^k\right)\right\|. \tag{38}$$

We start from the descent inequality for (L_0, L_1)-smooth functions with constants $L_0 = \mathcal{L}_0$ and $L_1 = \sqrt{\rho}\mathcal{L}_1$ (see equation 12 in Section B.2):

$$\begin{aligned} f\left(x^{k+1}\right) - f\left(x^k\right) &\overset{①}{\leq} \left\langle \nabla f\left(x^k\right), x^{k+1} - x^k \right\rangle + \frac{\mathcal{L}_0 + \sqrt{\rho}\mathcal{L}_1 \left\|\nabla f\left(x^k\right)\right\|}{2} \left\|x^{k+1} - x^k\right\|^2 \\ &\overset{②}{=} -\eta \left\langle \nabla f\left(x^k\right), \mathrm{clip}_c\left(\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right) \right\rangle \\ &\quad + \frac{\eta^2\left(\mathcal{L}_0 + \sqrt{\rho}\mathcal{L}_1 \left\|\nabla f\left(x^k\right)\right\|\right)}{2} \left\|\mathrm{clip}_c\left(\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right)\right\|^2, \end{aligned} \tag{39}$$

Here, in ① we used the descent inequality equation 9; its locality condition $\left\|x^{k+1} - x^k\right\| \leq \frac{1}{\sqrt{\rho}\mathcal{L}_1}$ is verified in Remark C.2. In ② we used the ClipSGD update $x^{k+1} = x^k - \eta \cdot \mathrm{clip}_c\left(\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right)$.

We now consider two cases depending on the deterministic gradient norm.

1) The case $\left\|\nabla f\left(x^k\right)\right\| \geq c$.

Define the bad event

$$\delta = \mathbb{1}_{\left\{\|\nabla f(x^k, \boldsymbol{\xi}^k) - \nabla f(x^k)\| > 3\sqrt{\frac{\rho-1}{B}}\|\nabla f(x^k)\|\right\}}.$$

We decompose the first term in equation 39 according to this event:

$$\mathbb{E}\left[-\eta \left\langle \nabla f\left(x^k\right), \mathrm{clip}_c\left(\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right)\right\rangle \mid x^k\right] \overset{⑧}{=} \mathbb{E}\left[-\eta\alpha_{\boldsymbol{\xi}^k} \left\langle \nabla f\left(x^k\right), \nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\rangle \mid x^k\right]$$

$$\leq \mathbb{P}\left(\delta = 0 \mid x^k\right) \underbrace{\mathbb{E}\left[-\eta\alpha_{\boldsymbol{\xi}^k}\left\langle \nabla f\left(x^k\right), \nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\rangle \mid x^k, \delta = 0\right]}_{=:T_1}$$
$$+ \mathbb{P}\left(\delta = 1 \mid x^k\right) \underbrace{\mathbb{E}\left[-\eta\alpha_{\boldsymbol{\xi}^k}\left\langle \nabla f\left(x^k\right), \nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\rangle \mid x^k, \delta = 1\right]}_{=:T_2}, \tag{40}$$

where in $\circledast$ we used $\text{clip}_c\left(\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right) = \alpha_{\boldsymbol{\xi}^k} \cdot \nabla f\left(x^k, \boldsymbol{\xi}^k\right)$. We first estimate the probability of the bad event. By Markov's inequality,

$$\begin{aligned}\mathbb{P}\left(\delta = 1 \mid x^k\right) &= \mathbb{P}\left(\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right) - \nabla f\left(x^k\right)\right\|^2 > 9\frac{\rho-1}{B}\left\|\nabla f\left(x^k\right)\right\|^2 \mid x^k\right) \\ &\leq \frac{\mathbb{E}\left[\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right) - \nabla f\left(x^k\right)\right\|^2 \mid x^k\right]}{9(\rho-1)\left\|\nabla f\left(x^k\right)\right\|^2} \cdot B \\ &\overset{equation\ 24}{\leq} \frac{1}{9}.\end{aligned} \tag{41}$$

Consequently, $\mathbb{P}\left(\delta = 0 \mid x^k\right) = 1 - \mathbb{P}\left(\delta = 1 \mid x^k\right) \geq \frac{8}{9}$. We next bound the contribution of the good event:

$$\begin{aligned}&\mathbb{E}\left[-\eta\alpha_{\boldsymbol{\xi}^k}\left\langle \nabla f\left(x^k\right), \nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\rangle \mid x^k, \delta = 0\right] \\ &= \mathbb{E}\left[-\eta\alpha_{\boldsymbol{\xi}^k}\left\langle \nabla f\left(x^k\right), \nabla f\left(x^k, \boldsymbol{\xi}^k\right) \pm \nabla f\left(x^k\right)\right\rangle \mid x^k, \delta = 0\right] \\ &= \mathbb{E}\left[-\eta\alpha_{\boldsymbol{\xi}^k}\left\|\nabla f\left(x^k\right)\right\|^2 - \eta\alpha_{\boldsymbol{\xi}^k}\left\langle \nabla f\left(x^k\right), \nabla f\left(x^k, \boldsymbol{\xi}^k\right) - \nabla f\left(x^k\right)\right\rangle \mid x^k, \delta = 0\right] \\ &\overset{equation\ 7}{\leq} \mathbb{E}\left[-\eta\alpha_{\boldsymbol{\xi}^k}\left\|\nabla f\left(x^k\right)\right\|^2 + \eta\alpha_{\boldsymbol{\xi}^k}\left\|\nabla f\left(x^k\right)\right\|\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right) - \nabla f\left(x^k\right)\right\| \mid x^k, \delta = 0\right] \\ &\overset{①}{\leq} \mathbb{E}\left[-\eta\alpha_{\boldsymbol{\xi}^k}\left\|\nabla f\left(x^k\right)\right\|^2 + 3\eta\alpha_{\boldsymbol{\xi}^k}\sqrt{\frac{\rho-1}{B}}\left\|\nabla f\left(x^k\right)\right\|^2 \mid x^k, \delta = 0\right] \\ &\overset{②}{\leq} \mathbb{E}\left[-\eta\alpha_{\boldsymbol{\xi}^k}\left\|\nabla f\left(x^k\right)\right\|^2 + \frac{\eta\alpha_{\boldsymbol{\xi}^k}}{2}\left\|\nabla f\left(x^k\right)\right\|^2 \mid x^k, \delta = 0\right] \\ &= \mathbb{E}\left[-\frac{\eta\alpha_{\boldsymbol{\xi}^k}}{2}\left\|\nabla f\left(x^k\right)\right\|^2 \mid x^k, \delta = 0\right] \\ &\overset{③}{\leq} -\frac{\eta c}{4}\left\|\nabla f\left(x^k\right)\right\|,\end{aligned} \tag{42}$$

Here, in ① we used the definition of the event $\{\delta = 0\}$, in ② we used $B \geq 36(\rho-1)$, and in ③ we used $\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\| \leq 2\left\|\nabla f\left(x^k\right)\right\|$, which implies $\alpha_{\boldsymbol{\xi}^k} \geq c/(2\left\|\nabla f\left(x^k\right)\right\|)$. On the complementary event, we only use the boundedness induced by clipping:

$$\begin{aligned}\mathbb{E}\left[-\eta\alpha_{\boldsymbol{\xi}^k}\left\langle \nabla f\left(x^k\right), \nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\rangle \mid x^k, \delta = 1\right] &\overset{equation\ 7}{\leq} \eta\left\|\nabla f\left(x^k\right)\right\|\mathbb{E}\left[\alpha_{\boldsymbol{\xi}^k}\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\| \mid x^k, \delta = 1\right] \\ &\overset{①}{\leq} \eta c\left\|\nabla f\left(x^k\right)\right\|,\end{aligned} \tag{43}$$

where in ① we used $\left\|\text{clip}_c\left(\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right)\right\| \leq c$.

Combining the good and bad events yields

$$\begin{aligned}\mathbb{E}\left[-\eta\left\langle \nabla f\left(x^k\right), \text{clip}_c\left(\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right)\right\rangle \mid x^k\right] &\overset{equation\ 40}{\leq} \mathbb{P}\left(\delta = 0 \mid x^k\right)\underbrace{\mathbb{E}\left[-\eta\alpha_{\boldsymbol{\xi}^k}\left\langle \nabla f\left(x^k\right), \nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\rangle \mid x^k, \delta = 0\right]}_{=:T_1} \\ &+ \mathbb{P}\left(\delta = 1 \mid x^k\right)\underbrace{\mathbb{E}\left[-\eta\alpha_{\boldsymbol{\xi}^k}\left\langle \nabla f\left(x^k\right), \nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\rangle \mid x^k, \delta = 1\right]}_{=:T_2} \\ &\overset{equation\ 42, equation\ 43}{\leq} -\frac{\eta c}{4}\left\|\nabla f\left(x^k\right)\right\| \cdot \mathbb{P}\left(\delta = 0 \mid x^k\right) + \eta c\left\|\nabla f\left(x^k\right)\right\| \cdot \mathbb{P}\left(\delta = 1 \mid x^k\right)\end{aligned}$$

$$
\begin{aligned}
&= -\frac{\eta c}{4}\left\|\nabla f\left(x^k\right)\right\| \cdot \left(1 - \mathbb{P}\left(\delta = 1 \mid x^k\right)\right) + \eta c \left\|\nabla f\left(x^k\right)\right\| \cdot \mathbb{P}\left(\delta = 1 \mid x^k\right) \\
&\overset{equation\ 41}{\leq} -\eta c\left(\frac{8}{9}\cdot\frac{1}{4} - \frac{1}{9}\right)\left\|\nabla f\left(x^k\right)\right\| \\
&= -\frac{\eta c}{9}\left\|\nabla f\left(x^k\right)\right\|. \qquad (44)
\end{aligned}
$$

Substituting this estimate into equation 39, we obtain:

$$
\begin{aligned}
\mathbb{E}\left[F_{k+1} \mid x^k\right] - F_k &\overset{equation\ 39}{\leq} -\mathbb{E}\left[\eta\left\langle \nabla f\left(x^k\right), \text{clip}_c\left(\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right)\right\rangle \mid x^k\right] \\
&\quad + \frac{\eta^2\left(\mathcal{L}_0 + \sqrt{\rho}\mathcal{L}_1\left\|\nabla f\left(x^k\right)\right\|\right)}{2}\mathbb{E}\left[\left\|\text{clip}_c\left(\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right)\right\|^2 \mid x^k\right] \\
&\overset{equation\ 44}{\leq} -\frac{\eta c}{9}\left\|\nabla f\left(x^k\right)\right\| + \frac{\eta^2\left(\mathcal{L}_0 + \sqrt{\rho}\mathcal{L}_1\left\|\nabla f\left(x^k\right)\right\|\right)}{2}\mathbb{E}\left[\left\|\text{clip}_c\left(\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right)\right\|^2 \mid x^k\right] \\
&\overset{①}{\leq} -\frac{\eta c}{9}\left\|\nabla f\left(x^k\right)\right\| + \frac{\eta^2\left(\mathcal{L}_0 + \sqrt{\rho}\mathcal{L}_1\left\|\nabla f\left(x^k\right)\right\|\right)}{2}\cdot c^2 \\
&\overset{②}{\leq} -\frac{\eta c}{9}\left\|\nabla f\left(x^k\right)\right\| + \frac{\eta^2\left(\mathcal{L}_0 + \sqrt{\rho}\mathcal{L}_1 c\right)}{2}\cdot c \cdot \left\|\nabla f\left(x^k\right)\right\| \\
&\overset{③}{\leq} -\frac{\eta c}{9}\left\|\nabla f\left(x^k\right)\right\| + \frac{\eta c}{18}\left\|\nabla f\left(x^k\right)\right\| \\
&= -\frac{\eta c}{18}\left\|\nabla f\left(x^k\right)\right\| \\
&\overset{equation\ 38}{\leq} -\frac{\eta c}{18R_0}F_k, \qquad (45)
\end{aligned}
$$

where in ① we used $\left\|\text{clip}_c\left(\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right)\right\| \leq c$, in ② we used $\left\|\nabla f\left(x^k\right)\right\| \geq c$, in ③ we chose $\eta \leq \left[9\left(\mathcal{L}_0 + \sqrt{\rho}\mathcal{L}_1 c\right)\right]^{-1}$.

Hence, in the first case,

$$
\mathbb{E}\left[F_{k+1} \mid x^k\right] \leq \left(1 - \frac{\eta c}{18R_0}\right)F_k. \qquad (46)
$$

2) The case $\left\|\nabla f\left(x^k\right)\right\| < c$.

In this case, $\nabla f\left(x^k\right) = \text{clip}_c\left(\nabla f\left(x^k\right)\right)$. From equation 39, we have:

$$
\begin{aligned}
\mathbb{E}\left[f\left(x^{k+1}\right) \mid x^k\right] - f\left(x^k\right) &\overset{equation\ 39}{\leq} -\eta\mathbb{E}\left[\left\langle \nabla f\left(x^k\right), \text{clip}_c\left(\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right)\right\rangle \mid x^k\right] \\
&\quad + \frac{\eta^2\left(\mathcal{L}_0 + \sqrt{\rho}\mathcal{L}_1\left\|\nabla f\left(x^k\right)\right\|\right)}{2}\mathbb{E}\left[\left\|\text{clip}_c\left(\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right)\right\|^2 \mid x^k\right] \\
&\overset{①}{\leq} -\eta\mathbb{E}\left[\left\langle \nabla f\left(x^k\right), \text{clip}_c\left(\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right)\right\rangle \mid x^k\right] \\
&\quad + \frac{\eta^2\left(\mathcal{L}_0 + \sqrt{\rho}\mathcal{L}_1 c\right)}{2}\mathbb{E}\left[\left\|\text{clip}_c\left(\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right)\right\|^2 \mid x^k\right] \\
&\overset{②}{=} -\frac{\eta}{2}\left\|\nabla f\left(x^k\right)\right\|^2 - \frac{\eta}{2}\mathbb{E}\left[\left\|\text{clip}_c\left(\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right)\right\|^2 \mid x^k\right] \\
&\quad + \frac{\eta}{2}\mathbb{E}\left[\left\|\text{clip}_c\left(\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right) - \nabla f\left(x^k\right)\right\|^2 \mid x^k\right] \\
&\quad + \frac{\eta^2\left(\mathcal{L}_0 + \sqrt{\rho}\mathcal{L}_1 c\right)}{2}\mathbb{E}\left[\left\|\text{clip}_c\left(\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right)\right\|^2 \mid x^k\right] \\
&\overset{③}{=} -\frac{\eta}{2}\left\|\nabla f\left(x^k\right)\right\|^2 + \frac{\eta}{2}\mathbb{E}\left[\left\|\text{clip}_c\left(\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right) - \text{clip}_c\left(\nabla f\left(x^k\right)\right)\right\|^2 \mid x^k\right] \\
&\quad - \frac{\eta}{2}\left(1 - \eta\left(\mathcal{L}_0 + \sqrt{\rho}\mathcal{L}_1 c\right)\right)\mathbb{E}\left[\left\|\text{clip}_c\left(\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right)\right\|^2 \mid x^k\right] \\
&\overset{④}{\leq} -\frac{\eta}{2}\left\|\nabla f\left(x^k\right)\right\|^2 + \frac{\eta}{2}\mathbb{E}\left[\left\|\text{clip}_c\left(\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right) - \text{clip}_c\left(\nabla f\left(x^k\right)\right)\right\|^2 \mid x^k\right],
\end{aligned}
$$

where in ① we used $\left\|\nabla f\left(x^k\right)\right\| < c$, in ② we used inequality equation 6 with $a = \nabla f\left(x^k\right)$ and $b = \text{clip}_c\left(\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right)$, in ③ we used $\nabla f\left(x^k\right) = \text{clip}_c\left(\nabla f\left(x^k\right)\right)$, in ④ we chose $\eta \leq \left[9\left(\mathcal{L}_0 + \sqrt{\rho}\mathcal{L}_1 c\right)\right]^{-1}$.

Since clipping is the projection onto the Euclidean ball of radius c, it is a Lipschitz operator with Lipschitz constant 1. Therefore,

$$\begin{aligned}
\mathbb{E}\left[f\left(x^{k+1}\right) \mid x^k\right] - f\left(x^k\right) &\leq -\frac{\eta}{2}\left\|\nabla f\left(x^k\right)\right\|^2 + \frac{\eta}{2}\mathbb{E}\left[\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right) - \nabla f\left(x^k\right)\right\|^2 \mid x^k\right] \\
&\overset{equation\ 24}{\leq} -\frac{\eta}{2}\left\|\nabla f\left(x^k\right)\right\|^2 + \frac{\eta(\rho-1)}{2B}\left\|\nabla f\left(x^k\right)\right\|^2 \\
&\overset{①}{\leq} -\frac{\eta}{2}\left\|\nabla f\left(x^k\right)\right\|^2 + \frac{\eta}{72}\left\|\nabla f\left(x^k\right)\right\|^2 \\
&\leq -\frac{\eta}{2}\left\|\nabla f\left(x^k\right)\right\|^2 + \frac{\eta}{9}\left\|\nabla f\left(x^k\right)\right\|^2 \\
&= -\frac{7\eta}{18}\left\|\nabla f\left(x^k\right)\right\|^2 \\
&\leq -\frac{\eta}{9}\left\|\nabla f\left(x^k\right)\right\|^2 \\
&\overset{equation\ 38}{\leq} -\frac{\eta}{9R_0^2}\left(f\left(x^k\right) - f^*\right)^2, \qquad (47)
\end{aligned}$$

where in ① we chose $B \geq 36(\rho - 1)$.

In summary. Let $\mathcal{F}_k$ be the sigma-algebra generated by the history up to x^k, and define

$$G_k := \left\|\nabla f\left(x^k\right)\right\|, \qquad I_k := \mathbb{1}_{\{G_k \geq c\}}.$$

The random variable I_k is $\mathcal{F}_k$-measurable. Set

$$a := \frac{\eta c}{18R_0}, \qquad b := \frac{\eta}{9R_0^2}.$$

The two conditional estimates derived above can be combined as

$$\mathbb{E}\left[F_{k+1} \mid \mathcal{F}_k\right] \leq F_k - aF_kI_k - bF_k^2(1 - I_k). \qquad (48)$$

Indeed, on the event $I_k = 1$ this is exactly the large-gradient estimate, whereas on the event $I_k = 0$ it is exactly the small-gradient estimate.

