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

Aleksandar Armacki
CMU, PA, USA

Shuhua Yu
CMU, PA, USA

Pranay Sharma
CMU, PA, USA

Gauri Joshi
CMU, PA, USA

Dragana Bajović
UNS, Serbia

Dušan Jakovetić
UNS, Serbia

Soumya Kar
CMU, PA, USA

Abstract

We study high-probability convergence in online learning, in the presence of heavy-tailed noise. To combat the heavy tails, a general framework of nonlinear SGD methods is considered, subsuming several popular nonlinearities like sign, quantization, component-wise and joint clipping. In our work the nonlinearity is treated in a black-box manner, allowing us to establish unified guarantees for a broad range of nonlinear methods. For symmetric noise and non-convex costs we establish convergence of gradient norm-squared, at a rate $\tilde{\mathcal{O}}(t^{-1/4})$, while for the last iterate of strongly convex costs we establish convergence to the population optima, at a rate $\mathcal{O}(t^{-\zeta})$, where $\zeta \in (0, 1)$ depends on noise and problem parameters. Further, if the noise is a (biased) mixture of symmetric and non-symmetric components, we show convergence to a neighbourhood of stationarity, whose size depends on the mixture coefficient, nonlinearity and noise. Compared to state-of-the-art, who only consider clipping and require unbiased noise with bounded p -th moments, $p \in (1, 2]$, we provide guarantees for a broad class of nonlinearities, without any assumptions on noise moments. While the rate exponents in state-of-the-art depend on noise moments and vanish as $p \rightarrow 1$, our exponents are constant and strictly better whenever $p < 6/5$ for non-convex and $p < 8/7$ for strongly convex costs. Experiments validate our theory, showing that clipping is

not always the optimal nonlinearity, further underlining the value of a general framework.

1 INTRODUCTION

Stochastic optimization is a well-studied problem, e.g., Robbins and Monro (1951); Nemirovski et al. (2009), where the goal is to minimize an expected cost, without knowing the underlying probability distribution. Formally, the problem is cast as

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

where $\mathbf{x} \in \mathbb{R}^d$ represents model parameters, $\ell : \mathbb{R}^d \times \mathcal{V} \mapsto \mathbb{R}$ is a loss function, $v \in \mathcal{V}$ is a random sample distributed according to the unknown distribution Υ , while $f : \mathbb{R}^d \mapsto \mathbb{R}$ is commonly known as the *population cost*. Many modern machine learning applications, such as classification and regression, are modeled using (1).

Perhaps the most popular method to solve (1) is stochastic gradient descent (SGD) Robbins and Monro (1951), whose popularity stems from low computation cost and incredible empirical success Bottou (2010); Hardt et al. (2016). Convergence guarantees of SGD have been studied extensively Moulines and Bach (2011); Rakhlin et al. (2012); Bottou et al. (2018). Classical convergence results are mostly concerned with *mean-squared error* (MSE) convergence, characterizing the average performance across many runs of the algorithm. However, due to significant computational cost of a single run of an algorithm in many modern machine learning applications, it is often infeasible to perform multiple runs Harvey et al. (2019); Davis et al. (2021). As such, many applications require more fine-grained results, such as *high-probability convergence*, which characterize the behaviour of an algorithm with respect to a single run.

Another striking feature of existing works is the assumption that the gradient noise has *light-tails* or

Table 1: High-probability guarantees of SGD methods under heavy-tailed noise. Online indicates whether a method uses a time-varying step-size and is applicable in the online setting (indicated by lower-case t), or if it uses a fixed step-size and requires a preset time horizon which is optimized to achieve the best rate and works only in the offline setting (indicated by upper-case T). The value $\beta \in (0, 1)$ represents the failure probability, while $\tilde{\mathcal{O}}(\cdot)$ hides factors poly-logarithmic in time t . All the works achieve a poly-logarithmic dependence on the failure probability β (i.e., contain a multiplicative factor of $\log(1/\beta)$ in the bound), which is hidden under the big O notation, for ease of presentation.

COST	WORK	NONLINEARITY	NOISE	ONLINE	RATE
NON-CONVEX	NGUYEN ET AL. (2023A)	CLIPPING ONLY	UNBIASED, BOUNDED MOMENT OF ORDER $p \in (1, 2]$	✓	$\tilde{\mathcal{O}}(t^{2(1-p)/(3p-2)})$
	SADIEV ET AL. (2023)			✗	$\mathcal{O}(T^{(1-p)/p})$
	THIS PAPER	COMPONENT-WISE AND JOINT	SYMMETRIC PDF, POSITIVE AROUND ZERO	✓	$\tilde{\mathcal{O}}(t^{-1/4})^\dagger$
STRONGLY CONVEX	SADIEV ET AL. (2023)	CLIPPING ONLY	UNBIASED, BOUNDED MOMENT OF ORDER $p \in (1, 2]$	✗	$\mathcal{O}(T^{2(1-p)/p})$
	THIS PAPER - WEIGHTED AVERAGE OF ITERATES	COMPONENT-WISE AND JOINT	SYMMETRIC PDF, POSITIVE AROUND ZERO	✓	$\tilde{\mathcal{O}}(t^{-1/4})^\dagger$
	THIS PAPER - LAST ITERATE			✓	$\mathcal{O}(t^{-\zeta})^\S$

[†] We derive convergence guarantees for a wide range of step-sizes of the form $\alpha_t = a/(t+1)^\delta$, where $a > 0$, $\delta \in (2/3, 1)$, with the resulting convergence rate depending on δ . The best rate, shown in the table, is achieved for the choice $\delta = 3/4$.

[§] The rate $\zeta \in (0, 1)$ depends on the choice of nonlinearity, noise and problem related parameters, see Section 3 and Appendix S4. We provide examples of noise for which $\zeta > 2(p-1)/p$, see Examples 1-5 ahead.

uniformly bounded variance Rakhlin et al. (2012); Ghadimi and Lan (2012, 2013), which represents a major limitation in many modern applications, see Simsekli et al. (2019a,b). For example, Zhang et al. (2020) show that the gradient noise distribution during training of large attention models resembles a Levy α -stable distribution, with $\alpha < 2$, which has unbounded variance. To better model this phenomena, the authors propose the *bounded p -th moment* assumption, i.e.,

$$\mathbb{E}_{v \sim \Upsilon} \|\nabla \ell(\mathbf{x}, v) - \nabla f(\mathbf{x})\|^p \leq \sigma^p, \quad (\text{BM})$$

for every $\mathbf{x} \in \mathbb{R}^d$ and some $p \in (1, 2]$, $\sigma > 0$, subsuming the bounded variance case for $p = 2$. Under this assumption, Zhang et al. (2020) show that SGD fails to converge for any fixed step-size. The clipped variant of SGD solves this problem and achieves *optimal* MSE convergence rate for smooth non-convex losses. Along with addressing heavy-tailed noise, clipped SGD also addresses non-smoothness of the cost Zhang et al. (2019), ensures differential privacy Zhang et al. (2022) and robustness to malicious nodes in distributed learning Yu and Kar (2023). While popular, clipping is not the only nonlinearity employed in practice. Sign and quantized variants of SGD improve communication efficiency in distributed learning 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 Crawshaw et al. (2022), and is robust to faulty and malicious users Bernstein et al. (2018b). Normalized SGD is empirically observed to accelerate neural network training Hazan et al. (2015); You et al. (2019); Cutkosky and Mehta (2020) and facilitates privacy Das et al. (2021); Yang et al. (2022), while Zhang et al. (2020) empirically observe that component-wise clipping converges faster than

the joint one, showing better dependence on problem dimension. Although assumption (BM) helps bridge the gap between theory and practice, the downside is that the resulting convergence rates have exponents which explicitly depend on the noise moment and vanish as $p \rightarrow 1$. This seems to contradict the strong performance of nonlinear SGD methods observed in practice and fails to explain the empirical success of nonlinear SGD, e.g., during training of models such as neural networks, in the presence of heavy-tailed noise. A growing body of works recently provided strong evidence that the stochastic noise during training of neural networks is *symmetric*, by studying the empirical distribution of gradient noise during training. For example, Bernstein et al. (2018a,b) show that histograms of gradient noise during training of different Resnet architectures on CIFAR-10 and Imagenet data exhibit strong symmetry under various batch sizes, see their Figures 2 (in both works). Similarly, Chen et al. (2020) demonstrate strong symmetry of gradient distributions during training of convolutional neural networks (CNN) on CIFAR-10 and MNIST data, see their Figures 1-3. Barsbey et al. (2021) show that the histograms of weights of a CNN layer trained on MNIST data almost identically match samples simulated from a symmetric α -stable distribution, see their Figure 2. Finally, Battash et al. (2024) show that a heavy-tailed symmetric α -stable distribution is a much better fit for the stochastic gradient noise than a Gaussian, for a myriad of deep learning architectures and datasets, see their Tables 1-3. Relying on a generalization of the central limit theorem (CLT), Simsekli et al. (2019b); Peluchetti et al. (2020); Gurbuzbalaban et al. (2021); Barsbey et al. (2021) theoretically show that symmetric heavy-tailed noises are appropriate models in many

practical settings, e.g., when training neural networks with mini-batch SGD using a large batch size. In contrast, works using assumption (BM) are inherently oblivious to this widely observed phenomena. The goal of this paper is to study high-probability guarantees of nonlinear SGD methods in the presence of symmetric heavy-tailed noise and the benefits symmetry brings.

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 and include Nemirovski et al. (2009); Lan (2012); Hazan and Kale (2014); Harvey et al. (2019); Ghadimi and Lan (2013); Li and Orabona (2020). Subsequent works Gorbunov et al. (2020, 2021); Parletta et al. (2022) generalized these results to noise with bounded variance. Tsai et al. (2022) study clipped SGD, assuming the variance is bounded by iterate distance, while Li and Liu (2022); Eldowa and Paudice (2023); Madden et al. (2024) consider sub-Weibull noise. Recent works Liu et al. (2023a); Eldowa and Paudice (2023) remove restrictive assumptions, like bounded stochastic gradients and domain. Sadiev et al. (2023) show that even with bounded variance and smooth, strongly-convex functions, vanilla SGD cannot achieve an exponential tail decay, implying that the complexity of achieving a high-probability bound for SGD can be much worse than that of the corresponding MSE bound. As such, nonlinear SGD is used to handle tails heavier than sub-Gaussian. Recent works consider a class of heavy-tailed noises satisfying (BM), 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 losses. It is worth mentioning Puchkin et al. (2023), who show that clipped SGD achieves the optimal $\mathcal{O}(T^{-1})^1$ rate for smooth, strongly convex costs, under a class of heavy-tailed noises with possibly unbounded first moments. However, their noise assumption is difficult to verify, as it requires computing convolutions of order k , for all $k \in \mathbb{N}$. Additionally, they use a median-of-means gradient estimator, which requires evaluating multiple stochastic gradients per iteration and is not applicable in the online setting considered in this paper.

The works closest to ours are Nguyen et al. (2023a)

¹We use lower-case t to indicate an online method, using a time-varying step-size, whereas upper-case T indicates an offline method, which uses a fixed-step size and a predefined time horizon T . While an online method can clearly be used in the offline setting, the converse is not true.

for online non-convex and Sadiev et al. (2023) for offline strongly convex problems. We present a detailed comparison in Table 1. Both works study only the clipping operator and use assumption (BM). For non-convex costs, Nguyen et al. (2023a) achieve the optimal rate $\tilde{\mathcal{O}}(t^{2(1-p)/(3p-2)})$, while Sadiev et al. (2023) achieve the optimal rate $\mathcal{O}(T^{2(1-p)/p})$ for strongly convex costs. Compared to them, we consider a much broader class of nonlinearities in the presence of noise with symmetric density with no moment requirements, achieving the near-optimal rate $\tilde{\mathcal{O}}(t^{-1/4})$ for non-convex costs and extending it to the weighted average of iterates for strongly convex costs. Crucially, our rate exponent *is independent of noise and problem parameters*, which is not the case with Nguyen et al. (2023a); Sadiev et al. (2023). Our rates are strictly better whenever $p < 6/5$ for non-convex and $p < 8/7$ for strongly convex costs.² Additionally, we establish convergence of the *last iterate* for strongly convex costs, with rate $\mathcal{O}(t^{-\zeta})$, where $\zeta \in (0, 1)$ depends on 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 1-5) and demonstrate numerically that *clipping is not always the best nonlinearity* (see Section 4), further highlighting the importance and usefulness of our general framework. Finally, it is worth mentioning Jakovetić et al. (2023), who provide MSE, asymptotic normality and almost sure guarantees of the same nonlinear framework for strongly convex costs and noises with symmetric PDF, positive around zero and bounded first moments. Our work differs in that we study high-probability convergence, relax the moment conditions 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 gradient (see Lemma 3.2).

Contributions. Our contributions are as follows.

1. We study convergence in high probability of a unified framework of nonlinear SGD, in the presence of heavy-tailed noise and widely observed noise symmetry, making no assumptions on noise moments. The nonlinear map is treated in a black-box manner, subsuming many popular nonlinearities, like sign, normalization, clipping and quantization. *To the best of our knowledge, we provide the first high-probability results under heavy-tailed noise for methods such as sign, quantized and component-wise clipped SGD.*

²This does not contradict the optimality of the rates in Nguyen et al. (2023a); Sadiev et al. (2023), as their assumptions differ from ours. While Nguyen et al. (2023a); Sadiev et al. (2023) require bounded noise moment of order $p \in (1, 2]$, we study noise with symmetric density, without making any moment requirements. As such, we show that symmetry leads to improved results and allows for relaxed moment conditions and heavier tails (see Examples 1-4).

2. For non-convex costs, we show convergence of gradient norm-squared, at a near-optimal rate $\tilde{\mathcal{O}}(t^{-1/4})$. The exponent in our rate is constant, independent of noise and problem parameters, which is not the case with state-of-the-art Nguyen et al. (2023a). Our rate is strictly better than state-of-the-art whenever the noise has bounded moments of order $p < \frac{6}{5}$.
3. For strongly convex costs we show convergence of the weighted average of iterates, at the same rate $\tilde{\mathcal{O}}(t^{-1/4})$. Our rate dominates the state-of-the-art Sadiev et al. (2023) whenever the noise has bounded moments of order $p < \frac{8}{7}$, while being applicable in the online setting, which is not the case for Sadiev et al. (2023). For the last iterate we show convergence at a rate $\mathcal{O}(t^{-\zeta})$, where $\zeta \in (0, 1)$ depends on noise, nonlinearity and problem parameters, but remains bounded away from zero even for unbounded noise moments.
4. We extend our results beyond symmetric noise, by considering a mixture of symmetric and non-symmetric components. For non-convex costs we show convergence to a neighbourhood of stationarity, at a rate $\tilde{\mathcal{O}}(t^{-1/4})$, where the size of the neighbourhood depends on the mixture coefficient, nonlinearity and noise. While Nguyen et al. (2023a) achieve convergence under condition (BM), which does not require symmetry, they explicitly require *unbiased noise*, which is not the case for our mixture noise, allowing it to be *biased*.
5. Compared to state-of-the-art Nguyen et al. (2023a); Sadiev et al. (2023), who only consider clipping, require bounded noise moments of order $p \in (1, 2]$ and whose rates vanish as $p \rightarrow 1$, we consider a much broader class of nonlinearities, relax the moment condition and provide convergence rates with constant exponents. Finally, we provide 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 proposed framework. Section 3 presents the main results. Section 4 provides numerical results. Section 5 concludes the paper. Supplement contains additional experiments and proofs omitted from the main body. The remainder of this section introduces the notation.

Notation. The set of positive integers is denoted by \mathbb{N} . For $a \in \mathbb{N}$, the set of integers up to and including a is denoted by $[a] = \{1, \dots, a\}$. The sets of real numbers and d -dimensional vectors are denoted by \mathbb{R} and \mathbb{R}^d . Regular and bold symbols denote scalars and vectors, i.e., $x \in \mathbb{R}$ and $\mathbf{x} \in \mathbb{R}^d$. The Euclidean inner product and induced norm are denoted by $\langle \cdot, \cdot \rangle$ and $\|\cdot\|$.

2 PROPOSED FRAMEWORK

To solve (1) in the online setting, under the presence of heavy-tailed noise, we use the *nonlinear SGD*

Algorithm 1 Online Nonlinear SGD

Require: Choice of nonlinearity $\Psi : \mathbb{R}^d \mapsto \mathbb{R}^d$, model initialization $\mathbf{x}^{(1)} \in \mathbb{R}^d$, step-size schedule $\{\alpha_t\}_{t \in \mathbb{N}}$;

- 1: **for** $t = 1, 2, \dots$ **do**:
- 2: Query the oracle and receive $\nabla \ell(\mathbf{x}^{(t)}; v^{(t)})$;
- 3: Update $\mathbf{x}^{(t+1)} \leftarrow \mathbf{x}^{(t)} - \alpha_t \Psi(\nabla \ell(\mathbf{x}^{(t)}; v^{(t)}))$;

framework. The algorithm starts by choosing a deterministic initial model $\mathbf{x}^{(1)} \in \mathbb{R}^d$,³ a step-size schedule $\{\alpha_t\}_{t \in \mathbb{N}}$ and a nonlinear map $\Psi : \mathbb{R}^d \mapsto \mathbb{R}^d$. In iteration $t = 1, 2, \dots$, the method performs as follows: a first-order oracle is queried, which returns the gradient of the loss ℓ evaluated at the current model $\mathbf{x}^{(t)}$ and a random sample $v^{(t)}$.⁴ Then, the model is updated as

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

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 Ψ .

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

1. $\mathcal{N}_1, \mathcal{N}_2$ are continuous almost everywhere,⁵ \mathcal{N}_1 is piece-wise differentiable and the map $a \mapsto a \mathcal{N}_2(a)$ is non-decreasing.
2. \mathcal{N}_1 is monotonically non-decreasing and odd, 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 $\|\Psi(\mathbf{x})\| \leq C$, where $C = C_1 \sqrt{d}$ or $C = C_2$, depending on the form of nonlinearity. We will use the general bound $\|\Psi(\mathbf{x})\| \leq C$ for ease of presentation, and specialize where appropriate. Assumption 1 is satisfied by a wide class of nonlinearities, including:

1. *Sign*: $[\Psi(\mathbf{x})]_i = \text{sign}(x_i)$, $i \in [d]$.
2. *Component-wise clipping*: $[\Psi(\mathbf{x})]_i = x_i$, for $|x_i| \leq m$, and $[\Psi(\mathbf{x})]_i = m \cdot \text{sign}(x_i)$, for $|x_i| > m$, $i \in [d]$, for user-specified $m > 0$.
3. *Component-wise quantization*: for each $i \in [d]$, let $[\Psi(\mathbf{x})]_i = r_j$, for $x_i \in (q_j, q_{j+1}]$, with $j = 0, \dots, J - 1$

³While the initial model is deterministically chosen, it can be any vector in \mathbb{R}^d . This distinction is required for the theoretical analysis in the next section.

