Tight Long-Term Tail Decay of (Clipped) SGD in Non-Convex Optimization

Aleksandar Armacki ALEKSANDAR.ARMACKI@EPFL.CH
École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

Dragana Bajović DBAJOVIC@UNS.AC.RS
Faculty of Technical Sciences, University of Novi Sad, Novi Sad, Serbia

Dušan Jakovetić DUSAN.JAKOVETIC@DMI.UNS.AC.RS
Faculty of Sciences, University of Novi Sad, Novi Sad, Serbia

Soummya Kar SOUMMYAK@ANDREW.CMU.EDU
Carnegie Mellon University, Pittsburgh, Pennsylvania, USA

Ali H. Sayed ALI.SAYED@EPFL.CH
École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

Editors: Steve Hanneke and Tor Lattimore

Abstract

The study of tail behaviour of **SGD**-induced processes has been attracting a lot of interest, due to offering strong guarantees with respect to individual runs of an algorithm. While many works provide high-probability guarantees, quantifying the error rate for a fixed probability threshold, there is a lack of work directly studying the probability of failure, i.e., quantifying the tail decay rate for a fixed error threshold. Moreover, existing results are of finite-time nature, limiting their ability to capture the true long-term tail decay which is more informative for modern learning models, typically trained for millions of iterations. Our work closes these gaps, by studying the long-term tail decay of **SGD**-based methods through the lens of large deviations theory, establishing several strong results in the process. First, we provide an upper bound on the tails of the gradient norm-squared of the best iterate produced by (vanilla) **SGD**, for non-convex costs and bounded noise, with long-term decay at rate $e^{-\frac{t}{\log(t)}}$. Next, we relax the noise assumption by considering clipped **SGD** (**c−SGD**) under heavy-tailed noise with bounded moment of order $p \in (1,2]$, showing an upper bound with long-term decay at rate $e^{-\frac{t^{\beta_p}}{\log(t)}}$, where $\beta_p = \frac{4(p-1)}{3p-2}$ for $p \in (1,2)$ and $e^{-\frac{t}{\log^2(t)}}$ for $p = 2$. Finally, we provide lower bounds on the tail decay, at rate e^{-t}, showing that our rates for both **SGD** and **c−SGD** are tight, up to poly-logarithmic factors. Notably, our results demonstrate *an order of magnitude faster* long-term tail decay compared to existing work based on finite-time bounds, which show rates $e^{-\sqrt{t}}$ and $e^{-t^{\beta_p/2}}$, $p \in (1,2]$, for **SGD** and **c−SGD**, respectively. As such, we uncover regimes where the tails decay much faster than previously known, providing stronger long-term guarantees for individual runs.

Keywords: non-convex, sgd, clipping, heavy-tails, tail bounds, large deviations, lower bounds

1. Introduction

Non-convex optimization is an integral part of modern machine learning (ML), as many practical settings, such as training large language models (LLMs), e.g., Zhang et al. (2022b), represent instances of the general problem of minimizing a non-convex cost $f : \mathbb{R}^d \mapsto \mathbb{R}$, given by

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

In practice, the problem (1) is solved using iterative methods, typically based on the stochastic gradient descent (**SGD**) algorithm Robbins and Monro (1951), making the study of convergence guarantees of **SGD**-based methods an integral part of ML theory. While classical guarantees, like mean-squared error (MSE) convergence, e.g., Sayed (2023), offer important indicators of average performance across many runs, the advent of large-scale models, such as transformers and LLMs, has shifted the focus on guarantees with respect to an *individual run* of an algorithm, as even a single training run of such models can be incredibly costly, both time- and resource-wise. This has led to an increased interest in *tail probabilities* and results of type $\mathbb{P}\big(F_t > \epsilon_t\big) \leq \delta_t$, where $F_t \coloneqq \min_{k\in[t]} \|\nabla f(x_k)\|^2$ is the standard metric of interest for non-convex costs, $\{x_t\}_{t\in\mathbb{N}}$ is the sequence of models generated by an iterative method, while $\epsilon_t > 0$ and $\delta_t \in (0, 1]$ are the *error* and *probability thresholds*, respectively. Conventional results quantify the tail behaviour for a *fixed probability threshold* and *decaying error threshold*, i.e., the goal is to show $\mathbb{P}\big(F_t > c(\delta)/n_t\big) \leq \delta$, for any fixed $\delta \in (0, 1)$, where $n_t \to \infty$ is the *decay rate* and $c(\delta)$ is a function of the probability threshold. Of particular interest are high-probability (HP) convergence results, where $c(\delta) = \log(1/\delta)$, e.g., Nemirovskiĭ et al. (2009); Ghadimi and Lan (2013), which can be used to establish sharp *tail bounds*, i.e., *exponentially decaying probability threshold* for a *fixed error threshold*, of the form $\mathbb{P}(F_t > \epsilon) \leq e^{-n_t\epsilon}$, for any fixed $\epsilon > 0$ (obtained from HP bounds by solving $\frac{\log(1/\delta)}{n_t} = \epsilon$ for δ). Tail bounds for fixed error thresholds are important, as in practice the goal is to reach a *ϵ-stationary point*, i.e., a point $x \in \mathbb{R}^d$ that satisfies $\|\nabla f(x)\| \leq \epsilon$, e.g., Arjevani et al. (2022), hence tail bounds quantify the *probability of failure* (i.e., not reaching an ϵ-stationary point)[1] for an individual run.

Another important observation is that classical tail bounds are of *finite-time* nature, designed to hold *for every* $t \in \mathbb{N}$. While useful for smaller models, trained for a moderate number of iterations (e.g., $t = 10^4$-10^5), such results might not be too informative for modern large-scale models that typically require a huge number of training iterations, e.g., `AlexNet` and `ResNet-50` are both trained for approximately $t = 5 \times 10^5$ iterations Krizhevsky et al. (2012); He et al. (2016), the original BERT model is trained for $t = 10^6$ iterations Devlin et al. (2019), while LLMs are estimated to require many millions of training iterations, see, e.g., Hoffmann et al. (2022). As such, bounds capturing the long-term tail behaviour, which holds *for t sufficiently large*, are more appropriate for models requiring an enormous number of training iterations. Although finite-time results derived for a fixed probability threshold, like HP guarantees, can imply exponentially decaying tail bounds, as we show later, they do so in a conservative manner, resulting in very loose bounds in the regime of large t. Motivated by these observations, we consider the following important question:

What is the sharpest long-term tail decay achievable for **SGD***-type methods in non-convex optimization, for a given, fixed error threshold*?

To answer this question, we take a large deviations principle (LDP) approach, e.g., Dembo and Zeitouni (2009), by directly studying the long-term tail behaviour of F_t, for a fixed error threshold. In Table 1 we provide an overview of long-term tail decay rates for **SGD**-type methods, where it can be seen that our results show *an order of magnitude faster* long-term tail decay rate of F_t than implied by prior results based on finite-time bounds, indicating that a strictly sharper tail decay, not properly captured by existing works, is achievable in the long run. We next state our full contributions.

1. It can be easily seen that if $\min_{k\in[t]} \|\nabla f(x_k)\| > \epsilon$, then an ϵ-stationary point has not been reached up to time t.

Table 1: Long-term tail decay of **SGD**-based methods in non-convex optimization. *Method* specifies the variant of **SGD**; *Cost* states the assumptions on the cost function; *Noise* provides the noise assumptions; *Decay rate* is the largest positive sequence $n_t \to \infty$ such that, for any $\epsilon > 0$ and some $C_\epsilon > 0$, we have $\limsup_{t\to\infty} n_t^{-1} \log \mathbb{P}(\min_{k\in[t]} \|\nabla f(x_k)\|^2 > \epsilon) \leq -C_\epsilon$.[§] The decay rates for Liu et al. (2023a); Cutkosky and Mehta (2021); Nguyen et al. (2023) stem from their finite-time HP bounds, whereas the decay rates from Armacki et al. (2026a,b)[†] and our work are obtained by establishing an LDP upper bound with a full rate function, see Section 2.2 for details.

WORK	METHOD	COST	NOISE	DECAY RATE
LIU ET AL. (2023A)	**SGD**	SMOOTH, BOUNDED FROM BELOW	SUB-GAUSSIAN	$\sqrt{t}$
THIS WORK (THEOREM 3)	**SGD**	SMOOTH, BOUNDED FROM BELOW, BOUNDED GRADIENTS	A.S. BOUNDED	$t/\log(t)$
CUTKOSKY AND MEHTA (2021)	**NORMALIZED c-SGD**	SMOOTH, BOUNDED FROM BELOW, BOUNDED GRADIENTS	BOUNDED MOMENT OF ORDER $p \in (1,2]$	$t^{\frac{2(p-1)}{3p-2}}$
NGUYEN ET AL. (2023)	**c-SGD**	SMOOTH, BOUNDED FROM BELOW	BOUNDED MOMENT OF ORDER $p \in (1,2]$	$t^{\frac{2(p-1)}{3p-2}} / \log^{\frac{2p}{3p-2}}(t)$
ARMACKI ET AL. (2026A,B)	**N-SGD**[‡]	SMOOTH, BOUNDED FROM BELOW	SYMMETRIC PDF, POSITIVE AROUND ORIGIN	$\sqrt{t}/\log(t)$
THIS WORK (THEOREM 6)	**c-SGD**	SMOOTH, BOUNDED FROM BELOW, BOUNDED GRADIENTS	BOUNDED MOMENT OF ORDER $p \in (1,2]$	$t^{\frac{4(p-1)}{3p-2}} / \log(t)$[¶]

[§] Although some works included in the table provide bounds on different quantities (e.g., $\frac{1}{t}\sum_{k=1}^t \|\nabla f(x_k)\|^2$), all of them imply a bound on $\min_{k\in[t]} \|\nabla f(x_k)\|^2$. As discussed in Section 3 and Appendix H, our results can be equivalently stated in terms of $\frac{1}{t}\sum_{k=1}^t \|\nabla f(x_k)\|^2$.

[†] While Armacki et al. (2026a) provide HP bounds for **N-SGD**, their results can be used to get LDP upper bounds, see Armacki (2025) for details.

[‡] **N-SGD** is a general nonlinear **SGD** framework which, for state-dependent noise, among others includes clipping, normalization, smooth sign and smooth component-wise clipping. If the noise is also independent, identically distributed (IID), in addition to the previous, the **N-SGD** framework includes non-smooth component-wise nonlinearities, like standard sign and component-wise clipping.

[¶] For the special case $p = 2$, our decay rate incurs an additional log factor, i.e., is of the form $n_t = t/\log^2(t)$, see Theorem 6 for details.

1.1. Contributions

Motivated by the need for sharp bounds on the failure probability of individual runs in applications like training modern large-scale models, we study the long-term tail behaviour of iterates induced by **SGD**-type methods through the lens of large deviations (LD) theory. Our contributions are as follows.

- We provide a sharp characterization of the failure probability of **SGD**-type methods in non-convex optimization, by studying the long-term tail behaviour for a fixed error threshold. First, we consider vanilla **SGD** under almost surely (a.s.) bounded noise and establish an LDP upper bound on F_t, with an exponential long-term tail decay, at rate $n_t = t/\log(t)$. In Table 1 we summarize existing long-term tail decay results. We can see that the decay rate for **SGD** in our work is an order of magnitude faster than the $n_t = \sqrt{t}$ rate resulting from finite-time HP result in Liu et al. (2023a).

- Next, we relax the noise condition, by considering clipped **SGD** (**c-SGD**) under heavy-tailed noise with bounded moment of order $p \in (1,2]$. We provide an LDP upper bound on F_t, for an appropriately chosen clipping threshold, with long-term exponential tail decay at rate $n_t = t/\log^2(t)$ for $p = 2$ and $n_t = t^{\frac{4(p-1)}{3p-2}}/\log(t)$ for $p \in (1,2)$. We can again see in Table 1 that the long-term tail decay rate for **c-SGD** in our work is an order of magnitude faster than the $n_t = t^{\frac{2(p-1)}{3p-2}}$ rate resulting from finite-time HP bounds in Cutkosky and Mehta (2021); Nguyen et al. (2023).

- Finally, we show that the long-term decay rates established in our LDP upper bounds are tight, by providing matching finite-time (and asymptotic) lower bounds on the tail probability induced by both vanilla **SGD** and **c-SGD**. To do so, we carefully construct an instance of cost, noise, model initialization and error threshold, under which we show that the tails induced by both methods

exhibit exponentially decaying lower bounds at rate $n_t = t$, demonstrating that our long-term upper bounds for **SGD** and **C-SGD** (when $p = 2$) are tight, up to poly-logarithmic factors.

As such, our results show that the long-term tail probability induced by **SGD**-type methods in non-convex optimization decays at a rate that is both tight and significantly faster than previously known, leading to much sharper bounds on the probability of failure in non-convex problems (i.e., not reaching an ϵ-stationary point) and stronger guarantees for individual runs of an algorithm.

1.2. Literature Review

We next review the literature on finite-time high-probability (HP) and asymptotic LD results for **SGD**-based methods. For an overview of other popular types of guarantees, see Appendix B.

High-probability guarantees. Initial HP results consider light-tailed noise and include Nemirovskiĭ et al. (2009); Lan (2012); Ghadimi and Lan (2013); Hazan et al. (2015); Harvey et al. (2019); Li and Orabona (2020); Liu and Zhou (2024), with Liu et al. (2023a) providing HP convergence of F_t for vanilla **SGD** and non-convex costs, with order-optimal rate $\mathcal{O}\left(\frac{\log(t/\delta)}{\sqrt{t}}\right)$. More recently, HP convergence of nonlinear **SGD** methods (e.g., clipping, sign, normalization) under noise with heavier tails has attracted attention, starting with Gorbunov et al. (2020); Parletta et al. (2024), who consider noise with bounded variance, while Li and Liu (2022); Eldowa and Paudice (2024); Madden et al. (2024) consider sub-Weibull noise. This is extended by Cutkosky and Mehta (2021); Nguyen et al. (2023); Sadiev et al. (2023); Liu et al. (2023b); Hübler et al. (2025); Kornilov et al. (2025), who consider various nonlinear **SGD** methods under noise with bounded moment of order $p \in (1, 2]$, and Armacki et al. (2025, 2026a), who consider a unified nonlinear **SGD** framework (dubbed **N-SGD**) under noise with symmetric probability density function (PDF), positive around zero and potentially unbounded moments. Among them, Cutkosky and Mehta (2021); Nguyen et al. (2023) show that F_t achieves the optimal rate $\mathcal{O}\left(\log(1/\delta)t^{\frac{2(1-p)}{3p-2}}\right)$ using **C-SGD**,[2] while Armacki et al. (2026a) show that **N-SGD** converges with rate $\mathcal{O}\left(\frac{\log(t/\delta)}{\sqrt{t}}\right)$, matching the rate in Liu et al. (2023a) established under light tails.[3] Translated into tail bounds, it follows that Liu et al. (2023a); Armacki et al. (2026a) imply an asymptotic exponential tail decay for vanilla **SGD** under light-tailed noise and general nonlinear **SGD** under noise with symmetric PDF, respectively, with decay rate $n_t = \sqrt{t}$. Similarly, Cutkosky and Mehta (2021); Nguyen et al. (2023) imply a long-term exponential tail decay of F_t for **C-SGD** under bounded p-th moment noise, with decay rate $n_t = t^{\frac{2(p-1)}{3p-2}}$.