Taking expectations in equation 48, we obtain

$$A_{k+1} \leq A_k - a\mathbb{E}\left[F_kI_k\right] - b\mathbb{E}\left[F_k^2(1 - I_k)\right], \qquad A_k := \mathbb{E}\left[F_k\right]. \qquad (49)$$

We now use the following elementary inequality. For any nonnegative random variable X and any indicator random variable I,

$$a\mathbb{E}\left[XI\right] + b\mathbb{E}\left[X^2(1 - I)\right] \geq \min\left\{\frac{a}{2}\mathbb{E}\left[X\right], b\left(\mathbb{E}\left[X\right]\right)^2\right\}. \qquad (50)$$

To prove it, let $A = \mathbb{E}\left[X\right]$ and $r = \mathbb{E}\left[X(1 - I)\right]$. Since

$$\mathbb{E}\left[X^2(1 - I)\right] = \mathbb{E}\left[\left(X(1 - I)\right)^2\right] \geq \left(\mathbb{E}\left[X(1 - I)\right]\right)^2 = r^2,$$

we have

$$a\mathbb{E}\left[XI\right] + b\mathbb{E}\left[X^2(1 - I)\right] \geq a(A - r) + br^2.$$

Minimizing the right-hand side over $r \in [0, A]$, we get: if $A \leq a/(2b)$, the minimum is attained at $r = A$ and equals bA^2; if $A \geq a/(2b)$, the minimum is attained at $r = a/(2b)$ and equals

$$aA - \frac{a^2}{4b} \geq \frac{a}{2}A.$$

This proves equation 50.

Applying equation 50 to $X = F_k$ and $I = I_k$, inequality equation 49 gives

$$A_{k+1} \leq A_k - \min\left\{\frac{a}{2}A_k, bA_k^2\right\}. \tag{51}$$

In particular, $(A_k)_{k\geq 0}$ is non-increasing.

Define the capped contraction parameter

$$\alpha_0 := \frac{a}{2} = \frac{\eta c}{36R_0}, \qquad \bar{\alpha} := \min\{\alpha_0, 1\}, \qquad \beta := b = \frac{\eta}{9R_0^2}.$$

Since $\bar{\alpha} \leq \alpha_0$, inequality equation 51 implies

$$A_{k+1} \leq A_k - \min\{\bar{\alpha}A_k, \beta A_k^2\}. \tag{52}$$

Let

$$\tau := \min\left\{k \in \{0, \ldots, N-1\} : A_k < \frac{\bar{\alpha}}{\beta}\right\},$$

with the convention $\tau = N$ if the set is empty. Since (A_k) is non-increasing, the indices $k < \tau$ correspond to the linear part of equation 52, and the indices $k \geq \tau$ correspond to the quadratic part.

If $\tau > N/2$, then for at least $N/2$ iterations,

$$A_{k+1} \leq (1 - \bar{\alpha})A_k.$$

Using monotonicity for the remaining iterations, we get

$$A_N \leq (1 - \bar{\alpha})^{N/2} A_0 = (1 - \bar{\alpha})^{N/2} F_0. \tag{53}$$

If $\tau \leq N/2$, then for $k = \tau, \ldots, N-1$,

$$A_{k+1} \leq A_k - \beta A_k^2.$$

If $A_N = 0$, the claim is trivial. Otherwise,

$$\frac{1}{A_{k+1}} - \frac{1}{A_k} = \frac{A_k - A_{k+1}}{A_k A_{k+1}} \geq \beta \frac{A_k}{A_{k+1}} \geq \beta.$$

Summing from $k = \tau$ to $N-1$ and using $N - \tau \geq N/2$, we obtain

$$A_N \leq \frac{2}{\beta N} = \frac{18R_0^2}{\eta N}. \tag{54}$$

Combining equation 53 and equation 54, we get

$$\mathbb{E}\left[f\left(x^N\right)\right] - f^* = A_N \leq \max\left\{(1 - \bar{\alpha})^{N/2} F_0, \frac{18R_0^2}{\eta N}\right\}.$$

□

Remark C.2. *In the first step of equation 39, we use the descent lemma for (L_0, L_1)-smooth functions equation 9 (see Appendix A.1 in Zhang et al., 2020a); the condition $\left\|x^{k+1} - x^k\right\| \leq \frac{1}{\sqrt{\rho}\mathcal{L}_1}$ required for its application is guaranteed by the chosen step size $\eta \leq \left[9\left(\mathcal{L}_0 + \sqrt{\rho}\mathcal{L}_1 c\right)\right]^{-1}$:*

$$\left\|x^{k+1} - x^k\right\| \overset{①}{=} \eta \left\|clip_c\left(\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right)\right\| \overset{②}{\leq} \eta c \overset{③}{\leq} \frac{c}{9\left(\mathcal{L}_0 + \sqrt{\rho}\mathcal{L}_1 c\right)} \overset{④}{\leq} \frac{c}{9\sqrt{\rho}\mathcal{L}_1 c} \overset{④}{\leq} \frac{1}{\sqrt{\rho}\mathcal{L}_1},$$

where in ① we used $x^{k+1} = x^k - \eta \cdot clip_c\left(\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right)$, in ② we used $\left\|clip_c\left(\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right)\right\| = \left\|\min\left\{1, \frac{c}{\|\nabla f(x^k, \boldsymbol{\xi}^k)\|}\right\} \nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\| \leq c$, in ③ we used $\eta \leq \left[9\left(\mathcal{L}_0 + \sqrt{\rho}\mathcal{L}_1 c\right)\right]^{-1}$, where in ④ we used the fact that, for $0 < b \leq a$, one has $\frac{1}{a} \leq \frac{1}{b}$.

C.2 Missing proofs for normalized stochastic gradient descent method

This section gives the corresponding proofs for NSGD. The hyperparameter $\lambda > 0$ plays a role analogous to the clipping radius, but it also appears directly in the normalization denominator. As a result, the admissible step size and the final error floor depend on λ.

C.2.1 Non-convex setup under (L_0, L_1)-smooth and generalized strong growth conditions

Proof. For simplicity of presentation, define

$$G_\lambda\left(x^k, \boldsymbol{\xi}^k\right) \coloneqq \frac{\nabla f\left(x^k, \boldsymbol{\xi}^k\right)}{\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\| + \lambda}, \qquad F_k \coloneqq f\left(x^k\right) - f^*.$$

In the one-step estimates below, expectations are taken with respect to the fresh mini-batch $\boldsymbol{\xi}^k$ and are written explicitly as conditional expectations given the current iterate x^k. After taking full expectation, we return to the notation $\mathbb{E}\left[\cdot\right]$.

We start from the descent inequality for (L_0, L_1)-smooth functions:

$$\begin{aligned}
f\left(x^{k+1}\right) - f\left(x^k\right) &\overset{①}{\leq} \left\langle \nabla f\left(x^k\right), x^{k+1} - x^k \right\rangle + \frac{L_0 + L_1\left\|\nabla f\left(x^k\right)\right\|}{2}\left\|x^{k+1} - x^k\right\|^2 \\
&\overset{②}{=} -\eta\left\langle \nabla f\left(x^k\right), G_\lambda\left(x^k, \boldsymbol{\xi}^k\right)\right\rangle + \frac{\eta^2\left(L_0 + L_1\left\|\nabla f\left(x^k\right)\right\|\right)}{2}\left\|G_\lambda\left(x^k, \boldsymbol{\xi}^k\right)\right\|^2 \\
&\overset{③}{\leq} -\eta\frac{\left\langle \nabla f\left(x^k\right), \nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\rangle}{\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\| + \lambda} + \frac{\eta^2\left(L_0 + L_1\left\|\nabla f\left(x^k\right)\right\|\right)}{2} \\
&= -\eta\frac{\left\langle \nabla f\left(x^k\right) \pm \nabla f\left(x^k, \boldsymbol{\xi}^k\right), \nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\rangle}{\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\| + \lambda} + \frac{\eta^2\left(L_0 + L_1\left\|\nabla f\left(x^k\right)\right\|\right)}{2} \\
&= -\eta\frac{\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\|^2}{\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\| + \lambda} - \eta\frac{\left\langle \nabla f\left(x^k\right) - \nabla f\left(x^k, \boldsymbol{\xi}^k\right), \nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\rangle}{\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\| + \lambda} \\
&\qquad + \frac{\eta^2\left(L_0 + L_1\left\|\nabla f\left(x^k\right)\right\|\right)}{2} \\
&\overset{④}{\leq} -\eta\frac{\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\|^2}{\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\| + \lambda} + \eta\frac{\left\|\nabla f\left(x^k\right) - \nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\|\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\|}{\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\| + \lambda} \\
&\qquad + \frac{\eta^2\left(L_0 + L_1\left\|\nabla f\left(x^k\right)\right\|\right)}{2} \\
&\overset{③}{\leq} -\eta\frac{\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\|^2}{\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\| + \lambda} + \eta\left\|\nabla f\left(x^k\right) - \nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\| \\
&\qquad + \frac{\eta^2\left(L_0 + L_1\left\|\nabla f\left(x^k\right)\right\|\right)}{2} \\
&= -\eta\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\| + \eta\frac{\lambda\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\|}{\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\| + \lambda} + \eta\left\|\nabla f\left(x^k\right) - \nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\| \\
&\qquad + \frac{\eta^2\left(L_0 + L_1\left\|\nabla f\left(x^k\right)\right\|\right)}{2} \\
&\overset{③}{\leq} -\eta\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\| + \eta\lambda + \eta\left\|\nabla f\left(x^k\right) - \nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\| \\
&\qquad + \frac{\eta^2\left(L_0 + L_1\left\|\nabla f\left(x^k\right)\right\|\right)}{2} \\
&\overset{⑤}{\leq} -\eta\left\|\nabla f\left(x^k\right)\right\| + 2\eta\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right) - \nabla f\left(x^k\right)\right\| \\
&\qquad + \frac{\eta^2\left(L_0 + L_1\left\|\nabla f\left(x^k\right)\right\|\right)}{2} + \eta\lambda \\
&\overset{⑥}{\leq} -\eta\left\|\nabla f\left(x^k\right)\right\| + 2\eta\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right) - \nabla f\left(x^k\right)\right\|
\end{aligned}$$

$$+\frac{\eta}{2}\left(\left\|\nabla f\left(x^k\right)\right\|+\lambda\right)+\eta\lambda$$
$$=-\frac{\eta}{2}\left\|\nabla f\left(x^k\right)\right\|+2\eta\left\|\nabla f\left(x^k,\boldsymbol{\xi}^k\right)-\nabla f\left(x^k\right)\right\|+\frac{3}{2}\eta\lambda, \tag{55}$$

Here, in ① we used the descent inequality equation 9; its locality condition $\left\|x^{k+1}-x^k\right\| \leq \frac{1}{L_1}$ is verified in Remark C.3. In ② we used the NSGD update $x^{k+1}=x^k-\eta\cdot G_\lambda\left(x^k,\boldsymbol{\xi}^k\right)$. In ③ we used $\left\|G_\lambda\left(x^k,\boldsymbol{\xi}^k\right)\right\| \leq 1$, in ④ the Cauchy–Schwarz inequality equation 7, and in ⑤ the triangle inequality. Finally, in ⑥ we used the step-size condition $\eta \leq \frac{\lambda}{L_0+L_1\lambda}$ together with $\frac{\lambda}{L_0+L_1\lambda} \leq \frac{\left\|\nabla f\left(x^k\right)\right\|+\lambda}{L_0+L_1\left\|\nabla f\left(x^k\right)\right\|}$, which follows from

$$\frac{\left\|\nabla f\left(x^k\right)\right\|+\lambda}{L_0+L_1\left\|\nabla f\left(x^k\right)\right\|}-\frac{\lambda}{L_0+L_1\lambda}=\frac{\left(\left\|\nabla f\left(x^k\right)\right\|+\lambda\right)(L_0+L_1\lambda)-\lambda(L_0+L_1\left\|\nabla f\left(x^k\right)\right\|)}{(L_0+L_1\left\|\nabla f\left(x^k\right)\right\|)(L_0+L_1\lambda)}$$
$$=\frac{L_0\left\|\nabla f\left(x^k\right)\right\|+\lambda^2 L_1}{(L_0+L_1\left\|\nabla f\left(x^k\right)\right\|)(L_0+L_1\lambda)} \geq 0,$$

where $L_0, L_1, \lambda, \left\|\nabla f\left(x^k\right)\right\| \geq 0$.

Taking conditional expectation in equation 55 given x^k, we obtain:

$$\begin{aligned}
\mathbb{E}\left[F_{k+1} \mid x^k\right]-F_k &\overset{\text{equation 55}}{\leq} -\frac{\eta}{2}\left\|\nabla f\left(x^k\right)\right\|+2\eta\mathbb{E}\left[\left\|\nabla f\left(x^k,\boldsymbol{\xi}^k\right)-\nabla f\left(x^k\right)\right\| \mid x^k\right]+\frac{3}{2}\eta\lambda \\
&= -\frac{\eta}{2}\left\|\nabla f\left(x^k\right)\right\|+2\eta\mathbb{E}\left[\left(\left\|\nabla f\left(x^k,\boldsymbol{\xi}^k\right)-\nabla f\left(x^k\right)\right\|^2\right)^{1/2} \mid x^k\right]+\frac{3}{2}\eta\lambda \\
&\overset{①}{\leq} -\frac{\eta}{2}\left\|\nabla f\left(x^k\right)\right\|+2\eta\left(\mathbb{E}\left[\left\|\nabla f\left(x^k,\boldsymbol{\xi}^k\right)-\nabla f\left(x^k\right)\right\|^2 \mid x^k\right]\right)^{1/2}+\frac{3}{2}\eta\lambda \\
&\overset{\text{equation 24}}{\leq} -\frac{\eta}{2}\left\|\nabla f\left(x^k\right)\right\|+2\eta\left(\frac{1}{B}\left[(\rho-1)\left\|\nabla f\left(x^k\right)\right\|^2+\sigma^2\right]\right)^{1/2}+\frac{3}{2}\eta\lambda \\
&\overset{②}{\leq} -\frac{\eta}{2}\left\|\nabla f\left(x^k\right)\right\|+2\eta\sqrt{\frac{(\rho-1)}{B}}\left\|\nabla f\left(x^k\right)\right\|+\frac{2\eta\sigma}{\sqrt{B}}+\frac{3}{2}\eta\lambda \\
&\overset{③}{\leq} -\frac{\eta}{2}\left\|\nabla f\left(x^k\right)\right\|+\frac{\eta}{4}\left\|\nabla f\left(x^k\right)\right\|+\frac{2\eta\sigma}{\sqrt{B}}+\frac{3}{2}\eta\lambda \\
&= -\frac{\eta}{4}\left\|\nabla f\left(x^k\right)\right\|+\frac{2\eta\sigma}{\sqrt{B}}+\frac{3}{2}\eta\lambda,
\end{aligned} \tag{56}$$

where in ① we applied Jensen's inequality, in ② we used $\sqrt{a^2+b^2} \leq a+b$ when $a, b \geq 0$, in ③ we chose $B \geq 64(\rho-1)$. Thus, we have obtained the inequality:

$$\left\|\nabla f\left(x^k\right)\right\| \overset{\text{equation 56}}{\leq} \frac{4\left(F_k-\mathbb{E}\left[F_{k+1} \mid x^k\right]\right)}{\eta}+\frac{8\sigma}{\sqrt{B}}+6\lambda.$$

Taking full expectation and summing this inequality over all indices $0 \leq k \leq N-1$ gives:

$$\min_{k\in[0,N-1]}\mathbb{E}\left[\left\|\nabla f\left(x^k\right)\right\|\right] \leq \frac{1}{N}\sum_{k=0}^{N-1}\mathbb{E}\left[\left\|\nabla f\left(x^k\right)\right\|\right] \leq \underbrace{\frac{4F_0}{\eta N}+\frac{8\sigma}{\sqrt{B}}}_{①}+\underbrace{6\lambda}_{②}.$$

To guarantee convergence to the desired accuracy $\min_{k\in[0,N-1]}\mathbb{E}\left[|\nabla f(x^k)|^2\right] \leq \varepsilon$, it is sufficient to bound the first two terms and the third term by $\varepsilon/2$. This yields the constraint:

$$②: \quad \lambda \leq \frac{\varepsilon}{12}.$$

Taking $\lambda=\frac{\varepsilon}{12}$ and the maximum admissible step size $\eta=\frac{\lambda}{L_0+L_1\lambda}=\left[\frac{12L_0}{\varepsilon}+L_1\right]^{-1}$, we obtain the following sufficient condition under batch size $B \geq 64(\rho-1)$:

$$①: \quad \frac{4F_0}{\eta N}+\frac{8\sigma}{\sqrt{B}}=\frac{48L_0}{\varepsilon N}F_0+\frac{4L_1}{N}F_0+\frac{8\sigma}{\sqrt{B}} \leq \frac{\varepsilon}{2}.$$

In the case $\sigma = 0$, the above conditions are satisfied by

$$\lambda = \frac{\varepsilon}{12}; \quad N \geq \frac{48L_0}{\varepsilon^2}F_0 + \frac{4L_1}{\varepsilon}F_0; \quad B \geq 64(\rho - 1),$$

and the corresponding number of oracle calls $\#T$:

$$T = N \cdot B = \mathcal{O}\left(\frac{(\rho-1)L_0}{\varepsilon^2}F_0 + \frac{(\rho-1)L_1}{\varepsilon}F_0\right).$$

□

Remark C.3. *In the first step of equation 55, we use the descent lemma for (L_0, L_1)-smooth functions equation 9 (see Appendix A.1 in Zhang et al., 2020a); the condition $\left\|x^{k+1} - x^k\right\| \leq 1/L_1$ required for its application is guaranteed by the chosen step size $\eta \leq \frac{\lambda}{L_0 + L_1\lambda}$:*

$$\left\|x^{k+1} - x^k\right\| \overset{①}{=} \eta \left\|G_\lambda\left(x^k, \boldsymbol{\xi}^k\right)\right\| \overset{②}{\leq} \eta \overset{③}{\leq} \frac{\lambda}{L_0 + L_1\lambda} \overset{④}{\leq} \frac{\lambda}{L_1\lambda} = \frac{1}{L_1},$$

where in ① we used $x^{k+1} = x^k - \eta \cdot G_\lambda\left(x^k, \boldsymbol{\xi}^k\right)$, in ② we used $\left\|G_\lambda\left(x^k, \boldsymbol{\xi}^k\right)\right\| = \left\|\frac{\nabla f\left(x^k, \boldsymbol{\xi}^k\right)}{\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\| + \lambda}\right\| \leq 1$, in ③ we used $\eta \leq \frac{\lambda}{L_0 + L_1\lambda}$, where in ④ we used the fact that, for $0 < b \leq a$, one has $\frac{1}{a} \leq \frac{1}{b}$.

