# CONVERGENCE AND RELIABILITY LIMITS OF STANDARD NONCONVEX SGD UNDER HEAVY-TAILED NOISE

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

## ABSTRACT

Whether standard SGD converges for smooth nonconvex objectives under heavy-tailed gradient noise remains unresolved under the basic combination of standard $L$-smoothness and a bounded $\alpha$-th central moment of the stochastic-gradient noise. We resolve this question and further characterize the high-probability reliability limit of the standard SGD recursion.

We study smooth nonconvex optimization under unbiased stochastic gradients whose noise has a bounded $\alpha$-th central moment with $\alpha \in (1, 2]$. We first prove that, under a confidence-calibrated learning rate, it holds $\frac{1}{T}\sum_{t<T} \|\nabla f(\boldsymbol{w}^t)\|^2 = \mathcal{O}\left(\frac{L\Delta}{T} + \frac{\sigma\sqrt{L\Delta}}{\delta^{1/\alpha}T^{(\alpha-1)/\alpha}}\right)$ with at least $1-\delta$ probability. We then characterize the corresponding limitation of standard nonconvex SGD for optimization with heavy-tailed noise. For any deterministic learning rate schedule satisfying $0 < \gamma_t \leq 1/L$, we establish a high-probability lower bound as $\Theta\left(\frac{1}{\delta^{1/\alpha}T^{(\alpha-1)/\alpha}}\right)$ in the nontrivial regime. Such a result shows that standard nonconvex SGD cannot attain the rate achieved by gradient clipping under heavy-tailed noise, which cannot be removed by learning rate tuning alone.

## 1 INTRODUCTION

Reliable artificial intelligence depends not only on the models themselves, but also on the training algorithms used. While Stochastic Gradient Descent (SGD) has been a central workhorse in modern AI tasks training (Robbins & Monro, 1951; Bottou et al., 2018), its classical guarantees often rely on finite-variance assumptions that may fail under heavy-tailed gradient noise observed in various applications (Gürbüzbalaban et al., 2021; Simsekli et al., 2019; Zhang et al., 2020). Although there has been a line of research on the heavy-tailed stochastic optimization, most of them study the convergence performance for smooth nonconvex problems rely on gradient clipping (normalization), mini-batching, or additional structural assumptions. Indeed, the convergence of single-sample standard nonconvex SGD under standard $L$-smoothness and only a bounded $\alpha$-th central moment of the stochastic gradient noise is still unresolved. This paper is devoted to this convergence question and further characterizes the resulting high-probability reliability limit of standard nonconvex SGD.

### 1.1 RELIABILITY QUESTION FOR STANDARD NONCONVEX SGD WITH LIPSCHITZ GRADIENT UNDER HEAVY-TAILED NOISE

**What Is Known and What Remains Open.** In (Zhang et al., 2020), the authors provided a simple example to show that standard SGD may have an infinite expected squared gradient norm under heavy-tailed noise. Recent positive results, however, do not fully resolve the standard smooth nonconvex setting. In particular, Fatkhullin et al. (2025) establish convergence of vanilla nonconvex SGD under a much stronger bounded $p$-th moment assumption on the stochastic gradient or mini-batching. Yamada et al. (2026) analyze vanilla SGD and SGDM under Hölder-continuous gradients with the compatibility condition $\nu + 1 \leq p$. However, standard $L$-smoothness ($\nu = 1$) is not covered when the noise is strictly heavy-tailed ($p < 2$). We list some very close works on this topic in Table 1. In summary, whether single-sample standard nonconvex SGD converges under standard $L$-smoothness and only bounded $\alpha$-th central noise moments remains a genuine question.



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.

Table 1: Closest results on the convergence of standard SGD for nonconvex optimization under heavy-tailed noise.

| Reference | Method | Regularity | Noise | Extra condition | Guarantee / rate | Covers standard setting? |
|---|---|---|---|---|---|---|
| Zhang et al. (2020) | SGD | $L$-smooth | $\alpha$-BCM | None | $\mathbb{E}\|\nabla f(\boldsymbol{w}^t)\|^2 = +\infty$ possible | No positive convergence result |
| Scaman & Malherbe (2020) | SGD | $L$-smooth | heavy-tail | $f$ Lipschitz or symmetric noise | $\mathcal{O}\left(\frac{1}{\delta^{(\alpha+2)/(2\alpha)}T^{(\alpha-1)/\alpha}}\right)$ | No |
| Li & Liu (2022) | SGD | sample-wise $L$-smooth | sub-Weibull | $\eta_t\|\nabla F_S(\boldsymbol{w}^t)\| \le G$ | $\widetilde{\mathcal{O}}(T^{-1/2})$ HP | No |
| Madden et al. (2024) | SGD | strong smoothness | sub-Weibull | $f$ Lipschitz for $\theta > 1/2$ | $\widetilde{\mathcal{O}}(T^{-1/2})$ HP | No |
| Fatkhullin et al. (2025) | SGD | Hölder-smooth | $\alpha$-BM | bounded stochastic-gradient moment | Exp. convergence | No |
| Fatkhullin et al. (2025) | Mini-batch SGD | $L$-smooth | $\alpha$-BCM | mini-batch | Exp. convergence | No |
| Yamada et al. (2026) | SGD / SGDM | Hölder-smooth | $\alpha$-BCM | $\nu + 1 \le \alpha$ | Exp. convergence | No for $\nu = 1$, $\alpha < 2$ |
| Zhu et al. (2026) | projected SGD | weakly convex | $\alpha$-BCM | bounded domain | Exp. convergence | No |
| **Ours** | **SGD** | **$L$-smooth** | **$\alpha$-BCM** | **None** | $\mathcal{O}\left(\frac{1}{\delta^{1/\alpha}T^{(\alpha-1)/\alpha}}\right)$ **HP** | **Yes** |

The displayed rates suppress problem-dependent constants and lower-order deterministic terms. "Exp. convergence" denotes convergence guarantees stated in expectation, rather than in high probability. The $\alpha$-BM condition $\mathbb{E}\|\boldsymbol{g}\|^\alpha \le G^\alpha$ implies $\|\nabla f(\boldsymbol{w})\| \le G$ by Jensen's inequality. Sub-Weibull noise has finite moments of every order and is therefore strictly more restrictive, for the present purpose, than an $\alpha$-BCM assumption with $\alpha < 2$, which permits infinite variance.

A common way to analyze SGD is through second-moment quantities, typically under the bounded-variance assumption. Under heavy-tailed noise, however, the second moment of the stochastic gradient may be infinite. In particular, prior examples in (Zhang et al., 2020) show that standard nonconvex SGD may fail to produce iterates or stochastic gradients with finite second moments under heavy-tailed noise. Such observations invalidate standard variance-based analyses, *but they do not by themselves rule out convergence in weaker senses, such as probability convergence.* This distinction motivates the main question studied in this paper:

> Can standard nonconvex SGD be guaranteed to converge with high-probability form uunder standard $L$-smoothness and only a bounded $\alpha$-th central moment assumption on the noise? If the answer is positive, what is the correct finite-time rate, and how does it compare with the rates achieved by clipping-based methods?

Our results answer these questions by giving both an upper bound and a matching algorithm-specific lower bound for standard nonconvex SGD. The upper bound shows that unclipped SGD does converge in high probability under heavy-tailed noise. The lower bound further shows that the resulting dependence on $(T, \delta)$ is essentially unavoidable for standard SGD, thereby explaining why clipping can still be beneficial even though it is not strictly necessary for high-probability convergence.

**Numerical Illustration: Convergence with Probability is Completely Possible.** We present numerical results to illustrate the behavior of standard nonconvex SGD under an explicitly injected infinite-variance noise component. We consider VGG16 (Simonyan & Zisserman, 2015) on CIFAR-10 (Krizhevsky et al., 2009), ResNet20 (He et al., 2016) on CIFAR-100, and ResNet18 on Tiny-ImageNet (Le et al., 2015). We train with a batch size of 128, an initial learning rate of 0.1 (decayed by 0.1 at the 41st and 81st epoch), and a weight decay of $5 \times 10^{-4}$ for 100 epochs.

However, training on image classification tasks alone does not necessarily ensure that the stochastic noise follows a heavy-tailed distribution. To address this limitation, we introduce an additional heavy-tailed component to the stochastic gradient, thereby explicitly incorporating the desired distributional properties into the noise. Consider the 1-dimensional stochastic variable $\xi \in \mathbb{R}$, whose

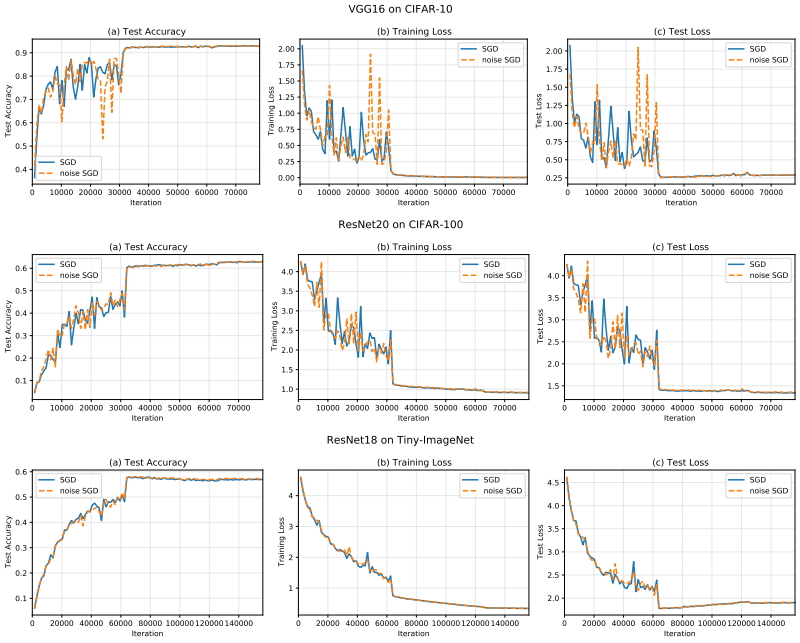



Figure 1: Test Accuracy, Training Loss, and Test Loss for SGD under different tasks, with and without heavy-tailed noise.

density function is $p(s) = \frac{3\sqrt{3}}{4\pi(1+|s|^3)}$. The symmetry of the distribution indicates the zero-mean expectation of the variable $\xi$, i.e., $\mathbb{E}\xi = 0$. On the other hand, direct computing shows that its variance is unbounded, that is, $\mathbb{E}|\xi|^2 = \frac{3\sqrt{3}}{2\pi}\int_0^{+\infty}\frac{s^2}{1+s^3}\,ds = +\infty$, whereas, for any $1 < \alpha < 2$, $\mathbb{E}|\xi|^\alpha = \frac{3\sqrt{3}}{2\pi}\int_0^{+\infty}\frac{s^\alpha}{1+s^3}\,ds < +\infty$. Hence, the noise $\xi$ is heavy-tailed. In this training, we conduct the SGD with the heavy-tailed noise $\xi$ as $\boldsymbol{w}^{t+1} = \boldsymbol{w}^t - \gamma(\boldsymbol{g}^t + 10^{-4}\xi \cdot \mathbf{1})$, where $\mathbf{1}$ denotes the vector whose elements are all 1s.

We compare the performance of standard nonconvex SGD with and without the added noise $\xi$, as shown in Figure 1. Our numerical experiments reveal that, while standard nonconvex SGD under heavy-tailed noise exhibits significantly higher instability compared to noise-free conditions, particularly in the early stages of training, it ultimately achieves comparable performance to its noise-free counterpart after a sufficient number of iterations. The observation suggests that convergence with probability is completely possible, supporting the distinction discussed above that the failure of second-moment guarantees does not by itself rule out convergence in weaker senses.

### 1.2 Contributions

This paper re-examines standard nonconvex SGD under heavy-tailed stochastic gradient noise. Our main contributions are as follows.

- **High-probability convergence under standard smoothness.** We prove that standard nonconvex SGD, without clipping or normalization, converges in high probability under standard $L$-smoothness and only a bounded $\alpha$-th moment assumption. This distinguishes the failure of second-moment bounds from the absence of probabilistic convergence.
- **A matching limitation of standard nonconvex SGD.** We establish an algorithm-specific lower bound showing that the $(T, \delta)$-dependence of our upper bound is essentially unavoidable for standard SGD. This identifies an intrinsic limitation of the standard update rather than merely a weakness of the proof technique.
- **The price of remaining unclipped.** The high-probability guarantee has degraded dependence on both the failure probability and the iteration number. Hence, the two main results presented in this paper show that clipping or normalization is not required for convergence

itself, but standard SGD pays an unavoidable price in its dependence on the iteration horizon and confidence level.