Large deviations guarantees. LD studies have a long history, see, e.g., Varadhan (2008); Dembo and Zeitouni (2009) and references therein, with a wide range of applications, including statistical mechanics Ellis (2005); Touchette (2009), distributed detection Bajović et al. (2011, 2012); Braca et al. (2014); Matta et al. (2016a,b), social learning Bordignon et al. (2021); Bajović (2024); Matta et al. (2025) and general ML Braca et al. (2022); Lindhe (2023). In the context of **SGD**-type methods,

2. Technically, Cutkosky and Mehta (2021) consider a variant of momentum **SGD**, using both clipping and normalization.
3. Hübler et al. (2025); Kornilov et al. (2025) show that normalized and sign **SGD** match the oracle complexity of clipped **SGD**, using an increasing batch size. Since we study the long-term behaviour, this implies an infinite batch size as $t \to \infty$, hence we focus our comparison on works that use a fixed batch size. Next, while the rate in Armacki et al. (2026a) is better than the one in Cutkosky and Mehta (2021); Nguyen et al. (2023) for any $p < 2$, it does not invalidate the optimality of their rate, as the two are derived under different conditions on the noise.

LDs are studied in Hu et al. (2019); Bajović et al. (2023); Jongeneel et al. (2024); Azizian et al. (2024, 2025); Gürbüzbalaban et al. (2025); Armacki et al. (2026b). Hu et al. (2019); Azizian et al. (2024, 2025) use a LD approach to study the behaviour of **SGD** iterates with fixed step-size for non-convex problems, in the limit as the step-size goes to zero. Hu et al. (2019) show that iterates can escape local minimizers in a number of iterations exponentially depending on the inverse of the step-size, while Azizian et al. (2024, 2025) show that the iterates concentrate around local minima and establish a full LDP for the time it takes to reach a global minima, also exponentially depending on the inverse of the step-size.[4] Bajović et al. (2023) provide an LDP upper bound for the last iterate of **SGD** and strongly convex costs, while Jongeneel et al. (2024) extend this result to costs satisfying the PŁ condition, with applications to reinforcement learning. Gürbüzbalaban et al. (2025) study a class of generalized momentum methods for strongly convex costs, establishing an LDP upper bound for the average cost sub-optimality. Finally, Armacki et al. (2026b) consider non-convex costs and a general nonlinear **SGD** framework dubbed **N-SGD**, under noise with symmetric PDF, positive around zero and unbounded moments, showing an LDP upper bound for F_t, with decay rate $n_t = \sqrt{t}/\log(t)$. As mentioned, Hu et al. (2019); Azizian et al. (2024, 2025) study the tail behaviour of iterates of **SGD** for non-convex costs using a fixed step-size, in the limit as the step-size goes to zero. On the other hand, we study the tail behaviour of F_t and consider the more natural variant of **SGD**-based methods using a time-varying step-size, in the limit as the number of iterations goes to infinity.

Technical challenges and novelty. In order to establish our results, we needed to overcome a number of challenges. First, to show LDP-type results and accelerated tail decay rates, we provide tight bounds on the moment-generating function (MGF) of F_t and use the Gärtner-Ellis theorem (see Proposition 9 in Appendix D). Next, in order to be able to apply the Gärtner-Ellis theorem, we needed to show that the MGF is finite everywhere, which is significantly stronger than the finite-time HP analysis, where it suffices to show that the MGF is bounded locally, e.g., over a compact interval or even at a single point (see Section 4 for a more detailed discussion). To resolve this challenge, we consider uniformly bounded noise for **SGD**, while for **C-SGD** we show via a careful analysis and tuning of the clipping threshold that the MGF is bounded everywhere, even under general heavy-tailed noise. Finally, we provide a lower bound on the tail probability, by carefully constructing an instance of cost, noise and model initialization for which the tails of F_t decay at least exponentially fast and which is of independent interest in HP-type studies.

Paper organization. Section 2 provides preliminaries, Section 3 presents the main results, Section 4 provides comparison with existing works, Section 5 concludes the paper, while Appendix contains results omitted from the main body. The remainder of this section introduces some notation.

Notation. We denote by $\mathbb{N}, \mathbb{R}$ and $\mathbb{R}^d$ the sets of positive integers, real numbers and d-dimensional real vectors. For any $m \in \mathbb{N}$, we denote by $[m]$ the set $[m] = \{1, 2, \ldots, m\}$. The Euclidean inner product and induced norm are denoted by $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$. For a set B, we denote by B° and $\overline{B}$ its topological interior and closure. We use $o(\cdot)$ and $\mathcal{O}(\cdot)$ as the standard "little o" and "big O", i.e., for two positive sequences $\{a_t\}_{t\in\mathbb{N}}, \{b_t\}_{t\in\mathbb{N}}$, such that $\lim_{t\to\infty} a_t = \lim_{t\to\infty} b_t = \infty$, we say that $a_t = o(b_t)$ ($a_t = \mathcal{O}(b_t)$), if $\lim_{t\to\infty} \frac{a_t}{b_t} = 0$ ($\limsup_{t\to\infty} \frac{a_t}{b_t} < \infty$), unless stated otherwise.

4. A full LDP means that matching lower and upper bounds are provided, see Section 2.2 for details.

Algorithm 1 Vanilla and clipped **SGD**

Require: Model initialization $x_1 \in \mathbb{R}^d$, step-size $\{\alpha_t\}_{t\in\mathbb{N}}$, clipping threshold $\{\gamma_t\}_{t\in\mathbb{N}}$;

for $t = 1, 2, \ldots$ **do**
  Query the $\mathcal{SFO}$ with input x_t and obtain g_t;
  Perform the following model update:
    Option 1: $x_{t+1} = x_t - \alpha_t g_t$; (Vanilla **SGD**)
    Option 2: $x_{t+1} = x_t - \alpha_t \min\Big\{1, \frac{\gamma_t}{\|g_t\|}\Big\} g_t$; (Clipped **SGD**)
end for

2. Preliminaries

In this section we provide the preliminaries. Subsection 2.1 specifies the oracle model and reviews the vanilla and clipped **SGD** methods, while Subsection 2.2 provides a primer on LD theory.

2.1. The Oracle Model and SGD-based Methods

We assume access to a Stochastic First-order Oracle ($\mathcal{SFO}$), which, when queried with input x, returns a stochastic estimate g of the gradient $\nabla f(x)$. The $\mathcal{SFO}$ subsumes the following paradigms.

1. Batch (i.e., offline) learning: for a finite dataset $\{\xi^i\}_{i\in[m]}$ and loss $\ell : \mathbb{R}^d \times \Xi \mapsto \mathbb{R}$, the cost is given by $f(x) = \frac{1}{m}\sum_{i\in[m]} \ell(x;\xi^i)$. When queried, the $\mathcal{SFO}$ chooses a sample of indices $S \subset [m]$ uniformly at random and outputs $g = \frac{1}{|S|}\sum_{j\in S} \nabla\ell(x;\xi^j)$, where $1 \leq |S| < m$.

2. Streaming (i.e., online) learning: for a random variable ξ following an unknown distribution $\mathcal{D}$, the cost is given by $f(x) = \mathbb{E}_{\xi\sim\mathcal{D}}\big[\ell(x;\xi)\big]$. When queried, the $\mathcal{SFO}$ generates a mini-batch $\{\xi^j\}_{j\in S}$ of IID copies of ξ and outputs $g = \frac{1}{|S|}\sum_{j\in S} \nabla\ell(x;\xi^j)$, where $|S| \geq 1$.

Next, we describe the two methods considered in our work, namely (vanilla) **SGD** and **C−SGD**. The general update rule for iterative **SGD**-based methods can be represented as

$$x_{t+1} = x_t - \alpha_t \Psi_t(g_t), \tag{2}$$

where $\alpha_t > 0$ is the step-size, while $\Psi_t : \mathbb{R}^d \mapsto \mathbb{R}^d$ is a (possibly) nonlinear mapping. The first method considered in our work, **SGD**, where Ψ_t is the linear identity map $\Psi_t(x) \equiv x$, is perhaps the most well-known and widely used algorithm, lauded for its ease of implementation and strong performance. However, the advent of deep learning and LLMs resulted in exponentially increasing model complexity and phenomena such as heavy-tailed noise and exploding gradients, e.g., Pascanu et al. (2013); Simsekli et al. (2019); Zhang et al. (2020c); Gurbuzbalaban et al. (2021), requiring nonlinear modifications to **SGD**, with clipping being a very popular choice. It is known to bring many benefits, like stabilizing and accelerating training Zhang et al. (2020b), ensuring convergence under heavy-tailed noise Sadiev et al. (2023) and providing differential privacy Zhang et al. (2022c). The resulting method, **C−SGD**, is widely used for LLM training Zhang et al. (2022b); Touvron et al. (2023); Liu et al. (2024), and represents an instance of (2) with $\Psi_t(x) = \min\Big\{1, \frac{\gamma_t}{\|x\|}\Big\}x$, for a user-specified clipping threshold $\gamma_t > 0$. The two methods are summarized in Algorithm 1.

2.2. Large Deviations Principle: a Background

The goal of LD studies is to quantify the long-term probability of (rare) events, by providing sharp, exponentially decaying long-term tail bounds. In particular, for a process of interest $\{F_t\}_{t\in\mathbb{N}}$, the LDP aims to find a lower semi-continuous function $I : \mathbb{R} \mapsto [0, \infty]$ and a positive sequence $\{n_t\}_{t\in\mathbb{N}}$ satisfying $\lim_{t\to\infty} n_t = \infty$, such that, for any (Borel) measurable $B \subseteq \mathbb{R}$

$$-\inf_{x\in B^\circ} I(x) \leq \liminf_{t\to\infty} \frac{1}{n_t} \log \mathbb{P}(F_t \in B) \leq \limsup_{t\to\infty} \frac{1}{n_t} \log \mathbb{P}(F_t \in B) \leq -\inf_{x\in\overline{B}} I(x). \tag{3}$$

If (3) holds, the process $\{F_t\}_{t\in\mathbb{N}}$ is said to satisfy the *(full) LDP*, with decay rate n_t and rate function I, see, e.g., Dembo and Zeitouni (2009). The relation (3) provides a tight characterization of the asymptotic, long-term behaviour of F_t. While the full LDP is desirable, the lower bound in (3) is often difficult to obtain. Instead, one typically aims to establish (only) the upper bound, in which case $\{F_t\}_{t\in\mathbb{N}}$ is said to satisfy the *LDP upper bound*. Note that the upper bound in (3) implies

$$\mathbb{P}(F_t \in B) \leq e^{-n_t \inf_{x\in\overline{B}} I(x) + o(n_t)} \approx e^{-n_t \inf_{x\in\overline{B}} I(x)}, \tag{4}$$

where the second relation holds for all t sufficiently large, hence the LDP upper bound alone is a very strong indicator of the long-term behaviour of F_t, establishing exponentially decaying long-term probability of F_t ending up in any set B, such that $\inf_{x\in\overline{B}} I(x) > 0$.

3. Main Results

In this section we provide the main results. Subsection 3.1 states the assumptions, Subsection 3.2 provides results for **SGD**, Subsection 3.3 presents results for **c−SGD** under heavy-tailed noise, while Subsection 3.4 establishes a (nearly) matching lower bound on the tail probability.

3.1. Assumptions

We start by stating a technical condition on the model initialization, used for analysis purposes.

Assumption 1 *The model initialization $x_1 \in \mathbb{R}^d$ is selected in a deterministic manner.*

Assumption 1 allows the initialization to be any real vector, as long as it is a deterministic quantity.

Assumption 2 *The cost is bounded from below, has uniformly bounded gradients and is L-smooth, i.e., it holds that $f_\star := \inf_{x\in\mathbb{R}^d} f(x) > -\infty$ and for some $G > 0$ and any $x, y \in \mathbb{R}^d$, we have*

$$\|\nabla f(x)\| \leq G \;\text{ and }\; f(x) \leq f(y) + \langle \nabla f(y), x - y\rangle + \frac{L}{2}\|x - y\|^2.$$

Assumption 2 specifies conditions on the cost. Boundedness from below and L-smoothness are standard for general non-convex costs, e.g., Ghadimi and Lan (2013). The bounded gradient condition is also widely used for non-convex costs, e.g., Berkenkamp et al. (2017); Mertikopoulos et al. (2020); Cutkosky and Mehta (2021) and is satisfied, e.g., by any G-Lipschitz continuous cost, which includes a wide class of convolutional and deep neural networks, e.g., Fazlyab et al. (2019); Zou et al. (2020); Combettes and Pesquet (2020); Zhang et al. (2022a), as well as transformer-based models, e.g., Kim et al. (2021). For ease of notation, let $z_t := g_t - \nabla f(x_t)$ denote the gradient noise and let $\mathcal{F}_t := \sigma(\{x_1, \ldots, x_t\})$ be the natural filtration, with $\mathcal{F}_1 := \sigma(\{\emptyset, \Omega\})$ being the trivial σ-algebra. The next two assumptions state the noise regimes considered in our work.

Assumption 3 *The gradient estimator is unbiased and the noise is uniformly bounded, i.e., for all $t \geq 1$ and some $M > 0$, it holds that $\mathbb{E}\big[g_t \mid \mathcal{F}_t\big] = \nabla f(x_t)$ and $\|z_t\| \leq M$, a.s.*

Assumption 4 *The gradient estimator is unbiased and the noise has bounded moment of order $p \in (1, 2]$, i.e., $\mathbb{E}\big[g_t \mid \mathcal{F}_t\big] = \nabla f(x_t)$ and $\mathbb{E}\big[\|z_t\|^p \mid \mathcal{F}_t\big] \leq \sigma^p$, a.s., for all $t \geq 1$ and some $\sigma > 0$.*

Assumption 3 is often used in the context of adaptive methods and relaxed smoothness, e.g., Harvey et al. (2019); Zhang et al. (2020b,a); Li et al. (2023b); Carmon and Hinder (2024). For example, in batch learning, where $f(x) = \frac{1}{m}\sum_{i\in[m]} \ell(x;\xi^i)$, the $\mathcal{SFO}$ estimator automatically satisfies Assumption 3 if ℓ has bounded gradients (see Appendix C for details). Assumption 4 is the standard heavy-tailed noise condition, e.g., Nguyen et al. (2023); Sadiev et al. (2023); Hübler et al. (2025). Moreover, many works empirically show that noise satisfying Assumption 4 frequently arises during training of neural networks and transformers, across a wide range of model architectures and datasets, see, e.g., Simsekli et al. (2019); Zhang et al. (2020c) and references therein.

3.2. Large Deviations Principle Upper Bound for SGD

In this subsection, we establish an LDP upper bound on $F_t = \min_{k\in[t]} \|\nabla f(x_k)\|^2$. Before stating the main theorem, we define an important concept and provide a useful technical result.

Definition 1 *A random vector $z \in \mathbb{R}^d$ is said to be σ-sub-Gaussian if $\mathbb{E}\Big[\exp\Big(\frac{\|z\|^2}{\sigma^2}\Big)\Big] \leq \exp(1)$.*

Sub-Gaussian (i.e., light-tailed) noise is widely used for deriving HP convergence, e.g., Ghadimi and Lan (2013); Li and Orabona (2020); Liu et al. (2023a). We then have the following result.

