NONLINEAR GRADIENT MAPPINGS AND STOCHASTIC OPTIMIZATION: A GENERAL FRAMEWORK WITH APPLICATIONS TO HEAVY-TAIL NOISE

DUŠAN JAKOVETIĆ*, DRAGANA BAJOVIĆ†, ANIT KUMAR SAHU‡, SOUMMYA KAR§, NEMANJA MILOŠEVIĆ*, AND DUŠAN STAMENKOVIĆ*

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 general nonlinearities that can be decoupled component wise and a class of joint nonlinearities, 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.

Key words. Stochastic optimization; stochastic gradient descent; nonlinear mapping; heavy-tail noise; convergence rate; mean square analysis; asymptotic normality; stochastic approximation.

AMS subject classifications. 90C15, 90C25, 65K05, 62L20, 68T05

1. Introduction.

Stochastic gradient descent (SGD) and its variants, e.g., [27, 16, 23, 35, 25, 12, 24, 7], are popular and standard methods for large scale optimization and training of various machine learning models, e.g., [5, 6, 31, 8]. Recently, there have been several studies that demonstrate that the gradient noise in SGD is heavy-tailed, e.g., when training deep learning models [32, 17, 37].

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., [28, 36], the sign gradient, e.g., [4, 2], and (component-wise) quantized gradient, e.g., [1, 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

*University of Novi Sad, Faculty of Sciences, Department of Mathematics and Informatics (dusan.jakovetic@dmi.uns.ac.rs, nmilosev@dmi.uns.ac.rs, dusan.stamenkovic@dmi.uns.ac.rs)

†University of Novi Sad, Faculty of Technical Sciences, Department of Power, Electronic and Communication Engineering, (dbajovic@uns.ac.rs)

‡Amazon Alexa AI (anit.sahu@gmail.com)

§Department of Electrical and Computer Engineering, Carnegie Mellon University (soummyak@andrew.cmu.edu). The work of D. Bajovic and D. Jakovetic is partially supported by the European Union's Horizon 2020 Research and Innovation program under grant agreement No 957337. The paper reflects only the view of the authors and the Commission is not responsible for any use that may be made of the information it contains.

[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., [36], or communication efficiency, [2, 4].

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 the $O(1/t)$ global MSE rate.

Next, for the general nonlinearities with bounded outputs that can be decoupled component-wise and a restricted class of joint nonlinearities with bounded outputs, 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., [26], following the noise and nonlinearities assumptions framework similar to [30]. The paper [30] 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 [37] 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 paper [32] establishes convergence of the *linear* SGD assuming that the gradient noise follows a heavy-tailed α-stable distribution.

It is worth noting that, in addition to the MSE (expected optimality gap) results achieved here, it is also of interest to derive high probability bounds. Specifically, given a target accuracy $\epsilon > 0$ and a confidence level $1-\beta$, $\beta \in (0,1)$, we would like

to find $T(\epsilon,\beta)$ such that $f(\mathbf{x}^t) - f(\mathbf{x}^\star) \leq \epsilon$ with probability at least $1-\beta$, for all iterations $t \geq T(\epsilon,\beta)$. Application of the Markov inequality to our result $\mathbb{E}[\, f(\mathbf{x}^t) - f(\mathbf{x}^\star)\,] = O(1/t^\zeta)$ yields, abstracting dependencies on other system parameters, a bound $T(\epsilon,\beta) \sim \frac{1}{(\beta\,\epsilon)^{1/\zeta}}$. This involves a strong dependence on β, on the order $1/\beta^{1/\zeta}$. Several works, e.g., [13, 14, 19, 15, 11], establish high probability bounds where $T(\epsilon,\beta)$ depends *logarithmically* on β for the settings therein. For example, references [13, 14] establish high probability bounds for the stochastic gradient methods therein assuming that the gradient noise has light tails (sub-Gaussian noise). The authors of [19] establish the corresponding bounds for the basic SGD and the mirror descent that utilize a gradient truncation technique. They relax the noise sub-Gaussianity assumption and assume a finite noise variance. Very recently, [15] establishes high probability bounds for accelerated SGD with a clipping nonlinearity, but assuming a finite variance of the gradient noise. Reference [11] proposes a procedure called proxBoost and establishes for the procedure high probability bounds, again assuming a finite noise variance (without the sub-Gaussianity assumption). It is highly relevant to investigate high probability bounds for the problem setting and the algorithmic class considered in this paper. Of special interest is to provide high probability bounds for a broader class of nonlinearities than the usually studied clipping-type nonlinearities; this is an interesting future work direction.

In summary, with respect to existing work, our framework is more general 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 [15, 37, 11, 32]. 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. See also Examples 3.2 and 3.3. ahead for further rate comparisons with existing work.

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., [22, 33, 10]. The paper [22] introduces a trigonometric nonlinearity into a standard linear consensus dynamics and shows an improved dependence of the method on initial conditions. References [33] and [10] 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 [34] 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 Section 4 explain our results on nonlinear SGD for component-wise and joint nonlinearities, respectively. Section 5 and Section 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 $\text{trace}(\mathbf{A})$ 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 $\text{sign}(a)$ the sign function, i.e., $\text{sign}(a) = 1$, for $a > 0$, $\text{sign}(a) = -1$, for $a < 0$, and $\text{sign}(0) = 0$. 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 $L \geq \mu$.*

For asymptotic results (see ahead Theorems 3.1 and 3.3), we will also require the following assumption.

ASSUMPTION 2. *Function $f : \mathbb{R}^d \mapsto \mathbb{R}$ is twice continuously differentiable.*

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, f can correspond to the risk function, i.e.,

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

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 that is smooth and convex in its first argument for any fixed value of the second argument; and $\mathcal{R} : \mathbb{R}^d \mapsto \mathbb{R}$ is a smooth 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 Assumptions 1–2, including, e.g., ℓ_2-regularized quadratic and logistic losses.

We introduce a general framework for *nonlinear* SGD methods to solve problem (1); an algorithm within the framework takes the following form:

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

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; $\alpha_t > 0$ 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: $[\boldsymbol{\Psi}(\mathbf{w})]_i = \text{sign}(w_i)$, $i = 1, ..., d$;
2. Component-wise clipping: $[\boldsymbol{\Psi}(\mathbf{w})]_i = w_i$, for $|w_i| \leq m$; $[\boldsymbol{\Psi}(\mathbf{w})]_i = m$, for $w_i > m$, and $[\boldsymbol{\Psi}(\mathbf{w})]_i = -m$, for $w_i < -m$, for some constant $m > 0$.
3. Component-wise quantization: for each $i = 1, ..., d$, we let $[\boldsymbol{\Psi}(\mathbf{w})]_i = r_j$, 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., $[\boldsymbol{\Psi}(\mathbf{w})]_i = -[\boldsymbol{\Psi}(-\mathbf{w})]_i$, for each i and for each $\mathbf{w}$;
4. Normalized gradient: $\boldsymbol{\Psi}(\mathbf{w}) = \frac{\mathbf{w}}{\|\mathbf{w}\|}$, for $\mathbf{w} \neq 0$, and $\boldsymbol{\Psi}(0) = 0$;
5. Clipped gradient: $\boldsymbol{\Psi}(\mathbf{w}) = \mathbf{w}$, for $\|\mathbf{w}\| \leq M$, and $\boldsymbol{\Psi}(\mathbf{w}) = \frac{\mathbf{w}}{\|\mathbf{w}\|} M$, 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 $\boldsymbol{\nu} = \mathbf{g}_i(\mathbf{x}) - \nabla f(\mathbf{x})$. Several recent studies indicate that noise $\boldsymbol{\nu}$ exhibits heavy tails on many real data sets, e.g, [32, 17, 37]. (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., [29]. 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 $\boldsymbol{\Psi} : \mathbb{R}^d \mapsto \mathbb{R}^d$ of the form $\boldsymbol{\Psi}(w_1, ..., w_d) = (\Psi(w_1), ..., \Psi(w_d))^\top$, 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 a probability space $(\Omega, \mathcal{F}, P)$, where $\omega \in \Omega$ is a canonical element. For each $t = 0, 1, ...$, $\boldsymbol{\nu}^t : \Omega \mapsto \mathbb{R}^d$ is a random vector defined on $(\Omega, \mathcal{F}, P)$. We also denote by $\mathcal{F}_t$, $t = 0, 1, ...$, the σ-algebra generated by random vectors $\{\boldsymbol{\nu}^s\}$, $s = 0, ..., t$. Clearly, in view of (2.3), $\mathbf{x}^{t+1}$ is measurable with respect to $\mathcal{F}_t$, $t = 0, 1, ...$ We make the following assumptions; they follow the noise and nonlinearity framework similar to [30].

ASSUMPTION 3 (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. *The sequence of random vectors $\{\boldsymbol{\nu}^t\}$ is independent identically distributed*

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

(i.i.d.) Also, random variables ν_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 $p(u)$, $p : \mathbb{R} \mapsto \mathbb{R}_+$.*
3. *The pdf p is symmetric, i.e., $p(u) = p(-u)$, for any $u \in \mathbb{R}$ with $\int |u| p(u) du < +\infty$, and $p(u) > 0$ for $|u| \leq c_p$, for some constant $c_p > 0$.*

Note that Assumption 3 implies that $\boldsymbol{\nu}^t$ is zero-mean, for all t, and that $\boldsymbol{\nu}^t$ and $\mathbf{x}^t$ are mutually independent, for all t. For a class of unbounded nonlinearities Ψ that obey Assumption 6 ahead, we will additionally require the following.

ASSUMPTION 4. *The gradient noise variance $\sigma_\nu^2 = \int_{-\infty}^{+\infty} u^2 p(u) du < +\infty$.*

Assumption 3 requires that the noise vector is i.i.d. across its components $i = 1, ..., d$ which may be restrictive in certain scenarios. For the global MSE analysis, these assumptions can be relaxed; see ahead the remark after Theorem 3.2 and Appendix C.