⁴Equivalently, the oracle directly sends the random sample $v^{(t)}$, which we use to compute the gradient of ℓ .

⁵With respect to the Lebesgue measure.

and $-\infty = q_0 < q_1 < \dots < q_J = +\infty$, where r_j, q_j are chosen such that each component of Ψ is odd, and we have $\max_{j \in \{0, \dots, J-1\}} |r_j| < R$, for user-specified $R > 0$.
4. *Normalization*: $\Psi(\mathbf{x}) = \frac{\mathbf{x}}{\|\mathbf{x}\|}$ and $\Psi(\mathbf{x}) = \mathbf{0}$, if $\mathbf{x} = \mathbf{0}$.
5. *Clipping*: $\Psi(\mathbf{x}) = \min\{1, M/\|\mathbf{x}\|\}\mathbf{x}$, for user-specified $M > 0$.

3 THEORETICAL GUARANTEES

In this section we present the main results of the paper. Subsection 3.1 presents the preliminaries, Subsection 3.2 presents the results for symmetric noises, while Subsection 3.3 presents the results for non-symmetric noises. The proofs can be found in the Supplement.

3.1 Preliminaries

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

Assumption 2. The cost 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 $\nabla f(\mathbf{x}^*) = \mathbf{0}$, and $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|$, for some $L > 0$ and every $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$.

Remark 1. Boundedness from below and Lipschitz continuous gradients are standard for non-convex losses, e.g., Ghadimi and Lan (2013). Since the goal in non-convex optimization is to reach a stationary point, it is natural to assume at least one such point exists, see Liu et al. (2023a); Madden et al. (2024).

Remark 2. It can be shown that Lipschitz continuous gradients imply the L -smoothness inequality, i.e., $f(\mathbf{y}) \leq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{L}{2}\|\mathbf{x} - \mathbf{y}\|^2$, for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, see Nesterov (2018); Wright and Recht (2022).

In addition to Assumption 2, we will sometimes use the following assumption.

Assumption 3. The cost f is strongly convex, i.e., $f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{\mu}{2}\|\mathbf{x} - \mathbf{y}\|^2$, for some $\mu > 0$ and every $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$.

Denote the infimum of f by $f^* \triangleq \inf_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x})$. Denote the set of stationary points of f by $\mathcal{X} \triangleq \{\mathbf{x}^* \in \mathbb{R}^d : \nabla f(\mathbf{x}^*) = \mathbf{0}\}$. By Assumption 2, it follows that $\mathcal{X} \neq \emptyset$. If in addition Assumption 3 holds, we have $\mathcal{X} = \{\mathbf{x}^*\}$ and $f^* = f(\mathbf{x}^*)$, for some $\mathbf{x}^* \in \mathbb{R}^d$. Denote the distance of the initial model from the set of stationary points by $D_{\mathcal{X}} \triangleq \inf_{\mathbf{x} \in \mathcal{X}} \|\mathbf{x}^{(1)} - \mathbf{x}\|^2$. Next, rewrite the update (2) as

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

where $\mathbf{z}^{(t)} \triangleq \nabla \ell(\mathbf{x}^{(t)}; v^{(t)}) - \nabla f(\mathbf{x}^{(t)})$ is the stochastic noise at iteration t . To simplify the notation, we use

the shorthand $\Psi^{(t)} \triangleq \Psi(\nabla f(\mathbf{x}^{(t)}) + \mathbf{z}^{(t)})$. We make the following assumption on the noise vectors $\{\mathbf{z}^{(t)}\}_{t \in \mathbb{N}}$.

Assumption 4. The noise vectors $\{\mathbf{z}^{(t)}\}_{t \in \mathbb{N}}$ are independent, identically distributed, with symmetric probability density function (PDF) $P : \mathbb{R}^d \mapsto \mathbb{R}$, positive around zero, i.e., $P(-\mathbf{z}) = P(\mathbf{z})$, for all $\mathbf{z} \in \mathbb{R}^d$ and $P(\mathbf{z}) > 0$, for all $\|\mathbf{z}\| \leq B_0$ and some $B_0 > 0$.

Remark 3. Assumption 4 imposes no moment conditions, at the expense of requiring a symmetric PDF, positive in a neighborhood of zero. Symmetry and positivity around zero are mild assumptions, satisfied by many noise distributions, such as Gaussian, the ones in Examples 1-4 below, and a broad class of heavy-tailed symmetric α -stable distributions, e.g., Bercovici et al. (1999); Nair et al. (2022).

Remark 4. As discussed in the introduction, heavy-tailed symmetric noise has been widely observed during training of deep learning models, across different architectures, datasets and batch sizes, e.g., Bernstein et al. (2018a,b); Chen et al. (2020); Barsbey et al. (2021); Battash et al. (2024). Building on the generalized CLT, Simsekli et al. (2019b); Peluchetti et al. (2020); Gurbuzbalaban et al. (2021); Barsbey et al. (2021) provide theoretical justification for this phenomena, e.g., when training neural nets with a large batch size.

Remark 5. The independent, identically distributed requirement in Assumption 4 can be significantly relaxed, to allow for noises which are not identically distributed, and in the case of joint nonlinearities, potentially depend on the current model. The reader is referred to the Supplement for a detailed discussion.

Remark 6. Positivity around zero of the PDF is a technical condition, ensuring that the “denoised nonlinearity” (see Section 3 ahead) is well-behaved. As such, the magnitude of the neighborhood B_0 does not directly affect the bounds established in Section 3.

Remark 7. While the noise assumption used in our work and the p -th bounded moment assumption (BM) are different, neither set of assumptions is uniformly weaker, with both having some advantages and disadvantages. For a detailed comparison between the two sets of assumptions, the reader is referred to the Supplement.

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

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

Example 2. The noise PDF $P(\mathbf{z}) = \rho(z_1) \times \dots \times \rho(z_d)$, 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 the noise has a finite first moment, but for any $p \in (1, 2]$, the p -th moments do not exist.

Example 3. The noise PDF $P(\mathbf{z}) = \rho(z_1) \times \dots \times \rho(z_d)$, where $\rho(z) = \frac{\gamma}{\pi\gamma^2 + \pi(x-x_0)^2}$, for some $x_0 \in \mathbb{R}$ and $\gamma > 0$, i.e., each component is distributed according to the Cauchy distribution. In this case, even the mean of the noise does not exist.

Example 4. 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 the noise does not have finite p -th moments, for any $p > 1$, while if ρ is the PDF of the Cauchy distribution, then the noise does even not have the first moment.

Remark 8. While noise in Example 1 satisfies moment condition (BM), noises in Examples 2 and 3 do not.

Next, define the function $\Phi : \mathbb{R}^d \mapsto \mathbb{R}^d$, given by $\Phi(\mathbf{x}) \triangleq \mathbb{E}_{\mathbf{z}}[\Psi(\mathbf{x} + \mathbf{z})] = \int \Psi(\mathbf{x} + \mathbf{z})P(\mathbf{z})d\mathbf{z}$,⁶ where the expectation is taken with respect to the gradient noise at a random sample. We use the shorthand $\Phi^{(t)} \triangleq \mathbb{E}_{\mathbf{z}^{(t)}}[\Psi(\nabla f(\mathbf{x}^{(t)}) + \mathbf{z}^{(t)}) \mid \mathcal{F}_t]$,⁷ where \mathcal{F}_t is the natural filtration, i.e., $\mathcal{F}_1 \triangleq \sigma(\{\emptyset, \Omega\})$ and $\mathcal{F}_t \triangleq \sigma(\{\mathbf{x}^{(2)}, \dots, \mathbf{x}^{(t)}\})$, for $t \geq 2$.⁸ The vector $\Phi^{(t)}$ can be seen as the “denoised” version of $\Psi^{(t)}$. Using $\Phi^{(t)}$, we can rewrite the update rule (3) as

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

where $\mathbf{e}^{(t)} \triangleq \Phi^{(t)} - \Psi^{(t)}$ represents the *effective noise* term. As we show next, the effective noise is light-tailed, even though the original noise may not be.

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

1. $\mathbb{E}[\mathbf{e}^{(t)} \mid \mathcal{F}_t] = 0$ and $\|\mathbf{e}^{(t)}\| \leq 2C$.
2. The effective noise is sub-Gaussian, i.e., for any $\mathbf{x} \in \mathbb{R}^d$, we have $\mathbb{E}[\exp(\langle \mathbf{x}, \mathbf{e}^{(t)} \rangle) \mid \mathcal{F}_t] \leq \exp(4C^2 \|\mathbf{x}\|^2)$.

3.2 Main Results

In this section we establish convergence in high probability of the proposed framework. Our results are facilitated by a novel result on the interplay of $\Phi(\mathbf{x})$ and the original vector \mathbf{x} , which is presented next.

Lemma 3.2. *Let Assumptions 1 and 4 hold. Then $\langle \Phi(\mathbf{x}), \mathbf{x} \rangle \geq \min\{\eta_1 \|\mathbf{x}\|, \eta_2 \|\mathbf{x}\|^2\}$, for any $\mathbf{x} \in \mathbb{R}^d$, where $\eta_1, \eta_2 > 0$, are noise, nonlinearity and problem dependent constants.*

Lemma 3.2 provides a novel characterization of the inner product between the “denoised” nonlinearity Φ

⁶If Ψ is a component-wise nonlinearity, then Φ is a vector with components $\phi_i(x_i) = \mathbb{E}_{z_i}[\Psi_1(x_i + z_i)]$, where \mathbb{E}_{z_i} is the marginal expectation with respect to the i -th noise component, $i \in [d]$ (see Lemma S3.1 ahead).

⁷Conditioning on \mathcal{F}_t ensures that the quantity $\nabla f(\mathbf{x}^{(t)})$ is deterministic and $\Phi^{(t)}$ is well defined.

⁸Recall that in our setup, the initialization $\mathbf{x}^{(1)} \in \mathbb{R}^d$ is an arbitrary, but deterministic quantity.

at vector \mathbf{x} and the vector \mathbf{x} itself. We specialize the value of constants η_1, η_2 for different nonlinearities in the Supplement. We are now ready to state our high-probability convergence bounds for non-convex costs.

Theorem 1. *Let Assumptions 1, 2 and 4 hold. Let $\{\mathbf{x}^{(t)}\}_{t \in \mathbb{N}}$ be the sequence generated by Algorithm 1, with step-size $\alpha_t = \frac{a}{(t+1)^\delta}$, for any $\delta \in (2/3, 1)$ and $a > 0$. Then, for any $t \geq 1$ and $\beta \in (0, 1)$, with probability at least $1 - \beta$, the following hold.*

1. For the choice $\delta \in (2/3, 3/4)$, we have

$$\begin{aligned} \min_{k \in [t]} \|\nabla f(\mathbf{x}^{(k)})\|^2 &= \frac{2R_1(\beta)/\eta_2}{(t+2)^{1-\delta} - 2^{1-\delta}} \\ &+ \frac{2R_2/\eta_2}{(t+2)^{3\delta-2} - 2^{3\delta-2}} + \left(\frac{2R_1(\beta)/\eta_1}{(t+2)^{1-\delta} - 2^{1-\delta}} \right)^2 \\ &+ \left(\frac{2R_2/\eta_1}{(t+2)^{3\delta-2} - 2^{3\delta-2}} \right)^2, \end{aligned}$$

where $R_1(\beta) \triangleq (1 - \delta) [(f(\mathbf{x}^{(1)}) - f^* + \log(1/\beta)) / a + aLC^2(1/2 + 8LD_{\mathcal{X}})/(2\delta - 1)]$ and $R_2 \triangleq \frac{8a^3C^4L^2}{(1-\delta)(3-4\delta)}$.

2. For the choice $\delta = 3/4$, we have

$$\begin{aligned} \min_{k \in [t]} \|\nabla f(\mathbf{x}^{(k)})\|^2 &= \frac{2R_3(t, \beta)/\eta_2}{(t+2)^{1/4} - 2^{1/4}} \\ &+ \left(\frac{\sqrt{2}R_3(t, \beta)/\eta_1}{(t+2)^{1/4} - 2^{1/4}} \right)^2, \end{aligned}$$

where $R_3(t, \beta) \triangleq (f(\mathbf{x}^{(1)}) - f^* + \log(1/\beta)) / 4a + aLC^2(1/4 + 4LD_{\mathcal{X}}) + 32a^3C^4L^2 \log(t+1)$.

3. For the choice $\delta \in (3/4, 1)$, we have

$$\begin{aligned} \min_{k \in [t]} \|\nabla f(\mathbf{x}^{(k)})\|^2 &= \frac{2R_4(\beta)/\eta_2}{(t+2)^{1-\delta} - 2^{1-\delta}} \\ &+ \left(\frac{\sqrt{2}R_4(\beta)/\eta_1}{(t+2)^{1-\delta} - 2^{1-\delta}} \right)^2, \end{aligned}$$

where $R_4(\beta) \triangleq (1 - \delta) [(f(\mathbf{x}^{(1)}) - f^* + \log(1/\beta)) / a + aLC^2(1/2 + 8LD_{\mathcal{X}})/(2\delta - 1)] + 8a^3C^4L^2/(1-\delta)(4\delta-3)$.

Remark 9. Theorem 1 provides convergence in high-probability of nonlinear SGD in the online setting, for a broad range of nonlinearities and step-sizes, with the best rate achieved for the choice $\delta = 3/4$. Compared to Nguyen et al. (2023a), who achieve the rate $\mathcal{O}(t^{2(1-p)/(3p-2)} \log(t/\beta))$ for clipped SGD, with the exponent explicitly depending on p and vanishing as $p \rightarrow 1$, our results apply to a broad range of nonlinearities and are strictly better whenever $p < 6/5$.

Remark 10. Note that for both step-size choices $\delta_1 \in (2/3, 3/4)$ and $\delta_2 \in (3/4, 1)$, we can get arbitrarily close to the rate $t^{-1/4}$, by choosing $\delta_1 = 3/4 - \epsilon_1$, for $\epsilon_1 \in (0, 1/12)$ and $\delta_2 = 3/4 + \epsilon_2$, for $\epsilon_2 \in (0, 1/4)$. In both cases, the rate incurs a constant multiplicative factor $1/\epsilon_i$, $i \in [2]$.

Remark 11. For the choice of $\delta = 3/4$, our rate incurs an additional $\log(t+1)$ factor. This additional factor is unavoidable in online learning, where the time horizon is unknown and a time-varying step-size is required. The rate in Nguyen et al. (2023a) incurs the same logarithmic factor for the “unknown T ” regime (see Theorem B.2 in their work). The logarithmic factor can be removed if the time horizon $T \in \mathbb{N}$ is preset, by using a fixed step-size inversely proportional to the time horizon, i.e., $\alpha_t \equiv \alpha \propto T^{-3/4}$. Our analysis readily goes through when a fixed step-size is used.

Remark 12. The guarantees in Theorem 1 (and Theorem 3 ahead) are stated in terms of the metric $\min_{k \in [t]} \|\nabla f(\mathbf{x}^{(k)})\|^2$, widely used in non-convex optimization. However, in our proof, we provide convergence guarantees of the same order for the metric $\frac{1}{t} \sum_{k=1}^t \min\{\|\nabla f(\mathbf{x}^{(k)})\|, \|\nabla f(\mathbf{x}^{(k)})\|^2\}$, which is more general, in the sense that the high-probability bounds on this metric imply the bounds on the metric $\min_{k \in [t]} \|\nabla f(\mathbf{x}^{(k)})\|^2$. For details, as well as comparisons with the metric used in Nguyen et al. (2023a), see the Supplement.

Remark 13. The convergence bounds in Theorem 1 depend on standard quantity, such as the initialization gap (through $f(\mathbf{x}^{(1)}) - f^*$ and $D_{\mathcal{X}}$) and smoothness parameter L , as well as nonlinearity and noise dependent quantities, such as C , η_1 and η_2 . These constants can be specialized for specific nonlinearities and noises. For example, consider the noise from Example 1, sign nonlinearity and step-size parameter $\delta = 3/4$. It can then be shown that $C = \sqrt{d}$, $\eta_1 = (\alpha-1)/2\alpha\sqrt{d}$, $\eta_2 = (\alpha-1)/d$ (see the Supplement), resulting in the following problem related constant (up to a logarithmic factor) in the leading term $\frac{ad^2L(1/4+4LD_{\mathcal{X}})}{\alpha-1} + \frac{32a^3d^3L^2\log(t+1)}{\alpha-1} + \frac{d(f(\mathbf{x}^{(1)})-f^*+\log(1/\beta))}{4a(\alpha-1)}$, where we recall that $\alpha > 2$. Choosing $\alpha = d^{-1/2}$, reduces the overall dependence on problem dimension to $d^{3/2}$.

We specialize the rates from Theorem 1 for specific choices of nonlinearities and noise in the Supplement, showing that our rates predict that *clipping is not always the optimal choice of nonlinearity* and confirm the findings of Zhang et al. (2020), namely that component-wise clipping demonstrates better dimension dependence than joint clipping, for some noise instances. This is further validated in our numerical experiments in Section 4.

Next, if the cost is strongly convex, results of Theorem 1 can be improved. Define $\tilde{\alpha}_k \triangleq \alpha_k / \sum_{j=1}^t \alpha_j$, $k \in [t]$, so that $\sum_{k=1}^t \tilde{\alpha}_k = 1$ and define a weighted average of iterates as $\hat{\mathbf{x}}^{(t)} \triangleq \sum_{k=1}^t \tilde{\alpha}_k \mathbf{x}^{(k)}$. The estimator $\hat{\mathbf{x}}^{(t)}$ can be seen as generalized Polyak-Ruppert averaging, e.g., Ruppert (1988); Polyak (1990); Polyak and Juditsky (1992). We then have the following result.

Corollary 1.1. Let Assumptions 1-4 hold. Let $\{\mathbf{x}^{(t)}\}_{t \in \mathbb{N}}$ be the sequence generated by Algorithm 1, with step-size $\alpha_t = \frac{a}{(t+1)^\delta}$, for any $\delta \in (2/3, 1)$ and $a > 0$. Then, for any $t \geq 1$ and any $\beta \in (0, 1)$, with probability at least $1 - \beta$, the following hold.