Lemma 2 *Let Assumption 3 hold. Then the following are true, for any $t \geq 1$.*

1. *The noise is M-sub-Gaussian, i.e., we have $\mathbb{E}\Big[\exp\Big(\frac{\|z_t\|^2}{M^2}\Big) \mid \mathcal{F}_t\Big] \leq \exp(1)$.*
2. *For any $\mathcal{F}_t$-measurable vector $x \in \mathbb{R}^d$, we have $\mathbb{E}\big[\exp\big(\langle x, z_t\rangle\big) \mid \mathcal{F}_t\big] \leq \exp\Big(\frac{3M^2\|x\|^2}{4}\Big)$.*

Lemma 2 provides some useful bounds on the MGF of the noise. We next state the main result.

Theorem 3 *Let Assumptions 1-3 hold and let $\{x_t\}_{t\in\mathbb{N}}$ be the sequence generated by SGD, using the step-size $\alpha_t = \frac{a}{\sqrt{t+1}}$, where $a \leq \frac{1}{L}$. Then the sequence $\{F_t\}_{t\in\mathbb{N}}$ satisfies an LDP upper bound, with decay rate $n_t = \frac{t}{\log(t)}$ and rate function $I_v : \mathbb{R} \mapsto [0,\infty]$, i.e., for any measurable set $B \subseteq \mathbb{R}$*

$$\limsup_{t\to\infty} \frac{\log(t)}{t} \log \mathbb{P}(F_t \in B) \leq -\inf_{x\in\overline{B}} I_v(x),$$

where the rate function is given by $I_v(x) = \begin{cases} \frac{x^2}{24M^2G^2}, & x \geq 0 \\ +\infty, & x < 0 \end{cases}$.

For the special case $B = (\epsilon, \infty)$, where $\epsilon > 0$, we immediately have the following corollary.

Corollary 4 *Let conditions of Theorem 3 hold. We then have, for any* $\epsilon > 0$

$$\limsup_{t\to\infty} \frac{\log(t)}{t} \log \mathbb{P}(F_t > \epsilon) \leq -\frac{\epsilon^2}{24M^2G^2}.$$

Theorem 3 and Corollary 4 establish the long-term tail decay for **SGD** under bounded noise, at rate $e^{-\frac{t}{\log(t)}}$. We next discuss several different aspects of our results.

On the rate function. Note that the rate function I_v has two distinct regimes: for $x \geq 0$, it is of the form $I_v(x) = \frac{x^2}{24M^2G^2}$, while for any $x < 0$, it takes the value $I_v(x) = +\infty$. In practice, this means that for any set $B \subseteq \mathbb{R}$ whose closure contains a non-negative value, there is a probability that F_t visits the set, but it decays exponentially at rate $t/\log(t)$. On the other hand, for any set $B \subseteq (-\infty, -a)$, for $a > 0$, we can see that $\inf_{x\in\overline{B}} I(x) = +\infty$, hence $\limsup_{t\to\infty} \frac{\log(t)}{t} \log \mathbb{P}(F_t \in B) = -\infty$, implying that $\mathbb{P}(F_t \in B) = 0$ in the long run, which is expected, as $F_t \geq 0$.

Dependence on problem parameters. We can see that the rate function depends on two problem parameters, the noise and gradient bounds M and G, decreasing as either grows. This is again expected, as the convergence of **SGD** slows down with more noise (measured by M) and a more complex cost (measured by G), resulting in a smaller leading constant in the tail decay rate, meaning that it takes more time for F_t to escape a set B. Finally, while the leading term in finite-time bounds depends on parameters like the initial optimality gap $f(x_1) - f_\star$ and smoothness L (see Section 4 for details), our results indicate that these parameters do not affect the asymptotic decay.

On the metric. While our results are presented in terms of $F_t = \min_{k\in[t]} \|\nabla f(x_k)\|^2$, they continue to hold for the average norm-squared, i.e., $A_t = \frac{1}{t}\sum_{k=1}^t \|\nabla f(x_k)\|^2$, which is more general, as $F_t \leq A_t$. In our proofs, the results are first established in terms of A_t, which then implies the same bounds on F_t, while the choice of presenting the results in terms of F_t stems from it being standard in non-convex optimization and its ease of interpretability, see Appendix H for details.

On the noise. Compared to existing HP results for vanilla **SGD**, e.g., Liu et al. (2023a), which consider sub-Gaussian noise, we impose a slightly stronger condition of a.s. bounded noise. As discussed in the introduction, this is a byproduct of the fact that we need to bound the MGF $\mathbb{E}[\exp(\lambda F_t)]$ over the entire domain, i.e., for every $\lambda \in \mathbb{R}$, while HP results only require bounding the MGF locally, e.g., over a compact domain, or at $\lambda = 1$ (see the discussion in Section 4 for details). As we show in the next subsection, this condition can be significantly relaxed, by applying a (bounded) nonlinear operator to **SGD**, ensuring that the effective noise remains bounded.

3.3. Large Deviations Principle Upper Bound Under Heavy-Tailed Noise

In this subsection, we relax the noise assumption by considering the **c−SGD** method and provide an LDP upper bound under heavy-tailed noise. For ease of notation, let $\widetilde{g}_t := \min\left\{1, \frac{\gamma_t}{\|g_t\|}\right\} g_t$ denote the clipped stochastic gradient and let $\theta_t^u := \widetilde{g}_t - \mathbb{E}\left[\widetilde{g}_t \mid \mathcal{F}_t\right]$ and $\theta_t^b := \mathbb{E}\left[\widetilde{g}_t \mid \mathcal{F}_t\right] - \nabla f(x_t)$ respectively be the unbiased and biased components of the difference of the clipped stochastic gradient and the true gradient (i.e., *clipping bias*), noting that $\widetilde{g}_t - \nabla f(x_t) = \theta_t^u + \theta_t^b$. Decomposing the clipping bias into an unbiased and biased component is standard when analyzing **c−SGD**, see, e.g., Sadiev et al. (2023); Nguyen et al. (2023). We then have the following important result.

Lemma 5 *Let Assumptions 2 and 4 hold and let the clipping threshold be chosen as*

$$\gamma_t = \begin{cases} 2G(t+1)^{\frac{2-p}{6p-4}}, & p \in (1,2) \\ 2G\sqrt{\log(t+1)}, & p = 2. \end{cases} \tag{5}$$

Then the following are true, for any $t \geq 1$.

1. $\|\theta_t^b\| \leq 4\sigma^p \gamma_t^{1-p}$.
2. *For any $\mathcal{F}_t$-measurable $x \in \mathbb{R}^d$, we have* $\mathbb{E}\big[\exp\left(\langle x, \theta_t^u \rangle\right) \mid \mathcal{F}_t\big] \leq \exp\left(3\gamma_t^2 \|x\|^2\right)$.

Lemma 5 provides a bound on the biased component and establishes sub-Gaussian concentration of the unbiased one, facilitating the rest of our analysis. We note that γ_t can be tuned without requiring knowledge of G, with the choice in (5) simplifying the exposition (see the discussion after Theorem 6 and Appendix I for more details). We next state the main result.

Theorem 6 *Let Assumptions 1, 2 and 4 hold and let $\{x_t\}_{t\in\mathbb{N}}$ be the sequence generated by* **c−SGD** *using the step-size $\alpha_t = (t+1)^{-\frac{p}{3p-2}}$ and clipping threshold γ_t given in (5). Then the sequence $\{F_t\}_{t\in\mathbb{N}}$ satisfies an LDP upper bound, with the following decay rate and rate function.*

1. *If $p \in (1,2)$, the decay rate is $n_t = \frac{t^{\frac{4(p-1)}{3p-2}}}{\log(t)}$, with rate function given by* $I_c(x) = \begin{cases} \frac{x^2}{768G^4}, & x \geq 0 \\ +\infty, & x < 0. \end{cases}$
2. *If $p = 2$, the decay rate is $n_t = \frac{t}{\log^2(t)}$, with rate function given by* $I_c(x) = \begin{cases} \frac{x^2}{384G^4}, & x \geq 0 \\ +\infty, & x < 0. \end{cases}$

Similarly to the previous section, if $B = (\epsilon, \infty)$, where $\epsilon > 0$, we have the following result.

Corollary 7 *Let conditions of Theorem 6 hold. Then the following are true, for any $\epsilon > 0$.*

1. *If $p \in (1,2)$, then* $\limsup_{t\to\infty} \frac{\log(t)}{t^{\beta_p}} \log \mathbb{P}(F_t > \epsilon) \leq -\frac{\epsilon^2}{768G^4}$, *where* $\beta_p = \frac{4(p-1)}{3p-2}$.
2. *If $p = 2$, then* $\limsup_{t\to\infty} \frac{\log^2(t)}{t} \log \mathbb{P}(F_t > \epsilon) \leq -\frac{\epsilon^2}{384G^4}$.

Theorem 6 and Corollary 7 establish the long-term tail decay for **c−SGD** under heavy-tailed noise, notably showing the rate $e^{-t/\log^2(t)}$ for noise with bounded variance. We next discuss the results.

On the decay rate. We can see that the decay rate in Theorem 6 has two distinct regimes: for $p \in (1,2)$, the decay rate is of order $e^{-t^{\beta_p}/\log(t)}$, where $\beta_p = \frac{4(p-1)}{3p-2}$, while for $p = 2$, we incur an additional $\log(t)$ factor, showing the decay rate $e^{-t/\log^2(t)}$. This is consistent with MSE and HP convergence results established under Assumption 4, in the sense that the convergence rate exponent explicitly depends on the noise moment p, see, e.g., Zhang et al. (2020c); Nguyen et al. (2023).

On the rate function and dependence on problem parameters. Similarly to the discussion in the previous subsection, the rate function I_c has two distinct regimes for negative and non-negative values. On the other hand, while it exhibits dependence on problem parameters, it does so only through the gradient bound G. We note that this stems from our choice of clipping threshold in (5) and the fact that the unbiased component of the clipping bias is γ_t-sub-Gaussian (recall Lemma 5).

On the metric. Similarly to the previous subsection, our results can be equally stated in terms of the average norm-squared of the gradients, $\frac{1}{t}\sum_{k=1}^{t}\|\nabla f(x_k)\|^2$, see Appendix H for details.

On the clipping threshold. We use an increasing clipping threshold in (5), which is consistent with existing works Cutkosky and Mehta (2021); Nguyen et al. (2023). However, compared to Cutkosky and Mehta (2021); Nguyen et al. (2023), who use the clipping threshold $\gamma_t = \widetilde{\mathcal{O}}\left(t^{\frac{1}{3p-2}}\right)$,[5] our clipping threshold increases at a strictly slower rate (e.g., for $p=2$ our threshold increases at rate $\sqrt{\log(t)}$, while in the said works it increases at rate $t^{1/4}$), making the likelihood of clipping higher, further closing the gap on how clipping is used in practice.[6] Next, we assume knowledge of noise moment p and gradient bound G to tune the clipping threshold in (5), which is on par with Cutkosky and Mehta (2021), while Nguyen et al. (2023) relax the bounded gradient condition, at the expense of require knowledge of noise moment p and parameter σ, the initial optimality gap $f(x_1)-f_\star$ and smoothness constant L. Finally, we note that knowledge of G is not necessary in (5) for our results to hold. In particular, our results continue to hold if the clipping threshold is selected as

$$\gamma_t = \begin{cases} C(t+1)^{\frac{2-p}{6p-4}}, & p\in(1,2) \\ C\sqrt{\log(t+1)}, & p=2, \end{cases}$$

where $C>0$ is any value. Using this threshold, it can be shown that the resulting rate function will be of the form $I_c(x)=\mathcal{O}\left(\frac{x^2}{C^2G^2}\right)$ for $x\geq 0$, and $I_c(x)=+\infty$ otherwise, where $\mathcal{O}(\cdot)$ hides global constants. The reader is referred to Appendix I for a formal statement and derivations.

3.4. A Lower Bound on the Tail Probability

In this subsection, we show that the results in Theorems 3 and 6 are tight, by establishing a lower bound on the tail probability induced by **SGD**/**c−SGD**. To that end, we have the following result.

Theorem 8 *There exist a cost, initialization and $\mathcal{SFO}$ obeying Assumptions 1-3, as well as a problem dependent constant $b>0$ and global constants $a_1,a_2>0$, such that, for any $\epsilon\in(0,b)$ and all $t\geq 1$, the tail probability induced by either* **SGD** *or* **c−SGD**, *satisfies* $\mathbb{P}(F_t>\epsilon)\geq a_1 e^{-a_2 t}$.

It readily follows that Theorem 8 implies the following asymptotic lower bound, for any $\epsilon\in(0,b)$

$$\liminf_{t\to\infty}\frac{1}{t}\log\mathbb{P}(F_t>\epsilon)\geq -a_2.$$

Theorem 8 shows that there exist hard problem instances for which the tail probability of both **SGD** and **c−SGD** *can not decay faster than* e^{-t}. We next discuss the result from several perspectives.

On the tightness of decay rates. Combining the results of Theorem 8 with Corollary 4, it follows that for any $\epsilon\in(0,b)$, the tails induced by **SGD** satisfy

$$-a_2\leq\liminf_{t\to\infty}\frac{1}{t}\log\mathbb{P}(F_t>\epsilon)\ \text{ and }\ \limsup_{t\to\infty}\frac{\log(t)}{t}\log\mathbb{P}(F_t>\epsilon)\leq-\frac{\epsilon^2}{24M^2G^2}. \tag{6}$$

5. Here, we use $\widetilde{\mathcal{O}}(\cdot)$ to hide problem related constants and terms poly-logarithmic in t.
6. Contrary to the increasing threshold used in theory, in practice clipping is used with a small, constant threshold, see, e.g., Zhang et al. (2022b); Touvron et al. (2023); Liu et al. (2024).

Equation (6) demonstrates that the long-term tail decay rate of **SGD** in Theorem 3 is tight up to a $\log(t)$ factor. Similarly, combined with Corollary 7 for $p = 2$, it can be seen that for any $\epsilon \in (0, b)$

$$-a_2 \leq \liminf_{t\to\infty} \frac{1}{t} \log \mathbb{P}(F_t > \epsilon) \text{ and } \limsup_{t\to\infty} \frac{\log^2(t)}{t} \log \mathbb{P}(F_t > \epsilon) \leq -\frac{\epsilon^2}{384G^4}, \tag{7}$$

underlining the tightness of the asymptotic tail decay rate of **c−SGD** for $p = 2$, up to a $\log^2(t)$ factor.

On the rate function optimality. While (6) and (7) indicate the the tightness of the decay rates, they are weaker than a full LDP of the form in (3), in the sense that they do not establish matching upper and lower bounds on the asymptotic probability of failure. As such, it is unclear if the rate functions in Theorems 3 and 6 depend optimally on the problem parameters. However, as we discuss in the next section, our rate functions exhibit dependence on similar problem parameters as the corresponding decay constants obtained from finite-time high-probability bounds.

4. Comparison with State-Of-The-Art

In this section, we provide detailed comparison with state-of-the-art (SOTA) long-term tail decay results for **SGD**-type methods in non-convex optimization stemming from either HP or LDP guarantees, as well as a further discussion on the differences in the assumptions.