C.2.2 Convex setup under $(\mathcal{L}_0, \mathcal{L}_1)$-smooth and strong growth conditions

We now prove the convex NSGD result. The proof follows the non-convex argument up to the one-step descent estimate, and then uses convexity to obtain a recursion for the optimality gap.

Proof. For simplicity of presentation, define

$$G_\lambda\left(x^k, \boldsymbol{\xi}^k\right) := \frac{\nabla f\left(x^k, \boldsymbol{\xi}^k\right)}{\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\| + \lambda}, \qquad F_k := f\left(x^k\right) - f^*.$$

In the one-step estimates below, expectations are taken with respect to the fresh mini-batch $\boldsymbol{\xi}^k$ and are written explicitly as conditional expectations given the current iterate x^k. After taking full expectation, we return to the notation $\mathbb{E}\left[\cdot\right]$. By convexity, the Cauchy–Schwarz inequality, and the monotonicity of the distance to the optimum from Lemma B.4, we have

$$F_k \leq \left\langle \nabla f\left(x^k\right), x^k - x^* \right\rangle \leq \left\|\nabla f\left(x^k\right)\right\| \left\|x^k - x^*\right\| \leq R_0 \left\|\nabla f\left(x^k\right)\right\|. \tag{57}$$

We start from the descent inequality for (L_0, L_1)-smooth functions with constants $L_0 = \mathcal{L}_0$ and $L_1 = \sqrt{\rho}\mathcal{L}_1$ (see Section B.2):

$$\begin{aligned}
f\left(x^{k+1}\right) - f\left(x^k\right) &\overset{①}{\leq} \left\langle \nabla f\left(x^k\right), x^{k+1} - x^k \right\rangle + \frac{\mathcal{L}_0 + \sqrt{\rho}\mathcal{L}_1 \left\|\nabla f\left(x^k\right)\right\|}{2} \left\|x^{k+1} - x^k\right\|^2 \\
&\overset{②}{=} -\eta \left\langle \nabla f\left(x^k\right), G_\lambda\left(x^k, \boldsymbol{\xi}^k\right) \right\rangle + \frac{\eta^2\left(\mathcal{L}_0 + \sqrt{\rho}\mathcal{L}_1 \left\|\nabla f\left(x^k\right)\right\|\right)}{2} \left\|G_\lambda\left(x^k, \boldsymbol{\xi}^k\right)\right\|^2 \\
&\overset{③}{\leq} -\eta \frac{\left\langle \nabla f\left(x^k\right), \nabla f\left(x^k, \boldsymbol{\xi}^k\right) \right\rangle}{\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\| + \lambda} + \frac{\eta^2\left(\mathcal{L}_0 + \sqrt{\rho}\mathcal{L}_1 \left\|\nabla f\left(x^k\right)\right\|\right)}{2} \\
&= -\eta \frac{\left\langle \nabla f\left(x^k\right) \pm \nabla f\left(x^k, \boldsymbol{\xi}^k\right), \nabla f\left(x^k, \boldsymbol{\xi}^k\right) \right\rangle}{\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\| + \lambda} + \frac{\eta^2\left(\mathcal{L}_0 + \sqrt{\rho}\mathcal{L}_1 \left\|\nabla f\left(x^k\right)\right\|\right)}{2} \\
&= -\eta \frac{\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\|^2}{\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\| + \lambda} - \eta \frac{\left\langle \nabla f\left(x^k\right) - \nabla f\left(x^k, \boldsymbol{\xi}^k\right), \nabla f\left(x^k, \boldsymbol{\xi}^k\right) \right\rangle}{\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\| + \lambda} \\
&\qquad + \frac{\eta^2\left(\mathcal{L}_0 + \sqrt{\rho}\mathcal{L}_1 \left\|\nabla f\left(x^k\right)\right\|\right)}{2} \\
&\overset{④}{\leq} -\eta \frac{\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\|^2}{\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\| + \lambda} + \eta \frac{\left\|\nabla f\left(x^k\right) - \nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\| \left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\|}{\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\| + \lambda}
\end{aligned}$$

$$
\begin{aligned}
&\quad + \frac{\eta^2\left(\mathcal{L}_0 + \sqrt{\rho}\mathcal{L}_1\left\|\nabla f\left(x^k\right)\right\|\right)}{2} \\
&\overset{③}{\leq} -\eta\frac{\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\|^2}{\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\| + \lambda} + \eta\left\|\nabla f\left(x^k\right) - \nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\| \\
&\quad + \frac{\eta^2\left(\mathcal{L}_0 + \sqrt{\rho}\mathcal{L}_1\left\|\nabla f\left(x^k\right)\right\|\right)}{2} \\
&= -\eta\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\| + \eta\frac{\lambda\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\|}{\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\| + \lambda} + \eta\left\|\nabla f\left(x^k\right) - \nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\| \\
&\quad + \frac{\eta^2\left(\mathcal{L}_0 + \sqrt{\rho}\mathcal{L}_1\left\|\nabla f\left(x^k\right)\right\|\right)}{2} \\
&\overset{③}{\leq} -\eta\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\| + \eta\lambda + \eta\left\|\nabla f\left(x^k\right) - \nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\| \\
&\quad + \frac{\eta^2\left(\mathcal{L}_0 + \sqrt{\rho}\mathcal{L}_1\left\|\nabla f\left(x^k\right)\right\|\right)}{2} \\
&\overset{⑤}{\leq} -\eta\left\|\nabla f\left(x^k\right)\right\| + 2\eta\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right) - \nabla f\left(x^k\right)\right\| \\
&\quad + \frac{\eta^2\left(\mathcal{L}_0 + \sqrt{\rho}\mathcal{L}_1\left\|\nabla f\left(x^k\right)\right\|\right)}{2} + \eta\lambda \\
&\overset{⑥}{\leq} -\eta\left\|\nabla f\left(x^k\right)\right\| + 2\eta\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right) - \nabla f\left(x^k\right)\right\| \\
&\quad + \frac{\eta}{2}\left(\left\|\nabla f\left(x^k\right)\right\| + \lambda\right) + \eta\lambda \\
&= -\frac{\eta}{2}\left\|\nabla f\left(x^k\right)\right\| + 2\eta\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right) - \nabla f\left(x^k\right)\right\| + \frac{3}{2}\eta\lambda, \qquad (58)
\end{aligned}
$$

Here, in ① we used the descent inequality equation 9; its locality condition $\left\|x^{k+1} - x^k\right\| \leq \frac{1}{L_1}$ with $L_1 = \sqrt{\rho}\mathcal{L}_1$ is verified in Remark C.4. In ② we used the NSGD update $x^{k+1} = x^k - \eta \cdot G_\lambda\left(x^k, \boldsymbol{\xi}^k\right)$. In ③ we used $\left\|G_\lambda\left(x^k, \boldsymbol{\xi}^k\right)\right\| \leq 1$, in ④ the Cauchy–Schwarz inequality equation 7, and in ⑤ the triangle inequality. Finally, in ⑥ we used the step-size condition $\eta \leq \frac{\lambda}{\mathcal{L}_0 + \sqrt{\rho}\mathcal{L}_1\lambda}$ together with $\frac{\lambda}{\mathcal{L}_0 + \sqrt{\rho}\mathcal{L}_1\lambda} \leq \frac{\|\nabla f(x^k)\| + \lambda}{\mathcal{L}_0 + \sqrt{\rho}\mathcal{L}_1\|\nabla f(x^k)\|}$, which follows from

$$
\begin{aligned}
\frac{\left\|\nabla f\left(x^k\right)\right\| + \lambda}{\mathcal{L}_0 + \sqrt{\rho}\mathcal{L}_1\left\|\nabla f\left(x^k\right)\right\|} - \frac{\lambda}{\mathcal{L}_0 + \sqrt{\rho}\mathcal{L}_1\lambda} &= \frac{\left(\left\|\nabla f\left(x^k\right)\right\| + \lambda\right)\left(\mathcal{L}_0 + \sqrt{\rho}\mathcal{L}_1\lambda\right) - \lambda\left(\mathcal{L}_0 + \sqrt{\rho}\mathcal{L}_1\left\|\nabla f\left(x^k\right)\right\|\right)}{\left(\mathcal{L}_0 + \sqrt{\rho}\mathcal{L}_1\left\|\nabla f\left(x^k\right)\right\|\right)\left(\mathcal{L}_0 + \sqrt{\rho}\mathcal{L}_1\lambda\right)} \\
&= \frac{\mathcal{L}_0\left\|\nabla f\left(x^k\right)\right\| + \lambda^2\sqrt{\rho}\mathcal{L}_1}{\left(\mathcal{L}_0 + \sqrt{\rho}\mathcal{L}_1\left\|\nabla f\left(x^k\right)\right\|\right)\left(\mathcal{L}_0 + \sqrt{\rho}\mathcal{L}_1\lambda\right)} \geq 0,
\end{aligned}
$$

where $\mathcal{L}_0, \mathcal{L}_1, \lambda, \left\|\nabla f\left(x^k\right)\right\| \geq 0$ and $\rho \geq 1$.

Taking conditional expectation in equation 58 given x^k, we obtain:

$$
\begin{aligned}
\mathbb{E}\left[F_{k+1} \mid x^k\right] - F_k &\overset{\text{equation } 58}{\leq} -\frac{\eta}{2}\left\|\nabla f\left(x^k\right)\right\| + 2\eta\mathbb{E}\left[\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right) - \nabla f\left(x^k\right)\right\| \mid x^k\right] + \frac{3}{2}\eta\lambda \\
&= -\frac{\eta}{2}\left\|\nabla f\left(x^k\right)\right\| + 2\eta\mathbb{E}\left[\left(\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right) - \nabla f\left(x^k\right)\right\|^2\right)^{1/2} \mid x^k\right] + \frac{3}{2}\eta\lambda \\
&\overset{①}{\leq} -\frac{\eta}{2}\left\|\nabla f\left(x^k\right)\right\| + 2\eta\left(\mathbb{E}\left[\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right) - \nabla f\left(x^k\right)\right\|^2 \mid x^k\right]\right)^{1/2} + \frac{3}{2}\eta\lambda \\
&\overset{\text{equation } 24}{\leq} -\frac{\eta}{2}\left\|\nabla f\left(x^k\right)\right\| + 2\eta\left(\frac{1}{B}\left[(\rho - 1)\left\|\nabla f\left(x^k\right)\right\|^2 + \sigma^2\right]\right)^{1/2} + \frac{3}{2}\eta\lambda \\
&\overset{②}{\leq} -\frac{\eta}{2}\left\|\nabla f\left(x^k\right)\right\| + 2\eta\sqrt{\frac{(\rho - 1)}{B}}\left\|\nabla f\left(x^k\right)\right\| + \frac{2\eta\sigma}{\sqrt{B}} + \frac{3}{2}\eta\lambda \\
&\overset{③}{\leq} -\frac{\eta}{2}\left\|\nabla f\left(x^k\right)\right\| + \frac{\eta}{4}\left\|\nabla f\left(x^k\right)\right\| + \frac{2\eta\sigma}{\sqrt{B}} + \frac{3}{2}\eta\lambda
\end{aligned}
$$

$$= -\frac{\eta}{4}\left\|\nabla f\left(x^k\right)\right\| + \frac{2\eta\sigma}{\sqrt{B}} + \frac{3}{2}\eta\lambda, \tag{59}$$

where in ① we applied Jensen's inequality, in ② we used $\sqrt{a^2+b^2} \le a+b$ when $a, b \ge 0$, in ③ we chose $B \ge 64(\rho-1)$. Thus, we have obtained the inequality:

$$\begin{aligned}\mathbb{E}\left[F_{k+1} \mid x^k\right] &\overset{equation\ 59}{\le} F_k - \frac{\eta}{4}\left\|\nabla f\left(x^k\right)\right\| + \frac{2\eta\sigma}{\sqrt{B}} + \frac{3\eta}{2}\lambda \\ &\overset{equation\ 57}{\le} F_k - \frac{\eta}{4R_0}F_k + \frac{2\eta\sigma}{\sqrt{B}} + \frac{3\eta}{2}\lambda \\ &= \left(1 - \frac{\eta}{4R_0}\right)F_k + \frac{2\eta\sigma}{\sqrt{B}} + \frac{3\eta}{2}\lambda. \end{aligned} \tag{60}$$

Define

$$\gamma := \min\left\{\frac{\eta}{4R_0}, 1\right\}.$$

Since $\gamma \le \eta/(4R_0)$, taking full expectation in equation 60 gives the valid contraction

$$\mathbb{E}\left[F_{k+1}\right] \le (1-\gamma)\mathbb{E}\left[F_k\right] + \frac{2\eta\sigma}{\sqrt{B}} + \frac{3\eta\lambda}{2}, \qquad 0 < \gamma \le 1.$$

Thus, even without the additional cap $\eta \le 4R_0$, unrolling yields

$$\mathbb{E}\left[F_N\right] \le (1-\gamma)^N F_0 + \frac{2\eta\sigma/\sqrt{B} + 3\eta\lambda/2}{\gamma}. \tag{61}$$

Under the step-size condition of Theorem 4.7, $\eta \le 4R_0$ and hence $\gamma = \eta/(4R_0)$. Therefore,

$$\mathbb{E}\left[F_N\right] \le \left(1 - \frac{\eta}{4R_0}\right)^N F_0 + \frac{8\sigma R_0}{\sqrt{B}} + 6\lambda R_0.$$

In particular, for $\sigma = 0$,

$$\mathbb{E}\left[f\left(x^N\right)\right] - f^* \le \underbrace{\left(1 - \frac{\eta}{4R_0}\right)^N F_0}_{①} + \underbrace{6\lambda R_0}_{②}.$$

To guarantee convergence to the desired accuracy $\mathbb{E}\left[f\left(x^N\right)\right] - f^* \le \varepsilon$, it is sufficient to bound the two terms above by $\varepsilon/2$. This yields the constraint:

$$②: \quad \lambda \le \frac{\varepsilon}{12R_0}.$$

Taking $\lambda = \frac{\varepsilon}{12R_0}$ and the maximum admissible step size

$$\eta = \min\left\{\left[\frac{12\mathcal{L}_0 R_0}{\varepsilon} + \sqrt{\rho}\,\mathcal{L}_1\right]^{-1}, 4R_0\right\},$$

we obtain the following sufficient condition under batch size $B \ge 64(\rho-1)$:

$$①: \quad \left(1 - \frac{\eta}{4R_0}\right)^N F_0 \le \exp\left(-\frac{\eta N}{4R_0}\right) F_0 \le \frac{\varepsilon}{2}.$$

Thus, NSGD has the following values for the hyperparameter λ, the number of iterations $\#N$, and the batch size $\#B$ for convergence to the desired accuracy ε:

$$\lambda = \frac{\varepsilon}{12R_0}; \quad N \ge \max\left\{1, \frac{48\mathcal{L}_0 R_0^2}{\varepsilon} + 4\sqrt{\rho}\mathcal{L}_1 R_0\right\}\log\frac{2F_0}{\varepsilon}; \quad B \ge 64(\rho-1),$$

and the corresponding number of oracle calls $\#T$:

$$T = N \cdot B = \tilde{\mathcal{O}}\left((\rho-1)\max\left\{1, \frac{\mathcal{L}_0 R_0^2}{\varepsilon} + \sqrt{\rho}\mathcal{L}_1 R_0\right\}\right),$$

where we have used the notation $\tilde{\mathcal{O}}(\cdot)$ to hide logarithmic factors.

□

Remark C.4. *In the first step of equation 58, we use the descent lemma for $(\mathcal{L}_0, \mathcal{L}_1)$-smooth functions equation 9 (see Appendix A.1 in Zhang et al., 2020a); the condition $\left\|x^{k+1} - x^k\right\| \leq \frac{1}{\sqrt{\rho}\mathcal{L}_1}$ required for its application is guaranteed by the chosen step size $\eta \leq \frac{\lambda}{\mathcal{L}_0 + \sqrt{\rho}\mathcal{L}_1\lambda}$:*

$$\left\|x^{k+1} - x^k\right\| \overset{①}{=} \eta \left\|G_\lambda\left(x^k, \boldsymbol{\xi}^k\right)\right\| \overset{②}{\leq} \eta \overset{③}{\leq} \frac{\lambda}{\mathcal{L}_0 + \sqrt{\rho}\mathcal{L}_1\lambda} \overset{④}{\leq} \frac{\lambda}{\sqrt{\rho}\mathcal{L}_1\lambda} = \frac{1}{\sqrt{\rho}\mathcal{L}_1},$$

where in ① we used $x^{k+1} = x^k - \eta \cdot G_\lambda\left(x^k, \boldsymbol{\xi}^k\right)$, in ② we used $\left\|G_\lambda\left(x^k, \boldsymbol{\xi}^k\right)\right\| \leq 1$, in ③ we used $\eta \leq \frac{\lambda}{\mathcal{L}_0 + \sqrt{\rho}\mathcal{L}_1\lambda}$, and in ④ we used the fact that, for $0 < b \leq a$, one has $\frac{1}{a} \leq \frac{1}{b}$.

C.2.3 Modification to the proof to include heavy-tailed noise

If we assume that the stochastic noise is heavy-tailed, then the generalized strong growth condition equation 3 takes the following form for $\rho, \sigma \geq 0$ and $p \in (1, 2)$:

$$\begin{aligned}
\mathbb{E}\left[\|\nabla f(x, \xi)\|^p\right] &= \mathbb{E}\left[\|\nabla f(x, \xi) - \nabla f(x) + \nabla f(x)\|^p\right] \\
&\overset{①}{\leq} 2^{p-1}\mathbb{E}\left[\|\nabla f(x, \xi) - \nabla f(x)\|^p\right] + 2^{p-1}\|\nabla f(x)\|^p \\
&\overset{\text{Ass. 2.5}}{\leq} 2^{p-1}(\rho - 1)\|\nabla f(x)\|^p + 2^{p-1}\sigma^p + 2^{p-1}\|\nabla f(x)\|^p \\
&= 2^{p-1}\rho\|\nabla f(x)\|^p + 2^{p-1}\sigma^p.
\end{aligned}$$

where in ① we applied Finite form of Jensen's inequality.