1. For the choice $\delta \in (2/3, 3/4)$, we have $\|\hat{\mathbf{x}}^{(t)} - \mathbf{x}^*\|^2 = \mathcal{O}((t^{\delta-1} + t^{2-3\delta})\log(1/\beta))$.
2. For the choice $\delta = 3/4$, we have $\|\hat{\mathbf{x}}^{(t)} - \mathbf{x}^*\|^2 = \mathcal{O}(t^{-1/4}\log(t/\beta))$.
3. For the choice $\delta \in (3/4, 1)$, we have $\|\hat{\mathbf{x}}^{(t)} - \mathbf{x}^*\|^2 = \mathcal{O}(t^{\delta-1}\log(1/\beta))$.

Remark 14. Corollary 1.1 maintains the rates from Theorem 1, while improving on the metric of interest, providing guarantees for the generalized Polyak-Ruppert average $\hat{\mathbf{x}}^{(t)}$. Compared to Sadiev et al. (2023), who show convergence of the last iterate for clipped SGD in the offline setting, with a rate $\mathcal{O}(T^{2(1-p)/p})$, our results again apply to a much broader range of nonlinearities and the online setting, beating the rate from Sadiev et al. (2023) whenever $p < 8/7$.

For strongly convex functions it is of interest to characterize the convergence guarantees of the last iterate Harvey et al. (2019); Tsai et al. (2022); Sadiev et al. (2023). To that end, we first characterize the interplay between $\Phi^{(t)}$ and $\nabla f(\mathbf{x}^{(t)})$.

Lemma 3.3. Let Assumptions 1-4 hold and $\{\mathbf{x}^{(t)}\}_{t \in \mathbb{N}}$ be the sequence generated by Algorithm 1, with step-size $\alpha_t = \frac{a}{(t+1)^\delta}$, for any $\delta \in (1/2, 1)$ and $a > 0$. Then $\langle \Phi^{(t)}, \nabla f(\mathbf{x}^{(t)}) \rangle \geq \gamma(t+2)^{\delta-1} \|\nabla f(\mathbf{x}^{(t)})\|^2$, for some $\gamma = \gamma(a) > 0$ and any $t \geq 1$, almost surely.

We then have the following result.

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

$$\|\mathbf{x}^{(t+1)} - \mathbf{x}^*\|^2 = \mathcal{O}(\log(1/\beta)(t+1)^{-\zeta}),$$

where $\zeta = \min\{2\delta - 1, a\mu\gamma/2\}$.

We specialize $\gamma = \gamma(a)$ for different nonlinearities and discuss its impact on the rate in the Supplement. The value of ζ can be explicitly calculated for specific choices of nonlinearities and noise, as we show next.

Example 5. For the noise from Example 1 and sign nonlinearity, it can be shown that $\zeta \approx \min\{2\delta - 1, \frac{\mu}{L} \frac{1-\delta}{\sqrt{d}} \frac{\alpha-1}{\alpha}\}$, while for component-wise clipping it can be shown that $\zeta \approx \min\{2\delta - 1, \frac{\mu}{L\sqrt{d}} \frac{(1-\delta)(m-1)(1-(m+1)^{-\alpha})}{m}\}$, see Jakovetić et al. (2023). On the other hand, the rate from Sadiev et al. (2023) for joint clipping, adapted to the same noise, is

$2(r-1)/r$, where $r \leq \min\{\alpha - 1, 2\}$. If $\alpha = 2 + \epsilon$, where $\epsilon \in (0, 1]$, i.e., very heavy tails, then moments of order $1 < p < 1 + \epsilon \leq 2$ exist, satisfying the bounded p -th moment condition from Sadiev et al. (2023), with their best-case rate given by $2(r-1)/r = 2\epsilon/(1+\epsilon) < 2\epsilon$. On the other hand, consider the sign nonlinearity with step-size parameter $\delta = 3/4$. In this case, our rate is given by $\zeta = \min\{\frac{1}{2}, \frac{\mu(1+\epsilon)}{4L\sqrt{d}(2+\epsilon)}\} = \frac{\mu(1+\epsilon)}{4L\sqrt{d}(2+\epsilon)} > \frac{\mu}{8L\sqrt{d}}$, where the last inequality follows from $\epsilon \in (0, 1]$. Using the corresponding lower and upper bounds, it follows that our rate is strictly better than the one from Sadiev et al. (2023), i.e., $\zeta > 2(r-1)/r$, whenever $\epsilon < \frac{\mu}{16L\sqrt{d}}$. Therefore, for any heavy-tailed noise of the form given in Example 1, such that $\alpha = 2 + \epsilon$, with $0 < \epsilon < \frac{\mu}{16L\sqrt{d}}$, the noise condition in both our work and Sadiev et al. (2023) is satisfied, with our rate being strictly better. Additionally, we can see that our rate for noises of this form is uniformly bounded below by a quantity constant with respect to α , i.e., $\zeta > \frac{\mu}{8L\sqrt{d}}$. On the other hand, the rate from Sadiev et al. (2023), specialized to noises from Example 1 with $\alpha = 2 + \epsilon$ and $\epsilon \in (0, 1]$, is upper-bounded by a quantity strictly depending on the noise, i.e., $2(r-1)/r < 2\epsilon$, with $2(r-1)/r \rightarrow 0$ as $\alpha \rightarrow 2$ (i.e., as $\epsilon \rightarrow 0$). Similar results can be shown to hold for component-wise clipping.

3.3 Beyond Symmetric Noise

In this section we extend our results to the case when the noise is not necessarily symmetric. In particular, we make the following assumption on the noise vectors.

Assumption 5. The noise vectors $\{\mathbf{z}^{(t)}\}_{t \in \mathbb{N}}$ are independent, identically distributed, drawn from a mixture distribution with PDF $P(\mathbf{z}) = (1 - \lambda)P_1(\mathbf{z}) + \lambda P_2(\mathbf{z})$, where $P_1(\mathbf{z})$ is symmetric and $\lambda \in (0, 1)$ is the mixture coefficient. Additionally, P_1 is positive around zero, i.e., $P_1(\mathbf{z}) > 0$, for all $\|\mathbf{z}\| \leq B_0$ and some $B_0 > 0$.

Remark 15. Assumption 5 relaxes Assumption 4, by allowing for a mixture of symmetric and non-symmetric noises. The resulting noise is non-symmetric and in general does not have to be zero mean, i.e., we allow for the oracle to send *biased* gradient estimators. We again make no assumptions on noise moments.

Remark 16. Assumption 5 arises naturally in scenarios like training with a large batch size, in which the generalized CLT implies that the noise becomes more symmetric as the batch size grows Simsekli et al. (2019b); Peluchetti et al. (2020); Gurbuzbalaban et al. (2021); Barsbey et al. (2021). In such scenarios, the effect of the non-symmetric part decreases with batch size, resulting in small λ for a large enough batch size.

We then have the following result.

Theorem 3. *Let Assumptions 1, 2 and 5 hold. Let $\{\mathbf{x}^{(t)}\}_{t \in \mathbb{N}}$ be the sequence generated by Algorithm 1,*

with step-size $\alpha_t = \frac{a}{(t+1)^\delta}$, for any $\delta \in (2/3, 1)$ and $a > 0$. If $\lambda < \frac{\eta_1}{C+\eta_1}$, then for any $t \geq 1$ and $\beta \in (0, 1)$, with probability at least $1 - \beta$, the following hold.

1. *For the choice $\delta \in (2/3, 3/4)$, we have $\min_{k \in [t]} \|\nabla f(\mathbf{x}^{(k)})\|^2 = \mathcal{O}(t^{\delta-1} + t^{2-3\delta}) + \frac{\eta_1 \lambda C}{\eta_2^2(1-\lambda)}$.*
2. *For the choice $\delta = 3/4$, we have $\min_{k \in [t]} \|\nabla f(\mathbf{x}^{(k)})\|^2 = \mathcal{O}(\log(t)/t^{1/4}) + \frac{\eta_1 \lambda C}{\eta_2^2(1-\lambda)}$.*
3. *For the choice $\delta \in (3/4, 1)$, we have $\min_{k \in [t]} \|\nabla f(\mathbf{x}^{(k)})\|^2 = \mathcal{O}(t^{\delta-1}) + \frac{\eta_1 \lambda C}{\eta_2^2(1-\lambda)}$.*

Remark 17. All three step-size regimes in Theorem 3 again achieve exponential tail decay, i.e., a $\log(1/\beta)$ dependence on the failure probability β , which is hidden under the big O notation, for ease of exposition.

Remark 18. Theorem 3 provides convergence guarantees to a neighbourhood of stationarity, for mixtures of symmetric and non-symmetric components, if the contribution of the non-symmetric component is sufficiently small. As discussed in Remark 16, this can be guaranteed, e.g., by using a sufficiently large batch size. As the neighborhood size is determined by the mixture component via $\lambda/(1-\lambda)$, it follows that, as the noise becomes more symmetric (i.e., $\lambda \rightarrow 0$), we recover exact convergence from Theorem 1.

Remark 19. While Nguyen et al. (2023a) guarantee convergence of gradient norm-squared to zero under condition (BM), which allows for non-symmetric noises, it is important to note that they explicitly require that the oracle sends *unbiased* gradient estimators. On the other hand, we allow for the oracle to send *biased* gradient estimators. Without incorporating a correction mechanism (e.g., momentum or error-feedback), in general, it is not possible to guarantee exact convergence with a biased oracle.

The size of the neighbourhood and the condition on the mixture coefficient provided in Theorem 3 are both determined by the noise and choice of nonlinearity. We can specialize the constants η_1, η_2 and C for specific choices of nonlinearities and noises. We now give some examples. For full derivations, see the Supplement.

Example 6. Consider the noise from Example 1. For the sign nonlinearity it can be shown that $\eta_1 = (\alpha-1)/2\alpha\sqrt{d}$, $\eta_2 = (\alpha-1)/2d$ and $C = \sqrt{d}$, resulting in convergence to a neighborhood of size $2d^2\lambda/[\alpha(\alpha-1)(1-\lambda)]$ and $\lambda < (\alpha-1)/[\alpha(2d+1)-1]$. As $\alpha > 2$, we can see that $\lambda < 1/(2d+1)$, at best and $\lambda < 1/(4d+1)$, at worst.

Example 7. For component-wise clipping with $m > 1$, it can be shown that $\eta_1 = [1-(m+1)^{-\alpha}]/(m-1)/2\sqrt{d}$, $\eta_2 = [1-(m+1)^{-\alpha}]/2d$ and $C = m\sqrt{d}$, resulting in convergence to a neighborhood of size $2d^2m(m-1)\lambda/[1-(m+1)^{-\alpha}](1-\lambda)$ and $\lambda < (m-1)[1-(m+1)^{-\alpha}]/[(m-1)[1-(m+1)^{-\alpha}] + 2md]$. While taking $m \rightarrow 1$ results in full convergence, we can see that this simultaneously implies $\lambda \rightarrow 0$, i.e., requiring the noise to be symmetric.

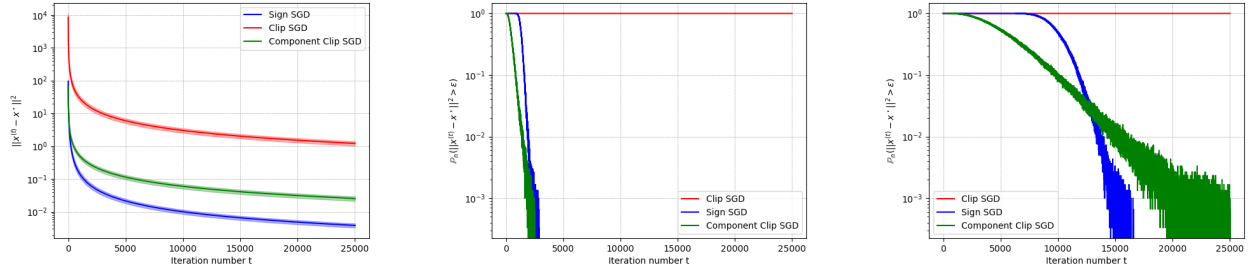


Figure 1: Performance of sign, component-wise clipping and joint clipping. Left to right: MSE performance and high-probability performance for $\varepsilon = \{0.1, 0.01\}$, respectively. We can see that both component-wise nonlinearities converge faster in the MSE sense and achieve exponential tail decay. Note that clipping does not achieve exponentially decaying tails in second and third figures, as it does not reach the required accuracy in the allocated number of iterations.

Example 8. For joint clipping with threshold $M > 0$, we have $\eta_1 = [(\alpha-1)/2]^d \min\{1, M\}/2$, $\eta_2 = [(\alpha-1)/2]^d \min\{1, M\}$ and $C = M$, resulting in convergence to a neighborhood of size $2^{d-1}\lambda M / [(\alpha-1)^d(1-\lambda) \min\{1, M\}]$ and $\lambda < \frac{(\alpha-1)^d \min\{1, M\}}{2(\alpha-1)^d \min\{1, M\} + 2^{d+1}M}$. Choosing $M \leq 1$, results in converging to a neighborhood of size $2^{d-1}\lambda / (\alpha-1)^d(1-\lambda)$ and condition $\lambda < (\alpha-1)^d / 2(\alpha-1)^d + 2^{d+1}$. Similar observations hold for $M > 1$.

4 NUMERICAL RESULTS

In this section we present numerical results. We consider a quadratic problem $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top \mathbf{A} \mathbf{x} + \mathbf{b}^\top \mathbf{x}$, where $\mathbf{A} \in \mathbb{R}^{d \times d}$ is positive definite and $\mathbf{b} \in \mathbb{R}^d$ is fixed. We set $d = 100$. The stochastic noise is generated according to the PDF from Example 1, with $\alpha = 2.05$. We compare the performance of sign, component-wise and joint clipped SGD, with all three using the step-size schedule $\alpha_t = \frac{1}{t+1}$. We choose the clipping thresholds m and M for which component-wise and joint clipping performed the best, those 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:

1. *Mean-squared error:* we present the MSE of the algorithms, by evaluating the gap $\|\mathbf{x}^{(t)} - \mathbf{x}^*\|^2$ in each iteration, averaged across all runs, i.e., the final estimator at iteration $t = 1, \dots, T$, is given by $MSE^t = \frac{1}{R} \sum_{r=1}^R \|\mathbf{x}_r^{(t)} - \mathbf{x}^*\|^2$, where $\mathbf{x}_r^{(t)}$ is the t -th iterate in the r -th run, generated by the algorithms.
2. *Tail estimate:* we evaluate the tail behaviour of the methods as follows. We consider the events $A^t = \{\|\mathbf{x}^{(t)} - \mathbf{x}^*\|^2 > \varepsilon\}$, for a fixed $\varepsilon > 0$. To estimate the probability of A^t , for each $t = 1, \dots, T$, we construct a Monte Carlo estimator, by sampling $n = 3000$ instances from the $R = 5000$ runs, uniformly with replacement. We compute the empirical probability as

$\mathbb{P}_n(A^t) = \frac{1}{n} \sum_{i=1}^n \mathbb{I}(\{\|\mathbf{x}_i^{(t)} - \mathbf{x}^*\|^2 > \varepsilon\})$, where $\mathbb{I}(\cdot)$ is the indicator and $\mathbf{x}_i^{(t)}$ is the i -th Monte Carlo sample.

The results are presented in Figure 1. We can see that component-wise nonlinearities outperform joint clipping, both in terms of MSE and high-probability performance and demonstrate exponential tail decay, validating our theoretical results and further underlining the usefulness of considering a general framework beyond only clipping. Additional experiments can be found in the Supplement.

5 CONCLUSION

We present high-probability guarantees for a broad class of nonlinear SGD algorithms in the online setting and the presence of heavy-tailed noise with symmetric PDF. We establish near-optimal $\tilde{\mathcal{O}}(t^{-1/4})$ convergence rates for non-convex costs, and similar rates for the weighted average of iterates for strongly convex costs. Additionally, for the last iterate of strongly convex costs we establish convergence at a rate $\mathcal{O}(t^{-\zeta})$, where $\zeta \in (0, 1)$ depends on noise and other problem parameters. We extend our analysis to noises that are mixtures of symmetric and non-symmetric components, showing convergence to a neighbourhood of stationarity, where the size of the neighborhood depends on choice of nonlinearity, noise and mixture coefficient. Compared to state-of-the-art works Nguyen et al. (2023a); Sadiev et al. (2023), we extend the high-probability convergence guarantees to a broad class of nonlinearities, relax the noise moment condition, and demonstrate regimes in which our convergence rates are strictly better. This is made possible by a novel result on the interplay between the “denoised” nonlinearity and the gradient. Numerical results confirm the theory and demonstrate that clipping, exclusively considered in prior works, is not always the optimal choice of nonlinearity, further highlighting the importance and usefulness of our general framework.

Acknowledgments. The work of AA, SY and SK is supported in part by NSF under grant ECCS-2330196. The work of PS and GJ is supported in part by NSF grants CCF 2045694, CNS-2112471, CPS-2111751, SHF-2107024, and ONR N00014-23-1-2149. The work of DB is supported by the Ministry of Science, Technological Development and Innovation (No. 451-03-65/2024-03/200156), by Faculty of Technical Sciences, University of Novi Sad, project “Scientific and Artistic Research Work of Researchers in Teaching and Associate Positions at the Faculty of Technical Sciences, University of Novi Sad” (No. 01-3394/1), and by the European Union’s Horizon Europe program, grant agreement 101135916. The work of DJ is supported by the Science Fund of Republic of Serbia, project “LASCADO”, grant No. 7359, by Provincial Secretariat for Higher Education and Scientific Research, grant No. 142-451- 2593/2021-01/2, and by the European Union’s Horizon Europe program, grant agreements No. 101093006 and 101135775. Correspondence: A. Armacki (aarmacki@andrew.cmu.edu).