SOTA results for SGD. Liu et al. (2023a) provide SOTA finite-time HP convergence guarantees of **SGD**, for a L-smooth, bounded from below cost, and B-sub-Gaussian noise. Using the time-varying step-size $\alpha_t = \frac{1}{L\sqrt{t}}$, they show the following result, for any $\delta \in (0, 1)$ and any $t \geq 1$

$$\mathbb{P}\left(F_t > \frac{2\Delta L + 3B^2(1 + \log(t)) + 12B^2 \log(1/\delta)}{\sqrt{t}}\right) \leq \delta,$$

where $\Delta \coloneqq f(x_1) - f_\star$ is the optimality gap of the initial model. The above bound leads to the following tail result $\mathbb{P}\big(F_t > \epsilon\big) \leq \exp\left(-\frac{\epsilon\sqrt{t}}{12B^2} + \frac{\Delta L}{6B^2} + \frac{1+\log(t)}{4}\right)$, for any $\epsilon > 0$ and any $t \geq 1$. Taking the logarithm, dividing everything by $\sqrt{t}$ and taking the $\limsup$, we get

$$\limsup_{t\to\infty} \frac{1}{\sqrt{t}} \log \mathbb{P}\big(F_t > \epsilon\big) \leq -\frac{\epsilon}{12B^2}. \tag{8}$$

Comparing (8) to the bound in Corollary 4, we can see that both results depend on the noise (via B and M), with our result further depending on G. More importantly, the long-term decay rate in (8) is of the order $\sqrt{t}$, while the one in Corollary 4 is of the order $\frac{t}{\log(t)}$, an order of magnitude faster. This improvement stems from the different approach taken in our work, focusing directly on long-term tail bounds, while on the other hand, Liu et al. (2023a) focus on finite-time HP results, that hold for any t, but result in loose long-term bounds. Finally, we note that while our rate function depends on the gradient bound G, which can be large in some applications, this becomes negligible relative to the gain in the decay rate, especially in the long-term regime $t \to \infty$ considered in our work.

SOTA results for c−SGD. Nguyen et al. (2023) provide SOTA finite-time HP convergence guarantees of **c−SGD**, for a L-smooth, bounded from below cost, and noise with bounded p-th moment.

Using the time varying step-size $\alpha_t = \widetilde{\mathcal{O}}\Big(t^{-\frac{p}{3p-2}}\Big)$ and clipping threshold $\gamma_t = \widetilde{\mathcal{O}}\Big(t^{\frac{1}{3p-2}}\Big)$,[7] they show the following result, for any $\delta \in (0, 1/e)$ and $t \geq 1$

$$\mathbb{P}\left(F_t > \frac{720\sigma\sqrt{\Delta L}\log^{\frac{2p}{3p-2}}(t)\log(1/\delta)}{t^{\beta_p/2}}\right) \leq \delta,$$

where we recall that $\beta_p = \frac{4(p-1)}{3p-2}$.[8] The above bound leads to the following tail result $\mathbb{P}\big(F_t > \epsilon\big) \leq \exp\Big(-\frac{\epsilon t^{\beta_p/2}}{720\sigma\sqrt{\Delta L}\log^{2p/(3p-2)}(t)}\Big)$, for any $\epsilon > 0$ and any $t \geq 1$. Taking the logarithm, dividing everything by $t^{\beta_p/2}/\log^{\frac{2p}{3p-2}}(t)$ and taking the $\limsup$, we get

$$\limsup_{t\to\infty} \frac{\log^{\frac{2p}{3p-2}}}{t^{\beta_p/2}} \log \mathbb{P}\big(F_t > \epsilon\big) \leq -\frac{\epsilon}{720\sigma\sqrt{\Delta L}}. \tag{9}$$

Comparing (9) to the bound in Corollary 7, we can see that our result depends on G, while the result in (9) depends on the noise, initial optimality gap and smoothness (with the latter two stemming from the choice of the clipping threshold in Nguyen et al. (2023)). Ignoring the $\log$ factors, we can again see that the long-term decay rate in (9) is of order $t^{\beta_p/2}$, with the one in Corollary 7 being of order t^{β_p}, an order of magnitude faster. Similarly, Nguyen et al. (2023) provide bounds that hold for any t and do not require bounded gradients, however, the resulting long-term tail bounds are very loose. Importantly, the finite-time HP bound established in Nguyen et al. (2023) matches that of **c−SGD** from Cutkosky and Mehta (2021), who additionally require bounded gradients. As such, the assumptions used in Cutkosky and Mehta (2021) are the same as in our work, while their tail decay rate is the same as in Nguyen et al. (2023) and an order of magnitude slower than ours. This further highlights that the improved decay rates shown in our work are not simply a result of stronger assumptions, but of a *fundamentally different approach in studying the long-term tail decay*.

Finally, Armacki et al. (2026b) provide an LDP upper bound for a family of nonlinear **SGD** methods, which, among others, includes clipping. Under L-smooth, lower bounded cost, noise with a symmetric PDF that is strictly positive around the origin, using the step-size $\alpha_t = \frac{1}{\sqrt{t+1}}$ and a constant threshold $\gamma_t = C$, the authors show that

$$\limsup_{t\to\infty} \frac{\log(t)}{\sqrt{t}} \log \mathbb{P}(F_t > \epsilon) \leq -\frac{\min\{\epsilon, \sqrt{\epsilon}\}}{16C^4L^2}, \tag{10}$$

implying a long-term tail decay rate $\frac{\sqrt{t}}{\log(t)}$, which is strictly worse than the rate in Corollary 7 for any $p > 6/5$. Crucially, the LDP upper bound in Armacki et al. (2026b) is derived for a black-box nonlinear framework, under very different noise conditions, namely IID noise with symmetric PDF.

On the assumptions. As was already mentioned, we require uniformly bounded noise for vanilla **SGD**, which is stronger than the sub-Gaussian condition used in Liu et al. (2023a). This stronger requirement stems from the fact that, in order to invoke the Gärtner-Ellis theorem, we need to show that the *scaled MGF is bounded everywhere*, i.e., that $\limsup_{t\to\infty} \frac{1}{n_t}\log \mathbb{E}[\exp(n_t\lambda F_t)] < \infty$, for all $\lambda \in \mathbb{R}$ and $n_t = \frac{t}{\log(t)}$. This is much stronger than the bound for HP results, where it typically

7. Here $\widetilde{\mathcal{O}}(\cdot)$ hides global and problem related constants, as well as terms poly-logarithmic in t and $1/\delta$.
8. The bound from Nguyen et al. (2023) is simplified, for ease of presentation. It can be shown that the tail decay rate stemming from their bounds actually deteriorates to a slower one as $p \to 1$, which we omit and present the faster rate.

suffices to show $\mathbb{E}[\exp(F_t)] \leq \exp\left(\mathcal{O}\left(\sum_{k=1}^t \alpha_k^2\right)\right)$, i.e., for $n_t = \lambda = 1$, making it much easier to control the MGF. This stronger requirement necessitates uniformly bounded noise for vanilla **SGD**, for the following reason. Starting from (17) in the Appendix, pushing M^2 inside the sum and replacing it by $\|z_k\|^2$, we have $F_t = A_t + C\sum_{k=1}^t \alpha_k^2\|z_k\|^2$. Ignoring the term A_t and assuming that the noise is sub-Gaussian, we then get $\mathbb{E}[\exp(n_t\lambda F_t)] \leq \mathbb{E}[\exp(n_t\lambda\sum_{k=1}^t \alpha_k^2\|z_k\|^2)] = \mathbb{E}[\exp(n_t\lambda M^2\sum_{k=1}^t \alpha_k^2\|z_k\|^2/M^2)]$. To control this term, we need $h_{t,\lambda} := n_t\alpha_k^2\lambda M^2 \leq 1$ to hold for all $\lambda \in \mathbb{R}$ and $t \geq k \geq 1$. Recalling that $n_t = \frac{t}{\log(t)}$ and λ can be arbitrarily large, one can see that $h_{t,\lambda}$ is unbounded, necessitating uniformly bounded noise. This is not required in the analysis of **c−SGD**, as the resulting operator is bounded, facilitating heavy-tailed noise. Similarly, the gradient bound is required to control the terms $\langle\nabla f(x_t), z_t\rangle$ and $\langle\nabla f(x_t), \theta_t^u\rangle$ for **SGD** and **c−SGD**, respectively. While we believe that it might be possible to fully remove the bounded gradient condition for **c−SGD**, it is beyond the scope of the current work and represents an important future direction.

5. Conclusion

We studied the long-term tail decay and probability of failure of **SGD**-type methods in non-convex optimization, demonstrating that, in the long run, the tails decay at an order of magnitude faster rate than suggested by existing works. Further, we show that our results are tight, by establishing lower bounds on the tail probability, which match our upper bounds up to poly-logarithmic factors. As such, our results provide much sharper bounds on the probability of failure in non-convex optimization, implying stronger guarantees for individual runs of **SGD**-based methods in applications that require an enormous number of iterations, such as training of deep learning models and LLMs. Future work includes removing assumptions like the uniformly bounded gradients for **c−SGD**, as well as establishing a full LDP with matching lower tail decay rate and rate function.

Acknowledgments

The work of A. Armacki and S. Kar is partially supported by NSF, Grant No. 2330196. The work of D. Bajović is supported by the European Union's Horizon Europe Research and Innovation program under grant agreement No. 101135916. The work of D. Jakovetić was supported by the Ministry of Science, Technological Development and Innovation of the Republic of Serbia (Grants No. 451-03-33/2026-03/200125 & 451-03-34/2026-03/200125), and by the Science Fund of the Republic of Serbia, Grant No. 7359, Project title LASCADO.

References

Yossi Arjevani, Yair Carmon, John C Duchi, Dylan J Foster, Nathan Srebro, and Blake Woodworth. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, 199(1): 165–214, 2022. ISSN 1436-4646. doi: 10.1007/s10107-022-01822-7. URL https://doi.org/10.1007/s10107-022-01822-7.

Aleksandar Armacki. *High-Probability and Large Deviations Techniques for Design and Analysis of Large-Scale and Distributed Learning Systems*. PhD thesis, Carnegie Mellon University, 2025. URL https://www.proquest.com/dissertations-theses/high-probability-large-deviations-techniques/docview/3238247992/se-2.

Aleksandar Armacki, Dragana Bajović, Dušan Jakovetić, and Soummya Kar. Gradient Based Clustering. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 929–947. PMLR, 2022. URL https://proceedings.mlr.press/v162/armacki22a.html.

Aleksandar Armacki, Shuhua Yu, Pranay Sharma, Gauri Joshi, Dragana Bajović, Dušan Jakovetić, and Soummya Kar. High-probability Convergence Bounds for Online Nonlinear Stochastic Gradient Descent under Heavy-tailed Noise. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pages 1774–1782. PMLR, 2025. URL https://proceedings.mlr.press/v258/armacki25a.html.

Aleksandar Armacki, Dragana Bajović, Dušan Jakovetić, and Soummya Kar. Sharp High-Probability Rates for Nonlinear SGD under Heavy-Tailed Noise via Symmetrization. *IEEE Transactions on Information Theory*, pages 1–23, 2026a. doi: 10.1109/TIT.2026.3682577.

Aleksandar Armacki, Shuhua Yu, Dragana Bajović, Dušan Jakovetić, and Soummya Kar. Large Deviation Upper Bounds and Improved MSE Rates of Nonlinear SGD: Heavy-Tailed Noise and Power of Symmetry. *SIAM Journal on Optimization*, 36(1):32–59, 2026b. doi: 10.1137/24M1704154. URL https://doi.org/10.1137/24M1704154.

Waïss Azizian, Franck Iutzeler, Jerome Malick, and Panayotis Mertikopoulos. What is the Long-Run Distribution of Stochastic Gradient Descent? A Large Deviations Analysis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 2168–2229. PMLR, 2024. URL https://proceedings.mlr.press/v235/azizian24a.html.

Waïss Azizian, Franck Iutzeler, Jerome Malick, and Panayotis Mertikopoulos. The Global Convergence Time of Stochastic Gradient Descent in Non-Convex Landscapes: Sharp Estimates via Large Deviations. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 1982–2044. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/azizian25a.html.

Dragana Bajović. Inaccuracy Rates for Distributed Inference Over Random Networks With Applications to Social Learning. *IEEE Transactions on Information Theory*, 70(1):415–435, 2024. doi: 10.1109/TIT.2023.3324866.

Dragana Bajović, Dus˘an Jakovetić, João Xavier, Bruno Sinopoli, and José M. F. Moura. Distributed Detection via Gaussian Running Consensus: Large Deviations Asymptotic Analysis. *IEEE Transactions on Signal Processing*, 59(9):4381–4396, 2011. doi: 10.1109/TSP.2011.2157147.

Dragana Bajović, Dušan Jakovetić, José M. F. Moura, João Xavier, and Bruno Sinopoli. Large Deviations Performance of Consensus+Innovations Distributed Detection With Non-Gaussian Observations. *IEEE Transactions on Signal Processing*, 60(11):5987–6002, 2012. doi: 10.1109/TSP.2012.2210885.

Dragana Bajović, Dušan Jakovetić, and Soummya Kar. Large deviations rates for stochastic gradient descent with strongly convex functions. In *Proceedings of The 26th International Conference on*

Artificial Intelligence and Statistics, volume 206 of *Proceedings of Machine Learning Research*, pages 10095–10111. PMLR, 2023. URL https://proceedings.mlr.press/v206/bajovic23a.html.

Felix Berkenkamp, Matteo Turchetta, Angela Schoellig, and Andreas Krause. Safe Model-based Reinforcement Learning with Stability Guarantees. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/766ebcd59621e305170616ba3d3dac32-Paper.pdf.

Dimitri P. Bertsekas and John N. Tsitsiklis. Gradient Convergence in Gradient methods with Errors. *SIAM Journal on Optimization*, 10(3):627–642, 2000. doi: 10.1137/S1052623497331063. URL https://doi.org/10.1137/S1052623497331063.

Virginia Bordignon, Vincenzo Matta, and Ali H. Sayed. Adaptive Social Learning. *IEEE Transactions on Information Theory*, 67(9):6053–6081, 2021. doi: 10.1109/TIT.2021.3094633.

Paolo Braca, Stefano Marano, Vincenzo Matta, and Ali H. Sayed. Large deviations analysis of adaptive distributed detection. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6112–6116, 2014. doi: 10.1109/ICASSP.2014.6854778.

Paolo Braca, Leonardo M. Millefiori, Augusto Aubry, Stefano Marano, Antonio De Maio, and Peter Willett. Statistical Hypothesis Testing Based on Machine Learning: Large Deviations Analysis. *IEEE Open Journal of Signal Processing*, 3:464–495, 2022. doi: 10.1109/OJSP.2022.3232284.

Yair Carmon and Oliver Hinder. The Price of Adaptivity in Stochastic Convex Optimization. In *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pages 772–774. PMLR, 2024. URL https://proceedings.mlr.press/v247/carmon24a.html.

Jianshu Chen and Ali H. Sayed. Distributed Pareto Optimization via Diffusion Strategies. *IEEE Journal of Selected Topics in Signal Processing*, 7(2):205–220, 2013. doi: 10.1109/JSTSP.2013.2246763.

