Nonlinear Gradient Mappings and Stochastic Optimization: A General Framework with Applications to Heavy-tail Noise

Dušan Jakovetić
Faculty of Sciences
University of Novi Sad
dusan.jakovetic@dmi.uns.ac.rs

Dragana Bajović
Faculty of Technical Sciences
University of Novi Sad
dbajovic@uns.ac.rs

Anit Kumar Sahu
Amazon Alexa AI
anit.sahu@gmail.com

Soummya Kar
Carnegie Mellon University
soummyak@andrew.cmu.edu

Nemanja Milošević
Faculty of Sciences
University of Novi Sad
nmilosev@dmi.uns.ac.rs

Dušan Stamenković
Faculty of Sciences
University of Novi Sad
dusan.stamenkovic@dmi.uns.ac.rs

February 11, 2022

Abstract

We introduce a general framework for nonlinear stochastic gradient descent (SGD) for the scenarios when gradient noise exhibits heavy tails. The proposed framework subsumes several popular nonlinearity choices, like clipped, normalized, signed or quantized gradient, but we also consider novel nonlinearity choices. We establish for the considered class of methods strong convergence guarantees assuming a strongly convex cost function with Lipschitz continuous gradients under very general assumptions on the gradient noise. Most notably, we show that, for a nonlinearity with bounded outputs and for the gradient noise that may not have finite moments of order greater than one, the nonlinear SGD's mean squared error (MSE), or equivalently, the expected cost function's optimality gap, converges to zero at rate $O(1/t^{\zeta})$, $\zeta \in (0,1)$. In contrast, for the same noise setting, the linear SGD generates a sequence with unbounded variances. Furthermore, for the nonlinearities that can be decoupled component wise, like, e.g., sign gradient or component-wise clipping, we show that the nonlinear SGD asymptotically (locally) achieves a $O(1/t)$ rate in the weak convergence sense and explicitly quantify the corresponding asymptotic variance. Experiments show that, while our framework is more general than existing studies of SGD under heavy-tail noise, several easy-to-implement nonlinearities from our framework are competitive with state of the art alternatives on real data sets with heavy tail noises.

1 Introduction

Stochastic gradient descent (SGD) and its variants, e.g., [1, 2, 3, 4, 5, 6, 7, 8], are popular and standard methods for large scale optimization and training of various machine learning models, e.g., [9, 10, 11, 12]. Recently, there have been several studies that demonstrate that the gradient noise in SGD that arises, e.g., when training deep learning models, is heavy-tailed, e.g., [13, 14, 15].

Motivated by these studies, we introduce a general analytical framework for *nonlinear* SGD when the gradient evaluation is subject to a heavy-tailed noise. We combat the gradient noise with a generic nonlinearity that is applied on the noisy gradient to effectively reduce the noise effect. The resulting class of nonlinear methods subsumes several popular choices in training machine learning models, including normalized gradient descent and clipped gradient descent, e.g., [16], the sign gradient, e.g., [17], and (component-wise) quantized gradient, e.g., [18].[1]

We establish for the considered class of methods several results that demonstrate a high degree of robustness to noise under very general assumptions on the nonlinearity and on the gradient noise, assuming a strongly convex cost with Lipschitz continuous gradient. First, for a nonlinearity with bounded outputs (e.g., a sign, normalized, or clipped gradient) and the gradient noise that may have infinite moments of order greater than one, assuming that the noise probability density function (pdf) is symmetric, we show that the nonlinear SGD converges almost surely to the solution, and, moreover, achieves a global $O(1/t^{\zeta})$ mean squared error (MSE) convergence rate, where we explicitly quantify the degree $\zeta \in (0,1)$. In the same setting, the linear SGD generates a sequence with unbounded variances at each iteration t. Furthermore, assuming the gradient noise with finite variance, we show – for the unbounded nonlinearities that are lower bounded by a linear function – almost sure convergence and he $O(1/t)$ global MSE rate.

Next, for the nonlinearities with bounded outputs that can be decoupled component-wise (e.g., a sign or component-wise clipping), we show under the heavy-tail noise a local (asymptotic) $O(1/t)$ rate in the weak convergence sense. More precisely, we show that the sequence generated by the nonlinear SGD is asymptotically normal and explicitly quantify the asymptotic variance. Finally, we illustrate the results on several examples of the nonlinearity and the gradient noise pdf, highlighting and quantifying the noise regimes and the corresponding gains of the nonlinear SGD over the linear SGD scheme. In more detail, the asymptotic variance expression reveals an interesting tradeoff that the nonlinearity makes on the algorithm performance: on the one hand, the nonlinearity suppresses the noise effect to a certain degree, but on the other hand it also reduces the "useful information flow" and hence slows down convergence with respect to the noiseless case. We explicitly quantify this tradeoff and demonstrate through examples that an appropriately chosen nonlinearity strictly improves performance over the linear scheme in a high noise setting. Finally, we carry out numerical experiments on several real data sets that exhibit heavy tail gradient noise effects. The experiments show that, while our analytical framework is more general than usual studies of SGD under heavy-tail noise, several easy-to-implement example nonlinearities of our framework – including those not previously used – are competitive with state of the art alternatives.

Technically, for component-wise nonlinearities and the asymptotic analysis, we develop proofs based on stochastic approximation arguments, e.g., [20], following the noise and nonlinearities assumptions framework similar to [21]. The paper [21] is concerned with a related but different problem than ours: it considers linear estimation of a vector parameter observed through a sequence of scalar observation equations, and it is not concerned with a global MSE rate analysis that we provide here. For the MSE analysis and for the nonlinearities that cannot be expressed component-wise, like the clipped and normalized gradient, we develop novel analysis techniques.

There have been several works that study robustness of stochastic gradient descent under certain variants of heavy-tailed noises. Reference [15] consider an adaptive gradient clipping method and establish convergence rates in expectation for the considered method under a heavy-tailed noise. For this, the authors assume that the expected value of the norm of the gradient noise raised to power α is finite, for $\alpha \in (1,2]$. They also provide lower complexity bounds for SGD methods assuming in addition that the expected α-power of the norm of the *stochastic gradient* is finite. The authors of [22] consider an accelerated SGD with gradient clipping. They establish high probability bounds for the considered method under the noise that has finite second moment but that does not have to satisfy the sub-Gaussianity assumption. Reference [23] proposes a method called proxBoost and establishes for the method high probability bounds, again assuming a finite noise variance and relaxing the sub-Gaussianity assumption. The paper [13] establishes convergence of the *linear* SGD assuming that the gradient noise follows a heavy-tailed α-stable distribution. In summary, with respect to existing work, our framework establishes results for the more general setting with respect to both the adopted nonlinearity in SGD and the "thickness" of the gradient noise tail, assuming in addition that the noise pdf is a symmetric function. For example, current works usually assume a single choice for the nonlinearity, e.g., gradient clipping, while we consider a general nonlinearity that subsumes many popular choices. Also, provided that the nonlinearity's output is bounded (which is true for many popular choices like the clipped, signed, and normalized gradient), we establish a sublinear MSE convergence rate $O(1/t^{\zeta})$ assuming only that the expected norm of the gradient noise is finite, an assumption weaker than those considered in the works of [22, 15, 23, 13]. On the other hand, we assume a strongly convex smooth cost function, which is equivalent to or stronger than the assumptions made in these works. The MSE rate we establish ζ may be slower than that in the work of [15]. However, the rate ζ holds uniformly for a general class of nonlinearities, holds for a "thicker" noise tail than that in [15], and holds uniformly irrespective of the assumed "thickness" of the noise tail. In contrast, the MSE rate in [15] holds for a specific adaptive gradient clipping scheme and is dependent on the degree α of the assumed finite noise moment.

[1] Interestingly, some of these nonlinear methods are usually introduced with a different motivation than robustness, like, e.g., speeding up training, see, e.g., [16], or communication efficiency, [17, 19].

The idea of employing a nonlinearity into a "baseline" linear scheme has also been used in other contexts. Most notably, several works consider nonlinear versions of the standard consensus algorithm to evaluate average of scalar values in a distributed fashion, e.g., [24, 25, 26]. The paper [24] introduces a trigonometric nonlinearity into a standard linear consensus dynamics and shows an improved dependence of the method on initial conditions. References [25] and [26] employ a general nonlinearity in the linear consensus dynamics and show that it improves the method's resilience to additive communication noise. The authors of [27] modify the linear consensus by taking out from the averaging operation the maximal and minimal estimates among the estimates from all neighbors of a node. The above works are different from ours as they focus on the specific consensus problem that can be translated into minimizing a convex quadratic cost function in a distributed way over a generic, connected network. In contrast, we consider general strongly convex costs, and we are not directly concerned with distributed systems.

Paper organization. Section 2 describes the problem model and the nonlinear SGD framework that we assume. Section 3 and 4 explain our results on nonlinear SGD for component-wise and joint nonlinearities, respectively. Sections 5 and 6 then provide proofs of the corresponding results. Section 7 illustrates the performance of several example methods from our nonlinear SGD framework on real data sets that have heavy-tail gradient noise. Finally, Section 8 concludes the paper. Some auxiliary results and proofs are delegated to the Appendix.

Notation. We denote by $\mathbb{R}$ and $\mathbb{R}_+$, respectively, the set of real numbers and real nonnegative numbers, and by $\mathbb{R}^m$ the m-dimensional Euclidean real coordinate space. We use normal (lower-case or upper-case) letters for scalars, lower-case boldface letters for vectors, and upper case boldface letters for matrices. Further, we denote by: a_i or $[\mathbf{a}]_i$, as appropriate, the i-th element of vector $\mathbf{a}$; $\mathbf{A}_{ij}$ or $[\mathbf{A}]_{ij}$, as appropriate, the entry in the i-th row and j-th column of a matrix $\mathbf{A}$; $\mathbf{A}^\top$ the transpose of a matrix $\mathbf{A}$; and $\mathrm{trace}(\mathbf{A})$ the sum of diagonal elements of $\mathbf{A}$. Further, we use either $\mathbf{a}^\top\mathbf{b}$ or $\langle \mathbf{a},\, \mathbf{b}\rangle$ for the inner product of vectors $\mathbf{a}$ and $\mathbf{b}$. Next, we let $\mathbf{I}$ and $\mathbf{0}$ be, respectively, the identity matrix and the zero matrix; $\|\cdot\| = \|\cdot\|_2$ the Euclidean (respectively, spectral) norm of its vector (respectively, matrix) argument; $\phi'(w)$ the first derivative evaluated at w of a function $\phi:\mathbb{R}\to\mathbb{R}$; $\nabla h(\mathbf{w})$ and $\nabla^2 h(\mathbf{w})$ the gradient and Hessian, respectively, evaluated at $\mathbf{w}$ of a function $h:\mathbb{R}^m\to\mathbb{R}$; $\mathbb{P}(\mathcal{A})$ and $\mathbb{E}[u]$ the probability of an event $\mathcal{A}$ and expectation of a random variable u, respectively; and by $\mathrm{sign}(a)$ the sign function, i.e., $\mathrm{sign}(a)=1$, for $a>0$, $\mathrm{sign}(a)=-1$, for $a<0$, and $\mathrm{sign}(0)=0$. Finally, for two positive sequences η_n and χ_n, we have: $\eta_n = O(\chi_n)$ if $\limsup_{n\to\infty}\frac{\eta_n}{\chi_n}<\infty$; $\eta_n=\Omega(\chi_n)$ if $\liminf_{n\to\infty}\frac{\eta_n}{\chi_n}>0$; and $\eta_n=\Theta(\chi_n)$ if $\eta_n=O(\chi_n)$ and $\eta_n=\Omega(\chi_n)$.

2 Problem Model and the nonlinear SGD Framework

We consider the following unconstrained problem:

$$\text{minimize}\quad f(\mathbf{x}), \tag{2.1}$$

where $f:\mathbb{R}^d\mapsto\mathbb{R}$ is a convex function.

We make the following standard assumption.

Assumption 1 *Function $f:\mathbb{R}^d\mapsto\mathbb{R}$ is strongly convex with strong convexity parameter $\mu>0$, and it has Lipschitz continuous gradient with Lipschitz constant $L\geq\mu$.*

Under Assumption 1, problem (2.1) has a unique solution, which we denote by $\mathbf{x}^\star\in\mathbb{R}^d$.

In machine learning settings, function f can correspond to the risk function, i.e.,

$$f(\mathbf{x}) = \mathbb{E}_{d\sim P}\left[\,\ell\left(\mathbf{x};\mathbf{d}\right)\right] + \mathcal{R}(\mathbf{x}). \tag{2.2}$$

Here, P is the (unknown) distribution from which the data samples $\mathbf{d}\in\mathbb{R}^q$ are drawn; $\ell(\cdot;\cdot)$ is a loss function, convex in its first argument for any fixed value of the second argument; and $\mathcal{R}:\mathbb{R}^d\mapsto\mathbb{R}$ is a strongly convex regularizer. Similarly, f can be empirical risk, i.e., $f(\mathbf{x})=\frac{1}{n}\left(\sum_{j=1}^n \ell\left(\mathbf{x};\mathbf{d}_j\right)\right)+\mathcal{R}(\mathbf{x})$, where $\mathbf{d}_j$, $j=1,...,n$, is the set of training data points. Several machine learning models fall within the described framework under Assumption 1, including, e.g., ℓ_2-regularized quadratic and logistic losses.

We introduce a general framework for *nonlinear* SGD methods to solve problem 2.1; an algorithm within the framework takes the following form:

$$\mathbf{x}^{t+1} = \mathbf{x}^t - \alpha_t\boldsymbol{\Psi}(\nabla f(\mathbf{x}^t)+\boldsymbol{\nu}^t). \tag{2.3}$$

Here, $\mathbf{x}^t$ denotes the solution estimate at iteration t, $t=0,1,...$; $\boldsymbol{\Psi}:\mathbb{R}^d\mapsto\mathbb{R}^d$ is a general nonlinear map; α_t is the employed step size; $\boldsymbol{\nu}^t\in\mathbb{R}^d$ is a zero-mean gradient noise; and $\mathbf{x}^0$ is an arbitrary deterministic point in $\mathbb{R}^d$.