**Notation.** We use $\mathbb{E}[\cdot]$ to denote expectation over the underlying probability space. The Euclidean norm of a vector is denoted by $\|\cdot\|$. For two positive sequences $\{a_t\}$ and $\{b_t\}$, we write $a_t = \mathcal{O}(b_t)$ if there exists a constant $C > 0$ such that $a_t \leq Cb_t$ for all sufficiently large $t$. We write $a_t = \Theta(b_t)$ if there exist constants $C_1, C_2 > 0$ such that $C_2 b_t \leq a_t \leq C_1 b_t$ for all sufficiently large $t$. We write $a_t \asymp b_t$ when $a_t = \Theta(b_t)$. We denote $\Delta := f(\boldsymbol{w}^0) - \min f + \frac{1}{2}$.

## 2 Related Works

Recent studies have empirically shown that the gradient noise in SGD for DNN training exhibits distinctly non-Gaussian characteristics, often manifesting as heavy-tailed distributions or even extreme values (Sutskever et al., 2013; Şimşekli et al., 2019; Zhang et al., 2020). This deviation from the typical Gaussian noise assumption has significant implications for understanding and optimizing the training process. Gürbüzbalaban et al. (2021) provide a theoretical explanation for this heavy-tailed behavior observed in SGD during training, shedding light on the underlying mechanisms. The subsequent work Gürbüzbalaban et al. (2024) extend this analysis to decentralized SGD, demonstrating that heavy-tailed noise is also prevalent in this distributed setting and underscoring the need to address this phenomenon across various stochastic optimization scenarios. Barsbey et al. (2021) formally prove that SGD converges to a heavy-tailed stationary distribution when the ratio of learning rate to batch-size is large, providing a crucial link between these training parameters and the resulting noise distribution. Furthermore, the impact of heavy-tailed noise on generalization performance has been actively investigated (Simsekli et al., 2020; Raj et al., 2023b;a), highlighting the practical consequences of this phenomenon for model performance on unseen data. Hodgkinson & Mahoney (2021) model stochastic optimization as discrete random recurrences and reveal that multiplicative noise, in particular, contributes to heavy-tailed parameter distributions. This insight provides a more nuanced understanding of the sources of heavy-tailed noise in optimization. Wang et al. (2021) establish convergence guarantees for SGD under state-dependent and heavy-tailed noise for a specific class of strongly convex objectives, demonstrating that convergence can still be achieved despite these challenging noise conditions. More recently, Wan et al. (2024) propose a simple yet effective modification to SGD that guarantees compressible outputs without requiring any non-trivial assumptions. This contribution offers a practical approach to managing the complexity of learned models in the presence of heavy-tailed noise.

A growing body of research explores the performance of SGD with gradient clipping under heavy-tailed noise. The authors of (Zhang et al., 2020) establish the convergence of SGD with gradient clipping in such settings (Assumption 2). Significantly, they demonstrate a convergence rate that matches the lower bound for nonconvex stochastic optimization under heavy-tailed noise, a bound they also prove within the same work for cases where the global gradient is Lipschitz continuous. Wang et al. (2022) further demonstrates the advantages of gradient clipping, showing that it can effectively eliminate sharp local minima from the SGD training trajectory. In (Menon et al., 2020), the authors show that basic gradient clipping is not robust to label noise, but a simple variant is, effectively creating a robust modified cross-entropy loss. Qian et al. (2021) study the convergence of gradient clipping for incremental methods. Gorbunov et al. (2020) introduce an accelerated stochastic algorithm designed specifically for heavy-tailed noise, offering a novel approach to optimization in these challenging conditions. Koloskova et al. (2023) provide tight convergence guarantees with precise dependence on arbitrary clipping thresholds c for both deterministic and stochastic gradients. Expanding the scope of application, Sadiev et al. (2023) investigate the use of stochastic algorithms for solving variational inequalities under heavy-tailed noise, extending these methods beyond traditional stochastic optimization problems. Liu et al. (2022) investigate a relaxed-smoothness property of the loss landscape, previously validated for LSTMs, to develop a communication-efficient gradient clipping algorithm. Cutkosky & Mehta (2021) prove a high-probability convergence rate when combining gradient normalization and clipping. This result is further refined by (Liu et al., 2023), who eliminate the logarithmic factor in the convergence rate. Puchkin et al. (2024) achieve enhanced convergence results by assuming a specific structure for the heavy-tailed noise. Meanwhile, Vural et al. (2022) analyze the convergence rate of the stochastic mirror descent algorithm in stochastic convex optimization with heavy-tailed noise, providing lower bounds that highlight the algorithm's efficiency without requiring gradient clipping. Nguyen et al. (2023) build upon this work, offer-

ing an improved convergence rate for SGD with clipping under heavy-tailed noise specifically for convex cases. Sun et al. (2025) delve into the underlying mechanisms of combining gradient normalization and gradient clipping in SGD under heavy-tailed noise, providing valuable insights into the interplay of these techniques. In (Hübler et al., 2025), the authors demonstrate that applying gradient normalization with mini-batches also ensure the convergence of SGD under heavy-tailed noise, while attaining the optimal lower bound on sample complexity. They further establish high-probability guarantees for the convergence rate. Meanwhile, the authors of (Liu & Zhou, 2024) extend the analysis by incorporating momentum into mini-batch gradient normalization for SGD under heavy-tailed noise and deriving convergence rates in expectation. He & Lu (2025) study accelerated stochastic first-order methods in the convex setting under heavy-tailed noise.

**Vanilla nonconvex SGD under heavy-tailed noise.** The works most closely related to ours are (Fatkhullin et al., 2025; Yamada et al., 2026). Fatkhullin et al. (2025) established the convergence of standard nonconvex SGD in under Hölder smoothness and bounded $p$-th moments of the stochastic gradients. They proved the convergence of SGD under assumptions of bounded central moments and standard smoothness when mini-batch size is sufficiently large. Yamada et al. (2026) established convergence results for standard SGD and SGDM where the gradients satisfy Hölder continuity and the condition $\nu + 1 \leq p$, which does not cover the case of standard $L$-smoothness. In contrast, our results demonstrate that single-sample standard non-convex SGD converges with high probability under standard $L$-smoothness and bounded $\alpha$-th central moments, and we further provide matching algorithm-specific lower bounds.

# 3 High-Probability Reliability of Standard Nonconvex SGD

## 3.1 Problem Formulation

Nonconvex smooth optimization is a crucial aspect of machine learning research, as various tasks can be formulated as minimizing a nonconvex smooth function as

$$\min_{\boldsymbol{w}\in\mathbb{R}^d} f(\boldsymbol{w}) : \mathbb{R}^d \mapsto \mathbb{R}.$$

SGD is a widely used and effective algorithm for these problems, which uses unbiased stochastic gradients to update parameters rather than the full gradient due to memory limitations and large dataset sizes. Standard SGD generates $\boldsymbol{w}^{t+1} = \boldsymbol{w}^t - \gamma_t \boldsymbol{g}^t$, where $\boldsymbol{g}^t$ is a stochastic gradient evaluated at $\boldsymbol{w}^t$ and $\gamma_t$ is the learning rate used in the $t$-th iteration.

**Assumption 1** *Let $\{\mathcal{F}_t\}_{t\geq 0}$ be the filtration defined by $\mathcal{F}_t := \sigma(\boldsymbol{w}^0, \boldsymbol{g}^0, \boldsymbol{g}^1, \ldots, \boldsymbol{g}^{t-1})$, so that the iterate $\boldsymbol{w}^t$ is $\mathcal{F}_t$-measurable. There exist constants $\sigma > 0$ and $\alpha \in (1, 2]$ such that, for every $t \geq 0$, the stochastic gradient $\boldsymbol{g}^t$ satisfies*

$$\mathbb{E}[\boldsymbol{g}^t \mid \mathcal{F}_t] = \nabla f(\boldsymbol{w}^t), \quad \mathbb{E}\left[\|\boldsymbol{g}^t - \nabla f(\boldsymbol{w}^t)\|^\alpha \middle| \mathcal{F}_t\right] \leq \sigma^\alpha.$$

When $\alpha = 2$, Assumption 1 simplifies to the standard bounded variance assumption, which is a common condition in many optimization algorithms. However, when $\alpha$ is strictly smaller than 2, the noise is considered heavy-tailed, and the situation becomes more complicated. In this case, Zhang et al. (2020) provide a simple example that demonstrates the failure of standard nonconvex SGD to produce results with finite variance under any learning rate setting. This limitation has sparked a line of research focused on exploring how gradient clipping can ensure the convergence of SGD under heavy-tailed noise.

The convergence analysis presented in this paper is based on the commonly employed smoothness assumption in stochastic optimization, as outlined below.

**Assumption 2** *There exists a constant $L > 0$, $\forall\, \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$ such that*

$$\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\| \leq L\|\boldsymbol{x} - \boldsymbol{y}\|. \tag{1}$$

*The function $f$ is lower bounded and minimum value can be attained, i.e.,* $\min f > -\infty$.

### 3.2 High-Probability Reliability Bound

The convergence result of standard nonconvex SGD is expressed in probabilistic form as follows.

**Theorem 1 (High-probability convergence of standard nonconvex SGD)** *Let $(\boldsymbol{w}^t)_{t\geq 0}$ be generated by SGD, and Assumptions 1 and 2 hold. Given any $0<\delta<1$, define $\Delta := f(\boldsymbol{w}^0)-\min f+\frac{1}{2}$, $u := 4\Delta$, $G := \sqrt{\frac{u}{L}}$, and $C_\alpha := 2^{\frac{7}{2}+\frac{2}{\alpha}}$. Let $(\gamma_t)_{t=0}^{T-1}$ be any deterministic learning-rate schedule satisfying $0<\gamma_t\leq\frac{1}{L}$, $\left(\sum_{t=0}^{T-1}\gamma_t^\alpha\right)^{1/\alpha}\leq\frac{G\delta^{1/\alpha}}{C_\alpha\sigma}$. Then, with probability at least $1-\delta$,*

$$\frac{\sum_{t=0}^{T-1}\gamma_t\|\nabla f(\boldsymbol{w}^t)\|^2}{\sum_{t=0}^{T-1}\gamma_t}\leq\frac{4\Delta}{\sum_{t=0}^{T-1}\gamma_t}. \tag{2}$$

*Moreover, by choosing $\gamma_t\equiv\gamma=\min\left\{\frac{1}{L},\frac{G\,\delta^{1/\alpha}}{C_\alpha\sigma T^{1/\alpha}}\right\}$, we obtain, with probability at least $1-\delta$,*

$$\frac{1}{T}\sum_{t<T}\|\nabla f(\boldsymbol{w}^t)\|^2\leq\frac{4L\Delta}{T}+\frac{2C_\alpha\sigma\sqrt{L\Delta}}{\delta^{1/\alpha}T^{(\alpha-1)/\alpha}}=\mathcal{O}\left(\frac{L\Delta}{T}+\frac{\sigma\sqrt{L\Delta}}{\delta^{1/\alpha}T^{(\alpha-1)/\alpha}}\right). \tag{3}$$

Theorem 1 is indeed also a reliability statement, not just a convergence result. Ignoring problem-dependent constants, the confidence-calibrated learning rate in Theorem 1 is $\gamma\asymp\delta^{1/\alpha}T^{-1/\alpha}$. Hence, for a fixed $T$, reducing the probability of target failure $\delta$ requires a smaller learning rate, and the resulting stationarity error increases with $\delta^{-1/\alpha}$. Therefore, the confidence level directly affects the performance of standard nonconvex SGD.

When $\alpha=2$, Assumption 1 reduces to the standard bounded variance condition, in which Theorem 1 then gives $\frac{1}{T}\sum_{t<T}\|\nabla f(\boldsymbol{w}^t)\|^2=\mathcal{O}\left(\frac{1}{\sqrt{\delta T}}\right)$ with at least $1-\delta$ probability; the corresponding learning rate in this case is then $\gamma=\Theta\left(\sqrt{\frac{\delta}{T}}\right)$. This dependence is also consistent with the response-dependent lower-bound mechanism for SGD established by Jin et al. (2026) (a detailed explanation is presented in Appendix F.1). Such a bound is sharper in its confidence dependence than the $\mathcal{O}(1/(\delta\sqrt{T}))$ bound obtained by applying Markov's inequality directly to the standard expectation guarantee. The improvement comes from controlling the stochastic trajectory before applying a terminal Markov inequality.

