High-probability Convergence Bounds for Nonlinear Stochastic Gradient Descent Under Heavy-tailed Noise

Aleksandar Armacki[1], Pranay Sharma[1], Gauri Joshi[1], Dragana Bajović[2], Dušan Jakovetić[3], and Soummya Kar[1]

[1]Carnegie Mellon University, Pittsburgh, PA, USA,
{aarmacki,pranaysh,gaurij,soummyak}@andrew.cmu.edu
[2]Faculty of Technical Sciences, University of Novi Sad, Novi Sad, Serbia,
dbajovic@uns.ac.rs
[3]Faculty of Sciences, University of Novi Sad, Novi Sad, Serbia,
dusan.jakovetic@dmi.uns.ac.rs

Abstract

We study high-probability convergence guarantees of learning on streaming data in the presence of heavy-tailed noise. In the proposed scenario, the model is updated in an online fashion, as new information is observed, without storing any additional data. To combat the heavy-tailed noise, we consider a general framework of nonlinear stochastic gradient descent (SGD), providing several strong results. First, for non-convex costs and component-wise nonlinearities, we establish a convergence rate arbitrarily close to $\mathcal{O}\left(t^{-1/4}\right)$, *whose exponent is independent of noise and problem parameters*. Second, for strongly convex costs and component-wise nonlinearities, we establish a rate arbitrarily close to $\mathcal{O}\left(t^{-1/2}\right)$ for the weighted average of iterates, with exponent again independent of noise and problem parameters. Finally, for strongly convex costs and a broader class of nonlinearities, we establish convergence of the last iterate, with a rate $\mathcal{O}\left(t^{-\zeta}\right)$, where $\zeta \in (0,1)$ depends on problem parameters, noise and nonlinearity. As we show analytically and numerically, ζ can be used to inform the preferred choice of nonlinearity for given problem settings. Compared to state-of-the-art, who only consider clipping, require bounded noise moments of order $\eta \in (1,2]$, and establish convergence rates whose exponents go to zero as $\eta \to 1$, we provide high-probability guarantees for a much broader class of nonlinearities and symmetric density noise, with convergence rates whose exponents are bounded away from zero, even when the noise has *finite first moment only*. Moreover, in the case of strongly convex functions, we demonstrate analytically and numerically that clipping is not always the optimal nonlinearity, further underlining the value of our general framework.

1 Introduction

Learning on streaming data is a paradigm in which incoming samples are processed incrementally, while using limited memory and time, e.g., Gama (2012); Krawczyk et al. (2017). Formally, it can be represented as an optimization problem, with the goal of solving

$$\arg\min_{\mathbf{x}\in\mathbb{R}^d} \left\{ f(\mathbf{x}) \triangleq \mathbb{E}_\omega[\ell(\mathbf{x};\omega)] \right\}, \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^d$ represents model parameters, $\ell : \mathbb{R}^d \times \mathcal{W} \mapsto \mathbb{R}$ is a loss function, while $\omega \in \mathcal{W}$ is a random sample. The function $f : \mathbb{R}^d \mapsto \mathbb{R}$ is commonly known as the *population loss*. Many modern machine learning applications, such as classification and regression, can be modeled using (1). Learning on streaming data can be seen as a special case of *stochastic optimization* (SO), e.g., Robbins and Monro (1951); Nemirovski et al. (2009); Ghadimi and Lan (2013), with some important constraints. While the goal of both is to solve (1), SO approaches utilize both stochastic and mini-batch estimators, e.g., Bottou et al. (2018); Woodworth et al. (2020a,b), while streaming algorithms use only stochastic estimators and update the model parameters as each new sample arrives in the stream, without storing any additional information[1], e.g., Harvey et al. (2019); Nguyen et al. (2023a); Jakovetić et al. (2023).

Perhaps the most popular method to solve (1) is Stochastic Gradient Descent (SGD) Robbins and Monro (1951), with its popularity stemming from low computation cost and incredible empirical success, e.g., Bottou (2010); Hardt et al. (2016). Theoretical convergence guarantees of SGD have been studied extensively, e.g., Moulines and Bach (2011); Rakhlin et al. (2012); Ghadimi and Lan (2012, 2013); Bottou et al. (2018). The classical convergence results are mostly *in expectation*[2], characterizing the average performance across many runs of the algorithm. However, due to significant computational costs of a single run of an algorithm in many modern large-scale machine learning applications, it is often infeasible to perform multiple runs, e.g., Harvey et al. (2019); Davis et al. (2021). As such, the classical in expectation results do not fully capture the convergence behavior of these algorithms. More fine-grained results, like *high-probability convergence*, characterizing the behaviour of an algorithm with respect to a single run, offer better guarantees when multiple runs are infeasible.

Another striking feature of existing works is assuming that the gradient noise is *light-tailed* or has uniformly *bounded variance*, e.g., Rakhlin et al. (2012); Ghadimi and Lan (2012, 2013). As we discuss next, this is a major limitation in many modern applications. It has been observed in applications like training attention models that SGD performs worse than adaptive methods, even after extensive hyperparameter tuning, e.g., Simsekli et al. (2019); Zhang et al. (2020). In Zhang et al. (2020), it is shown that the gradient noise distribution during training of large attention models resembles a Levy α-stable distribution with $\alpha < 2$, which has unbounded variance. The authors show that SGD fails to converge due to presence of large stochastic gradients. The clipped variant of SGD solves the problem and achieves *optimal* convergence rate in expectation for smooth non-convex costs. Subsequently, clipped stochastic methods have been extensively analyzed in recent years to solve stochastic minimization, e.g., Gorbunov et al. (2020); Sadiev et al. (2023). Along with addressing heavy-tailed noise, clipped SGD also helps address non-smoothness of the objective function, e.g., Zhang et al. (2019), achieve differential privacy, e.g., Chen et al. (2020); Zhang et al. (2022); Yang et al. (2022) and robustness to malicious nodes in distributed learning, e.g., Yu and Kar (2023).

Despite its popularity, clipping is not the only nonlinear transformation of SGD employed in practice. Sign and quantized variants of SGD improve communication efficiency in distributed learning, e.g., Alistarh et al. (2017); Bernstein et al. (2018a); Gandikota et al. (2021). Sign SGD achieves performance on par with state-of-the-art adaptive methods, e.g., Crawshaw et al. (2022), and is robust to faulty/malicious users, e.g., Bernstein et al. (2018b). Normalized SGD is empirically observed to accelerate neural network training, e.g., Hazan et al. (2015); You et al. (2019); Cutkosky and Mehta (2020), facilitate private learning, e.g., Das et al.

[1] Such as the samples themselves, or stochastic gradients.

[2] Also commonly referred to as the *mean-squared error*.

(2021); Yang et al. (2022), and is effective in solving distributionally robust optimization, e.g., Jin et al. (2021). Zhang et al. (2020) observed empirically that component-wise clipping converges faster than joint clipping, exhibiting a better dependence on problem dimension.

Table 1: Comparison of streaming SGD methods. Lower-case t indicates a streaming method, upper-case T indicates a preset time horizon is needed, $\beta \in (0,1)$ is the failure probability, while $\widetilde{\mathcal{O}}$ hides logarithmic factors.

Cost	Work	Nonlinearity	Noise	Rate
Non-convex	Nguyen et al. (2023a)	Clipping only	bounded moment of order $\eta \in (1,2]$	$\widetilde{\mathcal{O}}\left(t^{\frac{\eta-2\eta}{3\eta-2}}\right)$
	Sadiev et al. (2023)	Clipping only	bounded moment of order $\eta \in (1,2]$	$\mathcal{O}\left(T^{\frac{1-\eta}{\eta}}\right)$[1]
	This paper	Component-wise	bounded first moment pdf symmetric, positive around zero	$\mathcal{O}\left(t^{\delta-1}\right)$[2]
Strongly convex	Tsai et al. (2022)	Clipping only	bounded growth second moment[3]	$\mathcal{O}\left(t^{-1}\right)$
	Sadiev et al. (2023)	Clipping only	bounded moment of order $\eta \in (1,2]$	$\mathcal{O}\left(T^{\frac{2(1-\eta)}{\eta}}\right)$
	This paper - weighted average of iterates	Component-wise	bounded first moment pdf symmetric, positive around zero	$\mathcal{O}\left(t^{2(\delta-1)}\right)$[4]
	This paper - last iterate	Component-wise and joint	bounded first moment pdf symmetric, positive around zero	$\mathcal{O}\left(t^{-\zeta}\right)$[5]

[1] The method in Sadiev et al. (2023) uses a preset time horizon T to tune algorithm parameters, (e.g., step-size). As such, it is not a streaming algorithm per se, however, it can be adapted to the streaming setting with minor tweaks (e.g., time-varying step-size).
[2] The parameter $\delta \in (3/4, 1)$ is user-specified. As such, our rate can be made arbitrarily close to $\mathcal{O}\left(t^{-1/4}\right)$, by setting $\delta = \frac{3}{4} + \epsilon$, for $\epsilon < \frac{1}{4}$ small.
[3] The gradient noise $\mathbf{z}$ at a point $\mathbf{x}$ satisfies $\mathbb{E}\|\mathbf{z}\|^2 \leq C + B\|\mathbf{x} - \mathbf{x}^\star\|^2$, where $\mathbf{x}^\star$ is the solution to (1) and $C, B > 0$ are constants.
[4] The parameter $\delta \in (3/4, 1)$ is user-specified. As such, our rate can be made arbitrarily close to $\mathcal{O}\left(t^{-1/2}\right)$, by setting $\delta = \frac{3}{4} + \epsilon$, for $\epsilon < \frac{1}{4}$ small.
[5] The rate $\zeta \in (0,1)$ depends on the choice of nonlinearity, noise and problem related parameters, see Sections 3, 4 and Appendix D. We provide examples of noise for which $\zeta > 2(\eta-1)/\eta$, see Examples 1-4 ahead.

Literature review. We now review the literature on high-probability convergence of SGD and its variants. Initial works on high-probability convergence of stochastic gradient methods considered light-tailed noise (see Definition 1) and include Nemirovski et al. (2009); Lan (2012); Hazan and Kale (2014); Harvey et al. (2019) for convex, and Ghadimi and Lan (2013); Li and Orabona (2020) for non-convex costs. Subsequent works Gorbunov et al. (2020, 2021); Parletta et al. (2022) generalized these results to noise with bounded variance. Tsai et al. (2022) study the behaviour of clipped SGD under the assumption that the noise variance is bounded by iterate distance (see Table 1), while Li and Liu (2022); Eldowa and Paudice (2023) considered sub-Weibull noise. Recent works Liu et al. (2023a); Eldowa and Paudice (2023) remove restrictive assumptions, like bounded stochastic gradients and bounded domain. Sadiev et al. (2023) show that even with bounded variance and smooth, strongly-convex functions, vanilla SGD cannot achieve a *logarithmic* dependence on the failure probability $\beta \in (0,1)$[3], implying that the complexity of achieving a high-probability bound for SGD is much worse than the complexity of converging in expectation. As such, nonlinear SGD is required to handle tails heavier than sub-Gaussian. Recent works consider a broad class of heavy-tailed noises with bounded moments of order $\eta \in (1,2]$, e.g., Nguyen et al. (2023a,b); Sadiev et al. (2023); Liu et al. (2023b). Nguyen et al. (2023a,b) study high-probability convergence of clipped SGD for convex and non-convex minimization, Sadiev et al. (2023) study clipped SGD for optimization and variational inequality problems, while Liu et al. (2023b) study accelerated variants of clipped SGD for smooth costs. It is worth mentioning a recent work Puchkin

[3] For high-probability convergence guarantees of type "with probability at least $1 - \beta$".

et al. (2023), which shows clipped SGD can achieve the optimal $\mathcal{O}\left(T^{-1}\right)$ rate for smooth strongly convex costs and a class of heavy-tailed noises with possibly unbounded first moments. However, they use a median-of-means gradient estimator, which requires storing multiple stochastic gradients prior to performing the update and it is not clear how such an approach can be extended to the streaming setting considered in this paper.

The works closest to ours are Nguyen et al. (2023a); Sadiev et al. (2023) for non-convex and Tsai et al. (2022); Sadiev et al. (2023) for strongly convex costs. For non-convex costs, the optimal rate $\widetilde{\mathcal{O}}\left(t^{\frac{2-2\eta}{3\eta-2}}\right)$ is achieved in Nguyen et al. (2023a). For the same costs, we consider a broad class of component-wise nonlinearities (e.g., sign, cclip and quantization) in the presence of noise with symmetric density and bounded first moment, achieving the rate $\mathcal{O}\left(t^{-1/4+\epsilon}\right)$, for any $\epsilon > 0$. As such, our rate exponent *is independent of noise or problem related parameters*, which is not the case with Nguyen et al. (2023a); Sadiev et al. (2023), with our rate dominating that from Nguyen et al. (2023a) whenever $\eta < \frac{6+8\epsilon}{5+12\epsilon}$.[4] Furthermore, Nguyen et al. (2023a); Sadiev et al. (2023) require knowledge of noise moment η and other problem parameters to tune the step-size and clipping radius, whereas our analysis *requires no knowledge of problem or noise related parameters*. For strongly convex costs, Tsai et al. (2022); Sadiev et al. (2023) study the convergence of the last iterate for clipped SGD, with Sadiev et al. (2023) allowing for a more general class of noise, achieving the rate $\mathcal{O}\left(T^{2(1-\eta)/\eta}\right)$. Compared to them, we consider a more general framework for nonlinear SGD (both component-wise and joint), establishing a rate $\mathcal{O}\left(t^{-\zeta}\right)$, for some $\zeta \in (0, 1)$ that depends on the noise, nonlinearity and other problem parameters. We give examples of noise regimes where our rate is better than the one in Sadiev et al. (2023) (see Examples 1-4). Moreover, we demonstrate both analytically and numerically that ζ is informative for choosing the best nonlinearity for given problem and noise settings and that *clipping is not always the best chocice of nonlinearity* (see Section 4), further highlighting the importance and usefulness of our general framework. The comparison is summed up in Table 1. Finally, it is worth mentioning Jakovetić et al. (2023), who study the same general framework for nonlinear SGD and strongly convex costs, in expectation and almost sure sense. Our work differs in that we study high-probability convergence and allow for non-convex costs. The latter is achieved by providing a novel characterization of the interplay of the "denoised" nonlinear gradient and the true, noiseless gradient for component-wise nonlinearities (see Section 3). For strongly convex costs, naively combining the in expectation bound from Jakovetić et al. (2023) and Markov inequality results in a sub-optimal $1/\beta$ dependence on the failure probability β, with our work closing this gap, by showing the optimal $\log(1/\beta)$ dependence.

Contributions. Our contributions can be summarized as follows.

- We provide a unified framework for studying convergence in high probability of nonlinear streaming SGD under heavy-tailed noise. In the proposed framework, the nonlinearity is treated in a black-box manner, subsuming many popular nonlinearities, like sign, normalization, cclip and quantization. *To the best of our knowledge, we provide the first*

[4] This does not contradict the optimality of the rate achieved in Nguyen et al. (2023a), as their assumptions slightly differ from ours. Whereas Nguyen et al. (2023a) require bounded noise moment of order $\eta \in (1, 2]$, we require noise with symmetric density, but allow for bounded first moment only. As such, our work shows that additional structure in the noise leads to improved results, while allowing for relaxed moment conditions and heavier tails (see Examples 1-3).

high-probability results under heavy-tailed noise for methods such as sign, quantized and component-wise clipped SGD.

- For non-convex costs and component-wise nonlinearities, we show a convergence rate of $\mathcal{O}\left(t^{-1/4+\epsilon}\right)$, for any $\epsilon > 0$. The exponent in our rate is independent of noise and problem parameters, which is not the case for state-of-the-art rate in Nguyen et al. (2023a), with our rate dominating the state-of-the-art whenever the noise has bounded moments of order $\eta < \frac{6+8\epsilon}{5+12\epsilon}$. Additionally, our analysis requires no knowledge of problem parameters to tune the step-size, whereas the analysis in Nguyen et al. (2023a) requires knowledge of noise moments and problem parameters to tune the step-size and clipping radius.