We will specify further ahead the assumptions that we make on the step size α_t, the map $\boldsymbol{\Psi}$ and the noise $\boldsymbol{\nu}^t$. Some examples of commonly used maps $\boldsymbol{\Psi}$ that fall within our framework are the following:

1. Sign gradient: $[\mathbf{\Psi}(\mathbf{w})]_i = \mathrm{sign}(w_i)$, $i = 1, ..., d$;
2. Component-wise clipping: $[\mathbf{\Psi}(\mathbf{w})]_i = w_i$, for $|w_i| \leq m$; $[\mathbf{\Psi}(\mathbf{w})]_i = m$, for $w_i > m$, and $[\mathbf{\Psi}(\mathbf{w})]_i = -m$, for $w_i < -m$, for some constant $m > 0$.
3. Component-wise quantization: for each $i = 1, ..., d$, we let $[\mathbf{\Psi}(\mathbf{w})]_i = r_j$, for $w_i \in (q_{j-1}, q_j]$, $j = 1, ..., J$, where $-\infty = q_0 < q_1 < ... < q_J = +\infty$, J is a positive integer, and the r_j's and q_j's are chosen such that each component nonlinearity is an odd function, i.e., $[\mathbf{\Psi}(\mathbf{w})]_i = -[\mathbf{\Psi}(-\mathbf{w})]_i$, for each i and for each $\mathbf{w}$;
4. Normalized gradient: $\mathbf{\Psi}(\mathbf{w}) = \frac{\mathbf{w}}{\|\mathbf{w}\|}$, for $\mathbf{w} \neq 0$, and $\mathbf{\Psi}(0) = 0$;
5. Clipped gradient: $\mathbf{\Psi}(\mathbf{w}) = \mathbf{w}$, for $\|\mathbf{w}\| \leq M$, and $\mathbf{\Psi}(\mathbf{w}) = \frac{\mathbf{w}}{\|\mathbf{w}\|} M$, for $\|\mathbf{w}\| > M$, for some constant $M > 0$.

Other nonlinearity choices are also introduced ahead (see Section 7).

We next discuss the various possible sources of the gradient noise $\boldsymbol{\nu}^t$. First, the noise may arise due to utilizing a search direction with respect to a data sample. That is, a common search direction in machine learning algorithms is the gradient of the loss with respect to a single data point $\mathbf{d}_i$[2]: $\mathbf{g}_i(\mathbf{x}) = \nabla \ell\,(\mathbf{x}; \mathbf{d}_i) + \nabla \mathcal{R}(\mathbf{x})$. In case of the risk function (2.2), $\boldsymbol{d}_i$ is drawn from distribution P; in case of the empirical risk, $\boldsymbol{d}_i$ can be, e.g., drawn uniformly at random from the set of data points $\boldsymbol{d}_j$, $j = 1, ..., n$, with repetition along iterations. In both cases, the corresponding gradient noise equals $\boldsymbol{\nu} = \mathbf{g}_i(\mathbf{x}) - \nabla f_i(\mathbf{x})$. Several recent studies indicate that noise $\boldsymbol{\nu}$ exhibits heavy tails on many real data sets, e.g, [13, 14, 15], (See also Section 7.)

We also comment on other possible sources of gradient noise. The noise may be added on purpose to the gradient $\nabla f(\mathbf{x})$ for improving privacy of an SGD-based learning process, e.g., [28]. Also, the noise $\boldsymbol{\nu}^t$ may model random computational perturbations or inexact calculations in evaluating a gradient $\nabla f(\mathbf{x})$.

3 Main results: Component-wise Nonlinearities

Section 3 provides analysis of the nonlinear SGD method for component-wise nonlinearities. That is, we consider here maps $\mathbf{\Psi} : \mathbb{R}^d \mapsto \mathbb{R}^d$ of the form $\mathbf{\Psi}(w_1, ..., w_d))^\top = (\Psi(w_1), ..., \Psi(w_d))^\top$, for any $\mathbf{w} \in \mathbb{R}^d$, where (somewhat abusing notation) we denote by $\Psi :\, \mathbb{R} \mapsto \mathbb{R}$ the component-wise nonlinearity. In this setting, we establish for (2.3) almost sure convergence and evaluate the MSE convergence rate and the asymptotic covariance of the method.

In more detail, we consider algorithm (2.3) under Assumptions 2 and 3 below; they follow the noise and nonlinearity framework similar to [21].

Assumption 2 (Gradient noise) *For the gradient noise random vector sequence $\{\boldsymbol{\nu}^t\}$ in (2.3), $t = 0, 1, ...$, $\boldsymbol{\nu}^t \in \mathbb{R}^d$, we assume the following:*

1. *$\{\boldsymbol{\nu}^t\}$ is independent identically distributed (i.i.d.) across iterations, and, for any fixed t, $\boldsymbol{\nu}^t$ is independent of $\mathbf{x}^t$. Also, random variables $\boldsymbol{\nu}_i^t$ are mutually independent across $i = 1, ...d;$*
2. *Each component ν_i^t, $i = 1, ..., d$, of vector $\boldsymbol{\nu}^t = (\nu_1^t, ..., \nu_d^t)^\top$ has a probability density function $p(u)$, $p : \mathbb{R} \mapsto \mathbb{R}_+$. The pdf p is symmetric, i.e., $p(u) = p(-u)$, for any $u \in \mathbb{R}$ with $\int |u| p(u) du < +\infty$.*
3. *The pdf $p(u)$ is strictly unimodal, i.e., $p(0) < +\infty$ and $p(v_1) < p(v_2)$ for $|v_1| > |v_2|$.*
4. *Function Ψ is strictly increasing, and, for the cumulative distribution function (cdf) associated with pdf p, $\Phi(u) = \int_{-\infty}^{u} p(v) dv$, it holds that Φ and Ψ have a common growth point, i.e., $\Phi(v + \epsilon) > \Phi(v - \epsilon)$ and $\Psi(v + \epsilon) > \Psi(v - \epsilon)$ for a certain $v \in \mathbb{R}$ and for all $\epsilon > 0$.*

We assume that at least one of the conditions 3. or 4. hold.

Conditions 1. and 2. in Assumption 2 concerning the requirement that the noise vector is i.i.d. across its components $i = 1, ..., d$ may be restrictive. For the global MSE analysis, these assumptions can be relaxed; see ahead the remark after Theorem 3.2 and Appendix C.

Assumption 3 (Nonlinearity Ψ) *Function $\Psi : \mathbb{R} \mapsto \mathbb{R}$ has the following properties:*

1. *Function Ψ is a continuous (except possibly on a point set with Lebesgue measure of zero), piece-wise differentiable, monotonically nondecreasing and odd function, i.e., $\Psi(-w) = -\Psi(w)$, for any $w \in \mathbb{R}$;*

[2] Similar considerations hold for a loss with respect to a mini-batch of data points; this discussion is abstracted for simplicity.

2. $|\Psi(w)| \leq C_1\,(1 + |w|)$*, for any* $w \in \mathbb{R}$*, for some constant* $C_1 > 0$*.*

3. $|\Psi(w)| \leq C_2$*, for some constant* $C_2 > 0$*.*

Here, we assume that either 2. or 3. holds. If 2. holds, then we additionally require a finite variance for the gradient noise, i.e., there holds $\int |u|^2 p(d)du < +\infty$.[3]

Note that, provided that condition 3. in Assumption 3 holds, we require only a finite first moment of the gradient noise, while the moments of α-order, $\alpha > 1$, may be infinite, hence allowing for heavy-tail noise distributions. For example, the gradient noise variance can be infinite. Condition 3. in Assumption 3 holds for several interesting component-wise nonlinearities, like, e.g., the sign gradient, component-wise clipping, and quantization schemes introduced in Section 2. Note also that Assumption 3 encompasses a broad range of component-wise nonlinearities, beyond the examples in Section 2. (For example, see Section 7 for the $\tanh$ and a bi-level quantization nonlinearity.)

Let us define function $\phi : \mathbb{R} \mapsto \mathbb{R}$, as follows. For a fixed (deterministic) point $w \in \mathbb{R}$, $\phi(w)$ is defined by:

$$\phi(w) = \mathbb{E}\left[\Psi(w + \nu_1^0)\right] = \int \Psi(w + u)p(u)du, \tag{3.1}$$

where the expectation is taken with respect to the distribution of a single entry of the gradient noise at any iteration, i.e., with respect to pdf $p(u)$. Intuitively, the nonlinearity ϕ is a convolution-like transformation of the nonlinearity Ψ, where the convolution is taken with respect to the gradient noise pdf $p(u)$. As we will see ahead, the nonlinearity ϕ plays an effective role in determinining the performance of algorithm (2.3).

We have the following Theorem.

Theorem 3.1 (Almost sure convergence: Component-wise nonlinearity) *Consider algorithm* (2.3) *for solving optimization problem* (2.1)*, and let Assumptions 1–3 hold. Assume in addition that f is twice continuously differentiable. Further, let the step-size sequence* $\{\alpha_t\}$ *be square summable, non-summable:* $\sum \alpha_t = +\infty$*;* $\sum \alpha_t^2 < +\infty$*. Then, the sequence of iterates* $\{\mathbf{x}^t\}$ *generated by algorithm* (2.3) *converges almost surely to the solution* $\mathbf{x}^\star$ *of the optimization problem* (2.1).

Theorem 3.1 establishes a.s. convergence of the nonlinear SGD scheme (2.3) under a general setting for the component-wise nonlinearities and gradient noise. For example, provided that the output of the nonlinearity Ψ is bounded, algorithm (2.3) converges even when the gradient noise may not have a finite α-moment, for any $\alpha > 1$. (Hence it may have an infinite variance). In contrast, as shown in Appendix 8, the linear SGD (algorithm (2.3) with $\mathbf{\Psi}$ being the identity function) generates a sequence of solution estimates with infinite variances, provided that the variance of $p(u)$ is infinite.

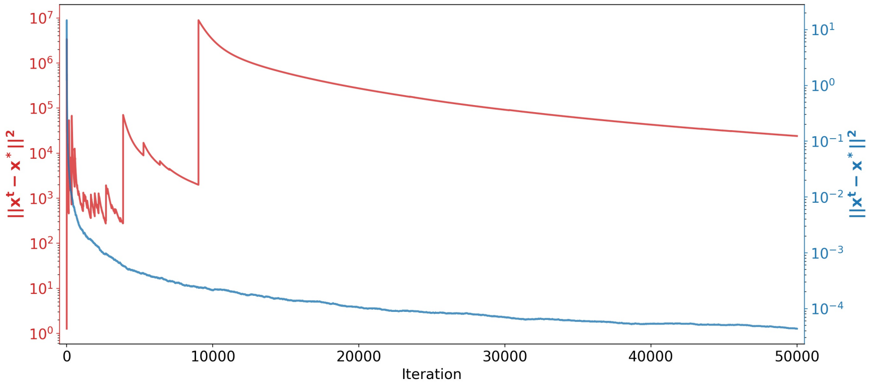


Figure 3.1: Illustration of Theorem 3.1: estimated MSE versus iteration counter for the nonlinear SGD in (2.3) with component-wise sign nonlinearity (blue line) and the linear SGD (red line).

[3] As it will be seen in subsequent text, several statements of results and several proofs treat separately the following two scenarios: 1) condition 2. in Assumption 3 holds (and the gradient noise may not have finite variance); and 2) condition 3. in Assumption 3 holds, but the gradient noise has finite variance. We clearly indicate ahead when we want to distinguish between the two scenarios. For example, when we say that condition 3. in Assumption 3 holds, we refer to the second scenario above. See, e.g., Theorem 3.2. If the result holds for either of the two scenarios, we do not make specific mention of condition 2. or 3. in Assumption 3. See, e.g., Theorem 3.1 .

Example 3.1 *Figure 3.1 illustrates Theorem 3.1 with a simulation example. We consider a strongly convex quadratic function* $f : \mathbb{R}^d \mapsto \mathbb{R}$*,* $f(\mathbf{x}) = \mathbf{x}^\top \mathbf{A}\mathbf{x} + \mathbf{b}^\top \mathbf{x}$*, where* $\mathbf{A} \in \mathbb{R}^{d\times d}$ *is a (symmetric) positive definite matrix,* $d = 16$*, and quantities* $\mathbf{A}, \mathbf{b}$ *are generated at random. We consider algorithm* (2.3) *with the component-wise sign nonlinearity and the linear SGD. The gradient noise is i.i.d. across iterations and across components and has the following pdf:*

$$p(u) = \frac{\alpha - 1}{2(1 + |u|)^\alpha}, \tag{3.2}$$

for $u \in \mathbb{R}$ *and* $\alpha > 2$*. Note that the distribution* (3.2) *does not have a finite* $\alpha - 1$ *moment and has finite moments of* r*-th order for* $r < \alpha - 1$*. We set in simulation* $\alpha = 2.05$*. Note that, in this case, the gradient noise has infinite variance. We initialize both the linear and nonlinear algorithm with* $\mathbf{x}^0 = 0$*, and we let step size* $\alpha_t = \frac{1}{t+1}$*. Figure 1 shows an estimate of MSE, i.e., of the quantity* $\mathbb{E}[\|\mathbf{x}^t - \mathbf{x}^\star\|^2]$*, obtained by averaging results from 100 sample paths. The red line corresponds to the linear SGD, while the blue line corresponds to the nonlinear SGD with the component-wise sign nonlinearity. As predicted by Theorem 3.1, the nonlinear SGD drives the MSE to zero, while the linear SGD does not seem to provide a meaningful solution estimate sequence.*

We next establish the mean square error (MSE) convergence rate of algorithm (2.3).

Theorem 3.2 (MSE convergence: Component-wise nonlinearity) *Consider algorithm* (2.3) *for solving optimization problem* (2.1)*, and let Assumptions 1, 2, and Assumption 3 with condition 3. hold Further, let the step-size sequence* $\{\alpha_t\}$ *be* $\alpha_t = a/(t+1)^\delta$*,* $a > 0$*,* $\delta \in [0.5, 1)$*. Then, for the sequence of iterates* $\{\mathbf{x}^t\}$ *generated by algorithm* (2.3)*, it holds that* $\mathbb{E}\left[\|\mathbf{x}^t - \mathbf{x}^\star\|^2\right] = O(1/t^\zeta)$*, or equivalently,* $\mathbb{E}\left[f(\mathbf{x}^t) - f^\star\right] = O(1/t^\zeta)$*. Here,* $\zeta < 1$ *is any positive number such that* $\zeta < \min\left(2\delta - 1, \frac{a\,(1-\delta)\xi\,\phi'(0)\mu}{L\,(a\,C_2\,\sqrt{d} + \|\mathbf{x}^0\| + \|\mathbf{x}^\star\|)}\right)$*, and constant* $\xi > 0$ *is such that* $\phi(a) \geq \frac{\phi'(0)\xi}{2}\,a$*, for any* $a \in (0, \xi)$*. Furthermore, let Assumptions 1, 2, and Assumption 3 with condition 2. hold, let* $\alpha_t = \frac{a}{(t+1)^\delta}$*,* $\delta \in [0.5, 1]$*, and assume that* $\inf_{a\neq 0} \frac{|\Psi(a)|}{|a|} > 0$*. Then, there holds that* $\mathbb{E}\left[\|\mathbf{x}^t - \mathbf{x}^\star\|^2\right] = O(1/t^\delta)$*, or equivalently,* $\mathbb{E}\left[f(\mathbf{x}^t) - f^\star\right] = O(1/t^\delta)$*. In particular, for* $\delta = 1$*, we obtain the* $O(1/t)$ *MSE rate.*

Remark. The MSE convergence $O(1/t^{\zeta'})$, for some $\zeta' \in (0, 1)$, continues to hold under the same set of assumptions as in Theorem 3.2 but with a relaxed version of Assumption 2, where we no longer require that the gradient noise vector has mutually independent components. More precisely, we allow for an i.i.d. noise vector sequence $\{\boldsymbol{\nu}^t\}$, $\boldsymbol{\nu}^t \in \mathbb{R}^d$, that has a symmetric joint pdf $p : \mathbb{R}^d \mapsto \mathbb{R}$, $p(\mathbf{u}) = p(-\mathbf{u})$, for any $\mathbf{u} \in \mathbb{R}^d$. In that case, effectively, the role of function ϕ in Theorem 3.2 is replaced by functions $w \mapsto \phi_i(w)$, $w \in \mathbb{R}$, $i = 1, ..., d$, where $\phi_i(w) = \int \Psi(w + u)p_i(u)du$, and $p_i : \mathbb{R} \mapsto \mathbb{R}$ is the marginal pdf of the i-th component associated with the joint pdf $p : \mathbb{R}^d \mapsto \mathbb{R}$. (See Appendix C.)