Regarding noise pdf $p(u)$, except for strictly positivite values in the vicinity of zero (a very mild assumption), we require that the noise pdf is symmetric. Examples of the distributions that satisfy Assumption 3 include, e.g., a Gaussian zero-mean pdf or a Laplace zero-mean pdf with strictly positive variances, and heavy-tail zero-mean symmetric α-stable distributions [3]. [3] On the other hand, $p(u)$ may not be symmetric if, e.g., it is a mixture of some standard distributions. For example, consider random variable ν that is sampled from $\mathbb{N}(-m_1, \sigma^2)$ with probability $p = \frac{m_2}{m_1+m_2}$ and it is sampled from $\mathbb{N}(m_2, \sigma^2)$ with probability $1 - p$, for some $m_1 \neq m_2$, $m_1, m_2 > 0$, and $\sigma > 0$. Then, clearly, ν is zero-mean but does not have a symmetric pdf.

ASSUMPTION 5 (Nonlinearity Ψ). *Function $\Psi : \mathbb{R} \mapsto \mathbb{R}$ is a continuous (except possibly on a point set with Lebesgue measure of zero), monotonically non-decreasing and odd function, i.e., $\Psi(-w) = -\Psi(w)$, for any $w \in \mathbb{R}$. Moreover, Ψ is piece-wise differentiable. Finally, Ψ is either discontinuous at zero, or $\Psi(u)$ is strictly increasing for $u \in (-c_\Psi, c_\Psi)$, for some $c_\Psi > 0$.*

In addition, we impose one of the Assumptions 6 or 7 below.

ASSUMPTION 6. *$|\Psi(w)| \leq C_1 (1 + |w|)$, for any $w \in \mathbb{R}$, for some constant $C_1 > 0$.*

ASSUMPTION 7. *$|\Psi(w)| \leq C_2$, for some constant $C_2 > 0$.*

Assumption 3 and Assumption 5 are imposed throughout the paper. Assumption 4 is imposed when Assumption 6 holds, i.e., for the nonlinearities Ψ that can have unbounded outputs. When Assumption 7 is imposed, then Assumption 4 is not required.

Note that, provided that Assumption 7 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. Assumption 5 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 5 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:

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

[3] A random variable Z has a symmetric α-stable zero-mean distribution with scale parameter $\sigma > 0$ if its characteristic function takes the form: $\mathbb{E}\left[\exp(i\, u\, Z)\right] = \exp(-\sigma^\alpha |u|^\alpha)$, $u \in \mathbb{R}$, $\alpha \in [0, 2]$.

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 now state the main results on (2.3) with component-wise nonlinearities, including the results on a.s. convergence, MSE rate, and asymptiotic normality. We start with the following Theorem that establishes a.s. convergence.

THEOREM 3.1 (Almost sure convergence: Component-wise nonlinearity). *Consider algorithm (2.3) for solving optimization problem (2.1), and let Assumptions 1, 2, 3, 5, and 7 hold. Further, let the positive 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). Moreover, the result holds if Assumption 7 is replaced with Assumption 6, and Assumption 4 is additionally imposed.*

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 B, 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.

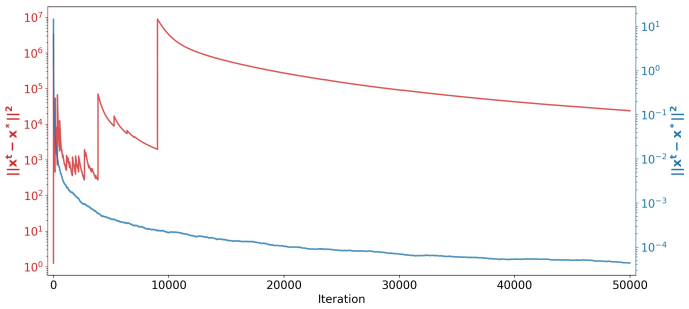


Fig. 3.1: Illustration of Theorem 3.1: estimated MSE versus iteration counter for the nonlinear SGD in (2.3) with component-wise sign nonlinearity (blue line) and the linear SGD (red line).

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 has a heavy-tailed pdf given by:

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

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 3.1 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, 3, 5, and 7 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 $\phi(a) \geq \frac{\phi'(0)}{2}\,a$, for any $a \in [0, \xi)$. Furthermore, let Assumptions 1, 3, 5, and 6, and 4 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 3, 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 $\mathbf{u} \in \mathbb{R}^d$, that is strictly positive for $\|\mathbf{u}\| \leq u_0$, for some $u_0 > 0$. 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 given in (3.2). To be specific, consider the sign nonlinearity $\Psi(w) = \text{sign}(w)$. 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 $\min\{2\delta - 1,\ \frac{\mu}{L}\frac{1-\delta}{\sqrt{d}}\frac{\alpha-1}{\alpha}\}$.

We also compare the rate ζ with the analysis in [37] that is closest to our setting

with respect to existing work. Modulo the differences in the assumptions of the assumed settings here and in [37], the rate in [37], when adapted to the noise pdf in Example 3.1, reads as follows: $\frac{2(r-1)}{r}$, where r is any number such that $r \leq \min\{\alpha-1, 2\}$. When compared with ζ, the rate in [37] is clearly better for α above a threshold. However, as α decreases and approaches the value 2, the rate achieved here stays bounded away from zero and approaches the quantity: $\min\left\{2\delta-1, \frac{1}{2}\frac{\mu}{L}\frac{1-\delta}{\sqrt{d}}\right\}$. In contrast, the rate in in [37] approaches zero as α approaches 2. [4]

Example 3.3. We continue to assume the noise pdf in (3.2), but here we consider the component-wise clipping nonlinearity Ψ with saturation value m. For simplicity, we take $m > 1$, while similar bounds can be obtained for $m \leq 1$ as well. It can be shown that the rate ζ can be estimated as (see Appendix E):

$$\min\left\{2\delta-1, \frac{\mu}{L\sqrt{d}}\frac{(1-\delta)(m-1)(1-(m+1)^{-\alpha})}{m}\right\}. \tag{3.3}$$

The above α-dependent estimate can be replaced with a more conservative rate that holds for any $\alpha > 2$: $\min\{2\delta-1, \frac{\mu}{L\sqrt{d}}\frac{(1-\delta)(m-1)(1-(m+1)^{-2})}{m}\}$. We again compare the rate achieved by the proposed method with the rate from [37] that equals: $\frac{2(r-1)}{r}$, $r < \min\{\alpha-1, 2\}$. We can see that the rate in [37] is better than (3.3) for α above a threshold. On the other hand, when α decreases to 2, the rate of [37] approaches zero, while (3.3) becomes better and stays bounded away from zero.

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, 2, 3, 5, and 7 hold. 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:*

$$\sqrt{t+1}(\mathbf{x}^t - \mathbf{x}^\star) \xrightarrow{d} \mathbb{N}(0, \mathcal{S}), \tag{3.4}$$

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:

$$\mathcal{S}_0 = \sigma_\Psi^2\,\mathbf{I},\ \ \sigma_\Psi^2 = \int |\Psi(v)|^2 p(v) dv,\ \ \Sigma = \frac{1}{2}\mathbf{I} - a\,\phi'(a)\nabla^2 f(\mathbf{x}^\star). \tag{3.5}$$

Moreover, the same result holds when Assumption 7 is replaced with Assumption 6, and Assumption 4 is additionally imposed.

[4] It is worth noting that reference [37] establishes certain tightness results on the rate achieved therein, by providing a "hard" problem example where the mean squared error after t iterations is $\Omega(1/t^{\frac{2(r-1)}{r}})$. However, this does not contradict our results due to the different sets of Assumptions made here and in [37]. Most notably, [37] assumes bounded moments of *gradients* and allow for *dependence* between the current point $\mathbf{x}^t$ and the gradient noise $\boldsymbol{\nu}^t$. In fact, the "hard example" construction in the proof of Theorem 5 in [37] constructs $\boldsymbol{\nu}^t$ as an explicit function of $\mathbf{x}^t$.

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

Example 3.4 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.

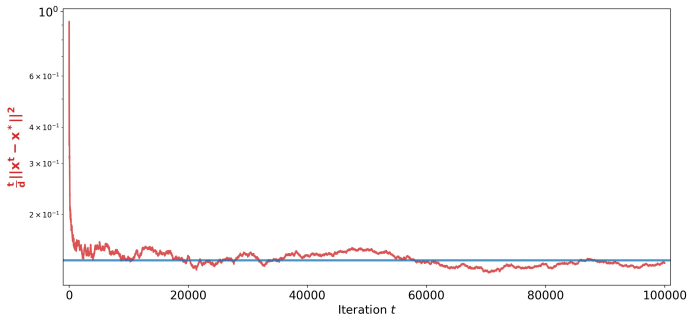


Fig. 3.2: Illustration of Theorem 3.3: Monte Carlo estimate of per-entry asymptotic variance (red line) and the theoretical per-entry asymptotic variance in Theorem 3.3 (blue line).

Example 3.5. 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:

$$\inf_{a>\frac{1}{2\phi'(0)}} \text{trace}\,(\mathcal{S}) = \frac{\sigma_\Psi^2}{(\phi'(0))^2}. \tag{3.6}$$

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 (3.6) 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 $\phi'(0) = 2\int_0^m p(v)dv$. (See Appendix F for the derivation.) 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 (3.6) (it suppresses the noise effect). On the other hand, reducing m has a negative effect on the denominator in (3.6); 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. Note also that, when σ_ν^2 is infinite, the linear SGD has an infinite asymptotic variance in (3.6), while the nonlinear SGD with any $m \in (0, \infty)$ has a finite asymptotic variance.