References

- 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.)
- Barsbey, M., Sefidgaran, M., Erdogdu, M. A., Richard, G., and Simsekli, U. (2021). Heavy tails in sgd and compressibility of overparametrized neural networks. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*, volume 34, pages 29364–29378. Curran Associates, Inc. (Cited on pages 2, 5, and 8.)
- Battash, B., Wolf, L., and Lindenbaum, O. (2024). Revisiting the noise model of stochastic gradient descent. In Dasgupta, S., Mandt, S., and Li, Y., editors, *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 4780–4788. PMLR. (Cited on pages 2 and 5.)
- 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 5.)
- 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 pages 2 and 5.)
- 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 pages 2 and 5.)
- Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT’2010: 19th International Conference on Computational Statistics Paris France, August 22-27, 2010 Keynote, Invited and Contributed Papers*, pages 177–186. Springer. (Cited on page 1.)
- 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 1.)
- Chen, X., Wu, S. Z., and Hong, M. (2020). Understanding gradient clipping in private sgd: A geometric perspective. *Advances in Neural Information Processing Systems*, 33:13773–13782. (Cited on pages 2, 5, 26, and 30.)
- 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 1.)
- Eldowa, K. and Paudice, A. (2023). General tail bounds for non-smooth stochastic mirror descent. *arXiv preprint arXiv:2312.07142*. (Cited on page 3.)
- 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 5.)
- 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 page 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.)
- Gurbuzbalaban, M., Simsekli, U., and Zhu, L. (2021). The heavy-tail phenomenon in sgd. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 3964–3975. PMLR. (Cited on pages 2, 5, and 8.)
- 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 1.)
- 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 1, 3, 7, and 22.)
- 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 19.)
- Hübner, F., Fatkhullin, I., and He, N. (2024). From gradient clipping to normalization for heavy tailed sgd. *arXiv preprint arXiv:2410.13849*. (Cited on page 30.)
- 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 3, 7, 15, 16, 20, and 25.)
- Lan, G. (2012). An optimal method for stochastic composite optimization. *Mathematical Programming*, 133(1-2):365–397. (Cited on page 3.)
- LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, 521(7553):436–444. (Cited on page 26.)
- LeCun, Y., Cortes, C., and Burges, C. J. (1998). The mnist database of handwritten digits. <http://yann.lecun.com/exdb/mnist/>. (Cited on page 26.)
- 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 5.)
- 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. (2024). High probability convergence bounds for non-convex stochastic gradient descent with sub-weibull noise. *Journal of Machine Learning Research*, 25(241):1–36. (Cited on pages 3 and 5.)
- 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 1.)
- Nair, J., Wierman, A., and Zwart, B. (2022). *The Fundamentals of Heavy Tails: Properties, Emergence, and Estimation*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press. (Cited on page 5.)
- 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 1 and 3.)

- Nesterov, Y. (2018). *Lectures on Convex Optimization*. Springer Publishing Company, Incorporated, 2nd edition. (Cited on pages 5, 14, and 19.)
- 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, 6, 7, 8, 9, and 30.)
- 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.)
- Peluchetti, S., Favaro, S., and Fortini, S. (2020). Stable behaviour of infinitely wide deep neural networks. In Chiappa, S. and Calandra, R., editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 1137–1146. PMLR. (Cited on pages 2, 5, and 8.)
- Polyak, B. (1990). New stochastic approximation type procedures. *Avtomatica i Telemekhanika*, 7:98–107. (Cited on page 7.)
- Polyak, B. and Juditsky, A. (1992). Acceleration of Stochastic Approximation by Averaging. *SIAM Journal on Control and Optimization*, 30:838–855. (Cited on page 7.)
- Polyak, B. and Tsytkin, Y. (1979). Adaptive estimation algorithms: Convergence, optimality, stability. *Automation and Remote Control*, 1979. (Cited on page 15.)
- 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 Conference on Machine Learning*, pages 1571–1578. (Cited on pages 1 and 2.)
- Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407. (Cited on page 1.)
- 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 7.)
- 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, 7, 8, and 9.)
- Simsekli, U., Gurbuzbalaban, M., Nguyen, T. H., Richard, G., and Sagun, L. (2019a). On the heavy-tailed theory of stochastic gradient descent for deep neural networks. *arXiv preprint arXiv:1912.00018*. (Cited on page 2.)
- Simsekli, U., Sagun, L., and Gurbuzbalaban, M. (2019b). A tail-index analysis of stochastic gradient noise in deep neural networks. In *International Conference on Machine Learning*, pages 5827–5837. PMLR. (Cited on pages 2, 5, and 8.)
- 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 and 7.)
- 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 14.)
- Wright, S. J. and Recht, B. (2022). *Optimization for Data Analysis*. Cambridge University Press. (Cited on page 5.)
- 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 page 2.)
- 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.)
- Zeiler, M. D. (2012). Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*. (Cited on page 26.)
- 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, 7, 25, and 30.)

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.)

Checklist

1. For all models and algorithms presented, check if you include:
 - (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
 - (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
 - (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [No]
2. For any theoretical claim, check if you include:
 - (a) Statements of the full set of assumptions of all theoretical results. [Yes]
 - (b) Complete proofs of all theoretical results. [Yes]
 - (c) Clear explanations of any assumptions. [Yes]
3. For all figures and tables that present empirical results, check if you include:
 - (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes] While we do not provide the specific code for the experiments, we provide all the details necessary to fully reproduce our results.
 - (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
 - (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
 - (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [No] Our experiments were run on a laptop, requiring no special computing infrastructure.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
 - (a) Citations of the creator If your work uses existing assets. [Not Applicable]
 - (b) The license information of the assets, if applicable. [Not Applicable]
 - (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
 - (d) Information about consent from data providers/curators. [Not Applicable]
 - (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
 - (a) The full text of instructions given to participants and screenshots. [Not Applicable]
 - (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
 - (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

High-probability Convergence Bounds for Online Nonlinear Stochastic Gradient Descent Under Heavy-tailed Noise: Supplementary Materials

S1 INTRODUCTION

The Supplement presents results omitted from the main body. Section S2 provides some useful facts and results used in the proofs, Section S3 provides the proofs omitted from the main body, Section S4 specializes the rate exponent ζ from Theorem 2 for component-wise and joint nonlinearities, Section S5 details the derivations for Examples 6-8, Section S6 provides an analytical example for which our rates predict that joint clipping is not the optimal choice of nonlinearity, Section S7 provides additional experiments, Section S8 provides a detailed discussion on the noise assumption used in our work, while Section S9 provides a discussion on the metric used in our work.

S2 USEFUL FACTS

In this section we present some useful facts and results, concerning L -smooth, μ -strongly convex 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 $\mathbf{x}^* = \arg \min_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x})$. Then, for any $\mathbf{x} \in \mathbb{R}^d$, we have

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

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

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 $\sigma\sqrt{2}$ -sub-Gaussian, i.e., for any $v \in \mathbb{R}^d$, we have

$$\mathbb{E}e^{\langle X, v \rangle} \leq e^{\sigma^2 \|v\|^2}.$$

Proof. The proof follows a similar idea to the one of proving sub-Gaussian properties in, e.g., Vershynin (2018). Using the general inequality $e^x \leq x + e^{x^2}$, which holds for any $x \in \mathbb{R}$, setting $x = \langle X, v \rangle$, we get

$$\mathbb{E}[\exp(\langle X, v \rangle)] \leq \mathbb{E}[\exp(\langle X, v \rangle^2)] \leq e^{\sigma^2 \|v\|^2},$$

where the first inequality follows from the fact that X is zero mean, while the second follows from the Cauchy-Schwartz inequality and $\|X\| \leq \sigma$. \square

S3 MISSING PROOFS

In this section we provide proofs omitted from the main body. Subsection S3.1 proves results pertinent to Theorem 1, Subsection S3.2 proves results relating to Theorem 2, while Subsection S3.3 proves Theorem 3.

S3.1 Proof of Theorem 1

In this section we prove Lemmas 3.1, 3.2, Theorem 1 and Corollary 1.1. We begin by proving Lemma 3.1.

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

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

Moreover, by Assumption 1, we have $\|\mathbf{e}^{(t)}\| = \|\Phi^{(t)} - \Psi^{(t)}\| \leq \|\Phi^{(t)}\| + \|\Psi^{(t)}\| \leq \mathbb{E}\|\Psi^{(t)}\| + C \leq 2C$, 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. \square

Prior to proving Lemma 3.2, we state two results used in the proof. The first result, due to Polyak and Tsypkin (1979), provides some properties of the mapping Φ for component-wise nonlinearities under symmetric noise.

Lemma S3.1. *Let Assumptions 1 and 4 hold, with the nonlinearity $\Psi : \mathbb{R}^d \mapsto \mathbb{R}^d$ being component wise, i.e., of the form $\Psi(\mathbf{x}) = [\mathcal{N}_1(x_1), \dots, \mathcal{N}_1(x_d)]^\top$. Then, the function $\Phi : \mathbb{R}^d \mapsto \mathbb{R}^d$ is of the form $\Phi(\mathbf{x}) = [\phi_1(x_1), \dots, \phi_d(x_d)]^\top$, where $\phi_i(x_i) = \mathbb{E}_{z_i} [\mathcal{N}_1(x_i + z_i)]$ 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 $\phi'_i(0) > 0$.

The second result, due to Jakovetić et al. (2023), gives a useful property of Φ for joint nonlinearities.

Lemma S3.2. *Let Assumption 1 hold, with the nonlinearity $\Psi : \mathbb{R}^d \mapsto \mathbb{R}^d$ being joint, i.e., of the form $\Psi(\mathbf{x}) = \mathbf{x} \mathcal{N}_2(\|\mathbf{x}\|)$. Then for any $\mathbf{x}, \mathbf{z} \in \mathbb{R}^d$ such that $\|\mathbf{z}\| > \|\mathbf{x}\|$*

$$|\mathcal{N}_2(\|\mathbf{x} + \mathbf{z}\|) - \mathcal{N}_2(\|\mathbf{x} - \mathbf{z}\|)| \leq \|\mathbf{x}\|/\|\mathbf{z}\| [\mathcal{N}_2(\|\mathbf{x} + \mathbf{z}\|) + \mathcal{N}_2(\|\mathbf{x} - \mathbf{z}\|)].$$

Define $\phi'(0) \triangleq \min_{i \in [d]} \phi'_i(0)$ and $p_0 \triangleq P(0)$. We are now ready to prove Lemma 3.2. For convenience, we restate the full lemma below.

Lemma S3.3. *Let Assumptions 1 and 4 hold. Then, for any $\mathbf{x} \in \mathbb{R}^d$, we have $\langle \Phi(\mathbf{x}), \mathbf{x} \rangle \geq \min \{\eta_1 \|\mathbf{x}\|, \eta_2 \|\mathbf{x}\|^2\}$, where $\eta_1, \eta_2 > 0$, are noise, nonlinearity and problem dependent constants. In particular, if the nonlinearity Ψ is component-wise, we have $\eta_1 = \phi'(0)\xi/2\sqrt{d}$ and $\eta_2 = \phi'(0)/2d$, where $\xi > 0$ is a constant that depends only on the noise and choice of nonlinearity. If Ψ is a joint nonlinearity, then $\eta_1 = p_0 \mathcal{N}_2(1)/2$ and $\eta_2 = p_0 \mathcal{N}_2(1)$.*

Proof of Lemma S3.3. First, consider the case when $\Phi(\mathbf{x}) = [\mathcal{N}_1(x_1), \dots, \mathcal{N}_1(x_d)]^\top$ is component-wise. From Lemma S3.1 it follows that, 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 $\lim_{x \rightarrow 0} h_i(x) = 0$. Recalling that $\phi'(0) = \min_{i \in [d]} \phi'_i(0) > 0$, it follows that there exists a $\xi > 0$ (depending only on the nonlinearity \mathcal{N}_1) 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 $\phi_i(x) \geq \frac{\phi'(0)x}{2}$. On the other hand, for $x > \xi$, since ϕ_i is non-decreasing, we have from the previous relation that $\phi_i(x) \geq \phi_i(\xi) \geq \frac{\phi'(0)\xi}{2}$. Therefore, it follows that $\phi_i(x) \geq \frac{\phi'(0)}{2} \min\{x, \xi\}$, for any $x \geq 0$. Combined with the oddity of ϕ_i , we get $x\phi_i(x) = |x|\phi_i(|x|) \geq \frac{\phi'(0)}{2} \min\{\xi|x|, x^2\}$, for any $x \in \mathbb{R}$. Using the previously established relations, we then have, for any vector $\mathbf{x} \in \mathbb{R}^d$

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

where the last inequality follows from the fact that $\|\mathbf{x}\|_\infty \geq \|\mathbf{x}\|/\sqrt{d}$. Next, consider the case when $\Phi(\mathbf{x}) = \mathbf{x}\mathcal{N}_2(\|\mathbf{x}\|)$ is joint. The proof follows a similar idea to the one in (Jakovetić et al., 2023, Lemma 6.2), with some important differences due to the different noise assumption. Fix an arbitrary $\mathbf{x} \in \mathbb{R}^d \setminus \{\mathbf{0}\}$. By the definition of Ψ , we have

$$\langle \Phi(\mathbf{x}), \mathbf{x} \rangle = \int_{\mathbf{z} \in \mathbb{R}^d} \underbrace{(\mathbf{x} + \mathbf{z})^\top \mathbf{x} \mathcal{N}_2(\|\mathbf{x} + \mathbf{z}\|)}_{\triangleq M(\mathbf{x}, \mathbf{z})} P(\mathbf{z}) d\mathbf{z} = \int_{\{\mathbf{z} \in \mathbb{R}^d : \langle \mathbf{z}, \mathbf{x} \rangle \geq 0\} \cup \{\mathbf{z} \in \mathbb{R}^d : \langle \mathbf{z}, \mathbf{x} \rangle < 0\}} M(\mathbf{x}, \mathbf{z}) P(\mathbf{z}) d\mathbf{z}.$$

Next, by symmetry of P , it readily follows that $\langle \Phi(\mathbf{x}), \mathbf{x} \rangle = \int_{J_1(\mathbf{x})} M_2(\mathbf{x}, \mathbf{z}) P(\mathbf{z}) d\mathbf{z}$, where $J_1(\mathbf{x}) \triangleq \{\mathbf{z} \in \mathbb{R}^d : \langle \mathbf{z}, \mathbf{x} \rangle \geq 0\}$ and $M_2(\mathbf{x}, \mathbf{z}) \triangleq (\|\mathbf{x}\|^2 + \langle \mathbf{z}, \mathbf{x} \rangle) \mathcal{N}_2(\|\mathbf{x} + \mathbf{z}\|) + (\|\mathbf{x}\|^2 - \langle \mathbf{z}, \mathbf{x} \rangle) \mathcal{N}_2(\|\mathbf{x} - \mathbf{z}\|)$. Consider the set $J_2(\mathbf{x}) \triangleq \left\{ \mathbf{z} \in \mathbb{R}^d : \frac{\langle \mathbf{z}, \mathbf{x} \rangle}{\|\mathbf{z}\|\|\mathbf{x}\|} \in [0, 0.5] \right\} \cup \{\mathbf{0}\}$. Clearly $J_2(\mathbf{x}) \subset J_1(\mathbf{x})$. Note that on $J_1(\mathbf{x})$ we have $\|\mathbf{x} + \mathbf{z}\| \geq \|\mathbf{x} - \mathbf{z}\|$, which, together with the fact that \mathcal{N}_2 is non-increasing, implies

$$\mathcal{N}_2(\|\mathbf{x} - \mathbf{z}\|) - \mathcal{N}_2(\|\mathbf{x} + \mathbf{z}\|) = |\mathcal{N}_2(\|\mathbf{x} - \mathbf{z}\|) - \mathcal{N}_2(\|\mathbf{x} + \mathbf{z}\|)|, \quad (5)$$

for any $\mathbf{z} \in J_1(\mathbf{x})$. For any $\mathbf{z} \in J_2(\mathbf{x})$ such that $\|\mathbf{z}\| > \|\mathbf{x}\|$, we then have

$$\begin{aligned} M_2(\mathbf{x}, \mathbf{z}) &= \|\mathbf{x}\|^2 [\mathcal{N}_2(\|\mathbf{x} - \mathbf{z}\|) + \mathcal{N}_2(\|\mathbf{x} + \mathbf{z}\|)] - \langle \mathbf{z}, \mathbf{x} \rangle [\mathcal{N}_2(\|\mathbf{x} - \mathbf{z}\|) - \mathcal{N}_2(\|\mathbf{x} + \mathbf{z}\|)] \\ &\stackrel{(a)}{=} \|\mathbf{x}\|^2 [\mathcal{N}_2(\|\mathbf{x} - \mathbf{z}\|) + \mathcal{N}_2(\|\mathbf{x} + \mathbf{z}\|)] - \langle \mathbf{z}, \mathbf{x} \rangle |\mathcal{N}_2(\|\mathbf{x} - \mathbf{z}\|) - \mathcal{N}_2(\|\mathbf{x} + \mathbf{z}\|)| \\ &\stackrel{(b)}{\geq} \|\mathbf{x}\|^2 [\mathcal{N}_2(\|\mathbf{x} - \mathbf{z}\|) + \mathcal{N}_2(\|\mathbf{x} + \mathbf{z}\|)] - \langle \mathbf{z}, \mathbf{x} \rangle \|\mathbf{x}\|/\|\mathbf{z}\| [\mathcal{N}_2(\|\mathbf{x} - \mathbf{z}\|) + \mathcal{N}_2(\|\mathbf{x} + \mathbf{z}\|)] \\ &\stackrel{(c)}{\geq} 0.5\|\mathbf{x}\|^2 [\mathcal{N}_2(\|\mathbf{x} - \mathbf{z}\|) + \mathcal{N}_2(\|\mathbf{x} + \mathbf{z}\|)], \end{aligned}$$

where (a) follows from (5), (b) follows from Lemma S3.2, while (c) follows from the definition of $J_2(\mathbf{x})$. Next, consider any $\mathbf{z} \in J_2(\mathbf{x})$, such that $0 < \|\mathbf{z}\| \leq \|\mathbf{x}\|$. We have

$$\begin{aligned} M_2(\mathbf{x}, \mathbf{z}) &= \|\mathbf{x}\|^2 [\mathcal{N}_2(\|\mathbf{x} - \mathbf{z}\|) + \mathcal{N}_2(\|\mathbf{x} + \mathbf{z}\|)] - \langle \mathbf{z}, \mathbf{x} \rangle [\mathcal{N}_2(\|\mathbf{x} - \mathbf{z}\|) - \mathcal{N}_2(\|\mathbf{x} + \mathbf{z}\|)] \\ &\stackrel{(a)}{=} \|\mathbf{x}\|^2 [\mathcal{N}_2(\|\mathbf{x} - \mathbf{z}\|) + \mathcal{N}_2(\|\mathbf{x} + \mathbf{z}\|)] - \langle \mathbf{z}, \mathbf{x} \rangle |\mathcal{N}_2(\|\mathbf{x} - \mathbf{z}\|) - \mathcal{N}_2(\|\mathbf{x} + \mathbf{z}\|)| \\ &\stackrel{(b)}{\geq} \|\mathbf{x}\|^2 [\mathcal{N}_2(\|\mathbf{x} - \mathbf{z}\|) + \mathcal{N}_2(\|\mathbf{x} + \mathbf{z}\|)] - 0.5\|\mathbf{x}\|^2 |\mathcal{N}_2(\|\mathbf{x} - \mathbf{z}\|) - \mathcal{N}_2(\|\mathbf{x} + \mathbf{z}\|)| \\ &\stackrel{(c)}{\geq} 0.5\|\mathbf{x}\|^2 [\mathcal{N}_2(\|\mathbf{x} - \mathbf{z}\|) + \mathcal{N}_2(\|\mathbf{x} + \mathbf{z}\|)], \end{aligned}$$