For the bounded nonlinearity case (e.g., sign gradient, component-wise clipping, quantization nonlinearity) and the heavy-tail noise (only the first noise moment assumed to be finite), the nonlinear SGD (2.3) achieves a global sublinear MSE rate $O(1/t^\zeta)$, $\zeta \in (0, 1)$. On the other hand, for the finite variance case and an unbounded nonlinearity, the nonlinear SGD (2.3) achieves a global MSE rate $O(1/t)$ provided that $\inf_{w\neq 0} \frac{|\Psi(w)|}{|w|} > 0$. This is the best achievable rate and equal to that of the linear SGD in the same setting. Furthermore, by Theorem 3.3 ahead, the nonlinear SGD (2.3) with bounded outputs under the heavy-tail noise achieves *locally*, in the weak convergence sense, the faster $O(1/t)$ rate. This is again in the setting where the linear SGD fails.

Example 3.2 *We next illustrate the value* ζ *in Theorem 3.2 on the family of heavy-tailed pdfs in* (3.2)*. To be specific, consider the sign nonlinearity* $\Psi(w) = \mathrm{sign}(w)$*. Then, it is easy to show that:*

$$\phi(w) = 2\int_0^w p(u)du,\ \phi'(0) = 2\,p(0),\ \xi \geq 2^{1/\alpha} - 1 \approx \frac{1}{\alpha}.$$

Using the above calculations, we can see that, for a large a*,* ζ *can be approximated as* $\min\left(2\delta - 1, \frac{\mu}{L}\frac{1-\delta}{\sqrt{d}}\right) > 0$*,* $\delta \in (0.5, 1)$*.*

We also compare the rate ζ *with the analysis in [15] that is closest to our setting with respect to existing work. The authors of [15] assume for the MSE upper bound and strongly convex functions analysis that, in our notation, both the quantities* $\mathbb{E}[\|\nabla f(\mathbf{x}^t) + \boldsymbol{\nu}^t\|^\alpha]$ *(a restrictive bounded gradients assumption) and* $\mathbb{E}[\|\boldsymbol{\nu}^t\|^\alpha]$ *are finite for some* $\alpha \in (1, 2]$*. They then show a rate* $O(1/t^{2(\alpha-1)/\alpha})$ *for a specific adaptive clipping scheme. This rate can be faster than the* ζ *rate we establish, but ours holds uniformly for a general class of nonlinearities and for any symmetric noise pdf with a finite first moment.*

We next establish asymptotic normality of (2.3).

Theorem 3.3 (Asymptotic normality: Component-wise nonlinearity) *Consider algorithm* (2.3) *for solving optimization problem* (2.1)*, and let Assumptions 1–3 hold. Assume in addition that f is twice continuously differentiable. Further, let the step-size sequence $\{\alpha_t\}$ equal: $\alpha_t = a/(t+1)$, $t = 0, 1, ...$, with parameter $a > \frac{1}{2\phi'(0)\,\mu}$. Then, the sequence of iterates $\{\mathbf{x}^t\}$ generated by algorithm* (2.3) *is asymptotically normal, and there holds:*

$$\sqrt{t+1}(\mathbf{x}^t - \mathbf{x}^\star) \xrightarrow{d} \mathbb{N}(0, \mathcal{S}), \tag{3.3}$$

where $\xrightarrow{d}$ designates convergence in distribution. The asymptotic covariance $\mathcal{S}$ of the multivariate normal distribution $\mathbb{N}(0, \mathcal{S})$ is given by:

$$\mathcal{S} = a^2 \int_{\nu=0}^{\infty} e^{\nu \mathbf{\Sigma}} \mathcal{S}_0 e^{\nu \mathbf{\Sigma}} d\nu = a^2 \sigma_\psi^2 \left[2a\phi'(0)\nabla^2 f(x^\star) - \mathbf{I}\right]^{-1},$$

where:

$$\mathcal{S}_0 = \sigma_\Psi^2\,\mathbf{I},\;\; \sigma_\Psi^2 = \int |\Psi(v)|^2 p(v) dv,\;\; \Sigma = \frac{1}{2}\mathbf{I} - a\,\phi'(a)\nabla^2 f(\mathbf{x}^\star). \tag{3.4}$$

Theorem 3.3 establishes asymptotic normality of (2.3) and, moreover, it gives an exact expression for the asymptotic covariance $\mathcal{S}$ in (3.4), that basically corresponds to the constant in the $1/t$ variance decay near the solution. The asymptotic covariance value (3.4) reveals an interesting tradeoff with respect to the effect of the nonlinearity Ψ. We provide some insights into the tradeoff through examples below.

Example 3.3 *We compare the linear SGD and the nonlinear SGD with component wise clipping. For illustration and simplification of calculations, we consider the special case when $\nabla^2 f(\mathbf{x}^\star)$ is a symmetric matrix with all eigenvalues equal to one. Then, it is straightforward to show that the per-entry asymptotic variance for the best choice of parameter a over the admissible set of values equals:*

$$\inf_{a > \frac{1}{2\phi'(0)}} \text{trace}\,(\mathcal{S}) = \frac{\sigma_\Psi^2}{(\phi'(0))^2}. \tag{3.5}$$

Here, for the linear SGD i.e., when $\Psi(a) = a$, we have that $\sigma_\Psi^2 = \int a^2 p(a) da$ equals the gradient noise (per component) variance σ_ν^2, and $\phi'(0) = 1$, and so (3.5) *equals σ_ν^2. Now, consider the coordinate-wise clipping, with $\Psi(a) = a$ for $|a| \leq m$ and $\Psi(a) = \text{sign}(a)\,m$, for $|a| > m$, for some $m > 0$. Then, we have: $\sigma_\Psi^2 = m^2 - 2\int_0^m (m^2 - v^2) p(v) dv$, and $\phi'(0) = 2\int_0^m p(v) dv$. Note that the case $m \to \infty$ corresponds to the linear SGD case. Consider now the tradeoff with respect to the choice of m. Clearly, taking a smaller m has a positive effect on the numerator in* (3.5) *(it suppresses the noise effect). On the other hand, reducing m has a negative effect on the denominator in* (3.5)*; that is, it reduces the value $\phi'(0)$ – intuitively, it "lowers the quality" of the search direction utilized with* (2.3)*. One needs to choose the nonlinearity, i.e., the parameter m, optimally, to strike the best balance here. Clearly, for larger gradient noise σ_ν^2, we should pick a smaller value of m. It can be shown that, for any finite σ_ν^2, there is an optimal value $m^\star \in (0, \infty)$ that minimizes* (3.5).

Example 3.4 *We continue to assume the simplified setting when the per-entry asymptotic variance equals* (3.5). *We consider the sign gradient nonlinearity and the class of heavy-tail gradient noise distributions in* (3.2). *It can be shown that here: $\sigma_\Psi^2 = 1$; $\sigma_\nu^2 = \frac{2}{(\alpha-3)(\alpha-2)}$, for $\alpha > 3$ and $\sigma_\nu^2 = \infty$, else; and $\phi'(0) = \alpha - 1$. Therefore, for the sign gradient, the best achievable per entry asymptotic variance equals $\frac{1}{(\alpha-1)^2}$, while for the linear SGD it equals $\frac{2}{(\alpha-2)(\alpha-3)}$ for $\alpha > 3$, and is infinite for $\alpha \in (2, 3]$. Hence, we can see for the considered example that the sign gradient outperforms the linear SGD for any $\alpha > 2$, and the gap becomes larger as α gets smaller.*

Example 3.5 *We still consider the simplified setting of* (3.5). *If the noise pdf $p(u)$ is known, then, following [21], we can find a globally optimal nonlinearity that minimizes 5.14 that takes the form: $\Psi(a) = -\frac{d}{da}\ln(p(a))$. The corresponding optimal asymptotic variance equals the Fisher information associated with the pdf $p(u)$.*

Example 3.6 *Figure 3.2 illustrates Theorem 3.3 for the nonlinear SGD in* (2.3) *with component-wise sign nonlinearity and the same simulation setting used for the numerical illustration of Theorem 3.1 and step-size $\alpha_t = \frac{10}{t+1}$. The red line plots quantity $\frac{t}{d}\|\mathbf{x}^t - \mathbf{x}^\star\|^2$ estimated through 100 sample path runs. This quantity estimates the constant in the $1/t$ per-entry asymptotic variance decay, i.e., it is a numerical estimate of the per-entry asymptotic variance $\frac{\text{trace}(\mathcal{S})}{d}$, where $\mathcal{S}$ is given in Theorem 3.3. The blue horizontal line marks the value $\frac{\text{trace}(\mathcal{S})}{d}$. We can see that the simulation matches well the theory.*

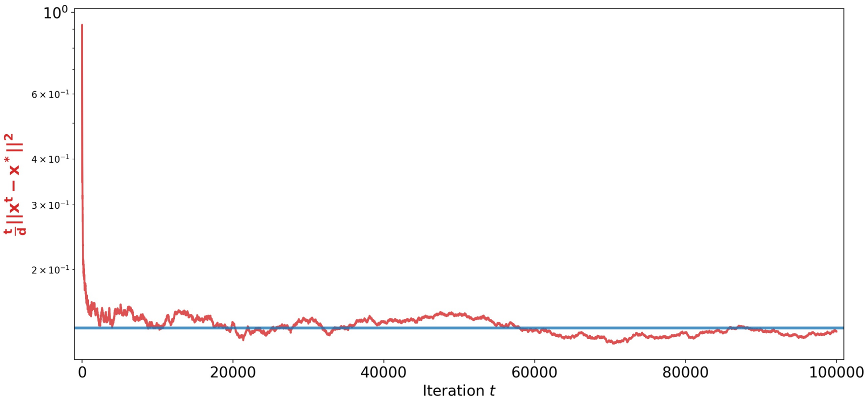


Figure 3.2: Illustration of Theorem 3.3: Monte Carlo estimate of per-entry asymptotic variance (red line) and the theoretical per-entry asymptotic variance in Theorem 3.3 (blue line).

4 Main results: Joint Nonlinearities

We now consider algorithm (2.3) for a nonlinearity $\mathbf{\Psi} : \mathbb{R}^d \mapsto \mathbb{R}^d$ that cannot be decoupled into (equal) component wise nonlinearities $\Psi : \mathbb{R} \mapsto \mathbb{R}$, as it was possible before. More precisely, we make the following assumptions on the gradient noise $\boldsymbol{\nu}^t$ and the nonlinear map $\mathbf{\Psi} : \mathbb{R}^d \mapsto \mathbb{R}^d$.

Assumption 4 *[Gradient noise] For the gradient noise sequence $\{\boldsymbol{\nu}^t\}$, we assume the following:*

1. *The sequence of random vectors $\{\boldsymbol{\nu}^t\}$ is i.i.d. and for any $t = 0, 1, ...,$ $\boldsymbol{\nu}^t$ is independent of $\mathbf{x}^t$. Moreover, $\boldsymbol{\nu}^t$ has a joint symmetric pdf $p(\mathbf{u})$, $p : \mathbb{R}^d \mapsto \mathbb{R}$, i.e., $p(\mathbf{u}) = p(-\mathbf{u})$, for any $\mathbf{u} \in \mathbb{R}^d$ with $\int \|\mathbf{u}\| p(\mathbf{u}) d\mathbf{u} < \infty$;*
2. *There exists a positive constant B_0 such that, for any $\mathbf{x} \in \mathbb{R}^d$, $\mathbf{x} \neq 0$, for any $A \in (0, 1]$, there exists $\lambda = \lambda(A) > 0$, such that*[4] $\int_{\{\mathbf{u} \in \mathbb{R}^d : \, \frac{\mathbf{u}^\top \mathbf{x}}{\|\mathbf{u}\| \|\mathbf{x}\|} \in [0,A], \, \|u\| \leq B_0\}} p(\mathbf{u}) d\mathbf{u} > \lambda(A)$.

Assumption 4 allows for a heavy-tailed noise vector whose components can be mutually dependent. Condition 2. in Assumption 4 is mild; it says that the joint pdf $p(\mathbf{u})$ is "non-degenerate" in the sense that, along each "direction" (determined by arbitrary nonzero vector $\mathbf{x}$), the intersection of the set $\{\frac{\mathbf{u}^\top \mathbf{x}}{\|\mathbf{u}\| \|\mathbf{x}\|} \in [0, A]\}$ and the ball $\{\|\mathbf{u}\| \leq B_0\}$ consumes a positive mass of the joint pdf $p(\mathbf{u})$.