- For strongly convex costs and component-wise nonlinearities we show convergence of the weighted average of iterates with the same rate $\mathcal{O}\left(t^{-1/2+\epsilon}\right)$, with our rate dominating the state-of-the-art Sadiev et al. (2023), whenever the noise has bounded moments of order $\eta < \frac{4}{3+2\epsilon}$. Next, we show convergence of the last iterate for a broader class of nonlinearities, with rate $\mathcal{O}\left(t^{-\zeta}\right)$, where $\zeta \in (0,1)$ depends on noise, nonlinearity and other problem parameters. As such, the exponent ζ is shown to be informative in choosing the best nonlinearity for the given problem and noise settings, both analytically and numerically. We provide examples of noise regimes in which our exponent ζ dominates the one instate-of-the-art Sadiev et al. (2023).

- Compared to state-of-the-art Nguyen et al. (2023a); Sadiev et al. (2023), who only consider clipping, require bounded noise moments of order $\eta \in (1,2]$ and vanishing rates as $\eta \to 1$, we consider a much broader class of nonlinearities under noise with symmetric density, while relaxing the moment condition and providing non-vanishing convergence rates even for noise with *finite first moment only*. Moreover, for strongly convex costs, we provide analytical and numerical results that show *clipping is not always the optimal choice of nonlinearity*, further reinforcing the importance of our general framework.

Paper organization. The rest of the paper is organized as follows. Section 2 outlines the nonlinear streaming SGD framework. Section 3 presents the main results of the paper. Section 4 provides analytical and numerical results demonstrating that clipping is not always the optimal choice of nonlinearity. Section 5 concludes the paper. The Appendix presents some useful facts and results omitted from the main body. The remainder of this section introduces the notation.

Notation. We denote the set of positive integers by $\mathbb{N}$, while $\mathbb{N}_0$ denotes the set of non-negative integers, i.e., $\mathbb{N}_0 \triangleq \mathbb{N} \cup \{0\}$. We use $\mathbb{R}$, $\mathbb{R}_+$ and $\mathbb{R}^d$ to denote the sets of real numbers, non-negative real numbers and the d-dimensional real vector space, respectively. For $a \in \mathbb{N}$, $[a]$ denotes the set of integers up to and including a, i.e., $[a] = \{1, \dots, a\}$. Regular and bold symbols denote scalars and vectors, respectively, i.e., $x \in \mathbb{R}$ and $\mathbf{x} \in \mathbb{R}^d$. The Euclidean inner product is denoted by $\langle \cdot, \cdot \rangle$, while $\| \cdot \|$ denotes the induced norm.

2 Proposed Framework

To solve (1) in the streaming setting and in the presence of heavy-tailed noise, we use the *nonlinear SGD* framework. The algorithm starts by choosing a deterministic initial model

Algorithm 1 Nonlinear SGD on Streaming Data

Require: Choice of nonlinearity $\mathbf{\Psi} : \mathbb{R}^d \mapsto \mathbb{R}^d$, model initialization $\mathbf{x}^0 \in \mathbb{R}^d$;

1: **at** time t = 0,1,2,... **do**:
2: Observe a new sample $\omega^{(t)}$ and compute the gradient $\nabla \ell(\mathbf{x}^{(t)}; \omega^{(t)})$;
3: Update the model $\mathbf{x}^{(t+1)} \leftarrow \mathbf{x}^{(t)} - \alpha_t \mathbf{\Psi}\left(\nabla \ell(\mathbf{x}^{(t)}; \omega^{(t)})\right)$;

$\mathbf{x}^{(0)} \in \mathbb{R}^d$ and a nonlinear map $\mathbf{\Psi} : \mathbb{R}^d \mapsto \mathbb{R}^d$. In iteration $t = 0, 1, \ldots$, the method performs as follows: a new sample $\omega^{(t)}$ is observed and the gradient of the loss ℓ at the current model $\mathbf{x}^{(t)}$ and sample $\omega^{(t)}$ is computed[5]. Then, the model is updated as

$$\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - \alpha_t \mathbf{\Psi}\left(\nabla \ell(\mathbf{x}^{(t)}; \omega^{(t)})\right), \tag{2}$$

where $\alpha_t > 0$ is the step-size at iteration t. The method is summed up in Algorithm 1. We make the following assumption on the nonlinear map $\mathbf{\Psi}$.

Assumption 1. The nonlinear map $\mathbf{\Psi} : \mathbb{R}^d \mapsto \mathbb{R}^d$ is either of the form $\mathbf{\Psi}(\mathbf{x}) = \mathbf{\Psi}(x_1, \ldots, x_d) = [\mathcal{N}_1(x_1), \ldots, \mathcal{N}_1(x_d)]^\top$ or $\mathbf{\Psi}(\mathbf{x}) = \mathbf{x}\mathcal{N}_2(\|\mathbf{x}\|)$, where $\mathcal{N}_1$, $\mathcal{N}_2 : \mathbb{R} \mapsto \mathbb{R}$ satisfy

1. $\mathcal{N}_1, \mathcal{N}_2$ are continuous almost everywhere (with respect to the Lebesgue measure), with $\mathcal{N}_1$ piece-wise differentiable, while the mapping $a \mapsto a\mathcal{N}_2(a)$ is non-decreasing.
2. $\mathcal{N}_1$ is monotonically non-decreasing and odd function, while $\mathcal{N}_2$ is non-increasing.
3. $\mathcal{N}_1$ is either discontinuous at zero, or strictly increasing on $(-c_1, c_1)$, for some $c_1 > 0$, with $\mathcal{N}_2(a) > 0$, for any $a > 0$.
4. $\mathcal{N}_1$ and $\mathbf{x}\mathcal{N}_2(\|\mathbf{x}\|)$ are uniformly bounded, i.e., $|\mathcal{N}_1(x)| \leq C_1$ and $\|\mathbf{x}\mathcal{N}_2(\|\mathbf{x}\|)\| \leq C_2$, for some C_1, $C_2 > 0$, and all $x \in \mathbb{R}$, $\mathbf{x} \in \mathbb{R}^d$.

Note that the fourth property implies $\|\mathbf{\Psi}(\mathbf{x})\| \leq C$, where $C = C_1\sqrt{d}$ or $C = C_2$, depending on the form of nonlinearity. We will use the general bound $\|\mathbf{\Psi}(\mathbf{x})\| \leq C$ for ease of presentation, and specialize where appropriate. Assumption 1 is satisfied by a wide class of nonlinearities, including:

1. *Sign*: $[\mathbf{\Psi}(\mathbf{x})]_i = \text{sign}(x_i)$, $i = 1, \ldots, d$,
2. *Component-wise clip (cclip)*: $[\mathbf{\Psi}(\mathbf{x})]_i = x_i$, for $|x_i| \leq m$, and $[\mathbf{\Psi}(\mathbf{x})]_i = m \cdot \text{sign}(x_i)$, for $|x_i| > m$, $i = 1, \ldots, d$, for some constant $m > 0$.
3. *Component-wise quantization*: for each $i = 1, \ldots, d$, let $[\mathbf{\Psi}(\mathbf{x})]_i = r_j$, for $x_i \in (q_j, q_{j+1}]$, with $j = 0, \ldots, J-1$ and $-\infty = q_0 < q_1 < \ldots < q_J = +\infty$, where r_j's and q_j's are chosen such that each component of $\mathbf{\Psi}$ is an odd function, and we have $\max_{j \in \{0,\ldots,J-1\}} |r_j| < R$, for $R > 0$.
4. *Normalization*: $\mathbf{\Psi}(\mathbf{x}) = \frac{\mathbf{x}}{\|\mathbf{x}\|}$, with $\mathbf{\Psi}(\mathbf{x}) = 0$ if $\mathbf{x} = \mathbf{0}$.
5. *Clipping*: $\mathbf{\Psi}(\mathbf{x}) = \min\left\{1, \frac{M}{\|\mathbf{x}\|}\right\}\mathbf{x}$, for some constant $M > 0$.

[5]Equivalently, we have access to a first-order oracle that directly streams gradients of ℓ, instead of samples.

3 Main Results

In this section we present the main results of the paper. Subsection 3.1 presents the preliminaries and assumptions, Subsection 3.2 presents the main theoretical results for general non-convex costs, while Subsection 3.3 presents the main theoretical results for strongly convex costs. All the proofs can be found in Appendix C.

3.1 Preliminaries

In this section we provide the preliminaries and assumptions used in the analysis. To begin with, we state the assumptions on the behaviour of the population loss f used in the paper.

Assumption 2. The population loss f is bounded from below, has at least one stationary point and Lipschitz continuous gradients, i.e., $\inf_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) > -\infty$, there exists a $\mathbf{x} \in \mathbb{R}^d$ such that $\nabla f(\mathbf{x}) = 0$, and for some $L > 0$ and every $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|.$$

Remark 1. Boundedness from below and Lipschitz continuous gradients are standard for non-convex costs, e.g., Ghadimi and Lan (2013); Madden et al. (2020); Liu et al. (2023a). Since the goal in non-convex optimization is to reach a stationary point, it is natural to assume at least one such point exists.

Remark 2. It can be readily shown that Lipschitz continuous gradients imply, for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x}\rangle + \frac{L}{2}\|\mathbf{x} - \mathbf{y}\|^2,$$

know as *L-smoothness inequality/property*, see, e.g., Nesterov (2018); Wright and Recht (2022).

Assumption 3. The population loss f is strongly convex, i.e., for some $\mu > 0$ and every $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$

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

Remark 3. Combined with Assumption 3, it follows that $\mu \leq L$.

Denote the infimum of f by $f^\star$, i.e., $f^\star \triangleq \inf_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x})$. Denote by $\mathcal{X} \subset \mathbb{R}^d$ the set of stationary points of f, i.e., $\mathcal{X} \triangleq \{\mathbf{x} \in \mathbb{R}^d : \nabla f(\mathbf{x}) = 0\}$. Then, by Assumption 2, it follows that $\mathcal{X} \neq \emptyset$. Assumption 3 implies the existence of a unique global minimizer, i.e., we have $f^\star = f(\mathbf{x}^\star)$, for some unique $\mathbf{x}^\star \in \mathbb{R}^d$. Equivalently, we have $\mathcal{X} = \{\mathbf{x}^\star\}$. Next, rewrite the update (2) as

$$\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - \alpha_t \mathbf{\Psi}(\nabla f(\mathbf{x}^{(t)}) + \mathbf{z}^{(t)}), \tag{3}$$

where $\mathbf{z}^{(t)} \triangleq \nabla \ell(\mathbf{x}^{(t)}; \omega^{(t)}) - \nabla f(\mathbf{x}^{(t)})$ is the stochastic noise at iteration t. To simplify the notation, we use the shorthand $\mathbf{\Psi}^{(t)} \triangleq \mathbf{\Psi}(\nabla f(\mathbf{x}^{(t)}) + \mathbf{z}^{(t)})$. We make the following assumption on the sequence of noise vectors $\{\mathbf{z}^{(t)}\}_{t\in\mathbb{N}_0}$.

Assumption 4. Noise vectors $\{\mathbf{z}^{(t)}\}_{t\in\mathbb{N}_0}$ are zero mean, independent, identically distributed and integrable, with symmetric probability density function (PDF) $p : \mathbb{R}^d \mapsto \mathbb{R}_+$, strictly positive in a neighborhood of zero, i.e., $p(\mathbf{z}) > 0$, for all $\|\mathbf{z}\| \leq B_0$ and some $B_0 > 0$.

Remark 4. Assumption 4 relaxes the noise moment condition made in Nguyen et al. (2023a); Sadiev et al. (2023), at the expense of requiring a symmetric PDF, positive in a neighborhood of zero. PDF symmetry and positivity around zero are mild assumptions, satisfied by many noise distributions, such as Gaussian, as well as a broad class of heavy-tailed zero-mean α-stable distributions, Bercovici et al. (1999); Şimşekli et al. (2019), or the ones in Examples 1-3 below.

We now give some examples of noise PDFs satisfying Assumption 4.

Example 1. The noise PDF $p(\mathbf{z}) = \rho(z_1) \times \ldots \times \rho(z_d)$, where $\rho(z) = \frac{\alpha-1}{2(1+|z|)^\alpha}$, for some $\alpha > 2$. It can be shown that the PDF only has finite η-th moments for $\eta < \alpha - 1$.

Example 2. The noise PDF $p(\mathbf{z}) = \rho(z_1) \times \ldots \times \rho(z_d)$, where $\rho(z) = \frac{c}{(z^2+1)\log^2(|z|+2)}$, with $c = \int 1/[(z^2+1)\log^2(|z|+2)]dz$ being the normalizing constant. It can be shown that p has a finite first moment, but for any $\eta \in (1, 2]$, the η-th moments do not exist.

Example 3. The PDF $p : \mathbb{R}^d \mapsto \mathbb{R}_+$ with "radial symmetry", i.e., $p(\mathbf{z}) = \rho(\|\mathbf{z}\|)$, where $\rho : \mathbb{R} \mapsto \mathbb{R}_+$ is itself a PDF. If ρ is the PDF from Example 2, then p inherits the properties of ρ, i.e., it does not have finite η-th moments, for any $\eta > 1$.

Note that, while the noise from Example 1 satisfies the moment condition from Nguyen et al. (2023a); Sadiev et al. (2023), the noise from Example 2 clearly does not. Next, define the function $\mathbf{\Phi} : \mathbb{R}^d \mapsto \mathbb{R}^d$, given by $\mathbf{\Phi}(\mathbf{x}) \triangleq \mathbb{E}_{\mathbf{z}}[\mathbf{\Psi}(\mathbf{x}+\mathbf{z})] = \int \mathbf{\Psi}(\mathbf{x}+\mathbf{z})p(\mathbf{z})d\mathbf{z}$,[6] where the expectation is taken with respect to the gradient noise at a random sample, i.e., $\mathbf{z} \triangleq \nabla\ell(\mathbf{x};\omega) - \nabla f(\mathbf{x})$. We use the shorthand $\mathbf{\Phi}^{(t)} \triangleq \mathbf{\Phi}(\nabla f(\mathbf{x}^{(t)}))$. The vector $\mathbf{\Phi}^{(t)}$ can be seen as the denoised version of $\mathbf{\Psi}^{(t)}$. Using $\mathbf{\Phi}^{(t)}$, we can rewrite the update rule (3) as

$$\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - \alpha_t \mathbf{\Phi}^{(t)} + \alpha_t \mathbf{e}^{(t)}, \tag{4}$$

where $\mathbf{e}^{(t)} \triangleq \mathbf{\Phi}^{(t)} - \mathbf{\Psi}^{(t)}$, represents the *effective noise* term. As we show next, the effective noise is light-tailed, even though the original noise may not be. This allows us to establish exponential concentration inequalities and tight control of the behaviour of MGFs. Prior to that, we define the concept of *sub-Gaussianity*, e.g., Vershynin (2018); Jin et al. (2019).

Definition 1. A zero-mean random vector $\mathbf{v} \in \mathbb{R}^d$ is said to be sub-Gaussian, if there exists a constant $N > 0$, such that, for any $\mathbf{x} \in \mathbb{R}^d$

$$\mathbb{E}\left[\exp\left(\langle \mathbf{x}, \mathbf{v}\rangle\right)\right] \leq \exp\left(N\|\mathbf{x}\|^2\right).$$

It can be shown that any bounded random variable is sub-Gaussian, a fact used in the following sections (see Appendix B for a proof). Define $\mathcal{F}_t$ to be the natural filtration, i.e., $\mathcal{F}_t \triangleq \sigma\left(\{\mathbf{z}^{(0)}, \ldots, \mathbf{z}^{(t-1)}\}\right)$, with $\mathcal{F}_0 \triangleq \sigma(\{\emptyset, \Omega\})$ being the trivial σ-algebra. Next, we state some properties of the effective noise $\mathbf{e}^{(t)}$.