where (a) again follows from (5), (b) follows from the definition of $J_2(\mathbf{x})$ and the fact that $0 < \|\mathbf{z}\| \leq \|\mathbf{x}\|$, while (c) follows from the fact that \mathcal{N}_2 is non-negative and $|\mathcal{N}_2(\|\mathbf{x} - \mathbf{z}\|) - \mathcal{N}_2(\|\mathbf{x} + \mathbf{z}\|)| \leq \mathcal{N}_2(\|\mathbf{x} - \mathbf{z}\|) + \mathcal{N}_2(\|\mathbf{x} + \mathbf{z}\|)$. Finally, if $\mathbf{z} = \mathbf{0}$, we have $M_2(\mathbf{x}, \mathbf{0}) = 2\|\mathbf{x}\|^2 \mathcal{N}_2(\|\mathbf{x}\|) > 0.5\|\mathbf{x}\|^2 [\mathcal{N}_2(\|\mathbf{x} + \mathbf{0}\|) + \mathcal{N}_2(\|\mathbf{x} - \mathbf{0}\|)]$. Therefore, for any $\mathbf{z} \in J_2(\mathbf{x})$, we have $M_2(\mathbf{x}, \mathbf{z}) \geq 0.5\|\mathbf{x}\|^2 [\mathcal{N}_2(\|\mathbf{x} - \mathbf{z}\|) + \mathcal{N}_2(\|\mathbf{x} + \mathbf{z}\|)] \geq \|\mathbf{x}\|^2 \mathcal{N}_2(\|\mathbf{x}\| + \|\mathbf{z}\|)$, where the second inequality follows from the fact that \mathcal{N}_2 is non-increasing and $\|\mathbf{x} \pm \mathbf{z}\| \leq \|\mathbf{x}\| + \|\mathbf{z}\|$. Note that following a similar argument as above, it can be shown that $M_2(\mathbf{x}, \mathbf{z}) \geq 0$, for any $\mathbf{x} \in \mathbb{R}^d$ and $\mathbf{z} \in J_1(\mathbf{x})$. Combining everything, it readily follows that

$$\langle \Phi(\mathbf{x}), \mathbf{x} \rangle \geq \int_{J_2(\mathbf{x})} M_2(\mathbf{x}, \mathbf{z}) P(\mathbf{z}) d\mathbf{z} \geq \|\mathbf{x}\|^2 \int_{J_2(\mathbf{x})} \mathcal{N}_2(\|\mathbf{x}\| + \|\mathbf{z}\|) P(\mathbf{z}) d\mathbf{z}, \quad (6)$$

where the first inequality follows from the fact that $M_2(\mathbf{x}, \mathbf{z}) \geq 0$ on $J_1(\mathbf{x})$. Define $C_0 \triangleq \min\{B_0, 0.5\}$ and consider the set $J_3(\mathbf{x}) \subset J_2(\mathbf{x})$, defined as $J_3(\mathbf{x}) \triangleq \left\{ \mathbf{z} \in \mathbb{R}^d : \frac{\langle \mathbf{z}, \mathbf{x} \rangle}{\|\mathbf{z}\|\|\mathbf{x}\|} \in [0, 0.5], \|\mathbf{z}\| \leq C_0 \right\} \cup \{\mathbf{0}\}$. Since $a\mathcal{N}_2(a)$ is non-decreasing, it follows that $\mathcal{N}_2(a) \geq \mathcal{N}_2(1) \min\{a^{-1}, 1\}$, for any $a > 0$. For any $\mathbf{z} \in J_3(\mathbf{x})$, it then holds that $\mathcal{N}_2(\|\mathbf{z}\| + \|\mathbf{x}\|) \geq \mathcal{N}_2(1) \min\{1/(\|\mathbf{x}\| + C_0), 1\}$. Plugging in (6), we then have

$$\begin{aligned} \langle \Phi(\mathbf{x}), \mathbf{x} \rangle &\geq \|\mathbf{x}\|^2 \int_{J_3(\mathbf{x})} \mathcal{N}_2(\|\mathbf{x}\| + \|\mathbf{z}\|) P(\mathbf{z}) d\mathbf{z} \\ &\geq \|\mathbf{x}\|^2 \mathcal{N}_2(1) \min\{(\|\mathbf{x}\| + C_0)^{-1}, 1\} \int_{J_3(\mathbf{x})} P(\mathbf{z}) d\mathbf{z} \geq \|\mathbf{x}\|^2 \mathcal{N}_2(1) \min\{(\|\mathbf{x}\| + C_0)^{-1}, 1\} p_0. \end{aligned} \quad (7)$$

If $\|\mathbf{x}\| \leq C_0$, it follows that $\|\mathbf{x}\| + C_0 \leq 2C_0$, therefore $\min\{1/(\|\mathbf{x}\| + C_0), 1\} \geq \min\{1/(2C_0), 1\} \triangleq \kappa$. If $\|\mathbf{x}\| \geq C_0$, it follows that $\|\mathbf{x}\| + C_0 \leq 2\|\mathbf{x}\|$, therefore $\min\{1/(\|\mathbf{x}\| + C_0), 1\} \geq \min\{1/(2\|\mathbf{x}\|), 1\} \geq \min\{1/(2\|\mathbf{x}\|), \kappa\}$. Combining, we get $\langle \Phi(\mathbf{x}), \mathbf{x} \rangle \geq p_0 \mathcal{N}_2(1) \min\{\|\mathbf{x}\|/2, \kappa\|\mathbf{x}\|^2\}$. Consider the constant $\kappa = \min\{1/(2C_0), 1\}$. If $B_0 \geq 0.5$, it follows that $C_0 = 0.5$ and therefore $\kappa = 1$. On the other hand, if $B_0 < 0.5$, it follows that $C_0 = B_0$ and therefore $\kappa = \min\{1/(2B_0), 1\} = 1$, as $2B_0 < 1$. \square

We are now ready to prove Theorem 1.

Proof of Theorem 1. For ease of notation, let $Z_t \triangleq \min\{\eta_1 \|\nabla f(\mathbf{x}^{(t)})\|, \eta_2 \|\nabla f(\mathbf{x}^{(t)})\|^2\}$. Applying the L -smoothness property of f and the update rule (4), to get

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

where the second inequality follows from Lemma 3.2 and Assumption 1. Rearranging and summing up the first t terms, we get

$$\sum_{k=1}^t \alpha_k Z_k \leq \underbrace{f(\mathbf{x}^{(1)}) - f^* + \frac{L C^2}{2} \sum_{k=1}^t \alpha_k^2}_{\triangleq B_1} + \underbrace{\sum_{k=1}^t \alpha_k \langle \nabla f(\mathbf{x}^{(k)}), \mathbf{e}^{(k)} \rangle}_{\triangleq B_2}. \quad (8)$$

Denote the left-hand side of (8) by G_t , i.e., $G_t \triangleq \sum_{k=1}^t \alpha_k Z_k$ and note that B_1 is independent of the noise, i.e., is a deterministic quantity. We then have

$$\mathbb{E}[\exp(G_t)] \stackrel{(8)}{\leq} \mathbb{E}[\exp(B_1 + B_2)] = \exp(B_1) \mathbb{E}[\exp(B_2)].$$

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

$$\begin{aligned} \mathbb{E}[\exp(B_2)] &= \mathbb{E} \left[\exp \left(\sum_{k=1}^t \alpha_k \langle \nabla f(\mathbf{x}^{(k)}), \mathbf{e}^{(k)} \rangle \right) \right] \\ &= \mathbb{E} \left[\exp \left(\sum_{k=1}^{t-1} \alpha_k \langle \nabla f(\mathbf{x}^{(k)}), \mathbf{e}^{(k)} \rangle \right) \mathbb{E}_t \left[\exp(\alpha_t \langle \nabla f(\mathbf{x}^{(t)}), \mathbf{e}^{(t)} \rangle) \right] \right] \\ &\leq \mathbb{E} \left[\exp \left(\sum_{k=1}^{t-1} \alpha_k \langle \nabla f(\mathbf{x}^{(k)}), \mathbf{e}^{(k)} \rangle \right) \exp \left(4C^2 \alpha_t^2 \|\nabla f(\mathbf{x}^{(t)})\|^2 \right) \right], \end{aligned} \quad (9)$$

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

$$\begin{aligned} \|\nabla f(\mathbf{x}^{(k)})\| &\leq L \|\mathbf{x}^{(k)} - \mathbf{x}^*\| = L \|\mathbf{x}^{(k-1)} - \alpha_{k-1} \Psi^{(k-1)} - \mathbf{x}^*\| \leq L \left(\|\mathbf{x}^{(k-1)} - \mathbf{x}^*\| + \alpha_{k-1} C \right) \\ &\leq \dots \leq L \left(\|\mathbf{x}^{(1)} - \mathbf{x}^*\| + C \sum_{s=1}^{k-1} \alpha_s \right) \leq L \left(\|\mathbf{x}^{(1)} - \mathbf{x}^*\| + C A_k \right), \end{aligned} \quad (10)$$

where we recall that $\mathbf{x}^* \in \mathcal{X}$ is any stationary point of f . Combining (9) and (10), we get

$$\mathbb{E}[\exp(B_2)] \leq \exp(8C^2 L^2 D_{\mathcal{X}} \alpha_t^2 + 8C^4 L^2 \alpha_t^2 A_t^2) \mathbb{E} \left[\exp \left(\sum_{k=1}^{t-1} \alpha_k \langle \nabla f(\mathbf{x}^{(k)}), \mathbf{e}^{(k)} \rangle \right) \right],$$

where $D_{\mathcal{X}} = \inf_{\mathbf{x}^* \in \mathcal{X}} \|\mathbf{x}^{(1)} - \mathbf{x}^*\|^2$ is the distance of the initial model estimate from the set of stationary points. Repeating the same arguments recursively, we then get

$$\mathbb{E}[\exp(B_2)] \leq \exp \left(8C^2 L^2 D_{\mathcal{X}} \sum_{k=1}^t \alpha_k^2 + 8C^4 L^2 \sum_{k=1}^t \alpha_k^2 A_k^2 \right),$$

Combining everything, we get

$$\mathbb{E}[\exp(G_t)] \leq \exp\left(f(\mathbf{x}^{(1)}) - f^* + LC^2(1/2 + 8LD_{\mathcal{X}}) \sum_{k=1}^t \alpha_k^2 + 8C^4L^2 \sum_{k=1}^t \alpha_k^2 A_k^2\right).$$

Define $N_t \triangleq f(\mathbf{x}^{(1)}) - f^* + LC^2(1/2 + 8LD_{\mathcal{X}}) \sum_{k=1}^t \alpha_k^2 + 8C^4L^2 \sum_{k=1}^t \alpha_k^2 A_k^2$. Using Markov's inequality, it then follows that, for any $\epsilon > 0$

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

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

$$G_t \leq \log(1/\beta) + N_t \iff A_t^{-1}G_t \leq A_t^{-1}(\log(1/\beta) + N_t). \quad (11)$$

Note that for the step-size schedule $\alpha_t = \frac{a}{(t+1)^\delta}$ and any $\delta \in (2/3, 1)$, using lower and upper Darboux sums, we have

$$\begin{aligned} \frac{a}{1-\delta}((t+2)^{1-\delta} - 2^{1-\delta}) &\leq A_t \leq \frac{a}{1-\delta}((t+1)^{1-\delta} - 1), \\ \frac{a^2}{2\delta-1}(2^{1-2\delta} - (t+2)^{1-2\delta}) &\leq \sum_{k=1}^t \alpha_k^2 \leq \frac{a^2}{2\delta-1}(1 - (t+1)^{1-2\delta}). \end{aligned} \quad (12)$$

Plugging (12) in (11), we then get, with probability at least $1 - \beta$

$$\begin{aligned} \sum_{k=1}^t \tilde{\alpha}_k Z_k &\leq \frac{(1-\delta)(f(\mathbf{x}^{(1)}) - f^* + \log(1/\beta))}{a((t+2)^{1-\delta} - 2^{1-\delta})} \\ &\quad + \frac{a(1-\delta)LC^2(1/2 + 8LD_{\mathcal{X}})}{(2\delta-1)((t+2)^{1-\delta} - 2^{1-\delta})} + \frac{8a^3C^4L^2 \sum_{k=1}^t (k+1)^{2-4\delta}}{(1-\delta)((t+2)^{1-\delta} - 2^{1-\delta})}. \end{aligned} \quad (13)$$

To bound the last sum, we consider different step-size schedules.

1. First, consider $\alpha_t = \frac{a}{(t+1)^\delta}$, for $\delta \in (2/3, 3/4)$. Using the lower Darboux sum, we have

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

Combining with (13), we get

$$\sum_{k=1}^t \tilde{\alpha}_k Z_k \leq \frac{R_1}{(t+2)^{1-\delta} - 2^{1-\delta}} + \frac{R_2(t+1)^{3-4\delta}}{(t+2)^{1-\delta} - 2^{1-\delta}} \leq \frac{R_1}{(t+2)^{1-\delta} - 2^{1-\delta}} + \frac{R_2}{(t+2)^{3\delta-2} - 2^{3\delta-2}}, \quad (14)$$

where $R_1 \triangleq (1-\delta) \left[\frac{(f(\mathbf{x}^{(1)}) - f^* + \log(1/\beta))}{a} + \frac{aLC^2(1/2 + 8LD_{\mathcal{X}})}{(2\delta-1)} \right]$ and $R_2 \triangleq \frac{8a^3C^4L^2}{(1-\delta)(3-4\delta)}$.

2. Next, consider $\alpha_t = \frac{a}{(t+1)^\delta}$, for $\delta = 3/4$. Using the lower Darboux sum, we have

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

Combining with (13), we get

$$\sum_{k=1}^t \tilde{\alpha}_k Z_k \leq \frac{R_1 + R_3 \log(t+1)}{(t+2)^{1/4} - 2^{1/4}}, \quad (15)$$

where $R_3 \triangleq 32a^3C^4L^2$.

3. Finally, for $\alpha_t = \frac{a}{(t+1)^\delta}$, where $\delta \in (3/4, 1)$, we have

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

therefore, combining with (13), we get

$$\sum_{k=1}^t \tilde{\alpha}_k Z_k \leq \frac{R_1 + R_4}{(t+2)^{1-\delta} - 2^{1-\delta}}, \quad (16)$$

where $R_4 \triangleq \frac{8a^3 C^4 L^2}{(1-\delta)(4\delta-3)}$.

Finally, to obtain a bound on the quantity of interest $\min_{k \in [t]} \|\nabla f(\mathbf{x}^{(k)})\|^2$, we proceed as follows. Notice that the bounds in (14)-(16) can be represented in a unified manner as

$$\sum_{k=1}^t \tilde{\alpha}_k Z_k \leq M t^{-\kappa}, \quad (17)$$

for appropriately selected constants $M, \kappa > 0$.⁹ Next, define $U \triangleq \{k \in [t] : \|\nabla f(\mathbf{x}^{(k)})\| \leq \eta_1/\eta_2\}$, with $U^c \triangleq [t] \setminus U$. From (17), we then have

$$\sum_{k \in U^c} \tilde{\alpha}_k \|\nabla f(\mathbf{x}^{(k)})\| \leq M_1 t^{-\kappa} \text{ and } \sum_{k \in U} \tilde{\alpha}_k \|\nabla f(\mathbf{x}^{(k)})\|^2 \leq M_2 t^{-\kappa},$$

where $M_1 = M/\eta_1$, $M_2 = M/\eta_2$. It then readily follows that

$$\min_{k \in [t]} \|\nabla f(\mathbf{x}^{(k)})\| \leq \sum_{k \in U} \tilde{\alpha}_k \|\nabla f(\mathbf{x}^{(k)})\| + \sum_{k \in U^c} \tilde{\alpha}_k \|\nabla f(\mathbf{x}^{(k)})\| \leq \sum_{k=1}^t \tilde{\alpha}_k z_k + M_1 t^{-\kappa},$$

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

$$\min_{k \in [t]} \|\nabla f(\mathbf{x}^{(k)})\| \leq \sqrt{\sum_{k=1}^t \tilde{\alpha}_k z_k^2} + M_1 t^{-\kappa} = \sqrt{\sum_{k \in U} \tilde{\alpha}_k \|\nabla f(\mathbf{x}^{(k)})\|^2} + M_1 t^{-\kappa} \leq \sqrt{M_2 t^{-\kappa}} + M_1 t^{-\kappa}.$$

Squaring both sides and using $(a+b)^2 \leq 2a^2 + 2b^2$, gives the desired result. \square

We next prove Corollary 1.1.