If we use a mini-batch stochastic gradient $\nabla f(x, \boldsymbol{\xi}) = \frac{1}{B}\sum_{i=1}^B \nabla f(x, \xi_i)$, then Lemma C.5 should be applied.

Lemma C.5 (Kornilov et al. (2023); Hübler et al. (2025)). *Let $p \in [1, 2]$, and $X_1, \ldots, X_n \in \mathbb{R}^d$ be a martingale difference sequence (MDS), i.e.,*

$$\mathbb{E}[X_j \mid X_{j-1}, \ldots, X_1] = 0 \quad \text{a.s. for all } j = 1, \ldots, n,$$

satisfying $\mathbb{E}[\|X_j\|^p] < \infty$ for all $j = 1, \ldots, n$. Define $S_n := \sum_{j=1}^n X_j$, then

$$\mathbb{E}[\|S_n\|^p] \leq 2\sum_{j=1}^n \mathbb{E}[\|X_j\|^p]. \tag{62}$$

Historical background. The result of Lemma C.5 was initially proven by von Bahr & Esseen (1965) for $d = 1$. It was later extended to Banach spaces (Pisier, 1975, Proposition 2.4), and optimal constants were derived Cox (1982). An alternative proof for the one-dimensional i.i.d. case was rediscovered by Cherapanamjeri et al. (2022) and later extended to higher dimensions by Kornilov et al. (2023). Later, Hübler et al. (2025) corrected inaccuracies in the extension to higher dimensions.

Thus, applying Lemma C.5, we obtain:

$$\begin{aligned}
\mathbb{E}\left[\|\nabla f(x, \boldsymbol{\xi}) - \nabla f(x)\|^p\right] &\overset{①}{=} \mathbb{E}\left[\left\|\frac{1}{B}\sum_{i=1}^B \nabla f(x, \xi_i) - \nabla f(x)\right\|^p\right] \\
&\overset{②}{=} \mathbb{E}\left[\left\|\frac{1}{B}\sum_{i=1}^B (\nabla f(x, \xi_i) - \nabla f(x))\right\|^p\right] \\
&= \frac{1}{B^p}\mathbb{E}\left[\left\|\sum_{i=1}^B (\nabla f(x, \xi_i) - \nabla f(x))\right\|^p\right] \\
&\overset{\text{equation 62}}{\leq} \frac{2}{B^p}\sum_{i=1}^B \mathbb{E}\left[\|\nabla f(x, \xi_i) - \nabla f(x)\|^p\right] \\
&\overset{\text{Ass. 2.5}}{\leq} \frac{2}{B^p}\sum_{i=1}^B \left((\rho - 1)\|\nabla f(x)\|^p + \sigma^p\right)
\end{aligned}$$

$$= \frac{2}{B^{p-1}} \left((\rho - 1) \left\| \nabla f(x) \right\|^p + \sigma^p \right)$$
$$= \frac{2(\rho - 1)}{B^{p-1}} \left\| \nabla f(x) \right\|^p + \frac{2\sigma^p}{B^{p-1}}, \tag{63}$$

where in ① we used $\nabla f(x, \boldsymbol{\xi}) = \frac{1}{B} \sum_{i=1}^{B} \nabla f(x, \xi_i)$, in ② we used $\nabla f(x) = \frac{1}{B} \sum_{i=1}^{B} \nabla f(x)$. Then, in the proofs for NSGD, we use the following:

$$\begin{aligned}
\mathbb{E}\left[\left\| \nabla f(x, \boldsymbol{\xi}) - \nabla f(x) \right\| \right] &= \mathbb{E}\left[\left(\left\| \nabla f(x, \boldsymbol{\xi}) - \nabla f(x) \right\|^p \right)^{\frac{1}{p}} \right] \\
&\overset{①}{\leq} \left(\mathbb{E}\left[\left\| \nabla f(x, \boldsymbol{\xi}) - \nabla f(x) \right\|^p \right] \right)^{\frac{1}{p}} \\
&\overset{equation\ 63}{\leq} \left(\frac{2(\rho - 1)}{B^{p-1}} \left\| \nabla f(x) \right\|^p + \frac{2\sigma^p}{B^{p-1}} \right)^{\frac{1}{p}} \\
&\overset{②}{\leq} \left(\frac{1}{8^p} \left\| \nabla f(x) \right\|^p + \frac{2\sigma^p}{B^{p-1}} \right)^{\frac{1}{p}} \\
&\overset{③}{\leq} \frac{1}{8} \left\| \nabla f(x) \right\| + \frac{2^{\frac{1}{p}} \sigma}{B^{\frac{p-1}{p}}},
\end{aligned}$$

where in ① we applied Jensen's inequality, in ② we chose $B \geq 2^{\frac{3p+1}{p-1}} (\rho - 1)^{\frac{1}{p-1}}$, in ③ we applied inequality $|a+b|^{1/p} \leq |a|^{1/p} + |b|^{1/p}$, and by performing the remaining steps of the proof identically, we obtain the following final convergence rates for *non-convex setup under* (L_0, L_1)*-smooth and generalized strong growth conditions* with $\eta \leq \frac{\lambda}{L_0 + L_1 \lambda}$:

$$\min_{k \in [0, N-1]} \mathbb{E}\left[\left\| \nabla f\left(x^k\right) \right\| \right] \lesssim \frac{F_0}{\eta N} + \frac{\sigma}{B^{\frac{p-1}{p}}} + \lambda,$$

and for the *convex setup under* $(\mathcal{L}_0, \mathcal{L}_1)$*-smooth and generalized strong growth conditions* with $\eta \leq \min\left\{ \frac{\lambda}{\mathcal{L}_0 + \sqrt{\rho} \mathcal{L}_1 \lambda}, 4R_0 \right\}$:

$$\mathbb{E}\left[f\left(x^N\right) \right] - f^* \leq \left(1 - \frac{\eta}{4R_0} \right)^N F_0 + \frac{2^{3 + \frac{1}{p}} \sigma R_0}{B^{\frac{p-1}{p}}} + 6\lambda R_0,$$

with batch size $B \geq 2^{\frac{3p+1}{p-1}} (\rho - 1)^{\frac{1}{p-1}}$ and hyperparameter $\lambda > 0$.

D Extending the Analysis to $(\mathcal{H}_0, \mathcal{H}_1)$-Smoothness

In this section, we extend the analysis of Section 4.2 to functions satisfying $(\mathcal{H}_0, \mathcal{H}_1)$-smoothness Vaswani & Babanezhad (2025); Liu et al. (2025); Alimisis et al. (2025). We show that, under stochastic $(\mathcal{H}_0, \mathcal{H}_1)$-smoothness and the interpolation condition, one can recover deterministic (H_0, H_1)-smoothness with the same constants, i.e., $H_0 = \mathcal{H}_0$ and $H_1 = \mathcal{H}_1$. We also complement this extension with an analysis of (H_0, H_1)-smoothness in the deterministic setting, focusing on gradient descent under different convexity assumptions.

D.1 Assumptions on the objective function

Smoothness. We begin by introducing the stochastic $(\mathcal{H}_0, \mathcal{H}_1)$-smoothness assumption. This condition relates the local smoothness of each stochastic realization to its suboptimality gap.

Assumption D.1 (Optional)**.** Let $f(\cdot, \xi)$ be $(\mathcal{H}_0, \mathcal{H}_1)$-smooth for almost every $\xi \sim \mathcal{D}$. That is, for almost every $\xi \sim \mathcal{D}$ and all $x, y \in \mathbb{R}^d$ such that $\|y - x\| \leq \frac{1}{\sqrt{\mathcal{H}_1}}$, we have

$$\left\| \nabla f(y, \xi) - \nabla f(x, \xi) \right\| \leq \left(\mathcal{H}_0 + \mathcal{H}_1 \left[f(x, \xi) - f_\xi^* \right] \right) \|y - x\|,$$

where $f_\xi^* := \inf_{x \in \mathbb{R}^d} f(x, \xi)$.

We compare Assumption D.1 with its deterministic counterpart.

Assumption D.2 (Optional)**.** Let f be (H_0, H_1)-smooth. That is, for all $x, y \in \mathbb{R}^d$ such that $\|y - x\| \leq \frac{1}{\sqrt{H_1}}$, we have

$$\|\nabla f(y) - \nabla f(x)\| \leq (H_0 + H_1 [f(x) - f^*]) \|y - x\|.$$

The following lemma shows that, under the interpolation condition, the stochastic $(\mathcal{H}_0, \mathcal{H}_1)$-smoothness assumption implies its deterministic counterpart with the same constants.

Lemma D.3. *Let Assumption 2.2 hold and suppose that Assumption D.1 is satisfied. Assume additionally that the interpolation condition from Definition 2.3 holds. Then f satisfies Assumption D.2 with the same constants*

$$H_0 = \mathcal{H}_0, \qquad H_1 = \mathcal{H}_1.$$

That is, for all $x, y \in \mathbb{R}^d$ such that $\|y - x\| \leq \frac{1}{\sqrt{H_1}}$, we have

$$\|\nabla f(y) - \nabla f(x)\| \leq (H_0 + H_1 [f(x) - f^*]) \|y - x\|.$$

Proof. By Assumption 2.2, for any $x \in \mathbb{R}^d$,

$$\mathbb{E}[\nabla f(x, \xi)] = \nabla f(x).$$

Therefore, for any $x, y \in \mathbb{R}^d$, we have

$$\begin{aligned} \|\nabla f(y) - \nabla f(x)\| &= \|\mathbb{E}[\nabla f(y, \xi) - \nabla f(x, \xi)]\| \\ &\leq \mathbb{E}[\|\nabla f(y, \xi) - \nabla f(x, \xi)\|], \end{aligned} \tag{64}$$

where we used Jensen's inequality.

Now take arbitrary $x, y \in \mathbb{R}^d$ such that $\|y - x\| \leq \frac{1}{\sqrt{\mathcal{H}_1}}$. Applying Assumption D.1 inside the expectation in equation 64, we obtain

$$\begin{aligned} \|\nabla f(y) - \nabla f(x)\| &\leq \mathbb{E}\left[\left(\mathcal{H}_0 + \mathcal{H}_1 \left[f(x, \xi) - f_\xi^*\right]\right) \|y - x\|\right] \\ &= \left(\mathcal{H}_0 + \mathcal{H}_1 \mathbb{E}\left[f(x, \xi) - f_\xi^*\right]\right) \|y - x\| \\ &= \left(\mathcal{H}_0 + \mathcal{H}_1 \left[f(x) - \mathbb{E}\left[f_\xi^*\right]\right]\right) \|y - x\|. \end{aligned} \tag{65}$$

It remains to show that $\mathbb{E}\left[f_\xi^*\right] = f^*$. Let $x^* \in \arg\min f$. By the interpolation condition from Definition 2.3, we have

$$x^* \in \arg\min f(\cdot, \xi) \quad \text{a.s.}$$

Hence,

$$f_\xi^* = f(x^*, \xi) \quad \text{a.s.}$$

Taking expectation with respect to ξ, we get

$$\mathbb{E}\left[f_\xi^*\right] = \mathbb{E}[f(x^*, \xi)] = f(x^*) = f^*.$$

Substituting this identity into equation 65, we obtain

$$\|\nabla f(y) - \nabla f(x)\| \leq (\mathcal{H}_0 + \mathcal{H}_1 [f(x) - f^*]) \|y - x\|.$$

Finally, setting $H_0 = \mathcal{H}_0$ and $H_1 = \mathcal{H}_1$, we obtain

$$\|\nabla f(y) - \nabla f(x)\| \leq (H_0 + H_1 [f(x) - f^*]) \|y - x\|,$$

for all $x, y \in \mathbb{R}^d$ such that $\|y - x\| \leq \frac{1}{\sqrt{H_1}}$. This completes the proof. □

Assumption D.2 is strictly weaker than Assumption 2.1, and consequently also strictly weaker than classical L-smoothness. To see this, Proposition B.1 of Alimisis et al. (2025) establishes that every (L_0, L_1)-smooth function is also (H_0, H_1)-smooth with $H_0 = L_0 + L_0 L_1$ and $H_1 = \frac{4L_1^2 + L_1}{2}$. Moreover, Example 1 of Liu et al. (2025) provides a simple function, motivated by deep neural network training, that satisfies (H_0, H_1)-smoothness but does not satisfy (L_0, L_1)-smoothness. Thus, (H_0, H_1)-smoothness captures a broader class of objectives than the (L_0, L_1)-smoothness condition.

Similarly to (L_0, L_1)-smoothness, Assumption D.2 yields the following descent-type inequality: for all $x, y \in \mathbb{R}^d$ such that $\|y - x\| \leq \frac{1}{\sqrt{H_1}}$,

$$f(y) - f(x) \leq \langle \nabla f(x), y - x \rangle + \frac{H_0 + H_1 (f(x) - f^*)}{2} \|y - x\|^2. \tag{66}$$

Convexity. We also impose convexity at the level of stochastic realizations. The next lemma shows that this assumption immediately implies convexity of the expected objective.

Assumption D.4 (Optional)**.** Let $f(\cdot, \xi)$ be convex for almost every $\xi \sim \mathcal{D}$. That is, for almost every $\xi \sim \mathcal{D}$ and all $x, y \in \mathbb{R}^d$, we have

$$f(y, \xi) \geq f(x, \xi) + \langle \nabla f(x, \xi), y - x \rangle .$$

Lemma D.5. *Let Assumptions D.4 and 2.2 hold. Then f is convex, i.e., for all $x, y \in \mathbb{R}^d$, we have*

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle .$$

Proof. By Assumption D.4, for almost every $\xi \sim \mathcal{D}$ and all $x, y \in \mathbb{R}^d$, we have

$$f(y, \xi) \geq f(x, \xi) + \langle \nabla f(x, \xi), y - x \rangle .$$

Taking expectation with respect to ξ on both sides, we obtain

$$\begin{aligned} \mathbb{E}[f(y, \xi)] &\geq \mathbb{E}[f(x, \xi)] + \mathbb{E}[\langle \nabla f(x, \xi), y - x \rangle] \\ &= \mathbb{E}[f(x, \xi)] + \langle \mathbb{E}[\nabla f(x, \xi)], y - x \rangle . \end{aligned} \tag{67}$$

Using the definitions $f(x) = \mathbb{E}[f(x, \xi)]$ and $f(y) = \mathbb{E}[f(y, \xi)]$, together with Assumption 2.2, we get

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle .$$

Therefore, f is convex. □

D.2 Implication of convexity and monotone behavior of the distance to optimum

In this subsection, we derive a useful consequence of convexity and show that, under the corresponding step-size conditions, the distance to the optimum is non-increasing for both ClipSGD and NSGD in the $(\mathcal{H}_0, \mathcal{H}_1)$-smooth setting.

Implication of convexity. If Assumptions D.4 and 2.2 hold, then by Lemma D.5, the expected objective f is convex. Therefore, for all $x, y \in \mathbb{R}^d$, we have

$$f(x) - f(y) \leq \langle \nabla f(x), x - y \rangle . \tag{68}$$

Taking $y = x^*$, where $x^* \in \arg\min f$, we obtain the following relation between the function suboptimality gap and the gradient norm:

$$f\left(x^k\right) - f^* \overset{\text{equation } 68}{\leq} \left\langle \nabla f\left(x^k\right), x^k - x^* \right\rangle \overset{\text{equation } 7}{\leq} \left\|\nabla f\left(x^k\right)\right\| \left\|x^k - x^*\right\| \overset{①}{\leq} \left\|\nabla f\left(x^k\right)\right\| \underbrace{\left\|x^0 - x^*\right\|}_{=: R_0} . \tag{69}$$

Here, in ① we used the monotonicity of the distance to the optimum established below for ClipSGD and NSGD.

For the proofs below, we use the following standard self-bounding consequence of convexity and $(\mathcal{H}_0, \mathcal{H}_1)$-smoothness. For any mini-batch $\boldsymbol{\xi}^k$, define

$$F_k(\boldsymbol{\xi}^k) := f\left(x^k, \boldsymbol{\xi}^k\right) - f\left(x^*, \boldsymbol{\xi}^k\right) .$$

Under interpolation, x^* minimizes each realization almost surely, and hence also minimizes the mini-batch objective $f\left(\cdot, \boldsymbol{\xi}^k\right)$. Therefore, $F_k(\boldsymbol{\xi}^k) \geq 0$. Moreover,

$$\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\|^2 \leq 2\left(\mathcal{H}_0 + \mathcal{H}_1 F_k(\boldsymbol{\xi}^k)\right) F_k(\boldsymbol{\xi}^k). \tag{70}$$

We also use convexity of $f\left(\cdot, \boldsymbol{\xi}^k\right)$ and interpolation, which imply

$$F_k(\boldsymbol{\xi}^k) \leq \left\langle \nabla f\left(x^k, \boldsymbol{\xi}^k\right), x^k - x^* \right\rangle . \tag{71}$$

Lemma D.6. *Let Assumptions D.1, D.4, and the interpolation condition from Definition 2.3 hold. Let $H_0 = \mathcal{H}_0$ and $H_1 = \mathcal{H}_1$. Then ClipSGD from Algorithm 1 with step size*

$$\eta \leq \left[9\left(H_0 + \max\left\{\sqrt{H_1}, H_1 R_0\right\} c\right)\right]^{-1},$$

where $R_0 := \left\|x^0 - x^*\right\|$, guarantees

$$\left\|x^{k+1} - x^*\right\|^2 \leq \left\|x^k - x^*\right\|^2.$$

Consequently,

$$\left\|x^k - x^*\right\| \leq \left\|x^0 - x^*\right\| = R_0 \quad \textit{for all } k \geq 0.$$

Proof. We prove the statement by induction. The claim is trivial for $k = 0$. Assume that

$$\left\|x^k - x^*\right\| \leq R_0.$$

We show that

$$\left\|x^{k+1} - x^*\right\| \leq \left\|x^k - x^*\right\|.$$

For a mini-batch $\boldsymbol{\xi}^k$, define

$$F_k(\boldsymbol{\xi}^k) := f\left(x^k, \boldsymbol{\xi}^k\right) - f\left(x^*, \boldsymbol{\xi}^k\right).$$

By the interpolation condition, x^* minimizes each stochastic realization almost surely, and hence it also minimizes the mini-batch objective. Therefore,

$$F_k(\boldsymbol{\xi}^k) \geq 0.$$

Moreover, by convexity,

$$F_k(\boldsymbol{\xi}^k) \leq \left\langle \nabla f\left(x^k, \boldsymbol{\xi}^k\right), x^k - x^* \right\rangle. \tag{72}$$

Using the induction hypothesis and Cauchy's inequality, we also get

$$F_k(\boldsymbol{\xi}^k) \leq \left\langle \nabla f\left(x^k, \boldsymbol{\xi}^k\right), x^k - x^* \right\rangle \leq \left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\| \left\|x^k - x^*\right\| \leq \left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\| R_0. \tag{73}$$

The ClipSGD update has the form

$$x^{k+1} = x^k - \eta \alpha_{\boldsymbol{\xi}^k} \nabla f\left(x^k, \boldsymbol{\xi}^k\right),$$

where

$$\alpha_{\boldsymbol{\xi}^k} = \min\left\{1, \frac{c}{\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\|}\right\}.$$

Then

$$\begin{aligned}
\left\|x^{k+1} - x^*\right\|^2 &= \left\|x^k - x^* - \eta \alpha_{\boldsymbol{\xi}^k} \nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\|^2 \\
&= \left\|x^k - x^*\right\|^2 - 2\eta \alpha_{\boldsymbol{\xi}^k} \left\langle \nabla f\left(x^k, \boldsymbol{\xi}^k\right), x^k - x^* \right\rangle + \eta^2 \alpha_{\boldsymbol{\xi}^k}^2 \left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\|^2 \\
&\overset{\text{equation 72}}{\leq} \left\|x^k - x^*\right\|^2 - 2\eta \alpha_{\boldsymbol{\xi}^k} F_k(\boldsymbol{\xi}^k) + \eta^2 \alpha_{\boldsymbol{\xi}^k}^2 \left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\|^2.
\end{aligned} \tag{74}$$

By the self-bounding property of convex (H_0, H_1)-smooth functions applied to the mini-batch objective, we have

$$\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\|^2 \leq 2\left(H_0 + H_1 F_k(\boldsymbol{\xi}^k)\right) F_k(\boldsymbol{\xi}^k). \tag{75}$$

Substituting equation 75 into equation 74, we obtain

$$\begin{aligned}
\left\|x^{k+1} - x^*\right\|^2 \leq \left\|x^k - x^*\right\|^2 &- 2\eta \alpha_{\boldsymbol{\xi}^k} F_k(\boldsymbol{\xi}^k) \\
&+ 2\eta^2 \alpha_{\boldsymbol{\xi}^k}^2 \left(H_0 + H_1 F_k(\boldsymbol{\xi}^k)\right) F_k(\boldsymbol{\xi}^k).
\end{aligned} \tag{76}$$

We now consider two cases.