The rate derived for standard nonconvex SGD under heavy-tailed noise, with respect to iteration time $T$, in our analysis does not align with the lower bound of first-order under heavy-tailed noise presented in (Zhang et al., 2020), which can be achieved with gradient clipping. These findings highlight the crucial role of clipping in effectively handling heavy-tailed noise. On the other hand, we emphasize that our rate is derived without the use of gradient clipping. Vural et al. (2022) consider stochastic mirror descent, while our analysis is for the standard nonconvex SGD update. Thus their bound is not a direct counterpart of ours. The point of comparison is only the dependence on $T$: the rate we obtain for nonconvex SGD has the same $T$-order as their convex SMD result. This makes the loss due to heavy-tailed noise appear less tied to mirror descent or to convexity, and more related to the noise model itself. The same $T$-dependence is also in line with convex results for clipping-based methods (Sadiev et al., 2023; Nguyen et al., 2023). In fact, our result is also comparable to the rate derived for convex clipping methods (Sadiev et al., 2023; Nguyen et al., 2023). In the subsequent section, we also demonstrate that the rate concerning $T$ is tight.

The dependence on the confidence parameter $\delta$ is also tight for standard nonconvex SGD. In Section 4, we show that, there exists a smooth stochastic optimization instance for which any constant-learning-rate SGD method incurs $\Omega\left(\frac{1}{\delta^{1/\alpha}T^{(\alpha-1)/\alpha}}\right)$ with probability larger than $\delta$. Therefore, the factor $\delta^{-1/\alpha}$ in Theorem 1 is not an artifact of the techniques used in the proof.

## 4 Reliability Lower Bounds for Deterministic Learning Rate Schedules

This section presents an algorithm-specific lower bound for standard nonconvex SGD under the heavy-tailed noise. Unlike oracle-complexity lower bounds, characterizing the best possible per-

formance over a broad class of stochastic first-order algorithms, our goal is to demonstrate that the convergence rate of nonconvex standard SGD cannot be improved due to the scheme of the algorithm rather than techniques. Mathematically, we construct a smooth lower-bounded nonconvex function together with an unbiased stochastic oracle satisfying the bounded $\alpha$-th moment assumption. Under these settings, we show that standard SGD incurs a convergence error of order $\frac{1}{\delta^{1/\alpha}T^{(\alpha-1)/\alpha}}$ with probability larger than $\delta$. Therefore, the dependence on both the confidence parameter $\delta$ and the iteration number $T$ in Theorem 1 is unimprovable.

**Algorithm-specific lower bound rather than oracle lower bound.** We once stress that what we studied is the algorithm-specific lower bound rather than oracle ones for heavy-tailed noise (Zhang et al., 2020; Vural et al., 2022; Liu et al., 2023; Fradin et al., 2026). Our objective here is therefore different. We fix the algorithm to be nonconvex standard SGD and ask whether its dependence on $(T, \delta)$ in Theorem 1 can be improved by a sharper analysis or by retuning an arbitrary deterministic learning rate. We now prove that the dependence on $(T, \delta)$ cannot be improved for nonconvex standard SGD under heavy-tailed noise.

**Theorem 2 (Lower bound for arbitrary deterministic learning rates)** *Let $\alpha \in (1, 2]$, $T \geq 8$, and $\delta \in (0, 1/16)$. For every deterministic stepsize sequence $(\gamma_0, \gamma_1, \ldots, \gamma_{T-1}) \in (0, 1]^T$, there exist a 1-smooth lower-bounded nonconvex objective $f$, an initialization satisfying $f(\boldsymbol{w}^0) - \min f \leq \frac{1}{2}$, and an unbiased stochastic-gradient oracle satisfying the bounded conditional $\alpha$-th moment assumption such that*

$$\mathbb{P}\left(\frac{1}{T}\sum_{t<T}\|\nabla f(\boldsymbol{w}^t)\|^2 \geq \; c_\alpha \min\left\{1, \frac{1}{\delta^{1/\alpha}T^{(\alpha-1)/\alpha}}\right\}\right) \geq \delta,$$

*where $c_\alpha$ is a constant that is only dependent of $\alpha$.*

In the nontrivial regime where $\delta^{-1/\alpha}T^{-(\alpha-1)/\alpha} \leq 1$, the lower bound is $\Theta\left(\frac{1}{\delta^{1/\alpha}T^{(\alpha-1)/\alpha}}\right)$. Theorem 2 then present a result that matches the stochastic term in the upper bound of Theorem 1 in both $T$ and $\delta$. Hence, neither the factor $\delta^{-1/\alpha}$ nor the iteration dependence $T^{-(\alpha-1)/\alpha}$ can be improved in general without changing the standard SGD update when using the deterministic learning rates.

We want to stress that Theorem 2 only applies to the deterministic learning rate schedule, and therefore does not rule out improvements obtained by learning rates that adapt to the stochastic gradient. This distinction is important. For example, NSGD can be written as $\boldsymbol{w}^{t+1} = \boldsymbol{w}^t - \gamma_t \boldsymbol{g}^t$ with $\gamma_t := \frac{\eta}{\|\boldsymbol{g}^t\|}$ and $\eta > 0$. Hence, NSGD may be viewed as SGD equipped with a learning rate schedule dependent on stochastic gradient. Such a case is outside the class covered by Theorem 2 because the learning rate $\gamma_t$ depends on the stochastic gradient and is thus not deterministic. Indeed, existing results have shown that NSGD can achieve the lower bound of first-order algorithms under heavy-tailed noise (Hübler et al., 2025; Sun et al., 2025). Therefore, our lower bound is not showing that any learning rate adaptation can never overcome the reliability limitation of standard nonconvex SGD; rather, it shows that this limitation cannot be removed by any deterministic retuning.

**The Reliability Frontier of Standard Nonconvex SGD.** The upper and lower bounds above together yield a tight characterization of the best high-probability reliability attainable by standard nonconvex SGD under the heavy-tailed noise. To make this statement formal, we consider the problem class $\mathfrak{P}_\alpha$ consisting of all objective-oracle pairs satisfying Assumptions 1 and 2 with $L = 1, \sigma \leq 1, f(\boldsymbol{w}^0) - \min f \leq \frac{1}{2}$. Given a deterministic learning-rate schedule $\boldsymbol{\gamma} := (\gamma_0, \ldots, \gamma_{T-1}) \in (0, 1]^T$, define $\mathcal{E}_T(\boldsymbol{\gamma}) := \frac{1}{T}\sum_{t=0}^{T-1}\|\nabla f(\boldsymbol{w}^t)\|^2$. We define the minimax reliability frontier of standard SGD by

$$\mathfrak{R}_\alpha(T, \delta) := \inf_{\boldsymbol{\gamma} \in (0,1]^T} \inf\left\{r > 0 : \sup_{(f,\mathcal{O}) \in \mathfrak{P}_\alpha} \mathbb{P}\left(\mathcal{E}_T(\boldsymbol{\gamma}) > r\right) \leq \delta\right\}. \tag{4}$$

Thus, $\mathfrak{R}_\alpha(T, \delta)$ is the smallest stationarity level that standard nonconvex SGD can guarantee uniformly over the problem class with failure probability at most $\delta$.

**Corollary 1 (Reliability frontier of standard SGD)** *Let $\alpha \in (1, 2]$, $T \geq 8$, and $\delta \in (0, 1/32)$. In the nontrivial regime $\frac{1}{\delta^{1/\alpha}T^{(\alpha-1)/\alpha}} \leq 1$, the reliability frontier of standard SGD satisfies*

$$\mathfrak{R}_\alpha(T, \delta) = \Theta\left(\frac{1}{\delta^{1/\alpha}T^{(\alpha-1)/\alpha}}\right).$$

## 5 Numerical Illustration of $T$ and $\delta$ Scaling

We provide the numerical illustration of the dependence of the high-probability reliability of standard SGD on the iteration horizon $T$ and the target failure probability $\delta$. We consider the 1-smooth nonconvex objective as $f(x, y) = \frac{1}{2}x^2 + 1 - \cos(y)$, initialized at $(x^0, y^0) = (0, 0)$. At every iteration $t = 0, \dots, T-1$, we consider the stochastic gradient as $\boldsymbol{g}^t = \nabla f(\boldsymbol{w}^t) + \xi^t \boldsymbol{e}_1$, where $\boldsymbol{e}_1$ is the vector whose first coordinate is 1 and second one is zero. To generate controlled heavy-tailed stochastic gradients, we first consider the variable Here $Z_t$ has density $q_\alpha(z) = c_\alpha \frac{e^{-\alpha z}}{z^2}$ as $z \geq 1$, and $c_\alpha^{-1} = \int_1^\infty \frac{e^{-\alpha z}}{z^2}\, dz$. Then, we let $\xi^t = p^{-1/\alpha} B_t H_t$ with $p = \frac{4\delta}{T}$, where $B_t \sim \text{Bernoulli}(p)$ and $H_t = S_t \frac{\exp(Z_t)}{c_\alpha^{1/\alpha}}$ with $\mathbb{P}(S_t = \pm 1) = \frac{1}{2}$. By this construction, we have $\mathbb{E}[\xi^t] = 0$ and $\mathbb{E}|\xi^t|^\alpha = 1$, whereas $\mathbb{E}|\xi^t|^q = +\infty$ for every $q > \alpha$. In particular, for all $\alpha < 2$ considered below, the stochastic gradient noise has infinite variance.

We measure stationarity by $R_T := \frac{1}{T}\sum_{t<T} \|\nabla f(\boldsymbol{w}^t)\|^2$ and use the confidence-calibrated learning rate $\gamma = 4\delta^{1/\alpha} T^{-1/\alpha}$. For each configuration, we perform $N$ independent runs and report the empirical $(1-\delta)$-quantile of $R_T$. Specifically, if $R_{T,(1)} \leq \cdots \leq R_{T,(N)}$ are the ordered values obtained from the $N$ runs, we define

$$Q_{1-\delta}(R_T) := R_{T,(\lceil (1-\delta)N \rceil)}.$$

Thus, approximately a fraction $1-\delta$ of the runs attain a stationarity error no larger than $Q_{1-\delta}(R_T)$, making this quantity the empirical counterpart of our high-probability guarantee.

Figure 2 reports the results for $\alpha \in \{1.5, 1.7, 1.9\}$. We can find that heavier-tailed noise simultaneously results in slower improvement with the iteration horizon and a stronger penalty for requiring a smaller failure probability, consistently with the $(T, \delta)$ dependence characterized by our upper and lower bounds.

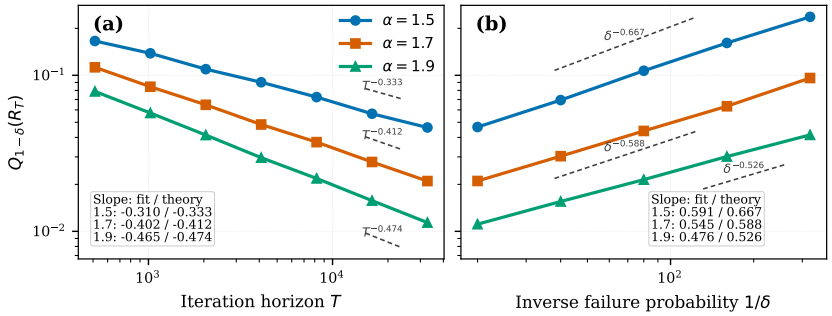



Figure 2: **Empirical scaling of the reliability limit.** We report the empirical $(1-\delta)$-quantile of $R_T = T^{-1}\sum_{t<T} \|\nabla f(\boldsymbol{w}^t)\|^2$. **(a)** Scaling with the iteration horizon $T$ at $\delta = 0.05$. **(b)** Scaling with the inverse failure probability $1/\delta$ at $T = 2^{15}$. Solid curves show empirical quantiles and dashed lines indicate the theoretical slopes $-(\alpha-1)/\alpha$ and $1/\alpha$, respectively.

## 6 Conclusion

Our results establish that standard nonconvex SGD converges in high probability under standard $L$-smoothness and bounded $\alpha$-th central moments of the stochastic-gradient noise, without gradient clipping or normalization. The resulting stationarity convergence rate is as fast as $\mathcal{O}\big(\delta^{-1/\alpha} T^{-(\alpha-1)/\alpha}\big)$. We further prove that this dependence on $(T, \delta)$ is unavoidable for standard nonconvex SGD with arbitrary deterministic learning-rate schedules. Therefore, the slower heavy-tailed rate is an intrinsic limitation of the standard nonconvex SGD update rather than the analysis used in this paper. Our results also clarify the role of robustification: clipping or normalization is unnecessary for convergence itself, but can overcome limitations that cannot be removed by learning rate tuning alone.

### AI USE STATEMENT

In this work, we used generative AI tools to assist with language editing and checking, primarily to improve grammar, wording, clarity, readability, and checking proof arguments. The core theoretical ideas, mathematical analysis, assumptions, and final proofs were developed and independently verified by the authors. We did not use generative AI tools to develop the theoretical models or conceptual framework, propose research hypotheses. All AI-assisted text was manually reviewed and revised by the authors. The authors take full responsibility for the final content of this work, including any text or other material produced with the aid of generative AI.