Assumption 5 (Nonlinearity $\mathbf{\Psi}$) *The nonlinear map $\mathbf{\Psi} : \mathbb{R}^d \mapsto \mathbb{R}^d$ takes the folloing form: $\mathbf{\Psi}(\mathbf{w}) = \mathbf{w} \mathcal{N}(\|\mathbf{w}\|)$, where function $\mathcal{N} : \mathbb{R}_+ \mapsto \mathbb{R}_+$ satisfies the following:*

1. *Function $\mathcal{N}$ is non-increasing and continuous except possibly on a point set with Lebesgue measure of zero with $\mathcal{N}(q) > 0$, for any $q > 0$. The function $q\mathcal{N}(q)$ is non-decreasing;*
2. *$\|\mathbf{\Psi}(\mathbf{w})\| \leq C_1'(1 + \|\mathbf{w}\|)$, for any $\mathbf{w} \in \mathbb{R}^d$, for some $C_1' > 0$;*
3. *$\|\mathbf{\Psi}(\mathbf{w})\| \leq C_2'$, for any $\mathbf{w} \in \mathbb{R}^d$, for some $C_2' > 0$.*

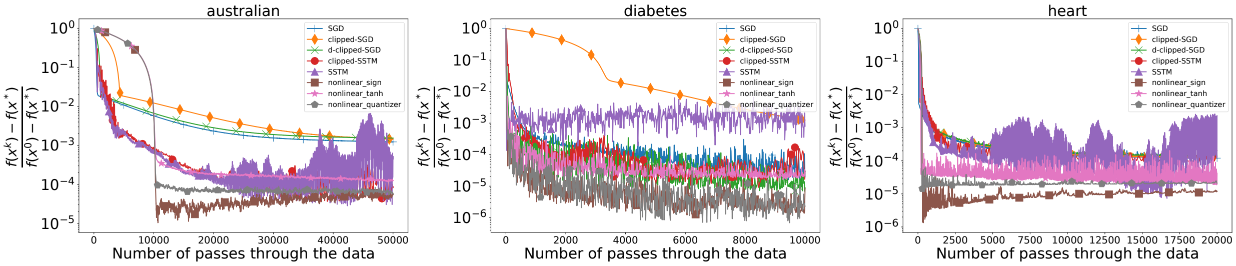
Here, we assume that either 2 or 3 holds. If 2 holds, then we additionally require that the second moment of $\boldsymbol{\nu}^t$ is bounded, i.e., $\int \|\mathbf{u}\|^2 p(\mathbf{u}) d\mathbf{u} < \infty$.[5]

There are many nonlinearities that satisfy Assumption 5, including, the normalized gradient and the clipped gradient discussed in Section 2.

Theorem 4.1 (MSE and a.s. convergence: Joint nonlinearity) *Consider algorithm* (2.3) *for solving optimization problem* (2.1)*, and let Assumptions 1, 4, and Assumption 5 with condition 3. hold. Further, let the step-size sequence*

[4] The integration set $\{\mathbf{u} \in \mathbb{R}^d : \frac{\mathbf{u}^\top \mathbf{x}}{\|\mathbf{u}\| \|\mathbf{x}\|} \in [0, A], \, \|u\| \leq B_0\}$ also includes the point $\mathbf{u} = 0$. In other words, for compact notation here and throughout the paper, we write $\frac{\mathbf{u}^\top \mathbf{x}}{\|\mathbf{u}\| \|\mathbf{x}\|} \in [0, A]$ instead of $0 \leq \mathbf{u}^\top \mathbf{x} \leq A \, \|\mathbf{u}\| \|\mathbf{x}\|$.

[5] Analogously to Assumption 3, we often refer to the two different scenarios, corresponding to conditions 2. and 3. in Assumption, 5, respecticvely. See also the footnote at the end of Assumption 3.



(a) (b) (c)

Figure 4.1: Comparison of the optimization algorithms across different datasets

$\{\alpha_t\}$ be $\alpha_t = a/(t+1)^\delta$, $a > 0$, $\delta \in (0.5, 1)$. Then, for the sequence of iterates $\{\mathbf{x}^t\}$ generated by algorithm (2.3)*, it holds that $\mathbb{E}\left[\|\mathbf{x}^t - \mathbf{x}^\star\|^2\right] = O(1/t^\zeta)$, or equivalently, $\mathbb{E}\left[f(\mathbf{x}^t) - f^\star\right] = O(1/t^\zeta)$, where $\zeta \in (0,1)$. In alternative, let Assumptions 1, 4, and Assumption 5 with condition 2. hold, let $\alpha_t = \frac{a}{(t+1)^\delta}$, $\delta \in (0.5, 1]$, and assume that $\inf_{\mathbf{w}\neq 0} \frac{\|\boldsymbol{\Psi}(\mathbf{w})\|}{\|\mathbf{w}\|} > 0$. Then, $\mathbb{E}\left[\|\mathbf{x}^t - \mathbf{x}^\star\|^2\right] = O(1/t^\delta)$, or equivalently, $\mathbb{E}\left[f(\mathbf{x}^t) - f^\star\right] = O(1/t^\delta)$. In particular, for $\delta = 1$ and a sufficiently large parameter a, we obtain the $O(1/t)$ MSE rate. Finally, under Assumptions 1, 4, and 5, $\delta \in (0.5, 1]$ and f that is in addition twice continuously differentiable, we have that $\mathbf{x}^t$ converges to $\mathbf{x}^\star$, a.s.*

Theorem 4.1 establishes a.s. convergence and a global sublinear MSE rate of algorithm (2.3) for a nonlinearity with bounded outputs (e.g., a normalized or clipped gradient) in the presence of heavy-tailed noise that may have infinite moments of order greater than one. See also the discussion after Theorem 3.2 for analogous interpretations and comparisons with existing work. See ahead (6.23) in the proof of Theorem 4.1 for the obtained bound on rate ζ.

5 Intermediate results and proofs: Component-wise nonlinearities

This section provides proofs of Theorems 3.1–3.3, accompanied with the required intermediate results. Subsection 5.1 deals with the asymptotic analysis (Theorems 3.1 and 3.3), while Subsection 5.2 considers MSE analysis (Theorem 3.2).

5.1 Asymptotic analysis: Proofs of Theorems 3.1 and 3.3

The next Lemma, due to [21], establishes structural properties of function ϕ in (3.1). The Lemma says that essentially, the convolution-like transofrmation of the nonlinearity preserves the structural properties of the nonlinearity.

Lemma 5.1 *[21] Consider function ϕ in* (3.1)*, where function $\Psi : \mathbb{R} \mapsto \mathbb{R}$ satisfies Assumption 3. Then, the following holds.*

1. *ϕ is odd;*
2. *If $|\Psi(\nu)| \leq C_2$, for any $\nu \in \mathbb{R}$, then $|\phi(a)| \leq K_2$, for any $a \in \mathbb{R}$, for some constant $K_2 > 0$;*
3. *If $|\Psi(\nu)| \leq C_1(1 + |\nu|)$, for any $\nu \in \mathbb{R}$, then $|\phi(a)| \leq K_1(1 + |a|)$, for any $a \in \mathbb{R}$, for some constant $K_1 > 0$;*
4. *$\phi(a)$ is monotonically nondecreasing;*
5. *$\phi(a) > 0$, for any $a > 0$.*
6. *ϕ is continuous at zero;*
7. *ϕ is differentiable at zero, with a strictly positive derivative at zero, equal to:*

$$\phi'(0) = \sum_{i=1}^{s} \left(\Psi(\nu_i + 0) + \Psi(\nu_i - 0)\right) p(\nu_i) + \sum_{i=0}^{s} \int_{\nu_i}^{\nu_{i+1}} \Psi'(\nu) p(\nu) d\nu, \tag{5.1}$$

where $\nu_i, i = 1, ..., s$ are points of discontinuity of Ψ such that $\nu_0 = -\infty$ and $\nu_{s+1} = +\infty$

We proceed by setting up the proof of Theorem 3.1. The proof relies on convergence analysis of single-time scale stochastic approximation methods from [20]; more precisely, we utilize Theorem 8.1 in the Appendix; see also [29].

We first put algorithm (2.3) in the format that complies with Theorem 8.1. Namely, algorithm (2.3) can be written as:

$$\mathbf{x}^{t+1} = \mathbf{x}^t + \alpha_t \left[\mathbf{r}(\mathbf{x}^t) + \boldsymbol{\gamma}(t+1, x^t, \omega)\right]. \tag{5.2}$$

Here, ω denotes an element of the underlying probability space, and

$$\mathbf{r}(\mathbf{x}) = -\boldsymbol{\phi}(\nabla f(\mathbf{x})), \tag{5.3}$$

where, abusing notation, $\boldsymbol{\phi} : \mathbb{R}^d \mapsto \mathbb{R}^d$ is defined by $\left(\boldsymbol{\phi}(a_1, ..., a_d)\right)^\top = \left(\phi(a_1), ..., \phi(a_d)\right)^\top$. That is, we have that:

$$\mathbf{r}(\mathbf{x}) = -\left(\phi[\nabla f(x))_1], ..., \phi[\nabla f(x))_d]\right)^\top \tag{5.4}$$

and

$$\boldsymbol{\gamma}(t+1, \mathbf{x}, \omega) = \boldsymbol{\phi}(\nabla f(\mathbf{x})) - \Psi(\nabla f(\mathbf{x}) + \boldsymbol{\nu}^t). \tag{5.5}$$

We provide an intuition behind the algorithmic format (5.2). Quantity $\mathbf{r}(x)$ is a deterministic, "useful", progress direction with respect to the evolution of $\mathbf{x}^t$; quantity $\boldsymbol{\gamma}(t+1, x, \omega)$ is the stochastic component that plays a role of a noise in the system.

We adopt the following Lyapunov function: $V(x) = f(x) - f^\star$, $V : \mathbb{R}^d \mapsto \mathbb{R}$, where $f^\star = \inf_{x \in \mathbb{R}^d} f(x) = f(x^\star)$. We are ready to prove Theorem 3.1.

Proof (Proof of Theorem 3.1). We now verify conditions B1-B5 from Theorem 8.1. It can be shown that, under Assumptions 2 and 3, $\phi(a)$ continuous for any $a \in \mathbb{R}$(see the proof of Lemma 5 and Theorem 1 in [21]), and therefore, in view of Assumptions 1–3, for each t, function $\boldsymbol{\gamma}(t+1, \cdot, \cdot)$ is measurable. Hence, condition B1 holds. Consider the filtration $\mathcal{H}_t$, $t = 1, 2, ...$, where $\mathcal{H}_t$ is the σ-algebra generated with random vectors $\boldsymbol{\nu}^s$, $s = 0, ..., t-1$. Then, the family of random vectors $\{\boldsymbol{\gamma}\left(t, \mathbf{x}, \omega\right)\}_{\mathbf{x} \in \mathbb{R}^d}$ is $\mathcal{H}_t$ measurable, zero-mean and independent of $\mathcal{H}_{t-1}$. Thus, condition B2 holds.

For B3, we need to prove that

$$\sup_{\mathbf{x}: \|\mathbf{x} - \mathbf{x}^\star\| \in (\epsilon, \frac{1}{\epsilon})} \langle \mathbf{r}(\mathbf{x}), \frac{\partial V}{\partial x}(\mathbf{x}) \rangle < 0, \text{ for any } \epsilon > 0, \tag{5.6}$$

where $\frac{\partial V}{\partial x}(\mathbf{x}) = \nabla f(\mathbf{x})$. Let us fix an $\epsilon > 0$. Consider arbitrary $\mathbf{x} \in \{\mathbf{y} \in \mathbb{R}^d : \; \|\mathbf{y} - \mathbf{x}^\star\| \in (\epsilon, \frac{1}{\epsilon})\}$. Then, we have:

$$\begin{aligned} \langle \mathbf{r}(\mathbf{x}), \frac{\partial V}{\partial x} \rangle &= -\boldsymbol{\phi}(\nabla f(\mathbf{x}))^\top (\nabla f(\mathbf{x})) \\ &= -\sum_{j=1}^{d} \phi([\nabla f(\mathbf{x})]_j)[\nabla f(\mathbf{x})]_j = -\sum_{j=1}^{d} |\phi([\nabla f(x)]_j)| \, |[\nabla f(x)]_j|, \end{aligned}$$

where the last inequality holds because ϕ is an odd function. Since $\|\mathbf{x} - \mathbf{x}^\star\| > \epsilon$ and $\|\nabla f(\mathbf{x})\|^2 > \frac{\mu}{2}\|\mathbf{x} - \mathbf{x}^\star\|^2$ (due to strong convexity of f), we have $\|\nabla f(\mathbf{x})\| > \sqrt{\frac{\mu}{2}}\epsilon$, where we recall that μ is the strong convexity constant of f. Therefore, there exists an index $i \in \{1, ..., d\}$ such that $|[\nabla f(\mathbf{x})]_i| > \frac{1}{d}\sqrt{\frac{\mu}{2}}\epsilon =: \epsilon'$. Next, because $\phi'(0) > 0$, and ϕ is continuous at 0 and is nondecreasing (by Lemma 5.1), we have that $|\phi(b)| > \delta$ for some $\delta = \delta(\epsilon) > 0$, for all $b \in (\epsilon, 1/\epsilon)$. Finally, we have that: $\leq -\epsilon'\delta(\epsilon)$, for any $\mathbf{x}$ such that $\|\mathbf{x} - \mathbf{x}^\star\| \in (\epsilon, \frac{1}{\epsilon})$, and therefore $\sup_{\mathbf{x}: \|\mathbf{x} - \mathbf{x}^\star\| \in (\epsilon, \frac{1}{\epsilon})}$ $\langle \mathbf{r}(x), \frac{\partial V}{\partial x} \rangle < 0$, hence verifying condition B3.

We next verify condition B4. Consider quantity $\mathbf{r}(\mathbf{x})$ in (5.3). By Lemma 5.1 and the fact that f has Lipschitz gradient and is strongly convex (Assumption 1), it follows that:

$$\|\mathbf{r}(\mathbf{x})\|^2 \leq C_1 + C_2 V(\mathbf{x}), \tag{5.7}$$

for some positive constants C_1 and C_2. Also, since

$$\|\boldsymbol{\gamma}(\mathbf{x}, t+1, \omega)\|^2 \leq 2\|\phi(\nabla f(\mathbf{x}))\|^2 + 2\|\Psi(\nabla f(\mathbf{x}) + \boldsymbol{\nu}^t)\|^2, \tag{5.8}$$

and it holds that either 1) Ψ is bounded or 2) $|\Psi(a)| \leq C_2\left(1 + |a|\right)$ and ν_i^t has a finite variance, we have:

$$\mathbb{E}\left[\|\boldsymbol{\gamma}(\mathbf{x}, t+1, \omega)\|^2\right] \leq C_3 + C_4 \, V(\mathbf{x}), \tag{5.9}$$

for some positive constants C_3, C_4. Now, we finally have:

$$\|\mathbf{r}(\mathbf{x})\|^2 + \mathbb{E}\left[\|\boldsymbol{\gamma}(\mathbf{x}, t+1, \omega)\|^2\right] \le C_5 + C_6\, V(\mathbf{x}), \tag{5.10}$$

for some positive constants C_5, C_6, and hence condition B4 holds. Condition B5 holds by the choice of the step size sequence $\{\alpha_t\}$ in the Theorem statement. Summarizing, all conditions B1-B5 hold true, and hence $\mathbf{x}^t \to \mathbf{x}^\star$, almost surely. □

We continue by proving Theorem 3.3.