Example 3.6. We continue to assume the simplified setting when the per-entry asymptotic variance equals (3.6). 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 $\phi'(0) = \alpha - 1$. (See Appendix G.) 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 3.7. We still consider the simplified setting of (3.6). If the noise pdf $p(u)$ is known, then, following [30], we can find a globally optimal nonlinearity that minimizes (3.6) 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 $p(u)$.

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 $\mathbf{\Psi} : \mathbb{R}^d \mapsto \mathbb{R}^d$. Recall also filtration $\mathcal{F}_t$ in Section 3.

ASSUMPTION 8. *[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. 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 $\int_{\mathcal{J}_A} p(\mathbf{u}) d\mathbf{u} > \lambda(A)$, where $\mathcal{J}_A = \{\mathbf{u} \in \mathbb{R}^d : \frac{\mathbf{u}^\top \mathbf{x}}{\|\mathbf{u}\|\|\mathbf{x}\|} \in [0, A], \|u\| \leq B_0\}$.*[5]

Assumption 8 allows for a heavy-tailed noise vector whose components can be mutually dependent. Condition 2. in Assumption 8 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 $p(\mathbf{u})$.

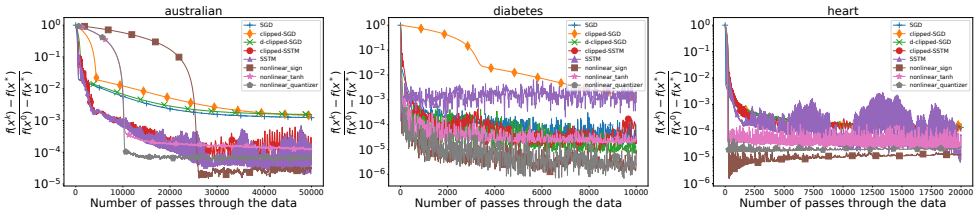
We make the following assumption on the joint nonlinearity.

ASSUMPTION 9 (Nonlinearity $\mathbf{\Psi}$). *The nonlinear map $\mathbf{\Psi} : \mathbb{R}^d \mapsto \mathbb{R}^d$ takes the following form: $\mathbf{\Psi}(\mathbf{w}) = \mathbf{w}\mathcal{N}(\|\mathbf{w}\|)$, where function $\mathcal{N} : \mathbb{R}_+ \mapsto \mathbb{R}_+$ satisfies the following: $\mathcal{N}$ 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.*

In addition, we assume that either Assumption 10 or Assumption 11 holds.

ASSUMPTION 10. *$\|\mathbf{\Psi}(\mathbf{w})\| \leq C_2'$, for any $\mathbf{w} \in \mathbb{R}^d$, for some $C_2' > 0$.*

[5]The integration set $\mathcal{J}_A$ 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 $0 \leq \mathbf{u}^\top \mathbf{x} \leq A\|\mathbf{u}\|\|\mathbf{x}\|$.



(a) (b) (c)

Fig. 4.1: Comparison of the optimization algorithms across different datasets

ASSUMPTION 11. $\|\mathbf{\Psi}(\mathbf{w})\| \leq C_1'(1+\|\mathbf{w}\|)$, *for any* $\mathbf{w} \in \mathbb{R}^d$, *for some* $C_1' > 0$

There are many nonlinearities that satisfy the above Assumptions, including the normalized gradient and the clipped gradient discussed in Section 2. If Assumption 11 holds, then we additionally require the following.

ASSUMPTION 12. *There holds:* $\int \|\mathbf{u}\|^2 p(\mathbf{u}) d\mathbf{u} < \infty$.

For asymptotic normality in the joint nonlinearity case, we additionally impose the following.

ASSUMPTION 13. *Function* $\mathcal{N} : \mathbb{R}_+ \mapsto \mathbb{R}$ *is differentiable for any positive argument, i.e.,* $\mathcal{N}'(a)$ *exists for any* $a > 0$. *Furthermore,* $\sup_{a>0} \mathcal{N}(a) < +\infty$.

We first state Theorem 4.1 and Theorem 4.2 on the a.s. convergence and the MSE rate of algorithm (2.3), respectively; we then illustrate the results with examples.

THEOREM 4.1 (A.s. convergence: Joint nonlinearity). *Consider algorithm* (2.3) *for solving optimization problem* (2.1), *and let Assumptions 1, 2, 8, 9, and 10 hold. Further, let the step-size sequence* $\{\alpha_t\}$ *be square-summable, non-summable. .Then, for the sequence of iterates* $\{\mathbf{x}^t\}$ *generated by algorithm* (2.3), *it holds that* $\mathbf{x}^t \to \mathbf{x}^\star$, *a.s. Moreover, the result continues to hold if Assumption 10 is replaced with Assumption 11, and Assumption 12 is additionally imposed.*

We now state our MSE rate result for the joint nonlinearity case.

THEOREM 4.2 (MSE convergence rate: Joint nonlinearity). *Consider algorithm* (2.3) *for solving optimization problem* (2.1), *and let Assumptions 1, 8, 9, and 10 hold. Further, let the step-size sequence* $\{\alpha_t\}$ *be* $\alpha_t = a/(t+1)$, $a > 0$, $\delta \in (0.5, 1)$. *Then,* $\mathbb{E}\left[\|\mathbf{x}^t - \mathbf{x}^\star\|^2\right] = O(1/t^\zeta)$, *or equivalently,* $\mathbb{E}\left[f(\mathbf{x}^t) - f^\star\right] = O(1/t^\zeta)$. *Here,* $\zeta \in (0,1)$ *is any positive number smaller than:* $\min\left\{2\delta - 1, \frac{4\,a\,\mu\,(1-\kappa)\lambda(\kappa)(1-\delta)\mathcal{N}(1)}{L\,(\,a\,C_2' + \|\mathbf{x}^0\| + \|\mathbf{x}^\star\|\,) + B_0}\right\}$, *where* κ *is an arbitrary constant in* $(0,1)$, *and we recall quantities* B_0 *and* $\lambda(\kappa)$ *in Assumption 8;* μ *and* L *in Assumption 1; and* C_2' *in Assumption 9. In alternative, let Assumptions 1, 8, 9, 11, and 12 hold. Let* $\alpha_t = \frac{a}{(t+1)^\delta}$, $\delta \in (0.5, 1]$, *and assume that* $\inf_{\mathbf{w}\neq 0} \frac{\|\mathbf{\Psi}(\mathbf{w})\|}{\|\mathbf{w}\|} > 0$. *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.*

Example 4.1. We illustrate the rate ζ in Theorem 4.2 for the gradient clipping nonlinearity with floor level $M > 0$. We consider an arbitrary joint pdf $p : \mathbb{R}^d \mapsto \mathbb{R}_+$ that has "radial symmetry", i.e., $p(\mathbf{u}) = q(\|\mathbf{u}\|)$, where $q : \mathbb{R}_+ \mapsto \mathbb{R}_+$ is a given

function. For example, we let:

$$p(\mathbf{u}) = q(\|\mathbf{u}\|),\ q(\rho) = \frac{(\alpha-2)(\alpha-1)}{2\pi}\frac{1}{(1+\rho)^{\alpha}},\ \rho \geq 0,\ \ \alpha > 3. \tag{4.1}$$

It can be shown that $p(\mathbf{u})$ in (4.1) has finite moments of order r, $r < \alpha - 2$, and it has infinite moments for $r \geq \alpha - 2$. It holds that (see Appendix H for derivations) the rate ζ can be estimated as: $\min\left\{2\delta - 1,\ (1-\delta)\frac{0.68\,\mu}{L}\right\}$. Hence, up to universal constants, the rate ζ is approximated as $\min\left\{2\delta - 1,\ (1-\delta)\frac{\mu}{L}.\right\}$. It is easy to see that the same rate estimate can be obtained for the normalized gradient nonlinearity, under the same gradient noise setting.

We compare the rate estimate here with the rate for component-wise nonlinearities (e.g., component-wise clipping in Example 3.3) that is, up to universal constants, of order $\min\left\{2\delta - 1,\ (1-\delta)\frac{\mu}{\sqrt{d}\,L}.\right\}$. We can see that, with the joint nonlinearity examples here, the rate is improved with respect to the component-wise nonlinearities by a factor $\sqrt{d}$. In other words, the rate estimate for the joint nonlinearities does not deteriorate with the dimension d increase. This may be intuitively explained by considering the sign component-wise nonlinearity and the normalized gradient. These two functions coincide for $d = 1$ (and this is reflected by the identical rate estimates we obtain here), but they become different for $d > 1$ (as also reflected by our obtained rate estimates). Intuitively, in the noiseless case, the normalized gradient preserves "more information" about the exact gradient ("true search direction") than the component-wise sign function; hence, the difference in the estimated rates.

We now examine asymptotic normality for the joint nonlinearities case. We have the following theorem.