### ETHICS STATEMENT

This work is theoretical in nature and studies stochastic optimization algorithms and their convergence properties under heavy-tailed gradient noise. It does not involve human subjects, personal or sensitive data, data collection or release, or experiments requiring institutional ethical approval. To the best of our knowledge, this work does not raise additional ethical concerns specific to the present study.

### REPRODUCIBILITY STATEMENT

The assumptions, algorithms, and settings have been specified in the main text. Complete proofs of the theoretical results are provided in the appendix to facilitate independent verification of our results.

## REFERENCES

Melih Barsbey, Milad Sefidgaran, Murat A Erdogdu, Gael Richard, and Umut Simsekli. Heavy tails in sgd and compressibility of overparametrized neural networks. *Advances in neural information processing systems*, 34:29364–29378, 2021.

Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *Siam Review*, 60(2):223–311, 2018.

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

Ilyas Fatkhullin, Florian Hübler, and Guanghui Lan. Can sgd handle heavy-tailed noise? In *OPT 2025: Optimization for Machine Learning*, 2025.

Adrien Fradin, Abdurakhmon Sadiev, Laurent Condat, and Peter Richtárik. Tight lower bounds and optimal algorithms for stochastic nonconvex optimization with heavy-tailed noise. In *The 29th International Conference on Artificial Intelligence and Statistics*, 2026. URL https://openreview.net/forum?id=7Oyy8esOwq.

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

Mert Gürbüzbalaban, Umut Simsekli, and Lingjiong Zhu. The heavy-tail phenomenon in sgd. In *International Conference on Machine Learning*, pp. 3964–3975. PMLR, 2021.

Mert Gürbüzbalaban, Yuanhan Hu, Umut Şimşekli, Kun Yuan, and Lingjiong Zhu. Heavy-tail phenomenon in decentralized sgd. *IISE Transactions*, (just-accepted):1–23, 2024.

Chuan He and Zhaosong Lu. Accelerated stochastic first-order method for convex optimization under heavy-tailed noise. *arXiv preprint arXiv:2510.11676*, 2025.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Liam Hodgkinson and Michael Mahoney. Multiplicative noise and heavy tails in stochastic optimization. In *International Conference on Machine Learning*, pp. 4262–4274. PMLR, 2021.

Florian Hübler, Ilyas Fatkhullin, and Niao He. From gradient clipping to normalization for heavy tailed SGD. In *The 28th International Conference on Artificial Intelligence and Statistics*, 2025. URL https://openreview.net/forum?id=nuG0FYyWRT.

Ruinan Jin, Yingbin Liang, and Shaofeng Zou. Why adam can beat sgd: Second-moment normalization yields sharper tails. *arXiv preprint arXiv:2603.03099*, 2026.

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, 2024.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

Yann Le, Xuan Yang, et al. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

Shaojie Li and Yong Liu. High probability guarantees for nonconvex stochastic gradient descent with heavy tails. In *International Conference on Machine Learning*, pp. 12931–12963. PMLR, 2022.

Mingrui Liu, Zhenxun Zhuang, Yunwen Lei, and Chunyang Liao. A communication-efficient distributed gradient clipping algorithm for training deep neural networks. *Advances in Neural Information Processing Systems*, 35:26204–26217, 2022.

Zijian Liu and Zhengyuan Zhou. Nonconvex stochastic optimization under heavy-tailed noises: Optimal convergence without gradient clipping. *arXiv preprint arXiv:2412.19529*, 2024.

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.

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

Aditya Krishna Menon, Ankit Singh Rawat, Sashank J Reddi, and Sanjiv Kumar. Can gradient clipping mitigate label noise? In *International Conference on Learning Representations*, 2020.

Ta Duy Nguyen, Thien H Nguyen, Alina Ene, and Huy Nguyen. Improved convergence in high probability of clipped gradient methods with heavy tailed noise. *Advances in Neural Information Processing Systems*, 36:24191–24222, 2023.

Iosif Pinelis. Best possible bounds of the von bahr–esseen type. *Annals of Functional Analysis*, 6 (4):1–29, 2015.

Iosif Pinelis. Multidimensional probability inequalities via spherical symmetry. *arXiv preprint arXiv:2210.04391*, 2022.

Nikita Puchkin, Eduard Gorbunov, Nickolay Kutuzov, and Alexander Gasnikov. Breaking the heavy-tailed noise barrier in stochastic optimization problems. In *International Conference on Artificial Intelligence and Statistics*, pp. 856–864. PMLR, 2024.

Jiang Qian, Yuren Wu, Bojin Zhuang, Shaojun Wang, and Jing Xiao. Understanding gradient clipping in incremental gradient methods. In *International Conference on Artificial Intelligence and Statistics*, pp. 1504–1512. PMLR, 2021.

Anant Raj, Melih Barsbey, Mert Gürbüzbalaban, Lingjiong Zhu, Umut Şim, et al. Algorithmic stability of heavy-tailed stochastic gradient descent on least squares. In *International Conference on Algorithmic Learning Theory*, pp. 1292–1342. PMLR, 2023a.

Anant Raj, Lingjiong Zhu, Mert Gürbüzbalaban, and Umut Simsekli. Algorithmic stability of heavy-tailed sgd with general loss functions. In *International Conference on Machine Learning*, pp. 28578–28597. PMLR, 2023b.

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

Abdurakhmon Sadiev, Marina Danilova, Eduard Gorbunov, Samuel Horváth, Gauthier Gidel, Pavel Dvurechensky, Alexander Gasnikov, and Peter Richtárik. High-probability bounds for stochastic optimization and variational inequalities: the case of unbounded variance. In *International Conference on Machine Learning*, pp. 29563–29648. PMLR, 2023.

Kevin Scaman and Cédric Malherbe. Robustness analysis of non-convex stochastic gradient descent using biased expectations. *Advances in Neural Information Processing Systems*, 33:16377–16387, 2020.

K Simonyan and A Zisserman. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations (ICLR 2015)*. Computational and Biological Learning Society, 2015.

Umut Şimşekli, Mert Gürbüzbalaban, Thanh Huy Nguyen, Gaël Richard, and Levent Sagun. On the heavy-tailed theory of stochastic gradient descent for deep neural networks. *arXiv preprint arXiv:1912.00018*, 2019.

Umut Simsekli, Levent Sagun, and Mert Gürbüzbalaban. A tail-index analysis of stochastic gradient noise in deep neural networks. In *International Conference on Machine Learning*, pp. 5827–5837. PMLR, 2019.

Umut Simsekli, Ozan Sener, George Deligiannidis, and Murat A Erdogdu. Hausdorff dimension, heavy tails, and generalization in neural networks. *Advances in Neural Information Processing Systems*, 33:5138–5151, 2020.

Tao Sun, Xinwang Liu, and Kun Yuan. Revisiting gradient normalization and clipping for nonconvex sgd under heavy-tailed noise: Necessity, sufficiency, and acceleration. *Journal of Machine Learning Research*, 26(237):1–42, 2025.

Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pp. 1139–1147. PMLR, 2013.

Nuri Mert Vural, Lu Yu, Krishna Balasubramanian, Stanislav Volgushev, and Murat A Erdogdu. Mirror descent strikes again: Optimal stochastic convex optimization under infinite noise variance. In *Conference on Learning Theory*, pp. 65–102. PMLR, 2022.

Yijun Wan, Melih Barsbey, Abdellatif Zaidi, and Umut Simsekli. Implicit compressibility of overparametrized neural networks trained with heavy-tailed sgd. In *Forty-first International Conference on Machine Learning*, 2024.

Hongjian Wang, Mert Gürbüzbalaban, Lingjiong Zhu, Umut Simsekli, and Murat A Erdogdu. Convergence rates of stochastic gradient descent under infinite noise variance. *Advances in Neural Information Processing Systems*, 34:18866–18877, 2021.

Xingyu Wang, Sewoong Oh, and Chang Han Rhee. Eliminating sharp minima from sgd with truncated heavy-tailed noise. In *10th International Conference on Learning Representations, ICLR 2022*, 2022.

Ryusei Yamada, Naoki Sato, and Hideaki Iiduka. Vanilla SGD with momentum survives heavy-tailed noise: Convergence analysis without gradient clipping or normalization. In Emilija Perković and Daniel Malinsky (eds.), *Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence*, volume 337 of *Proceedings of Machine Learning Research*, pp. 7670–7700. PMLR, 17–21 Aug 2026.

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, 2020.

Tianxi Zhu, Yi Xu, Qi Wang, and Xiangyang Ji. Stochastic gradient methods under heavy-tailed noises in weakly convex optimization. In *Forty-third International Conference on Machine Learning*, 2026.

# Appendices for

## *Reliability Limits of Standard Nonconvex SGD in AI Training under Heavy-Tailed Gradient Noise*

## A Distributed Nonconvex Standard SGD with Heavy-Tailed Noise

In this part, we consider the distributed training setting with heavy-tailed noise in the workers and show that the high-probability reliability of is governed by an explicit system-level effective noise scale. Consider the following distributed objective

$$f(\boldsymbol{w}) := \sum_{i=1}^{m} p_i f_i(\boldsymbol{w}), \qquad p_i \geq 0, \quad \sum_{i=1}^{m} p_i = 1, \tag{5}$$

where $m$ is the number of workers and $p_i$ is the aggregation weight of worker $i$. At iteration $t$, each worker computes a local mini-batch stochastic gradient $\boldsymbol{g}_i^t := \frac{1}{b_i}\sum_{j=1}^{b_i} \boldsymbol{g}_{i,j}^t$, where $b_i$ is the local batch size of worker $i$. The server aggregates the local gradients as $\boldsymbol{g}^t := \sum_{i=1}^{m} p_i \boldsymbol{g}_i^t$, and updates $\boldsymbol{w}^{t+1} = \boldsymbol{w}^t - \gamma \boldsymbol{g}^t$. Standard mini-batch SGD is recovered as the degenerate single-worker case with $m = 1$, $p_1 = 1$, and $b_1 = B$.

We use the following moment inequality for sums of independent random vectors.

**Lemma 1** *Pinelis (2022); Kornilov et al. (2024) Let $X_1, \ldots, X_n$ be independent mean-zero random vectors in $\mathbb{R}^d$. For any $\alpha \in (1, 2]$ and any scalars $a_1, \ldots, a_n$, if $\mathbb{E}\|X_i\|^\alpha < +\infty$, then*

$$\mathbb{E}\left\|\sum_{i=1}^{n} a_i X_i\right\|^\alpha \leq 2\sum_{i=1}^{n} |a_i|^\alpha \mathbb{E}\|X_i\|^\alpha.$$

We impose the following worker-wise heavy-tailed noise assumption.

**Assumption 3** *Let $\{\mathcal{F}_t\}_{t\geq 0}$ be the filtration generated by all randomness before iteration $t$. For each worker $i \in \{1, \ldots, m\}$ and each local sample $j \in \{1, \ldots, b_i\}$, the stochastic gradient $\boldsymbol{g}_{i,j}^t$ satisfies*

$$\mathbb{E}\left[\boldsymbol{g}_{i,j}^t \mid \mathcal{F}_t\right] = \nabla f_i(\boldsymbol{w}^t), \ \mathbb{E}\left[\left\|\boldsymbol{g}_{i,j}^t - \nabla f_i(\boldsymbol{w}^t)\right\|^\alpha \mid \mathcal{F}_t\right] \leq \sigma_i^\alpha$$

*for constants $\sigma_i > 0$. Conditioned on $\mathcal{F}_t$, the local stochastic gradient noises are independent across $i$ and $j$.*

The following lemma shows that distributed aggregation preserves the bounded $\alpha$-th moment structure, but with an effective noise scale depending on the worker weights, local batch sizes, and worker-wise tail magnitudes.

**Lemma 2 (Effective heavy-tailed noise scale of distributed aggregation)** *Suppose Assumption 3 holds. Define*

$$\sigma_{\mathrm{dist}} := \left[4\sum_{i=1}^{m} p_i^\alpha \frac{\sigma_i^\alpha}{b_i^{\alpha-1}}\right]^{\frac{1}{\alpha}}. \tag{6}$$

*Then the aggregated stochastic gradient $\boldsymbol{g}^t$ satisfies*

$$\mathbb{E}\left[\boldsymbol{g}^t \mid \mathcal{F}_t\right] = \nabla f(\boldsymbol{w}^t), \ \mathbb{E}\left[\left\|\boldsymbol{g}^t - \nabla f(\boldsymbol{w}^t)\right\|^\alpha \mid \mathcal{F}_t\right] \leq \sigma_{\mathrm{dist}}^\alpha.$$

We now state the high-probability convergence result for the distributed standard nonconvex SGD.