Proof (Proof of Theorem 3.3). We prove the Theorem by verifying conditions C1-C5 in Theorem 8.1. To verify condition C1, consider $\mathbf{r}(\mathbf{x})$ in (5.3) and note that, using the mean value theorem, it can be expressed as follows:

$$\begin{aligned} \mathbf{r}(\mathbf{x}) &= -\phi(\nabla f(\mathbf{x}) - \nabla f(\mathbf{x}^\star)) \\ &= -\phi\left(\underbrace{\left[\int_0^1 \nabla^2 f(\mathbf{x}^\star + t(\mathbf{x} - \mathbf{x}^\star)dt\right]}_{\mathbf{H}_t} (\mathbf{x} - \mathbf{x}^\star) \right) \\ &= -\phi\left(\mathbf{H}_t(\mathbf{x} - \mathbf{x}^*)\right) = -\phi'(0)\nabla^2 f(\mathbf{x}^\star)(\mathbf{x} - \mathbf{x}^\star) + \delta(\mathbf{x}), \end{aligned} \tag{5.11}$$

where $\lim_{\mathbf{x}\to\mathbf{x}^\star} \frac{\|\delta(x)\|}{\|x - x^*\|} = 0$. Hence, in the notation of Theorem 8.1, we have that $\mathbf{B} = -\phi'(0)\nabla^2 f(\mathbf{x}^\star)$. Hence, C1 holds. Also, C2 holds, by assumptions of Theorem 3.3. Now, we consider C3, which requires that the matrix $\mathbf{\Sigma} = a\,\mathbf{B} + \frac{1}{2}\mathbf{I}$ is stable, where $\mathbf{B} = -\phi'(0)\nabla^2 f(\mathbf{x}^\star)$. Note that $\mathbf{\Sigma} = \frac{1}{2}\mathbf{I} - a\,\phi'(0)\nabla^2 f(\mathbf{x}^\star)$. Clearly, $\mathbf{\Sigma}$ is stable for a small enough a, because the matrix $\phi'(0)\nabla^2 f(\mathbf{x}^\star)$ is positive definite. More precisely, $\mathbf{\Sigma}$ is stable for $a > 1/(2\mu)$. Therefore, condition C3 holds, provided that $a > 1/(2\mu)$. We next consider condition C4. In the notation of Theorem 8.1, consider the following quantity:

$$\mathbf{A}(t, \mathbf{x}) := \mathbb{E}\left[\boldsymbol{\gamma}(t+1, \mathbf{x}, \omega)\boldsymbol{\gamma}(t+1, \mathbf{x}, \omega)^\top\right] \tag{5.12}$$

$$= \mathbb{E}\left[\left(\phi(\nabla f(\mathbf{x})) - \Psi(\nabla f(\mathbf{x}) + \boldsymbol{\nu}^t)\right)\left((\phi(\nabla f(\mathbf{x})) - \Psi(\nabla f(\mathbf{x}) + \boldsymbol{\nu}^t)\right)^\top\right] \tag{5.13}$$

Now, because $\nabla f(\mathbf{x}^\star) = 0$; $\boldsymbol{\phi}(0) = 0$, and the entries of $\boldsymbol{\nu}^t$ are mutually independent, with pdf $p(u)$, we have that:

$$\lim_{t\to\infty, \mathbf{x}\to\mathbf{x}^\star} \mathbf{A}(t, \mathbf{x}) =: \mathcal{S}_0 = \mathbb{E}\left[\Psi(\boldsymbol{\nu}^t)\cdot\Psi(\boldsymbol{\nu}^t)^\top\right] = \sigma_\Psi^2 \cdot \mathbf{I}, \tag{5.14}$$

where $\sigma_\Psi^2 = \int |\Psi(a)|^2 p(a) da$. Therefore, condition C4 holds. We finally verify condition C5. We follow the arguments analogous to those in Theorem 10 in [29]. Condition C5 means uniform integrability of the family $\{\|\boldsymbol{\gamma}(t+1, \mathbf{x}, \omega)\|^2\}_{t=0,1,\ldots,\,\|\mathbf{x}-\mathbf{x}^\star\|<\epsilon}$. We have:

$$\|\boldsymbol{\gamma}(t+1, \mathbf{x}, \omega)\|^2 \le 2\|\phi(\nabla f(\mathbf{x}))\|^2 + 2\|\psi(\nabla f(\mathbf{x}) + \boldsymbol{\nu}^t)\|^2. \tag{5.15}$$

We consider separately the cases when condition 2. or condition 3. hold in Assumption 3. If condition 2. holds, then:

$$\|\boldsymbol{\gamma}(t+1, \mathbf{x}, \omega)\|^2 \;\le\; C_7 + C_8\|\mathbf{x}^t - \mathbf{x}^\star\|^2 + C_9\|\boldsymbol{\nu}^t\|^2 \tag{5.16}$$

$$\le\; C_7 + C_8\,\epsilon^2 + C_9\|\boldsymbol{\nu}^t\|^2, \tag{5.17}$$

for some positive constants C_7, C_8, C_9. Consider next the family $\{\widetilde{\gamma}(t+1, \mathbf{x}, \omega)\}_{t=0,1,\ldots,\|\mathbf{x}-\mathbf{x}^\star\|<\epsilon}$, with

$$\widetilde{\gamma}(t+1, \mathbf{x}, \omega) = C_7 + C_8\,\epsilon^2 + C_9\|\nu^t\|^2. \tag{5.18}$$

The family $\{\widetilde{\gamma}(t+1, \mathbf{x}, \omega)\}_{t=0,1,\ldots,\|\mathbf{x}-\mathbf{x}^\star\|<\epsilon}$ is i.i.d. and hence it is uniformly integrable. The family $\{\|\boldsymbol{\gamma}(t+1, \mathbf{x}, \omega)\|^2\}_{t=0,1,\ldots,\|\mathbf{x}-\mathbf{x}^\star\|<\epsilon}$ is dominated by
$\{\widetilde{\gamma}(t+1, \mathbf{x}, \omega)\}_{t=0,1,\ldots,\|\mathbf{x}-\mathbf{x}^\star\|<\epsilon}$ that is uniformly integrable, and hence
$\{\|\boldsymbol{\gamma}(t+1, \mathbf{x}, \omega)\|^2\}_{t=0,1,\ldots,\|\mathbf{x}-\mathbf{x}^\star\|<\epsilon}$ is also uniformly integrable. Hence, C5 holds under condition 2. of Assumption 3. Now, let condition 3. in Assumption 3) hold. Then:

$$\|\boldsymbol{\gamma}(t+1, \mathbf{x}, \omega)\|^2 \;\le\; C_{10} + C_{11}\|\mathbf{x} - \mathbf{x}^\star\|^2 \tag{5.19}$$

$$\le\; C_{10} + C_{11}\,\epsilon^2. \tag{5.20}$$

Consider the family $\{\widehat{\gamma}(t+1, \mathbf{x}, \omega)\}_{t=0,1,\ldots,\|\mathbf{x}-\mathbf{x}^\star\|<\epsilon}$, with

$$\widehat{\gamma}(t+1, \mathbf{x}, \omega) = C_{10} + C_{11}\,\epsilon^2. \tag{5.21}$$

The family $\{\widehat{\gamma}(t+1, \mathbf{x}, \omega)\}_{t=0,1,\ldots,\|\mathbf{x}-\mathbf{x}^\star\|<\epsilon}$ is uniformly integrable, and condition C5 is verified analogously to the previous case.

Summarizing, we have established that all conditions C1-C5 of Theorem 8.1 hold true, thus the proof of Theorem 3.3. □.

5.2 MSE analysis: Proof of Theorem 3.2

We start with the following Lemma that shows that, with algorithm (2.3), almost surely, $\nabla f(\mathbf{x}^t)$ can be at most $O(\ln(t))$.

Lemma 5.2 *Let Assumptions 1, 2, and Assumption 3 with condition 3. hold (the nonlinearity with bounded outputs case). Then, for each* $t = 1, 2, ...$, *we have:*

$$\|\nabla f(\mathbf{x}^t)\| \leq G_t := L \left(a\, C_2 \sqrt{d}\, \frac{t^{1-\delta}}{1-\delta} + \|\mathbf{x}^0\| + \|\mathbf{x}^\star\| \right). \tag{5.22}$$

Proof. Consider (2.3). Because the output of each component nonlinearity Ψ is bounded in the absolute value by C_2 (Assumption 3), we have, for each $t \geq 1$:

$$\begin{aligned} \|\mathbf{x}^t\| &\leq \|\mathbf{x}^0\| + a\sqrt{d}\, C_2 \sum_{s=0}^{t-1} \frac{1}{(s+1)^\delta} \\ &\leq \|\mathbf{x}^0\| + a\, C_2 \sqrt{d} \left(\frac{t^{1-\delta}}{1-\delta} \right). \end{aligned} \tag{5.23}$$

Next, because ∇f is L-Lipschitz, we have: $\|\nabla f(\mathbf{x}^t)\| \leq L\, \|\mathbf{x}^t - \mathbf{x}^\star\|$. Applying this inequality to (5.23), the result follows. □

We will also make use of the following Lemma.

Lemma 5.3 *There exists a positive constant* ξ *such that, for any* $t = 1, 2, ...$, *there holds, almost surely, for each* $j = 1, ..., d$, *that:*

$$|\phi([\nabla f(\mathbf{x}^t)]_j)| \geq |[\nabla f(\mathbf{x}^t)]_j| \, \frac{\phi'(0)\, \xi}{2\, G_t},$$

where G_t *is defined in* (5.22).

Proof. Consider function ϕ in (3.1). By Lemma 5.1, we have that $\phi'(0) > 0$ and ϕ is continuous at zero.[6] Therefore, there exists a positive constant ξ such that:

$$\phi(a) \geq \frac{\phi'(0)}{2}\, a,$$

for any $a \in [0, \xi)$. Now, because ϕ is non-decreasing (by Lemma 5.1), it holds for any $a' > \xi$ that

$$\phi(a) \geq \frac{\phi'(0)\, \xi\, a}{2\, a'}, \text{ for any } a \in [0, a'). \tag{5.24}$$

Consider now $\nabla f(\mathbf{x}^t)$. By Lemma 5.2, we have that $\|\nabla f(\mathbf{x}^t)\| \leq G_t$, a.s., and so, for any $j = 1, ..., d$, $|[\nabla f(\mathbf{x}^t)]_j| \leq G_t$. Therefore, in view of (5.24), setting $a' = G_t$, the Lemma follows. □

We are now ready to prove Theorem 3.2.

Proof (Proof of Theorem 3.2). Consider algorithm (2.3). By the Lipschitz property of ∇f, we have, for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, that:

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

and so, almost surely:

$$\begin{aligned} f(\mathbf{x}^{t+1}) \leq f(\mathbf{x}^t) &+ \left(\nabla f(\mathbf{x}^t)\right)^\top (-\alpha_t \Psi(\nabla f(\mathbf{x}^t) + \boldsymbol{\nu}^t)) \\ &+ \frac{L}{2} \alpha_t^2 \|\Psi(\nabla f(\mathbf{x}^t) + \boldsymbol{\nu}^t)\|^2. \end{aligned} \tag{5.25}$$

Next, letting $\boldsymbol{\eta}^t = \Psi(\nabla f(\mathbf{x}^t) + \boldsymbol{\nu}^t) - \boldsymbol{\phi}(\nabla f(\mathbf{x}^t))$, and using the fact that Ψ has bounded outputs, we obtain:

$$f(\mathbf{x}^{t+1}) \leq f(\mathbf{x}^t) + \left(\nabla f(\mathbf{x}^t)\right)^\top (-\alpha_t \boldsymbol{\phi}(\nabla f(\mathbf{x}^t))) + \frac{L}{2} \alpha_t^2\, d^2 C_{12}^2 - \alpha_t \left(\nabla f(\mathbf{x}^t)\right)^\top \boldsymbol{\eta}^t, \text{ a.s.,} \tag{5.26}$$

[6] As ϕ is an odd function, for simplicity, in the proof we consider only nonnegative arguments of ϕ, while analogous analysis applies for negative arguments of ϕ.

for some positive constant C_{12}. Let $\mathcal{F}_t$ be the history of (2.3) up to iteration t. Then, taking conditional expectation, and using the fact that $\mathbb{E}[\boldsymbol{\eta}^t \,|\, \mathcal{F}_t] = 0$, we get that, almost surely:

$$\mathbb{E}[f(\mathbf{x}^{t+1}) \,|\, \mathcal{F}_t] \leq f(\mathbf{x}^t) - \alpha_t \left(\nabla f(\mathbf{x}^t)\right)^\top \boldsymbol{\phi}(\nabla f(\mathbf{x}^t)) + \frac{L}{2}\alpha_t^2 \, d^2 \, C_{12}^2. \tag{5.27}$$

Next, using Lemma 5.3, the fact that $\alpha_t = a/(t+1)^\delta$, and the fact that $G_t = O(t^\epsilon)$, for $\epsilon > 0$ we obtain that. a.s.:

$$\mathbb{E}[f(\mathbf{x}^{t+1}) \,|\, \mathcal{F}_t] \leq f(x^t) - \frac{c'}{(t+1)}\|\nabla f(\mathbf{x}^t)\|^2 + \frac{L}{2}\frac{a^2 \, d^2 \, C_2^2}{(t+1)^{2\delta}}, \tag{5.28}$$

where $c' = \frac{a\,(1-\delta)\xi\,\phi'(0)}{2\,L\,(a\,C_2\,\sqrt{d}+\|\mathbf{x}^0\|+\|\mathbf{x}^\star\|)}$. Next, by strong convexity of f, we have that $\|\nabla f(\mathbf{x}^t) - \nabla f(\mathbf{x}^\star)\|^2 \geq 2\,\mu\,(f(\mathbf{x}^t) - f^\star)$. Using the latter inequality, subtracting $f^\star$ from both sides of the inequality, taking expectation, and applying Theorem 8.2, claims (2) and (3), we get the MSE rate result under condition 3. in Assumption 5).