Case 1: $\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\| \leq c$**.** In this case, $\alpha_{\boldsymbol{\xi}^k} = 1$. From equation 73, we have

$$F_k(\boldsymbol{\xi}^k) \leq cR_0.$$

Therefore,

$$H_0 + H_1 F_k(\boldsymbol{\xi}^k) \leq H_0 + H_1 cR_0 \leq H_0 + \max\left\{\sqrt{H_1}, H_1 R_0\right\} c.$$

Using this bound in equation 76, we obtain

$$\begin{aligned}
\left\|x^{k+1} - x^*\right\|^2 &\leq \left\|x^k - x^*\right\|^2 - 2\eta F_k(\boldsymbol{\xi}^k) \\
&\quad + 2\eta^2 \left(H_0 + \max\left\{\sqrt{H_1}, H_1 R_0\right\} c\right) F_k(\boldsymbol{\xi}^k) \\
&\leq \left\|x^k - x^*\right\|^2 - 2\eta F_k(\boldsymbol{\xi}^k) + \frac{2}{9}\eta F_k(\boldsymbol{\xi}^k) \\
&= \left\|x^k - x^*\right\|^2 - \frac{16}{9}\eta F_k(\boldsymbol{\xi}^k) \\
&\leq \left\|x^k - x^*\right\|^2.
\end{aligned}$$

Case 2: $\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\| > c$**.** In this case,

$$\alpha_{\boldsymbol{\xi}^k} = \frac{c}{\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\|}.$$

From equation 73, we get

$$\alpha_{\boldsymbol{\xi}^k} F_k(\boldsymbol{\xi}^k) \leq cR_0.$$

Since $\alpha_{\boldsymbol{\xi}^k} \leq 1$, we have

$$\begin{aligned}
\alpha_{\boldsymbol{\xi}^k}\left(H_0 + H_1 F_k(\boldsymbol{\xi}^k)\right) &= \alpha_{\boldsymbol{\xi}^k} H_0 + H_1 \alpha_{\boldsymbol{\xi}^k} F_k(\boldsymbol{\xi}^k) \\
&\leq H_0 + H_1 cR_0 \\
&\leq H_0 + \max\left\{\sqrt{H_1}, H_1 R_0\right\} c.
\end{aligned}$$

Using this bound in equation 76, we obtain

$$\begin{aligned}
\left\|x^{k+1} - x^*\right\|^2 &\leq \left\|x^k - x^*\right\|^2 - 2\eta\alpha_{\boldsymbol{\xi}^k} F_k(\boldsymbol{\xi}^k) \\
&\quad + 2\eta^2 \alpha_{\boldsymbol{\xi}^k}\left[\alpha_{\boldsymbol{\xi}^k}\left(H_0 + H_1 F_k(\boldsymbol{\xi}^k)\right)\right] F_k(\boldsymbol{\xi}^k) \\
&\leq \left\|x^k - x^*\right\|^2 - 2\eta\alpha_{\boldsymbol{\xi}^k} F_k(\boldsymbol{\xi}^k) \\
&\quad + 2\eta^2 \alpha_{\boldsymbol{\xi}^k}\left(H_0 + \max\left\{\sqrt{H_1}, H_1 R_0\right\} c\right) F_k(\boldsymbol{\xi}^k) \\
&\leq \left\|x^k - x^*\right\|^2 - 2\eta\alpha_{\boldsymbol{\xi}^k} F_k(\boldsymbol{\xi}^k) + \frac{2}{9}\eta\alpha_{\boldsymbol{\xi}^k} F_k(\boldsymbol{\xi}^k) \\
&= \left\|x^k - x^*\right\|^2 - \frac{16}{9}\eta\alpha_{\boldsymbol{\xi}^k} F_k(\boldsymbol{\xi}^k) \\
&\leq \left\|x^k - x^*\right\|^2.
\end{aligned}$$

Combining both cases, we conclude that

$$\left\|x^{k+1} - x^*\right\|^2 \leq \left\|x^k - x^*\right\|^2.$$

Thus, by induction,

$$\left\|x^k - x^*\right\| \leq \left\|x^0 - x^*\right\| = R_0 \quad \text{for all } k \geq 0.$$

Finally, we note that the step-size condition also implies

$$\eta c \leq \frac{c}{9\left(H_0 + \sqrt{H_1} c\right)} \leq \frac{1}{\sqrt{H_1}},$$

which ensures that all ClipSGD updates remain within the locality radius required by the (H_0, H_1)-smoothness condition. This completes the proof. □

Lemma D.7. *Let Assumptions D.1, D.4, and the interpolation condition from Definition 2.3 hold. Let $H_0 = \mathcal{H}_0$ and $H_1 = \mathcal{H}_1$. Then NSGD from Algorithm 1 with any $\lambda > 0$ and step size*

$$\eta \leq \frac{\lambda}{H_0 + \max\left\{\sqrt{H_1}, H_1 R_0\right\}\lambda},$$

where $R_0 := \left\|x^0 - x^*\right\|$, guarantees

$$\left\|x^{k+1} - x^*\right\|^2 \leq \left\|x^k - x^*\right\|^2.$$

Consequently,

$$\left\|x^k - x^*\right\| \leq \left\|x^0 - x^*\right\| = R_0 \quad \textit{for all } k \geq 0.$$

Proof. We prove the statement by induction. The claim is trivial for $k = 0$. Assume that

$$\left\|x^k - x^*\right\| \leq R_0.$$

We show that

$$\left\|x^{k+1} - x^*\right\| \leq \left\|x^k - x^*\right\|.$$

For a mini-batch $\boldsymbol{\xi}^k$, define

$$F_k(\boldsymbol{\xi}^k) := f\left(x^k, \boldsymbol{\xi}^k\right) - f\left(x^*, \boldsymbol{\xi}^k\right).$$

By the interpolation condition, x^* minimizes each stochastic realization almost surely, and hence it also minimizes the mini-batch objective. Therefore,

$$F_k(\boldsymbol{\xi}^k) \geq 0.$$

Moreover, by convexity,

$$F_k(\boldsymbol{\xi}^k) \leq \left\langle \nabla f\left(x^k, \boldsymbol{\xi}^k\right), x^k - x^*\right\rangle. \tag{77}$$

Using the induction hypothesis and Cauchy's inequality, we also get

$$F_k(\boldsymbol{\xi}^k) \leq \left\langle \nabla f\left(x^k, \boldsymbol{\xi}^k\right), x^k - x^*\right\rangle \leq \left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\| \left\|x^k - x^*\right\| \leq \left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\| R_0. \tag{78}$$

Introduce the NSGD direction

$$G_\lambda\left(x^k, \boldsymbol{\xi}^k\right) := \frac{\nabla f\left(x^k, \boldsymbol{\xi}^k\right)}{\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\| + \lambda}.$$

Then the update rule of NSGD can be written as

$$x^{k+1} = x^k - \eta G_\lambda\left(x^k, \boldsymbol{\xi}^k\right).$$

Hence,

$$\begin{aligned}
\left\|x^{k+1} - x^*\right\|^2 &= \left\|x^k - x^* - \eta G_\lambda\left(x^k, \boldsymbol{\xi}^k\right)\right\|^2 \\
&= \left\|x^k - x^*\right\|^2 - 2\eta\left\langle G_\lambda\left(x^k, \boldsymbol{\xi}^k\right), x^k - x^*\right\rangle + \eta^2\left\|G_\lambda\left(x^k, \boldsymbol{\xi}^k\right)\right\|^2 \\
&= \left\|x^k - x^*\right\|^2 - \frac{2\eta}{\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\| + \lambda}\left\langle \nabla f\left(x^k, \boldsymbol{\xi}^k\right), x^k - x^*\right\rangle \\
&\qquad + \eta^2\left(\frac{\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\|}{\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\| + \lambda}\right)^2 \\
&\overset{\text{equation } 77}{\leq} \left\|x^k - x^*\right\|^2 - \frac{2\eta}{\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\| + \lambda} F_k(\boldsymbol{\xi}^k) \\
&\qquad + \eta^2\left(\frac{\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\|}{\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\| + \lambda}\right)^2.
\end{aligned} \tag{79}$$

By the self-bounding property of convex (H_0, H_1)-smooth functions applied to the mini-batch objective, we have

$$\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\|^2 \leq 2\left(H_0 + H_1 F_k(\boldsymbol{\xi}^k)\right) F_k(\boldsymbol{\xi}^k). \tag{80}$$

Substituting equation 80 into equation 79, we obtain

$$\begin{aligned}\left\|x^{k+1}-x^*\right\|^2 &\le \left\|x^k-x^*\right\|^2 - \frac{2\eta}{\|\nabla f\left(x^k,\boldsymbol{\xi}^k\right)\|+\lambda}F_k(\boldsymbol{\xi}^k)\\ &\quad+\frac{2\eta^2\left(H_0+H_1F_k(\boldsymbol{\xi}^k)\right)}{\left(\|\nabla f\left(x^k,\boldsymbol{\xi}^k\right)\|+\lambda\right)^2}F_k(\boldsymbol{\xi}^k)\\ &= \left\|x^k-x^*\right\|^2 - \frac{2\eta F_k(\boldsymbol{\xi}^k)}{\|\nabla f\left(x^k,\boldsymbol{\xi}^k\right)\|+\lambda}\left(1-\eta\frac{H_0+H_1F_k(\boldsymbol{\xi}^k)}{\|\nabla f\left(x^k,\boldsymbol{\xi}^k\right)\|+\lambda}\right). \end{aligned} \tag{81}$$

It remains to show that the term in parentheses is non-negative. From equation 78, we have

$$F_k(\boldsymbol{\xi}^k) \le R_0\left\|\nabla f\left(x^k,\boldsymbol{\xi}^k\right)\right\|.$$

Therefore,

$$\begin{aligned}\frac{H_0+H_1F_k(\boldsymbol{\xi}^k)}{\|\nabla f\left(x^k,\boldsymbol{\xi}^k\right)\|+\lambda} &\le \frac{H_0+H_1R_0\left\|\nabla f\left(x^k,\boldsymbol{\xi}^k\right)\right\|}{\|\nabla f\left(x^k,\boldsymbol{\xi}^k\right)\|+\lambda}\\ &\le \frac{H_0+H_1R_0\lambda}{\lambda}. \end{aligned} \tag{82}$$

Indeed, the last inequality follows from

$$\begin{aligned}&\frac{H_0+H_1R_0\lambda}{\lambda} - \frac{H_0+H_1R_0\left\|\nabla f\left(x^k,\boldsymbol{\xi}^k\right)\right\|}{\|\nabla f\left(x^k,\boldsymbol{\xi}^k\right)\|+\lambda}\\ &= \frac{H_0\left\|\nabla f\left(x^k,\boldsymbol{\xi}^k\right)\right\|+H_1R_0\lambda^2}{\lambda\left(\|\nabla f\left(x^k,\boldsymbol{\xi}^k\right)\|+\lambda\right)} \ge 0.\end{aligned}$$

Using the step-size condition

$$\eta \le \frac{\lambda}{H_0+\max\left\{\sqrt{H_1},H_1R_0\right\}\lambda} \le \frac{\lambda}{H_0+H_1R_0\lambda},$$

together with equation 82, we obtain

$$1-\eta\frac{H_0+H_1F_k(\boldsymbol{\xi}^k)}{\|\nabla f\left(x^k,\boldsymbol{\xi}^k\right)\|+\lambda} \ge 0.$$

Substituting this into equation 81, and using $F_k(\boldsymbol{\xi}^k) \ge 0$, we conclude that

$$\left\|x^{k+1}-x^*\right\|^2 \le \left\|x^k-x^*\right\|^2.$$

Thus, by induction,

$$\left\|x^k-x^*\right\| \le \left\|x^0-x^*\right\| = R_0 \quad \text{for all } k \ge 0.$$

Finally, we note that the step-size condition also implies

$$\eta \le \frac{\lambda}{H_0+\sqrt{H_1}\lambda} \le \frac{1}{\sqrt{H_1}}.$$

Moreover,

$$\left\|x^{k+1}-x^k\right\| = \eta\left\|\frac{\nabla f\left(x^k,\boldsymbol{\xi}^k\right)}{\|\nabla f\left(x^k,\boldsymbol{\xi}^k\right)\|+\lambda}\right\| \le \eta \le \frac{1}{\sqrt{H_1}}.$$

Thus, all NSGD updates remain within the locality radius required by the (H_0, H_1)-smoothness condition. This completes the proof. □

We next consider two settings: a stochastic setting, where we analyze ClipSGD and NSGD (see Algorithm 1) for convex (H_0, H_1)-smooth objectives, and a deterministic setting, where we analyze GD (see Algorithm 2) under various convexity assumptions.

D.3 Stochastic setting

Below, we present the convergence result for ClipSGD.

Theorem D.8 (Convex, ClipSGD). *Let $f(\cdot,\xi)$ be $(\mathcal{H}_0,\mathcal{H}_1)$-smooth (Assumption D.1) and convex (Assumption D.4). Suppose that the gradient oracle is unbiased (Assumption 2.2) and satisfies SGC (Assumption 2.4 with $\sigma = 0$). Then ClipSGD (Algorithm 1) with step size $\eta \leq \left[9\left(H_0 + \max\left\{\sqrt{H_1}, H_1 R_0\right\} c\right)\right]^{-1}$, where $H_0 = \mathcal{H}_0$, $H_1 = \mathcal{H}_1$, and with batch size $B \geq 36(\rho-1)$ guarantees*

$$\mathbb{E}\left[f\left(x^N\right)\right] - f^* \leq \max\left\{(1-\bar{\alpha})^{N/2} F_0, \frac{18 R_0^2}{N\eta}\right\},$$

where $R_0 = \left\|x^0 - x^*\right\| > 0$, $F_0 = f\left(x^0\right) - f^*$, and $\bar{\alpha} := \min\{\eta c/(36 R_0), 1\}$.

If the step size is chosen as

$$\eta = \left[9\left(H_0 + \max\left\{\sqrt{H_1}, H_1 R_0\right\} c\right)\right]^{-1},$$

then the iteration complexity N takes the following form:

$$\max\left\{\tilde{\mathcal{O}}\left(\max\left\{1, \frac{\left[H_0 + \max\left\{\sqrt{H_1}, H_1 R_0\right\} c\right] R_0}{c}\right\}\right), \mathcal{O}\left(\frac{\left[H_0 + \max\left\{\sqrt{H_1}, H_1 R_0\right\} c\right] R_0^2}{\varepsilon}\right)\right\}.$$

Comparison to the prior work. The prior best-known convergence guarantee for ClipSGD under convex (H_0, H_1)-smoothness Alimisis et al. (2025) is stated in terms of a non-standard metric. In contrast, Theorem D.8 provides a bound in the standard expected optimality gap $\mathbb{E}\left[f\left(x^N\right)\right] - f^*$.

Proof. Throughout the proof, we use the shorthand

$$F_k := f\left(x^k\right) - f^*.$$

By Lemma D.3, Assumption D.1 together with the interpolation condition implies deterministic (H_0, H_1)-smoothness of f with $H_0 = \mathcal{H}_0$ and $H_1 = \mathcal{H}_1$.

We start by applying the descent inequality for (H_0, H_1)-smooth functions:

$$\begin{aligned}
f\left(x^{k+1}\right) - f\left(x^k\right) &\overset{①}{\leq} \left\langle \nabla f\left(x^k\right), x^{k+1} - x^k \right\rangle + \frac{H_0 + H_1\left[f\left(x^k\right) - f^*\right]}{2}\left\|x^{k+1} - x^k\right\|^2 \\
&\overset{②}{=} -\eta\left\langle \nabla f\left(x^k\right), \mathrm{clip}_c\left(\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right)\right\rangle \\
&\quad + \frac{\eta^2\left(H_0 + H_1 F_k\right)}{2}\left\|\mathrm{clip}_c\left(\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right)\right\|^2 \\
&\overset{\text{equation } 69}{\leq} -\eta\left\langle \nabla f\left(x^k\right), \mathrm{clip}_c\left(\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right)\right\rangle \\
&\quad + \frac{\eta^2\left(H_0 + H_1 R_0\left\|\nabla f\left(x^k\right)\right\|\right)}{2}\left\|\mathrm{clip}_c\left(\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right)\right\|^2. \qquad (83)
\end{aligned}$$

Here, in ① we used equation 66; its locality condition $\left\|x^{k+1} - x^k\right\| \leq 1/\sqrt{H_1}$ is verified in Remark D.9. In ② we used the ClipSGD update $x^{k+1} = x^k - \eta \mathrm{clip}_c\left(\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right)$.