THEOREM 4.3 (Asymptotic normality: Joint nonlinearity). *Consider algorithm (2.3) for solving optimization problem (2.1), and let Assumptions 1, 2, 8, 9, 10, and 13 hold. Further, let the step-size sequence $\{\alpha_t\}$ equal $\alpha_t = a/(t+1)$, $a > 0$. Then: $\sqrt{t+1}(\mathbf{x}^t - \mathbf{x}^\star) \xrightarrow{d} \mathbb{N}(0, \mathcal{S})$. The asymptotic covariance $\mathcal{S}$ is given by $\mathcal{S} = a^2 \int_0^\infty e^{v\,\boldsymbol{\Sigma}} \mathcal{S}_0 e^{v\,\boldsymbol{\Sigma}} dv$, where $\mathcal{S}_0 = \int \mathbf{u}\mathbf{u}^\top \left(\mathcal{N}(\|\mathbf{u}\|)\right)^2 p(\mathbf{u}) d\mathbf{u}$; $\boldsymbol{\Sigma} = 1_{\overline{2\mathbf{I} + a\,\mathbf{B}}}$; $\mathbf{B} = -(\int \mathcal{N}(\|\mathbf{u}\|) p(\mathbf{u}) d\mathbf{u} + \int_{\mathbf{u}\neq 0} \frac{\mathbf{u}\mathbf{u}^\top}{\|\mathbf{u}\|} \mathcal{N}'(\|\mathbf{u}\|) p(\mathbf{u}) d\mathbf{u}) \nabla^2 f(\mathbf{x}^\star)$, and constant $a > 0$ in the step-size sequence is taken large enough such that matrix $\boldsymbol{\Sigma}$ is stable. Moreover, the result continues to hold if Assumption 10 is replaced with Assumption 11, and Assumption 12 is additionally imposed.*

Theorem 4.3 shows that asymptotic normality continues to hold for the joint nonlinearity case as well, provided that $\mathcal{N}(a)$ is differentiable for any $a > 0$ and that $\mathcal{N}$ is uniformly bounded from above.

5. Intermediate results and proofs: Component-wise nonlinearities. This section provides proofs of Theorem 3.1, Theorem 3.2, and Theorem 3.3, accompanied with the required intermediate results. Subsection 5.1 presents some useful intermediate results on stochastic approximation and deterministic time-varying sequences; Subsection 5.2 deals with the asymptotic analysis (Theorem 3.1 and Theorem 3.3); and Subsection 5.3 considers MSE analysis (Theorem 3.2).

5.1. Stochastic approximation and time-varying sequences. We present a useful result on single time scale stochastic approximation; see [26], Theorems 4.4.4 and 6.6.1.

THEOREM 5.1. *Let* $\left\{\mathbf{x}^t \in \mathbb{R}^d\right\}$ *be a random sequence that satisfies:*

$$\mathbf{x}^{t+1} = \mathbf{x}^t + \alpha_t \left[\mathbf{r}(\mathbf{x}^t) + \boldsymbol{\gamma}\left(t+1, \mathbf{x}^t, \omega\right)\right], \tag{5.1}$$

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* $V(\mathbf{x})$ *with bounded second order partial derivatives and a point* $\mathbf{x}^\star \in \mathbb{R}^d$ *satisfying:* $V(\mathbf{x}^\star) = 0$*,* $V(\mathbf{x}) > 0$*,* $\mathbf{x} \neq \mathbf{x}^\star$*,* $\lim_{\|\mathbf{x}\|\to\infty} V(\mathbf{x}) = \infty$*,* $\sup_{\epsilon<\|\mathbf{x}-\mathbf{x}^\star\|<\frac{1}{\epsilon}} \langle \mathbf{r}(\mathbf{x}), \nabla V(\mathbf{x})\rangle < 0$*, for any* $\epsilon > 0$*.*
(B4)*: There exist constants* $k_1, k_2 > 0$*, such that,*

$$\|\mathbf{r}(\mathbf{x})\|^2 + \mathbb{E}\left[\|\boldsymbol{\gamma}\left(t+1, \mathbf{x}, \omega\right)\|^2\right] \leq k_1\left(1 + V(\mathbf{x})\right) - \\ - k_2\langle \mathbf{r}(\mathbf{x}), \nabla V(\mathbf{x})\rangle.$$

(B5)*: The weight sequence* $\{\alpha_t\}$ *satisfies* $\alpha_t > 0$*,* $\sum_{t\geq 0}\alpha_t = \infty$*,* $\sum_{t\geq 0}\alpha_t^2 < \infty$*.*
(C1)*: The function* $\mathbf{r}(\mathbf{x})$ *admits the representation*

$$\mathbf{r}(\mathbf{x}) = \boldsymbol{B}\left(\mathbf{x} - \mathbf{x}^\star\right) + \delta(\mathbf{x}), \tag{5.2}$$

where $\lim_{\mathbf{x}\to\mathbf{x}^\star} \frac{\|\delta(\mathbf{x})\|}{\|\mathbf{x}-\mathbf{x}^\star\|} = 0$*.*
(C2)*: The step-size sequence,* $\{\alpha_t\}$ *is of the form,* $\alpha_t = \frac{a}{t+1}$*, for any* $t \geq 0$*, where* $a > 0$ *is a constant.*
(C3)*: Let* $\mathbf{I}$ *be the* $d \times d$ *identity matrix and* $a, \mathbf{B}$ *as in C2 and C1, respectively. Then, the matrix* $\Sigma = a\mathbf{B} + \frac{1}{2}\mathbf{I}$ *is stable.*
(C4)*: The entries of the matrices, for any* $t \geq 0, \mathbf{x} \in \mathbb{R}^d$*,* $\mathbf{A}(t, \mathbf{x}) = \mathbb{E}[\,\boldsymbol{\gamma}\left(t+1, \mathbf{x}, \omega\right) \boldsymbol{\gamma}^\top\left(t+1, \mathbf{x}, \omega\right)]$ *are finite, and the following limit exists:* $\lim_{t\to\infty,\ \mathbf{x}\to\mathbf{x}^\star} \mathbf{A}(t, \mathbf{x}) = \mathcal{S}_0$*.*
(C5)*: There exists* $\epsilon > 0$*, such that*

$$\lim_{R\to\infty} \sup_{\|\mathbf{x}-\mathbf{x}^\star\|<\epsilon} \sup_{t\geq 0} \int_{\|\boldsymbol{\gamma}(t+1,\mathbf{x},\omega)\|>R} \|\boldsymbol{\gamma}\left(t+1, \mathbf{x}, \omega\right)\|^2 dP = 0. \tag{5.3}$$

Then we have the following:

Let Assumptions ***(B1)-(B5)*** *hold for* $\{\mathbf{x}^t\}$ *in* (5.1)*. 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:* $\sqrt{t}\left(\mathbf{x}^t - \mathbf{x}^\star\right) \xrightarrow{d} \mathbb{N}(0, \mathcal{S})$*. Also, the asymptotic covariance* $\mathcal{S}$ *of the multivariate Gaussian distribution* $\mathbb{N}(0, \mathcal{S})$ *is* $\mathcal{S} = a^2 \int_0^\infty e^{v\,\boldsymbol{\Sigma}}\, \mathcal{S}_0 e^{v\,\boldsymbol{\Sigma}^\top} dv$*.*

Proof. For a proof see [26] (c.f. Theorems 4.4.4, 6.6.1). ∎

We also make use of the following Theorem, proved in Appendix A; see also Lemmas 4 and 5 in [21].

THEOREM 5.2. *Let* z^t *be a nonnegative (deterministic) sequence satisfying:*

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

for all $t \geq t'$, *for some* $t' > 0$, *with some* $z^{t'} \geq 0$. *Here,* $\{r_1^t\}$ *and* $\{r_2^t\}$ *are deterministic sequences with* $\frac{a_1}{(t+1)^{\delta_1}} \leq r_1^t \leq 1$ *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* $a_1 \leq \delta_2 - 1$, *then* $z^t = O(\frac{1}{t^{\zeta}})$, *for any* $\zeta < a_1$.

5.2. Asymptotic analysis: Proofs of Theorem 3.1 and Theorem 3.3. The next Lemma, due to [30], 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. For a proof of the Lemma, see Appendix D.

LEMMA 5.3. *[30] Consider function* ϕ *in* (3.1), *where function* $\Psi : \mathbb{R} \mapsto \mathbb{R}$ *satisfies Assumption 5, and noise pdf* $p : \mathbb{R} \mapsto \mathbb{R}_+$ *satisfies Assumption 3. Then, the following holds.*

1. ϕ *is odd;*
2. *If in addition Assumption 7 holds, then* $|\phi(a)| \leq K_2$, *for any* $a \in \mathbb{R}$, *for some constant* $K_2 > 0$;
3. *If in addition Assumption 6 holds, 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. *If in addition either Assumption 6 or Assumption 7 holds, then* ϕ *is differentiable at zero, with a strictly positive derivative at zero, equal to:*

$$\phi'(0) = \sum_{i=1}^{s} \left(\Psi(\nu_i + 0) - \Psi(\nu_i - 0)\right) p(\nu_i) + \sum_{i=0}^{s} \int_{\nu_i}^{\nu_{i+1}} \Psi'(\nu) p(\nu) d\nu, \tag{5.4}$$

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

Remark. In view of (5.4), we highlight the need that $p(u)$ is strictly positive in the vicinity of zero and that Ψ is either discontinuous at zero or strictly increasing in the vicinity of zero, in order for $\phi'(0)$ to be strictly positive. (see Assumptions 3 and 5.) Consider the following counterexample: $\Psi(u) = \text{sign}(u)$, where p corresponds to the uniform distribution on the set $(-u_2, -u_1) \cup (u_1, u_2)$, for $0 < u_1 < u_2$. Note that p is zero in the vicinity of zero. Then, by (5.4), $\phi'(0) = 0$.

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 [26]; more precisely, we utilize Theorem 5.1 in the Appendix; see also [20].

We first put algorithm (2.3) in the format that complies with Theorem 5.1. Namely, algorithm (2.3) can be written as:

$$\mathbf{x}^{t+1} = \mathbf{x}^t + \alpha_t \left[\mathbf{r}(\mathbf{x}^t) + \boldsymbol{\gamma}(t+1, \mathbf{x}^t, \omega)\right]. \tag{5.5}$$

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

$$\mathbf{r}(\mathbf{x}) = -\boldsymbol{\phi}(\nabla f(\mathbf{x})), \tag{5.6}$$

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

(5.7)

$$\mathbf{r}(\mathbf{x}) = -\left(\phi([\nabla f(x)]_1), ..., \phi([\nabla f(x)]_d)\right)^\top, \quad \boldsymbol{\gamma}(t+1, \mathbf{x}, \omega) = \boldsymbol{\phi}(\nabla f(\mathbf{x})) - \Psi(\nabla f(\mathbf{x}) + \boldsymbol{\nu}^t).$$

We provide an intuition behind the algorithmic format (5.5). Quantity $\mathbf{r}(\mathbf{x})$ 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 $f^\star = \inf_{x\in\mathbb{R}^d} f(x) = f(x^\star)$. By Assumptions 1 and 2, V is twice continuously differentiable and has uniformly bounded second order partial derivatives, as required by Theorem 5.1. We are ready to prove Theorem 3.1.