We next consider the case when condition 2. in Assumption 5) holds, and we have the bounded second moment of $\boldsymbol{\nu}^t$. Following analogous arguments as in the first part of the proof, it can be shown that, a.s.:

$$\mathbb{E}[f(\mathbf{x}^{t+1}) \,|\, \mathcal{F}_t] \leq f(\mathbf{x}^t) - \alpha_t \, \boldsymbol{\phi}(\nabla f(\mathbf{x}^t))^\top \nabla f(\mathbf{x}^t) + \frac{L}{2}\alpha_t^2 \left(C_{13} + C_{14}\|\boldsymbol{\nu}^t\|^2\right), \tag{5.29}$$

for some positive constants C_{13}, C_{14}. Next, because $\inf_{a \neq 0} \frac{|\phi(a)|}{|a|} > 0$, we have that $\phi(\nabla f(\mathbf{x}^t))^\top \nabla f(\mathbf{x}^t) \geq C_{15}\,\|\nabla f(\mathbf{x}^t)\|^2$, for some constant $C_{15} > 0$. Using the latter bound in (5.29), subtracting $f^\star$ from both sides of the inequality, taking expectation, and applying Theorem 8.2, claim (1) and (2), the result follows. □

6 Intermediate results and proofs: Joint nonlinearities

Subsection 6.1 provides the required intermediate results, while Subsection 6.2 proves Theorem 4.1.

6.1 Intermediate results: Joint nonlinearities

Recall function $\mathcal{N} : \mathbb{R}_+ \mapsto \mathbb{R}_+$ in Assumption 5. We first state and prove the following Lemma on the properties of function $\mathcal{N}$.

Lemma 6.1 *Under Assumption 5, for any* $\mathbf{x}, \mathbf{u} \in \mathbb{R}^d$*, such that* $\|\mathbf{u}\| > \|\mathbf{x}\|$*, there holds:*

$$|\mathcal{N}(\|\mathbf{x}+\mathbf{u}\|) - \mathcal{N}(\|\mathbf{x}-\mathbf{u}\|)| \leq \frac{\|\mathbf{x}\|}{\|\mathbf{u}\|}\left[\mathcal{N}(\|\mathbf{x}+\mathbf{u}\|) + \mathcal{N}(\|\mathbf{x}-\mathbf{u}\|)\right]. \tag{6.1}$$

Proof. Fix a pair $\mathbf{x}, \mathbf{u} \in \mathbb{R}^d$, such that $\|\mathbf{u}\| > \|\mathbf{x}\|$, and assume without loss of generality that $\mathcal{N}(\|\mathbf{x}+\mathbf{u}\|) \geq \mathcal{N}(\|\mathbf{x}-\mathbf{u}\|)$. Then, (6.1) is equivalent to:

$$(\|\mathbf{u}\| - \|\mathbf{x}\|)\mathcal{N}(\|\mathbf{x}+\mathbf{u}\|) \leq (\|\mathbf{u}\| + \|\mathbf{x}\|)\mathcal{N}(\|\mathbf{x}-\mathbf{u}\|). \tag{6.2}$$

Denote by $\rho = \|\mathbf{u}\|$. Notice that:

$$\rho - \|\mathbf{x}\| \leq \|\mathbf{x}+\mathbf{u}\| \leq \|\mathbf{x}\| + \|\mathbf{u}\| = \|\mathbf{x}\| + \rho, \tag{6.3}$$

and similarly,

$$\rho + \|\mathbf{x}\| \geq \|\mathbf{x}-\mathbf{u}\| \geq \rho - \|\mathbf{x}\|.$$

As $\mathcal{N}$ is non-increasing, it follows that:

$$\mathcal{N}(\|\mathbf{x}+\mathbf{u}\|) \leq \mathcal{N}(\rho - \|\mathbf{x}\|), \;\; \mathcal{N}(\|\mathbf{x}-\mathbf{u}\|) \geq \mathcal{N}(\rho + \|\mathbf{x}\|).$$

Now, we have:

$$(\|\mathbf{u}\| - \|\mathbf{x}\|)\mathcal{N}(\|\mathbf{x}+\mathbf{u}\|) \leq (\rho - \|\mathbf{x}\|)\mathcal{N}(\rho - \|\mathbf{x}\|), \tag{6.4}$$

and similarly:

$$(\|\mathbf{u}\| + \|\mathbf{x}\|)\mathcal{N}(\|\mathbf{x}-\mathbf{u}\|) \geq (\rho + \|\mathbf{x}\|)\mathcal{N}(\rho + \|\mathbf{x}\|). \tag{6.5}$$

By assumption, function $a \mapsto a\mathcal{N}(a)$, $a > 0$, is non-decreasing, and so $(\rho - \|\mathbf{x}\|)\mathcal{N}(\rho - \|\mathbf{x}\|) \leq (\rho + \|\mathbf{x}\|)\mathcal{N}(\|\mathbf{x}\| + \rho)$. Thus, combining (6.4) and (6.5), we have that (6.2) holds, which is in turn equivalent to the claim of the Lemma.

We now define map $\boldsymbol{\phi} : \mathbb{R}^d \mapsto \mathbb{R}^d$, as follows. For a fixed (deterministic) point $\mathbf{w} \in \mathbb{R}^d$, we let:

$$\boldsymbol{\phi}(\mathbf{w}) = \int \boldsymbol{\Psi}(\mathbf{w}+\mathbf{u})p(\mathbf{u})d\mathbf{u} = \mathbb{E}[\boldsymbol{\Psi}(\mathbf{w}+\boldsymbol{\nu}^0)], \tag{6.6}$$

where the expectation is taken with respect to the joint pdf of the gradient noise at any iteration t, e.g., $t = 0$. The map $\boldsymbol{\phi} : \mathbb{R}^d \mapsto \mathbb{R}^d$ is, abusing notation, a counterpart of the component-wise map $\phi : \mathbb{R} \mapsto \mathbb{R}$ in (3.1). We have the following Lemma.

Lemma 6.2 *The following holds:*

$$\boldsymbol{\phi}(\mathbf{x})^\top \mathbf{x} \geq 2(1-\kappa)\|\mathbf{x}\|^2 \int\limits_{\mathcal{J}(\mathbf{x})} \mathcal{N}(\|\mathbf{x}\| + \|\mathbf{u}\|)p(\mathbf{u})d\mathbf{u}, \tag{6.7}$$

where $\mathcal{J}(\mathbf{x}) = \{\mathbf{u} : \frac{\mathbf{u}^\top \mathbf{x}}{\|\mathbf{u}\|\|\mathbf{x}\|} \in [0, \kappa]\}$*, and* κ *is any constant in the interval* $(0, 1)$.

Proof. Let us fix arbitrary $\mathbf{x} \in \mathbb{R}^d, \mathbf{x} \neq 0$. As $\mathbf{\Psi}(\mathbf{a}) = \mathbf{a}\mathcal{N}(\|\mathbf{a}\|)$, we have:

$$\boldsymbol{\phi}(\mathbf{x})^\top \mathbf{x} = \int_{\mathbf{u}\in\mathbb{R}^d} \underbrace{(\mathbf{x}+\mathbf{u})^\top \mathbf{x}\mathcal{N}(\|\mathbf{x}+\mathbf{u}\|)}_{:=\mathcal{M}(\mathbf{x},\mathbf{u})} p(\mathbf{u})d\mathbf{u} \tag{6.8}$$

$$= \int_{J_1(\mathbf{x})=\{\mathbf{u}:\, \mathbf{u}^\top\mathbf{x}\geq 0\}} \mathcal{M}(\mathbf{x},\mathbf{u})p(\mathbf{u})d\mathbf{u} \tag{6.9}$$

$$+ \int_{J_2(\mathbf{x})=\{\mathbf{u}:\, \mathbf{u}^\top\mathbf{x}< 0\}} \mathcal{M}(\mathbf{x},\mathbf{u})p(\mathbf{u})d\mathbf{u}. \tag{6.10}$$

Note also that there holds:

$$\mathcal{M}(\mathbf{x},\mathbf{u}) = (\|\mathbf{x}\|^2 + \mathbf{u}^\top\mathbf{x})\mathcal{N}(\|\mathbf{x}+\mathbf{u}\|).$$

Similarly,

$$\mathcal{M}(\mathbf{x},-\mathbf{u}) = (\|\mathbf{x}\|^2 - \mathbf{u}^\top\mathbf{x})\mathcal{N}(\|\mathbf{x}-\mathbf{u}\|).$$

Therefore, using the fact that $p(\mathbf{u}) = p(-\mathbf{u})$, for all $u \in \mathbb{R}^d$, we obtain:

$$\boldsymbol{\phi}(\mathbf{x})^\top \mathbf{x} = \int_{J_1(\mathbf{x})} \mathcal{M}_2(\mathbf{x},\mathbf{u})p(\mathbf{u})d\mathbf{u}, \tag{6.11}$$

where $\mathcal{M}_2(\mathbf{x},\mathbf{u}) = [(\|\mathbf{x}\|^2 + \mathbf{u}^\top\mathbf{x})\mathcal{N}(\|\mathbf{x}+\mathbf{u}\|) + (\|\mathbf{x}\|^2 - \mathbf{u}^\top\mathbf{x})\mathcal{N}(\|\mathbf{x}-\mathbf{u}\|)]$. There holds:

$$\mathcal{M}_2(\mathbf{x},\mathbf{u}) \geq \|\mathbf{x}\|^2[\mathcal{N}(\|\mathbf{x}+\mathbf{u}\|) + \mathcal{N}(\|\mathbf{x}-\mathbf{u}\|)] - \|\mathbf{u}\|\|\mathbf{x}\||\mathcal{N}(\|\mathbf{x}+\mathbf{u}\|) - \mathcal{N}(\|\mathbf{x}-\mathbf{u}\|)|. \tag{6.12}$$

Since $\mathbf{u} \in J_1(\mathbf{x})$, there holds $\|\mathbf{x}+\mathbf{u}\| \geq \|\mathbf{x}-\mathbf{u}\|$. Now, using Lemma 6.1, we have:

$$\mathcal{M}_2(\mathbf{x},\mathbf{u}) \geq \|\mathbf{x}\|^2[\mathcal{N}(\|\mathbf{x}+\mathbf{u}\|) + \mathcal{N}(\|\mathbf{x}-\mathbf{u}\|)] - \quad \|\mathbf{u}\|\|\mathbf{x}\|\frac{\|\mathbf{x}\|}{\|\mathbf{u}\|}|\mathcal{N}(\|\mathbf{x}+\mathbf{u}\|) + \mathcal{N}(\|\mathbf{x}-\mathbf{u}\|)| = 0. \tag{6.13}$$

Therefore, we have:

$$\mathcal{M}_2(\mathbf{x},\mathbf{u}) \geq 0, \text{ for any } \mathbf{u} \in J_1(\mathbf{x}),\ \|\mathbf{u}\| > \|\mathbf{x}\|. \tag{6.14}$$

Now, consider $\mathcal{J}(\mathbf{x}) = \{\mathbf{u} \in \mathbb{R}^d :\ \mathbf{u}^\top\mathbf{x} \geq 0,\ \frac{\mathbf{u}^\top\mathbf{x}}{\|\mathbf{u}\|\|\mathbf{x}\|} \in [0,\kappa]\}$, where $\kappa \in (0, 1)$. Let us consider $\mathbf{u} \in \mathcal{J}(\mathbf{x})$ such that $\|\mathbf{u}\| > \|\mathbf{x}\|$. Then, using Lemma 6.1, we get:

$$\begin{aligned}\mathcal{M}_2(\mathbf{x},\mathbf{u}) &\geq \|\mathbf{x}\|^2[\mathcal{N}(\|\mathbf{x}+\mathbf{u}\|) + \mathcal{N}(\|\mathbf{x}-\mathbf{u}\|)] \\ &\quad - \|\mathbf{u}\|\|\mathbf{x}\|\kappa\underbrace{|\mathcal{N}(\|\mathbf{x}+\mathbf{u}\|) - \mathcal{N}(\|\mathbf{x}-\mathbf{u}\|)|} \\ &\geq (1-\kappa)\|\mathbf{x}\|^2(\mathcal{N}(\|\mathbf{x}+\mathbf{u}\|) + \mathcal{N}(\|\mathbf{x}-\mathbf{u}\|)).\end{aligned} \tag{6.15}$$

Now, consider $\mathbf{u} \in \mathcal{J}(\mathbf{x})$ such that $\|\mathbf{u}\| \leq \|\mathbf{x}\|$. Then, there holds:

$$\begin{aligned}\mathcal{M}_2(\mathbf{x},\mathbf{u}) &\geq \|\mathbf{x}\|^2[\mathcal{N}(\|\mathbf{x}+\mathbf{u}\|) + \mathcal{N}(\|\mathbf{x}-\mathbf{u}\|)] - \\ &\quad \underbrace{\|\mathbf{u}\|}_{\leq\|\mathbf{x}\|}\ \|\mathbf{x}\|\kappa|\underbrace{\mathcal{N}(\|\mathbf{x}+\mathbf{u}\|)}_{\geq 0} + \underbrace{\mathcal{N}(\|\mathbf{x}-\mathbf{u}\|)}_{\geq 0}| \\ &\geq (1-\kappa)\|\mathbf{x}\|^2(\mathcal{N}(\|\mathbf{x}+\mathbf{u}\|) + \mathcal{N}(\|\mathbf{x}-\mathbf{u}\|)).\end{aligned} \tag{6.16}$$

where the last inequality holds due to the fact that $|a - b| \leq |a| + |b|$, for any $a, b \in \mathbb{R}$. Now, we have:

$$\begin{aligned}\mathcal{M}_2(\mathbf{x},\mathbf{u}) &\geq (1-\kappa)\|\mathbf{x}\|^2(\underbrace{\mathcal{N}(\|\mathbf{x}+\mathbf{u}\|)}_{\geq\mathcal{N}(\|x\|+\|u\|)} + \underbrace{\mathcal{N}(\|\mathbf{x}-\mathbf{u}\|)}_{\geq\mathcal{N}(\|\mathbf{x}\|+\|\mathbf{u}\|)}) \\ &\geq 2(1-\kappa)\|\mathbf{x}\|^2\mathcal{N}(\|\mathbf{x}\| + \|\mathbf{u}\|), \text{ for any } \mathbf{u} \in \mathcal{J}(\mathbf{x}).\end{aligned} \tag{6.17}$$

Combining (6.15) and (6.17), we finally get:

$$\begin{aligned}\boldsymbol{\phi}(\mathbf{x})^\top \mathbf{x} &\geq \int\limits_{\mathcal{J}(\mathbf{x})} 2(1-\kappa)\|\mathbf{x}\|^2 \mathcal{N}(\|\mathbf{x}\| + \|\mathbf{u}\|) p(\mathbf{u}) d\mathbf{u} \\ &= 2(1-\kappa)\|\mathbf{x}\|^2 \int\limits_{\mathcal{J}(\mathbf{x})} \mathcal{N}(\|\mathbf{x}\| + \|\mathbf{u}\|) p(\mathbf{u}) d\mathbf{u}.\end{aligned} \tag{6.18}$$

Lemma 6.3 *Let Assumptions 1, 4, and Assumption 5 with condition 3. hold (the nonlinearity with bounded outputs case). Then, for each $t = 1, 2, ...,$ we have:*

$$\|\nabla f(\mathbf{x}^t)\| \leq G'_t := L\,(\,a\,C'_2\,\frac{t^{1-\delta}}{1-\delta} + \|\mathbf{x}^0\| + \|\mathbf{x}^\star\|\,). \tag{6.19}$$

Proof. The proof is analogous to the proof of Lemma 5.2.