Patrick L. Combettes and Jean-Christophe Pesquet. Lipschitz Certificates for Layered Network Structures Driven by Averaged Activation Operators. *SIAM Journal on Mathematics of Data Science*, 2(2):529–557, 2020. doi: 10.1137/19M1272780. URL https://doi.org/10.1137/19M1272780.

Ashok Cutkosky and Harsh Mehta. High-probability Bounds for Non-Convex Stochastic Optimization with Heavy Tails. In *Advances in Neural Information Processing Systems*, volume 34, pages 4883–4895. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/26901debb30ea03f0aa833c9de6b81e9-Paper.pdf.

Amir Dembo and Ofer Zeitouni. *Large Deviations Techniques and Applications*. Springer Berlin, Heidelberg, 2 edition, 2009. ISBN 978-3-642-03310-0.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 4171–4186. Association for Computational Linguistics, 2019. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

DLMF. *NIST Digital Library of Mathematical Functions*. https://dlmf.nist.gov/, Release 1.2.5 of 2025-12-15, 2025. URL https://dlmf.nist.gov/. F. W. J. Olver, A. B. Olde Daalhuis, D. W. Lozier, B. I. Schneider, R. F. Boisvert, C. W. Clark, B. R. Miller, B. V. Saunders, H. S. Cohl, and M. A. McClain, eds.

Khaled Eldowa and Andrea Paudice. General Tail Bounds for Non-Smooth Stochastic Mirror Descent. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 3205–3213. PMLR, 2024. URL https://proceedings.mlr.press/v238/eldowa24a.html.

Richard Ellis. *Entropy, Large Deviations, and Statistical Mechanics*. Springer Berlin, Heidelberg, 11 2005. ISBN 978-3-540-29059-9. doi: 10.1007/3-540-29060-5.

Mahyar Fazlyab, Alexander Robey, Hamed Hassani, Manfred Morari, and George Pappas. Efficient and Accurate Estimation of Lipschitz Constants for Deep Neural Networks. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/95e1533eb1b20a97777749fb94fdb944-Paper.pdf.

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. doi: 10.1137/110848864. URL https://doi.org/10.1137/110848864.

Saeed Ghadimi and Guanghui Lan. Stochastic First- and Zeroth-Order Methods for Nonconvex Stochastic Programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013. doi: 10.1137/120880811. URL https://doi.org/10.1137/120880811.

Eduard Gorbunov, Marina Danilova, and Alexander Gasnikov. Stochastic Optimization with Heavy-Tailed Noise via Accelerated Gradient Clipping. In *Advances in Neural Information Processing Systems*, volume 33, pages 15042–15053. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/abd1c782880cc59759f4112fda0b8f98-Paper.pdf.

Mert Gurbuzbalaban, Umut Simsekli, and Lingjiong Zhu. The Heavy-Tail Phenomenon in SGD. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 3964–3975. PMLR, 2021. URL https://proceedings.mlr.press/v139/gurbuzbalaban21a.html.

Mert Gürbüzbalaban, Yasa Syed, and Necdet Serhat Aybat. Accelerated Gradient Methods with Biased Gradient Estimates: Risk Sensitivity, High-Probability Guarantees, and Large Deviation Bounds. *arXiv preprint arXiv:2509.13628*, 2025.

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

Elad Hazan, Kfir Levy, and Shai Shalev-Shwartz. Beyond Convexity: Stochastic Quasi-Convex Optimization. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper_files/paper/2015/file/934815ad542a4a7c5e8a2dfa04fea9f5-Paper.pdf.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Thomas Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karén Simonyan, Erich Elsen, Oriol Vinyals, Jack Rae, and Laurent Sifre. Training Compute-Optimal Large Language Models. In *Advances in Neural Information Processing Systems*, volume 35, pages 30016–30030. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/c1e2faff6f588870935f114ebe04a3e5-Paper-Conference.pdf.

Wenqing Hu, Chris Junchi Li, Lei Li, and Jian-Guo Liu. On the diffusion approximation of nonconvex stochastic gradient descent. *Annals of Mathematical Sciences and Applications*, 4(1):3–32, 2019.

Peter J. Huber. Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics*, 35(1):73 – 101, 1964. doi: 10.1214/aoms/1177703732. URL https://doi.org/10.1214/aoms/1177703732.

Florian Hübler, Ilyas Fatkhullin, and Niao He. From Gradient Clipping to Normalization for Heavy Tailed SGD. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pages 2413–2421. PMLR, 03–05 May 2025. URL https://proceedings.mlr.press/v258/hubler25a.html.

Dušan Jakovetić, Dragana Bajović, Anit Kumar Sahu, Soummya Kar, Nemanja Milošević, and Dušan Stamenković. Nonlinear Gradient Mappings and Stochastic Optimization: A General Framework with Applications to Heavy-Tail Noise. *SIAM Journal on Optimization*, 33(2):394–423, 2023. doi: 10.1137/21M145896X. URL https://doi.org/10.1137/21M145896X.

Wouter Jongeneel, Daniel Kuhn, and Mengmeng Li. A large deviations perspective on policy gradient algorithms. In *Proceedings of the 6th Annual Learning for Dynamics and Control Conference*, volume 242 of *Proceedings of Machine Learning Research*, pages 916–928. PMLR, 2024. URL https://proceedings.mlr.press/v242/jongeneel24a.html.

Ahmed Khaled and Peter Richtárik. Better Theory for SGD in the Nonconvex World. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=AU4qHN2VkS.

Hyunjik Kim, George Papamakarios, and Andriy Mnih. The Lipschitz Constant of Self-Attention. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5562–5571. PMLR, 2021. URL https://proceedings.mlr.press/v139/kim21i.html.

Nikita Kornilov, Philip Zmushko, Andrei Semenov, Alexander Gasnikov, and Alexander Beznosikov. Sign Operator for Coping with Heavy-Tailed Noise in Non-Convex Optimization: High Probability Bounds Under (L_0, L_1)-Smoothness. *arXiv preprint*, 2025.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

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

Haochuan Li, Jian Qian, Yi Tian, Alexander Rakhlin, and Ali Jadbabaie. Convex and Non-convex Optimization Under Generalized Smoothness. In *Advances in Neural Information Processing Systems*, volume 36, pages 40238–40271. Curran Associates, Inc., 2023a. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/7e8bb8d17bb1cb24dfe972a2f8ff2500-Paper-Conference.pdf.

Haochuan Li, Alexander Rakhlin, and Ali Jadbabaie. Convergence of Adam Under Relaxed Assumptions. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 52166–52196. Curran Associates, Inc., 2023b. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/a3cc50126338b175e56bb3cad134db0b-Paper-Conference.pdf.

Shaojie Li and Yong Liu. High Probability Guarantees for Nonconvex Stochastic Gradient Descent with Heavy Tails. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 12931–12963. PMLR, 2022. URL https://proceedings.mlr.press/v162/li22q.html.

Xiaoyu Li and Francesco Orabona. On the Convergence of Stochastic Gradient Descent with Adaptive Stepsizes. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 983–992. PMLR, 16–18 Apr 2019. URL https://proceedings.mlr.press/v89/li19c.html.

Xiaoyu Li and Francesco Orabona. A high probability analysis of adaptive sgd with momentum. *Workshop on "Beyond first-order methods in ML systems", 37th International Conference on Machine Learning*, 2020.

Adam Lindhe. *Topics on Large Deviations in Artificial Intelligence*. PhD thesis, KTH Royal Institute of Technology, 2023.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. DeepSeek-V3 Technical Report. *arXiv preprint arXiv:2412.19437*, 2024.

Yanli Liu, Yuan Gao, and Wotao Yin. An Improved Analysis of Stochastic Gradient Descent with Momentum. In *Advances in Neural Information Processing Systems*, volume 33, pages 18261–18271. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/d3f5d4de09ea19461dab00590df91e4f-Paper.pdf.

Zijian Liu and Zhengyuan Zhou. Revisiting the Last-Iterate Convergence of Stochastic Gradient Methods. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=xxaEhwC1I4.

Zijian Liu and Zhengyuan Zhou. Nonconvex Stochastic Optimization under Heavy-Tailed Noises: Optimal Convergence without Gradient Clipping. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=NKotdPUc3L.

Zijian Liu, Ta Duy Nguyen, Thien Hang Nguyen, Alina Ene, and Huy Nguyen. High Probability Convergence of Stochastic Gradient Methods. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 21884–21914. PMLR, 2023a. URL https://proceedings.mlr.press/v202/liu23aa.html.

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 *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 2266–2290. PMLR, 2023b. URL https://proceedings.mlr.press/v195/liu23c.html.

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. URL http://jmlr.org/papers/v25/23-0466.html.

Vincenzo Matta, Paolo Braca, Stefano Marano, and Ali H. Sayed. Diffusion-Based Adaptive Distributed Detection: Steady-State Performance in the Slow Adaptation Regime. *IEEE Transactions on Information Theory*, 62(8):4710–4732, 2016a. doi: 10.1109/TIT.2016.2580665.

Vincenzo Matta, Paolo Braca, Stefano Marano, and Ali H. Sayed. Distributed Detection Over Adaptive Networks: Refined Asymptotics and the Role of Connectivity. *IEEE Transactions on Signal and Information Processing over Networks*, 2(4):442–460, 2016b. doi: 10.1109/TSIPN.2016.2613682.

Vincenzo Matta, Virginia Bordignon, and Ali H. Sayed. *Social Learning: Opinion Formation and Decision-Making over Graphs*. Emerald Publishing Limited, 2025. ISBN 978-1-63828-472-7. doi: 10.1561/978-1-63828-473-4. URL https://doi.org/10.1561/978-1-63828-473-4.

Panayotis Mertikopoulos, Nadav Hallak, Ali Kavis, and Volkan Cevher. On the Almost Sure Convergence of Stochastic Gradient Descent in Non-Convex Problems. In *Advances in Neural Information Processing Systems*, volume 33, pages 1117–1128. Curran Associates, Inc.,

2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/0cb5ebb1b34ec343dfe135db691e4a85-Paper.pdf.

Arkadi Nemirovskiĭ, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust Stochastic Approximation Approach to Stochastic Programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009. doi: 10.1137/070704277. URL https://doi.org/10.1137/070704277.

Ta Duy Nguyen, Thien H Nguyen, Alina Ene, and Huy Nguyen. Improved Convergence in High Probability of Clipped Gradient Methods with Heavy Tailed Noise. In *Advances in Neural Information Processing Systems*, volume 36, pages 24191–24222. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/4c454d34f3a4c8d6b4ca85a918e5d7ba-Paper-Conference.pdf.

Daniela Angela Parletta, Andrea Paudice, Massimiliano Pontil, and Saverio Salzo. High probability bounds for stochastic subgradient schemes with heavy tailed noise. *SIAM Journal on Mathematics of Data Science*, 6(4):953–977, 2024. doi: 10.1137/22M1536558. URL https://doi.org/10.1137/22M1536558.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1310–1318. PMLR, 2013.

Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Making Gradient Descent Optimal for Strongly Convex Stochastic Optimization. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pages 1571–1578, 2012.

Herbert Robbins and Sutton Monro. A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951. doi: 10.1214/aoms/1177729586. URL https://doi.org/10.1214/aoms/1177729586.

Abdurakhmon Sadiev, Marina Danilova, Eduard Gorbunov, Samuel Horváth, Gauthier Gidel, Pavel Dvurechensky, Alexander Gasnikov, and Peter Richtárik. High-Probability Bounds for Stochastic Optimization and Variational Inequalities: the Case of Unbounded Variance. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 29563–29648. PMLR, 2023. URL https://proceedings.mlr.press/v202/sadiev23a.html.

Ali H. Sayed. *Inference and Learning from Data: Foundations*, volume 1. Cambridge University Press, 2023. ISBN 978-1-009-21814-6. doi: https://doi.org/10.1017/9781009218146.

Othmane Sebbouh, Robert M Gower, and Aaron Defazio. Almost sure convergence rates for Stochastic Gradient Descent and Stochastic Heavy Ball. In *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 3935–3971. PMLR, 2021. URL https://proceedings.mlr.press/v134/sebbouh21a.html.

Umut Simsekli, Levent Sagun, and Mert Gurbuzbalaban. A Tail-Index Analysis of Stochastic Gradient Noise in Deep Neural Networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5827–5837. PMLR, 2019. URL https://proceedings.mlr.press/v97/simsekli19a.html.

Tao Sun, Xinwang Liu, and Kun Yuan. Revisiting Gradient Normalization and Clipping for Non-convex SGD under Heavy-Tailed Noise: Necessity, Sufficiency, and Acceleration. *Journal of Machine Learning Research*, 26(237):1–42, 2025. URL http://jmlr.org/papers/v26/24-1991.html.

Hugo Touchette. The large deviation approach to statistical mechanics. *Physics Reports*, 478(1): 1–69, 2009. ISSN 0370-1573. doi: https://doi.org/10.1016/j.physrep.2009.05.002. URL https://www.sciencedirect.com/science/article/pii/S0370157309001410.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint arXiv:2302.13971*, 2023.

S. R. S. Varadhan. Large deviations. *The Annals of Probability*, 36(2):397 – 419, 2008. doi: 10.1214/07-AOP348. URL https://doi.org/10.1214/07-AOP348.

Bohang Zhang, Jikai Jin, Cong Fang, and Liwei Wang. Improved Analysis of Clipping Algorithms for Non-convex Optimization. In *Advances in Neural Information Processing Systems*, volume 33, pages 15511–15521. Curran Associates, Inc., 2020a. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/b282d1735283e8eea45bce393cefe265-Paper.pdf.

Bohang Zhang, Du Jiang, Di He, and Liwei Wang. Rethinking Lipschitz Neural Networks and Certified Robustness: A Boolean Function Perspective. In *Advances in Neural Information Processing Systems*, volume 35, pages 19398–19413. Curran Associates, Inc., 2022a. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/7b04ec5f2b89d7f601382c422dfe07af-Paper-Conference.pdf.

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. URL https://openreview.net/forum?id=BJgnXpVYwS.

Jingzhao Zhang, Sai Praneeth Karimireddy, Andreas Veit, Seungyeon Kim, Sashank Reddi, Sanjiv Kumar, and Suvrit Sra. Why are Adaptive Methods Good for Attention Models? In *Advances in Neural Information Processing Systems*, volume 33, pages 15383–15393. Curran Associates, Inc., 2020c. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/b05b57f6add810d3b7490866d74c0053-Paper.pdf.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. OPT: Open Pre-trained Transformer Language Models. *arXiv preprint arXiv:2205.01068*, 2022b.

Xinwei Zhang, Xiangyi Chen, Mingyi Hong, Steven Wu, and Jinfeng Yi. Understanding Clipping for Federated Learning: Convergence and Client-Level Differential Privacy. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 26048–26067. PMLR, 2022c. URL https://proceedings.mlr.press/v162/zhang22b.html.

Dongmian Zou, Radu Balan, and Maneesh Singh. On Lipschitz Bounds of General Convolutional Neural Networks. *IEEE Transactions on Information Theory*, 66(3):1738–1759, 2020. doi: 10.1109/TIT.2019.2961812.

Appendix A. Introduction

The appendix contains results omitted from the main body. Appendix B reviews the literature on other useful performance guarantees, Appendix C discusses when noise satisfying Assumption 3 arises naturally, Appendix D provides some important intermediary results, Appendix E contains the proof of Theorem 3, Appendix F provides the proof of Theorem 6, Appendix G contains the proof of Theorem 8, Appendix H discusses the choice of metric, while Appendix I gives results for **c−SGD** with a clipping threshold that does not requiring knowledge of the gradient bound.