Lemma 3.1. *Let Assumptions 1 and 4 hold. Then, the effective noise vectors* $\{\mathbf{e}^{(t)}\}_{t\in\mathbb{N}_0}$ *satisfy*

1. $\mathbb{E}[\mathbf{e}^{(t)} | \mathcal{F}_t] = 0$ *and* $\|\mathbf{e}^{(t)}\| \leq 2C$,
2. *The effective noise is sub-Gaussian, i.e., for some* $N > 0$ *any* $\mathbf{x} \in \mathbb{R}^d$, *we have* $\mathbb{E}\left[\exp\left(\langle \mathbf{x}, \mathbf{e}^{(t)}\rangle\right) \mid \mathcal{F}_t\right] \leq \exp\left(N\|x\|^2\right)$

Note that the bound on the effective noise $\mathbf{e}^{(t)}$ comes from the nonlinearity, therefore, the constant N depends only on the nonlinearity, not the noise.

[6] If $\mathbf{\Psi}$ is a component-wise nonlinearity, then $\mathbf{\Phi}$ is a vector with components $\phi_i(x_i) = \mathbb{E}_{z_i}[\mathcal{N}_1(x_i + z_i)]$, where $\mathbb{E}_{z_i}$ is the marginal expectation with respect to the i-th noise component, $i \in [d]$ (see Lemma 3.2 ahead).

3.2 Non-convex Costs

In this section we establish the convergence in high-probability for general non-convex functions and component-wise nonlinearities. This is made possible by establishing a novel characterization of the behaviour of $\mathbf{\Phi}(\mathbf{x})$ with respect to the original (noiseless) vector $\mathbf{x}$. To begin with, we state a result from Polyak and Tsypkin (1979), that provides some basic properties of the mapping $\mathbf{\Phi}$ for component-wise nonlinearities.

Lemma 3.2. *Let Assumptions 1 and 4 hold, with the nonlinearity $\mathbf{\Psi} : \mathbb{R}^d \mapsto \mathbb{R}^d$ being component wise, i.e., of the form $\mathbf{\Psi}(\mathbf{x}) = \big[\mathcal{N}_1(x_1), \ldots, \mathcal{N}_1(x_d)\big]^\top$. Then, the function $\mathbf{\Phi} : \mathbb{R}^d \mapsto \mathbb{R}^d$ is of the form $\mathbf{\Phi}(\mathbf{x}) = \big[\phi_1(x_1), \ldots, \phi_d(x_d)\big]^\top$, where $\phi_i(x_i) = \mathbb{E}_{z_i}\left[\mathcal{N}_1(x_i + z_i)\right]$ is the marginal expectation of the i-th noise component, $i \in [d]$, with the following properties:*

1. *ϕ_i is non-decreasing and odd, with $\phi_i(0) = 0$;*
2. *ϕ_i is differentiable in zero, with $\phi_i'(0) > 0$.*

Define $\phi'(0) \triangleq \min_{i\in[d]} \phi_i'(0)$. We then have the following result on the interplay of the "denoised" nonlinearity $\mathbf{\Phi}(\mathbf{x})$ and $\mathbf{x}$.

Lemma 3.3. *Let Assumptions 1 and 4 hold, with the nonlinearity $\mathbf{\Psi} : \mathbb{R}^d \mapsto \mathbb{R}^d$ being component wise, i.e., of the form $\mathbf{\Psi}(\mathbf{x}) = \big[\mathcal{N}_1(x_1), \ldots, \mathcal{N}_1(x_d)\big]^\top$. Then, for any $\mathbf{x} \in \mathbb{R}^d$, we have*

$$\langle \mathbf{\Phi}(\mathbf{x}), \mathbf{x}\rangle \geq \frac{\phi'(0)}{2} \min\{\xi\|\mathbf{x}\|/\sqrt{d}, \|\mathbf{x}\|^2/d\},$$

where $\xi > 0$ is a global constant depending only on the choice of nonlinearity and noise.

Lemma 3.3 provides a novel characterization of the inner product of the "denoised" nonlinearity $\mathbf{\Phi}$ at vector $\mathbf{x}$ and vector $\mathbf{x}$ itself. For any $k = 0, \ldots, t-1$, define $\widetilde{\alpha}_k \triangleq \frac{\alpha_k}{\sum_{j=0}^{t-1}\alpha_j}$, so that $\sum_{k=0}^{t-1}\widetilde{\alpha}_k = 1$, and for any $\mathbf{x} \in \mathbb{R}^d$ define $D_{\mathcal{X}}(\mathbf{x}^{(0)}) \triangleq \inf_{\mathbf{x}\in\mathcal{X}} \|\mathbf{x}^{(0)} - \mathbf{x}\|^2$ to be the set distance function. We are now ready to state our high-probability convergence bound of component-wise nonlinear SGD for non-convex costs.

Theorem 1. *Let Assumptions 1, 2 and 4 hold, with the nonlinearity $\mathbf{\Psi} : \mathbb{R}^d \mapsto \mathbb{R}^d$ being component-wise, i.e., of the form $\mathbf{\Psi}(\mathbf{x}) = \big[\mathcal{N}_1(x_1), \ldots, \mathcal{N}_1(x_d)\big]^\top$. Let $\{\mathbf{x}^{(t)}\}_{t\in\mathbb{N}_0}$ be the sequence generated by (3), with step-size $\alpha_t = \frac{a}{(t+2)^\delta}$, for any $\delta \in (3/4, 1)$ and $a > 0$. Then, for any $t \in \mathbb{N}_0$, and any $\beta \in (0,1)$, with probability at least $1-\beta$, it holds that*

$$\sum_{k=0}^{t-1} \widetilde{\alpha}_k \min\{\xi\|\nabla f(\mathbf{x}^{(k)})\|/\sqrt{d}, \|\nabla f(\mathbf{x}^{(k)})\|^2/d\} \leq \frac{R(a,\beta,\delta)}{(t+2)^{1-\delta} - 2^{1-\delta}},$$

where $R(a,\beta,\delta) \triangleq \frac{2(1-\delta)}{\phi'(0)}\left[\big(f(\mathbf{x}^{(0)}) - f^\star + \log(1/\beta)\big)/a + \frac{a(dLC_1^2/2 + 2NL^2 D_{\mathcal{X}}(\mathbf{x}^{(0)}))}{(2\delta-1)} + \frac{2a^3 dNC_1^2L^2}{(1-\delta)^2(4\delta-3)}\right]$.

Some remarks are now in order.

Remark 5. The bound from Theorem 1 can be used to provide a bound on the best iterate, i.e., on the quantity $\min_{0\leq k\leq t-1}\|\nabla f(\mathbf{x}^{(k)})\|$. In particular, one can show that Theorem 1 implies

$$\min_{0\leq k\leq t-1}\|\nabla f(\mathbf{x}^{(k)})\| = \mathcal{O}\left(\sqrt{\frac{dR(a,\beta,\delta)}{(t+2)^{1-\delta} - 2^{1-\delta}}}\right),$$

i.e., in order to ensure we reach an ϵ-stationary point[7] (with high probability), we require at least $\mathcal{O}\left(\epsilon^{-\frac{2}{1-\delta}}\right)$ iterations. Derivations connecting the bound from Theorem 1 to the best iterate can be found in Appendix F.

Remark 6. The rate achieved by our analysis is of the order $\mathcal{O}\left(t^{\delta-1}\right)$, where $\delta \in (3/4, 1)$ is user-specified. As such, the exponent in our convergence rate is *independent of any problem related parameters*, like the choice of nonlinearity, noise, etc. This is not the case with state-of-the-art methods, e.g., Nguyen et al. (2023a); Sadiev et al. (2023), whose rate exponent explicitly depends on noise moment $\eta \in (1, 2]$ and vanishes as $\eta \to 1$. By choosing $\delta = 3/4 + \epsilon$, for some $\epsilon \in (0, 1/4)$, it follows that our method can achieve the rate $\mathcal{O}\left(t^{-1/4+\epsilon}\right)$, which can be made arbitrarily close to $t^{-1/4}$, for small ϵ. In this case $R(a, \beta, 3/4 + \epsilon) = \mathcal{O}\left(\epsilon^{-1}\right)$, with the convergence rate being $\mathcal{O}\left(\epsilon^{-1} t^{-1/4+\epsilon}\right)$. We can see an inherent trade-off with respect to ϵ, where smaller value of ϵ results in a convergence rate closer to $t^{-1/4}$, at the cost of a larger constant factor (i.e., ϵ^{-1}). Compared to the convergence rate $\mathcal{O}\left(t^{\frac{2-2\eta}{3\eta-2}}\right)$, achieved in Nguyen et al. (2023a), our rate is better if $\frac{2\eta-2}{3\eta-2} < \frac{1}{4} - \epsilon$, i.e., whenever $\eta < \frac{6+8\epsilon}{5+12\epsilon}$.

Remark 7. The constant $R(a, \beta, \delta)$ depends on multiple problem related quantities, such as the problem dimension d, the optimality gap $f(\mathbf{x}^{(0)}) - f^\star$, the distance of the initial model from the set of stationary points $D_{\mathcal{X}}(\mathbf{x}^{(0)})$, choice of nonlinearity and noise through C_1, N and $\phi'(0)$. The step-size parameter $a > 0$ offers an inherent trade-off, as choosing $a \approx 0$ alleviates the dependence of R on multiple problem parameters (e.g., initial model from set of stationary points, or the effect of small ϵ discussed in Remark 6), making it approximately $R(a, \beta, \delta) \approx \frac{2(1-\delta)}{\phi'(0)} \left(f(\mathbf{x}^{(0)}) - f^\star + \log(1/\beta)\right)/a$, while simultaneously resulting in small step-sizes and larger constant factor (of the order $1/a$), slowing convergence down.

Proof of Theorem 1. For ease of notation, let $Z(\|\nabla f(\mathbf{x}^{(t)})\|) \triangleq \min\{\xi\|\nabla f(\mathbf{x}^{(t)})\|/\sqrt{d}, \|\nabla f(\mathbf{x}^{(t)})\|^2/d\}$. Applying the L-smoothness property of f and the update rule (4), to get

$$\begin{aligned} f(\mathbf{x}^{(t+1)}) &\leq f(\mathbf{x}^{(t)}) - \alpha_t \langle \nabla f(\mathbf{x}^{(t)}), \mathbf{\Phi}^{(t)} - \mathbf{e}^{(t)} \rangle + \frac{\alpha_t^2 L}{2} \|\mathbf{\Psi}^{(t)}\|^2 \\ &\leq f(\mathbf{x}^{(t)}) - \frac{\alpha_t \phi'(0)}{2} Z(\|\nabla f(\mathbf{x}^{(t)})\|) + \alpha_t \langle \nabla f(\mathbf{x}^{(t)}), \mathbf{e}^{(t)} \rangle + \frac{\alpha_t^2 d L C_1^2}{2}, \end{aligned}$$

where the second inequality follows from Lemma 3.3 and Assumption 1. Rearranging and summing up the first t terms, we get

$$\frac{\phi'(0)}{2} \sum_{k=0}^{t-1} \alpha_k Z(\|\nabla f(\mathbf{x}^{(k)})\|) \leq \underbrace{f(\mathbf{x}^{(0)}) - f^\star + \frac{d L C_1^2}{2} \sum_{k=1}^{t} \alpha_k^2}_{=: B_1} + \underbrace{\sum_{k=1}^{t} \alpha_k \langle \nabla f(\mathbf{x}^{(k)}), \mathbf{e}^{(k)} \rangle}_{=: B_2}. \tag{5}$$

Denote the right-hand side of (5) by Z_t, i.e., $Z_t \triangleq \frac{\phi'(0)}{2} \sum_{k=1}^{t} \alpha_k Z(\|\nabla f(\mathbf{x}^{(k)})\|)$ and note that B_1 is independent of the noise, i.e., is a deterministic quantity. We then have

$$\mathbb{E}\left[\exp(Z_t)\right] \overset{(5)}{\leq} \mathbb{E}\left[\exp\left(B_1 + B_2\right)\right] = \exp\left(B_1\right) \mathbb{E}\left[\exp(B_2)\right].$$

[7] A point $\mathbf{x} \in \mathbb{R}^d$ such that $\|\nabla f(\mathbf{x})\| \leq \epsilon$.

We now bound $\mathbb{E}[\exp(B_2)]$. Denote by $\mathbb{E}_t[\cdot] \triangleq \mathbb{E}[\cdot \mid \mathcal{F}_t]$ the expectation conditioned on history up to time t. We then have

$$\begin{aligned}
\mathbb{E}[\exp(B_2)] &= \mathbb{E}\left[\exp\left(\sum_{k=0}^{t-1}\alpha_k\langle\nabla f(\mathbf{x}^{(k)}),\mathbf{e}^{(k)}\rangle\right)\right] \\
&= \mathbb{E}\left[\exp\left(\sum_{k=0}^{t-2}\alpha_k\langle\nabla f(\mathbf{x}^{(k)}),\mathbf{e}^{(k)}\rangle\right)\mathbb{E}_{t-1}\left[\exp(\alpha_{t-1}\langle\nabla f(\mathbf{x}^{(t-1)}),\mathbf{e}^{(t-1)}\rangle)\right]\right] \\
&\leq \mathbb{E}\left[\exp\left(\sum_{k=0}^{t-2}\alpha_k\langle\nabla f(\mathbf{x}^{(k)}),\mathbf{e}^{(k)}\rangle\right)\exp\left(N\alpha_{t-1}^2\|\nabla f(\mathbf{x}^{(t-1)})\|^2\right)\right] \\
&\leq \ldots \leq \mathbb{E}\left[\exp\left(N\sum_{k=0}^{t-1}\alpha_k^2\|\nabla f(\mathbf{x}^{(k)})\|^2\right)\right],
\end{aligned}$$

where we repeatedly use Lemma 3.1. Next, consider $\|\nabla f(\mathbf{x}^{(k)})\|$, for any $k \geq 0$. Define $A_t \triangleq \sum_{k=0}^{t-1}\alpha_k$ and use L-smoothness, to get

$$\begin{aligned}
\|\nabla f(\mathbf{x}^{(k)})\| \leq L\|\mathbf{x}^{(k)} - x^\star\| = L\|\mathbf{x}^{(k-1)} - \alpha_{k-1}\mathbf{\Psi}^{(k-1)} - \mathbf{x}^\star\| &\leq L\left(\|\mathbf{x}^{(k-1)} - \mathbf{x}^\star\| + \alpha_{k-1}\sqrt{d}C_1\right) \\
&\leq \ldots \leq L\left(\|\mathbf{x}^0 - \mathbf{x}^\star\| + \sqrt{d}C_1\sum_{s=0}^{k-1}\alpha_s\right) = L\left(\|\mathbf{x}^{(0)} - \mathbf{x}^\star\| + \sqrt{d}C_1A_k\right),
\end{aligned}$$

where we recall that $\mathbf{x}^\star \in \mathcal{X} = \left\{\mathbf{x} \in \mathbb{R}^d : \|\nabla f(\mathbf{x})\| = 0\right\}$ is any stationary point of f. Therefore, we have

$$\mathbb{E}\left[\exp(B_2)\right] \leq \exp\left(2NL^2D_{\mathcal{X}}(\mathbf{x}^{(0)})\sum_{k=0}^{t-1}\alpha_k^2 + 2dNC_1^2L^2\sum_{k=1}^{t}\alpha_k^2A_k^2\right),$$

where $D_{\mathcal{X}}(\mathbf{x}^{(0)}) = \inf_{\mathbf{x}\in\mathcal{X}}\|\mathbf{x}^{(0)} - \mathbf{x}\|^2$ is the distance of the initial model estimate from the set of stationary points. Combining everything, we get

$$\mathbb{E}\left[\exp(Z_t)\right] \leq \exp\left(f(\mathbf{x}^{(0)}) - f^\star + \left(dLC_1^2/2 + 2NL^2D_{\mathcal{X}}(\mathbf{x}^{(0)})\right)\sum_{k=1}^{t}\alpha_k^2 + 2dNC_1^2L^2\sum_{k=1}^{t}\alpha_k^2A_k^2\right).$$