**Proposition 1** *Assume Assumption 3 holds, and suppose that the global objective $f$ is $L$-smooth and lower bounded. Given any $0 < \delta < 1$, define $\Delta := f(\boldsymbol{w}^0) - \min f + \frac{1}{2}$, $u := 4\Delta$, $G := \sqrt{\frac{u}{L}}$, $C_\alpha := 2^{\frac{7}{2}+\frac{2}{\alpha}}$. By choosing*

$$\gamma = \min\left\{\frac{1}{L}, \frac{G\delta^{1/\alpha}}{C_\alpha \sigma_{\mathrm{dist}} T^{1/\alpha}}\right\},$$

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

$$\frac{1}{T}\sum_{t<T}\left\|\nabla f(\boldsymbol{w}^t)\right\|^2 \leq \frac{4L\Delta}{T} + \frac{2C_\alpha\sigma_{\text{dist}}\sqrt{L\Delta}}{\delta^{1/\alpha}T^{(\alpha-1)/\alpha}} = \mathcal{O}\left(\frac{L\Delta}{T} + \frac{\sigma_{\text{dist}}\sqrt{L\Delta}}{\delta^{1/\alpha}T^{(\alpha-1)/\alpha}}\right). \tag{7}$$

Proposition 1 shows that distributed training under heavy-tailed gradient noise is controlled by the effective noise scale $\sigma_{\text{dist}}$. This quantity explicitly captures the joint effect of worker weights, local batch sizes, and heterogeneous tail magnitudes.

We can directly present the result of mini-batch nonconvex SGD by Proposition 1.

**Proposition 2** *Let* $(\boldsymbol{w}^t)_{t\geq 0}$ *be generated by mini-batch SGD,* $\boldsymbol{w}^{t+1} = \boldsymbol{w}^t - \gamma \boldsymbol{g}^t$, $\boldsymbol{g}^t = \frac{1}{B}\sum_{i=1}^B \boldsymbol{g}^{t,i}$*, where* $\{\boldsymbol{g}^{t,i}\}_{i=1}^B$ *are conditionally independent unbiased stochastic gradients satisfying Assumption 1. Suppose also that Assumption 2 holds. Given any* $0<\delta<1$*, define* $\Delta := f(\boldsymbol{w}^0) - \min f + \frac{1}{2}$, $u := 4\Delta$, $G := \sqrt{\frac{u}{L}}$, $C_\alpha := 2^{\frac{7}{2}+\frac{2}{\alpha}}$*. By choosing*

$$\gamma = \min\left\{\frac{1}{L}, \frac{B^{(\alpha-1)/\alpha}G\delta^{1/\alpha}}{4^{1/\alpha}C_\alpha\sigma T^{1/\alpha}}\right\},$$

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

$$\frac{1}{T}\sum_{t<T}\left\|\nabla f(\boldsymbol{w}^t)\right\|^2 \leq \frac{4L\Delta}{T} + \frac{2\cdot 4^{1/\alpha}C_\alpha\sigma\sqrt{L\Delta}}{B^{(\alpha-1)/\alpha}\delta^{1/\alpha}T^{(\alpha-1)/\alpha}} = \mathcal{O}\left(\frac{L\Delta}{T} + \frac{\sigma\sqrt{L\Delta}}{\delta^{1/\alpha}(BT)^{(\alpha-1)/\alpha}}\right). \tag{8}$$

Proposition 2 shows that mini-batching improves the rate through the factor $B^{(\alpha-1)/\alpha}$. Equivalently, the bound depends on $(BT)^{(\alpha-1)/\alpha}$. This result has important implications for distributed training, which can be viewed as a special case of mini-batch SGD, where the batch is distributed across multiple processing units. Therefore, Proposition 2 suggests that distributed training also benefits from increasing the number of participating agents, yielding improved performance as the system scales, despite the potential presence of heavy-tailed noise in the training data. This robustness to heavy-tailed noise is a crucial aspect for real-world applications where such noise is often encountered.

# B Numerical Results for Distributed Settings

## B.1 Improvements with respect to the number of node

We present numerical experiments to demonstrate the mini-batch improvement of mini-batch nonconvex standard nonconvex SGD under heavy-tailed noise. We follow the same tasks used in the numerical tests in Section 1.1. To investigate the effect of batch size on the convergence of the algorithm, we choose four different batch sizes: 32, 64, 128, and 256. The initial learning rate for the smallest batch size 32 is set to 0.04, while for the larger batch sizes, we scale the learning rate by $\sqrt{2}$, 2, and $2\sqrt{2}$, respectively[1]. We adopt the same weight decay coefficient, iteration decay schedule, and total number of epochs as the same as those used in the numerical tests in Section 1.1.

The numerical results are presented in Figure 3, which illustrates the convergence of nonconvex SGD under heavy-tailed noise for different batch sizes. As expected, the results show that as the batch size increases, the convergence of nonconvex SGD under heavy-tailed noise indeed accelerates. This is evident from the fact that the larger batch sizes (e.g., 128 and 256) converge faster than the smaller batch sizes (e.g., 32 and 64). Furthermore, the results also suggest that the acceleration in convergence is consistent with the theoretical prediction that larger batches reduce the impact of heavy-tailed noise, although the theoretical dependence is $B^{(\alpha-1)/\alpha}$ rather than linear in $B$, which is consistent with our theoretical analysis.

[1]Here, we follow the learning rate usages in the traditional mini-batch size training for simplicity.

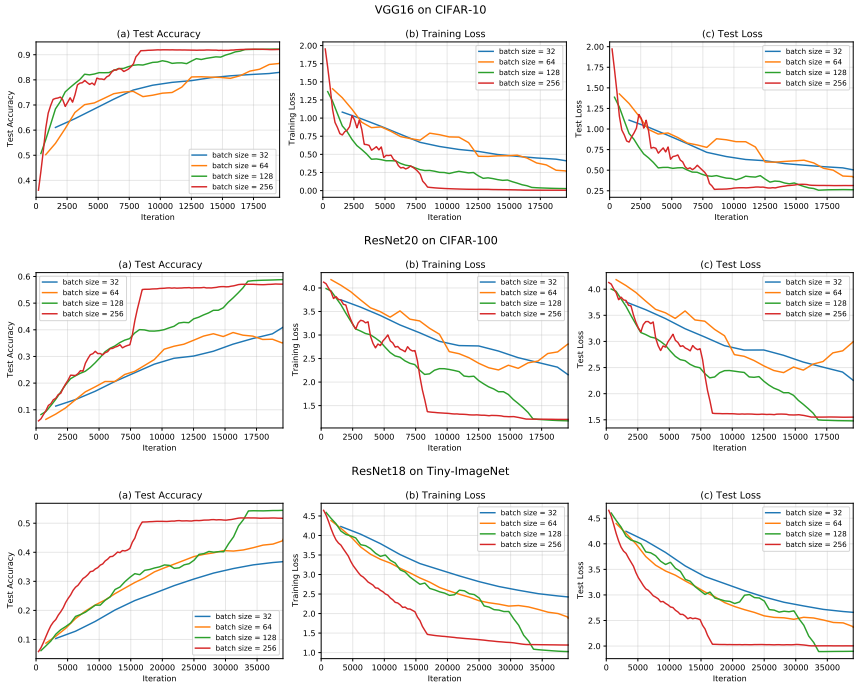



Figure 3: The test accuracy, training loss, and test loss with heavy-tailed noise for SGD under different batch sizes on different settings.

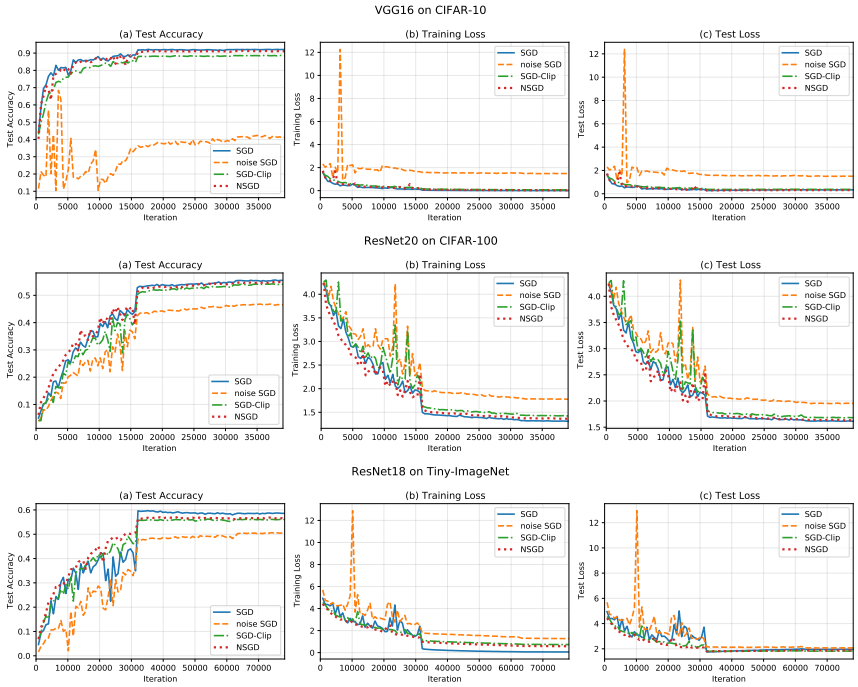



Figure 4: The test accuracy, training loss, and test loss with heavy-tailed noise for clipping methods under different settings.

### B.2 Comparisons with Clipping Methods

We compare SGD with gradient clipping (SGD-CLIP) and normalization method (NSGD). To highlight the effect of clipping and normalization, we consider a larger noise, i.e., using $10^{-3}$ for the noise added in the stochastic gradient. The batchsize is set as 128 and initial learning rate is set as 0.05. We also decay the learning rate at the 41st and 81st epoch with 0.1. For gradient clipping, the clipping scale is set as 5.0. All algorithms are performed with 100 epoches. The comparisons between clipping methods under different settings is presented in 4.

## C Proof of Theorem 1

In this subsection, let

$$\Delta := f(\boldsymbol{w}^0) - \min f + \frac{1}{2}, \qquad u := 4\Delta, \qquad G := \sqrt{\frac{u}{L}},$$

and define

$$\boldsymbol{\epsilon}^t := \boldsymbol{g}^t - \nabla f(\boldsymbol{w}^t).$$

We use the objective stopping time

$$\tau := \min\left\{t \geq 0 : f(\boldsymbol{w}^t) - \min f > u\right\} \wedge T. \tag{9}$$

The proof logic diagram is shown in Figure 5.

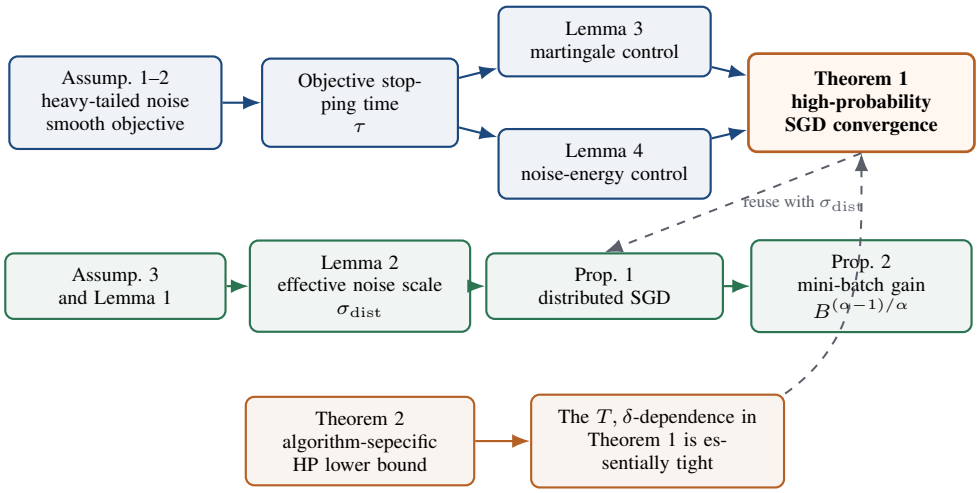



**Figure 5.** Proof roadmap of the paper. Solid arrows denote direct proof dependence, while dashed arrows denote reuse or interpretation of a result.