Proof (Proof of Theorem 3.1). We now verify conditions B1-B5 from Theorem 5.1. Recall from Section 3 $\mathcal{F}_t$, the σ-algebra generated with random vectors $\boldsymbol{\nu}^s$, $s = 0, ..., t$. Then, the family of random vectors $\{\boldsymbol{\gamma}(t+1,\mathbf{x},\omega)\}_{\mathbf{x}\in\mathbb{R}^d}$ is $\mathcal{F}_t$-measurable, zero-mean and independent of $\mathcal{F}_{t-1}$. Also, clearly, function $\boldsymbol{\gamma}(t+1,\cdot,\cdot)$ is measurable, for all t. Thus, conditions Bi and B2 hold.

For B3, we need to prove that $\sup_{\mathbf{x}:\|\mathbf{x}-\mathbf{x}^\star\|\in(\epsilon,\frac{1}{\epsilon})} \langle \mathbf{r}(\mathbf{x}), \nabla V(\mathbf{x})\rangle < 0$, for any $\epsilon > 0$, where $\nabla V(\mathbf{x}) = \nabla f(\mathbf{x})$. Let us fix an $\epsilon > 0$. Then, we have, for any $\mathbf{x} \in \mathbb{R}^d$:

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

where the last equality holds because ϕ is an odd function. Consider arbitrary $\mathbf{x}$ such that $\|\mathbf{x}-\mathbf{x}^\star\| \geq \epsilon$. As $\|\nabla f(\mathbf{x})\|^2 \geq \mu^2\|\mathbf{x}-\mathbf{x}^\star\|^2$ (due to strong convexity of f), we have $\|\nabla f(\mathbf{x})\| \geq \mu\epsilon$, where we recall that μ is the strong convexity constant of f. Therefore, there exists an index $i \in \{1, ..., d\}$ such that $|[\nabla f(\mathbf{x})]_i| \geq \frac{1}{d}\mu\epsilon =: \epsilon'$. Next, because $\phi'(0) > 0$, and ϕ is continuous at 0 and is non-decreasing (by Lemma 5.3), we have that $|\phi(b)| \geq \delta$ for some $\delta = \delta(\epsilon) > 0$, for all $b \in [\epsilon, 1/\epsilon]$. Finally, we have that: $\langle \mathbf{r}(\mathbf{x}), \nabla V(\mathbf{x})\rangle \leq -\epsilon'\delta(\epsilon)$, for any $\mathbf{x}$ such that $\|\mathbf{x}-\mathbf{x}^\star\| \in [\epsilon, \frac{1}{\epsilon}]$, and therefore $\sup_{\mathbf{x}:\|\mathbf{x}-\mathbf{x}^\star\|\in(\epsilon,\frac{1}{\epsilon})} \langle \mathbf{r}(x), \nabla V(\mathbf{x})\rangle \leq \sup_{\mathbf{x}:\|\mathbf{x}-\mathbf{x}^\star\|\in[\epsilon,\frac{1}{\epsilon}]} \langle \mathbf{r}(x), \nabla V(\mathbf{x})\rangle \leq -\delta(\epsilon)\,\epsilon'$ < 0, hence verifying condition B3.

We next verify condition B4. Consider quantity $\mathbf{r}(\mathbf{x})$ in (5.6). By Lemma 5.3 and the fact that f has Lipschitz gradient and is strongly convex (Assumption 1), it follows that: $\|\mathbf{r}(\mathbf{x})\|^2 \leq C_{r,1} + C_{r,2}V(\mathbf{x})$, for some positive constants $C_{r,1}$ and $C_{r,2}$. Also, since $\|\boldsymbol{\gamma}(\mathbf{x},t+1,\omega)\|^2 \leq 2\|\phi(\nabla f(\mathbf{x}))\|^2 + 2\|\Psi(\nabla f(\mathbf{x}) + \boldsymbol{\nu}^t)\|^2$, and it holds that either 1) Ψ is bounded or 2) $|\Psi(a)| \leq C_2\,(1+|a|)$ and ν_i^t has a finite variance, we have: $\mathbb{E}\left[\|\boldsymbol{\gamma}(\mathbf{x},t+1,\omega)\|^2\right] \leq C_3 + C_4\,V(\mathbf{x})$, for some positive constants C_3, C_4. Now, we finally have:

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

for some positive constants C_5, C_6, and hence condition B4 holds for a constant $k_1 > 0$ and $k_2 = 0$.[6] 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 5.1. To verify condition C1, consider $\mathbf{r}(\mathbf{x})$ in (5.6) and note that,

[6] Note that the term $-\langle \mathbf{r}(\mathbf{x}), \nabla V(\mathbf{x})\rangle$ in condition B4 of Theorem 5.1 equals $\langle \boldsymbol{\phi}(\mathbf{x}), \nabla f(\mathbf{x})\rangle$. This quantity is nonnegative, for any $\mathbf{x} \in \mathbb{R}^d$, and so k_2 can be taken to be any positive number. In other words, setting $k_2 = 0$ in B4 corresponds to a tighter inequality than the corresponding inequality for any $k_2 > 0$.

using the mean value theorem, it can be expressed as follows:

$$
\begin{aligned}
\mathbf{r}(\mathbf{x}) &= -\phi(\nabla f(\mathbf{x}) - \nabla f(\mathbf{x}^\star)) \\
&= -\phi\left(\underbrace{\left[\int_0^1 \nabla^2 f(\mathbf{x}^\star + t(\mathbf{x}-\mathbf{x}^\star)dt\right]}_{\mathbf{H}}(\mathbf{x}-\mathbf{x}^\star)\right) \\
&= -\phi\left(\mathbf{H}(\mathbf{x}-\mathbf{x}^\star)\right) = -\phi'(0)\nabla^2 f(\mathbf{x}^\star)(\mathbf{x}-\mathbf{x}^\star) + \delta(\mathbf{x}),
\end{aligned} \tag{5.8}
$$

where $\lim_{\mathbf{x}\to\mathbf{x}^\star} \frac{\|\delta(x)\|}{\|x-x^*\|} = 0$. Hence, in the notation of Theorem 5.1, we have that $\mathbf{B} = -\phi'(0)\nabla^2 f(\mathbf{x}^\star)$. Therefore, 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 (all its eigenvalues have negative real parts), 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 large enough a, because the matrix $\phi'(0)\nabla^2 f(\mathbf{x}^\star)$ is positive definite. More precisely, $\mathbf{\Sigma}$ is stable for $a > 1/(2\mu\phi'(0))$. Therefore, condition C3 holds, provided that $a > 1/(2\mu\phi'(0))$. We next consider condition C4. In the notation of Theorem 5.1, consider the following quantity:

$$
\begin{aligned}
&\mathbf{A}(t,\mathbf{x}) = \mathbb{E}\left[\boldsymbol{\gamma}(t+1,\mathbf{x},\omega)\boldsymbol{\gamma}(t+1,\mathbf{x},\omega)^\top\right] \\
&= \mathbb{E}\left[\left(\phi(\nabla f(\mathbf{x})) - \Psi(\nabla f(\mathbf{x}) + \boldsymbol{\nu}^t)\right)\left(\left(\phi(\nabla f(\mathbf{x})) - \Psi(\nabla f(\mathbf{x}) + \boldsymbol{\nu}^t)\right)^\top\right] \\
&= \mathbb{E}\left[\left(\phi(\nabla f(\mathbf{x})) - \Psi(\nabla f(\mathbf{x}) + \boldsymbol{\nu}^0)\right)\left(\left(\phi(\nabla f(\mathbf{x})) - \Psi(\nabla f(\mathbf{x}) + \boldsymbol{\nu}^0)\right)^\top\right] \qquad (5.9) \\
&= \mathbb{E}\left[\boldsymbol{\gamma}(1,\mathbf{x},\omega)\boldsymbol{\gamma}(1,\mathbf{x},\omega)^\top\right]. \qquad (5.10)
\end{aligned}
$$

Consider the set $\Omega^\star$ of all outcomes $\omega \in \Omega$ such that $\mathbf{\Psi}$ is continuous at $\nu^0(\omega)$. Clearly, the set $\Omega^\star$ has the probability one. For every $\omega \in \Omega^\star$, we have $\mathbf{\Upsilon}(\omega) := \lim_{t\to\infty,\mathbf{x}\to\mathbf{x}^\star} \boldsymbol{\gamma}(1,\mathbf{x},\omega)\boldsymbol{\gamma}(1,\mathbf{x},\omega)^\top = \Psi(\boldsymbol{\nu}^0)\Psi(\boldsymbol{\nu}^0)^\top$. Note that, for any $\epsilon > 0$, the random family $\|\boldsymbol{\gamma}(1,\mathbf{x},\omega)\boldsymbol{\gamma}(1,\mathbf{x},\omega)^\top\|$, $\|\mathbf{x}-\mathbf{x}^\star\| < \epsilon$ is dominated by an integrable random variable. (See ahead (5.12)–(5.13).) Therefore, by the dominated convergence theorem, and the fact that the entries of $\boldsymbol{\nu}^0$ are mutually independent with pdf $p(u)$, we have that:

$$
\lim_{t\to\infty,\mathbf{x}\to\mathbf{x}^\star} \mathbf{A}(t,\mathbf{x}) =: \mathcal{S}_0 = \mathbb{E}\left[\Psi(\boldsymbol{\nu}^0)\cdot\Psi(\boldsymbol{\nu}^0)^\top\right] = \sigma_\Psi^2\cdot\mathbf{I}, \tag{5.11}
$$