Define $G_t \triangleq f(\mathbf{x}^{(0)}) - f^\star + \left(dLC_1/2 + 2NL^2D_{\mathcal{X}}(\mathbf{x}^{(0)})\right)\sum_{k=1}^{t}\alpha_k^2 + 2dNC_1^2L^2\sum_{k=1}^{t}\alpha_k^2A_k^2$. Using Markov's inequality, it then follows that, for any $\epsilon > 0$

$$\mathbb{P}(Z_t > \epsilon) \leq \exp(-\epsilon)\mathbb{E}[\exp(Z_t)] \leq \exp(-\epsilon + G_t) \iff \mathbb{P}(Z_t > \epsilon + G_t) \leq \exp(-\epsilon).$$

Finally, for any $\beta \in (0,1)$, with probability at least $1-\beta$, we have

$$Z_t \leq \log(1/\beta) + G_t \iff A_t^{-1}Z_t \leq A_t^{-1}\left(\log(1/\beta) + G_t\right). \tag{6}$$

Consider the step-size schedule $\alpha_t = \frac{a}{(t+2)^\delta}$, for $\delta \in (3/4, 1)$. Using upper and lower Darboux sums, we get

$$\begin{aligned}
\frac{a}{1-\delta}((t+2)^{1-\delta} - 2^{1-\delta}) \leq A_t \leq \frac{a}{1-\delta}((t+1)^{1-\delta} - 1), \\
\frac{a^2}{2\delta-1}(2^{1-2\delta} - (t+2)^{1-2\delta}) \leq \sum_{k=1}^{t}\alpha_k^2 \leq \frac{a^2}{2\delta-1}(1 - (t+1)^{1-2\delta}).
\end{aligned}$$

Plugging in (6), we then get, with probability at least $1-\beta$

$$\frac{\phi'(0)}{2}\sum_{k=0}^{t-1}\widetilde{\alpha}_k Z(\|\nabla f(x_k)\|) \leq \frac{(1-\delta)\left(f(\mathbf{x}^{(0)})-f^\star+\log(1/\beta)\right)}{a((t+2)^{1-\delta}-2^{1-\delta})} + \frac{a(1-\delta)(dLC_1^2/2+2NL^2D_{\mathcal{X}}(\mathbf{x}^{(0)}))}{(2\delta-1)((t+2)^{1-\delta}-2^{1-\delta})} + \frac{2a^3dNC_1^2L^2\sum_{k=0}^{t-1}(k+2)^{2-4\delta}}{(1-\delta)((t+2)^{1-\delta}-2^{1-\delta})}.$$

Using the upper Darboux sum once more, we have

$$\sum_{k=0}^{t-1}(k+2)^{2-4\delta} \leq \int_1^{t+1} k^{2-4\delta}dk \leq \frac{1}{4\delta-3},$$

therefore, combining everything, we finally get

$$\sum_{k=0}^{t-1}\widetilde{\alpha}_k Z(\|\nabla f(x_k)\|) \leq \frac{R(a,\beta,\delta)}{(t+2)^{1-\delta}-2^{1-\delta}},$$

where $R(a,\beta,\delta) = \frac{2(1-\delta)}{\phi'(0)}\left[\left(f(\mathbf{x}^{(0)})-f^\star+\log(1/\beta)\right)/a + \frac{a\left(dLC_1^2/2+2NL^2D_{\mathcal{X}}(\mathbf{x}^{(0)})\right)}{(2\delta-1)} + \frac{2a^3dNC_1^2L^2}{(1-\delta)^2(4\delta-3)}\right]$. □

3.3 Strongly Convex Costs

In this section we establish the convergence in high-probability for strongly convex functions. First, recall the definition of the Huber loss $H_\lambda : \mathbb{R} \mapsto [0,\infty)$, parametrized by $\lambda > 0$, e.g., Huber (1964), which is given by

$$H_\lambda(x) \triangleq \begin{cases} \frac{1}{2}x^2, & |x| \leq \lambda, \\ \lambda|x| - \frac{\lambda^2}{2}, & |x| > \lambda. \end{cases}$$

By the definition of Huber loss, it is not hard to see that it is a convex, non-decreasing function on $[0,\infty)$. Moreover, it follows that

$$\min\{\xi\|\nabla f(\mathbf{x}^{(k)})\|/\sqrt{d}, \|\nabla f(\mathbf{x}^{(k)})\|^2/d\} \triangleq Z(\|\nabla f(\mathbf{x}^{(k)})\|) \geq H_\xi(\|\nabla f(\mathbf{x}^{(k)})\|/\sqrt{d}) = \frac{1}{d}H_{\xi\sqrt{d}}(\|\nabla f(\mathbf{x}^{(k)})\|), \tag{7}$$

where the last inequality follows by noticing that $H_\xi(x/d) = \frac{1}{d}H_{\xi\sqrt{d}}(x)$. Next, recall that μ-strongly convex costs satisfy the *gradient domination property*, i.e., $\mu\|\mathbf{x}^{(k)}-\mathbf{x}^\star\| \leq \|\nabla f(\mathbf{x})\|$, for any $\mathbf{x}\in\mathbb{R}^d$, e.g., Nesterov (2018). Combining (7) with the gradient domination property, we get

$$\sum_{k=0}^{t-1}\widetilde{\alpha}_k Z(\|\nabla f(\mathbf{x}^{(k)})\|) \geq \frac{1}{d}\sum_{k=0}^{t-1}\widetilde{\alpha}_k H_{\xi\sqrt{d}}(\mu\|\mathbf{x}^{(k)}-\mathbf{x}^\star\|) \geq \frac{\mu^2}{d}H_{\xi\sqrt{d}/\mu}(\|\widehat{\mathbf{x}}^{(t)}-\mathbf{x}^\star\|),$$

where $\widehat{\mathbf{x}}^{(t)} \triangleq \sum_{k=0}^{t-1}\widetilde{\alpha}_k\mathbf{x}^{(k)}$ is the weighted average of the first t iterates, the first inequality follows from (7), the gradient domination property and the fact that H is non-decreasing, while the second property follows from the fact that H is convex and non-decreasing H and applying Jensen's inequality twice. Therefore, we have just shown the following.

Theorem 2. *Let Assumptions 1-4 hold, with the nonlinearity* $\mathbf{\Psi}: \mathbb{R}^d \mapsto \mathbb{R}^d$ *being component-wise, i.e., of the form* $\mathbf{\Psi}(\mathbf{x}) = \left[\mathcal{N}_1(x_1), \ldots, \mathcal{N}_1(x_d)\right]^\top$. *Let* $\{\mathbf{x}^{(t)}\}_{t\in\mathbb{N}_0}$ *be the sequence generated by* (3)*, with step-size* $\alpha_t = \frac{a}{(t+2)^\delta}$*, for any* $\delta \in (3/4, 1)$ *and* $a > 0$*. Then, for any* $t \in \mathbb{N}_0$*, and any* $\beta \in (0,1)$*, with probability at least* $1-\beta$*, it holds that*

$$H_{\xi\sqrt{d}/\mu}(\|\widehat{\mathbf{x}}^{(t)} - \mathbf{x}^\star\|) \leq \frac{\widetilde{R}(a,\beta,\delta)}{(t+2)^{1-\delta} - 2^{1-\delta}},$$

where $\widetilde{R}(a,\beta,\delta) \triangleq \frac{2d(1-\delta)}{\mu^2\phi'(0)}\left[\left(f(\mathbf{x}^{(0)}) - f^\star + \log(1/\beta)\right)/a + \frac{a(dLC_1^2/2 + 2NL^2 D_{\mathcal{X}}(\mathbf{x}^{(0)}))}{(2\delta-1)} + \frac{2a^3 dNC_1^2L^2}{(1-\delta)^2(4\delta-3)}\right]$.

The quantity $\widehat{\mathbf{x}}^{(t)}$ can be seen as a generalized Polyak-Ruppert estimator, e.g., Ruppert (1988); Polyak (1990); Polyak and Juditsky (1992). Similarly to Remark 5, it can be shown that the bound from Theorem 2 implies

$$\|\widehat{\mathbf{x}}^{(t)} - \mathbf{x}^\star\|^2 = \mathcal{O}\left(t^{-2(1-\delta)}\right),$$

giving a rate $\mathcal{O}\left(t^{-1/2+\epsilon}\right)$, for any $\epsilon < 1/2$, with an *exponent independent of noise and problem parameters* (see Appendix F for details). Our rate improves on the state-of-the-art rate from Sadiev et al. (2023), whenever $\eta < \frac{4}{3+2\epsilon}$. Note that Theorem 2 provides a bound on the weighted average of past iterates and component-wise nonlinearities. However, for strongly convex functions it is of interest to characterize the convergence guarantees of the last iterate, e.g., Harvey et al. (2019); Tsai et al. (2022); Sadiev et al. (2023). Moreover, we would like to establish high-probability convergence guarantees for a wider class of nonlinearities, including joint ones, like clipping and normalization. To that end, we first characterize the behaviour of the mapping $\mathbf{\Phi}$ in the general nonlinearity case.

Lemma 3.4. *Let Assumptions 1-4 hold and* $\{\mathbf{x}^{(t)}\}_{t\in\mathbb{N}_0}$ *be the sequence generated by* (3)*, with step-size* $\alpha_t = \frac{a}{(t+2)^\delta}$*, for any* $\delta \in (0.5, 1)$*,* $a > 0$*. Then, for some* $\gamma = \gamma(a) > 0$ *and any* $t \in \mathbb{N}_0$

$$\langle \mathbf{\Phi}^{(t)}, \nabla f(\mathbf{x}^{(t)})\rangle \geq \gamma(t+2)^{\delta-1}\|\nabla f(\mathbf{x}^{(t)})\|^2.$$

We are now ready to state the main result.

Theorem 3. *Suppose Assumptions 1-4 hold and* $\{\mathbf{x}^{(t)}\}_{t\in\mathbb{N}_0}$ *is the sequence generated by* (3)*, with step-size* $\alpha_t = \frac{a}{(t+2)^\delta}$*, for any* $\delta \in (0.5, 1)$ *and* $a > 0$*. Then, for any* $t \in \mathbb{N}_0$*, and any* $\beta \in (0,1)$*, with probability at least* $1-\beta$*, it holds that*

$$\|\mathbf{x}^{(t+1)} - \mathbf{x}^\star\|^2 \leq \frac{2\log(e/\beta)}{\mu B(t+2)^\zeta},$$

where $\zeta = \min\left\{2\delta - 1, a\mu\gamma/2\right\}$, $B = \min\left\{\frac{1}{(f(\mathbf{x}^{(0)}) - f^\star)}, \frac{\mu\gamma}{aL(2N + C^2/2)}\right\}$.

We specialize the value of $\gamma = \gamma(a)$ for different nonlinearities and discuss its impact on the rate in Appendix D. The value of ζ can be explicitly calculated for specific choices of nonlinearities and noise. We now give some examples.

Example 4. For the noise from Example 1 and sign nonlinearity, it can be shown that $\zeta \approx \min\left\{2\delta - 1, \frac{\mu}{L}\frac{1-\delta}{\sqrt{d}}\frac{\alpha-1}{\alpha}\right\}$, see Jakovetić et al. (2023). For the same noise and cclip, it can

be shown that $\zeta \approx \min\left\{2\delta-1, \frac{\mu}{L\sqrt{d}}\frac{(1-\delta)(m-1)(1-(m+1)^{-\alpha})}{m}\right\}$. On the other hand, the rate from Sadiev et al. (2023) for gradient clipping, adapted to the same noise, is $2(r-1)/r$, where $r \leq \min\{\alpha-1, 2\}$. While $2(r-1)/r > \zeta$ for α above a certain threshold, as $\alpha \to 2$, we have $2(r-1)/r \to 0$, while our rate ζ stays strictly positive and bounded away from zero for both sign and cclip. Therefore, for α sufficiently close to 2, we have $\zeta > \frac{2(r-1)}{r}$, i.e., our rate is better than the one in state-of-the-art Sadiev et al. (2023).

Proof of Theorem 3. Using L-smoothness of f, the update rule (4) and Lemma 3.4, we have

$$\begin{aligned} f(\mathbf{x}^{(t+1)}) &\leq f(\mathbf{x}^{(t)}) - \alpha_t\langle\nabla f(\mathbf{x}^{(t)}), \mathbf{\Phi}^{(t)} - \mathbf{e}^{(t)}\rangle + \frac{\alpha_t^2 L}{2}\|\mathbf{\Psi}^{(t)}\|^2 \\ &\leq f(\mathbf{x}^{(t)}) - \frac{a\gamma\|\nabla f(\mathbf{x}^{(t)})\|^2}{(t+2)} + \frac{a\langle\nabla f(\mathbf{x}^{(t)}), \mathbf{e}^{(t)}\rangle}{(t+2)^{\delta}} + \frac{a^2LC^2}{2(t+2)^{2\delta}}. \end{aligned}$$

Subtracting $f^\star$ from both sides of the inequality, defining $F^{(t)} = f(\mathbf{x}^{(t)}) - f^\star$ and using μ-strong convexity of f, we get

$$F^{(t+1)} \leq \left(1 - \frac{2\mu a\gamma}{t+2}\right)F^{(t)} + \frac{a\langle\nabla f(\mathbf{x}^{(t)}), \mathbf{e}^{(t)}\rangle}{(t+2)^{\delta}} + \frac{a^2LC^2}{2(t+2)^{2\delta}}. \tag{8}$$

Let $\zeta = \min\{2\delta-1, a\gamma\mu/2\}$. Defining $Y^{(t+1)} \triangleq (t+2)^\zeta F^{(t+1)} = (t+2)^\zeta(f(\mathbf{x}^{(t+1)}) - f^\star)$, from (8) we get

$$Y^{(t+1)} \leq a_t Y^{(t)} + b_t\langle\nabla f(\mathbf{x}^{(t)}), \mathbf{e}^{(t)}\rangle + c_t V, \tag{9}$$

where $a_t = \left(1 - \frac{2\mu a\gamma}{t+2}\right)\left(\frac{t+2}{t+1}\right)^\zeta$, $b_t = \frac{a}{(t+2)^{\delta-\zeta}}$, $c_t = \frac{a^2}{(t+2)^{2\delta-\zeta}}$ and $V = \frac{LC^2}{2}$. Denote the MGF of $Y^{(t)}$ conditioned on $\mathcal{F}_t$ as $M_{t+1|t}(\nu) = \mathbb{E}\left[\exp\left(\nu Y^{(t+1)}\right)|\mathcal{F}_t\right]$. We then have, for any $\nu \geq 0$

$$\begin{aligned} M_{t+1|t}(\nu) &\overset{(a)}{\leq} \mathbb{E}\left[\exp\left(\nu(a_tY^{(t)} + b_t\langle\mathbf{e}^{(t)}, \nabla f(\mathbf{x}^{(t)})\rangle + c_tV)\right)\Big|\,\mathcal{F}_t\right] \\ &\overset{(b)}{\leq} \exp(\nu a_tY^{(t)} + \nu c_tV)\mathbb{E}\left[\exp(\nu b_t\langle\mathbf{e}^{(t)}, \nabla f(\mathbf{x}^{(t)})\rangle)\Big|\,\mathcal{F}_t\right] \\ &\overset{(c)}{\leq} \exp\left(\nu a_tY^{(t)} + \nu c_tV + \nu^2 b_t^2 N\|\nabla f(\mathbf{x}^{(t)})\|^2\right) \\ &\overset{(d)}{\leq} \exp\left(\nu a_tY^{(t)} + \nu c_tV + 2\nu^2 b_t'^2 LNY^{(t)}\right), \end{aligned} \tag{10}$$

where (a) follows from (9), (b) follows from the fact that $Y^{(t)}$ is $\mathcal{F}_t$ measurable, (c) follows from Lemma 3.1, in (d) we use $\|\nabla f(\mathbf{x})\|^2 \leq 2L(f(\mathbf{x}) - f^\star)$ and define $b_t' = a\frac{(t+1)^{\frac{-\zeta}{2}}}{(t+2)^{\delta-\zeta}}$, so that $b_t = (t+1)^{\frac{\zeta}{2}}b_t'$. For the choice $0 \leq \nu \leq B$, for some $B > 0$ (to be specified later), we get

$$M_{t+1|t}(\nu) \leq \exp\left(\nu(a_t + 2b_t'^2LNB)Y^{(t)}\right)\exp\left(\nu c_tV\right).$$