Proof of Corollary 1.1. Recall the definition of the Huber loss function $H_\lambda : \mathbb{R} \mapsto [0, \infty)$, parametrized by $\lambda > 0$, e.g., Huber (1964), 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, by the definition of Huber loss, we have, for any $k \geq 1$

$$Z_k = \min\{\eta_1 \|\nabla f(\mathbf{x}^{(k)})\|, \eta_2 \|\nabla f(\mathbf{x}^{(k)})\|^2\} \geq \eta_2 H_{\eta_1/\eta_2}(\|\nabla f(\mathbf{x}^{(k)})\|). \quad (18)$$

Next, recall that Assumption 3 implies the *gradient domination property*, i.e., $\|\nabla f(\mathbf{x})\|^2 \geq 2\mu(f(\mathbf{x}) - f^*)$, for any $\mathbf{x} \in \mathbb{R}^d$, see, e.g., Nesterov (2018). Combined with the definition of strong convexity, we have $\|\nabla f(\mathbf{x})\| \geq \mu\|\mathbf{x} - \mathbf{x}^*\|$, for any $\mathbf{x} \in \mathbb{R}^d$. Combining (18) with the gradient domination property, we get

$$\sum_{k=1}^t \tilde{\alpha}_k Z_k \geq \eta_2 \sum_{k=1}^t \tilde{\alpha}_k H_{\eta_1/\eta_2}(\mu\|\mathbf{x}^{(k)} - \mathbf{x}^*\|) \geq \mu^2 \eta_2 H_{\eta_1/(\eta_2\mu)}(\|\hat{\mathbf{x}}^{(t)} - \mathbf{x}^*\|),$$

⁹Note that for $\delta = 3/4$ we might have an additional factor of $\log(t)$ in the right-hand side of (17). However, this can be easily incorporated, by allowing M to depend on t , e.g., by defining $M_t = M \log(t)$.

where $\widehat{\mathbf{x}}^{(t)} \triangleq \sum_{k=1}^t \widetilde{\alpha}_k \mathbf{x}^{(k)}$ is the weighted average of the first t iterates, the first inequality follows from (18), the gradient domination property and the fact that H is non-decreasing, while the second inequality follows from the fact that H is convex and non-decreasing, applying Jensen's inequality twice and noticing that $H_\lambda(\mu x) = \mu^2 H_{\lambda/\mu}(x)$. Using (17), it readily follows that

$$H_{\eta_1/(\eta_2\mu)}(\|\widehat{\mathbf{x}}^{(t)} - \mathbf{x}^*\|) \leq \frac{M}{\eta_2^2 \mu^2 t^\kappa}, \quad (19)$$

where M, κ depend on the step-size schedule and other problem parameters. By the definition of Huber loss and (19), if $\|\widehat{\mathbf{x}}^{(t)} - \mathbf{x}^*\| \leq \eta_1/\eta_2\mu$, we have

$$\|\widehat{\mathbf{x}}^{(t)} - \mathbf{x}^*\|^2 \leq \frac{2M}{\eta_2^2 \mu^2 t^\kappa}. \quad (20)$$

Otherwise, if $\|\widehat{\mathbf{x}}^{(t)} - \mathbf{x}^*\| > \eta_1/\eta_2\mu$, by (19), we have

$$\frac{\eta_1 \|\widehat{\mathbf{x}}^{(t)} - \mathbf{x}^*\|}{2\eta_2\mu} < \frac{\eta_1 \|\widehat{\mathbf{x}}^{(t)} - \mathbf{x}^*\|}{\eta_2\mu} - \frac{\eta_1^2}{2\eta_2^2 \mu^2} = H_{\eta_1/(\eta_2\mu)}(\|\widehat{\mathbf{x}}^{(t)} - \mathbf{x}^*\|) \leq \frac{M}{\eta_2^2 \mu^2 t^\kappa},$$

implying that

$$\|\widehat{\mathbf{x}}^{(t)} - \mathbf{x}^*\|^2 \leq \frac{4M^2}{\eta_1^2 \eta_2^2 \mu^2 t^{2\kappa}}. \quad (21)$$

Combining (20) and (21), it then follows that

$$\|\widehat{\mathbf{x}}^{(t)} - \mathbf{x}^*\|^2 \leq \max \left\{ \frac{2M}{\eta_2^2 \mu^2 t^\kappa}, \frac{4M^2}{\eta_1^2 \eta_2^2 \mu^2 t^{2\kappa}} \right\},$$

completing the proof. \square

S3.2 Proof of Theorem 2

In this section we prove Lemma 3.3 and Theorem 2. In order to prove Lemma 3.3, we first state and prove some intermediate results.

Lemma S3.4. *Let Assumptions 1-4 hold, with the step-size given by $\alpha_t = \frac{a}{(t+1)^\delta}$, for any $\delta \in (0.5, 1)$ and $a > 0$. Then, for any $t \geq 1$, we have*

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

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

$$\begin{aligned} \|\nabla f(\mathbf{x}^{(t)})\| &\leq L\|\mathbf{x}^{(t)} - \mathbf{x}^*\| = L\|\mathbf{x}^{(t-1)} - \alpha_{t-1}\boldsymbol{\Psi}^{(t-1)} - \mathbf{x}^*\| \\ &\leq L \left(\|\mathbf{x}^{(t-1)} - \mathbf{x}^*\| + \alpha_{t-1}\|\boldsymbol{\Psi}^{(t-1)}\| \right) \\ &\leq L \left(\|\mathbf{x}^{(t-1)} - \mathbf{x}^*\| + \alpha_{t-1}C \right). \end{aligned} \quad (22)$$

Unrolling the recursion in (22), we get

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

completing the proof. \square

The next result characterizes the behaviour of the nonlinearity, when it takes the form $\boldsymbol{\Psi}(\mathbf{x}) = [\mathcal{N}_1(x_1), \dots, \mathcal{N}_1(x_d)]^\top$. 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 S3.5. *Let Assumptions 1-4 hold and the nonlinearity Ψ be component-wise, i.e., of the form $\Psi(\mathbf{x}) = [\mathcal{N}_1(x_1), \dots, \mathcal{N}_1(x_d)]^\top$. Then, there exists a positive constant ξ such that, for any $t \geq 1$, there holds almost surely for each $j = 1, \dots, d$, that $|\phi_i^{(t)}| \geq \|\nabla f(\mathbf{x}^{(t)})\|_i \frac{\phi_i'(0)\xi}{2H_t}$, where H_t is defined in Lemma S3.4, 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 characterizes the behaviour of the nonlinearity, when it takes the form $\Psi(\mathbf{x}) = \mathbf{x} \mathcal{N}_2(\|\mathbf{x}\|)$.

Lemma S3.6. *Let Assumptions 1-4 hold and the nonlinearity be of the form $\Psi(\mathbf{x}) = \mathbf{x} \mathcal{N}_2(\|\mathbf{x}\|)$. Then, for any $t \geq 1$, there holds almost surely that*

$$\langle \nabla f(\mathbf{x}^{(t)}), \Phi^{(t)} \rangle \geq \frac{\|\nabla f(\mathbf{x}^{(t)})\|^2 p_0 \mathcal{N}_2(1)}{H_t + C_0},$$

where $p_0 = P(\mathbf{0})$, $C_0 = \min\{0.5, B_0\}$ and H_t is defined in Lemma S3.4.

Proof. We start from (7), which tells us that, for any $t \geq 1$, almost surely

$$\langle \Phi^{(t)}, \nabla f(\mathbf{x}^{(t)}) \rangle \geq \|\nabla f(\mathbf{x}^{(t)})\|^2 p_0 \mathcal{N}_2(1) \min \left\{ \frac{1}{\|\nabla f(\mathbf{x}^{(t)})\| + C_0}, 1 \right\}.$$

Combining with Lemma S3.4 and the fact that $H_t \geq 1$, we get almost surely

$$\langle \Phi^{(t)}, \nabla f(\mathbf{x}^{(t)}) \rangle \geq \frac{\|\nabla f(\mathbf{x}^{(t)})\|^2 p_0 \mathcal{N}_2(1)}{H_t + C_0},$$

which completes the proof. \square

We are now ready to prove Lemma 3.3.

Proof of Lemma 3.3. First, consider the case when the nonlinearity is of the form $\Psi(\mathbf{x}) = [\mathcal{N}_1(x_1), \dots, \mathcal{N}_1(x_d)]^\top$. We then have

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

where $\gamma = \frac{(1-\delta)\phi'(0)\xi}{2L(\|\mathbf{x}^{(1)} - \mathbf{x}^*\| + aC)}$, (a) follows from the oddity of \mathcal{N}_1 , (b) follows from Lemma S3.5, (c) follows from $\phi'(0) = \min_{i=1, \dots, d} \phi_i'(0)$. On the other hand, if the nonlinearity is of the form $\Psi(\mathbf{x}) = \mathbf{x} \mathcal{N}_2(\|\mathbf{x}\|)$, we get

$$\langle \Phi^{(t)}, \nabla f(\mathbf{x}^{(t)}) \rangle \geq \frac{p_0 \mathcal{N}_2(1) \|\nabla f(\mathbf{x}^{(t)})\|^2}{H_t + C_0} \geq \gamma(t+1)^{\delta-1} \|\nabla f(\mathbf{x}^{(t)})\|^2,$$

where $\gamma = \frac{(1-\delta)p_0 \mathcal{N}_2(1)}{L(\|\mathbf{x}^{(1)} - \mathbf{x}^*\| + aC) + C_0}$, the first inequality follows from Lemma S3.6, while the second follows from the definition of H_t and the fact that $H_t + C_0 \leq (L(\|\mathbf{x}^{(1)} - \mathbf{x}^*\| + aC) + C_0) \frac{(t+1)^{1-\delta}}{1-\delta}$. This completes the proof. \square

We next prove Theorem 2.

Proof of Theorem 2. Using L -smoothness of f , the update rule (4) and Lemma 3.3, we have

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

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

$$F^{(t+1)} \leq \left(1 - \frac{2\mu a\gamma}{t+1}\right) F^{(t)} + \frac{a\langle \nabla f(\mathbf{x}^{(t)}), \mathbf{e}^{(t)} \rangle}{(t+1)^\delta} + \frac{a^2 LC^2}{2(t+1)^{2\delta}}. \quad (23)$$

Let $\zeta = \min\{2\delta - 1, a\gamma\mu/2\}$. Defining $Y^{(t)} \triangleq t^\zeta F^{(t)} = t^\zeta(f(\mathbf{x}^{(t)}) - f^*)$, from (23) we get

$$Y^{(t+1)} \leq a_t Y^{(t)} + b_t \langle \nabla f(\mathbf{x}^{(t)}), \mathbf{e}^{(t)} \rangle + c_t V, \quad (24)$$

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

$$\begin{aligned} M_{t+1|t}(\nu) &\stackrel{(a)}{\leq} \mathbb{E} \left[\exp \left(\nu (a_t Y^{(t)} + b_t \langle \mathbf{e}^{(t)}, \nabla f(\mathbf{x}^{(t)}) \rangle + c_t V) \right) \mid \mathcal{F}_t \right] \\ &\stackrel{(b)}{\leq} \exp(\nu a_t Y^{(t)} + \nu c_t V) \mathbb{E} \left[\exp(\nu b_t \langle \mathbf{e}^{(t)}, \nabla f(\mathbf{x}^{(t)}) \rangle) \mid \mathcal{F}_t \right] \\ &\stackrel{(c)}{\leq} \exp \left(\nu a_t Y^{(t)} + \nu c_t V + \nu^2 b_t^2 N \|\nabla f(\mathbf{x}^{(t)})\|^2 \right) \\ &\stackrel{(d)}{\leq} \exp \left(\nu a_t Y^{(t)} + \nu c_t V + 2\nu^2 b_t^2 L N Y^{(t)} \right), \end{aligned} \quad (25)$$

where (a) follows from (24), (b) follows from the fact that $Y^{(t)}$ is \mathcal{F}_t measurable, (c) follows from Lemma 3.1, in (d) we use $\|\nabla f(\mathbf{x})\|^2 \leq 2L(f(\mathbf{x}) - f^*)$ and define $b'_t = a \frac{t^{-\frac{\zeta}{2}}}{(t+1)^{\delta-\zeta}}$, so that $b_t = t^{\frac{\zeta}{2}} b'_t$. For the choice $0 \leq \nu \leq B$, for some $B > 0$ (to be specified later), we get

$$M_{t+1|t}(\nu) \leq \exp \left(\nu (a_t + 2b_t'^2 L N B) Y^{(t)} \right) \exp(\nu c_t V).$$

Taking the full expectation, we get

$$M_{t+1}(\nu) \leq M_t((a_t + 2b_t'^2 L N B)\nu) \exp(\nu c_t V). \quad (26)$$

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 $t \geq 1$. We proceed by induction. For $t = 1$, we have

$$M_1(\nu) = \exp(\nu Y^{(1)}) = \exp \left(\nu (f(\mathbf{x}^{(1)}) - f^*) \right),$$

where we simply used the definition of $Y^{(t)}$ and the fact that it is deterministic for $t = 1$. Choosing $B \leq (f(\mathbf{x}^{(1)}) - f^*)^{-1}$ ensures that $M_1(\nu) \leq e^{\frac{\nu}{B}}$. Next, assume that for some $t \geq 2$ it holds that $M_t(\nu) \leq e^{\frac{\nu}{B}}$. We then have

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

where we use (26) in the first and the induction hypothesis in the second inequality. For our claim to hold, it suffices to show $a_t + 2b_t'^2 L N B + c_t V B \leq 1$. Plugging in the values of a_t , b'_t and c_t , we have

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

Noticing that $2\delta - \zeta \geq 1$ and setting $B = \min \left\{ \frac{1}{(f(\mathbf{x}^{(1)}) - f^*)}, \frac{\mu\gamma}{2aLN+aV} \right\}$, gives

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

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 $M_t(\nu) \leq e^{\frac{\mu}{B}}$, for any $t \geq 1$ and any $0 \leq \nu \leq B$. By Markov's inequality, it readily follows that

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

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

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

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

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

completing the proof. \square

S3.3 Proof of Theorem 3

Proof of Theorem 3. Consider the “denoised” nonlinearity $\Phi^{(t)} \triangleq \mathbb{E}[\Psi(\nabla f(\mathbf{x}^{(t)}) + \mathbf{z}^{(t)}) \mid \mathcal{F}_t]$. From Assumption 5 and the linearity of expectation, it follows that $\Phi^{(t)}$ can be expressed as

$$\Phi^{(t)} = \lambda \Phi_1^{(t)} + (1 - \lambda) \Phi_2^{(t)}, \quad (27)$$

where $\Phi_i^{(t)} = \mathbb{E}_{\mathbf{z}^{(t)} \sim P_i}[\Psi(\nabla f(\mathbf{x}^{(t)}) + \mathbf{z}^{(t)}) \mid \mathcal{F}_t]$, $i \in [2]$ are the “denoised” nonlinearities with respect to each of the noise components. Defining the effective noise as $\mathbf{e}^{(t)} = \Phi^{(t)} - \Psi^{(t)}$, it can be readily seen that Lemma 3.1 still applies. Similarly, it can be seen that Lemma 3.2 holds for Φ_1 , as this represents the effective search direction with respect to the symmetric noise component. Apply the smoothness inequality and the update rule (4), to get

$$\begin{aligned} f(\mathbf{x}^{(t+1)}) &\leq f(\mathbf{x}^{(t)}) - \alpha_t \langle \nabla f(\mathbf{x}^{(t)}), \Phi^{(t)} - \mathbf{e}^{(t)} \rangle + \frac{\alpha_t^2 L}{2} \|\Psi^{(t)}\|^2 \\ &\leq f(\mathbf{x}^{(t)}) - \alpha_t (1 - \lambda) \langle \nabla f(\mathbf{x}^{(t)}), \Phi_1^{(t)} \rangle - \alpha_t \lambda \langle \nabla f(\mathbf{x}^{(t)}), \Phi_2^{(t)} \rangle + \alpha_t \langle \nabla f(\mathbf{x}^{(t)}), \mathbf{e}^{(t)} \rangle + \frac{\alpha_t^2 L C^2}{2} \\ &\leq f(\mathbf{x}^{(t)}) - \alpha_t (1 - \lambda) Z_t - \alpha_t \lambda \langle \nabla f(\mathbf{x}^{(t)}), \Phi_2^{(t)} \rangle + \alpha_t \langle \nabla f(\mathbf{x}^{(t)}), \mathbf{e}^{(t)} \rangle + \frac{\alpha_t^2 L C^2}{2}, \end{aligned} \quad (28)$$

where the first inequality follows from (27) and the boundedness of the nonlinearity, while the second inequality follows from Lemma 3.1, recalling that $Z_t \triangleq \min\{\eta_1 \|\nabla f(\mathbf{x}^{(t)})\|, \eta_2 \|\nabla f(\mathbf{x}^{(t)})\|^2\}$. To bound the inner product of the gradient and the non-symmetric component, we proceed as follows. For any $\mathbf{x} \in \mathbb{R}^d$, we have

$$\langle \mathbf{x}, \Phi_2(\mathbf{x}) \rangle \leq \|\mathbf{x}\| \|\Phi_2(\mathbf{x})\| \leq C \|\mathbf{x}\| \leq \begin{cases} C \|\mathbf{x}\|, & \|\mathbf{x}\| \geq B \\ C B, & \|\mathbf{x}\| < B \end{cases}, \quad (29)$$

where $B > 0$ is an arbitrary constant, to be specified later. Note that (29) is equivalent to

$$\langle \mathbf{x}, \Phi_2(\mathbf{x}) \rangle \leq C \max\{\|\mathbf{x}\|, B\}. \quad (30)$$

Plugging (30) in (28), we get

$$f(\mathbf{x}^{(t+1)}) \leq f(\mathbf{x}^{(t)}) - \alpha_t (1 - \lambda) Z_t + \alpha_t \lambda C \max\{\|\nabla f(\mathbf{x}^{(t)})\|, B\} + \alpha_t \langle \nabla f(\mathbf{x}^{(t)}), \mathbf{e}^{(t)} \rangle + \frac{\alpha_t^2 L C^2}{2}$$

Setting $B = \eta_1/\eta_2$, it can be readily seen that

$$(1 - \lambda) Z_t - \lambda C \max\{\|\nabla f(\mathbf{x}^{(t)})\|, \eta_1/\eta_2\} = \min\{(\eta_1(1 - \lambda) - \lambda C) \|\nabla f(\mathbf{x}^{(t)})\|, \eta_2(1 - \lambda) \|\nabla f(\mathbf{x}^{(t)})\|^2 - \lambda C \eta_1/\eta_2\}.$$

From the condition $\lambda < \frac{\eta_1}{\eta_1 + C}$, it follows that $\eta_1(1 - \lambda) - \lambda C > 0$. Next, define $\tilde{Z}_t \triangleq \min\{(\eta_1(1 - \lambda) - \lambda C) \|\nabla f(\mathbf{x}^{(t)})\|, \eta_2(1 - \lambda) \|\nabla f(\mathbf{x}^{(t)})\|^2 - \lambda C \eta_1/\eta_2\}$. Rearranging and summing up the first t terms, we get

$$\sum_{k=1}^t \alpha_k \tilde{Z}_k \leq f(\mathbf{x}^{(1)}) - f^* + \frac{L C^2}{2} \sum_{k=1}^t \alpha_k^2 + \sum_{k=1}^t \alpha_k \langle \nabla f(\mathbf{x}^{(k)}), \mathbf{e}^{(k)} \rangle.$$

Repeating the same steps as in the proof of Theorem 1, we get

$$\begin{aligned} \sum_{k=1}^t \tilde{\alpha}_k \tilde{Z}_k &\leq \frac{(1-\delta)(f(\mathbf{x}^{(1)}) - f^* + \log(1/\beta))}{a((t+2)^{1-\delta} - 2^{1-\delta})} \\ &\quad + \frac{a(1-\delta)LC^2(1/2 + 8LD_{\mathcal{X}})}{(2\delta-1)((t+2)^{1-\delta} - 2^{1-\delta})} + \frac{8a^3C^4L^2 \sum_{k=1}^t (k+1)^{2-4\delta}}{(1-\delta)((t+2)^{1-\delta} - 2^{1-\delta})}. \end{aligned} \quad (31)$$