### C.1 Technical Lemmas for the Proof of Theorem 1

**Lemma 3** *Define*

$$M_T := \sum_{t=0}^{T-1} \gamma_t(1 - L\gamma_t)\left\langle \nabla f(\boldsymbol{w}^t), \boldsymbol{\epsilon}^t \right\rangle \mathbf{1}_{\{t<\tau\}}.$$

*Suppose Assumptions 1 and 2 hold and* $\{\gamma_t\}_{t=0}^{T-1}$ *is a deterministic learning-rate schedule satisfying* $0 < \gamma_t \leq 1/L$. *Then,*

$$\mathbb{P}\left(|M_T| > 2^{\frac{1}{2}+\frac{2}{\alpha}}\sigma LG\left(\frac{\sum_{t=0}^{T-1}\gamma_t^\alpha}{\delta}\right)^{1/\alpha}\right) \leq \frac{\delta}{2}.$$

**Lemma 4** *Let*

$$Q_T := \sum_{t=0}^{T-1} \gamma_t^2 \|\boldsymbol{\epsilon}^t\|^2.$$

*Under Assumption 1, for any deterministic learning-rate schedule* $\{\gamma_t\}_{t=0}^{T-1}$*, we have*

$$\mathbb{P}\left( Q_T > 2^{2/\alpha}\sigma^2 \left( \frac{\sum_{t=0}^{T-1} \gamma_t^\alpha}{\delta} \right)^{2/\alpha} \right) \leq \frac{\delta}{2}.$$

### C.2 Proof of Theorem 1

By $L$-smoothness, we have

$$\begin{aligned} f(\boldsymbol{w}^{t+1}) - f(\boldsymbol{w}^t) \leq & -\gamma_t \left(1 - \frac{L\gamma_t}{2}\right) \|\nabla f(\boldsymbol{w}^t)\|^2 \\ & - \gamma_t(1 - L\gamma_t) \left\langle \nabla f(\boldsymbol{w}^t), \boldsymbol{\epsilon}^t \right\rangle + \frac{L\gamma_t^2}{2} \|\boldsymbol{\epsilon}^t\|^2. \end{aligned}$$

Since $\gamma_t \leq 1/L$, summing the above inequality over $t < \tau$ gives

$$f(\boldsymbol{w}^\tau) - \min f + \frac{1}{2} \sum_{t<\tau} \gamma_t \|\nabla f(\boldsymbol{w}^t)\|^2 \leq \; f(\boldsymbol{w}^0) - \min f + |M_T| + \frac{L}{2} Q_T. \tag{10}$$

Define

$$\mathcal{G}_M := \left\{ |M_T| \leq 2^{\frac{1}{2}+\frac{2}{\alpha}} \sigma L G \left( \frac{\sum_{t=0}^{T-1} \gamma_t^\alpha}{\delta} \right)^{1/\alpha} \right\},$$

and

$$\mathcal{G}_Q := \left\{ Q_T \leq 2^{2/\alpha}\sigma^2 \left( \frac{\sum_{t=0}^{T-1} \gamma_t^\alpha}{\delta} \right)^{2/\alpha} \right\}.$$

We also let

$$\mathcal{G} := \mathcal{G}_M \cap \mathcal{G}_Q.$$

Based on Lemmas 3 and 4, we are led to

$$\mathbb{P}(\mathcal{G}_M^c) \leq \frac{\delta}{2}, \quad \mathbb{P}(\mathcal{G}_Q^c) \leq \frac{\delta}{2}.$$

Hence, by the union bound, it follows

$$\mathbb{P}(\mathcal{G}) \geq 1 - \mathbb{P}(\mathcal{G}_M^c) - \mathbb{P}(\mathcal{G}_Q^c) \geq 1 - \delta.$$

Recall that our settings

$$\left( \sum_{t=0}^{T-1} \gamma_t^\alpha \right)^{1/\alpha} \leq \frac{G\delta^{1/\alpha}}{C_\alpha \sigma}, \quad C_\alpha = 2^{\frac{7}{2}+\frac{2}{\alpha}}.$$

Therefore, on $\mathcal{G}$, we have

$$|M_T| \leq 2^{\frac{1}{2}+\frac{2}{\alpha}} \sigma L G \left( \frac{\sum_{t=0}^{T-1} \gamma_t^\alpha}{\delta} \right)^{1/\alpha} \leq \frac{2^{\frac{1}{2}+\frac{2}{\alpha}}}{C_\alpha} LG^2 = \frac{LG^2}{8} = \frac{u}{8} = \frac{\Delta}{2}.$$

and

$$\frac{L}{2} Q_T \leq 2^{\frac{2}{\alpha}-1} L\sigma^2 \left( \frac{\sum_{t=0}^{T-1} \gamma_t^\alpha}{\delta} \right)^{2/\alpha} \leq \frac{2^{\frac{2}{\alpha}-1}}{C_\alpha^2} LG^2 \leq \frac{LG^2}{8} = \frac{u}{8} = \frac{\Delta}{2}.$$

Substituting the above bounds into (10), on $\mathcal{G}$ we obtain

$$f(\boldsymbol{w}^\tau) - \min f + \frac{1}{2}\sum_{t<\tau}\gamma_t\|\nabla f(\boldsymbol{w}^t)\|^2 \leq\ f(\boldsymbol{w}^0) - \min f + \Delta < 2\Delta = \frac{u}{2}.$$

If $\tau < T$, the definition of $\tau$ implies $f(\boldsymbol{w}^\tau) - \min f > u$, which contradicts the above inequality. Hence, we can conclude that $\mathcal{G} \subseteq \{\tau = T\}$.

Since $\tau = T$ on $\mathcal{G}$ and $f(\boldsymbol{w}^T) - \min f \geq 0$, we have

$$\frac{1}{2}\sum_{t=0}^{T-1}\gamma_t\|\nabla f(\boldsymbol{w}^t)\|^2 \leq 2\Delta \Longrightarrow \sum_{t=0}^{T-1}\gamma_t\|\nabla f(\boldsymbol{w}^t)\|^2 \leq 4\Delta.$$

Dividing both sides by $\sum_{t=0}^{T-1}\gamma_t$ yields

$$\frac{\sum_{t=0}^{T-1}\gamma_t\|\nabla f(\boldsymbol{w}^t)\|^2}{\sum_{t=0}^{T-1}\gamma_t} \leq \frac{4\Delta}{\sum_{t=0}^{T-1}\gamma_t},$$

which proves (2).

Now, we turn to the constant learning rate. Let

$$\gamma_t \equiv \gamma = \min\left\{\frac{1}{L}, \frac{G\delta^{1/\alpha}}{C_\alpha\sigma T^{1/\alpha}}\right\}.$$

Then, we can get

$$\left(\sum_{t=0}^{T-1}\gamma_t^\alpha\right)^{1/\alpha} = \gamma T^{1/\alpha} \leq \frac{G\delta^{1/\alpha}}{C_\alpha\sigma}.$$

Moreover,

$$\frac{\sum_{t=0}^{T-1}\gamma\|\nabla f(\boldsymbol{w}^t)\|^2}{\sum_{t=0}^{T-1}\gamma} = \frac{1}{T}\sum_{t=0}^{T-1}\|\nabla f(\boldsymbol{w}^t)\|^2 \Longrightarrow \frac{1}{T}\sum_{t=0}^{T-1}\|\nabla f(\boldsymbol{w}^t)\|^2 \leq \frac{4\Delta}{\gamma T}.$$

With the learning rate settings

$$\frac{1}{\gamma} = \max\left\{L, \frac{C_\alpha\sigma T^{1/\alpha}}{G\delta^{1/\alpha}}\right\} \leq L + \frac{C_\alpha\sigma T^{1/\alpha}}{G\delta^{1/\alpha}},$$

we obtain

$$\frac{4\Delta}{\gamma T} \leq \frac{4L\Delta}{T} + \frac{4C_\alpha\sigma\Delta}{G\delta^{1/\alpha}T^{(\alpha-1)/\alpha}}.$$

Using $G = 2\sqrt{\frac{\Delta}{L}}$ and $\frac{4\Delta}{G} = 2\sqrt{L\Delta}$, we finally obtain

$$\frac{1}{T}\sum_{t<T}\|\nabla f(\boldsymbol{w}^t)\|^2 \leq \frac{4L\Delta}{T} + \frac{2C_\alpha\sigma\sqrt{L\Delta}}{\delta^{1/\alpha}T^{(\alpha-1)/\alpha}},$$

which proves (3).

### C.3 Proof of Lemma 3

For $t < T$, we define

$$D_t := \gamma_t(1 - L\gamma_t)\left\langle\nabla f(\boldsymbol{w}^t), \boldsymbol{\epsilon}^t\right\rangle \mathbf{1}_{\{t<\tau\}}. \tag{11}$$

With the definition of the stopping time $\tau$, the event $\{t < \tau\}$ belongs to $\mathcal{F}_t$. Moreover, $\boldsymbol{w}^t$ is $\mathcal{F}_t$-measurable and $\gamma_t$ is deterministic. Therefore, by Assumption 1, we have

$$\mathbb{E}[D_t \mid \mathcal{F}_t] = \gamma_t(1 - L\gamma_t)\mathbf{1}_{\{t<\tau\}}\left\langle\nabla f(\boldsymbol{w}^t), \mathbb{E}[\boldsymbol{\epsilon}^t \mid \mathcal{F}_t]\right\rangle = 0.$$

Hence, $\{D_t\}_{t=0}^{T-1}$ is a martingale difference sequence and $M_T = \sum_{t=0}^{T-1} D_t$. On the event $\{t < \tau\}$, the definition of $\tau$ gives

$$f(\boldsymbol{w}^t) - \min f \le u.$$

Since $f$ is $L$-smooth, we have

$$\|\nabla f(\boldsymbol{w}^t)\|^2 \le 2L\big(f(\boldsymbol{w}^t) - \min f\big).$$

Consequently, on $\{t < \tau\}$,

$$\|\nabla f(\boldsymbol{w}^t)\|^2 \le 2Lu = 2L^2G^2, \tag{12}$$

where we used $G = \sqrt{u/L}$. Therefore, we can get

$$\|\nabla f(\boldsymbol{w}^t)\|\mathbf{1}_{\{t<\tau\}} \le \sqrt{2}\, LG. \tag{13}$$

With the setting that $0 < \gamma_t \le 1/L$, from (11) and (13), we have

$$\begin{aligned}|D_t|^\alpha &= \gamma_t^\alpha |1 - L\gamma_t|^\alpha \left|\left\langle \nabla f(\boldsymbol{w}^t), \boldsymbol{\epsilon}^t \right\rangle\right|^\alpha \mathbf{1}_{\{t<\tau\}} \\ &\le \gamma_t^\alpha \|\nabla f(\boldsymbol{w}^t)\|^\alpha \|\boldsymbol{\epsilon}^t\|^\alpha \mathbf{1}_{\{t<\tau\}} \le 2^{\alpha/2}\gamma_t^\alpha L^\alpha G^\alpha \|\boldsymbol{\epsilon}^t\|^\alpha \mathbf{1}_{\{t<\tau\}}.\end{aligned}$$

Taking conditional expectation with respect to $\mathcal{F}_t$ and using Assumption 1, we obtain

$$\mathbb{E}\left[|D_t|^\alpha \mid \mathcal{F}_t\right] \le 2^{\alpha/2}\gamma_t^\alpha L^\alpha G^\alpha \mathbf{1}_{\{t<\tau\}} \mathbb{E}\left[\|\boldsymbol{\epsilon}^t\|^\alpha \mid \mathcal{F}_t\right] \le 2^{\alpha/2}\gamma_t^\alpha \sigma^\alpha L^\alpha G^\alpha \mathbf{1}_{\{t<\tau\}}.$$

Taking full expectations,

$$\mathbb{E}|D_t|^\alpha \le 2^{\alpha/2}\gamma_t^\alpha \sigma^\alpha L^\alpha G^\alpha. \tag{14}$$

Using the von Bahr-Esseen inequality for martingale differences (Pinelis, 2015), together with (14), gives

$$\mathbb{E}|M_T|^\alpha = \mathbb{E}\left|\sum_{t=0}^{T-1} D_t\right|^\alpha \le 2\sum_{t=0}^{T-1}\mathbb{E}|D_t|^\alpha \le 2^{1+\alpha/2}\sigma^\alpha L^\alpha G^\alpha \sum_{t=0}^{T-1}\gamma_t^\alpha. \tag{15}$$

Now set

$$a := 2^{\frac{1}{2}+\frac{2}{\alpha}}\sigma LG\left(\frac{\sum_{t=0}^{T-1}\gamma_t^\alpha}{\delta}\right)^{1/\alpha}.$$

By the Markov's inequality and (15), we can have

$$\begin{aligned}\mathbb{P}(|M_T| > a) &= \mathbb{P}(|M_T|^\alpha > a^\alpha) \\ &\le \frac{\mathbb{E}|M_T|^\alpha}{a^\alpha} \le \frac{2^{1+\alpha/2}\sigma^\alpha L^\alpha G^\alpha \sum_{t=0}^{T-1}\gamma_t^\alpha}{2^{\alpha/2+2}\sigma^\alpha L^\alpha G^\alpha \left(\sum_{t=0}^{T-1}\gamma_t^\alpha\right)/\delta} = \frac{\delta}{2}.\end{aligned}$$