We now consider two cases depending on the deterministic gradient norm:

$$\left\|\nabla f\left(x^k\right)\right\| \geq c \qquad \text{and} \qquad \left\|\nabla f\left(x^k\right)\right\| < c.$$

1) The case $\left\|\nabla f\left(x^k\right)\right\| \geq c$.

Define

$$\delta := \mathbb{1}_{\left\{\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right) - \nabla f\left(x^k\right)\right\| > 3\sqrt{\frac{\rho-1}{B}}\left\|\nabla f\left(x^k\right)\right\|\right\}}.$$

Since

$$\mathrm{clip}_c\left(\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right) = \alpha_{\boldsymbol{\xi}^k} \nabla f\left(x^k, \boldsymbol{\xi}^k\right), \qquad \alpha_{\boldsymbol{\xi}^k} := \min\left\{1, \frac{c}{\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\|}\right\},$$

we have

$$\begin{aligned}
&\mathbb{E}\left[-\eta\left\langle\nabla f\left(x^k\right), \operatorname{clip}_c\left(\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right)\right\rangle \mid x^k\right] \\
&= \mathbb{P}\left(\delta=0 \mid x^k\right) \underbrace{\mathbb{E}\left[-\eta \alpha_{\boldsymbol{\xi}^k}\left\langle\nabla f\left(x^k\right), \nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\rangle \mid x^k, \delta=0\right]}_{=: T_1} \\
&\quad + \mathbb{P}\left(\delta=1 \mid x^k\right) \underbrace{\mathbb{E}\left[-\eta \alpha_{\boldsymbol{\xi}^k}\left\langle\nabla f\left(x^k\right), \nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\rangle \mid x^k, \delta=1\right]}_{=: T_2}. \qquad (84)
\end{aligned}$$

By Markov's inequality and equation 24,

$$\begin{aligned}
\mathbb{P}\left(\delta=1 \mid x^k\right) &= \mathbb{P}\left(\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)-\nabla f\left(x^k\right)\right\|^2 > 9 \frac{\rho-1}{B}\left\|\nabla f\left(x^k\right)\right\|^2 \mid x^k\right) \\
&\leq \frac{\mathbb{E}\left[\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)-\nabla f\left(x^k\right)\right\|^2 \mid x^k\right]}{9 \frac{\rho-1}{B}\left\|\nabla f\left(x^k\right)\right\|^2} \leq \frac{1}{9}. \qquad (85)
\end{aligned}$$

Hence,

$$\mathbb{P}\left(\delta=0 \mid x^k\right) \geq \frac{8}{9}.$$

We now estimate T_1. On the event $\delta=0$, we have

$$\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)-\nabla f\left(x^k\right)\right\| \leq 3 \sqrt{\frac{\rho-1}{B}}\left\|\nabla f\left(x^k\right)\right\|.$$

Therefore,

$$\begin{aligned}
T_1 &= \mathbb{E}\left[-\eta \alpha_{\boldsymbol{\xi}^k}\left\langle\nabla f\left(x^k\right), \nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\rangle \mid x^k, \delta=0\right] \\
&= \mathbb{E}\left[-\eta \alpha_{\boldsymbol{\xi}^k}\left\|\nabla f\left(x^k\right)\right\|^2-\eta \alpha_{\boldsymbol{\xi}^k}\left\langle\nabla f\left(x^k\right), \nabla f\left(x^k, \boldsymbol{\xi}^k\right)-\nabla f\left(x^k\right)\right\rangle \mid x^k, \delta=0\right] \\
&\leq \mathbb{E}\left[-\eta \alpha_{\boldsymbol{\xi}^k}\left\|\nabla f\left(x^k\right)\right\|^2+\eta \alpha_{\boldsymbol{\xi}^k}\left\|\nabla f\left(x^k\right)\right\|\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)-\nabla f\left(x^k\right)\right\| \mid x^k, \delta=0\right] \\
&\leq \mathbb{E}\left[-\eta \alpha_{\boldsymbol{\xi}^k}\left\|\nabla f\left(x^k\right)\right\|^2+3 \eta \alpha_{\boldsymbol{\xi}^k} \sqrt{\frac{\rho-1}{B}}\left\|\nabla f\left(x^k\right)\right\|^2 \mid x^k, \delta=0\right] \\
&\overset{①}{\leq} \mathbb{E}\left[-\frac{\eta \alpha_{\boldsymbol{\xi}^k}}{2}\left\|\nabla f\left(x^k\right)\right\|^2 \mid x^k, \delta=0\right] \\
&\overset{②}{\leq} -\frac{\eta c}{4}\left\|\nabla f\left(x^k\right)\right\|. \qquad (86)
\end{aligned}$$

Here, in ① we used $B \geq 36(\rho-1)$. In ②, we used that on the event $\delta=0$,

$$\begin{aligned}
\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\| &\leq\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)-\nabla f\left(x^k\right)\right\|+\left\|\nabla f\left(x^k\right)\right\| \\
&\leq\left(1+3 \sqrt{\frac{\rho-1}{B}}\right)\left\|\nabla f\left(x^k\right)\right\| \\
&\leq 2\left\|\nabla f\left(x^k\right)\right\|,
\end{aligned}$$

and, since $\left\|\nabla f\left(x^k\right)\right\| \geq c$,

$$\alpha_{\boldsymbol{\xi}^k} \geq \min\left\{1, \frac{c}{2\left\|\nabla f\left(x^k\right)\right\|}\right\} = \frac{c}{2\left\|\nabla f\left(x^k\right)\right\|}.$$

Next, we estimate T_2:

$$\begin{aligned}
T_2 &= \mathbb{E}\left[-\eta \alpha_{\boldsymbol{\xi}^k}\left\langle\nabla f\left(x^k\right), \nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\rangle \mid x^k, \delta=1\right] \\
&\leq \eta\left\|\nabla f\left(x^k\right)\right\| \mathbb{E}\left[\alpha_{\boldsymbol{\xi}^k}\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\| \mid x^k, \delta=1\right] \\
&\leq \eta c\left\|\nabla f\left(x^k\right)\right\|. \qquad (87)
\end{aligned}$$

Combining equation 84, equation 86, equation 87, and equation 85, we obtain

$$\begin{aligned}\mathbb{E}\left[-\eta\left\langle\nabla f\left(x^k\right), \text{clip}_c\left(\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right)\right\rangle \mid x^k\right] &\leq -\frac{\eta c}{4}\left\|\nabla f\left(x^k\right)\right\| \mathbb{P}\left(\delta = 0 \mid x^k\right) \\ &\quad + \eta c\left\|\nabla f\left(x^k\right)\right\| \mathbb{P}\left(\delta = 1 \mid x^k\right) \\ &\leq -\eta c\left(\frac{8}{9}\cdot\frac{1}{4} - \frac{1}{9}\right)\left\|\nabla f\left(x^k\right)\right\| \\ &= -\frac{\eta c}{9}\left\|\nabla f\left(x^k\right)\right\|. \end{aligned} \tag{88}$$

Taking conditional expectation in equation 83, using equation 88, and using $\left\|\text{clip}_c\left(\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right)\right\| \leq c$, we get

$$\begin{aligned}\mathbb{E}\left[F_{k+1} \mid x^k\right] - F_k &\leq -\frac{\eta c}{9}\left\|\nabla f\left(x^k\right)\right\| + \frac{\eta^2\left(H_0 + H_1 R_0\left\|\nabla f\left(x^k\right)\right\|\right)}{2} c^2 \\ &\overset{①}{\leq} -\frac{\eta c}{9}\left\|\nabla f\left(x^k\right)\right\| + \frac{\eta^2\left(H_0 + H_1 R_0 c\right)}{2} c\left\|\nabla f\left(x^k\right)\right\| \\ &\overset{②}{\leq} -\frac{\eta c}{9}\left\|\nabla f\left(x^k\right)\right\| + \frac{\eta^2\left(H_0 + \max\left\{\sqrt{H_1}, H_1 R_0\right\} c\right)}{2} c\left\|\nabla f\left(x^k\right)\right\| \\ &\overset{③}{\leq} -\frac{\eta c}{18}\left\|\nabla f\left(x^k\right)\right\| \\ &\overset{\text{equation } 69}{\leq} -\frac{\eta c}{18 R_0} F_k. \end{aligned} \tag{89}$$

Here, in ① we used $\left\|\nabla f\left(x^k\right)\right\| \geq c$, in ② we used $H_1 R_0 \leq \max\{\sqrt{H_1}, H_1 R_0\}$, and in ③ we used the step-size condition.

Thus, in the first case we obtain

$$\mathbb{E}\left[F_{k+1} \mid x^k\right] \leq \left(1 - \frac{\eta c}{18 R_0}\right) F_k. \tag{90}$$

2) The case $\left\|\nabla f\left(x^k\right)\right\| < c$.

In this case, $\nabla f\left(x^k\right) = \text{clip}_c\left(\nabla f\left(x^k\right)\right)$. From equation 83, we have

$$\begin{aligned}\mathbb{E}\left[F_{k+1} \mid x^k\right] - F_k &\leq -\eta\mathbb{E}\left[\left\langle\nabla f\left(x^k\right), \text{clip}_c\left(\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right)\right\rangle \mid x^k\right] \\ &\quad + \frac{\eta^2\left(H_0 + H_1 R_0\left\|\nabla f\left(x^k\right)\right\|\right)}{2}\mathbb{E}\left[\left\|\text{clip}_c\left(\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right)\right\|^2 \mid x^k\right] \\ &\overset{①}{\leq} -\eta\mathbb{E}\left[\left\langle\nabla f\left(x^k\right), \text{clip}_c\left(\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right)\right\rangle \mid x^k\right] \\ &\quad + \frac{\eta^2\left(H_0 + \max\left\{\sqrt{H_1}, H_1 R_0\right\} c\right)}{2}\mathbb{E}\left[\left\|\text{clip}_c\left(\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right)\right\|^2 \mid x^k\right] \\ &\overset{②}{=} -\frac{\eta}{2}\left\|\nabla f\left(x^k\right)\right\|^2 - \frac{\eta}{2}\mathbb{E}\left[\left\|\text{clip}_c\left(\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right)\right\|^2 \mid x^k\right] \\ &\quad + \frac{\eta}{2}\mathbb{E}\left[\left\|\text{clip}_c\left(\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right) - \nabla f\left(x^k\right)\right\|^2 \mid x^k\right] \\ &\quad + \frac{\eta^2\left(H_0 + \max\left\{\sqrt{H_1}, H_1 R_0\right\} c\right)}{2}\mathbb{E}\left[\left\|\text{clip}_c\left(\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right)\right\|^2 \mid x^k\right] \\ &= -\frac{\eta}{2}\left\|\nabla f\left(x^k\right)\right\|^2 + \frac{\eta}{2}\mathbb{E}\left[\left\|\text{clip}_c\left(\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right) - \text{clip}_c\left(\nabla f\left(x^k\right)\right)\right\|^2 \mid x^k\right] \\ &\quad - \frac{\eta}{2}\left(1 - \eta\left(H_0 + \max\left\{\sqrt{H_1}, H_1 R_0\right\} c\right)\right)\mathbb{E}\left[\left\|\text{clip}_c\left(\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right)\right\|^2 \mid x^k\right] \\ &\overset{③}{\leq} -\frac{\eta}{2}\left\|\nabla f\left(x^k\right)\right\|^2 + \frac{\eta}{2}\mathbb{E}\left[\left\|\text{clip}_c\left(\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right) - \text{clip}_c\left(\nabla f\left(x^k\right)\right)\right\|^2 \mid x^k\right]. \end{aligned} \tag{91}$$

Here, in ① we used $\left\|\nabla f\left(x^k\right)\right\| < c$ and $H_1 R_0 \le \max\{\sqrt{H_1}, H_1 R_0\}$; in ② we used the identity

$$-\eta \langle a, b\rangle = -\frac{\eta}{2}\|a\|^2 - \frac{\eta}{2}\|b\|^2 + \frac{\eta}{2}\|a-b\|^2$$

with $a = \nabla f\left(x^k\right)$ and $b = \mathrm{clip}_c\left(\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right)$; in ③ we used

$$\eta\left(H_0 + \max\left\{\sqrt{H_1}, H_1 R_0\right\} c\right) \le \frac{1}{9} \le 1.$$

Since clipping is the Euclidean projection onto the ball of radius c, it is non-expansive. Hence,

$$\begin{aligned}
\mathbb{E}\left[F_{k+1} \mid x^k\right] - F_k &\le -\frac{\eta}{2}\left\|\nabla f\left(x^k\right)\right\|^2 + \frac{\eta}{2}\mathbb{E}\left[\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right) - \nabla f\left(x^k\right)\right\|^2 \mid x^k\right] \\
&\overset{equation\ 24}{\le} -\frac{\eta}{2}\left\|\nabla f\left(x^k\right)\right\|^2 + \frac{\eta(\rho-1)}{2B}\left\|\nabla f\left(x^k\right)\right\|^2 \\
&\overset{①}{\le} -\frac{\eta}{2}\left\|\nabla f\left(x^k\right)\right\|^2 + \frac{\eta}{72}\left\|\nabla f\left(x^k\right)\right\|^2 \\
&\le -\frac{\eta}{9}\left\|\nabla f\left(x^k\right)\right\|^2 \\
&\overset{equation\ 69}{\le} -\frac{\eta}{9R_0^2}F_k^2.
\end{aligned} \tag{92}$$

where in ① we used $B \ge 36(\rho - 1)$.

Thus, in the second case we obtain

$$\mathbb{E}\left[F_{k+1} \mid x^k\right] \le F_k - \frac{\eta}{9R_0^2}F_k^2. \tag{93}$$

In summary. Let $\mathcal{F}_k$ be the sigma-algebra generated by the history up to x^k, and define

$$G_k := \left\|\nabla f\left(x^k\right)\right\|, \qquad I_k := \mathbb{1}_{\{G_k \ge c\}}.$$

The random variable I_k is $\mathcal{F}_k$-measurable. Set

$$a := \frac{\eta c}{18 R_0}, \qquad b := \frac{\eta}{9R_0^2}.$$

The two conditional estimates equation 90 and equation 93 can be combined as

$$\mathbb{E}\left[F_{k+1} \mid \mathcal{F}_k\right] \le F_k - aF_k I_k - bF_k^2(1 - I_k). \tag{94}$$

Indeed, on the event $I_k = 1$ this is exactly the large-gradient estimate, whereas on the event $I_k = 0$ it is exactly the small-gradient estimate.

Taking expectations in equation 94, we obtain

$$A_{k+1} \le A_k - a\mathbb{E}\left[F_k I_k\right] - b\mathbb{E}\left[F_k^2(1 - I_k)\right], \qquad A_k := \mathbb{E}\left[F_k\right]. \tag{95}$$

We use the following elementary inequality: for any nonnegative random variable X and any indicator random variable I,

$$a\mathbb{E}\left[XI\right] + b\mathbb{E}\left[X^2(1-I)\right] \ge \min\left\{\frac{a}{2}\mathbb{E}\left[X\right], b\left(\mathbb{E}\left[X\right]\right)^2\right\}. \tag{96}$$

Indeed, let $A = \mathbb{E}\left[X\right]$ and $r = \mathbb{E}\left[X(1-I)\right]$. Since

$$\mathbb{E}\left[X^2(1-I)\right] = \mathbb{E}\left[\left(X(1-I)\right)^2\right] \ge \left(\mathbb{E}\left[X(1-I)\right]\right)^2 = r^2,$$

we have

$$a\mathbb{E}\left[XI\right] + b\mathbb{E}\left[X^2(1-I)\right] \ge a(A-r) + br^2.$$

Minimizing the right-hand side over $r \in [0, A]$ gives

$$a(A-r) + br^2 \ge \min\left\{\frac{a}{2}A, bA^2\right\},$$

which proves equation 96.

Applying equation 96 to $X = F_k$ and $I = I_k$, inequality equation 95 yields

$$A_{k+1} \leq A_k - \min\left\{\frac{a}{2}A_k, bA_k^2\right\}. \tag{97}$$

In particular, $(A_k)_{k\geq 0}$ is non-increasing.

Define the capped contraction parameter

$$\alpha_0 := \frac{a}{2} = \frac{\eta c}{36R_0}, \qquad \bar{\alpha} := \min\{\alpha_0, 1\}, \qquad \beta := b = \frac{\eta}{9R_0^2}.$$

Since $\bar{\alpha} \leq \alpha_0$, inequality equation 97 implies

$$A_{k+1} \leq A_k - \min\{\bar{\alpha}A_k, \beta A_k^2\}. \tag{98}$$

Let

$$\tau := \min\left\{k \in \{0, \ldots, N-1\} : A_k < \frac{\bar{\alpha}}{\beta}\right\},$$

with the convention $\tau = N$ if the set is empty. Since (A_k) is non-increasing, the indices $k < \tau$ correspond to the linear part of equation 98, and the indices $k \geq \tau$ correspond to the quadratic part.

If $\tau > N/2$, then for at least $N/2$ iterations,

$$A_{k+1} \leq (1 - \bar{\alpha})A_k.$$

Using monotonicity for the remaining iterations, we get

$$A_N \leq (1 - \bar{\alpha})^{N/2} A_0 = (1 - \bar{\alpha})^{N/2} F_0. \tag{99}$$

If $\tau \leq N/2$, then for $k = \tau, \ldots, N-1$,

$$A_{k+1} \leq A_k - \beta A_k^2.$$

If $A_N = 0$, the claim is trivial. Otherwise,

$$\frac{1}{A_{k+1}} - \frac{1}{A_k} = \frac{A_k - A_{k+1}}{A_k A_{k+1}} \geq \beta \frac{A_k}{A_{k+1}} \geq \beta.$$

Summing from $k = \tau$ to $N-1$ and using $N - \tau \geq N/2$, we obtain

$$A_N \leq \frac{2}{\beta N} = \frac{18R_0^2}{N\eta}. \tag{100}$$

Combining equation 99 and equation 100, we conclude that

$$\mathbb{E}\left[f\left(x^N\right)\right] - f^* = A_N \leq \max\left\{(1 - \bar{\alpha})^{N/2} F_0, \frac{18R_0^2}{N\eta}\right\}.$$

Remark D.9. *In the first step of equation 83, we use the descent lemma for (H_0, H_1)-smooth functions equation 66. Its locality condition $\left\|x^{k+1} - x^k\right\| \leq 1/\sqrt{H_1}$ is guaranteed by the chosen step size. Indeed,*

$$\begin{aligned}\left\|x^{k+1} - x^k\right\| &= \eta \left\|clip_c\left(\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right)\right\| \leq \eta c \leq \frac{c}{9\left(H_0 + \max\left\{\sqrt{H_1}, H_1 R_0\right\} c\right)} \\ &\leq \frac{c}{9\sqrt{H_1} c} \leq \frac{1}{\sqrt{H_1}}.\end{aligned}$$

□

We now present the convergence result for NSGD.