Considering the different step-size schedules, we can similarly obtain a unified representation of the form

$$\sum_{k=1}^t \tilde{\alpha}_k \tilde{Z}_k \leq Mt^{-\kappa}, \quad (32)$$

for appropriately selected constants $M, \kappa > 0$. Using $U \triangleq \{k \in [t] : \|\nabla f(\mathbf{x}^{(k)})\| \leq \eta_1/\eta_2\}$, $U^c \triangleq [t] \setminus U$ and (32), we get

$$\eta_2(1-\lambda) \sum_{k \in U} \tilde{\alpha}_k \|\nabla f(\mathbf{x}^{(k)})\|^2 \leq Mt^{-\kappa} + \lambda C \eta_1/\eta_2 \text{ and } (\eta_1(1-\lambda) - \lambda C) \sum_{k \in U^c} \tilde{\alpha}_k \|\nabla f(\mathbf{x}^{(k)})\| \leq Mt^{-\kappa}.$$

It then readily follows that

$$\min_{k \in [t]} \|\nabla f(\mathbf{x}^{(k)})\| \leq \sum_{k \in U} \tilde{\alpha}_k \|\nabla f(\mathbf{x}^{(k)})\| + \sum_{k \in U^c} \tilde{\alpha}_k \|\nabla f(\mathbf{x}^{(k)})\| \leq \sum_{k=0}^{t-1} \tilde{\alpha}_k z_k + M_2 t^{-\kappa},$$

where $M_2 = M/(\eta_1(1-\lambda) - \lambda C)$, while $z_k = \|\nabla f(\mathbf{x}^{(k)})\|$, for $k \in U$, and $z_k = 0$, for $k \in U^c$. Using Jensen's inequality, we get

$$\min_{k \in [t]} \|\nabla f(\mathbf{x}^{(k)})\| \leq \sqrt{\sum_{k=1}^t \tilde{\alpha}_k z_k^2} + M_2 t^{-\kappa} = \sqrt{\sum_{k \in U} \tilde{\alpha}_k \|\nabla f(\mathbf{x}^{(k)})\|^2} + M_2 t^{-\kappa} \leq \sqrt{M_1 t^{-\kappa} + \frac{\lambda C \eta_1}{\eta_2^2(1-\lambda)}} + M_2 t^{-\kappa},$$

where $M_1 = \frac{M}{\eta_2(1-\lambda)}$. Squaring both sides and using $(a+b)^2 \leq 2a^2 + 2b^2$, gives the desired result. \square

S4 RATE ζ

Recalling Assumption 1 and the definition of C , it readily follows that $\gamma(a) = \frac{(1-\delta)\phi'(0)\xi}{2L(\|\mathbf{x}^{(1)} - \mathbf{x}^*\| + a\sqrt{d}C_1)}$ for nonlinearities of the form $\Psi(\mathbf{x}) = [\mathcal{N}_1(x_1), \dots, \mathcal{N}_1(x_d)]^\top$ (i.e., component-wise), while $\gamma(a) = \frac{(1-\delta)p_0\mathcal{N}_2(1)}{L(\|\mathbf{x}^{(1)} - \mathbf{x}^*\| + aC_2) + C_0}$, for nonlinearities of the form $\Psi(\mathbf{x}) = \mathbf{x}\mathcal{N}_2(\|\mathbf{x}\|)$ (i.e., joint). Combined with Theorem 2, it follows that the rate ζ is given by

$$\begin{aligned} \zeta_{joint} &= \min \left\{ 2\delta - 1, \frac{a\mu(1-\delta)p_0\mathcal{N}_2(1)}{2L(\|\mathbf{x}^{(1)} - \mathbf{x}^*\| + aC_2) + 2C_0} \right\}, \\ \zeta_{comp} &= \min \left\{ 2\delta - 1, \frac{a\mu\phi'(0)\xi(1-\delta)}{4L(\|\mathbf{x}^{(1)} - \mathbf{x}^*\| + aC_1\sqrt{d})} \right\}. \end{aligned}$$

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

- *Initialization* - starting farther from the minima results in smaller ζ (i.e., larger $\|\mathbf{x}^{(1)} - \mathbf{x}^*\|$). 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 or C_2 , and the value $\phi'(0)$ or $\mathcal{N}_2(1)$.
- *Problem dimension* - for component-wise nonlinearities through \sqrt{d} .
- *Noise* - in the form of $\phi'(0)$, ξ for component-wise and p_0 , $C_0 = \min\{0.5, B_0\}$ for joint ones.
- *Step-size* - both terms in the definition of ζ depend on the step-size parameter $\delta \in (0, 1)$.

S5 DERIVATIONS FOR EXAMPLES 6-8

Recall that the size of the neighborhood and condition on λ in Theorem 3 are given by $\frac{\eta_1 \lambda C}{\eta_2^2(1-\lambda)}$ and $\lambda < \frac{\eta_1}{C+\eta_1}$, where C is the bound on the nonlinearity, while η_1, η_2 are the constants from Lemma 3.2. From the full statement of Lemma 3.2 in the Supplement (i.e., Lemma S3.3), we know that $\eta_1 = \phi'(0)\xi/2\sqrt{d}$, $\eta_2 = \phi'(0)/2d$ for copmonent-wise and $\eta_1 = p_0\mathcal{N}_2(1)/2$, $\eta_2 = p_0\mathcal{N}_2(1)$ for joint nonlinearities. From the definition of PDF in Example 1, it follows that $p_0 = P(\mathbf{0}) = \left[\frac{\alpha-1}{2}\right]^d$. We now consider specific nonlinearities.

1. For sign, we have $C = \sqrt{d}$ and it can be shown that $\phi'(0) \approx \alpha - 1$, $\xi \approx \frac{1}{\alpha}$, see Jakovetić et al. (2023).
2. For component-wise clipping with parameter $m > 1$, we have $C = m\sqrt{d}$ and it can be shown that $\phi'(0) \approx 1 - (m+1)^{-\alpha}$, $\xi \approx m - 1$, see Jakovetić et al. (2023).
3. For joint clipping with parameter $M > 0$, we have $C = M$ and $\mathcal{N}_1(1) = \min\{1, M\}$.

Plugging in the said values completes the derivations.

S6 ANALYTICAL EXAMPLE

In this section we specialize the rates from Theorem 1 for specific choices of nonlinearity and noise, showing analytically that our theory predicts clipping is not always the optimal choice of nonlinearity and confirms the prior findings of Zhang et al. (2020), namely that for some noise instances, component-wise clipping shows better dimension dependence than joint clipping.

To that end, we consider the noise with PDF from Example 1, for some $\alpha > 2$ and choice of step-size with $\delta = 3/4$. We consider component-wise and joint clipping, with thresholds $m > 1$ and $M > 0$, respectively. As shown in the derivations from the previous section, in this case, we have $C_{cc} = m\sqrt{d}$, $\eta_{1,cc} = \frac{[1-(m+1)^{-\alpha}](m-1)}{2\sqrt{d}}$, $\eta_{2,cc} = \frac{1-(m+1)^{-\alpha}}{2d}$ for component-wise and $C_{jc} = M$, $\eta_{1,jc} = \left[\frac{\alpha-1}{2}\right]^d \min\{1/2, M/2\}$, $\eta_{2,jc} = \left[\frac{\alpha-1}{2}\right]^d \min\{1, M\}$ for joint clipping. For simplicity, we ignore the higher-order term in the bound of Theorem 1 and focus on the first, dominating term, which is ok to do, as the dependence on problem parameters and η_1, η_2 in both terms is almost identical. Similarly, we will only focus on the resulting problem related constants that figure in the leading term, ignoring the rate and global constants. To that end, we have the following problem related constants figuring in the leading terms

$$\begin{aligned} \text{Component clipping: } & \frac{d(f(\mathbf{x}^{(1)}) - f^* + \log(1/\beta)) + a^2 d^2 m^2 L(1 + LD_{\mathcal{X}}) + a^4 d^3 m^4 L^2}{a[1 - (m+1)^{-\alpha}]}, \\ \text{Joint clipping: } & \frac{(f(\mathbf{x}^{(1)}) - f^* + \log(1/\beta)) + a^2 M^2 L(1 + LD_{\mathcal{X}}) + a^4 M^4 L^2}{a[(\alpha-1)/2]^d \min\{1, M\}}. \end{aligned}$$

Note that the leading term for component clip shows a polynomial dependence on problem dimension, of order d^3 , while the leading term for the joint clip has an exponential dependence on d , via $[(\alpha-1)/2]^{-d}$. As α is an intrinsic property of the noise, whenever $\alpha \in (2, 3)$, (i.e., variance is unbounded and noise is heavy-tailed), we have $[(\alpha-1)/2]^{-d} \rightarrow \infty$, as $d \rightarrow \infty$, at an exponential rate, showing a much worse dependence on problem dimension than component clip, providing a theoretical confirmation of our numerical results (recall that we use $\alpha = 2.05$ in our simulations) and underlining the benefits of component clipping over joint one for certain noises and certain regimes, as noted in Zhang et al. (2020). The polynomial dependence of component clip on dimension d can be seen as a byproduct of our unified black-box analysis, wherein we provide a general bound C , which results in a factor \sqrt{d} when specialized to component-wise nonlinearities. This polynomial dependence is unavoidable, as, even by tuning the step-size parameter a and clipping threshold $m > 1$, we can at best remove the direct dependence on d in the numerator, while resulting in the denominator of the form $[1 - (m/d^\kappa + 1)^{-\alpha}]$, for some $\kappa > 0$, which still explodes as $d \rightarrow \infty$, again at a polynomial rate. Similarly, the exponential explosion of the bound in the joint clipping case and heavy-tailed noise (i.e., $\alpha \in (2, 3)$) is unavoidable, even under careful tuning of a and M . Therefore, our bounds confirm the observations from Zhang et al. (2020), that for some noise instances, component clipping shows better dimension dependence than the joint one. Finally, we note that the same dependence on problem dimension can be shown to hold for sign and normalized gradient, further underlining the benefit of component-wise nonlinearities for some noise instances.

S7 ADDITIONAL EXPERIMENTS

In this section we present additional numerical experiments. The first set of experiments demonstrates the noise symmetry phenomena on a deep learning model with real data. The second set of experiments provides results for the experiment from the main body, for a wider range of step-sizes and tolerance thresholds.

Noise Symmetry - Setup. We train a Convolutional Neural Network (CNN) LeCun et al. (2015) on the MNIST dataset LeCun et al. (1998) using PyTorch¹⁰. The CNN we use consists of two convolutional layers, followed by 2×2 max pooling with a stride of 2, and two fully connected layers. All layers use ReLU activations. The convolutional layers have 32 and 64 filters, with 3×3 kernels, respectively. The fully connected layers are of size 9216×168 and 168×10 , respectively. We apply dropout, with rates 0.25 and 0.5, respectively, applied after the max pooling layers and the first fully connected layer. We use a batch size of 64, set the learning rate to 1 and decrease it by a factor of 0.7 every epoch. For the optimizer, we use Adadelta Zeiler (2012) with ℓ_2 gradient clipping threshold $M = 1$. The experiments are done on MacOS 15.0 with M1 Pro processor using PyTorch 2.2.2 MPS backend.

Noise Symmetry - Visualization. Similar to the visualization method used in Chen et al. (2020), we evaluate the symmetry of gradient distribution by projecting all per-sample gradients into a 2-D space using random Gaussian matrices. For any symmetric distribution, its 2-D projection remains symmetric under any projection matrix. Conversely, if the projected gradient distribution is symmetric for every projection matrix, the original gradient distribution is also symmetric. In Figure 2, we present a 2-D plot of the random projections of all per-sample gradients after training for several epochs, with epoch 0 showing the gradient distribution at the initialization. We can observe that all the random projections exhibit a high degree of symmetry over the duration of the entire training process.

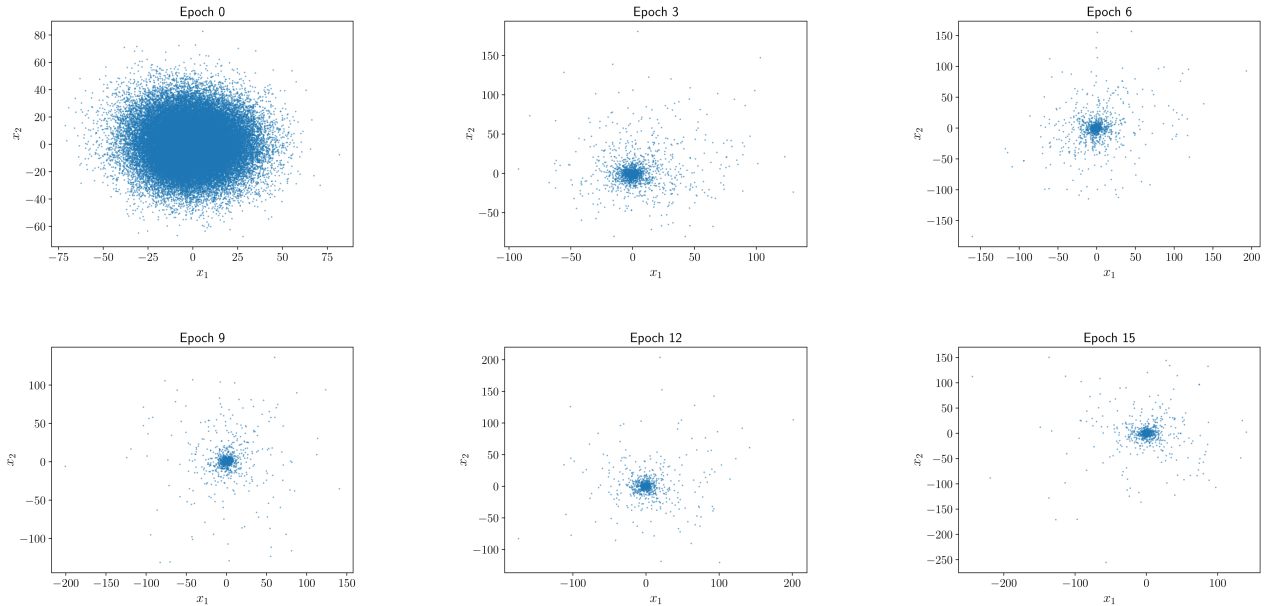


Figure 2: Random projections of per-sample gradients across epochs.

In Figure 3, we independently sample 6 Gaussian random projection matrices, and for each realization we plot the per-sample gradient projections, after training for 15 epochs. We can see that the noise projection is again highly symmetric for most random projections.

Comparisons of Nonlinear SGD Methods. We use the same MNIST dataset and CNN model as described above to test the performance of SGD with different nonlinearities under injected heavy-tailed noise. In particular, when computing mini-batch stochastic gradients, we inject random noise following a Levy stable distribution,

¹⁰<https://github.com/pytorch/examples/tree/main/mnist>

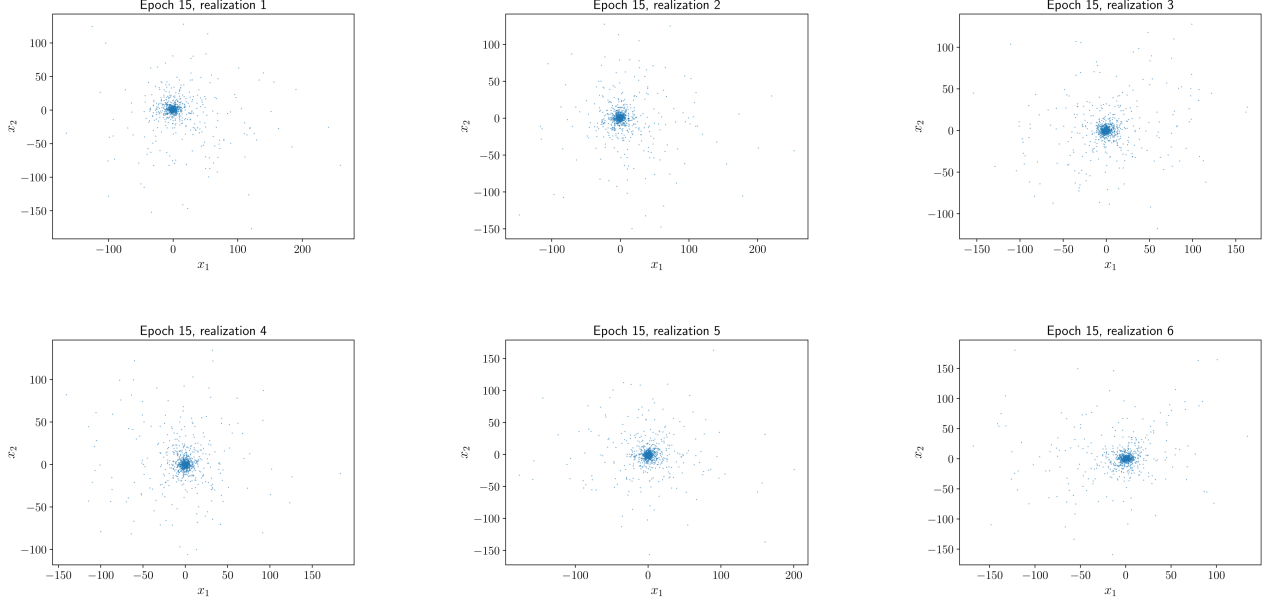


Figure 3: The distribution of gradient projections after training for 15 epochs, using 6 different projection matrices.

with the stability parameter 1.5, the skewness 1, location parameter 0, and scale 1. Note that this is a non-symmetric heavy-tailed distribution. We compare the test accuracies and test losses of baseline SGD method, SGD with component-wise and joint clipping, as well as normalized SGD. All algorithms use a varying step-size schedule $\alpha_t = \frac{a}{(t+1)^{3/4}}$, where a is a hyper-parameter chosen from $\{0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1.0\}$. For component-wise and joint clipped SGD, we pick the best clipping threshold from the set $\{0.1, 0.5, 1.0\}$. For the best hyper-parameter combination for each algorithm, we run the algorithm for 5 independent runs and plot the mean value with error bars. The results are presented in Figure 4, where it can be seen that all nonlinear SGD methods (fine-tuned) perform well, while the performance of vanilla SGD is significantly affected by the presence of heavy-tailed noise.