### C.4 Proof of Lemma 4

Recall that $Q_T = \sum_{t=0}^{T-1}\gamma_t^2\|\boldsymbol{\epsilon}^t\|^2$. Since $\alpha \in (1,2]$, we have $\alpha/2 \in (1/2, 1]$. Thus, the function $x \mapsto x^{\alpha/2}$ is subadditive on $\mathbb{R}_+$. Therefore, we derive

$$Q_T^{\alpha/2} = \left(\sum_{t=0}^{T-1}\gamma_t^2\|\boldsymbol{\epsilon}^t\|^2\right)^{\alpha/2} \le \sum_{t=0}^{T-1}\left(\gamma_t^2\|\boldsymbol{\epsilon}^t\|^2\right)^{\alpha/2} = \sum_{t=0}^{T-1}\gamma_t^\alpha\|\boldsymbol{\epsilon}^t\|^\alpha. \tag{16}$$

Taking full expectations and with Assumption 1, we obtain

$$\mathbb{E}[Q_T^{\alpha/2}] \le \sum_{t=0}^{T-1}\gamma_t^\alpha\mathbb{E}\|\boldsymbol{\epsilon}^t\|^\alpha = \sum_{t=0}^{T-1}\gamma_t^\alpha\mathbb{E}\left[\mathbb{E}\left[\|\boldsymbol{\epsilon}^t\|^\alpha \mid \mathcal{F}_t\right]\right] \le \sigma^\alpha\sum_{t=0}^{T-1}\gamma_t^\alpha. \tag{17}$$

Set

$$b := 2^{2/\alpha}\sigma^2\left(\frac{\sum_{t=0}^{T-1}\gamma_t^\alpha}{\delta}\right)^{2/\alpha} \Longrightarrow b^{\alpha/2} = \left[2^{2/\alpha}\sigma^2\left(\frac{\sum_{t=0}^{T-1}\gamma_t^\alpha}{\delta}\right)^{2/\alpha}\right]^{\alpha/2} = \frac{2\sigma^\alpha\sum_{t=0}^{T-1}\gamma_t^\alpha}{\delta}.$$

Therefore, by Markov's inequality and (17),

$$\mathbb{P}(Q_T > b) = \mathbb{P}\left(Q_T^{\alpha/2} > b^{\alpha/2}\right) \le \frac{\mathbb{E}[Q_T^{\alpha/2}]}{b^{\alpha/2}} \le \frac{\sigma^\alpha\sum_{t=0}^{T-1}\gamma_t^\alpha}{2\sigma^\alpha\left(\sum_{t=0}^{T-1}\gamma_t^\alpha\right)/\delta} = \frac{\delta}{2}.$$

## D PROOF OF THEOREM 2

This subsection defines

$$r_{T,\delta} := \min\left\{1, \frac{1}{\delta^{1/\alpha} T^{(\alpha-1)/\alpha}}\right\},\ A := T r_{T,\delta}.$$

Since $T \geq 8$, $\delta < 1/16$, and $1 < \alpha \leq 2$, we have $8 \leq A \leq T$ and

$$\delta \leq \frac{T}{A^\alpha}. \tag{18}$$

Set $m := \lfloor \frac{A}{8} \rfloor$. Due to $A \geq 8$, we have $\frac{A}{16} \leq\ m \leq \frac{A}{8} \leq \frac{T}{8}$.

We distinguish two cases.

**Case 1.** We assume that there exists an index $a \in \{0, \ldots, T-m\}$ such that

$$\sum_{t=a}^{a+m-1} \gamma_t \leq 1.$$

Consider the one-dimensional function

$$\phi(x) := \begin{cases} \frac{1}{4}, & x \leq 0, \\ \frac{1}{4} - \frac{x^2}{2}, & 0 \leq x \leq 1/2, \\ \frac{(x-1)^2}{2}, & 1/2 \leq x \leq 1, \\ 0, & x \geq 1. \end{cases}$$

Its derivative is then

$$\phi'(x) = \begin{cases} 0, & x \leq 0, \\ -x, & 0 \leq x \leq 1/2, \\ x-1, & 1/2 \leq x \leq 1, \\ 0, & x \geq 1. \end{cases}$$

Hence $\phi'$ is 1-Lipschitz, so $\phi$ is 1-smooth. Furthermore, $\phi$ is nonconvex due to $\phi''(x) = -1$ on $(0, 1/2)$.

Let $h := \frac{1}{4e}$ and choose the initialization $x^0 := \dfrac{h}{\prod_{j=0}^{a-1}(1+\gamma_j)}$, where the empty product is 1. Use the exact gradient oracle $g^t = \phi'(x^t)$. We first verify the trajectory before iteration $a$. For every $t = 0, \ldots, a$, with a simple mathematical induction, we can have $x^t = \dfrac{h}{\prod_{j=t}^{a-1}(1+\gamma_j)}$, where the empty product equals 1. Hence, we have $0 < x^t \leq h = \frac{1}{4e} < \frac{1}{2}$; an the identity $\phi'(x^t) = -x^t$ is valid throughout $t = 0, \ldots, a-1$, and in particular $x^a = h$. Thus the stochastic gradient noise is identically zero and satisfies the required conditional $\alpha$-moment assumption.

When $0 < x^t < 1/2$, we have $\phi'(x^t) = -x^t$, and therefore $x^{t+1} = x^t - \gamma_t \phi'(x^t) = (1+\gamma_t)x^t$. By the initialization, we have $x^a = h$. For every $t \in \{a, \ldots, a+m-1\}$, it follows

$$x^t = h \prod_{j=a}^{t-1}(1+\gamma_j) \leq\ h \exp\left(\sum_{j=a}^{t-1} \gamma_j\right) \leq\ he = \frac{1}{4} < \frac{1}{2}.$$

Thus the recursion is self-consistent throughout this interval. Moreover, we have

$$x^t \geq h, \quad t = a, \ldots, a+m-1.$$

Consequently, we are led to

$$|\phi'(x^t)|^2 = |x^t|^2 \geq\ h^2, \quad t = a, \ldots, a+m-1.$$

Therefore, we then get

$$\frac{1}{T}\sum_{t=0}^{T-1}|\phi'(x^t)|^2 \geq \frac{h^2 m}{T} \geq \frac{h^2 A}{16T} = \frac{1}{256e^2}\frac{A}{T} = \frac{1}{256e^2} r_{T,\delta}. \tag{19}$$

The bound holds with probability one.

With direct computation, we can check that $\phi(x^0) - \min \phi \leq \frac{1}{4} \leq \frac{1}{2}$. Hence, all conditions of the theorem are satisfied in Case 1.

**Case 2.** We now suppose that Case 1 does not occur. Thus, for every $a \in \{0, \ldots, T-m\}$, we have

$$\sum_{t=a}^{a+m-1} \gamma_t > 1. \tag{20}$$

We partition the learning rate schedule into consecutive renewal blocks according to cumulative learning rate. Let $\tau_0 := 0$. Starting from $\tau_k$, if $\sum_{t=\tau_k}^{T-1} \gamma_t < 1$, we stop. Otherwise, we define

$$\tau_{k+1} := \min\left\{ t > \tau_k : \sum_{j=\tau_k}^{t-1} \gamma_j \geq 1 \right\}. \tag{21}$$

Suppose that this construction produces $K$ complete blocks as $I_k := \{\tau_k, \ldots, \tau_{k+1} - 1\}$, $k = 0, \ldots, K-1$, and possibly a final incomplete block $I_{\mathrm{rem}} := \{\tau_K, \ldots, T-1\}$. We denote that

$$\ell_k := |I_k|, \quad \ell_{\mathrm{rem}} := |I_{\mathrm{rem}}|.$$

By the definition of $\tau_{k+1}$ and the fact that $\gamma_t \leq 1$, each complete block satisfies

$$1 \leq S_k := \sum_{t \in I_k} \gamma_t < 2. \tag{22}$$

Furthermore, we can conclude that

$$\ell_k \leq m, \ k = 0, \ldots, K-1.$$

Indeed, if $\ell_k > m$, then by the minimality in (21), $\sum_{t=\tau_k}^{\tau_k+m-1} \gamma_t < 1$, contradicting (20). For the same reason, $\ell_{\mathrm{rem}} < m$ because the total learning rate on the incomplete block is smaller than 1.

Since $m \leq T/8$, there are at least two complete blocks. Moreover, we have

$$\sum_{k=1}^{K-1} \ell_k = T - \ell_0 - \ell_{\mathrm{rem}} \geq T - 2m \geq \frac{3T}{4}. \tag{23}$$

By setting $p := \frac{\alpha}{2} \in \left(\frac{1}{2}, 1\right]$, for every $s \in I_k$ with $k = 0, \ldots, K-2$, we define

$$q_s := \gamma_s^2 \ell_{k+1}, \ \mathcal{I} := \bigcup_{k=0}^{K-2} I_k, W := \sum_{s \in \mathcal{I}} q_s^p. \tag{24}$$

We next turn to the lower bound of $W$. Consider the H"older's inequality as

$$\sum_{s \in I_k} \gamma_s = \sum_{s \in I_k} \gamma_s \cdot 1 \leq \left(\sum_{s \in I_k} \gamma_s^\alpha\right)^{1/\alpha} \left(\sum_{s \in I_k} 1^{\alpha/(\alpha-1)}\right)^{(\alpha-1)/\alpha} = \left(\sum_{s \in I_k} \gamma_s^\alpha\right)^{1/\alpha} \ell_k^{(\alpha-1)/\alpha},$$

where we used $|I_k| = \ell_k$. Hence, for every complete block $I_k$, we can have

$$\sum_{s \in I_k} \gamma_s^\alpha \geq \frac{\left(\sum_{s \in I_k} \gamma_s\right)^\alpha}{\ell_k^{\alpha-1}} \geq \frac{1}{\ell_k^{\alpha-1}} \Longrightarrow W = \sum_{k=0}^{K-2} \ell_{k+1}^p \sum_{s \in I_k} \gamma_s^{2p} \geq \sum_{k=0}^{K-2} \frac{\ell_{k+1}^p}{\ell_k^{2p-1}}.$$

Since both $\ell_k$ and $\ell_{k+1}$ are at most $m$, and $p \leq 1$, we have $\ell_{k+1}^p \geq \frac{\ell_{k+1}}{m^{1-p}}$, $\ell_k^{2p-1} \leq m^{2p-1}$. Therefore, we can get

$$\frac{\ell_{k+1}^p}{\ell_k^{2p-1}} \geq \frac{\ell_{k+1}}{m^p}. \tag{25}$$

Combining (23)-(25), we can get

$$W \geq \frac{1}{m^p} \sum_{k=1}^{K-1} \ell_k \geq \frac{3T}{4m^p} \geq \frac{3 \cdot 8^p}{4} \frac{T}{A^p}. \tag{26}$$

We now construct the stochastic hard instance. Use one coordinate $x_s$ for every $s \in \mathcal{I}$ and one additional coordinate $y$. We define the function as

$$f(\boldsymbol{x}, y) := \frac{1}{16} \sum_{s \in \mathcal{I}} x_s^2 + (1 - \cos y). \tag{27}$$

The function is 1-smooth and nonconvex since $\frac{\partial^2 f}{\partial y^2}(\boldsymbol{x}, \pi) = -1$. Consider the initialization as $\boldsymbol{x}^0 = \mathbf{0}$ and $y^0 = 0$, and thus $f(\boldsymbol{w}^0) - \min f = 0 \leq \frac{1}{2}$. For each $s \in \mathcal{I}$, define

$$\pi_s := \frac{4\delta q_s^p}{W}. \tag{28}$$

Since $\sum_{s \in \mathcal{I}} \pi_s = 4\delta < \frac{1}{4}$, we have $0 < \pi_s < 1$. Let $\{\boldsymbol{\xi}^t\}_{t=0}^{T-1}$ be independent random vectors. For $s \in \mathcal{I}$, we define

$$\boldsymbol{\xi}^s = \begin{cases} \pi_s^{-1/\alpha} \boldsymbol{e}_s, & \text{with probability } \pi_s/2, \\ -\pi_s^{-1/\alpha} \boldsymbol{e}_s, & \text{with probability } \pi_s/2, \\ \mathbf{0}, & \text{with probability } 1 - \pi_s, \end{cases}$$

where $\boldsymbol{e}_s$ denotes the unit vector in coordinate $x_s$. For $t \notin \mathcal{I}$, set $\boldsymbol{\xi}^t = \mathbf{0}$ almost surely. Defining the stochastic gradient oracle by $\boldsymbol{g}^t = \nabla f(\boldsymbol{w}^t) + \boldsymbol{\xi}^t$, we can easily see that $\mathbb{E}[\boldsymbol{g}^t \mid \mathcal{F}_t] = \nabla f(\boldsymbol{w}^t)$. Moreover, for every $s \in \mathcal{I}$, it follows

$$\mathbb{E}\left[\|\boldsymbol{\xi}^s\|^\alpha \mid \mathcal{F}_s\right] = \pi_s \left(\pi_s^{-1/\alpha}\right)^\alpha = 1,$$

indicating that the required conditional $\alpha$-moment assumption holds. We also denote that

$$\mathcal{E} := \{\boldsymbol{\xi}^s \neq \mathbf{0} \text{ for at least one } s \in \mathcal{I}\}.$$

By independence, we can get

$$\mathbb{P}(\mathcal{E}) = 1 - \prod_{s \in \mathcal{I}} (1 - \pi_s) \geq 1 - \exp\left(-\sum_{s \in \mathcal{I}} \pi_s\right) = 1 - e^{-4\delta} \geq 2\delta \geq \delta, \tag{29}$$

where we used that $4\delta < 1$ and the inequality $1 - e^{-x} \geq x/2$ for $x \in [0, 1]$.