Theorem D.10 (Convex, NSGD). *Let $f(\cdot,\xi)$ be $(\mathcal{H}_0,\mathcal{H}_1)$-smooth (Assumption D.1) and convex (Assumption D.4). Suppose that the gradient oracle is unbiased (Assumption 2.2) and satisfies SGC (Assumption 2.4 with $\sigma=0$). Then NSGD (Algorithm 1) with any $\lambda>0$, step size $\eta \le \min\left\{\frac{\lambda}{H_0+\max\left\{\sqrt{H_1},H_1R_0\right\}\lambda}, 4R_0\right\}$, where $H_0=\mathcal{H}_0$, $H_1=\mathcal{H}_1$, and batch size $B \ge 64(\rho-1)$ guarantees*

$$\mathbb{E}\left[f\left(x^N\right)\right]-f^* \le \left(1-\frac{\eta}{4R_0}\right)^N F_0 + 6\lambda R_0,$$

where $R_0=\left\|x^0-x^*\right\|>0$ and $F_0=f\left(x^0\right)-f^*$.

Theorem D.10 establishes the result for the case $\sigma=0$. The case $\sigma>0$ is handled directly in the proof below. The theorem shows that NSGD converges to an error floor of $6\lambda R_0$. To ensure convergence to a desired accuracy ε, i.e., $\mathbb{E}\left[f\left(x^N\right)\right]-f^*\le\varepsilon$, it is sufficient to choose

$$\lambda \le \frac{\varepsilon}{12R_0}.$$

Setting $\lambda=\frac{\varepsilon}{12R_0}$ and taking the maximal admissible step size

$$\eta=\min\left\{\frac{\lambda}{H_0+\max\left\{\sqrt{H_1},H_1R_0\right\}\lambda},4R_0\right\}=\min\left\{\left[\frac{12H_0R_0}{\varepsilon}+\max\left\{\sqrt{H_1},H_1R_0\right\}\right]^{-1},4R_0\right\},$$

we obtain the following iteration complexity:

$$N=\mathcal{O}\left(\max\left\{1,\left(\frac{H_0R_0^2}{\varepsilon}+\max\left\{\sqrt{H_1},H_1R_0\right\}R_0\right)\right\}\log\frac{F_0}{\varepsilon}\right).$$

Proof. For simplicity of presentation, let us introduce the notation $G_\lambda\left(x^k,\boldsymbol{\xi}^k\right)\coloneqq\frac{\nabla f\left(x^k,\boldsymbol{\xi}^k\right)}{\left\|\nabla f\left(x^k,\boldsymbol{\xi}^k\right)\right\|+\lambda}$. Then the NSGD update has the form $x^{k+1}=x^k-\eta G_\lambda\left(x^k,\boldsymbol{\xi}^k\right)$. Throughout the proof, we use the shorthand $F_k\coloneqq f\left(x^k\right)-f^*$. By Lemma D.3, Assumption D.1 together with the interpolation condition implies deterministic (H_0,H_1)-smoothness of f with $H_0=\mathcal{H}_0$ and $H_1=\mathcal{H}_1$.

Using the descent inequality for (H_0,H_1)-smooth functions, we obtain

$$\begin{aligned}
f\left(x^{k+1}\right)-f\left(x^k\right) &\overset{①}{\le} \left\langle\nabla f\left(x^k\right),x^{k+1}-x^k\right\rangle+\frac{H_0+H_1\left[f\left(x^k\right)-f^*\right]}{2}\left\|x^{k+1}-x^k\right\|^2\\
&\overset{②}{\le} \left\langle\nabla f\left(x^k\right),x^{k+1}-x^k\right\rangle\\
&\quad+\frac{H_0+\max\left\{\sqrt{H_1},H_1R_0\right\}\left\|\nabla f\left(x^k\right)\right\|}{2}\left\|x^{k+1}-x^k\right\|^2\\
&\overset{③}{=} -\eta\left\langle\nabla f\left(x^k\right),G_\lambda\left(x^k,\boldsymbol{\xi}^k\right)\right\rangle\\
&\quad+\frac{\eta^2\left(H_0+\max\left\{\sqrt{H_1},H_1R_0\right\}\left\|\nabla f\left(x^k\right)\right\|\right)}{2}\left\|G_\lambda\left(x^k,\boldsymbol{\xi}^k\right)\right\|^2\\
&\overset{④}{\le} -\eta\frac{\left\langle\nabla f\left(x^k\right),\nabla f\left(x^k,\boldsymbol{\xi}^k\right)\right\rangle}{\left\|\nabla f\left(x^k,\boldsymbol{\xi}^k\right)\right\|+\lambda}\\
&\quad+\frac{\eta^2\left(H_0+\max\left\{\sqrt{H_1},H_1R_0\right\}\left\|\nabla f\left(x^k\right)\right\|\right)}{2}\\
&= -\eta\frac{\left\langle\nabla f\left(x^k,\boldsymbol{\xi}^k\right)+\nabla f\left(x^k\right)-\nabla f\left(x^k,\boldsymbol{\xi}^k\right),\nabla f\left(x^k,\boldsymbol{\xi}^k\right)\right\rangle}{\left\|\nabla f\left(x^k,\boldsymbol{\xi}^k\right)\right\|+\lambda}\\
&\quad+\frac{\eta^2\left(H_0+\max\left\{\sqrt{H_1},H_1R_0\right\}\left\|\nabla f\left(x^k\right)\right\|\right)}{2}\\
&= -\eta\frac{\left\|\nabla f\left(x^k,\boldsymbol{\xi}^k\right)\right\|^2}{\left\|\nabla f\left(x^k,\boldsymbol{\xi}^k\right)\right\|+\lambda}-\eta\frac{\left\langle\nabla f\left(x^k\right)-\nabla f\left(x^k,\boldsymbol{\xi}^k\right),\nabla f\left(x^k,\boldsymbol{\xi}^k\right)\right\rangle}{\left\|\nabla f\left(x^k,\boldsymbol{\xi}^k\right)\right\|+\lambda}
\end{aligned}$$

$$
\begin{aligned}
&\quad + \frac{\eta^2 \left(H_0 + \max\left\{\sqrt{H_1}, H_1 R_0\right\} \left\|\nabla f\left(x^k\right)\right\|\right)}{2} \\
&\overset{⑤}{\leq} -\eta \frac{\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\|^2}{\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\| + \lambda} + \eta \frac{\left\|\nabla f\left(x^k\right) - \nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\| \left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\|}{\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\| + \lambda} \\
&\quad + \frac{\eta^2 \left(H_0 + \max\left\{\sqrt{H_1}, H_1 R_0\right\} \left\|\nabla f\left(x^k\right)\right\|\right)}{2} \\
&\leq -\eta \frac{\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\|^2}{\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\| + \lambda} + \eta \left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right) - \nabla f\left(x^k\right)\right\| \\
&\quad + \frac{\eta^2 \left(H_0 + \max\left\{\sqrt{H_1}, H_1 R_0\right\} \left\|\nabla f\left(x^k\right)\right\|\right)}{2} \\
&= -\eta \left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\| + \eta \frac{\lambda \left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\|}{\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\| + \lambda} \\
&\quad + \eta \left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right) - \nabla f\left(x^k\right)\right\| \\
&\quad + \frac{\eta^2 \left(H_0 + \max\left\{\sqrt{H_1}, H_1 R_0\right\} \left\|\nabla f\left(x^k\right)\right\|\right)}{2} \\
&\leq -\eta \left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right)\right\| + \eta\lambda + \eta \left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right) - \nabla f\left(x^k\right)\right\| \\
&\quad + \frac{\eta^2 \left(H_0 + \max\left\{\sqrt{H_1}, H_1 R_0\right\} \left\|\nabla f\left(x^k\right)\right\|\right)}{2} \\
&\overset{⑥}{\leq} -\eta \left\|\nabla f\left(x^k\right)\right\| + 2\eta \left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right) - \nabla f\left(x^k\right)\right\| \\
&\quad + \frac{\eta^2 \left(H_0 + \max\left\{\sqrt{H_1}, H_1 R_0\right\} \left\|\nabla f\left(x^k\right)\right\|\right)}{2} + \eta\lambda \\
&\overset{⑦}{\leq} -\frac{\eta}{2} \left\|\nabla f\left(x^k\right)\right\| + 2\eta \left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right) - \nabla f\left(x^k\right)\right\| + \frac{3}{2}\eta\lambda. \qquad (101)
\end{aligned}
$$

Here, in ① we used equation 66; its locality condition $\left\|x^{k+1} - x^k\right\| \leq 1/\sqrt{H_1}$ is verified in Remark D.11. In ② we used equation 69 and $H_1 R_0 \leq \max\left\{\sqrt{H_1}, H_1 R_0\right\}$. In ③ we used the NSGD update. In ④ we used $\left\|G_\lambda\left(x^k, \boldsymbol{\xi}^k\right)\right\| \leq 1$. In ⑤ we used the Cauchy–Schwarz inequality. In ⑥ we used the triangle inequality. Finally, in ⑦ we used the step-size condition and the fact that

$$
\eta \leq \frac{\lambda}{H_0 + \max\left\{\sqrt{H_1}, H_1 R_0\right\} \lambda} \leq \frac{\left\|\nabla f\left(x^k\right)\right\| + \lambda}{H_0 + \max\left\{\sqrt{H_1}, H_1 R_0\right\} \left\|\nabla f\left(x^k\right)\right\|}.
$$

Indeed,

$$
\begin{aligned}
&\frac{\left\|\nabla f\left(x^k\right)\right\| + \lambda}{H_0 + \max\left\{\sqrt{H_1}, H_1 R_0\right\} \left\|\nabla f\left(x^k\right)\right\|} - \frac{\lambda}{H_0 + \max\left\{\sqrt{H_1}, H_1 R_0\right\} \lambda} \\
&= \frac{H_0 \left\|\nabla f\left(x^k\right)\right\| + \max\left\{\sqrt{H_1}, H_1 R_0\right\} \lambda^2}{\left(H_0 + \max\left\{\sqrt{H_1}, H_1 R_0\right\} \left\|\nabla f\left(x^k\right)\right\|\right) \left(H_0 + \max\left\{\sqrt{H_1}, H_1 R_0\right\} \lambda\right)} \geq 0.
\end{aligned}
$$

Taking conditional expectation in equation 101, we get

$$
\begin{aligned}
\mathbb{E}\left[F_{k+1} \mid x^k\right] - F_k &\leq -\frac{\eta}{2} \left\|\nabla f\left(x^k\right)\right\| + 2\eta \mathbb{E}\left[\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right) - \nabla f\left(x^k\right)\right\| \mid x^k\right] + \frac{3}{2}\eta\lambda \\
&\leq -\frac{\eta}{2} \left\|\nabla f\left(x^k\right)\right\| + 2\eta \left(\mathbb{E}\left[\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right) - \nabla f\left(x^k\right)\right\|^2 \mid x^k\right]\right)^{1/2} + \frac{3}{2}\eta\lambda \\
&\overset{\text{equation 24}}{\leq} -\frac{\eta}{2} \left\|\nabla f\left(x^k\right)\right\| + 2\eta \left(\frac{(\rho - 1) \left\|\nabla f\left(x^k\right)\right\|^2 + \sigma^2}{B}\right)^{1/2} + \frac{3}{2}\eta\lambda \\
&\leq -\frac{\eta}{2} \left\|\nabla f\left(x^k\right)\right\| + 2\eta\sqrt{\frac{\rho - 1}{B}} \left\|\nabla f\left(x^k\right)\right\| + \frac{2\eta\sigma}{\sqrt{B}} + \frac{3}{2}\eta\lambda \\
&\overset{①}{\leq} -\frac{\eta}{4} \left\|\nabla f\left(x^k\right)\right\| + \frac{2\eta\sigma}{\sqrt{B}} + \frac{3}{2}\eta\lambda, \qquad (102)
\end{aligned}
$$

where in ① we used $B \geq 64(\rho - 1)$.

Thus,

$$\mathbb{E}\left[F_{k+1} \mid x^k\right] \leq F_k - \frac{\eta}{4}\left\|\nabla f\left(x^k\right)\right\| + \frac{2\eta\sigma}{\sqrt{B}} + \frac{3}{2}\eta\lambda$$
$$\overset{equation\ 69}{\leq} \left(1 - \frac{\eta}{4R_0}\right) F_k + \frac{2\eta\sigma}{\sqrt{B}} + \frac{3}{2}\eta\lambda. \tag{103}$$

Define $\gamma := \min\{\eta/(4R_0), 1\}$. Since $\gamma \leq \eta/(4R_0)$, taking full expectation in equation 103 and unrolling gives the general capped bound

$$\mathbb{E}\left[F_N\right] \leq (1-\gamma)^N F_0 + \frac{2\eta\sigma/\sqrt{B} + 3\eta\lambda/2}{\gamma}. \tag{104}$$

Under the step-size condition of Theorem D.10, $\eta \leq 4R_0$, so $\gamma = \eta/(4R_0)$ and

$$\mathbb{E}\left[F_N\right] \leq \left(1 - \frac{\eta}{4R_0}\right)^N F_0 + \frac{8\sigma R_0}{\sqrt{B}} + 6\lambda R_0.$$

For $\sigma = 0$, this gives the claimed result:

$$\mathbb{E}\left[f\left(x^N\right)\right] - f^* \leq \left(1 - \frac{\eta}{4R_0}\right)^N F_0 + 6\lambda R_0.$$

Remark D.11. *In the first step of equation 101, we use the descent lemma for (H_0, H_1)-smooth functions equation 66. Its locality condition $\left\|x^{k+1} - x^k\right\| \leq 1/\sqrt{H_1}$ is guaranteed by the chosen step size. Indeed,*

$$\left\|x^{k+1} - x^k\right\| = \eta\left\|G_\lambda\left(x^k, \boldsymbol{\xi}^k\right)\right\| \leq \eta \leq \frac{\lambda}{H_0 + \max\left\{\sqrt{H_1}, H_1 R_0\right\}\lambda} \leq \frac{\lambda}{\sqrt{H_1}\lambda} = \frac{1}{\sqrt{H_1}}.$$

□

D.3.1 Extension to heavy-tailed noise

We now extend Theorem D.10 to the heavy-tailed setting. The proof follows the same argument as in Theorem D.10; the only modification is the way we control the stochastic error term

$$\mathbb{E}\left[\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right) - \nabla f\left(x^k\right)\right\| \mid x^k\right].$$

Instead of the second-moment bound equation 24, we use the heavy-tailed mini-batch estimate from Appendix C.2.3. Namely, under Assumption 2.5, for $p \in (1, 2)$ and mini-batch size

$$B \geq 2^{\frac{3p+1}{p-1}}(\rho - 1)^{\frac{1}{p-1}},$$

we have

$$\mathbb{E}\left[\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right) - \nabla f\left(x^k\right)\right\| \mid x^k\right] \leq \frac{1}{8}\left\|\nabla f\left(x^k\right)\right\| + \frac{2^{1/p}\sigma}{B^{\frac{p-1}{p}}}. \tag{105}$$

Proposition D.12 (Heavy-tailed, convex NSGD)**.** *Let $f(\cdot, \xi)$ be $(\mathcal{H}_0, \mathcal{H}_1)$-smooth (Assumption D.1) and convex (Assumption D.4). Suppose that the gradient oracle is unbiased (Assumption 2.2) and satisfies the generalized p-th central moment condition (Assumption 2.5). Then NSGD (Algorithm 1) with any $\lambda > 0$, step size $\eta \leq \min\left\{\frac{\lambda}{H_0 + \max\left\{\sqrt{H_1}, H_1 R_0\right\}\lambda}, 4R_0\right\}$, where $H_0 = \mathcal{H}_0$, $H_1 = \mathcal{H}_1$, and batch size $B \geq 2^{\frac{3p+1}{p-1}}(\rho - 1)^{\frac{1}{p-1}}$, guarantees*

$$\mathbb{E}\left[f\left(x^N\right)\right] - f^* \leq \left(1 - \frac{\eta}{4R_0}\right)^N F_0 + \frac{2^{3+\frac{1}{p}}\sigma R_0}{B^{\frac{p-1}{p}}} + 6\lambda R_0,$$

where $R_0 = \left\|x^0 - x^*\right\| > 0$ and $F_0 = f\left(x^0\right) - f^*$. In particular, when $\sigma = 0$, this reduces to

$$\mathbb{E}\left[f\left(x^N\right)\right] - f^* \leq \left(1 - \frac{\eta}{4R_0}\right)^N F_0 + 6\lambda R_0.$$

Proof. The proof is identical to the proof of Theorem D.10 up to the estimate of the stochastic error term. Starting from equation 101, taking conditional expectation, and using equation 105, we obtain

$$\begin{aligned}
\mathbb{E}\left[F_{k+1} \mid x^k\right] - F_k &\leq -\frac{\eta}{2}\left\|\nabla f\left(x^k\right)\right\| + 2\eta\mathbb{E}\left[\left\|\nabla f\left(x^k, \boldsymbol{\xi}^k\right) - \nabla f\left(x^k\right)\right\| \mid x^k\right] + \frac{3}{2}\eta\lambda \\
&\leq -\frac{\eta}{2}\left\|\nabla f\left(x^k\right)\right\| + \frac{\eta}{4}\left\|\nabla f\left(x^k\right)\right\| + \frac{2^{1+\frac{1}{p}}\eta\sigma}{B^{\frac{p-1}{p}}} + \frac{3}{2}\eta\lambda \\
&= -\frac{\eta}{4}\left\|\nabla f\left(x^k\right)\right\| + \frac{2^{1+\frac{1}{p}}\eta\sigma}{B^{\frac{p-1}{p}}} + \frac{3}{2}\eta\lambda. \qquad (106)
\end{aligned}$$

Using the convexity implication equation 69, we further get

$$\mathbb{E}\left[F_{k+1} \mid x^k\right] \leq \left(1 - \frac{\eta}{4R_0}\right) F_k + \frac{2^{1+\frac{1}{p}}\eta\sigma}{B^{\frac{p-1}{p}}} + \frac{3}{2}\eta\lambda. \qquad (107)$$

Since the proposition assumes $\eta \leq 4R_0$, its contraction factor satisfies $0 \leq 1 - \eta/(4R_0) < 1$. Taking full expectation and unrolling the recursion yields

$$\begin{aligned}
\mathbb{E}\left[F_N\right] &\leq \left(1 - \frac{\eta}{4R_0}\right)^N F_0 + \left(\frac{2^{1+\frac{1}{p}}\eta\sigma}{B^{\frac{p-1}{p}}} + \frac{3}{2}\eta\lambda\right)\sum_{i=0}^{N-1}\left(1 - \frac{\eta}{4R_0}\right)^i \\
&\leq \left(1 - \frac{\eta}{4R_0}\right)^N F_0 + \left(\frac{2^{1+\frac{1}{p}}\eta\sigma}{B^{\frac{p-1}{p}}} + \frac{3}{2}\eta\lambda\right)\frac{4R_0}{\eta} \\
&= \left(1 - \frac{\eta}{4R_0}\right)^N F_0 + \frac{2^{3+\frac{1}{p}}\sigma R_0}{B^{\frac{p-1}{p}}} + 6\lambda R_0.
\end{aligned}$$

This completes the proof. □

D.4 Deterministic setting

Using the (H_0, H_1)-smoothness assumption allows us to improve existing convergence bounds of gradient descent with the warm-up stepsize strategy. Throughout this subsection, we assume $H_0, H_1 > 0$; the case $H_1 = 0$ reduces to the standard H_0-smooth setting.