Taking the full expectation, we get

$$M_{t+1}(\nu) \leq M_t((a_t + 2b_t'^2LNB)\nu)\exp(\nu c_tV). \tag{11}$$

Similarly to the approach in Harvey et al. (2019), we now want to show that $M_t(\nu) \leq e^{\frac{\nu}{B}}$, for any $0 \leq \nu \leq B$ and any $t \geq 0$. We proceed by induction. For $t = 0$, we have

$$M_0(\nu) = \exp(\nu Y^{(0)}) = \exp\left(\nu(f(\mathbf{x}^{(0)}) - f^\star)\right),$$

where we simply used the definition of $Y^{(t)}$ and the fact that it is deterministic for $t = 0$. Choosing $B \leq (f(\mathbf{x}^{(0)}) - f^\star)^{-1}$ ensures that $M_0(\nu) \leq e^{\frac{\nu}{B}}$. Next, assume that for some $t \geq 1$ it holds that $M_t(\nu) \leq e^{\frac{\nu}{B}}$. We then have

$$M_{t+1}(\nu) \leq M_t((a_t + 2b_t'^2 LNB)\nu)\exp(\nu c_t V) \leq \exp\left((a_t + 2b_t'^2 LNB + c_t VB)\frac{\nu}{B}\right),$$

where we use (11) in the first and the induction hypothesis in the second inequality. For our claim to hold, it suffices to show $a_t + 2b_t'^2 LNB + c_t VB \leq 1$. Plugging in the values of a_t, b_t' and c_t, we have

$$\begin{aligned} a_t + 2b_t'^2 LNB + c_t VB &= \left(1 - \frac{2\mu a\gamma}{t+2}\right)\left(\frac{t+2}{t+1}\right)^\zeta + \frac{2a^2 LNB}{(t+2)^{2\delta-2\zeta}(t+1)^\zeta} + \frac{a^2 VB}{(t+2)^{2\delta-\zeta}} \\ &\leq \left(\frac{t+2}{t+1}\right)^\zeta \left(1 - \frac{2\mu a\gamma}{t+2} + \frac{2a^2 LNB}{(t+2)^{2\delta-\zeta}} + \frac{a^2 VB(t+1)^\zeta}{(t+2)^{2\delta}}\right) \\ &\leq \left(\frac{t+2}{t+1}\right)^\zeta \left(1 - \frac{2\mu a\gamma}{t+2} + \frac{2a^2 LNB}{(t+2)^{2\delta-\zeta}} + \frac{a^2 VB}{(t+2)^{2\delta-\zeta}}\right). \end{aligned}$$

Noticing that $2\delta - \zeta \geq 1$ and setting $B = \min\left\{\frac{1}{(t_0-1)^\zeta(f(\mathbf{x}^{(0)})-f^\star)}, \frac{\mu\gamma}{aL(2N+C^2/2)}\right\}$, gives

$$a_t + 2b_t'^2 LNB + c_t VB \leq \left(\frac{t+2}{t+1}\right)^\zeta \left(1 - \frac{\mu a\gamma}{t+2}\right) \leq \exp\left(\frac{\zeta}{t+1} - \frac{a\mu\gamma}{t+2}\right) \leq 1,$$

where in the second inequality we use $1 + x \leq e^x$, while the third inequality follows from the choice of ζ. Therefore, we have shown that $M_t(\nu) \leq e^{\frac{\nu}{B}}$, for any $t \geq 0$ and any $0 \leq \nu \leq B$. By Markov's inequality, it readily follows that

$$\mathbb{P}(f(\mathbf{x}^{(t+1)}) - f^\star \geq \epsilon) = \mathbb{P}(Y_{t+1} \geq (t+2)^\zeta \epsilon) \leq e^{-\nu(t+2)^\zeta \epsilon} M_{t+1}(\nu) \leq e^{1 - B(t+2)^\zeta \epsilon},$$

where in the last inequality we set $\nu = B$. Finally, using strong convexity, we have

$$\mathbb{P}(\|\mathbf{x}^{(t+1)} - \mathbf{x}^\star\|^2 \geq \epsilon) \leq \mathbb{P}\left(f(\mathbf{x}^{(t+1)}) - f^\star \geq \frac{\mu}{2}\epsilon\right) \leq e e^{-B(t+2)^\zeta \frac{\mu}{2}\epsilon},$$

which implies that, for any $\beta \in (0,1)$, with probability at least $1 - \beta$,

$$\|\mathbf{x}^{(t+1)} - \mathbf{x}^\star\|^2 \leq \frac{2\log(e/\beta)}{\mu B(t+2)^\zeta},$$

completing the proof. □

4 Analytical and Numerical Studies

In this section we present analytical and numerical studies in the case of strongly convex costs, specializing the rate ζ from Theorem 3 for different nonlinearities. As we proceed to show, the rates provided by our theory, while general, are able to uncover important phenomena, namely which choice of nonlinearity is preferred for the given problem settings. Subsection 4.1 provides the analytical study, while Subsection 4.2 provides the accompanying numerical results.

4.1 Analytical Study

In this section we provide an analytical study, using the rate ζ obtained in Theorem 3, with the goal of showing that clipping, exclusively considered in prior works, is not always the optimal choice of nonlinearity. Recall the constants where $\phi'(0)$, ξ are constants defined in Section 3.2. We then have the following result, whose proof can be found in Appendix E.

Lemma 4.1. *For any strongly convex cost with Lipschitz continuous gradients and the noise from Example 1, a component-wise nonlinearity is preferred to joint clipping whenever*

$$\frac{\phi'(0)\xi}{C_1} \geq 8\sqrt{d}\left(1 - \frac{1}{(1 + {}^{B_0}/\sqrt{d})^{\alpha-1}}\right)^d. \tag{12}$$

For sign and cclip it can be shown that the value $\frac{\phi'(0)\xi}{C_1}$ is approximately $\frac{\alpha-1}{\alpha}$ and $\frac{m-1}{m}(1-(m+1)^{-\alpha})$, respectively, where $\alpha > 2$ is the constant from Example 1, while $m > 1$ is the clipping radius for cclip (see Jakovetić et al. (2023)). Clearly, if B_0 is fixed, one can find a large enough d for which the relation (12) holds for both sign and cclip. On the other hand, for a fixed d, if $B_0 \to +\infty$, then the relation does not hold. If both B_0 and d grow to infinity, we notice two regimes:

1. $B_0 = o(\sqrt{d})$, e.g., $B_0 = d^{1/4}$, the relation (12) holds for sufficiently large d.
2. $d = o(B_0)$, e.g., $B_0 = d^2$, the relation (12) does not hold.

Hence, for B_0 finite or growing at an appropriate rate, we can always find a d large enough, for which the relation from Lemma 4.1 holds. We verify this numerically in Figure 1, where we plot the behaviour of the right-hand side of (12) when $B_0 = d^2$, $B_0 = d^{1/4}$ and $B_0 = 100$ (i.e., constant) on the y-axis, versus the problem dimension d on the x-axis. The straight lines show the left-hand side of (12), when specialized to sign and cclip, i.e., $\frac{\alpha-1}{\alpha}$ and $\frac{m-1}{m}(1-(m+1)^{-\alpha})$, for $\alpha = 2.05$ and $m = 2$. We can clearly see that, as d increases, the right-hand side either blows up ($B_0 = d^2$), rapidly decreases to zero ($B_0 = d^{1/4}$), or initially blows up, but decreases to zero for d sufficiently large ($B_0 = const.$), as claimed. In the subplot we zoom in on the lines representing the behaviour of the left-hand side of (12) for sign and cclip, which suggests that sign is a better choice of nonlinearity in this instance. Therefore, there are regimes for which our theory suggests clipping is *not the optimal choice of nonlinearity*. This is in line with the observations made in Zhang et al. (2020), who noted that, compared to joint clipping, cclip converges faster and achieves a better dependence on problem dimension.

4.2 Numerical Results

In this section we verify our analytical findings numerically. We consider a quadratic problem $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top A\mathbf{x} + \mathbf{b}^\top\mathbf{x}$, where $A \in \mathbb{R}^{d\times d}$ is positive definite, with $\mathbf{b} \in \mathbb{R}^d$ a fixed vector. We set $d = 100$. The stochastic noise is generated according to the component-wise noise PDF from Example 1, with $\alpha = 2.05$. We compare the performance of sign, component-wise and joint clipped SGD, with all three algorithms using the step-size schedule $\alpha_t = \frac{1}{t+2}$. For clipping based algorithms, we choose the clipping thresholds m and M for which component-wise and joint clipped SGD performed the best, that being $m = 1$ and $M = 100$. All three algorithms are initialized at the zero vector and perform $T = 25000$ iterations, across $R = 5000$ runs. To evaluate the performance of the methods, we use the following criteria:

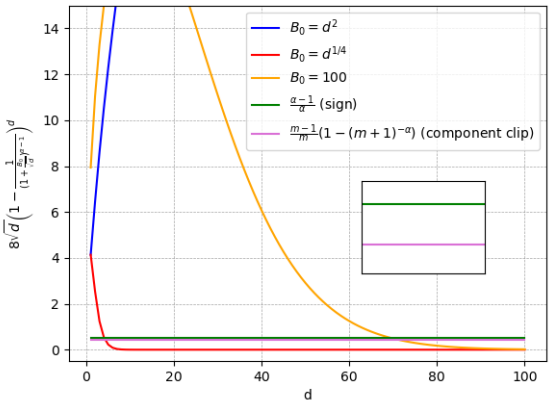

Figure 1: Numerical verification of the inequality from Lemma 4.1, for different values of d and B_0. We can see that the inequality holds for B_0 finite or growing at an appropriate rate.

1. *Mean-squared error (MSE)*: we present the MSE of the algorithms, by evaluating the model gap $\|\mathbf{x}^{(t)} - \mathbf{x}^\star\|^2$ in each iteration, averaged across all runs, i.e., the final estimator at iteration $t = 1, \ldots, T$, is given by $MSE^t = \frac{1}{R}\sum_{r=1}^{R} \|\mathbf{x}_r^{(t)} - \mathbf{x}^\star\|^2$, where $\mathbf{x}_r^{(t)}$ is the t-th iterate in the r-th run, generated by the nonlinear SGD algorithms.

2. *High-probability estimate*: we evaluate the high-probability behaviour of the methods, as follows. We consider the events $A^t = \{\|\mathbf{x}^{(t)} - \mathbf{x}^\star\|^2 > \varepsilon\}$, for a fixed $\varepsilon > 0$. To estimate the probability of A^t, for each $t = 1, \ldots, T$, we construct a Monte Carlo estimator of the empirical probability, by sampling $n = 3000$ instances from the $R = 5000$ runs, uniformly with replacement. We then obtain the empirical probability estimator as $\mathbb{P}_n(A^t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_i(A^t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\left(\{\|\mathbf{x}_i^{(t)} - \mathbf{x}^\star\|^2 > \varepsilon\}\right)$, where $\mathbb{I}(\cdot) \in \{0, 1\}$ is the indicator function, with $\mathbf{x}_i^{(t)}$ the i-th Monte Carlo sample.

The results are presented in Figure 2. We can see that component-wise nonlinearities outperform joint clipping, both in terms of MSE and high-probability performance, thus validating our analytical findings from Section 4, further underlining the usefulness of the exponent ζ.

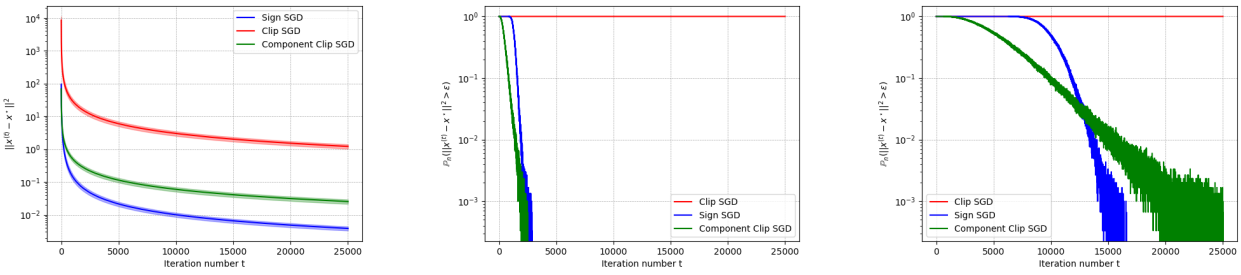

Figure 2: Performance of sign, cclip and joint clipping for $d = 10$. Left to right: MSE performance and high-probability performance for $\varepsilon = \{0.1, 0.01\}$, respectively.

5 Conclusion

We present high-probability convergence guarantees for a broad class of nonlinear streaming SGD algorithms, under heavy-tailed noise. Our results are built on a general framework, that encompasses many popular nonlinear versions of SGD, such as clipped, normalized, quantized and sign SGD, providing high-probability convergence guarantees for both non-convex and strongly convex functions. Compared to state-of-the-art works Nguyen et al. (2023a); Sadiev et al. (2023), we extend the high-probability convergence guarantees to novel nonlinearities, relax the noise moment condition, and demonstrate regimes in which our convergence rates are better than state-of-the-art. Moreover, for strongly convex functions we show that our rates are informative for the optimal choice of nonlinearity and that clipping, exclusively considered in prior works, is not always the optimal choice of nonlinearity, further highlighting the importance of our general framework. Numerical results confirm the theoretical findings and show the usefulness of our general framework for informing on the optimal choice of nonlinearity.

Bibliography

Alistarh, D., Grubic, D., Li, J., Tomioka, R., and Vojnovic, M. (2017). Qsgd: Communication-efficient sgd via gradient quantization and encoding. *Advances in neural information processing systems*, 30. (Cited on page 2.)

Bercovici, H., Pata, V., and Biane, P. (1999). Stable laws and domains of attraction in free probability theory. *Annals of Mathematics*, 149(3):1023–1060. (Cited on page 8.)

Bernstein, J., Wang, Y.-X., Azizzadenesheli, K., and Anandkumar, A. (2018a). signsgd: Compressed optimisation for non-convex problems. In *International Conference on Machine Learning*, pages 560–569. PMLR. (Cited on page 2.)

Bernstein, J., Zhao, J., Azizzadenesheli, K., and Anandkumar, A. (2018b). signsgd with majority vote is communication efficient and fault tolerant. *arXiv preprint arXiv:1810.05291*. (Cited on page 2.)

Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010: 19th International Conference on Computational StatisticsParis France, August 22-27, 2010 Keynote, Invited and Contributed Papers*, pages 177–186. Springer. (Cited on page 2.)

Bottou, L., Curtis, F. E., and Nocedal, J. (2018). Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311. (Cited on page 2.)

Chen, X., Wu, S. Z., and Hong, M. (2020). Understanding gradient clipping in private sgd: A geometric perspective. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 13773–13782. Curran Associates, Inc. (Cited on page 2.)

Crawshaw, M., Liu, M., Orabona, F., Zhang, W., and Zhuang, Z. (2022). Robustness to unbounded smoothness of generalized signsgd. *Advances in Neural Information Processing Systems*, 35:9955–9968. (Cited on page 2.)

Cutkosky, A. and Mehta, H. (2020). Momentum improves normalized sgd. In *International conference on machine learning*, pages 2260–2268. PMLR. (Cited on page 2.)

Das, R., Hashemi, A., Sanghavi, S., and Dhillon, I. S. (2021). Dp-normfedavg: Normalizing client updates for privacy-preserving federated learning. *arXiv preprint arXiv:2106.07094*. (Cited on page 2.)

Davis, D., Drusvyatskiy, D., Xiao, L., and Zhang, J. (2021). From low probability to high confidence in stochastic convex optimization. *The Journal of Machine Learning Research*, 22(1):2237–2274. (Cited on page 2.)