Appendix B. Other Performance Guarantees

In this section we provide a brief overview of other useful performance guarantees encountered in the literature on **SGD**-type methods. Perhaps the most frequent among them are *MSE guarantees*, which characterize the average behaviour across many runs, with classical results establishing convergence under the bounded variance assumption (i.e., $p = 2$ in Assumption 4), e.g., Rakhlin et al. (2012); Ghadimi and Lan (2012, 2013); Liu et al. (2020); Liu and Zhou (2024), with works like Chen and Sayed (2013); Khaled and Richtárik (2023) allowing the second noise moment to grow with the gradient norm and/or optimality gap, while Zhang et al. (2020c); Jakovetić et al. (2023); Hübler et al. (2025); Liu and Zhou (2025); Sun et al. (2025) study MSE guarantees under heavy-tailed noise. Another popular guarantee is *almost sure convergence*, where the goal is to show convergence with probability one, e.g., Bertsekas and Tsitsiklis (2000); Li and Orabona (2019); Mertikopoulos et al. (2020); Sebbouh et al. (2021); Jakovetić et al. (2023); Armacki et al. (2026b). Finally, some recent works focus on *low-probability convergence*, i.e., guarantees of the form $\mathbb{P}\big(F_t > \frac{1}{\delta n_t}\big) \leq \delta$, encountered in the context of convergence of adaptive methods or generalized smoothness, see, e.g., Li et al. (2023a,b). While important in their own rights, none of the said guarantees provide tight bounds on the tail probability, with decay at exponential scale, as is the case with HP and LD-style guarantees.

Appendix C. When Assumption 3 Arises Naturally

In this section, we elaborate on when the noise induced by the $\mathcal{SFO}$ satisfies Assumption 3. As mentioned in Section 3.1, we consider the batch setting, where the cost f is of the form $f(x) = \frac{1}{m}\sum_{i\in[m]} \ell(x;\xi^i)$, for some finite dataset $\{\xi^i\}_{i\in[m]}$ and the loss ℓ has G-bounded gradients, i.e., $\|\nabla\ell(x;\xi^i)\| \leq G$, for all $x \in \mathbb{R}^d$ and every $i \in [m]$ (e.g., satisfied by any G-Lipschitz loss, which, as discussed in Section 3.1, contains for a broad class of Lipschitz continuous deep neural networks and transformers Fazlyab et al. (2019); Zou et al. (2020); Combettes and Pesquet (2020); Zhang et al. (2022a); Kim et al. (2021)). It follows that the gradient of f is given by

$$\nabla f(x) = \frac{1}{m}\sum_{i\in[m]} \nabla\ell(x;\xi^i). \tag{11}$$

As discussed in Section 2.1, when queried in iteration t with input x_t, the $\mathcal{SFO}$ returns the estimator

$$g_t = \frac{1}{|S_t|}\sum_{j\in S_t}\nabla\ell(x_t;\xi^j), \tag{12}$$

where $S_t \subset [m]$ is a set of indices drawn uniformly at random. We now want to verify that g_t satisfies Assumption 3. By the definition of the $\mathcal{SFO}$, we have

$$\mathbb{E}[g_t \mid \mathcal{F}_t] = \frac{1}{|S_t|}\sum_{i\in[m]}\nabla\ell(x_t;\xi^i)\mathbb{P}(i\in S_t \mid \mathcal{F}_t) = \nabla f(x_t), \tag{13}$$

where the last equality follows from the fact that $\mathbb{P}(i \in S_t \mid \mathcal{F}_t) = \frac{|S_t|}{m}$ and (11). Moreover, using (11)-(12) and the triangle inequality, it can be seen that the noise $z_t = g_t - \nabla f(x_t)$ satisfies

$$\|z_t\| \le \|g_t\| + \|\nabla f(x_t)\| \le \frac{1}{|S_t|}\sum_{j\in S_t}\|\nabla\ell(x_t;\xi^j)\| + \frac{1}{m}\sum_{i\in[m]}\|\nabla\ell(x_t;\xi^i)\| \le 2G, \tag{14}$$

where the last inequality follows from the fact that the loss ℓ has G-bounded gradients. Equations (13) and (14) readily imply that Assumption 3 holds for the noise induced by the $\mathcal{SFO}$, as claimed.

Appendix D. Intermediate Results

We start by stating an important result, crucial to establishing LDP-style bounds, known as the Gärnter-Ellis theorem, see, e.g., Dembo and Zeitouni (2009).

Proposition 9 *Let $\Lambda_t : \mathbb{R} \mapsto \mathbb{R}$ be a sequence of log moment-generating functions induced by a sequence of measures $\mu_t : \mathcal{B}(\mathbb{R}) \mapsto [0,1]$, $t \in \mathbb{N}$. If for some $\{n_t\}_{t\in\mathbb{N}}$, such that $n_t > 0$ and $\lim_{t\to\infty} n_t = \infty$, and each $\lambda \in \mathbb{R}$, we have*

$$\limsup_{t\to\infty}\frac{\Lambda_t(n_t\lambda)}{n_t} \le \varphi(\lambda) < \infty,$$

then $\{\mu_t\}_{t\in\mathbb{N}}$ satisfies the LDP upper bound with decay rate n_t and rate function $I : \mathbb{R} \mapsto [0,\infty]$, given by the Fenchel-Legendre transform of φ, i.e., $I(x) = \varphi^\star(x) = \sup_{\lambda\in\mathbb{R}}\{x\lambda - \varphi(\lambda)\}$.

Next, we prove Lemma 2. For completeness, we restate it below.

Lemma 2 *Let Assumption 3 hold. Then the following are true, for any $t \ge 1$.*

1. *The noise is M-sub-Gaussian, i.e., we have $\mathbb{E}\Big[\exp\Big(\frac{\|z_t\|^2}{M^2}\Big) \mid \mathcal{F}_t\Big] \le \exp(1)$.*
2. *For any $\mathcal{F}_t$-measurable vector $x \in \mathbb{R}^d$, we have $\mathbb{E}\big[\exp\big(\langle x, z_t\rangle\big) \mid \mathcal{F}_t\big] \le \exp\Big(\frac{3M^2\|x\|^2}{4}\Big)$.*

Proof The first claim follows directly from Assumption 3 and Definition 1. To prove the second claim, we follow a similar approach to, e.g., Li and Orabona (2020, Lemma 1). For ease of notation, let $y_t := \frac{z_t}{M}$ and note that from the first claim we have

$$\mathbb{E}\big[\exp(\|y_t\|^2) \mid \mathcal{F}_t\big] \le \exp(1). \tag{15}$$

Next, let $x \in \mathbb{R}^d$ be $\mathcal{F}_t$-measurable and assume first that $\|x\| \leq \frac{4}{3}$. Using the inequality $\exp(a) \leq a + \exp(9a^2/16)$, which holds for any $a \in \mathbb{R}$, we then have

$$\begin{aligned}\mathbb{E}\left[\exp\left(\langle x, y_t\rangle\right) \mid \mathcal{F}_t\right] &\leq \mathbb{E}\left[\langle x, y_t\rangle + \exp\left(\frac{9\langle x, y_t\rangle^2}{16}\right) \mid \mathcal{F}_t\right] \overset{(a)}{=} \mathbb{E}\left[\exp\left(\frac{9\langle x, y_t\rangle^2}{16}\right) \mid \mathcal{F}_t\right] \\ &\overset{(b)}{\leq} \mathbb{E}\left[\exp\left(\frac{9\|x\|^2\|y_t\|^2}{16}\right) \mid \mathcal{F}_t\right] \overset{(c)}{\leq} \left(\mathbb{E}\left[\exp\left(\|y_t\|^2\right) \mid \mathcal{F}_t\right]\right)^{9\|x\|^2/16} \\ &\overset{(d)}{\leq} \exp\left(\frac{9\|x\|^2}{16}\right) \leq \exp\left(\frac{3\|x\|^2}{4}\right),\end{aligned}$$

where (a) follows from the facts that x is $\mathcal{F}_t$-measurable and the noise is unbiased, (b) follows from the Cauchy-Schwartz inequality, in (c) we use the fact that $\frac{9\|x\|^2}{16} \leq 1$ and Jensen's inequality, while (d) follows from (15). On the other hand, if $\|x\| > \frac{4}{3}$, we use Young's inequality, i.e., $ab \leq \frac{a^2}{2\epsilon} + \frac{\epsilon b^2}{2}$, with $\epsilon = \frac{4}{3}$, to get

$$\begin{aligned}\mathbb{E}[\exp(\langle x, y_t\rangle) \mid \mathcal{F}_t] &\leq \exp\left(\frac{3\|x\|^2}{8}\right) \mathbb{E}\left[\exp\left(\frac{2\|y_t\|^2}{3}\right) \mid \mathcal{F}_t\right] \\ &\leq \exp\left(\frac{2}{3} + \frac{3\|x\|^2}{8}\right) \leq \exp\left(\frac{3\|x\|^2}{4}\right),\end{aligned}$$

where the second inequality follows from Jensen's inequality and (15), while the third inequality follows from the fact that $\frac{2}{3} < \frac{3\|x\|^2}{8}$, since $\|x\| > \frac{4}{3}$. Combining both cases, we get

$$\mathbb{E}[\exp\left(\langle x, y_t\rangle\right) \mid \mathcal{F}_t] \leq \exp\left(\frac{3\|x\|^2}{4}\right). \tag{16}$$

The claim follows by noting that $\langle x, z_t\rangle = \langle Mx, y_t\rangle$ and applying (16). ■

Prior to proving Lemma 5, we provide a known technical result on the behaviour of the clipping operator, see, e.g., Sadiev et al. (2023, Lemma 5.1).

Proposition 11 *Let $X \in \mathbb{R}^d$ be a random vector and let $\widetilde{X} = \min\left\{1, \frac{\gamma}{\|X\|}\right\} X$ be its clipped version. If $\mathbb{E}[X] = x$, $\mathbb{E}\|x - X\|^p \leq \sigma^p$ for some $p \in (1, 2]$ and $\|x\| \leq \frac{\gamma}{2}$, then*

$$\left\|\mathbb{E}[\widetilde{X}] - x\right\| \leq 4\sigma^p \gamma^{1-p}.$$

We are now ready to prove Lemma 5, which we restate below, for the reader's convenience.

Lemma 5 *Let Assumptions 2 and 4 hold and let the clipping threshold be chosen as*

$$\gamma_t = \begin{cases} 2G(t+1)^{\frac{2-p}{6p-4}}, & p \in (1, 2) \\ 2G\sqrt{\log(t+1)}, & p = 2. \end{cases}$$

Then the following are true, for any $t \geq 1$.

1. $\|\theta_t^b\| \leq 4\sigma^p \gamma_t^{1-p}$.

2. *For any $\mathcal{F}_t$-measurable $x \in \mathbb{R}^d$, we have* $\mathbb{E}\big[\exp\left(\langle x, \theta_t^u\rangle\right) \mid \mathcal{F}_t\big] \leq \exp\left(3\gamma_t^2\|x\|^2\right)$.

Proof To prove the first claim, we use Assumption 2 and the fact that, for any $t \geq 1$

$$\|\nabla f(x_t)\| \leq G \leq \frac{\gamma_t}{2},$$

where the second inequality follows from the choice of clipping threshold in (5). The claim now follows by applying Proposition 11 to $\widetilde{g}_t$. To prove the second claim, note that, by the definition of θ_t^u, we have

$$\|\theta_t^u\| = \|\widetilde{g}_t - \mathbb{E}_t[\widetilde{g}_t]\| \leq \|\widetilde{g}_t\| + \mathbb{E}_t\|\widetilde{g}_t\| \leq 2\gamma_t.$$

The claim now readily follows from Lemma 2, setting $M = 2\gamma_t$. ■

Appendix E. Proof of Theorem 3

Using the L-smoothness and the **SGD** update rule (2) with $\Psi_t(x) = x$, we get

$$\begin{aligned}
f(x_{k+1}) &\leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k\rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2 \\
&\leq f(x_k) - \alpha_k\left(1 - \frac{\alpha_k L}{2}\right)\|\nabla f(x_k)\|^2 - \alpha_k(1-\alpha_k L)\langle \nabla f(x_k), z_k\rangle + \frac{\alpha_k^2 L}{2}\|z_k\|^2 \\
&\leq f(x_k) - \frac{\alpha_k}{2}\|\nabla f(x_k)\|^2 - \alpha_k(1-\alpha_k L)\langle \nabla f(x_k), z_k\rangle + \frac{\alpha_k^2 L M^2}{2},
\end{aligned}$$

where the third inequality follows from Assumption 3 and the choice $a \leq \frac{1}{L}$. Rearranging, summing up the first t iterates and using the fact that the step-sizes are non-increasing, we have

$$\frac{\alpha_t}{2}\sum_{k=1}^{t}\|\nabla f(x_k)\|^2 \leq f(x_1) - f_\star + \sum_{k=1}^{t}\alpha_k(\alpha_k L - 1)\langle \nabla f(x_k), z_k\rangle + \frac{LM^2}{2}\sum_{k=1}^{t}\alpha_k^2. \tag{17}$$

Multiplying both sides in (17) by $\frac{2}{\alpha_t t}$ and using the fact that $\min_{k\in[t]}\|\nabla f(x_k)\|^2 \leq \frac{1}{t}\sum_{k=1}^{t}\|\nabla f(x_k)\|^2$, we get

$$\min_{k\in[t]}\|\nabla f(x_k)\|^2 \leq \frac{2}{\alpha_t t}\left(f(x_1) - f_\star + \sum_{k=1}^{t}\alpha_k(\alpha_k L - 1)\langle \nabla f(x_k), z_k\rangle + \frac{LM^2}{2}\sum_{k=1}^{t}\alpha_k^2\right). \tag{18}$$

Recall that we use the shorthand $F_t = \min_{k\in[t]}\|\nabla f(x_k)\|^2$, let $\Delta \coloneqq f(x_1) - f_\star$ and denote by $\lambda_t \coloneqq n_t\lambda$, where $\lambda \in \mathbb{R}$ and $\{n_t\}_{t\in\mathbb{N}}$ is a positive sequence to be specified later. We now want to bound the log MGF of F_t, denoted by Λ_t, i.e., $\Lambda_t(\lambda) \coloneqq \log\mathbb{E}\left[\exp\left(\lambda F_t\right)\right]$, for any $\lambda \in \mathbb{R}$. First, note that for any $\lambda < 0$, we have $\Lambda_t(\lambda_t) \leq 0$ (since $n_t, F_t \geq 0$), hence

$$\limsup_{t\to\infty}\frac{\Lambda(n_t\lambda)}{n_t} \leq 0. \tag{19}$$

Next, consider any $\lambda \geq 0$. For ease of notation, let $b_t \coloneqq \frac{2\lambda_t}{\alpha_t t}$. From the definition of λ_t and (18), we have

$$\mathbb{E}\big[\exp\left(\lambda_t F_t\right)\big] \leq \mathbb{E}\left[\exp\left(b_t\left(\Delta + \sum_{k=1}^{t}\alpha_k(\alpha_k L - 1)\langle \nabla f(x_k), z_k\rangle + \frac{LM^2}{2}\sum_{k=1}^{t}\alpha_k^2\right)\right)\right]$$
$$= \exp\left(b_t\left(\Delta + \frac{LM^2}{2}\sum_{k=1}^{t}\alpha_k^2\right)\right)\mathbb{E}\left[\exp\left(b_t\sum_{k=1}^{t}\alpha_k(\alpha_k L - 1)\langle \nabla f(x_k), z_k\rangle\right)\right].$$