Algorithm 2 Gradient Descent Method (GD)

Input: $x_0 \in \mathbb{R}^d$, iterations number N, $H_0, H_1 \geq 0$, universal constant $\theta \in (0, 3/4]$
for $k = 0$ **to** $N - 1$ **do**
  $x^{k+1} \leftarrow x^k - \theta \min\left\{\frac{1}{H_0}, \frac{1}{3H_1(f(x^k) - f^*)}\right\}\nabla f(x^k)$
end for
Return: x^N

Here and below, if $F_k := f\left(x^k\right) - f^*) = 0$, we use the convention $\frac{1}{3H_1F_k} = +\infty$, so that $\eta_k = \theta/H_0$. The step size of Algorithm 2 corresponds to a warm-up strategy: as the gap F_k decreases, the admissible step size increases until it reaches the constant regime θ/H_0.

In what follows, we consider non-convex, convex (Assumption B.1), and strongly convex objectives.

Assumption D.13 (Strong-convexity)**.** There exists a constant $\mu > 0$ such that function f satisfies for all $x, y \in \mathbb{R}^d$:

$$f(x) - f(y) + \frac{\mu}{2}\|x - y\|^2 \leq \langle \nabla f(x), x - y\rangle. \qquad (108)$$

We first record two auxiliary facts that will be used in all deterministic proofs.

Lemma D.14 (Locality of the GD step)**.** *Let f be (H_0, H_1)-smooth (Assumption D.2) and let $F_k := f\left(x^k\right) - f^* \geq 0$. Then GD from Algorithm 2 satisfies*

$$\left\|x^{k+1} - x^k\right\| = \eta_k \left\|\nabla f\left(x^k\right)\right\| \leq \frac{1}{\sqrt{H_1}}.$$

Consequently, the local (H_0, H_1)-smoothness inequality can be applied to the pair (x^k, x^{k+1}).

Proof. Let $G_k := \left\|\nabla f\left(x^k\right)\right\|$. If $G_k = 0$, the claim is immediate. Otherwise, set

$$y^k := x^k - \frac{1}{\sqrt{H_1}} \frac{\nabla f\left(x^k\right)}{G_k}.$$

Since $\left\|y^k - x^k\right\| = 1/\sqrt{H_1}$, Assumption D.2 gives

$$\begin{aligned} f\left(y^k\right) &\leq f\left(x^k\right) + \left\langle \nabla f\left(x^k\right), y^k - x^k \right\rangle + \frac{H_0 + H_1 F_k}{2} \left\|y^k - x^k\right\|^2 \\ &= f\left(x^k\right) - \frac{G_k}{\sqrt{H_1}} + \frac{H_0 + H_1 F_k}{2H_1}. \end{aligned}$$

Using $f\left(y^k\right) \geq f^*$, we obtain

$$G_k \leq \frac{H_0}{2\sqrt{H_1}} + \frac{3\sqrt{H_1}}{2} F_k.$$

Therefore,

$$\begin{aligned} \eta_k G_k &\leq \theta \min\left\{\frac{1}{H_0}, \frac{1}{3H_1 F_k}\right\} \left(\frac{H_0}{2\sqrt{H_1}} + \frac{3\sqrt{H_1}}{2} F_k\right) \\ &\leq \theta \left(\frac{1}{2\sqrt{H_1}} + \frac{1}{2\sqrt{H_1}}\right) = \frac{\theta}{\sqrt{H_1}} \leq \frac{1}{\sqrt{H_1}}. \end{aligned}$$

□

Lemma D.15 (Basic descent)**.** *Let f be (H_0, H_1)-smooth (Assumption D.2). Then GD from Algorithm 2 satisfies*

$$F_{k+1} \leq F_k - \frac{\eta_k}{2} \left\|\nabla f\left(x^k\right)\right\|^2. \tag{109}$$

In particular, $(F_k)_{k \geq 0}$ is non-increasing.

Proof. By Lemma D.14, the update remains in the local smoothness region. Hence

$$\begin{aligned} F_{k+1} - F_k &= f\left(x^{k+1}\right) - f\left(x^k\right) \\ &\leq -\eta_k \left\|\nabla f\left(x^k\right)\right\|^2 + \frac{\eta_k^2}{2} \left(H_0 + H_1 F_k\right) \left\|\nabla f\left(x^k\right)\right\|^2. \end{aligned}$$

Moreover,

$$\eta_k (H_0 + H_1 F_k) \leq \theta \left(1 + \frac{1}{3}\right) = \frac{4\theta}{3} \leq 1.$$

Thus,

$$F_{k+1} - F_k \leq -\frac{\eta_k}{2} \left\|\nabla f\left(x^k\right)\right\|^2.$$

□

D.4.1 NON-CONVEX SETTING

Theorem D.16. *Let f be (H_0, H_1)-smooth (Assumption D.2). Then GD (Algorithm 2) guarantees*

$$\min_{0\le k\le N-1} \left\|\nabla f\left(x^k\right)\right\| \le \begin{cases} \sqrt{\frac{2H_0F_0}{\theta N}}, & M=0, \\ \min\left\{\sqrt{\frac{6H_1F_0\sum_{k=0}^{M-1}F_k}{\theta M^2}}, \sqrt{\frac{2H_0F_0}{\theta(N-M)}}\right\}, & 1\le M\le N-1, \\ \sqrt{\frac{6H_1F_0\sum_{k=0}^{N-1}F_k}{\theta N^2}}, & M=N, \end{cases}$$

where $F_k = f\left(x^k\right) - f^*$, $\theta = \frac{1}{4\sqrt{2}+4}$, and M is the smallest index $k \in \{0,\dots,N-1\}$ such that $F_k < \frac{H_0}{3H_1}$, with the convention $M = N$ if this set is empty.

Proof. From Lemma D.15, we have

$$F_k - F_{k+1} \ge \frac{\eta_k}{2}\left\|\nabla f\left(x^k\right)\right\|^2. \tag{110}$$

If $F_k \ge H_0/(3H_1)$, then

$$\eta_k = \frac{\theta}{3H_1F_k}.$$

Therefore,

$$\frac{\theta}{6H_1F_k}\left\|\nabla f\left(x^k\right)\right\|^2 \le F_k - F_{k+1}. \tag{111}$$

If $F_k < H_0/(3H_1)$, then

$$\eta_k = \frac{\theta}{H_0}.$$

Therefore,

$$\left\|\nabla f\left(x^k\right)\right\|^2 \le \frac{2H_0}{\theta}\left(F_k - F_{k+1}\right). \tag{112}$$

By Lemma D.15, the sequence (F_k) is non-increasing. Hence, by the definition of M, the indices $k = 0,\dots,M-1$ correspond to the large-gap phase, and the indices $k = M,\dots,N-1$ correspond to the small-gap phase.

Assume first that $1 \le M \le N-1$. Summing equation 111 for $k = 0,\dots,M-1$, we get

$$\sum_{k=0}^{M-1}\frac{\theta}{6H_1F_k}\left\|\nabla f\left(x^k\right)\right\|^2 \le \sum_{k=0}^{M-1}\left(F_k - F_{k+1}\right) \le F_0.$$

Using

$$\sum_{k=0}^{M-1}\frac{\theta}{6H_1F_k} \ge \frac{\theta M^2}{6H_1\sum_{k=0}^{M-1}F_k},$$

we obtain

$$\min_{0\le k\le M-1}\left\|\nabla f\left(x^k\right)\right\| \le \sqrt{\frac{6H_1F_0\sum_{k=0}^{M-1}F_k}{\theta M^2}}. \tag{113}$$

Summing equation 112 for $k = M,\dots,N-1$, we get

$$\sum_{k=M}^{N-1}\left\|\nabla f\left(x^k\right)\right\|^2 \le \frac{2H_0}{\theta}\sum_{k=M}^{N-1}\left(F_k - F_{k+1}\right) \le \frac{2H_0F_0}{\theta}.$$

Thus,

$$\min_{M\le k\le N-1}\left\|\nabla f\left(x^k\right)\right\| \le \sqrt{\frac{2H_0F_0}{\theta(N-M)}}. \tag{114}$$

Combining equation 113 and equation 114 gives the interior case.

If $M = 0$, then all indices are in the small-gap phase, and summing equation 112 for $k = 0, \dots, N-1$ gives

$$\min_{0 \le k \le N-1} \left\|\nabla f\left(x^k\right)\right\| \le \sqrt{\frac{2H_0F_0}{\theta N}}.$$

If $M = N$, then all indices are in the large-gap phase, and summing equation 111 for $k = 0, \dots, N-1$ gives

$$\min_{0 \le k \le N-1} \left\|\nabla f\left(x^k\right)\right\| \le \sqrt{\frac{6H_1F_0\sum_{k=0}^{N-1} F_k}{\theta N^2}}.$$

□

D.4.2 Convex setting

Lemma D.17 (Monotonicity of the distance to the optimum)**.** *Let f be (H_0, H_1)-smooth (Assumption D.2) and convex (Assumption B.1). Then GD from Algorithm 2 satisfies*

$$\left\|x^k - x^*\right\| \le \left\|x^0 - x^*\right\| = R \quad \text{for all } k \ge 0. \tag{115}$$

Proof. By Lemma D.15, we have

$$F_{k+1} \le F_k - \frac{\eta_k}{2}\left\|\nabla f\left(x^k\right)\right\|^2.$$

Since $F_{k+1} \ge 0$, this implies

$$\eta_k \left\|\nabla f\left(x^k\right)\right\|^2 \le 2F_k. \tag{116}$$

Using convexity, we obtain

$$\begin{aligned}
\left\|x^{k+1} - x^*\right\|^2 &= \left\|x^k - x^* - \eta_k \nabla f\left(x^k\right)\right\|^2 \\
&= \left\|x^k - x^*\right\|^2 - 2\eta_k\left\langle \nabla f\left(x^k\right), x^k - x^*\right\rangle + \eta_k^2\left\|\nabla f\left(x^k\right)\right\|^2 \\
&\le \left\|x^k - x^*\right\|^2 - 2\eta_k F_k + \eta_k^2\left\|\nabla f\left(x^k\right)\right\|^2 \\
&\overset{equation\ 116}{\le} \left\|x^k - x^*\right\|^2.
\end{aligned}$$

The result follows by induction. □

Theorem D.18. *Let f be (H_0, H_1)-smooth (Assumption D.2) and convex (Assumption B.1). Then GD (Algorithm 2) guarantees the following bound. Let*

$$T := \min\left\{k \in \{0, \dots, N-1\} : F_k < \frac{H_0}{3H_1}\right\},$$

with the convention $T = N$ if this set is empty. If $T < N$, then

$$f\left(x^N\right) - f^* \le \min\left\{\left(1 - \frac{\theta}{6H_1R^2}\right)^T F_0, \frac{2H_0R^2}{\theta(N-T)}\right\},$$

where $R = \left\|x^0 - x^*\right\|$ and $\theta = \frac{1}{4\sqrt{2}+4}$. If $T = N$, then

$$f\left(x^N\right) - f^* \le \left(1 - \frac{\theta}{6H_1R^2}\right)^N F_0.$$

Proof. By convexity and Lemma D.17,

$$F_k \le \left\langle \nabla f\left(x^k\right), x^k - x^*\right\rangle \le R\left\|\nabla f\left(x^k\right)\right\|. \tag{117}$$

Combining equation 117 with Lemma D.15, we get

$$F_{k+1} \le F_k - \frac{\eta_k}{2}\left\|\nabla f\left(x^k\right)\right\|^2$$

$$\leq F_k - \frac{\theta}{2R^2}\frac{F_k^2}{\max\{H_0, 3H_1F_k\}}. \tag{118}$$

If $F_k \geq H_0/(3H_1)$, then

$$\max\{H_0, 3H_1F_k\} = 3H_1F_k,$$

and equation 118 yields

$$F_{k+1} \leq \left(1 - \frac{\theta}{6H_1R^2}\right) F_k. \tag{119}$$

If $F_k < H_0/(3H_1)$, then

$$\max\{H_0, 3H_1F_k\} = H_0,$$

and equation 118 yields

$$F_{k+1} \leq F_k - \frac{\theta}{2H_0R^2}F_k^2. \tag{120}$$

By Lemma D.15, (F_k) is non-increasing. Therefore, by the definition of T, the recursion equation 119 applies for $k = 0, \ldots, T-1$, while equation 120 applies for $k = T, \ldots, N-1$ if $T < N$.

Unrolling equation 119, we get

$$F_T \leq \left(1 - \frac{\theta}{6H_1R^2}\right)^T F_0. \tag{121}$$

Since (F_k) is non-increasing, $F_N \leq F_T$, and hence

$$F_N \leq \left(1 - \frac{\theta}{6H_1R^2}\right)^T F_0. \tag{122}$$

If $T < N$, then from equation 120,

$$F_{k+1} \leq F_k - \frac{\theta}{2H_0R^2}F_k^2, \qquad k = T, \ldots, N-1.$$

If $F_N = 0$, the desired bound is immediate. Otherwise,

$$\frac{1}{F_{k+1}} - \frac{1}{F_k} = \frac{F_k - F_{k+1}}{F_kF_{k+1}} \geq \frac{\theta}{2H_0R^2}\frac{F_k}{F_{k+1}} \geq \frac{\theta}{2H_0R^2}.$$

Summing from $k = T$ to $N-1$, we obtain

$$F_N \leq \frac{2H_0R^2}{\theta(N-T)}. \tag{123}$$

Combining equation 122 and equation 123 gives the claim for $T < N$. If $T = N$, only the linear phase occurs, and equation 119 gives

$$F_N \leq \left(1 - \frac{\theta}{6H_1R^2}\right)^N F_0.$$

□

D.4.3 Strongly convex setting

Theorem D.19. *Let f be (H_0, H_1)-smooth (Assumption D.2) and μ-strongly convex (Assumption D.13). Then GD (Algorithm 2) needs at most*

$$N = \mathcal{O}\left(\frac{H_0}{\theta\mu}\log\left(\frac{F_0}{\varepsilon}\right) + \frac{H_1F_0}{\theta\mu}\right)$$

iterations to reach accuracy $f\left(x^N\right) - f^* \leq \varepsilon$, where $\theta = \frac{1}{4\sqrt{2}+4}$.

Proof. Since a strongly convex function is also convex, we apply Theorem D.18 with restarts. By strong convexity,

$$R^2 = \left\|x^0 - x^*\right\|^2 \overset{equation\ 108}{\leq} \frac{2}{\mu}\left(f\left(x^0\right) - f^*\right) = \frac{2F_0}{\mu}.$$

Therefore, the convex bound implies

$$\begin{aligned} f\left(x^N\right) - f^* &\leq \min\left\{\left(1 - \frac{\theta}{6H_1R^2}\right)^T F_0, \frac{2H_0R^2}{\theta(N-T)}\right\} \\ &\leq \min\left\{\exp\left(-\frac{\theta T}{6H_1R^2}\right) F_0, \frac{4H_0F_0}{\theta\mu(N-T)}\right\} \\ &\leq \min\left\{\exp\left(-\frac{\theta\mu T}{12H_1F_0}\right) F_0, \frac{4H_0F_0}{\theta\mu(N-T)}\right\}. \end{aligned}$$

Thus, to guarantee $f\left(x^N\right) - f^* \leq F_0/2$, it is sufficient to take

$$N \geq \frac{8H_0}{\theta\mu} + \frac{12H_1F_0}{\theta\mu}\log 2.$$

Indeed, if $T \geq \frac{12H_1F_0}{\theta\mu}\log 2$, the exponential term is at most $F_0/2$; otherwise, $N - T \geq \frac{8H_0}{\theta\mu}$ and the sublinear term is at most $F_0/2$.

Repeating this halving argument gives

$$N = \mathcal{O}\left(\frac{H_0}{\theta\mu}\log\left(\frac{F_0}{\varepsilon}\right) + \frac{H_1F_0}{\theta\mu}\right).$$

□

Remark D.20 (Non-convex + PL)**.** *If the Polyak-Łojasiewicz (PL) inequality*

$$\|\nabla f(x)\|^2 \geq 2\mu(f(x) - f^*)$$

holds for all $x \in \mathbb{R}^d$ with $\mu > 0$, then the same recursion as in Lemma D.15 gives

$$F_{k+1} \leq F_k - \eta_k\mu F_k.$$

Consequently, the same warm-up argument yields the complexity

$$\mathcal{O}\left(\frac{H_0}{\theta\mu}\log\left(\frac{F_0}{\varepsilon}\right) + \frac{H_1F_0}{\theta\mu}\right)$$

iterations to reach $f\left(x^N\right) - f^* \leq \varepsilon$.