6.2 Proof of Theorem 4.1: Joint nonlinearities

Proof. We first consider the case with bounded nonlinearity (condition 3. in Assumption 5). Analogously to the proof of 3.2, it can be shown that, a.s.:

$$\mathbb{E}[f(\mathbf{x}^{t+1})\,|\,\mathcal{F}_t] \leq f(\mathbf{x}^t) - \alpha_t\,\boldsymbol{\phi}(\nabla f(\mathbf{x}^t))^\top \nabla f(\mathbf{x}^t) + \alpha_t^2\,C_{17}, \tag{6.20}$$

for some positive constant C_{17}. By Lemma 6.2, there holds, for $\mathbf{a} := \nabla f(\mathbf{x}^t)$, a.s.:

$$\big(\boldsymbol{\phi}(\mathbf{a})\big)^\top \mathbf{a} \geq 2(1-\kappa)\|\mathbf{a}\|^2 \int_{\mathcal{J}} \mathcal{N}(\|\mathbf{a}\| + \|\mathbf{u}\|) p(\mathbf{u}) d\mathbf{u}, \tag{6.21}$$

where we recall $\mathcal{J} = \{\mathbf{u} : \frac{\mathbf{u}^\top \mathbf{a}}{\|\mathbf{u}\|\|\mathbf{a}\|} \in [0, \kappa]\}$, where $\kappa \in (0,1)$ is a constant. Note that, as $a \mapsto a\,\mathcal{N}(a)$ is non-decreasing, $\mathcal{N}$ satisfies: $\mathcal{N}(b) \geq \min\left(\frac{\mathcal{N}(1)}{b}, \mathcal{N}(1)\right)$ for any $b > 0$. Consider constant B_0 in condition 2. of Assumption 4. Then, for all $\mathbf{u}$ such that $\|\mathbf{u}\| \leq B_0$, there holds $\mathcal{N}(\|\mathbf{a}\| + \|\mathbf{u}\|) \geq \min\left(\frac{\mathcal{N}(1)}{\|\mathbf{a}\|+B_0}, \mathcal{N}(1)\right)$. Therefore, we have that, almost surely, for sufficiently large t:

$$\|\nabla f(\mathbf{x}^t)\|^2 \int_{J_4} \mathcal{N}(\|\nabla f(\mathbf{x}^t)\| + \|\mathbf{u}\|) p(\mathbf{u}) d\mathbf{u} \geq C_{18}\,\frac{\|\nabla f(\mathbf{x}^t)\|^2}{G'_t + B_0},$$

for some positive constant C_{18}. Here, $J_4 = \{u \in \mathbb{R}^d : \frac{\mathbf{u}^\top \nabla f(\mathbf{x}^t)}{\|\mathbf{u}\|\|\nabla f(\mathbf{x}^t)\|} \in [0, \kappa],\ \|\mathbf{u}\| \leq B_0\}$. Combining the last bound with Lemmas 6.2 and 6.3, in view of condition 2. in Assumption 4, we obtain that, for sufficiently large t, a.s.:

$$(\boldsymbol{\phi}(\nabla f(\mathbf{x}^t)))^\top \nabla f(\mathbf{x}^t) \geq C_{19}\,\frac{\|\nabla f(\mathbf{x}^t)\|^2}{B_0 + G'_t}, \tag{6.22}$$

where the positive constant C_{19} can be taken as $C_{19} = 2(1-\kappa)\lambda(\kappa)\mathcal{N}(1)$.

Applying the bound (6.22) to (6.20) we obtain an equivalent to 5.29. Therein, c' in 5.29 is replaced with a positive constant c'' that can be taken as $c'' = \frac{2\,a\,(1-\kappa)\lambda(\kappa)(1-\delta)\mathcal{N}(1)}{L\,(\,a\,C'_2 + \|\mathbf{x}^0\| + \|\mathbf{x}^\star\|\,) + B_0)}$. We now proceed analogously to the proof of Theorem 3.2, by applying claims (2) and (3) of Theorem 8.2. The result for the bounded nonlinearity $\boldsymbol{\Psi}$ follows, with the rate ζ being any positive number less than

$$\min\left\{2\delta - 1,\ \frac{2\,a\,\mu\,(1-\kappa)\lambda(\kappa)(1-\delta)\mathcal{N}(1)}{L\,(\,a\,C'_2 + \|\mathbf{x}^0\| + \|\mathbf{x}^\star\|\,) + B_0}\right\}. \tag{6.23}$$

We now prove the alternative case, for the nonlinearity $\boldsymbol{\Psi}$ with unbounded outputs and finite second moment of $\boldsymbol{\nu}^t$. We have that $\inf_{\mathbf{x}\neq 0} \frac{\|\boldsymbol{\Psi}(\mathbf{x})\|}{\|\mathbf{x}\|} > 0$. This is equivalent to saying that $\mathcal{N}$ is lower-bounded by a positive constant, i.e., $\mathcal{N}(a) \geq C_{20}$, for each a, for some constant $C_{20} > 0$. Then, it follows that, a.s.:

$$(\boldsymbol{\phi}(\nabla f(\mathbf{x}^t)))^\top \nabla f(\mathbf{x}^t) \geq C_{21}\,\|\nabla f(\mathbf{x}^t)\|^2, \tag{6.24}$$

for some positive constant C_{21}. The proof then proceeds analogously to the proof of Theorem 3.2 by applying the appropriate variant of Theorem 8.2.

It remains to prove a.s. convergence of (2.3). We do so again by verifying conditions B1-B5 in Theorem 8.1. Algorithm (2.3) admits again the representation in Theorem 8.1 with

$$\mathbf{r}(\mathbf{x}) = -\boldsymbol{\phi}(\nabla f(\mathbf{x})) \tag{6.25}$$

$$\boldsymbol{\gamma}(t+1, \mathbf{x}, \omega) = \boldsymbol{\phi}(\nabla f(\mathbf{x})) - \boldsymbol{\Psi}(\nabla f(\mathbf{x}) + \boldsymbol{\nu}^t). \tag{6.26}$$

Conditions B1, B2 clearly hold. Condition B3 follows from Lemma 6.2. Condition B4 holds analogously to the proof of Theorem 3.1. Finally, condition B5 follows from the definition of the step-size sequence in Theorem 4.1. Thus, the result. □

7 Experiments

In order to benchmark the proposed nonlinear SGD framework, we consider `Heart`, `Diabetes` and `Australian` datasets from the LibSVM library [30]. We consider the logistic regression loss function for binary classification, see, e.g., [22], where function f in Eq. (2.1) is the empirical loss, i.e., the sum of the logistic losses across all data points in a given dataset.

As it has been studied in [22] (see Figure 2 in [22]), we have, near the solution $\mathbf{x}^\star$, the following behavior with respect to gradient noise. (See also [22] for details how the gradient noise is evaluated in Figure 2 therein.) With the HEART dataset, tails of stochastic gradients are not heavy. On the other hand, for DIABETES and AUSTRALIAN datasets, the gradient noise has outliers and exhibits a heavy-tail behavior.

We consider three different nonlinearities to demonstrate the effectiveness of our nonlinear framework, namely, `tanh` (hyperbolic tangent), `sign` and a bi-level customization of `sign` with $\Psi(x) = -1, -0.5, 0.5, 1$, for $x \in (-\infty, -0.5], (-0.5, 0], (0, 0.5], (0.5, \infty]$, respectively (`nonlinear-quantizer` in figures). Note that the `tanh` function may be considered a smooth approximation of `sign`. We benchmark the above methods against the linear SGD, clipped-SGD and SSTM along with a clipped version of SSTM from [22]. For each of the methods, we use batch sizes of 50, 100 and 20 for the `Australian`, `Diabetes` and `Heart` datasets, respectively. We also consider clipped-SGD with periodically decreasing clipping level (`d-clipped-SGD` in Figures) as a baseline as introduced in [22]. This method starts with some initial clipping level and after every l epochs the clipping level is multiplied by some constant $c \in (0, 1)$. The step sizes α_t (learning rates) for each method from our framework were tuned after an experimentation. The learning rates for the baselines, i.e., SGD, clipped-SGD, SSTM and clipped-SSTM are also tuned and are selected to be as in [22]. In more detail, the learning rates for the proposed methods are of the form $a/(b\,(t+1) + L)$, where we recall that t is the iteration counter, L is the smoothness constant of ∇f, and parameters a, b are tuned via grid search. The value of a is chosen to be 1.0, 1.5 and 5.0, respectively, for `Heart`, `Diabetes` and `Australian` and for all the three non-linearities. The value of b is chosen to be 0.001, 7.0 and 7.0 respectively for `Australian`, `Heart` and `Diabetes` datasets for the `sign` nonlinearity. The value of b is chosen to be 0.0001, 2.0 and 3.0×10^{-6} respectively for `Australian`, `Heart` and `Diabetes` datasets for the `tanh` nonlinearity. The value of b is chosen to be 0.001, 5.0 and 5.0 respectively for `Australian`, `Heart` and `Diabetes` datasets for the `nonlinear-quantizer` nonlinearity.

We first note that (see Figure 4.1) `d-clipped-SGD` stabilizes the trajectory as compared to the linear SGD, even if the initial clipping level was high. At the same time, clipped-SGD with large clipping levels performs similarly as SGD. It is noteworthy, that SGD has the least oscillations for `Australian` and `Diabetes` datasets, despite the fact that these datasets have heavier or similar tails. This can be attributed to the fact that SGD does not get close to the solution in terms of functional value. SSTM in particular shows large oscillations, which can be attributed to it being a version of accelerated/momentum-based methods and its usage of small batch sizes. `Clipped-SSTM` on the other hand suffers less from oscillations and has a comparable convergence rate as SSTM. In comparison, all the three nonlinear schemes that have been proposed in this paper, have very little oscillations. While the `tanh` algorithm is outperformed by the algorithms with other nonlinearities from our framework, its performance is at par with the other baselines from [22]. In particular, the `sign` algorithm compares favorably to other baselines in terms of convergence for `Australian` and `Heart` datasets. The `nonlinear-quantizer` algorithm outperforms other baselines for the `Diabetes` dataset. The good behavior of `tanh` and `sign` on the heavy-tail data sets, specially relative to the linear SGD, also viewing `tanh` as a smooth approximation of `sign`, might also be related with the insights from Example 3.4. In summary, the three simple example nonlinearities from the proposed framework are comparable or favorable over the considered state of the art benchmarks on the studied datasets.

8 Conclusion

We proposed a general framework for nonlinear stochastic gradient descent (SGD) under heavy-tail gradient noise. Unlike existing studies of SGD under heavy-tail noise that focus on specific nonlinear functions (e.g., adaptive clipping), our framework includes a broad class of component-wise (e.g., sign gradient) and joint (e.g., gradient clipping) nonlinearities. We establish for the considered methods almost sure convergence, MSE convergence rate, and also asymptotic covariance for component-wise nonlinearities. We carry out numerical experiments on several real datasets that exhibit heavy tail gradient noise effects. The experiments show that, while our framework is more general than existing studies of SGD under heavy-tail noise, several easy-to-implement nonlinearities from our framework are competitive with state of the art alternatives.

Appendix

A. Some results in stochastic approximation

We present a useful result on single time scale stochastic approximation; see [20], Theorems 4.4.4 and 6.6.1.

Theorem 8.1 *Let $\left\{\mathbf{x}^t \in \mathbb{R}^d\right\}$ be a random sequence that satisfies:*

$$\mathbf{x}^{t+1} = \mathbf{x}^t + \alpha_t \left[\mathbf{r}(\mathbf{x}^t) + \boldsymbol{\gamma}\left(t+1, \mathbf{x}^t, \omega\right)\right], \tag{8.1}$$

where, $\mathbf{r}(\cdot) : \mathbb{R}^d \longmapsto \mathbb{R}^d$ is Borel measurable and $\{\boldsymbol{\gamma}(t, \mathbf{x}, \omega)\}_{t\geq 0,\ \mathbf{x}\in\mathbb{R}^d}$ is a family of random vectors in $\mathbb{R}^d$, defined on a probability space $(\Omega, \mathcal{F}, \mathcal{P})$, and $\omega \in \Omega$ is a canonical element. Let the following sets of assumptions hold:

- ***(B1)**: The function $\boldsymbol{\gamma}(t, \cdot, \cdot) : \mathbb{R}^d \times \Omega \longrightarrow \mathbb{R}^d$ is $\mathcal{B}^d \otimes \mathcal{F}$ measurable for every t; $\mathcal{B}^d$ is the Borel algebra of $\mathbb{R}^d$.*
- ***(B2)**: There exists a filtration $\{\mathcal{F}_t\}_{t\geq 0}$ of $\mathcal{F}$, such that, for each t, the family of random vectors $\{\boldsymbol{\gamma}\left(t, \mathbf{x}, \omega\right)\}_{\mathbf{x}\in\mathbb{R}^d}$ is $\mathcal{F}_t$ measurable, zero-mean and independent of $\mathcal{F}_{t-1}$.*
- ***(B3)**: There exists a twice continuously differentiable function $V\left(\mathbf{x}\right)$ with bounded second order partial derivatives and a point $\mathbf{x}^\star \in \mathbb{R}^d$ satisfying:*

  $$V\left(\mathbf{x}^\star\right) = 0,\ \ V\left(\mathbf{x}\right) > 0,\ \mathbf{x} \neq \mathbf{x}^\star,\ \ \lim_{\|\mathbf{x}\|\to\infty} V\left(\mathbf{x}\right) = \infty,$$
  $$\sup_{\epsilon<\|\mathbf{x}-\mathbf{x}^\star\|<\frac{1}{\epsilon}} \left(\mathbf{r}\left(\mathbf{x}\right), V_{\mathbf{x}}\left(\mathbf{x}\right)\right) < 0,\ \text{for any}\ \epsilon > 0$$

  where $V_{\mathbf{x}}\left(\mathbf{x}\right)$ denotes the gradient (vector) of $V(\cdot)$ at $\mathbf{x}$.
- ***(B4)**: There exist constants $k_1, k_2 > 0$, such that,*

  $$\left\|\mathbf{r}\left(\mathbf{x}\right)\right\|^2 + \mathbb{E}\left[\left\|\boldsymbol{\gamma}\left(t+1, \mathbf{x}, \omega\right)\right\|^2\right] \leq k_1\left(1 + V\left(\mathbf{x}\right)\right) - \\ - k_2\left(\mathbf{r}\left(\mathbf{x}\right), V_{\mathbf{x}}\left(\mathbf{x}\right)\right)$$
- ***(B5)**: The weight sequence $\{\alpha_t\}$ satisfies*

  $$\alpha_t > 0,\ \sum_{t\geq 0} \alpha_t = \infty,\ \sum_{t\geq 0} \alpha_t^2 < \infty. \tag{8.2}$$
- ***(C1)**: The function $\mathbf{r}\left(\mathbf{x}\right)$ admits the representation*

  $$\mathbf{r}\left(\mathbf{x}\right) = B\left(\mathbf{x} - \mathbf{x}^\star\right) + \delta\left(\mathbf{x}\right) \tag{8.3}$$

  where

  $$\lim_{\mathbf{x}\to\mathbf{x}^\star} \frac{\left\|\delta\left(\mathbf{x}\right)\right\|}{\left\|\mathbf{x} - \mathbf{x}^\star\right\|} = 0 \tag{8.4}$$
- ***(C2)**: The step-size sequence, $\{\alpha_t\}$ is of the form,*

  $$\alpha_t. = \frac{a}{t+1},\ \ \text{for any}\ t \geq 0, \tag{8.5}$$

  where $a > 0$ is a constant.