It remains to lower bound the gradient energy on $\mathcal{E}$. Suppose that a nonzero shock occurs at some $s \in I_k$, where $k \leq K - 2$. Before iteration $s$, the coordinate $x_s$ is identically zero, i.e., $x_s^{s+1} = -\gamma_s \xi_s^s$. No subsequent noise acts on the same coordinate, and therefore for every $t \geq s + 1$,

$$x_s^t = -\gamma_s \xi_s^s \prod_{j=s+1}^{t-1} (1 - \gamma_j/8). \tag{30}$$

Consider any $t \in I_{k+1}$. The indices $j = s + 1, \ldots, t - 1$ are contained in the remainder of $I_k$ and a prefix of $I_{k+1}$. By (22), we can get

$$\sum_{j=s+1}^{t-1} \gamma_j < S_k + S_{k+1} < 4 \Longrightarrow \sum_{j=s+1}^{t-1} \gamma_j/8 < \frac{1}{2}. \tag{31}$$

Using the inequality as $\prod_i (1 - a_i) \geq 1 - \sum_i a_i$ with $a_i \in [0, 1]$, we obtain $\prod_{j=s+1}^{t-1} (1 - \gamma_j/8) \geq \frac{1}{2}$. Consequently, on the event $\{\boldsymbol{\xi}^s \neq \mathbf{0}\}$, we have

$$\sum_{t=0}^{T-1} \|\nabla f(\boldsymbol{w}^t)\|^2 \geq \sum_{t \in I_{k+1}} \frac{1}{64} |x_s^t|^2 \geq \frac{1}{256} \gamma_s^2 \ell_{k+1} \pi_s^{-2/\alpha} = \frac{1}{256} q_s \pi_s^{-1/p}, \tag{32}$$

where we used the fact that $2/\alpha = 1/p$. By the definition of $\pi_s$, we also have $q_s \pi_s^{-1/p} = \left(\frac{W}{4\delta}\right)^{1/p}$. Thus every nonzero shock, regardless of its location, produces the same lower bound

$$\sum_{t=0}^{T-1} \|\nabla f(\boldsymbol{w}^t)\|^2 \geq \frac{1}{256} \left(\frac{W}{4\delta}\right)^{1/p}. \tag{33}$$

Using (26) and (18),

$$\left(\frac{W}{4\delta}\right)^{1/p} \geq \left(\frac{3 \cdot 8^p}{16} \frac{T}{\delta A^p}\right)^{1/p} \geq \left(\frac{3 \cdot 8^p}{16} A^p\right)^{1/p} = 8 \left(\frac{3}{16}\right)^{1/p} A. \tag{34}$$

Hence, on $\mathcal{E}$, we have

$$\frac{1}{T} \sum_{t=0}^{T-1} \|\nabla f(\boldsymbol{w}^t)\|^2 \geq \frac{1}{32} \left(\frac{3}{16}\right)^{1/p} \frac{A}{T} = \frac{1}{32} \left(\frac{3}{16}\right)^{2/\alpha} r_{T,\delta}. \tag{35}$$

Together with (29), this lower bound holds with probability at least $\delta$.

Combining Cases 1 and 2 proves

$$\mathbb{P}\left(\frac{1}{T} \sum_{t=0}^{T-1} \|\nabla f(\boldsymbol{w}^t)\|^2 \geq c_\alpha \min\left\{1, \frac{1}{\delta^{1/\alpha} T^{(\alpha-1)/\alpha}}\right\}\right) \geq \delta,$$

where

$$c_\alpha := \min\left\{\frac{1}{256e^2}, \frac{1}{32}\left(\frac{3}{16}\right)^{2/\alpha}\right\}.$$

## E Proof of Corollary 1

The proof is quite simple and direct. For every problem in $\mathfrak{P}_\alpha$, $\Delta = f(\boldsymbol{w}^0) - \min f + \frac{1}{2} \in \left[\frac{1}{2}, 1\right]$. Choose the constant learning rate $\gamma_t \equiv \gamma := \min\left\{1, \frac{\sqrt{2}\delta^{1/\alpha}}{C_\alpha T^{1/\alpha}}\right\}$ with $C_\alpha = 2^{\frac{7}{2}+\frac{2}{\alpha}}$. Because $G = \sqrt{4\Delta} \geq \sqrt{2}$ and $\sigma \leq 1$, this learning rate satisfies the condition of Theorem 1. Therefore, with probability at least $1-\delta$, $\mathcal{E}_T(\boldsymbol{\gamma}) \leq \frac{4\Delta}{\gamma T} = \mathcal{O}\left(\frac{1}{\delta^{1/\alpha} T^{(\alpha-1)/\alpha}}\right)$. Hence, we can have

$$\mathfrak{R}_\alpha(T, \delta) \leq C'_\alpha \frac{1}{\delta^{1/\alpha} T^{(\alpha-1)/\alpha}} \tag{36}$$

for a constant $C'_\alpha > 0$ depending only on $\alpha$.

For the converse, fix an arbitrary deterministic learning-rate schedule $\boldsymbol{\gamma} \in (0,1]^T$. Apply Theorem 2 with failure probability $2\delta$. Due to that $2\delta < 1/16$ and $\frac{1}{(2\delta)^{1/\alpha} T^{(\alpha-1)/\alpha}} \leq \frac{1}{\delta^{1/\alpha} T^{(\alpha-1)/\alpha}} \leq 1$, there exists an instance in $\mathfrak{P}_\alpha$ such that

$$\mathbb{P}\left(\mathcal{E}_T(\boldsymbol{\gamma}) \geq c_\alpha \frac{1}{(2\delta)^{1/\alpha} T^{(\alpha-1)/\alpha}}\right) \geq 2\delta. \tag{37}$$

Given any $r < c_\alpha \frac{1}{(2\delta)^{1/\alpha} T^{(\alpha-1)/\alpha}}$, then we can get $\left\{\mathcal{E}_T(\boldsymbol{\gamma}) \geq c_\alpha \frac{1}{(2\delta)^{1/\alpha} T^{(\alpha-1)/\alpha}}\right\} \subseteq \{\mathcal{E}_T(\boldsymbol{\gamma}) > r\}$. Therefore, by (37), it holds $\sup_{(f,\mathcal{O}) \in \mathfrak{P}_\alpha} \mathbb{P}\left(\mathcal{E}_T(\boldsymbol{\gamma}) > r\right) \geq 2\delta > \delta$. Since $\boldsymbol{\gamma}$ was arbitrary, we conclude that

$$\mathfrak{R}_\alpha(T, \delta) \geq \frac{c_\alpha}{2^{1/\alpha}} \frac{1}{\delta^{1/\alpha} T^{(\alpha-1)/\alpha}}. \tag{38}$$

where $c'_\alpha > 0$ depends only on $\alpha$. Combining (36) and (38) proves the result.

## F Missing Details

### F.1 Relation to Previous High-Probability Rates When $\alpha = 2$

When $\alpha = 2$, Theorem 1 reduces to the bounded-variance setting and gives the learning rate as $\gamma = \Theta\left(\sqrt{\frac{\delta}{T}}\right)$ and convergence rate as $\frac{1}{T}\sum_{t<T} \|\nabla f(\boldsymbol{w}^t)\|^2 = \mathcal{O}\left(\frac{1}{\sqrt{\delta T}}\right)$ with probability at

least $1-\delta$. At first sight, this rate may appear different from the lower bound for SGD reported in Jin et al. (2026). We next show that there is no discrepancy: when their more general response-dependent lower bound is evaluated at the confidence-calibrated learning rate $\gamma = \Theta\left(\sqrt{\frac{\delta}{T}}\right)$, it gives exactly the same $\delta^{-1/2}$ dependence as Theorem 1.

The key quantity in the lower-bound analysis of (Jin et al., 2026) is about

$$R_{\gamma,T} := \gamma^2 \sum_{r=0}^{T-\lfloor T/2 \rfloor - 1} (1-\gamma)^{2r} = \frac{\gamma}{2-\gamma}\left[1-(1-\gamma)^{2(T-\lfloor T/2 \rfloor)}\right]. \tag{39}$$

Hence, whenever $0 < \gamma \le 1/2$ and $\gamma T \ge 1$,

$$\frac{1-e^{-1}}{2}\gamma \le R_{\gamma,T} \le \gamma \Longrightarrow R_{\gamma,T} = \Theta(\gamma). \tag{40}$$

The response-dependent lower bound in (Jin et al., 2026) implies that there exists a bounded-variance stochastic oracle for which

$$\mathbb{P}\left(\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(\boldsymbol{w}^t)\|^2 \ge c\frac{R_{\gamma,T}}{\delta}\right) > \delta \tag{41}$$

for some absolute constant $c > 0$. Substituting our choice $\gamma = \Theta\left(\sqrt{\frac{\delta}{T}}\right)$ into (40) gives, in the nontrivial regime $\delta T \gtrsim 1$,

$$\frac{R_{\gamma,T}}{\delta} = \Theta\left(\frac{1}{\delta}\sqrt{\frac{\delta}{T}}\right) = \Theta\left(\frac{1}{\sqrt{\delta T}}\right). \tag{42}$$

This exactly matches the dependence on $(T, \delta)$ in our upper bound. Therefore, our result does not evade the lower bound mechanism identified in (Jin et al., 2026).

### F.2 Proof of Lemma 2

By the definition of $\boldsymbol{g}_i^t$ and Assumption 3, we have $\mathbb{E}\left[\boldsymbol{g}_i^t \mid \mathcal{F}_t\right] = \nabla f_i(\boldsymbol{w}^t)$. Therefore, we get $\mathbb{E}\left[\boldsymbol{g}^t \mid \mathcal{F}_t\right] = \sum_{i=1}^m p_i \nabla f_i(\boldsymbol{w}^t) = \nabla f(\boldsymbol{w}^t)$. By defining

$$\boldsymbol{\epsilon}_{i,j}^t := \boldsymbol{g}_{i,j}^t - \nabla f_i(\boldsymbol{w}^t), \quad \boldsymbol{\epsilon}_i^t := \boldsymbol{g}_i^t - \nabla f_i(\boldsymbol{w}^t),$$

we then have $\boldsymbol{\epsilon}_i^t = \frac{1}{b_i}\sum_{j=1}^{b_i}\boldsymbol{\epsilon}_{i,j}^t$. Using Lemma 1, we obtain $\mathbb{E}\left[\|\boldsymbol{\epsilon}_i^t\|^\alpha \mid \mathcal{F}_t\right] \le \frac{2\sigma_i^\alpha}{b_i^{\alpha-1}}$. Due to $\boldsymbol{g}^t - \nabla f(\boldsymbol{w}^t) = \sum_{i=1}^m p_i \boldsymbol{\epsilon}_i^t$, applying Lemma 1 again to the independent mean-zero random vectors $\{p_i\boldsymbol{\epsilon}_i^t\}_{i=1}^m$ gives

$$\mathbb{E}\left[\left\|\boldsymbol{g}^t - \nabla f(\boldsymbol{w}^t)\right\|^\alpha \mid \mathcal{F}_t\right] \le 2\sum_{i=1}^m p_i^\alpha \mathbb{E}\left[\left\|\boldsymbol{\epsilon}_i^t\right\|^\alpha \mid \mathcal{F}_t\right] \le 4\sum_{i=1}^m p_i^\alpha \frac{\sigma_i^\alpha}{b_i^{\alpha-1}}. \tag{43}$$

### F.3 Proof of Proposition 1

By Lemma 2, the aggregated stochastic gradient $\boldsymbol{g}^t$ is conditionally unbiased and satisfies

$$\mathbb{E}\left[\left\|\boldsymbol{g}^t - \nabla f(\boldsymbol{w}^t)\right\|^\alpha \mid \mathcal{F}_t\right] \le \sigma_{\mathrm{dist}}^\alpha. \tag{44}$$

Therefore, nonconvex standard distributed SGD satisfies Assumption 1 with $\sigma$ replaced by $\sigma_{\mathrm{dist}}$. Applying Theorem 1 directly yields the desired result.