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 [20]. 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: $\|\boldsymbol{\gamma}(t+1,\mathbf{x},\omega)\|^2 \le 2\|\phi(\nabla f(\mathbf{x}))\|^2 + 2\|\psi(\nabla f(\mathbf{x}) + \boldsymbol{\nu}^t)\|^2$. First, consider the case when Assumptions 6 and 4 hold. Then:

$$
\begin{aligned}
\|\boldsymbol{\gamma}(t+1,\mathbf{x},\omega)\|^2 &\le C_7 + C_8\|\mathbf{x}-\mathbf{x}^\star\|^2 + C_9\|\boldsymbol{\nu}^t\|^2 \\
&\le C_7 + C_8\,\epsilon^2 + C_9\|\boldsymbol{\nu}^t\|^2,
\end{aligned} \tag{5.12}
$$

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

Now, let Assumption 7 hold. Then:

$$\|\boldsymbol{\gamma}(t+1,\mathbf{x},\omega)\|^2 \leq C_{10} + C_{11}\|\mathbf{x}-\mathbf{x}^\star\|^2 \leq C_{10} + C_{11}\,\epsilon^2. \tag{5.13}$$

Consider the family $\{\widehat{\boldsymbol{\gamma}}(t+1,\mathbf{x},\omega)\}_{t=0,1,\ldots,\|\mathbf{x}-\mathbf{x}^\star\|<\epsilon}$, with $\widehat{\boldsymbol{\gamma}}(t+1,\mathbf{x},\omega) = C_{10} + C_{11}\,\epsilon^2$. The family $\{\widehat{\boldsymbol{\gamma}}(t+1,\mathbf{x},\omega)\}_{t=0,1,\ldots,\|\mathbf{x}-\mathbf{x}^\star\|<\epsilon}$ is uniformly integrable, and condition C5 is verified analogously to the previous case. Summarizing, we have established that all conditions C1-C5 of Theorem 5.1 hold true, thus the proof of Theorem 3.3 □.

5.3. MSE analysis: Proof of Theorem 3.2. We start with the following Lemma that upper boounds $\|\nabla f(\mathbf{x}^t)\|$.

LEMMA 5.4. *Let Assumptions 1, 3, 5, and 7 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 each $t=1,2,...$, we have, a.s.:*

$$\|\nabla f(\mathbf{x}^t)\| \leq G_t := L\left(a\,C_2\,\sqrt{d}\,\frac{t^{1-\delta}}{1-\delta} + \|\mathbf{x}^0-\mathbf{x}^\star\|\right). \tag{5.14}$$

Proof. Consider (2.3). Because the output of each component nonlinearity Ψ is bounded in the absolute value by C_2 (Assumption 7), we have, for each $t\geq 1$:

$$\begin{aligned}\|\mathbf{x}^t-\mathbf{x}^\star\| &\leq \|\mathbf{x}^0-\mathbf{x}^\star\| + a\,\sqrt{d}\,C_2\sum_{s=0}^{t-1}\frac{1}{(s+1)^\delta}\\ &\leq \|\mathbf{x}^0-\mathbf{x}^\star\| + a\,C_2\,\sqrt{d}\left(\frac{t^{1-\delta}}{1-\delta}\right).\end{aligned} \tag{5.15}$$

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 (5.15), the result follows. □

We will also make use of the following Lemma.

LEMMA 5.5. *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 (5.14).*

Proof. Consider function ϕ in (3.1). By Lemma 5.3, we have that $\phi'(0)>0$ and ϕ is continuous at zero.[7] Because ϕ is differentiable at zero, using first order Taylor series, there holds: $\phi(u) = \phi(0) + \phi'(0)u + h(u)u = \phi'(0)u + h(u)u$, $u\in\mathbb{R}$, where $h:\mathbb{R}\mapsto\mathbb{R}$ is a function such that $\lim_{u\to 0} h(u) = 0$. Due to the latter property of h, there exists a positive number ξ such that $|h(u)| \leq \frac{\phi'(0)}{2}$, for all $u\in[0,\xi)$. Using the latter bound, we obtain that $\phi(u)\geq\frac{1}{2}\phi'(0)u$, $u\in[0,\xi)$. Now, because ϕ is non-decreasing (by Lemma 5.3), it holds for any $a'>\xi$ that $\phi(a)\geq\frac{\phi'(0)\,\xi\,a}{2\,a'}$, for any $a\in[0,a')$. Consider now $\nabla f(\mathbf{x}^t)$. By Lemma 5.4, 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, 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) under Assumptions 1, 3, 5, and 7. By the Lipschitz property of ∇f, we have, for any $\mathbf{x},\mathbf{y}\in\mathbb{R}^d$, that:

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

[7] 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.

and so, almost surely:

$$f(\mathbf{x}^{t+1}) \leq f(\mathbf{x}^t) + \left(\nabla f(\mathbf{x}^t)\right)^\top (-\alpha_t \Psi(\nabla f(\mathbf{x}^t) + \boldsymbol{\nu}^t)) + \frac{L}{2}\alpha_t^2 \|\Psi(\nabla f(\mathbf{x}^t) + \boldsymbol{\nu}^t)\|^2. \tag{5.16}$$

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:

$$f(\mathbf{x}^{t+1}) \leq f(\mathbf{x}^t) + \left(\nabla f(\mathbf{x}^t)\right)^\top (-\alpha_t \boldsymbol{\phi}(\nabla f(\mathbf{x}^t))) + \frac{L}{2}\alpha_t^2\, d^2 C_2^2 - \alpha_t \left(\nabla f(\mathbf{x}^t)\right)^\top \boldsymbol{\eta}^t, \text{ a.s.} \tag{5.17}$$

Recall filtration $\mathcal{F}_t$. Taking conditional expectation, and using that $\mathbb{E}[\boldsymbol{\eta}^t \,|\, \mathcal{F}_t] = 0$, we get that, almost surely:

$$\mathbb{E}[f(\mathbf{x}^{t+1}) \,|\, \mathcal{F}_t] \leq f(\mathbf{x}^t) - \alpha_t \left(\nabla f(\mathbf{x}^t)\right)^\top \boldsymbol{\phi}(\nabla f(\mathbf{x}^t)) + \frac{L}{2}\alpha_t^2\, d^2\, C_2^2. \tag{5.18}$$

Next, using Lemma 5.5, and the fact that $\alpha_t = a/(t+1)^\delta$, we obtain that. a.s.:

$$\mathbb{E}[f(\mathbf{x}^{t+1}) \,|\, \mathcal{F}_t] \leq f(x^t) - \frac{c'}{(t+1)}\|\nabla f(\mathbf{x}^t)\|^2 + \frac{L}{2}\frac{a^2\, d^2\, C_2^2}{(t+1)^{2\delta}}, \tag{5.19}$$

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 5.2, claims (2) and (3), we obtain the desired MSE rate result.

We next consider the case when Assumption 7 is replaced with Assumption 6 and Assumption 4 is additionally imposed. Following analogous arguments as in the first part of the proof, it can be shown that, a.s.:

$$\mathbb{E}[f(\mathbf{x}^{t+1}) \,|\, \mathcal{F}_t] \leq f(\mathbf{x}^t) - \alpha_t\, \boldsymbol{\phi}(\nabla f(\mathbf{x}^t))^\top \nabla f(\mathbf{x}^t) + \frac{L}{2}\alpha_t^2 \left(C_{13} + C_{14}\mathbb{E}[\,\|\boldsymbol{\nu}^t\|^2 \,|\, \mathcal{F}_t\,]\right), \tag{5.20}$$

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 (5.20), subtracting $f^\star$ from both sides of the inequality, taking expectation, and applying Theorem 5.2, 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 9. We first state and prove the following Lemma on the properties of function $\mathcal{N}$.

LEMMA 6.1. *Under Assumption 9, 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: $\rho - \|\mathbf{x}\| \leq \|\mathbf{x}+\mathbf{u}\| \leq \|\mathbf{x}\| + \|\mathbf{u}\| = \|\mathbf{x}\| + \rho$, and similarly, $\rho + \|\mathbf{x}\| \geq \|\mathbf{x}-\mathbf{u}\| \geq \rho - \|\mathbf{x}\|$. As $\mathcal{N}$ is non-increasing, it follows that: $\mathcal{N}(\|\mathbf{x}+\mathbf{u}\|) \leq \mathcal{N}(\rho - \|\mathbf{x}\|)$, and $\mathcal{N}(\|\mathbf{x}-\mathbf{u}\|) \geq \mathcal{N}(\rho + \|\mathbf{x}\|)$. Now, we have:

$$(\|\mathbf{u}\| - \|\mathbf{x}\|)\mathcal{N}(\|\mathbf{x}+\mathbf{u}\|) \leq (\rho - \|\mathbf{x}\|)\mathcal{N}(\rho - \|\mathbf{x}\|), \tag{6.3}$$

and similarly:

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

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 (6.3) and (6.4), 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:

$$\boldsymbol{\phi}(\mathbf{w}) = \int \boldsymbol{\Psi}(\mathbf{w}+\mathbf{u})p(\mathbf{u})d\mathbf{u} = \mathbb{E}[\boldsymbol{\Psi}(\mathbf{w}+\boldsymbol{\nu}^0)], \tag{6.5}$$