- ***(C3)**: Let I be the $d \times d$ identity matrix and a, B as in* (8.5) *and* (8.3)*, respectively. Then, the matrix $\Sigma = aB + \frac{1}{2}I$ is stable.*
- ***(C4)**: The entries of the matrices, for any $t \geq 0, \mathbf{x} \in R^d$,*

$$\mathbf{A}\left(t, \mathbf{x}\right) = \mathbb{E}\left[\boldsymbol{\gamma}\left(t+1, \mathbf{x}, \omega\right) \boldsymbol{\gamma}^\top\left(t+1, \mathbf{x}, \omega\right)\right],$$

  are finite, and the following limit exists: $\lim_{t \to \infty,\ \mathbf{x} \to \mathbf{x}^\star} \mathbf{A}\left(t, \mathbf{x}\right) = \mathcal{S}_0$
- ***(C5)**: There exists $\epsilon > 0$, such that*

$$\lim_{R \to \infty} \sup_{\|\mathbf{x}-\mathbf{x}^\star\| < \epsilon} \sup_{t \geq 0} \int_{\|\boldsymbol{\gamma}(t+1,\mathbf{x},\omega)\| > R} \left\|\boldsymbol{\gamma}\left(t+1, \mathbf{x}, \omega\right)\right\|^2 dP = 0 \tag{8.6}$$

Then we have the following:

*Let Assumptions **(B1)-(B5)** hold for $\{\mathbf{x}^t\}$ in* (8.1)*. Then, starting from an arbitrary initial state, the process $\{\mathbf{x}^t\}$ converges a.s. to $\mathbf{x}^\star$.*

*The normalized process, $\left\{\sqrt{t}\left(\mathbf{x}^t - \mathbf{x}^\star\right)\right\}$, is asymptotically normal if, besides Assumptions **(B1)-(B5)**, Assumptions **(C1)-(C5)** are also satisfied. In particular, as $t \to \infty$, we have:*

$$\sqrt{t}\left(\mathbf{x}^t - \mathbf{x}^\star\right) \xrightarrow{d} .\mathbb{N}(0, \mathcal{S}). \tag{8.7}$$

Also, the asymptotic covariance $\mathcal{S}$ of the multivariate distribution $\mathbb{N}(0, \mathcal{S})$ is

$$\mathcal{S} = a^2 \int_0^\infty e^{v\,\Sigma} \mathcal{S}_0 e^{v\,\Sigma^\top} dv. \tag{8.8}$$

Proof. For a proof see [20] (c.f. Theorems 4.4.4, 6.6.1).

We also make use of the following Theorem that is a slight modification of Lemmas 4 and 5 in [31].

Theorem 8.2 *Let z^t be a nonnegative (deterministic) sequence satisfying:*

$$z^{t+1} \leq \left(1 - r_1^t\right) z_1^t + r_2^t,$$

where $\{r_1^t\}$ and $\{r_2^t\}$ are deterministic sequences with

$$\frac{a_1}{(t+1)^{\delta_1}} \leq r_1^t \leq 1 \ \text{ and } \ r_2^t \leq \frac{a_2}{(t+1)^{\delta_2}},$$

with $a_1, a_2 > 0$, and $\delta_2 > \delta_1 > 0$. Then, the following holds: (1) If $\delta_1 < 1$, then $z^t = O(\frac{1}{t^{\delta_2 - \delta_1}})$; (2) If $\delta_1 = 1$, then $z^t = O(\frac{1}{t^{\delta_2 - 1}})$ provided that $a_1 > \delta_2 - \delta_1$; (3) if $\delta_1 = 1$ and $a_1 < \delta_2 - 1$, then $z^t = O(\frac{1}{t^\zeta})$, for any $\zeta < a_1$.

B. A demonstration that the linear SGD's iterate sequence has infinite variance

We provide here a simple demonstration that the linear SGD's iterate sequence has infinite variance under the setting of Assumptions 1, 2, and Assumption 3, condition 3., holds.

More precisely, assume that the gradient noise ν^t has infinite variance. Consider algorithm (2.3) for solving problem (1) with $f : \mathbb{R} \mapsto \mathbb{R}$, $f(x) = \frac{x^2}{2}$, with Ψ being the identity function. Further, consider arbitrary sequence of positive step-sizes $\{\alpha_t\}$. Then, we have:

$$x^{t+1} = \left(1 - \alpha_t\right) x^t - \alpha_t\, \nu^t,\ t = 0, 1, ..., \tag{8.9}$$

with arbitrary deterministic initialization $x^0 \in \mathbb{R}$. Then, squaring (8.9), using the independence of x^t and ν^t, and the fact that ν^t has zero mean, we get: $\mathbb{E}\left[(x^{t+1})^2\right] = (1 - \alpha_t)^2\, \mathbb{E}\left[(x^t)^2\right] + \alpha_t^2\, \mathbb{E}[(\nu^t)^2] \geq \alpha_t^2\ \mathbb{E}[(\nu^t)^2],\ t = 0, 1, ...$ Taking expectation and using the fact that $\mathbb{E}[(\nu^t)^2] = +\infty$, we see that $\mathbb{E}\left[(x^t)^2\right] = +\infty$, for any $t \geq 1$.

for gradient noise vector with mutually dependent entries We show that Theorem 3.2 continues to hold if Assumptions 2, parts 2. and 3., are relaxed, i.e., when we have an i.i.d. zero mean noise vector sequence $\{\nu^t\}$ with a joint pdf $p:\, \mathbb{R}^d \mapsto \mathbb{R}$. In more detail, we provide an extension of Lemma 6.2 but for component-wise nonlinearities. Namely, as in Lemma 6.2, consider, for a fixed $y \neq 0$:

$$\int \psi(\mathbf{y}+\mathbf{u})^\top \mathbf{y}\, p(\mathbf{u})\, d\mathbf{u}. \tag{8.10}$$

As, for $\mathbf{a} \in \mathbb{R}^d$, we have $\Psi(\mathbf{a}) = (\Psi(a_1), ..., \Psi(a_d))^\top$ (component-wise nonlinearity), we have:

$$\begin{aligned}\int \psi(\mathbf{y}+\mathbf{u})^\top \mathbf{y}\, p(\mathbf{u})\, d\mathbf{u} &= \int \left(\sum_{i=1}^{d} \psi(y_i+u_i) y_i \right) p(\mathbf{u})\, d\mathbf{u} \\ &= \sum_{i=1}^{d} \int \left(\psi(y_i+u_i)y_i\right)\, p(\mathbf{u})\, d\mathbf{u} = \sum_{i=1}^{d} \int \left(\psi(y_i+u_i)y_i\right)\, p_i(u_i)\, du_i,\end{aligned}$$

where $p_i(u_i)$ is the marginal pdf of the i-th component of $\boldsymbol{\nu}^t$. It is easy to show, as $p(\mathbf{u}) = p(-\mathbf{u})$, $\mathbf{u} \in \mathbb{R}^d$, that, for any $i = 1, ..., d$, we have $p_i(u) = p_i(-u)$, $u \in \mathbb{R}$. Define $\phi_i(a) - \int \psi(a+u)p_i(u)du$. Note that $\phi_i(a)$ now obeys Lemma 5.1. In particular, ϕ_i is also odd, and hence: $\int \psi(\mathbf{y}+\mathbf{u})^\top\, \mathbf{u}\, p(\mathbf{u})\, d\mathbf{u} \geq \sum_{i=1}^{d} |\phi_i(y_i)|\, |y_i|$. The proof now proceeds analogously to that of Theorem 3.2.

References

[1] Feng Niu, Benjamin Recht, Christopher Ré, and Stephen J Wright. Hogwild!: A lock-free approach to parallelizing stochastic gradient descent. *arXiv preprint arXiv:1106.5730*, 2011.

[2] Eduard Gorbunov, Filip Hanzely, and Peter Richtárik. A unified theory of sgd: Variance reduction, sampling, quantization and coordinate descent. In *International Conference on Artificial Intelligence and Statistics*, pages 680–690. PMLR, 2020.

[3] Lihua Lei and Michael I Jordan. On the adaptivity of stochastic gradient-based optimization. *SIAM Journal on Optimization*, 30(2):1473–1500, 2020.

[4] Farzad Yousefian, Angelia Nedić, and Uday V Shanbhag. On stochastic gradient and subgradient methods with adaptive steplength sequences. *Automatica*, 48(1):56–67, 2012.

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

[6] Saeed Ghadimi and Guanghui Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization i: A generic algorithmic framework. *SIAM Journal on Optimization*, 22(4):1469–1492, 2012.

[7] Horia Mania, Xinghao Pan, Dimitris Papailiopoulos, Benjamin Recht, Kannan Ramchandran, and Michael I Jordan. Perturbed iterate analysis for asynchronous stochastic optimization. *SIAM Journal on Optimization*, 27(4):2202–2229, 2017.

[8] Richard H Byrd, Gillian M Chin, Will Neveitt, and Jorge Nocedal. On the use of stochastic hessian information in optimization methods for machine learning. *SIAM Journal on Optimization*, 21(3):977–995, 2011.

[9] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer, 2010.

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

[11] Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczynski. *Lectures on stochastic programming: modeling and theory*. SIAM, 2021.

[12] Volkan Cevher, Stephen Becker, and Mark Schmidt. Convex optimization for big data: Scalable, randomized, and parallel algorithms for big data analytics. *IEEE Signal Processing Magazine*, 31(5):32–43, Sept. 2014.

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

[14] Mert Gurbuzbalaban, Umut Simsekli, and Lingjiong Zhu. The heavy-tail phenomenon in sgd. In *International Conference on Machine Learning*, pages 3964–3975. PMLR, 2021.

[15] Jingzhao Zhang, Sai Praneeth Karimireddy, Andreas Veit, Seungyeon Kim, Sashank J Reddi, Sanjiv Kumar, and Suvrit Sra. Why are adaptive methods good for attention models? *arXiv preprint arXiv:1912.03194*, 2019.

[16] Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity. *arXiv preprint arXiv:1905.11881*, 2019.

[17] Lukas Balles, Fabian Pedregosa, and Nicolas Le Roux. The geometry of sign gradient descent. *arXiv preprint arXiv:2002.08056*, 2020.

[18] Samuel Horváth, Dmitry Kovalev, Konstantin Mishchenko, Sebastian Stich, and Peter Richtárik. Stochastic distributed learning with gradient quantization and variance reduction. *arXiv preprint arXiv:1904.05115*, 2019.

[19] Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar. signsgd: Compressed optimisation for non-convex problems. In *International Conference on Machine Learning*, pages 560–569. PMLR, 2018.

[20] Mikhail Borisovich Nevelson and Rafail Zalmanovich Khasminskiĭ. *Stochastic approximation and recursive estimation*, volume 47. American Mathematical Soc., 1976.

[21] Boris Teodorovich Polyak and Yakov Zalmanovich Tsypkin. Adaptive estimation algorithms: convergence, optimality, stability. *Avtomatika i Telemekhanika*, (3):71–84, 1979.

[22] Eduard Gorbunov, Marina Danilova, and Alexander Gasnikov. Stochastic optimization with heavy-tailed noise via accelerated gradient clipping. *arXiv preprint arXiv:2005.10785*, 2020.

[23] Damek Davis, Dmitriy Drusvyatskiy, Lin Xiao, and Junyu Zhang. From low probability to high confidence in stochastic convex optimization. *J. Mach. Learn. Res.*, 22:49–1, 2021.

[24] Usman A Khan, Soummya Kar, and José MF Moura. Distributed average consensus: Beyond the realm of linearity. In *2009 Conference Record of the Forty-Third Asilomar Conference on Signals, Systems and Computers*, pages 1337–1342. IEEE, 2009.

[25] Srdjan S Stanković, Marko Beko, and Miloš S Stanković. A robust consensus seeking algorithm. In *IEEE EUROCON 2019-18th International Conference on Smart Technologies*, pages 1–6. IEEE, 2019.

[26] Sivaraman Dasarathan, Cihan Tepedelenlioğlu, Mahesh K Banavar, and Andreas Spanias. Robust consensus in the presence of impulsive channel noise. *IEEE Transactions on Signal Processing*, 63(8):2118–2129, 2015.

[27] Shreyas Sundaram and Bahman Gharesifard. Consensus-based distributed optimization with malicious nodes. In *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 244–249. IEEE, 2015.

[28] Venkatadheeraj Pichapati, Ananda Theertha Suresh, Felix X Yu, Sashank J Reddi, and Sanjiv Kumar. Adaclip: Adaptive clipping for private sgd. *arXiv preprint arXiv:1908.07643*, 2019.

[29] Soummya Kar, José MF Moura, and Kavita Ramanan. Distributed parameter estimation in sensor networks: Nonlinear observation models and imperfect communication. *IEEE Transactions on Information Theory*, 58(6):3575–3605, 2012.

[30] Chih-Chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):1–27, 2011.

[31] S. Kar and J. M. F. Moura. Convergence rate analysis of distributed gossip (linear parameter) estimation: Fundamental limits and tradeoffs. *IEEE Jour. Sel. Top. Sig. Proc.*, 5(4):674–690, Aug. 2011.