Eldowa, K. and Paudice, A. (2023). General tail bounds for non-smooth stochastic mirror descent. *arXiv preprint arXiv:2312.07142*. (Cited on page 3.)

Gama, J. (2012). A survey on learning from data streams: current and future trends. *Progress in Artificial Intelligence*, 1(1):45–55. (Cited on page 1.)

Gandikota, V., Kane, D., Maity, R. K., and Mazumdar, A. (2021). vqsgd: Vector quantized stochastic gradient descent. In *International Conference on Artificial Intelligence and Statistics*, pages 2197–2205. PMLR. (Cited on page 2.)

Ghadimi, S. and Lan, G. (2012). Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization i: A generic algorithmic framework. *SIAM Journal on Optimization*, 22(4):1469–1492. (Cited on page 2.)

Ghadimi, S. and Lan, G. (2013). Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368. (Cited on pages 2, 3, and 7.)

Gorbunov, E., Danilova, M., and Gasnikov, A. (2020). Stochastic optimization with heavy-tailed noise via accelerated gradient clipping. *Advances in Neural Information Processing Systems*, 33:15042–15053. (Cited on pages 2 and 3.)

Gorbunov, E., Danilova, M., Shibaev, I., Dvurechensky, P., and Gasnikov, A. (2021). Near-optimal high probability complexity bounds for non-smooth stochastic optimization with heavy-tailed noise. *arXiv preprint arXiv:2106.05958*. (Cited on page 3.)

Hardt, M., Recht, B., and Singer, Y. (2016). Train faster, generalize better: Stability of stochastic gradient descent. In *International conference on machine learning*, pages 1225–1234. PMLR. (Cited on page 2.)

Harvey, N. J., Liaw, C., Plan, Y., and Randhawa, S. (2019). Tight analyses for non-smooth stochastic gradient descent. In *Conference on Learning Theory*, pages 1579–1613. PMLR. (Cited on pages 2, 3, 13, and 14.)

Hazan, E. and Kale, S. (2014). Beyond the regret minimization barrier: optimal algorithms for stochastic strongly-convex optimization. *The Journal of Machine Learning Research*, 15(1):2489–2512. (Cited on page 3.)

Hazan, E., Levy, K., and Shalev-Shwartz, S. (2015). Beyond convexity: Stochastic quasi-convex optimization. *Advances in neural information processing systems*, 28. (Cited on page 2.)

Huber, P. J. (1964). Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics*, 35(1):73 – 101. (Cited on page 12.)

Jakovetić, D., Bajović, D., Sahu, A. K., Kar, S., Milošević, N., and Stamenković, D. (2023). Nonlinear gradient mappings and stochastic optimization: A general framework with applications to heavy-tail noise. *SIAM Journal on Optimization*, 33(2):394–423. (Cited on pages 2, 4, 13, 16, and 25.)

Jin, C., Netrapalli, P., Ge, R., Kakade, S. M., and Jordan, M. I. (2019). A short note on concentration inequalities for random vectors with subgaussian norm. *arXiv preprint arXiv:1902.03736*. (Cited on page 8.)

Jin, J., Zhang, B., Wang, H., and Wang, L. (2021). Non-convex distributionally robust optimization: Non-asymptotic analysis. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*, volume 34, pages 2771–2782. Curran Associates, Inc. (Cited on page 3.)

Krawczyk, B., Minku, L. L., Gama, J., Stefanowski, J., and Woźniak, M. (2017). Ensemble learning for data stream analysis: A survey. *Information Fusion*, 37:132–156. (Cited on page 1.)

Lan, G. (2012). An optimal method for stochastic composite optimization. *Mathematical Programming*, 133(1-2):365–397. (Cited on page 3.)

Li, S. and Liu, Y. (2022). High probability guarantees for nonconvex stochastic gradient descent with heavy tails. In *International Conference on Machine Learning*, pages 12931–12963. PMLR. (Cited on page 3.)

Li, X. and Orabona, F. (2020). A high probability analysis of adaptive sgd with momentum. *arXiv preprint arXiv:2007.14294*. (Cited on page 3.)

Liu, Z., Nguyen, T. D., Nguyen, T. H., Ene, A., and Nguyen, H. (2023a). High probability convergence of stochastic gradient methods. In *International Conference on Machine Learning*, pages 21884–21914. PMLR. (Cited on pages 3 and 7.)

Liu, Z., Zhang, J., and Zhou, Z. (2023b). Breaking the lower bound with (little) structure: Acceleration in non-convex stochastic optimization with heavy-tailed noise. In Neu, G. and Rosasco, L., editors, *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 2266–2290. PMLR. (Cited on page 3.)

Madden, L., Dall'Anese, E., and Becker, S. (2020). High-probability convergence bounds for non-convex stochastic gradient descent. *arXiv preprint arXiv:2006.05610*. (Cited on page 7.)

Moulines, E. and Bach, F. (2011). Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc. (Cited on page 2.)

Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A. (2009). Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609. (Cited on pages 2 and 3.)

Nesterov, Y. (2018). *Lectures on Convex Optimization*. Springer Publishing Company, Incorporated, 2nd edition. (Cited on pages 7, 12, and 23.)

Nguyen, T. D., Nguyen, T. H., Ene, A., and Nguyen, H. (2023a). Improved convergence in high probability of clipped gradient methods with heavy tailed noise. In Oh, A., Neumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S., editors, *Advances in Neural Information Processing Systems*, volume 36, pages 24191–24222. Curran Associates, Inc. (Cited on pages 2, 3, 4, 5, 8, 10, and 18.)

Nguyen, T. D., Nguyen, T. H., Ene, A., and Nguyen, H. L. (2023b). High probability convergence of clipped-sgd under heavy-tailed noise. *arXiv preprint arXiv:2302.05437*. (Cited on page 3.)

Parletta, D. A., Paudice, A., Pontil, M., and Salzo, S. (2022). High probability bounds for stochastic subgradient schemes with heavy tailed noise. *arXiv preprint arXiv:2208.08567*. (Cited on page 3.)

Polyak, B. (1990). New stochastic approximation type procedures. *Avtomatica i Telemekhanika*, 7:98–107. (Cited on page 13.)

Polyak, B. and Juditsky, A. (1992). Acceleration of Stochastic Approximation by Averaging. *SIAM Journal on Control and Optimization*, 30:838–855. (Cited on page 13.)

Polyak, B. and Tsypkin, Y. (1979). Adaptive estimation algorithms: Convergence, optimality, stability. *Automation and Remote Control*, 1979. (Cited on page 9.)

Puchkin, N., Gorbunov, E., Kutuzov, N., and Gasnikov, A. (2023). Breaking the heavy-tailed noise barrier in stochastic optimization problems. *arXiv preprint arXiv:2311.04161*. (Cited on page 3.)

Rakhlin, A., Shamir, O., and Sridharan, K. (2012). 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. (Cited on page 2.)

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407. (Cited on page 2.)

Ruppert, D. (1988). Efficient Estimations from a Slowly Convergent Robbins-Monro Process. Technical Report 781, Cornell University Operations Research and Industrial Engineering. (Cited on page 13.)

Sadiev, A., Danilova, M., Gorbunov, E., Horváth, S., Gidel, G., Dvurechensky, P., Gasnikov, A., and Richtárik, P. (2023). High-probability bounds for stochastic optimization and variational inequalities: the case of unbounded variance. In *International Conference on Machine Learning*, pages 29563–29648. PMLR. (Cited on pages 2, 3, 4, 5, 8, 10, 13, 14, and 18.)

Şimşekli, U., Gürbüzbalaban, M., Nguyen, T. H., Richard, G., and Sagun, L. (2019). On the heavy-tailed theory of stochastic gradient descent for deep neural networks. *arXiv preprint arXiv:1912.00018*. (Cited on page 8.)

Simsekli, U., Sagun, L., and Gurbuzbalaban, M. (2019). A tail-index analysis of stochastic gradient noise in deep neural networks. In *International Conference on Machine Learning*, pages 5827–5837. PMLR. (Cited on page 2.)

Tsai, C.-P., Prasad, A., Balakrishnan, S., and Ravikumar, P. (2022). Heavy-tailed streaming statistical estimation. In Camps-Valls, G., Ruiz, F. J. R., and Valera, I., editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 1251–1282. PMLR. (Cited on pages 3, 4, and 13.)

Vershynin, R. (2018). *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press. (Cited on page 8.)

Woodworth, B., Patel, K. K., Stich, S., Dai, Z., Bullins, B., Mcmahan, B., Shamir, O., and Srebro, N. (2020a). Is local SGD better than minibatch SGD? In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 10334–10343. PMLR. (Cited on page 2.)

Woodworth, B. E., Patel, K. K., and Srebro, N. (2020b). Minibatch vs local sgd for heterogeneous distributed learning. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6281–6292. Curran Associates, Inc. (Cited on page 2.)

Wright, S. J. and Recht, B. (2022). *Optimization for Data Analysis*. Cambridge University Press. (Cited on page 7.)

Yang, X., Zhang, H., Chen, W., and Liu, T.-Y. (2022). Normalized/clipped sgd with perturbation for differentially private non-convex optimization. *arXiv preprint arXiv:2206.13033*. (Cited on pages 2 and 3.)

You, Y., Li, J., Hseu, J., Song, X., Demmel, J., and Hsieh, C.-J. (2019). Reducing bert pre-training time from 3 days to 76 minutes. *arXiv preprint arXiv:1904.00962*, 12. (Cited on page 2.)

Yu, S. and Kar, S. (2023). Secure distributed optimization under gradient attacks. *IEEE Transactions on Signal Processing*, 71:1802–1816. (Cited on page 2.)

Zhang, J., He, T., Sra, S., and Jadbabaie, A. (2019). Why gradient clipping accelerates training: A theoretical justification for adaptivity. In *International Conference on Learning Representations*. (Cited on page 2.)

Zhang, J., Karimireddy, S. P., Veit, A., Kim, S., Reddi, S., Kumar, S., and Sra, S. (2020). Why are adaptive methods good for attention models? *Advances in Neural Information Processing Systems*, 33:15383–15393. (Cited on pages 2, 3, and 16.)

Zhang, X., Chen, X., Hong, M., Wu, Z. S., and Yi, J. (2022). Understanding clipping for federated learning: Convergence and client-level differential privacy. In *International Conference on Machine Learning, ICML 2022*. (Cited on page 2.)

A Introduction

The Appendix presents results omitted from the main body. Appendix B provides some useful facts and results used in the proofs. Appendix C provides the proofs omitted from Section 3. Appendix D provides rate expressions for component-wise and joint nonlinearities. Appendix E presents results omitted from Section 4. Appendix F provides some further derivations.

B Useful Facts

In this section we present some useful facts and results, concerning L-smooth, μ-strongly functions, bounded random vectors and the behaviour of nonlinearities.

Fact 1. *Let $f : \mathbb{R}^d \mapsto \mathbb{R}$ be L-smooth, μ-strongly convex, and let $\mathbf{x}^\star = \arg\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x})$. Then, for any $\mathbf{x} \in \mathbb{R}^d$, we have*

$$2\mu \left(f(\mathbf{x}) - f(\mathbf{x}^\star)\right) \leq \|\nabla f(\mathbf{x})\|^2 \leq 2L \left(f(\mathbf{x}) - f(\mathbf{x}^\star)\right).$$

Proof. The proof of the upper bound follows by plugging $y = \mathbf{x}$, $x = \mathbf{x}^\star$ in equation (2.1.10) of Theorem 2.1.5 from Nesterov (2018). The proof of the lower bound similarly follows by plugging $y = \mathbf{x}$, $x = \mathbf{x}^\star$ in equation (2.1.24) of Theorem 2.1.10 from Nesterov (2018). □

Fact 2. *Let $X \in \mathbb{R}^d$ be a zero-mean, bounded random vector, i.e., $\mathbb{E}X = 0$ and $\|X\| \leq \sigma$, for some $\sigma > 0$. Then, X is sub-Gaussian, i.e., there exists a positive constant $N = N(\sigma)$ such that, for any $\lambda \in \mathbb{R}^d$, we have*

$$\mathbb{E}e^{\langle X,\lambda\rangle} \leq e^{\frac{N\|\lambda\|^2}{2}}.$$

Proof. We begin by first showing that a scalar Rademacher random variable is sub-Gaussian. Recall that a random variable ε is a Rademacher random variable, if ε takes the values -1 and 1, both with probability $1/2$. We then have, for any $t \in \mathbb{R}$

$$\mathbb{E}e^{\varepsilon t} \overset{(a)}{=} \frac{1}{2}\left(e^t + e^{-t}\right) \overset{(b)}{=} \sum_{k=0}^{\infty} \frac{t^{2k}}{(2k)!} \overset{(c)}{\leq} \sum_{k=0}^{\infty} \frac{t^{2k}}{2^k k!} = \sum_{k=0}^{\infty} \frac{(t^2/2)^k}{k!} = e^{\frac{t^2}{2}}, \tag{13}$$

where (a) follows from the definition of the Rademacher random variable, (b) follows from the fact that $e^t = \sum_{k=0}^{\infty} \frac{t^k}{k!}$, for any $t \in \mathbb{R}$, while (c) follows from the fact that $2k! \geq 2^k k!$, for any $k \in \mathbb{N}_0$. Let $X' \in \mathbb{R}^d$ be an independent, identically distributed copy of X. We then have, for any $\lambda \in \mathbb{R}^d$

$$\begin{aligned}\mathbb{E}e^{\langle X,\lambda\rangle} \overset{(a)}{=} \mathbb{E}e^{\langle X-\mathbb{E}X',\lambda\rangle} \overset{(b)}{\leq} \mathbb{E}_{X,X'}e^{\langle X-X',\lambda\rangle} \overset{(c)}{=} \mathbb{E}_{X,X'}\mathbb{E}_{\varepsilon}e^{\varepsilon\langle X-X',\lambda\rangle} \overset{(d)}{\leq} \mathbb{E}_{X,X'}e^{\frac{(\langle X-X',\lambda\rangle)^2}{2}} \\ \leq \mathbb{E}_{X,X'}e^{\frac{\|X-X'\|^2\|\lambda\|^2}{2}} \overset{(e)}{\leq} e^{2\sigma^2\|\lambda\|^2},\end{aligned}$$

where (a) follows from the fact that X is zero-mean, in (b) we use Jensen's inequality, (c) uses the fact that X, X' are i.i.d., therefore $X - X'$ has the same distribution as $\varepsilon(X - X')$, where ε is a Rademacher random variable, in (d) we use (13), while (e) follows from the boundedness of X. Choosing $N = 4\sigma^2$, the desired inequality follows. □

C Missing Proofs

In this section we provide proofs of Lemmas 3.1, 3.3 and 3.4, omitted from the main body. We begin by proving Lemma 3.1.

Proof of Lemma 3.1. Recall the definition of the error vector $\mathbf{e}^{(t)} \triangleq \mathbf{\Phi}^{(t)} - \mathbf{\Psi}^{(t)}$, where $\mathbf{\Phi}^{(t)} \triangleq \mathbb{E}_{\mathbf{z}^{(t)}}\left[\mathbf{\Psi}(\nabla f(\mathbf{x}^{(t)}) + \mathbf{z}^{(t)})\right]$ is the denoised version of $\mathbf{\Psi}^{(t)}$. By definition, it then follows that

$$\mathbb{E}\left[\mathbf{e}^{(t)} |\, \mathcal{F}_t\right] = \mathbb{E}\left[\mathbf{\Phi}^{(t)} - \mathbf{\Psi}^{(t)} |\, \mathcal{F}_t\right] = \mathbf{\Phi}^{(t)} - \mathbb{E}\left[\mathbf{\Psi}^{(t)} |\, \mathcal{F}_t\right] = 0,$$

where the last equality follows from the fact that $\mathbf{\Phi}^{(t)}$ is $\mathcal{F}_t$-measureable and $\mathbb{E}\left[\mathbf{\Psi}^{(t)} |\, \mathcal{F}_t\right] = \mathbb{E}_{\mathbf{z}^{(t)}}\left[\mathbf{\Psi}(\nabla f(\mathbf{x}^{(t)}) + \mathbf{z}^{(t)})\right] = \mathbf{\Phi}^{(t)}$. Moreover, by Assumption 1, we have

$$\|\mathbf{e}^{(t)}\| = \|\mathbf{\Phi}^{(t)} - \mathbf{\Psi}^{(t)}\| \leq \|\mathbf{\Phi}^{(t)}\| + \|\mathbf{\Psi}^{(t)}\| \leq \mathbb{E}\|\mathbf{\Psi}^{(t)}\| + C \leq 2C,$$

which proves the first claim. The second claim readily follows by using the fact that $\mathbf{e}^{(t)}$ is a bounded random variable and applying Fact 2. □

We next prove Lemma 3.3.