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. *Under Assumptions 8 and 9, the following holds:*

$$\boldsymbol{\phi}(\mathbf{x})^\top \mathbf{x} \geq 2(1-\kappa)\|\mathbf{x}\|^2 \int\limits_{\mathcal{J}(\mathbf{x})} \mathcal{N}(\|\mathbf{x}\| + \|\mathbf{u}\|)p(\mathbf{u})d\mathbf{u}, \tag{6.6}$$

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 $\boldsymbol{\Psi}(\mathbf{a}) = \mathbf{a}\mathcal{N}(\|\mathbf{a}\|)$, we have:

(6.7)

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

$$= \int_{J_1(\mathbf{x})=\{\mathbf{u}:\, \mathbf{u}^\top\mathbf{x}\geq 0\}} \mathcal{M}(\mathbf{x},\mathbf{u})p(\mathbf{u})d\mathbf{u} + \int_{J_2(\mathbf{x})=\{\mathbf{u}:\, \mathbf{u}^\top\mathbf{x}<0\}} \mathcal{M}(\mathbf{x},\mathbf{u})p(\mathbf{u})d\mathbf{u}. \tag{6.8}$$

Note also that there holds: $\mathcal{M}(\mathbf{x},\mathbf{u}) = (\|\mathbf{x}\|^2 + \mathbf{u}^\top\mathbf{x})\mathcal{N}(\|\mathbf{x}+\mathbf{u}\|)$; and $\mathcal{M}(\mathbf{x},-\mathbf{u}) = (\|\mathbf{x}\|^2 - \mathbf{u}^\top\mathbf{x})\mathcal{N}(\|\mathbf{x}-\mathbf{u}\|)$. Therefore, using the fact that $p(\mathbf{u}) = p(-\mathbf{u})$, for all $u \in \mathbb{R}^d$, we obtain: $\boldsymbol{\phi}(\mathbf{x})^\top\mathbf{x} = \int_{J_1(\mathbf{x})} \mathcal{M}_2(\mathbf{x},\mathbf{u})\, p(\mathbf{u})d\mathbf{u}$, 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:

$$\begin{aligned}\mathcal{M}_2(\mathbf{x},\mathbf{u}) \geq\ & \|\mathbf{x}\|^2[\mathcal{N}(\|\mathbf{x}+\mathbf{u}\|) + \mathcal{N}(\|\mathbf{x}-\mathbf{u}\|)] - \\ & - \|\mathbf{u}\|\|\mathbf{x}\|\,|\mathcal{N}(\|\mathbf{x}+\mathbf{u}\|) - \mathcal{N}(\|\mathbf{x}-\mathbf{u}\|)|.\end{aligned} \tag{6.9}$$

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:

$$\begin{aligned}\mathcal{M}_2(\mathbf{x},\mathbf{u}) &\geq \|\mathbf{x}\|^2[\mathcal{N}(\|\mathbf{x}+\mathbf{u}\|)+\mathcal{N}(\|\mathbf{x}-\mathbf{u}\|)]- \\ &\quad \|\mathbf{u}\|\|\mathbf{x}\|\frac{\|\mathbf{x}\|}{\|\mathbf{u}\|}|\mathcal{N}(\|\mathbf{x}+\mathbf{u}\|)+\mathcal{N}(\|\mathbf{x}-\mathbf{u}\|)| = 0.\end{aligned} \tag{6.10}$$

Therefore, we have: $\mathcal{M}_2(\mathbf{x},\mathbf{u}) \geq 0$, for any $\mathbf{u} \in J_1(\mathbf{x})$, $\|\mathbf{u}\| > \|\mathbf{x}\|$. 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:

$$\begin{aligned}\mathcal{M}_2(\mathbf{x},\mathbf{u}) &\geq \|\mathbf{x}\|^2[\mathcal{N}(\|\mathbf{x}+\mathbf{u}\|)+\mathcal{N}(\|\mathbf{x}-\mathbf{u}\|)] \\ &\quad - \|\mathbf{u}\|\|\mathbf{x}\|\kappa\underbrace{|\mathcal{N}(\|\mathbf{x}+\mathbf{u}\|)-\mathcal{N}(\|\mathbf{x}-\mathbf{u}\|)|} \\ &\geq (1-\kappa)\|\mathbf{x}\|^2(\mathcal{N}(\|\mathbf{x}+\mathbf{u}\|)+\mathcal{N}(\|\mathbf{x}-\mathbf{u}\|)).\end{aligned} \tag{6.11}$$

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

$$\begin{aligned}\mathcal{M}_2(\mathbf{x},\mathbf{u}) &\geq \|\mathbf{x}\|^2[\mathcal{N}(\|\mathbf{x}+\mathbf{u}\|)+\mathcal{N}(\|\mathbf{x}-\mathbf{u}\|)]- \\ &\quad \underbrace{\|\mathbf{u}\|}_{\leq\|\mathbf{x}\|}\|\mathbf{x}\|\kappa|\underbrace{\mathcal{N}(\|\mathbf{x}+\mathbf{u}\|)}_{\geq 0}+\underbrace{\mathcal{N}(\|\mathbf{x}-\mathbf{u}\|)|}_{\geq 0} \\ &\geq (1-\kappa)\|\mathbf{x}\|^2(\mathcal{N}(\|\mathbf{x}+\mathbf{u}\|)+\mathcal{N}(\|\mathbf{x}-\mathbf{u}\|)).\end{aligned} \tag{6.12}$$

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:

$$\begin{aligned}\mathcal{M}_2(\mathbf{x},\mathbf{u}) &\geq (1-\kappa)\|\mathbf{x}\|^2(\underbrace{\mathcal{N}(\|\mathbf{x}+\mathbf{u}\|)}_{\geq\mathcal{N}(\|x\|+\|u\|)}+\underbrace{\mathcal{N}(\|\mathbf{x}-\mathbf{u}\|)}_{\geq\mathcal{N}(\|\mathbf{x}\|+\|\mathbf{u}\|)}) \\ &\geq 2(1-\kappa)\|\mathbf{x}\|^2\mathcal{N}(\|\mathbf{x}\|+\|\mathbf{u}\|), \text{ for any } \mathbf{u} \in \mathcal{J}(\mathbf{x}).\end{aligned} \tag{6.13}$$

From (6.13), we finally get:

$$\begin{aligned}\boldsymbol{\phi}(\mathbf{x})^\top\mathbf{x} &\geq \int\limits_{\mathcal{J}(\mathbf{x})} 2(1-\kappa)\|\mathbf{x}\|^2\mathcal{N}(\|\mathbf{x}\|+\|\mathbf{u}\|)p(\mathbf{u})d\mathbf{u} \\ &= 2(1-\kappa)\|\mathbf{x}\|^2\int\limits_{\mathcal{J}(\mathbf{x})}\mathcal{N}(\|\mathbf{x}\|+\|\mathbf{u}\|)p(\mathbf{u})d\mathbf{u}.\end{aligned} \tag{6.14}$$

∎

LEMMA 6.3. *Let Assumptions 1, 8, and Assumption 9 with condition 3. hold (the nonlinearity with bounded outputs case). Then, for each $t = 1,2,...,$ we have:*

$$\|\nabla f(\mathbf{x}^t)\| \leq G'_t := L\,(\,a\,C'_2\,\frac{t^{1-\delta}}{1-\delta} + \|\mathbf{x}^0-\mathbf{x}^\star\|\,). \tag{6.15}$$

Proof. The proof is analogous to the proof of Lemma 5.4. ∎

6.2. Proofs of Theorems 4.1, 4.2, and 4.3: Joint nonlinearities. We are now ready to prove the results for the joint nonlinearities case.

Proof (Proof of Theorem 4.1) We carry out the proof again by verifying conditions B1-B5 in Theorem 5.1. Algorithm (2.3) admits again the representation in Theorem 5.1 with

$$\mathbf{r}(\mathbf{x}) = -\boldsymbol{\phi}(\nabla f(\mathbf{x})),\ \boldsymbol{\gamma}(t+1,\mathbf{x},\omega) = \boldsymbol{\phi}(\nabla f(\mathbf{x})) - \boldsymbol{\Psi}(\nabla f(\mathbf{x})+\boldsymbol{\nu}^t). \tag{6.16}$$

Conditions B1 and B2 hold analogously to the proof of Theorem 3.1. 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. □ We next prove Theorem 4.3. *Proof* (Proof of Theorem 4.3) We carry out the proof again by verifying conditions C1–C5 in Equation (8.2). The conditions C2–C5 are verified analogously as in the proof of Theorem 3.3. For condition C1, first fix an arbitrary $\mathbf{u} \neq 0$, and consider points $\mathbf{x}$ in the vicinity of $\mathbf{x}^\star$. Then, using the differentiability of $\mathcal{N}(a)$ for $a \neq 0$ and the differentiability of ∇f, it can be shown that:

$$\begin{aligned}\boldsymbol{\Psi}(\mathbf{u}+\nabla f(\mathbf{x})) &= \mathbf{u}\mathcal{N}(\|\mathbf{u}\|) + \mathcal{N}(\|\mathbf{u}\|)\nabla^2 f(\mathbf{x}^\star)(\mathbf{x}-\mathbf{x}^\star) \\ &+ \mathcal{N}'(\|\mathbf{u}\|)\frac{\mathbf{u}\mathbf{u}^\top}{\|\mathbf{u}\|}\nabla^2 f(\mathbf{x}^\star)(\mathbf{x}-\mathbf{x}^\star) + o(\|\mathbf{x}-\mathbf{x}^\star\|).\end{aligned}$$

We next integrate the above equality with respect to the joint pdf $p(\mathbf{u})$. For the first term above, note that $\int \mathcal{N}(\|\mathbf{u}\|)\mathbf{u}p(\mathbf{u})d\mathbf{u} = 0$, because $p(\mathbf{u}) = p(-\mathbf{u})$, for all $\mathbf{u}$. The second term is integrable as $\sup_{a>0}\mathcal{N}(a) < \infty$ (Assumption 13). The third term is integrable as function $a \mapsto a\mathcal{N}(a)$ is by assumptions non-decreasing; then, by taking its derivative, it follows that $|\mathcal{N}'(a)| \leq \mathcal{N}(a)/a$, $a > 0$, and so $\|\mathbf{u}\mathbf{u}^\top\mathcal{N}'(\|\mathbf{u}\|)\|/\|\mathbf{u}\| \leq \mathcal{N}(\|\mathbf{u}\|)$. Now, using the definition of $\mathbf{r}(\mathbf{x})$, it follows that $\mathbf{r}(\mathbf{x})$ admits the representation (5.2), with:

$$\mathbf{B} = -\left(\int \mathcal{N}(\|\mathbf{u}\|)p(\mathbf{u})d\mathbf{u} + \int_{\mathbf{u}\neq 0}\frac{\mathbf{u}\mathbf{u}^\top}{\|\mathbf{u}\|}\mathcal{N}'(\|\mathbf{u}\|)p(\mathbf{u})d\mathbf{u}\right)\nabla^2 f(\mathbf{x}^\star).$$

The conditions C1–C5 hold; thus, the result. □

We are now ready to prove Theorem 4.2.