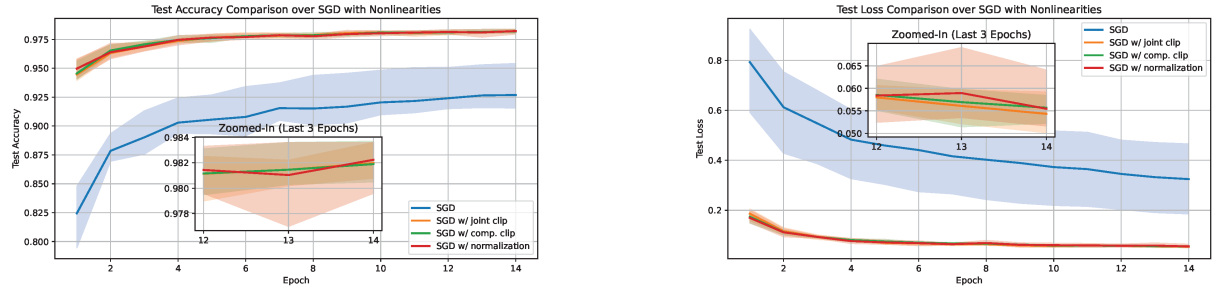


Figure 4: Comparisons of test accuracies and losses of SGD with different nonlinearities under Levy stable gradient noise.

Additional Experiments. Here, we present the results for the same setup as used in Section 4 in the main body, for a wider range of step-sizes and tail probability thresholds. Figure 5 provides the MSE behaviour of sign, joint and component-wise clipping for step-sizes $\alpha_t = \frac{1}{(t+1)^\delta}$, with $\delta \in \{17/24, 3/4, 7/8\}$, while Figure 6 presents the tail probability for all three methods, with step-size $\delta = 3/4$ and using thresholds $\varepsilon \in \{0.05, 0.1, 0.5, 5\}$. We can see that the results from Section 4 are consistent for different ranges of step-sizes, confirming that joint clipping is not always the optimal choice of nonlinearity. Moreover, we can see that all three methods achieve exponential tail decay, with joint clipping requiring a larger threshold, as it converges slower than the other two nonlinear methods, reaching a lower accuracy in the allocated number of iterations.

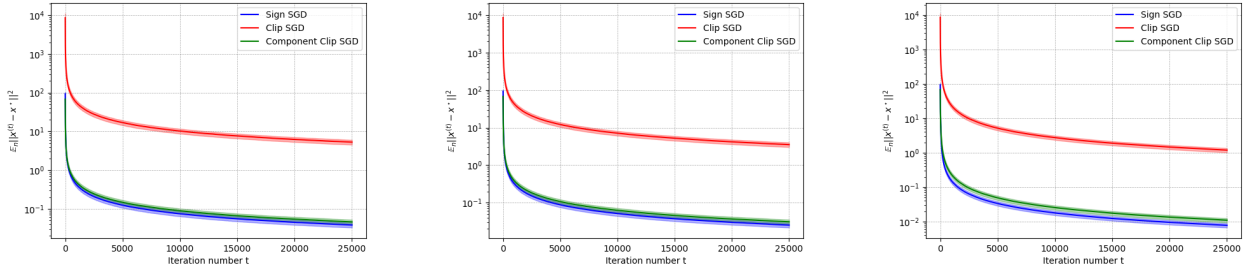


Figure 5: MSE performance of nonlinear SGD methods, using step-size policy $\alpha_t = 1/(t+1)^\delta$, for different values of $\delta \in (2/3, 1)$. Left to right: we choose the values $\delta \in \{17/24, 3/4, 7/8\}$, respectively. We can see that both component-wise nonlinearities converge faster in the MSE sense, independent of the step-size choice.

S8 ON THE NOISE ASSUMPTIONS

In this section we provide detailed discussions on the noise assumption used in our paper. In particular, we provide a detailed comparison with the bounded p -th moment assumption and discuss relaxations of the independent, identically distributed condition.

Comparison with Assumption (BM). As discussed in Remark 7, while the noise assumption in our work and in works assuming (BM) are different, it is important to note that neither set of assumptions is uniformly weaker and both come with some advantages and disadvantages, as we detail next. To begin with, both set of assumptions are concerned with heavy-tailed noises, with ours requiring no moment bounds, while assumption (BM) requires bounded moments of order $p \in (1, 2]$, uniformly for all $x \in \mathbb{R}^d$. This is a significant relaxation on our end and allows for considering extremely heavy-tailed noises, such as Cauchy noise, for which even the mean does not exist! On the other hand, in order to guarantee exact convergence, our work requires noise with symmetric PDF, positive around zero, whereas no such requirements are needed for (BM). However, we relax the symmetry requirement, allowing for mixtures of symmetric and non-symmetric components, resulting in potentially biased noise, for which convergence to a neighbourhood of stationarity is shown (which is in general the best possible guarantee for biased SGD without corrective mechanisms like momentum or error-feedback). Contrary to this, (BM) always requires the noise to be unbiased. Finally, while we require the noise vectors to be independent and identically distributed, which is not the case with (BM), this condition can be relaxed to include noises which are not identically distributed and depend on the current state (which we detail in the next paragraph), making the two sets of assumptions comparable on this point. Therefore, we can clearly see that both sets of noise assumptions come with advantages and disadvantages, with neither uniformly stronger than the other.

On the Independent, Identically Distributed Condition. As discussed in Remark 5, the independent, identically distributed condition can be significantly relaxed. First, the noise vectors need not be identically distributed. Instead, it suffices that in each iteration $t = 1, 2, \dots$, the noise vector $\mathbf{z}^{(t)}$ has a probability density function (PDF) P_t , where in addition to being symmetric, we make the following requirement: there exists a $B_0 > 0$, such that $\inf_{t=1,2,\dots} P_t(\mathbf{z}) > 0$, for each $\|\mathbf{z}\| \leq B_0$. This condition can be seen as a uniform positivity in a neighbourhood of zero requirement, which is a mild condition on the behaviour of the sequence of PDFs and is satisfied, e.g., if the PDFs are drawn from a finite family \mathcal{P} of symmetric PDFs, positive in a neighbourhood of zero (assuming a finite family is natural, as for our finite-time bounds, a weaker condition actually suffices, namely $\min_{t \in [T]} P_t(\mathbf{z}) > 0$, for all $\|\mathbf{z}\| \leq B_0$, which exactly corresponds to considering a finite family \mathcal{P} of symmetric distributions, positive in a neighbourhood of zero, with $|\mathcal{P}| = T$, for any finite time horizon T). Therefore, defining $\phi'(0) = \min_{i \in [d]} \inf_{t=1,2,\dots} \phi'_{i,t}(0) > 0$, where $\phi_{i,t}(x_i) = \mathbb{E}_{z_i \sim P_t} [\mathcal{N}_1(x_i + z_i)]$ is the marginal expectation of the i -th noise component at time t , and $p_0 = \inf_{t=1,2,\dots} P_t(\mathbf{0}) > 0$, our current analysis applies and our proofs readily go through. Second, for joint nonlinearities, the noise vectors need not be independent. Instead, in each iteration t , the noise vector $\mathbf{z}^{(t)}$ is allowed to depend on the history through current state $\mathbf{x}^{(t)}$. This is facilitated by assuming that, for each fixed $\mathbf{x} \in \mathbb{R}^d$, the noise vector $\mathbf{z} = \mathbf{z}(\mathbf{x})$ has a PDF $P_{\mathbf{x}}(\mathbf{z}) = P(\mathbf{z} | X = \mathbf{x})$, which is symmetric for each fixed $\mathbf{x} \in \mathbb{R}^d$, and that there exists a $B_0 > 0$, such that $\inf_{\mathbf{x} \in \mathbb{R}^d} P_{\mathbf{x}}(\mathbf{z}) > 0$, for all $\|\mathbf{z}\| \leq B_0$. The uniform positivity around zero for the conditional PDF $P_{\mathbf{x}}(\mathbf{z})$ is again a generalization of the positivity

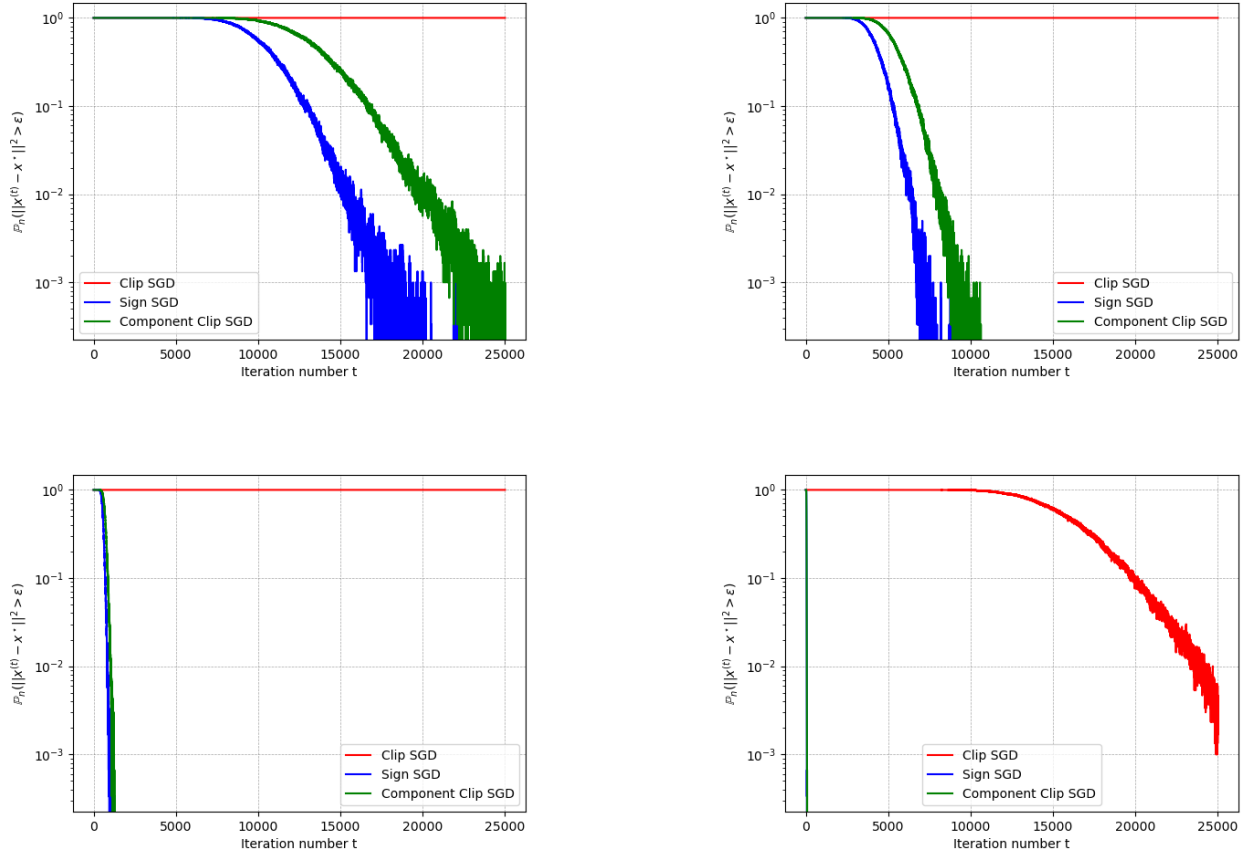


Figure 6: High-probability performance of nonlinear SGD methods, using step-size policy $\alpha_t = 1/(t+1)^\delta$, with $\delta = 3/4$. We use the thresholds $\epsilon \in \{0.05, 0.1, 0.5, 5\}$ to compute the tail probability, left to right and top to bottom. We can see that all the methods exhibit exponential tail decay, with joint clipping needing the largest threshold to achieve exponential decay, due to slower convergence.

around zero condition, and similar to the previous discussion, can be relaxed to a path-wise condition for our finite-time high-probability guarantees, namely, $\inf_{t \in [T]} P_{\mathbf{x}^{(t)}}(\mathbf{z}) > 0$, for all $\|\mathbf{z}\| \leq B_0$ and each fixed T . It can be shown, using the same steps of our proof, while replacing $P(\mathbf{z})$ with $P_{\mathbf{x}}(\mathbf{z})$, that Lemma 3.2 holds for joint nonlinearities (recall the proof of Lemma S3.2, with $p_0 = P(\mathbf{0})$ now replaced by $p_0 = \inf_{\mathbf{x} \in \mathbb{R}^d} P_{\mathbf{x}}(\mathbf{0})$). Similarly, the proofs of Theorems 1-3, which use the conditional moment-generating function, conditioned on the entire history of the algorithm, readily go through, requiring no further modification.

S9 ON THE METRIC

As discussed in Remark 12, it is possible to provide high-probability convergence guarantees of the same order as in Theorem 1, for the metric $\frac{1}{t} \sum_{k=1}^t \min\{\|\nabla f(\mathbf{x}^{(k)})\|, \|\nabla f(\mathbf{x}^{(k)})\|^2\}$. To do so, we proceed as follows. Recall equation (11) in the proof of Theorem 1 in Section S3, namely that, for any $\beta \in (0, 1)$, with probability at least $1 - \beta$, we have $G_t \leq \log(1/\beta) + N_t$, where $G_t \triangleq \sum_{k=1}^t \alpha_k \min\{\eta_1 \|\nabla f(\mathbf{x}^{(k)})\|, \eta_2 \|\nabla f(\mathbf{x}^{(k)})\|^2\}$ and $N_t \triangleq f(\mathbf{x}^{(1)}) - f^* + LC^2(1/2 + 8LD_{\mathcal{X}}) \sum_{k=1}^t \alpha_k^2 + 8C^4L^2 \sum_{k=1}^t \alpha_k^2 A_k^2$. Instead of dividing both sides of the inequality by $A_t = \sum_{k=1}^t \alpha_k$, as was originally done in (11), we divide both sides of the inequality by t and notice that the sequence of step-sizes is decreasing, to get, with probability at least $1 - \beta$

$$\frac{\alpha_t}{t} \sum_{k=1}^t \min\{\eta_1 \|\nabla f(\mathbf{x}^{(k)})\|, \eta_2 \|\nabla f(\mathbf{x}^{(k)})\|^2\} \leq \frac{\log(1/\beta) + N_t}{t}.$$

Dividing both sides of the above inequality by $\eta\alpha_t$, where $\eta = \min\{\eta_1, \eta_2\}$ and recalling that $\alpha_t = \frac{a}{(t+1)^\delta}$, we get

$$\frac{1}{t} \sum_{k=1}^t \min\{\|\nabla f(\mathbf{x}^{(k)})\|, \|\nabla f(\mathbf{x}^{(k)})\|^2\} \leq \frac{2^\delta (\log(1/\beta) + N_t)}{a\eta t^{1-\delta}},$$

with probability at least $1 - \beta$. Considering the different choices of step-size parameter $\delta \in (2/3, 1)$, we can obtain the same convergence rates as in Theorem 1. The same trick can be used to show convergence guarantees of the exact Polyak-Ruppert average $\tilde{\mathbf{x}}^{(t)} \triangleq \frac{1}{t} \sum_{k=1}^t \mathbf{x}^{(k)}$ in Corollary 1.1.

As discussed, the metric $\frac{1}{t} \sum_{k=1}^t \min\{\|\nabla f(\mathbf{x}^{(k)})\|, \|\nabla f(\mathbf{x}^{(k)})\|^2\}$ is a more general quantity than $\min_{k \in [t]} \|\nabla f(\mathbf{x}^{(k)})\|^2$, in the sense that in our proof of Theorem 1, we used the bounds on the metric $\frac{1}{t} \sum_{k=1}^t \min\{\|\nabla f(\mathbf{x}^{(k)})\|, \|\nabla f(\mathbf{x}^{(k)})\|^2\}$ to show that they imply the same rates on the more standard metric $\min_{k \in [t]} \|\nabla f(\mathbf{x}^{(k)})\|^2$.¹¹ The metric $\frac{1}{t} \sum_{k=1}^t \min\{\|\nabla f(\mathbf{x}^{(k)})\|, \|\nabla f(\mathbf{x}^{(k)})\|^2\}$ is directly comparable to the metric used in Nguyen et al. (2023a), namely $\frac{1}{t} \sum_{k=1}^t \|\nabla f(\mathbf{x}^{(k)})\|^2$. Moreover, the two metrics are asymptotically equivalent, in the sense that, for some $t_0 \in \mathbb{N}$ sufficiently large, we have $\min\{\|\nabla f(\mathbf{x}^{(k)})\|, \|\nabla f(\mathbf{x}^{(k)})\|^2\} = \|\nabla f(\mathbf{x}^{(k)})\|^2$, for all $k \geq t_0$, as the gradient norm converges to zero with high-probability, according to Theorem 1. The expression $\min\{\|\nabla f(\mathbf{x}^{(k)})\|, \|\nabla f(\mathbf{x}^{(k)})\|^2\}$ stems from our general, black-box analysis in Lemma 3.2, and was also previously used in works studying clipping, e.g., Zhang et al. (2020); Chen et al. (2020). We used the more standard metric $\min_{k \in [t]} \|\nabla f(\mathbf{x}^{(k)})\|^2$, to simplify the exposition in Theorem 1.

Finally, the reason why Nguyen et al. (2023a) are able to provide bounds on the quantity $\frac{1}{t} \sum_{k=1}^t \|\nabla f(\mathbf{x}^{(k)})\|^2$ stems from the fact that a large clipping threshold is used in their analysis, proportional to $t^{1/(3p-2)}$, allowing the authors to show that the norms of gradients of the sequence of iterates, i.e., $\|\nabla f(\mathbf{x}^{(k)})\|$, for all $k = 1, \dots, t$, are guaranteed to stay below the clipping threshold with high probability, i.e., that no clipping will be performed with high probability, in effect behaving like SGD with no clipping. As observed in a recent work Hübler et al. (2024), this is contrary to how clipping is used in practice, where clipping is typically deployed with a small, constant threshold, see Hübler et al. (2024) and references therein. On the other hand, our general black-box analysis provides convergence guarantees of (joint) clipped SGD for any constant value of the clipping threshold, bridging the existing gap between theory and practice.

¹¹Technically, we use the bounds on the metric $\sum_{k=1}^t \frac{\alpha_k}{\sum_{s=1}^t \alpha_s} \min\{\eta_1 \|\nabla f(\mathbf{x}^{(k)})\|, \eta_2 \|\nabla f(\mathbf{x}^{(k)})\|^2\}$, however, as we showed above, we can easily switch to the metric $\frac{1}{t} \sum_{k=1}^t \min\{\|\nabla f(\mathbf{x}^{(k)})\|, \|\nabla f(\mathbf{x}^{(k)})\|^2\}$.