Proof of Lemma 3.3. Using the results from Lemma 3.2, for any $x \in \mathbb{R}$, and any $i \in [d]$, we have

$$\phi_i(x) = \phi_i(0) + \phi_i'(0)x + h_i(x)x = \phi_i'(0)x + h_i(x)x,$$

where $h_i : \mathbb{R} \mapsto \mathbb{R}$ is such that $\lim_{x\to 0} h_i(x) = 0$. Recalling that $\phi'(0) \triangleq \min_{i\in[d]} \phi_i'(0) > 0$, it follows that there exists a $\xi > 0$ (depending only on $\mathbf{\Phi}$) such that, for each $x \in \mathbb{R}$ and all $i \in [d]$, we have $|h_i(x)| \leq \phi'(0)/2$, if $|x| \leq \xi$. Therefore, for any $0 \leq x \leq \xi$, we have

$$\phi_i(x) \geq \frac{\phi'(0)x}{2}. \tag{14}$$

On the other hand, for $x > \xi$, since ϕ_i is non-decreasing, we have from (14) that $\phi_i(x) \geq \phi_i(\xi) \geq \frac{\phi'(0)\xi}{2}$. Therefore, it follows that, for any $x \geq 0$

$$\phi_i(x) \geq \frac{\phi'(0)}{2}\min\{x, \xi\}. \tag{15}$$

Next, for any $a \in \mathbb{R}$, using oddity of ϕ_i, we get

$$a\phi_i(a) = |a|\phi_i(|a|). \tag{16}$$

Using (15) and (16), we then have, for any vector $\mathbf{x} \in \mathbb{R}^d$

$$\begin{aligned}\langle \mathbf{x}, \mathbf{\Phi}(\mathbf{x})\rangle = \sum_{i=1}^{d} x_i\phi_i(x_i) &\overset{(16)}{=} \sum_{i=1}^{d} |x_i|\phi(|x_i|) \geq \max_{i\in[d]} |x_i|\phi_i(|x_i|) \overset{(15)}{\geq} \frac{\phi'(0)}{2}\max_{i\in[d]} \min\{\xi|x_i|, |x_i|^2\} \\ &= \frac{\phi'(0)}{2}\min\{\xi\|\mathbf{x}\|_\infty, \|\mathbf{x}\|_\infty^2\} \geq \frac{\phi'(0)}{2}\min\{\xi\|\mathbf{x}\|/\sqrt{d}, \|\mathbf{x}\|^2/d\},\end{aligned} \tag{17}$$

where the last inequality follows from the fact that $\|\mathbf{x}\|_\infty \geq \|\mathbf{x}\|/\sqrt{d}$. □

In order to prove Lemma 3.4, we first state and prove some intermediate results.

Lemma C.1. *Let Assumptions 1-4 hold, with the step-size given by* $\alpha_t = \frac{a}{(t+t_0)^\delta}$*, for any* $\delta \in (0.5, 1)$ *and* $t_0 > 1$*. Then, for any* $t \in \mathbb{N}_0$*, we have*

$$\|\nabla f(\mathbf{x}^{(t)})\| \leq G_t \triangleq L\left(\|\mathbf{x}^{(0)} - \mathbf{x}^\star\| + aC\right)\frac{(t+t_0)^{1-\delta}}{1-\delta}.$$

Proof. Using L-smoothness of f and the update (3), we have

$$\begin{aligned}\|\nabla f(\mathbf{x}^{(t)})\| &\leq L\|\mathbf{x}^{(t)} - \mathbf{x}^\star\| = L\|\mathbf{x}^{(t-1)} - \alpha_{t-1}\mathbf{\Psi}^{(t-1)} - \mathbf{x}^\star\| \\ &\leq L\left(\|\mathbf{x}^{(t-1)} - \mathbf{x}^\star\| + \alpha_{t-1}\|\mathbf{\Psi}^{(t-1)}\|\right) \\ &\leq L\left(\|\mathbf{x}^{(t-1)} - \mathbf{x}^\star\| + \alpha_{t-1}C\right). \end{aligned} \tag{18}$$

Unrolling the recursion in (18), we get

$$\|\nabla f(\mathbf{x}^{(t)})\| \leq L\|\mathbf{x}^{(0)} - \mathbf{x}^\star\| + LC\sum_{k=1}^{t}\alpha_{k-1} \leq L\left(\|\mathbf{x}^{(0)} - \mathbf{x}^\star\| + aC\right)\frac{(t+t_0)^{1-\delta}}{1-\delta},$$

completing the proof. □

The next result characterize the behaviour of the nonlinearity, when it takes the form $\mathbf{\Psi}(\mathbf{x}) = [\mathcal{N}_1(x_1), \ldots, \mathcal{N}_1(x_d)]^\top$. It follows a similar idea to Lemma 5.5 from Jakovetić et al. (2023), with the main difference due to allowing for potentially different marginal PDFs of each noise component. Since the proof follows the same steps, we omit it for brevity.

Lemma C.2. *Let Assumptions 1-4 hold and the nonlinearity* $\mathbf{\Psi}$ *be component-wise, i.e., of the form* $\mathbf{\Psi}(\mathbf{x}) = [\mathcal{N}_1(x_1), \ldots, \mathcal{N}_1(x_d)]^\top$*. Then, there exists a positive constant* ξ *such that, for any* $t \in \mathbb{N}_0$*, there holds almost surely for each* $j = 1, \ldots, d$*, that* $|\phi_i^{(t)}| \geq |[\nabla f(\mathbf{x}^{(t)})]_i|\frac{\phi_i'(0)\xi}{2G_t}$*, where* G_t *is defined in Lemma C.1, while* $\phi_i'(0) = \frac{\partial}{\partial x_i}\mathbb{E}_{z_i}\mathcal{N}_1(x_i + z_i)\big|_{x_i=0}$.

The next result is a restatement of Lemma 6.2 in Jakovetić et al. (2023), in the case when the nonlinearity takes the form $\mathbf{\Psi}(\mathbf{x}) = \mathbf{x}\mathcal{N}_2(\|\mathbf{x}\|)$. We omit the proof for brevity.

Lemma C.3 (Lemma 6.2 from Jakovetić et al. (2023))**.** *Let Assumptions 1-4 hold and the nonlinearity be of the form* $\mathbf{\Psi}(\mathbf{x}) = \mathbf{x}\mathcal{N}_2(\|\mathbf{x}\|)$*. Then, the following holds*

$$\langle \mathbf{\Phi}(\mathbf{x}), \mathbf{x}\rangle \geq 2(1-\kappa)\|\mathbf{x}\|^2 \int_{\mathcal{J}(\mathbf{x})} \mathcal{N}_2(\|\mathbf{x}\| + \|\mathbf{z}\|)p(\mathbf{z})d\mathbf{z},$$

where $\mathcal{J}(\mathbf{x}) = \left\{\mathbf{z} \in \mathbb{R}^d : \frac{\langle \mathbf{z}, \mathbf{x}\rangle}{\|\mathbf{z}\|\|\mathbf{x}\|} \in [0, \kappa]\right\}$ *and* $\kappa \in (0,1)$ *is a constant.*

The next result characterizes the behaviour of the nonlinearity, when it takes the form $\mathbf{\Psi}(\mathbf{x}) = \mathbf{x}\mathcal{N}_2(\|\mathbf{x}\|)$. It builds on Lemma C.3 and we provide the full proof for completeness.

Lemma C.4. *Let Assumptions 1-4 hold and the nonlinearity be of the form* $\mathbf{\Psi}(\mathbf{x}) = \mathbf{x}\mathcal{N}_2(\|\mathbf{x}\|)$*. Then, there exists a constant* $\kappa \in (0,1)$ *such that for any* $t \in \mathbb{N}_0$*, there holds almost surely that*

$$\langle \nabla f(\mathbf{x}^{(t)}), \mathbf{\Phi}^{(t)}\rangle \geq \frac{2(1-\kappa)\lambda(\kappa)\mathcal{N}_2(1)\|\nabla f(\mathbf{x}^{(t)})\|^2}{B_0 + G_t},$$

where $\kappa \in (0,1)$ is the constant from Lemma C.3, $B_0 > 0$ is defined in Assumption 4, G_t is defined in Lemma C.1, with $\lambda(\kappa) > 0$ a constant such that $\int_{\left\{\mathbf{z}\in\mathbb{R}^d: \frac{\langle \mathbf{z},\mathbf{x}\rangle}{\|\mathbf{z}\|\|\mathbf{x}\|}\in[0,\kappa],\, \|\mathbf{z}\|\leq B_0\right\}} p(\mathbf{z})d\mathbf{z} > \lambda(\kappa)$, for any $\mathbf{x}$.

Proof. We start from Lemma C.3, which tells us that, for some $\kappa \in (0,1)$, almost surely

$$\langle \mathbf{\Phi}^{(t)}, \nabla f(\mathbf{x}^{(t)})\rangle \geq 2(1-\kappa)\|\nabla f(\mathbf{x}^{(t)})\|^2 \int_{\mathcal{J}(\nabla f(\mathbf{x}^{(t)}))} \mathcal{N}_2(\|\nabla f(\mathbf{x}^{(t)})\| + \|\mathbf{z}\|)p(\mathbf{z})d\mathbf{z}, \tag{19}$$

where $\mathcal{J}(\nabla f(\mathbf{x}^{(t)})) = \left\{\mathbf{z}\in\mathbb{R}^d : \frac{\langle \mathbf{z},\nabla f(\mathbf{x}^{(t)})\rangle}{\|\mathbf{z}\|\|\nabla f(\mathbf{x}^{(t)})\|} \in [0,\kappa]\right\}$. Note that, as $x\mathcal{N}_2(x)$ is a non-decreasing function (Assumption 1), $\mathcal{N}_2$ satisfies

$$\mathcal{N}_2(x) \geq \min\left\{\frac{\mathcal{N}_2(1)}{x}, \mathcal{N}_2(1)\right\}, \tag{20}$$

for any $x > 0$. In particular, for any $\mathbf{z}$ such that $\|\mathbf{z}\| \leq B_0$, we have $\mathcal{N}_2(\|\nabla f(\mathbf{x}^{(t)})\| + \|\mathbf{z}\|) \geq \min\left\{\frac{\mathcal{N}_2(1)}{\|\nabla f(\mathbf{x}^{(t)})\|+B_0}, \mathcal{N}_2(1)\right\}$. We then have

$$\begin{aligned}
\|\nabla f(\mathbf{x}^{(t)})\|^2 \int_{\mathcal{J}(\nabla f(\mathbf{x}^{(t)}))} \mathcal{N}_2(\|\nabla f(\mathbf{x}^{(t)})\| + \|\mathbf{z}\|)p(\mathbf{z})d\mathbf{z} &\geq \|\nabla f(\mathbf{x}^{(t)})\|^2 \int_{\mathcal{J}_1} \mathcal{N}_2(\|\nabla f(\mathbf{x}^{(t)})\| + \|\mathbf{z}\|)p(\mathbf{z})d\mathbf{z} \\
&\overset{(a)}{\geq} \|\nabla f(\mathbf{x}^{(t)})\|^2 \int_{\mathcal{J}_\kappa} \min\left\{\frac{\mathcal{N}_2(1)}{\|\nabla f(\mathbf{x}^{(t)})\| + B_0}, \mathcal{N}_2(1)\right\} p(\mathbf{z})d\mathbf{z} \\
&\overset{(b)}{\geq} \frac{\|\nabla f(\mathbf{x}^{(t)})\|^2 \mathcal{N}_2(1)}{G_t + B_0} \int_{\mathcal{J}_\kappa} p(\mathbf{z})d\mathbf{z} \\
&\overset{(c)}{\geq} \frac{\|\nabla f(\mathbf{x}^{(t)})\|^2 \lambda(\kappa)\mathcal{N}_2(1)}{G_t + B_0},
\end{aligned} \tag{21}$$

where $\mathcal{J}_\kappa = \left\{\mathbf{z}\in\mathbb{R}^d : \frac{\langle \mathbf{z},\nabla f(\mathbf{x}^{(t)})\rangle}{\|\mathbf{z}\|\|\nabla f(\mathbf{x}^{(t)})\|} \in [0,\kappa],\ \|\mathbf{z}\| \leq B_0\right\}$, (a) follows from (20), (b) follows from Lemma C.1 and the fact that $G_t > 1$, for every $t \geq 0$, while (c) follows from Assumption 4. Combining (19) and (21), we have that, almost surely

$$\langle \mathbf{\Phi}^{(t)}, \nabla f(\mathbf{x}^{(t)})\rangle \geq \frac{2(1-\kappa)\lambda(\kappa)\mathcal{N}_2(1)\|\nabla f(\mathbf{x}^{(t)})\|^2}{G_t + B_0},$$

which is what we wanted to show. □

We are now ready to prove Lemma 3.4.

Proof of Lemma 3.4. First, consider the case when the nonlinearity is of the form $\mathbf{\Psi}(\mathbf{x}) = [\mathcal{N}_1(x_1), \ldots, \mathcal{N}_1(x_d)]^\top$. We then have

$$\begin{aligned}
\langle \mathbf{\Phi}^{(t)}, \nabla f(\mathbf{x}^{(t)})\rangle = \sum_{i=1}^d \phi_i^{(t)}[\nabla f(\mathbf{x}^{(t)})]_i &\overset{(a)}{=} \sum_{i=1}^d |\phi_i^{(t)}||[\nabla f(\mathbf{x}^{(t)})]_i| \\
&\overset{(b)}{\geq} \sum_{i=1}^d |[\nabla f(\mathbf{x}^{(t)})]_i|^2 \frac{\phi_i'(0)\xi}{2G_t} \overset{(c)}{\geq} \frac{\phi'(0)\xi}{2G_t}\|\nabla f(\mathbf{x}^{(t)})\|^2 = \gamma(t+t_0)^{\delta-1}\|\nabla f(\mathbf{x}^{(t)})\|^2,
\end{aligned}$$

where $\gamma = \frac{(1-\delta)\phi'(0)\xi}{2L\left(\|\mathbf{x}^{(0)}-\mathbf{x}^\star\|+aC\right)}$, (a) follows from the oddity of $\mathcal{N}_1$, (b) follows from Lemma C.2, (c) follows from $\phi'(0) = \min_{i=1,\ldots,d}\phi_i'(0)$. On the other hand, if the nonlinearity is of the form $\mathbf{\Psi}(\mathbf{x}) = \mathbf{x}\mathcal{N}_2(\|\mathbf{x}\|)$, we get

$$\langle \mathbf{\Phi}^{(t)}, \nabla f(\mathbf{x}^{(t)})\rangle \geq \frac{2(1-\kappa)\lambda(\kappa)\mathcal{N}_2(1)\|\nabla f(\mathbf{x}^{(t)})\|^2}{G_t + B_0} \geq \gamma(t+t_0)^{\delta-1}\|\nabla f(\mathbf{x}^{(t)})\|^2,$$

where $\gamma = \frac{2(1-\delta)(1-\kappa)\lambda(\kappa)\mathcal{N}_2(1)}{L\left(\|\mathbf{x}^{(0)}-\mathbf{x}^\star\|+aC\right)+B_0}$, the first inequality follows from Lemma C.4, while the second follows from the definition of G_t and the fact that $G_t + B_0 \leq \left(L\left(\|\mathbf{x}^{(0)} - \mathbf{x}^\star\| + aC\right) + B_0\right)\frac{(t+t_0)^{1-\delta}}{1-\delta}$. This completes the proof. □