We now proceed to bound the last term. For ease of notation, let $Z_k \coloneqq \alpha_k(\alpha_k L - 1)\langle \nabla f(x_k), z_k\rangle$ and $\mathbb{E}_k[\cdot] \coloneqq \mathbb{E}[\cdot|\mathcal{F}_k]$. Using Lemma 2, we then get

$$\begin{aligned}
\mathbb{E}\left[\exp\left(b_t\sum_{k=1}^{t}Z_k\right)\right] &= \mathbb{E}\left[\exp\left(b_t\sum_{k=1}^{t-1}Z_k\right)\mathbb{E}_t\left[\exp\left(b_t\alpha_t(\alpha_t L - 1)\langle \nabla f(x_t), z_t\rangle\right)\right]\right] \\
&\leq \mathbb{E}\left[\exp\left(b_t\sum_{k=1}^{t-1}Z_k\right)\exp\left(\frac{3b_t^2\alpha_t^2(1-\alpha_t L)^2M^2\|\nabla f(x_t)\|^2}{4}\right)\right] \\
&\leq \exp\left(\frac{3b_t^2\alpha_t^2M^2G^2}{4}\right)\mathbb{E}\left[\exp\left(b_t\sum_{k=1}^{t-1}Z_k\right)\right] \\
&\leq \ldots \leq \exp\left(\frac{3b_t^2M^2G^2}{4}\sum_{k=1}^{t}\alpha_k^2\right),
\end{aligned}$$

where the second inequality follows from Assumption 2 and the fact that $(1-\alpha_k L)^2 \leq 1$, for all $k \geq 1$. Combining everything, we get

$$\mathbb{E}[\exp(\lambda_t F_t)] \leq \exp\left(\frac{2\lambda_t\Delta}{\alpha_t t} + \frac{3\lambda_t^2M^2G^2}{\alpha_t^2t^2}\sum_{k=1}^{t}\alpha_k^2 + \frac{2\lambda_t LM^2}{\alpha_t t}\sum_{k=1}^{t}\alpha_k^2\right).$$

Taking the logarithm, dividing by n_t and recalling that $\alpha_k = \frac{a}{\sqrt{k+1}}$, we have

$$\frac{\Lambda_t(\lambda_t)}{n_t} \leq \frac{2\sqrt{2}\lambda\Delta}{a\sqrt{t}} + \frac{6n_t\lambda^2M^2G^2}{t}\sum_{k=1}^{t}\frac{1}{k+1} + \frac{2\sqrt{2}a\lambda LM^2}{\sqrt{t}}\sum_{k=1}^{t}\frac{1}{k+1}. \tag{20}$$

Using the Darboux sum approximation, we then get

$$\sum_{k=1}^{t}\frac{1}{k+1} \leq \int_0^t \frac{1}{k+1}dk = \log(t+1).$$

Plugging in (20), it follows that

$$\frac{\Lambda_t(\lambda_t)}{n_t} \leq \frac{2\sqrt{2}\lambda\Delta}{a\sqrt{t}} + \frac{6n_t\lambda^2M^2G^2\log(t+1)}{t} + \frac{2\sqrt{2}a\lambda LM^2\log(t+1)}{\sqrt{t}}.$$

Choosing $n_t = \frac{t}{\log(t)}$ and taking the $\limsup$, we finally get

$$\limsup_{t\to\infty}\frac{\Lambda_t(n_t\lambda)}{n_t} \leq 6\lambda^2M^2G^2. \tag{21}$$

Define the continuous function $\varphi : \mathbb{R} \mapsto [0, \infty)$, given by

$$\varphi(\lambda) = \begin{cases} 6\lambda^2 M^2 G^2, & \lambda \geq 0 \\ 0, & \lambda < 0. \end{cases}$$

From (19) and (21), it follows that, for any $\lambda \in \mathbb{R}$

$$\limsup_{t\to\infty} \frac{\Lambda_t(n_t\lambda)}{n_t} \leq \varphi(\lambda) < +\infty.$$

The proof is then completed by invoking Proposition 9 and noting that the Fenchel-Legendre transform of φ is given by

$$\varphi^\star(x) = \begin{cases} \frac{x^2}{24M^2G^2}, & x \geq 0 \\ +\infty, & x < 0. \end{cases}$$

Appendix F. Proof of Theorem 6

Recall that the clipping bias is decomposed as $\widetilde{g}_k - \nabla f(x_k) = \theta_k^u + \theta_k^b$, where $\theta_k^u = \widetilde{g}_k - \mathbb{E}_k[\widetilde{g}_k]$ and $\theta_k^b = \mathbb{E}_k[\widetilde{g}_k] - \nabla f(x_k)$. Using L-smoothness and the **c-SGD** update rule (2), with $\Psi_t(x) = \min\left\{1, \frac{\gamma_t}{\|x\|}\right\}x$, we then have

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) - \alpha_k\langle \nabla f(x_k), \widetilde{g}_k\rangle + \frac{\alpha_k^2 L}{2}\|\widetilde{g}_k\|^2 \\ &\leq f(x_k) - \alpha_k\|\nabla f(x_k)\|^2 - \alpha_k\langle \nabla f(x_k), \theta_k^u + \theta_k^b\rangle + \frac{\alpha_k^2\gamma_k^2 L}{2} \\ &\leq f(x_k) - \alpha_k\|\nabla f(x_k)\|^2 - \alpha_k\langle \nabla f(x_k), \theta_k^u\rangle + \frac{\alpha_k}{2}\|\nabla f(x_k)\|^2 + \frac{\alpha_k}{2}\|\theta_k^b\|^2 + \frac{\alpha_k^2\gamma_k^2 L}{2} \\ &\leq f(x_k) - \frac{\alpha_k}{2}\|\nabla f(x_k)\|^2 - \alpha_k\langle \nabla f(x_k), \theta_k^u\rangle + 8\alpha_k\sigma^{2p}\gamma_k^{2(1-p)} + \frac{\alpha_k^2\gamma_k^2 L}{2}, \end{aligned}$$

where the third inequality follows from $\langle a, b\rangle \leq \frac{\|a\|^2}{2} + \frac{\|b\|^2}{2}$, while in the fourth inequality we use Lemma 5. Rearranging, summing up the first t iterates and using the fact that the step-sizes are non-increasing, we have

$$\frac{\alpha_t}{2}\sum_{k=1}^t \|\nabla f(x_k)\|^2 \leq f(x_1) - f_\star - \sum_{k=1}^t \alpha_k\langle \nabla f(x_k), \theta_k^u\rangle + \sum_{k=1}^t \left(8\alpha_k\sigma^{2p}\gamma_k^{2(1-p)} + \frac{\alpha_k^2\gamma_k^2 L}{2}\right). \quad (22)$$

Multiplying both sides in (22) by $\frac{2}{\alpha_t t}$ and using the fact that $F_t \leq \frac{1}{t}\sum_{k=1}^t \|\nabla f(x_k)\|^2$, with $F_t = \min_{k\in[t]} \|\nabla f(x_k)\|^2$, we get

$$F_t \leq \frac{2}{\alpha_t t}\left(f(x_1) - f_\star - \sum_{k=1}^t \alpha_k\langle \nabla f(x_k), \theta_k^u\rangle + \sum_{k=1}^t \left(8\alpha_k\sigma^{2p}\gamma_k^{2(1-p)} + \frac{\alpha_k^2\gamma_k^2 L}{2}\right)\right). \quad (23)$$

Let $\Delta \coloneqq f(x_1) - f_\star$ and denote by $\lambda_t \coloneqq n_t\lambda$, where $\lambda \in \mathbb{R}$ and $\{n_t\}_{t\in\mathbb{N}}$ is a positive sequence to be specified later. We again want to bound the log MGF of F_t evaluated at λ_t. First, if $\lambda < 0$, it follows that $\mathbb{E}[\exp(\lambda_t F_t)] \leq 1$, which readily implies

$$\limsup_{t\to\infty} \frac{\Lambda(n_t\lambda)}{n_t} \leq 0. \quad (24)$$

Next, let $\lambda \geq 0$ and denote by $b_t \coloneqq \frac{2\lambda_t}{\alpha_t t}$. From the definition of λ_t and (23), we have

$$\mathbb{E}\big[\exp\left(\lambda_t F_t\right)\big] \leq \exp\left(b_t\left(\Delta + \sum_{k=1}^{t}\left(8\alpha_k\sigma^{2p}\gamma_k^{2(1-p)} + \frac{\alpha_k^2\gamma_k^2 L}{2}\right)\right)\right) \times \mathbb{E}\left[\exp\left(-b_t\sum_{k=1}^{t}\alpha_k\langle\nabla f(x_k), \theta_k^u\rangle\right)\right]. \tag{25}$$

To bound the last term, we successively apply Lemma 5 and use Assumption 2, to get

$$\mathbb{E}\left[\exp\left(-b_t\sum_{k=1}^{t}\alpha_k\langle\nabla f(x_k), \theta_k^u\rangle\right)\right] \leq \exp\left(3b_t^2G^2\sum_{k=1}^{t}\alpha_k^2\gamma_k^2\right). \tag{26}$$

Combining (25) and (26), we get

$$\mathbb{E}[\exp(\lambda_t F_t)] \leq \exp\left(\frac{2\lambda_t}{\alpha_t t}\left(\Delta + \sum_{k=1}^{t}\left(8\alpha_k\sigma^{2p}\gamma_k^{2(1-p)} + \frac{\alpha_k^2\gamma_k^2 L}{2}\right)\right) + \frac{12\lambda_t^2G^2}{\alpha_t^2t^2}\sum_{k=1}^{t}\alpha_k^2\gamma_k^2\right).$$

Taking the logarithm and dividing by n_t, we have

$$\frac{\Lambda_t(\lambda_t)}{n_t} \leq \frac{2\lambda}{\alpha_t t}\left(\Delta + \sum_{k=1}^{t}\left(8\alpha_k\sigma^{2p}\gamma_k^{2(1-p)} + \frac{\alpha_k^2\gamma_k^2 L}{2}\right)\right) + \frac{12n_t\lambda^2G^2}{\alpha_t^2t^2}\sum_{k=1}^{t}\alpha_k^2\gamma_k^2. \tag{27}$$

We now consider two cases with respect to the noise moment condition. First, if $p \in (1,2)$, we set $\alpha_k = (k+1)^{-\frac{p}{3p-2}}$, $\gamma_k = 2G(k+1)^{\frac{2-p}{6p-4}}$ and note that $\alpha_t t \geq t^{\frac{2(p-1)}{3p-2}}/2$. Moreover, using the Darboux sum approximation, it follows that

$$\begin{aligned}\sum_{k=1}^{t}\alpha_k\gamma_k^{2(1-p)} &= 4^{1-p}G^{2(1-p)}\sum_{k=1}^{t}(k+1)^{\frac{-p+(2-p)(1-p)}{3p-2}} \leq 4^{1-p}G^{2(1-p)}\sum_{k=1}^{t}(k+1)^{\frac{2-4p+p^2}{3p-2}} \\ &\leq 4^{1-p}G^{2(1-p)}\int_1^{t+1}k^{\frac{2-4p+p^2}{3p-2}}dk \leq C_1 t^{\frac{p(p-1)}{3p-2}},\end{aligned} \tag{28}$$

where $C_1 = \frac{\sqrt{2}(3p-2)4^{1-p}G^{2(1-p)}}{p(p-1)}$, as well as

$$\sum_{k=1}^{t}\alpha_k^2\gamma_k^2 = 4G^2\sum_{k=1}^{t}(k+1)^{\frac{-2p+2-p}{3p-2}} = 4G^2\sum_{k=1}^{t}\frac{1}{k+1} \leq 4G^2\log(t+1). \tag{29}$$

Plugging (28) and (29) in (27) and choosing $n_t = \frac{t^{\frac{4(p-1)}{3p-2}}}{\log(t)}$, it then follows that

$$\frac{\Lambda_t(\lambda_t)}{n_t} \leq \frac{4\lambda\Delta}{t^{\frac{2(p-1)}{3p-2}}} + \frac{32C_1\sigma^{2p}\lambda}{t^{\frac{(p-1)(2-p)}{3p-2}}} + \frac{8\lambda LG^2\log(t+1)}{t^{\frac{2(p-1)}{3p-2}}} + \frac{192\lambda^2G^4\log(t+1)}{\log(t)}.$$

Noting that $(p-1)(2-p) > 0$ for any $p \in (1,2)$ and taking the $\limsup$, we get

$$\limsup_{t\to\infty}\frac{\Lambda_t(n_t\lambda)}{n_t} \leq 192\lambda^2G^4. \tag{30}$$

Next, if $p=2$, we set $\alpha_k = \frac{1}{\sqrt{k+1}}$, $\gamma_k = 2G\sqrt{\log(k+1)}$ and note that $\alpha_t t \geq \sqrt{\frac{t}{2}}$. Moreover, using Darboux sum approximation, it follows that

$$\sum_{k=1}^{t} \frac{\alpha_k}{\gamma_k^2} = \frac{1}{4G^2}\sum_{k=1}^{t} \frac{1}{\log(k+1)\sqrt{k+1}} = C_2 + \frac{1}{4G^2}\sum_{k=4}^{t} \frac{1}{\log(k+1)\sqrt{k+1}}$$
$$\leq C_2 + \frac{1}{4G^2}\int_4^{t+1} \frac{dk}{\log(k)\sqrt{k}} = C_2 + \frac{1}{4G^2}\int_2^{\sqrt{t+1}} \frac{ds}{\log(s)} \leq C_2 + \frac{\mathrm{li}(\sqrt{t+1})}{4G^2}, \quad (31)$$

where $C_2 = \frac{1}{4G^2}\sum_{k=1}^{3} \frac{1}{\log(k+1)\sqrt{k+1}}$, the fourth (in)equality follows by introducing the substitution $s=\sqrt{k}$, while $\mathrm{li} : (0,\infty) \mapsto \mathbb{R}$ is the logarithmic integral function, given by $\mathrm{li}(x) = \int_0^x \frac{dt}{\log(t)}$. Moreover, we have

$$\sum_{k=1}^{t} \alpha_k^2\gamma_k^2 = 4G^2\sum_{k=1}^{t} \frac{\log(k+1)}{k+1} \leq 4G^2\log(t+1)\sum_{k=1}^{t} \frac{1}{k+1} \leq 4G^2\log^2(t+1). \quad (32)$$

Plugging (31) and (32) in (27), using the fact that $\mathrm{li}(x) = \mathcal{O}\big(\frac{x}{\log(x)}\big)$ for x sufficiently large, see, e.g., Chapter 6 in DLMF, and choosing $n_t = \frac{t}{\log^2(t)}$, it then follows that for t sufficiently large

$$\frac{\Lambda_t(\lambda_t)}{n_t} \leq \frac{2\sqrt{2}\lambda\Delta}{\sqrt{t}} + \frac{16\sqrt{2}\sigma^4\lambda}{\sqrt{t}}\left(C_2 + \mathcal{O}\left(\frac{\sqrt{t+1}}{\log(t+1)}\right)\right)$$
$$+ \frac{4\sqrt{2}\lambda L G^2\log^2(t+1)}{\sqrt{t}} + \frac{96\lambda^2G^4\log^2(t+1)}{\log^2(t)}.$$

Taking the $\limsup$, we finally get

$$\limsup_{t\to\infty} \frac{\Lambda_t(n_t\lambda)}{n_t} \leq 96\lambda^2G^4. \quad (33)$$

Define the function $\varphi_p : [0,\infty) \mapsto [0,\infty)$, given by

$$\varphi_p(\lambda) = \begin{cases} 96\lambda^2G^4, & p=2 \\ 192\lambda^2G^4, & p\in(1,2), \end{cases}$$

and consider the continuous function $\varphi : \mathbb{R}\times(1,2] \mapsto [0,\infty)$

$$\varphi(\lambda,p) = \begin{cases} \varphi_p(\lambda), & \lambda \geq 0 \\ 0, & \lambda < 0. \end{cases}$$

From (24), (30) and (33), it follows that, for any $\lambda\in\mathbb{R}$ and $p\in(1,2]$

$$\limsup_{t\to\infty} \frac{\Lambda_t(n_t\lambda)}{n_t} \leq \varphi(\lambda,p) < +\infty.$$

The proof is complete by invoking Proposition 9 and noting that the Fenchel-Legendre transform of φ_p is given by

$$\varphi_p^\star(x) = \begin{cases} \frac{x^2}{768G^4}, & p\in(1,2) \\ \frac{x^2}{384G^4}, & p=2. \end{cases}$$

Appendix G. Proof of Theorem 8

To prove Theorem 8, we construct a specific instance of (1), which obeys our assumptions and show that there exists a problem related constant $b > 0$ and global constants $a_1, a_2 > 0$, such that, if the iterates $\{x_t\}_{t\in\mathbb{N}}$ are generated by either **SGD** or **c−SGD**, then for any $\epsilon \in (0, b)$ and $t \geq 1$

$$\mathbb{P}(F_t > \epsilon) \geq a_1 e^{-a_2 t}. \tag{34}$$

To that end, consider the Huber cost, e.g., Huber (1964), which is given by

$$f(x) = \begin{cases} \frac{\|x\|^2}{2}, & \|x\| \leq G \\ G\|x\| - \frac{G^2}{2}, & \|x\| > G \end{cases}, \tag{35}$$

for a user-specified threshold $G > 0$. It can be readily seen from the definition in (35) that f is lower bounded by zero and continuously differentiable, with the gradient given by

$$\nabla f(x) = \begin{cases} x, & \|x\| \leq G \\ \frac{Gx}{\|x\|}, & \|x\| > G. \end{cases} \tag{36}$$

It is easy to see from (36) that $\|\nabla f(x)\| \leq G$, for all $x \in \mathbb{R}^d$. Moreover, it can be shown that the gradient of f is Lipschitz continuous, with constant $L = 2$, see, e.g., Lemma B.2 in Armacki et al. (2022). As such, the Huber cost given in (35) satisfies Assumption 2. Next, let $x_1 \in \mathbb{R}^d$ be a deterministic initialization (hence satisfying Assumption 1), such that

$$0 < \|x_1\| \leq G. \tag{37}$$

Finally, consider a $\mathcal{SFO}$ which, when queried with x_t, returns $g_t = \nabla f(x_t) + z_t$, where at each time $t \geq 1$, the noise instances z_t are independent and identically distributed, according to the rule

$$z_t = \begin{cases} x_1, & \text{with probability } \frac{1}{2} \\ -x_1, & \text{with probability } \frac{1}{2} \end{cases}. \tag{38}$$

By definition, the noise is zero-mean. Moreover, using (37), it follows that $\|z_t\| = \|x_1\| \leq G$ almost surely, for all $t \geq 1$, hence the noise satisfies Assumption 3. Consider first the vanilla **SGD** method, with step-size $\alpha_t = \frac{1}{L\sqrt{t+1}} = \frac{1}{2\sqrt{t+1}}$. From (36)-(38), it can be readily verified that

$$x_2 = x_1 - \alpha_1 g_1 = (1-\alpha_1)x_1 - \alpha_1 z_1 = \begin{cases} x_1, & \text{with probability } \frac{1}{2} \\ \left(1 - \frac{1}{\sqrt{2}}\right)x_1, & \text{with probability } \frac{1}{2} \end{cases}.$$

Repeating the argument, using the independence of noise, (38), the **SGD** update and Bayes's rule, it can be inferred that, for any $t \geq 1$

$$\mathbb{P}(x_t = x_{t-1} = \ldots = x_1) = 2^{-t+1}. \tag{39}$$

Denote by B_t the event $B_t := \{\omega : x_t(\omega) = \ldots = x_2(\omega) = x_1\}$ and note that $\mathbb{P}(B_t) \overset{(39)}{=} 2^{1-t}$. Next, using (36)-(37), we have, for any $t \geq 1$, conditioned on B_t

$$F_t = \min_{k\in[t]} \|\nabla f(x_k)\|^2 = \|\nabla f(x_1)\|^2 = \|x_1\|^2. \tag{40}$$

Finally, using (39) and (40), it follows that, for any $\epsilon \in (0, \|x_1\|^2)$

$$\begin{aligned}\mathbb{P}(F_t > \epsilon) \geq \mathbb{P}(\{\omega : F_t(\omega) > \epsilon\} \cap B_t) &= \mathbb{P}(F_t > \epsilon \mid B_t)\mathbb{P}(B_t) \\ &= 2^{-t+1} \underbrace{\mathbb{P}(\|x_1\|^2 > \epsilon)}_{=1} = 2e^{-t\ln 2},\end{aligned}$$

which proves (34), with $b = \|x_1\|^2$, $a_1 = 2$ and $a_2 = \ln 2$. Consider next the **c−SGD** method, using the same step-size $\alpha_t = \frac{1}{2\sqrt{t+1}}$, with any clipping threshold satisfying $\gamma_t \geq 2G$, for all $t \geq 1$ (note that this is clearly satisfied for the clipping threshold specified in (5)). We now want to show, by induction, that for every $t \geq 1$

$$\|x_t\| \leq G, \tag{41}$$

which, together with the choice of oracle, cost and clipping threshold, would imply

$$\|g_t\| \leq \|\nabla f(x_t)\| + \|z_t\| \leq \|x_t\| + G \leq 2G \leq \gamma_t. \tag{42}$$

Equations (41) and (42) indicate that clipping is never performed and **c−SGD** reverts to vanilla **SGD**, implying that the same lower bound shown above holds for **c−SGD**. Condition (41) clearly holds for the case $t = 1$, by the design of the initialization in (37). Therefore, assume that (41) holds for some $t > 1$. As shown in (42), we then know that clipping will not be performed, i.e., $\widetilde{g}_t = g_t$, and clipping reverts to **SGD** in iteration $t + 1$, which implies that

$$\begin{aligned}\|x_{t+1}\| = \|x_t - \alpha_t \widetilde{g}_t\| = \|x_t - \alpha_t g_t\| &\overset{(i)}{=} \|(1-\alpha_t)x_t - \alpha_t z_t\| \\ &\leq (1-\alpha_t)\|x_t\| + \alpha_t\|z_t\| \leq (1-\alpha_t)G + \alpha_t\|x_1\| \overset{(ii)}{\leq} G,\end{aligned}$$

where (i) follows from (36) and the induction hypothesis (41), while (ii) follows from (37). Therefore, we have shown that (41) holds for every $t \geq 1$. Therefore, **c−SGD** reverts to **SGD** and the lower bound established above holds for **c−SGD** as well, completing the proof.

Appendix H. On the Metric

As discussed in the main body, while we use the metric $F_t = \min_{k\in[t]} \|\nabla f(x_k)\|^2$, our results continue to hold for the metric $A_t = \frac{1}{t}\sum_{k=1}^{t} \|\nabla f(x_k)\|^2$, which is stronger, seeing that $F_t \leq A_t$. To see why this is the case, note that in both the proof of Theorem 3 (recall (17)-(18)) and that of Theorem 6 (recall (22)-(23)), we start with a bound on A_t (in (17) and (22), respectively) and use the fact that $F_t \leq A_t$ to switch to the desired metric F_t (in (18) and (23), respectively). It can then be readily seen that skipping the inequality $F_t \leq A_t$, while using the exact same steps in the respective proofs, we would get the same results on the more general metric A_t. Similarly, the results of Theorem 8 hold for A_t, which can be seen by noting that, conditioned on the event B_t defined in the proof of Theorem 8, we have $A_t = \frac{1}{t}\sum_{k=1}^{t} \|\nabla f(x_k)\|^2 = \|\nabla f(x_1)\|^2 = \|x_1\|^2$, matching the value of F_t in (40). As mentioned in the main body, the reason for presenting our results in terms of the metric F_t instead of A_t stems from the fact that tail bounds on F_t have a more direct interpretation, as $F_t > \epsilon^2$ readily implies that an ϵ-stationary point has not been reached in any of the first t iterations. On the other hand, $A_t > \epsilon^2$ does not necessarily imply that an ϵ-stationary point has not been reached, as it is possible that $A_t > \epsilon^2 \geq F_t$, hence an ϵ-stationary point can be reached, while having $A_t > \epsilon^2$.

Appendix I. Clipping With a General Threshold

In this section, we discuss how our results for **c−SGD** can be extended when using a clipping threshold which does not requiring knowledge of the gradient bound G. In particular, we consider a clipping threshold of the form

$$\gamma_t = \begin{cases} C(t+1)^{\frac{2-p}{6p-4}}, & p \in (1,2) \\ C\sqrt{\log(t+1)}, & p = 2, \end{cases} \tag{43}$$

where $C > 0$ is any user-specified constant. We now want to establish a counterpart of Lemma 5, for the general clipping threshold in (43). To that end, we have the following result.

Lemma 13 *Let Assumptions 2 and 4 hold and let the clipping threshold be chosen as in (43). Then, the following are true.*

1. *For all $t \geq B_p$, we have $\|\theta_t^b\| \leq 4\sigma^p \gamma_t^{1-p}$, where $B_p = \begin{cases} \left(\frac{2G}{C}\right)^{\frac{6p-4}{2-p}}, & p \in (1,2) \\ 2^{4G^2/C^2} - 1, & p = 2 \end{cases}$.*
2. *For all $t \geq 1$ and any $\mathcal{F}_t$-measurable $x \in \mathbb{R}^d$, we have $\mathbb{E}\left[\exp\left(\langle x, \theta_t^u \rangle\right) | \mathcal{F}_t\right] \leq \exp\left(3\gamma_t^2 \|x\|^2\right)$.*

Proof To prove the first part, note that from Assumption 2, the choice of clipping threshold in (43) and the definition of B_p, we have, for any $t \geq B_p$

$$\|\nabla f(x_t)\| \leq G \leq \frac{\gamma_t}{2}.$$

The claim now readily follows by applying Proposition 11. The second part is proved in the same way as in the proof of Lemma 5. ■

Lemma 13 shows that the bound on the unbiased component established in Lemma 5 remains valid for all times t, while the bound on the biased component also becomes active, after a certain number of iterations. We then have the following result.

Theorem 14 *Let Assumptions 1, 2 and 4 hold and let $\{x_t\}_{t\in\mathbb{N}}$ be the sequence generated by **c−SGD** using the step-size $\alpha_t = (t+1)^{-\frac{p}{3p-2}}$ and clipping threshold γ_t given in (43). Then the sequence $\{F_t\}_{t\in\mathbb{N}}$ satisfies an LDP upper bound, with the following decay rate and rate function.*

1. *If $p \in (1,2)$, the decay rate is $n_t = \frac{t^{\frac{4(p-1)}{3p-2}}}{\log(t)}$, with rate function given by $I_c(x) = \begin{cases} \frac{x^2}{192C^2G^2}, & x \geq 0 \\ +\infty, & x < 0. \end{cases}$*
2. *If $p = 2$, the decay rate is $n_t = \frac{t}{\log^2(t)}$, with rate function given by $I_c(x) = \begin{cases} \frac{x^2}{96C^2G^2}, & x \geq 0 \\ +\infty, & x < 0. \end{cases}$*

Proof Recall that we showed the following inequality in Appendix F, for all $k \geq 1$

$$f(x_{k+1}) \leq f(x_k) - \frac{\alpha_k}{2}\|\nabla f(x_k)\|^2 - \alpha_k \langle \nabla f(x_k), \theta_k^u \rangle + \frac{\alpha_k}{2}\|\theta_k^b\|^2 + \frac{\alpha_k^2 \gamma_k^2 L}{2}.$$

Rearranging, summing up the first t iterations and using the fact that the step-sizes are non-increasing, we have

$$\frac{\alpha_t}{2}\sum_{k=1}^{t}\|\nabla f(x_k)\|^2 \leq f(x_1) - f_\star - \sum_{k=1}^{t}\alpha_k\langle \nabla f(x_k), \theta_k^u\rangle + \frac{1}{2}\sum_{k=1}^{t}\alpha_k\|\theta_k^b\|^2 + \sum_{k=1}^{t}\frac{\alpha_k^2\gamma_k^2 L}{2}. \quad (44)$$

We can see that the only difference between (44) and the corresponding equation (22) in the proof of Theorem 6 is the presence of $\sum_{k=1}^{t}\alpha_k\|\theta_k^b\|^2$. Therefore, we our aim is to bound this expression. To that end, let $t \geq B_p$ (this is fine, since we are interested in the limit behaviour $t \to \infty$), and notice that

$$\sum_{k=1}^{t}\alpha_k\|\theta_k^b\|^2 = \sum_{k=1}^{B_p}\alpha_k\|\theta_k^b\|^2 + \sum_{k=B_p}^{t}\alpha_k\|\theta_k^b\|^2 \leq \sum_{k=1}^{B_p}\alpha_k\|\theta_k^b\|^2 + 16\sum_{k=B_p}^{t}\alpha_k\sigma^{2p}\gamma_k^{2(1-p)},$$

where in the last inequality we used Lemma 13. What is left now is to show that the remaining term $\sum_{k=1}^{B_p}\alpha_k\|\theta_k^b\|^2$ stays bounded. Recalling that $\theta_k^b = \mathbb{E}\big[\widetilde{g}_k \mid \mathcal{F}_k\big] - \nabla f(x_k)$, where $\widetilde{g}_k$ is the clipped stochastic gradient, and using the fact that $\alpha_k \leq 1$, we get

$$\sum_{k=1}^{B_p}\alpha_k\|\theta_k^b\|^2 \leq 2\sum_{k=1}^{B_p}\left(\|\mathbb{E}\big[\widetilde{g}_k \mid \mathcal{F}_k\big]\|^2 + \|\nabla f(x_k)\|^2\right) \leq 2\sum_{k=1}^{B_p}\left(\gamma_k^2 + G^2\right) \leq 10G^2 B_p,$$

where the last inequality follows by noting that $\gamma_k \leq 2G$, for all $k \leq B_p$. The rest of the proof now follows the same steps as in Theorem 6, with $2G$ replaced by C, and is omitted, for brevity. ■