D Rate ζ

Recalling Assumption 1 and the definition of C, it readily follows that $\gamma = \frac{(1-\delta)\phi'(0)\xi}{2L\left(\|\mathbf{x}^{(0)}-\mathbf{x}^\star\|+a\sqrt{d}C_1\right)}$ for nonlinearities of the form $\mathbf{\Psi}(\mathbf{x}) = [\mathcal{N}_1(x_1), \ldots, \mathcal{N}_1(x_d)]^\top$ (i.e., component-wise), while $\gamma = \frac{2(1-\delta)(1-\kappa)\lambda(\kappa)\mathcal{N}_2(1)}{L\left(\|\mathbf{x}^{(0)}-\mathbf{x}^\star\|+aC_2\right)+B_0}$, for nonlinearities of the form $\mathbf{\Psi}(\mathbf{x}) = \mathbf{x}\mathcal{N}_2(\|\mathbf{x}\|)$ (i.e., joint). Combined with Theorem 3, it follows that the rate ζ is given by

$$\zeta_{joint} = \min\left\{2\delta - 1, \frac{2a\mu(1-\kappa)\lambda(\kappa)(1-\delta)\mathcal{N}_2(1)}{L\left(\|\mathbf{x}^{(0)} - \mathbf{x}^\star\| + aC_2\right) + B_0}\right\},$$

$$\zeta_{comp} = \min\left\{2\delta - 1, \frac{a\mu\phi'(0)\xi(1-\delta)}{4L\left(\|\mathbf{x}^{(0)} - \mathbf{x}^\star\| + aC_1\sqrt{d}\right)}\right\}.$$

We note that ζ depends on the following problem-specific parameters:

- *Initialization* - starting farther from the true minima $\mathbf{x}^\star$, results in smaller ζ. The effect of initialization can be eliminated by choosing sufficiently large a.
- *Condition number* - larger values of $\frac{L}{\mu}$ (i.e., a more difficult problem) result in smaller ζ.
- *Nonlinearity* - the dependence of ζ on the nonlinearity comes in the form of two terms: the uniform bound on the nonlinearity C_1 (C_2), and the value $\phi'(0)$ ($\mathcal{N}_2(1)$).
- *Problem dimension* - for component-wise nonlinearities directly in the form of $\sqrt{d}$; for joint ones indirectly, in the form of $\mathcal{N}_2(1)$. As we show in Section 4, the dependence on dimension for joint nonlinearities, while not explicit, can sometimes be worse than that of component-wise ones.
- *Noise* - in the form of $\phi'(0)$ and ξ for component-wise and B_0, $\lambda(\kappa)$ for joint ones.
- *Step-size* - both terms in the definition of ζ depend on the step-size parameter $\delta \in (0, 1)$.

E Miscellaneous

In this section we prove Lemma 4.1.

Proof of Lemma 4.1. Consider clipping and a generic component-wise nonlinearity. From Appendix D we know that the rate ζ is then given by

$$\zeta_{clip} = \min\left\{2\delta - 1, \frac{2a\mu(1-\kappa)\lambda(\kappa)(1-\delta)\mathcal{N}_2(1)}{L\left(\|\mathbf{x}^{(0)} - \mathbf{x}^\star\| + aC_2\right) + B_0}\right\},$$

$$\zeta_{comp} = \min\left\{2\delta - 1, \frac{a\mu\phi'(0)\xi(1-\delta)}{4L\left(\|\mathbf{x}^{(0)} - \mathbf{x}^\star\| + aC_1\sqrt{d}\right)}\right\},$$

where C_1, $C_2 > 0$ are the bounds on the nonlinearities, $\xi > 0$ is a constant depending on ϕ, $\kappa \in (0,1)$ and $\lambda(\kappa) > 0$ being a constant that satisfies

$$\int_{\left\{\mathbf{z}\in\mathbb{R}^d:\, \frac{\langle \mathbf{z},\mathbf{x}\rangle}{\|\mathbf{z}\|\|\mathbf{x}\|}\in[0,\kappa],\, \|\mathbf{z}\|\leq B_0\right\}} p(\mathbf{z})d\mathbf{z} > \lambda(\kappa).$$

Our goal is to show that clipping is not the best choice of nonlinearity. We will do so by showing that $\zeta_{clip} \leq \zeta_{comp}$. To guarantee this is the case, it suffices that

$$\frac{2a\mu(1-\kappa)\lambda(\kappa)(1-\delta)\mathcal{N}_2(1)}{L\left(\|\mathbf{x}^{(0)} - \mathbf{x}^\star\| + aC_2\right) + B_0} \leq \frac{a\mu\phi'(0)\xi(1-\delta)}{4L\left(\|\mathbf{x}^{(0)} - \mathbf{x}^\star\| + aC_1\sqrt{d}\right)},$$

or equivalently,

$$\frac{8(1-\kappa)\lambda(\kappa)\mathcal{N}_2(1)}{\|\mathbf{x}^{(0)} - \mathbf{x}^\star\| + aM + {}^{B_0}/_{L}} \leq \frac{\phi'(0)\xi}{\|\mathbf{x}^{(0)} - \mathbf{x}^\star\| + aC_1\sqrt{d}},$$

where we used the fact that $C_2 = M$, where $M > 0$, is the clipping threshold. Rearranging, we get

$$\phi'(0)\xi \geq 8(1-\kappa)\lambda(\kappa)\mathcal{N}_2(1)\frac{\|\mathbf{x}^{(0)} - \mathbf{x}^\star\| + aC_1\sqrt{d}}{\|\mathbf{x}^{(0)} - \mathbf{x}^\star\| + aM + {}^{B_0}/_{L}}.$$

Choosing $a > 0$ sufficiently large, it suffices to verify that the following holds

$$\frac{\phi'(0)\xi}{C_1} \geq (1-\kappa)\lambda(\kappa)\mathcal{N}_2(1)\frac{8\sqrt{d}}{M}. \tag{22}$$

Next, recall that the noise PDF from Example 1 is given by

$$p(x) = \frac{\alpha - 1}{2(1+|x|)^\alpha},$$

for some $\alpha > 2$. From the independence of noise components, we know that the joint PDF is given by

$$p(\mathbf{z}) = p(z_1) \times p(z_2) \times \ldots \times p(z_d),$$

which clearly satisfies Assumption 4 for any $B_0 > 0$. Next, for any $\mathbf{x} = \begin{bmatrix} x_1 & \ldots & x_d \end{bmatrix}^\top$ and $\kappa \in (0,1)$, define the set $\mathcal{J}(\kappa)$ as

$$\mathcal{J}(\kappa) = \left\{\mathbf{z} \in \mathbb{R}^d : \frac{\langle \mathbf{z}, \mathbf{x}\rangle}{\|\mathbf{z}\|\|\mathbf{x}\|} \in [0,\kappa],\, \|\mathbf{z}\| \leq B_0\right\}.$$

Consider the following sets

$$\mathcal{J}_i(\kappa) = \left\{ z \in \mathbb{R} : \ \frac{z \cdot x_i}{\|\mathbf{x}\|} \in [0, \kappa/d], \ |z| \leq \frac{B_0}{\sqrt{d}} \right\}, \text{ for all } i = 1, \ldots, d,$$

and define the set $\mathbf{B}_1 = \left\{ \mathbf{z} \in \mathbb{R}^d : \|z\| = 1 \right\}$. Then, clearly, for the set

$$\widetilde{\mathcal{J}}(\kappa) = \left\{ \mathbf{z} \in \mathbf{B}_1 : \ z_i \in \mathcal{J}_i(\kappa), \text{ for all } i = 1, \ldots, d \right\},$$

we have $\widetilde{\mathcal{J}}(\kappa) \subseteq \mathcal{J}(\kappa)$, hence

$$\int_{\mathcal{J}(\kappa)} p(\mathbf{z}) d\mathbf{z} > \int_{\widetilde{\mathcal{J}}(\kappa)} p(\mathbf{z}) d\mathbf{z} \triangleq \lambda(\kappa).$$

To prove our claim, it suffices to show that the relation (22) holds when $\lambda(\kappa)$ is replaced by some upper bound. To that end, by definition, we have $\widetilde{\mathcal{J}}(\kappa) \subseteq \left\{ \mathbf{z} \in \mathbb{R}^d : \ z_i \in \mathcal{J}_i(\kappa), \ i = 1, \ldots, d \right\}$, which implies

$$\lambda(\kappa) \leq \prod_{i=1}^{d} \int_{\mathcal{J}_i(\kappa)} p_i(z) dz \overset{(a)}{=} \left(\int_{\mathcal{J}_1(\kappa)} p(z_1) dz \right)^d \overset{(b)}{\leq} \left(\int_{|z| \leq \frac{B_0}{\sqrt{d}}} p(z) dz \right)^d \triangleq \widetilde{\lambda},$$

where (a) follows from the fact that all components of $\mathbf{z}$ are i.i.d., while (b) follows from the fact that $\mathcal{J}_1(\kappa) \subseteq \left\{ z \in \mathbb{R} : \ |z| \leq \frac{B_0}{\sqrt{d}} \right\}$. We can compute $\widetilde{\lambda}$ explicitly, by solving the integral, to get

$$\widetilde{\lambda} = \left(2 \int_0^{B_0/\sqrt{d}} \frac{\alpha - 1}{2(1+z)^\alpha} dz \right)^d = \left(1 - \frac{1}{(1 + B_0/\sqrt{d})^{\alpha - 1}} \right)^d.$$

Plugging $\widetilde{\lambda}$ in (22), we get

$$\frac{\phi'(0)\xi}{C_1} \geq 8(1-\kappa) \left(1 - \frac{1}{(1 + B_0/\sqrt{d})^{\alpha - 1}} \right)^d \mathcal{N}_2(1) \frac{\sqrt{d}}{M}.$$

Notice that $\mathcal{N}_2(1) = \min\{1, M\}$, which implies $\mathcal{N}_2(1)/M \leq 1$. Moreover, since the right-hand side (RHS) of the above equation is maximized for $\kappa \to 0$, we get the following relation

$$\frac{\phi'(0)\xi}{C_1} \geq 8\sqrt{d} \left(1 - \frac{1}{(1 + B_0/\sqrt{d})^{\alpha - 1}} \right)^d,$$

completing the proof. □

F Further derivations

In this Appendix we show how our bounds from Theorems 1 and 2 lead to bounds on the best iterate and weighted average of iterates, respectively. We begin by showing the former. In particular, from Theorem 1 we have

$$\sum_{k=0}^{t-1} \widetilde{\alpha}_k \min\left\{ \xi \|\nabla f(\mathbf{x}^{(k)})\| / \sqrt{d}, \ \|\nabla f(\mathbf{x}^{(k)})\|^2 / d \right\} \leq \frac{R(a, \beta, \delta)}{(t+2)^{1-\delta} - 2^{1-\delta}}. \tag{23}$$

Define $U \triangleq \{k \in \{0,\dots,t-1\} : \|\nabla f(\mathbf{x}^{(k)})\| \leq \xi\sqrt{d}\}$, with $U^c \triangleq \{0,1,\dots,t-1\} \setminus U$. From (23), we then have

$$\sum_{k\in U} \widetilde{\alpha}_k \|\nabla f(\mathbf{x}^{(k)})\|^2 \leq \frac{dR(a,\beta,\delta)}{(t+2)^{1-\delta}-2^{1-\delta}},$$
$$\sum_{k\in U^c} \widetilde{\alpha}_k \|\nabla f(\mathbf{x}^{(k)})\| \leq \frac{R(a,\beta,\delta)\sqrt{d}/\xi}{(t+2)^{1-\delta}-2^{1-\delta}}.$$

It then readily follows that

$$\min_{0\leq k\leq t-1} \|\nabla f(\mathbf{x}^{(k)})\| \leq \sum_{k\in U} \widetilde{\alpha}_k \|\nabla f(\mathbf{x}^{(k)})\| + \sum_{k\in U^c} \widetilde{\alpha}_k \|\nabla f(\mathbf{x}^{(k)})\| \leq \sum_{k=0}^{t-1} \widetilde{\alpha}_k z_k + \frac{R(a,\beta,\delta)\sqrt{d}/\xi}{(t+2)^{1-\delta}-2^{1-\delta}},$$

where $z_k = \|\nabla f(\mathbf{x}^{(k)})\|$, for $k \in U$, otherwise $z_k = 0$. Using Jensen's inequality, we get

$$\begin{aligned}\min_{0\leq k\leq t-1} \|\nabla f(\mathbf{x}^{(k)})\| &\leq \sqrt{\sum_{k=0}^{t-1} \widetilde{\alpha}_k z_k^2} + \frac{R(a,\beta,\delta)\sqrt{d}/\xi}{(t+2)^{1-\delta}-2^{1-\delta}} = \sqrt{\sum_{k\in U} \widetilde{\alpha}_k \|\nabla f(\mathbf{x}^{(k)})\|^2} + \frac{R(a,\beta,\delta)\sqrt{d}/\xi}{(t+2)^{1-\delta}-2^{1-\delta}} \\ &\leq \sqrt{\frac{dR(a,\beta,\delta)}{(t+2)^{1-\delta}-2^{1-\delta}}} + \frac{R(a,\beta,\delta)\sqrt{d}/\xi}{(t+2)^{1-\delta}-2^{1-\delta}},\end{aligned}$$

giving the bound $\min_{0\leq k\leq t-1} \|\nabla f(\mathbf{x}^{(k)})\| = \mathcal{O}\left(t^{(\delta-1)/2}\right)$. Next, from Theorem 2, we have

$$H_{\xi\sqrt{d}/\mu}(\|\widehat{\mathbf{x}}^{(t)} - \mathbf{x}^\star\|) \leq \frac{\widetilde{R}(a,\beta,\delta)}{(t+2)^{1-\delta}-2^{1-\delta}}, \tag{24}$$

By the definition of Huber loss, it then follows by (24) that, if $\|\widehat{\mathbf{x}}^{(t)} - \mathbf{x}^\star\| \leq \xi\sqrt{d}/\mu$, we have

$$\|\widehat{\mathbf{x}}^{(t)} - \mathbf{x}^\star\|^2 \leq \frac{2\widetilde{R}(a,\beta,\delta)}{(t+2)^{1-\delta}-2^{1-\delta}}. \tag{25}$$

Otherwise, if $\|\widehat{\mathbf{x}}^{(t)} - \mathbf{x}^\star\| > \xi\sqrt{d}/\mu$, by (24), we have

$$\frac{\xi\sqrt{d}\|\widehat{\mathbf{x}}^{(t)} - \mathbf{x}^\star\|}{2\mu} < \xi\sqrt{d}\|\widehat{\mathbf{x}}^{(t)} - \mathbf{x}^\star\|/\mu - \frac{\xi^2 d}{2\mu^2} \leq \frac{\widetilde{R}(a,\beta,\delta)}{(t+2)^{1-\delta}-2^{1-\delta}},$$

implying that

$$\|\widehat{\mathbf{x}}^{(t)} - \mathbf{x}^\star\|^2 \leq \frac{4\mu^2\widetilde{R}(a,\beta,\delta)^2/(\xi^2 d)}{\left((t+2)^{1-\delta}-2^{1-\delta}\right)^2}. \tag{26}$$

Combining (25) and (26), it then follows that

$$\|\widehat{\mathbf{x}}^{(t)} - \mathbf{x}^\star\|^2 \leq \min\left\{\frac{2\widetilde{R}(a,\beta,\delta)}{(t+2)^{1-\delta}-2^{1-\delta}}, \frac{4\mu^2\widetilde{R}(a,\beta,\delta)^2/(\xi^2 d)}{\left((t+2)^{1-\delta}-2^{1-\delta}\right)^2}\right\},$$

which, for t sufficiently large, implies $\|\widehat{\mathbf{x}}^{(t)} - \mathbf{x}^\star\|^2 = \mathcal{O}\left(t^{2(\delta-1)}\right)$.