Proof (Proof of Theorem 4.2) We first consider the case when Assumptions 1, 8, 9, and 10 hold. Analogously to the proof of 3.2, it can be shown that, a.s.:

$$\mathbb{E}[f(\mathbf{x}^{t+1}) \,|\, \mathcal{F}_t] \leq f(\mathbf{x}^t) - \alpha_t\, \boldsymbol{\phi}(\nabla f(\mathbf{x}^t))^\top \nabla f(\mathbf{x}^t) + \alpha_t^2\, C_{17}, \tag{6.17}$$

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

$$\big(\boldsymbol{\phi}(\mathbf{a})\big)^\top \mathbf{a} \geq 2(1-\kappa)\|\mathbf{a}\|^2 \int_{\mathcal{J}} \mathcal{N}(\|\mathbf{a}\| + \|\mathbf{u}\|)p(\mathbf{u})d\mathbf{u}, \tag{6.18}$$

where we recall $\mathcal{J} = \{\mathbf{u} : \frac{\mathbf{u}^\top\mathbf{a}}{\|\mathbf{u}\|\|\mathbf{a}\|} \in [0,\kappa]\}$, and $\kappa \in (0,1)$ 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 8. Then, for all $\mathbf{u}$ such that $\|\mathbf{u}\| \leq B_0$, there holds $\mathcal{N}(\|\mathbf{a}\| + \|\mathbf{u}\|) \geq \min\left\{\frac{\mathcal{N}(1)}{\|\mathbf{a}\|+B_0}, \mathcal{N}(1)\right\}$. We now have, a.s.:

$$\|\nabla f(\mathbf{x}^t)\|^2 \int_{\mathcal{J}} \mathcal{N}\left(\|\nabla f(\mathbf{x}^t)\| + \|\mathbf{u}\|\right)p(\mathbf{u})d\mathbf{u} \tag{6.19}$$

$$\geq \|\nabla f(\mathbf{x}^t)\|^2 \int_{\mathcal{J}_4} \min\left\{\frac{\mathcal{N}(1)}{B_0 + \|\nabla f(\mathbf{x}^t)\|}, \mathcal{N}(1)\right\}p(\mathbf{u})d\mathbf{u} \tag{6.20}$$

$$\geq \|\nabla f(\mathbf{x}^t)\|^2 \frac{\mathcal{N}(1)}{B_0 + G_t'}\int_{\mathcal{J}_4} p(\mathbf{u})d\mathbf{u}. \tag{6.21}$$

Here, $J_4 = \{u \in \mathbb{R}^d : \frac{\mathbf{u}^\top \nabla f(\mathbf{x}^t)}{\|\mathbf{u}\|\|\nabla f(\mathbf{x}^t)\|} \in [0, \kappa],\ \|\mathbf{u}\| \le B_0\}$. In (6.20), we used the fact that $\mathcal{N}(a)$ is non-negative for any $a \ge 0$, and in (6.21), we used Lemma 6.3.

Therefore, we have that, almost surely, for sufficiently large t:

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

for some positive constant C_{18}.

Combining the last bound with Lemmas 6.2 and 6.3, in view of condition 2. in Assumption 8, we obtain that, for sufficiently large t, a.s.:

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

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 (6.17) we obtain an equivalent to (5.19). Therein, c' in (5.19) is replaced with a positive constant c'' that can be taken as $c'' = \frac{4\,a\,(1-\kappa)\lambda(\kappa)(1-\delta)\mathcal{N}(1)}{L\,(\,a\,C'_2 + \|\mathbf{x}^0 - \mathbf{x}^\star\|\,) + B_0)}$. We now proceed analogously to the proof of Theorem 3.2, by applying claims (2) and (3) of Theorem 5.2. The desired MSE result now follows, with the rate ζ being any positive number less than

$$\min\left\{2\delta - 1,\ \frac{4\,a\,\mu\,(1-\kappa)\lambda(\kappa)(1-\delta)\mathcal{N}(1)}{L\,(\,a\,C'_2 + \|\mathbf{x}^0\| + \|\mathbf{x}^\star\|\,) + B_0}\right\}. \tag{6.23}$$

We now consider the case when Assumptions 1, 8, 9, 11, and 12 hold. We have, by assumption, 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., $\mathcal{N}(a) \ge C_{20}$, for each a, for some constant $C_{20} > 0$. Then, it follows that, a.s.:

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

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 5.2. □

7. Experiments. In order to benchmark the proposed nonlinear SGD framework, we consider `Heart`, `Diabetes` and `Australian` datasets from the LibSVM library [9]. We consider the logistic regression loss function for binary classification, see, e.g., [15], where function f in (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 [15] (see Figure 2 in [15]), we have, near the solution $\mathbf{x}^\star$, the following behavior with respect to gradient noise. (See also [15] 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 [15]. 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 [15]. 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 [15]. 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 [15]. 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.

REFERENCES

[1] D. Alistarh, D. Grubic, J. Li, R. Tomioka, and M. Vojnovic, *QSGD: Communication-efficient sgd via gradient quantization and encoding*, in Advances in Neural Information Processing Systems, 2017, pp. 1709–1720.

[2] L. Balles, F. Pedregosa, and N. L. Roux, *The geometry of sign gradient descent*, arXiv preprint arXiv:2002.08056, (2020).

[3] H. Bercovici and V. Pata, *Stable laws and domains of attraction in free probability theory*, Ann. of Math., 149 (1999), pp. 1023–1060.

[4] J. Bernstein, Y.-X. Wang, K. Azizzadenesheli, and A. Anandkumar, *signsgd: Compressed optimisation for non-convex problems*, in International Conference on Machine Learning, PMLR, 2018, pp. 560–569.

[5] L. Bottou, *Large-scale machine learning with stochastic gradient descent*, in Proceedings of COMPSTAT'2010, Springer, 2010, pp. 177–186.

[6] L. Bottou, F. E. Curtis, and J. Nocedal, *Optimization methods for large-scale machine learning*, Siam Review, 60 (2018), pp. 223–311.

[7] R. H. Byrd, G. M. Chin, W. Neveitt, and J. Nocedal, *On the use of stochastic hessian information in optimization methods for machine learning*, SIAM Journal on Optimization, 21 (2011), pp. 977–995.

[8] V. Cevher, S. Becker, and M. Schmidt, *Convex optimization for big data: Scalable, randomized, and parallel algorithms for big data analytics*, IEEE Signal Processing Magazine, 31 (2014), pp. 32–43.

[9] C.-C. Chang and C.-J. Lin, *Libsvm: a library for support vector machines*, ACM transactions on intelligent systems and technology (TIST), 2 (2011), pp. 1–27.

[10] S. Dasarathan, C. Tepedelenlioğlu, M. K. Banavar, and A. Spanias, *Robust consensus in the presence of impulsive channel noise*, IEEE Transactions on Signal Processing, 63 (2015), pp. 2118–2129.

[11] D. Davis, D. Drusvyatskiy, L. Xiao, and J. Zhang, *From low probability to high confidence in stochastic convex optimization.*, J. Mach. Learn. Res., 22 (2021), pp. 49–1.

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

[13] S. Ghadimi and G. Lan, *Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization I: A generic algorithmic framework*, SIAM J. Optim., 22 (2012), pp. 1469–1492.

[14] S. Ghadimi and G. Lan, *Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization, II: Shrinking procedures and optimal algorithms*, SIAM J. Optim., 23 (2013), pp. 2061–2089.

[15] E. Gorbunov, M. Danilova, and A. Gasnikov, *Stochastic optimization with heavy-tailed noise via accelerated gradient clipping*, arXiv preprint arXiv:2005.10785, (2020).

[16] E. Gorbunov, F. Hanzely, and P. Richtárik, *A unified theory of sgd: Variance reduction, sampling, quantization and coordinate descent*, in International Conference on Artificial Intelligence and Statistics, PMLR, 2020, pp. 680–690.

[17] M. Gurbuzbalaban, U. Simsekli, and L. Zhu, *The heavy-tail phenomenon in sgd*, in International Conference on Machine Learning, PMLR, 2021, pp. 3964–3975.

[18] S. Horváth, D. Kovalev, K. Mishchenko, S. Stich, and P. Richtárik, *Stochastic distributed learning with gradient quantization and variance reduction*, arXiv preprint arXiv:1904.05115, (2019).

[19] A. Juditsky, A. Nazin, A. Nemirovsky, and A. Tsybakov, *Algorithms of robust stochastic optimization based on mirror descent method*, arXiv:1907.02707, (2019).

[20] S. Kar, J. M. Moura, and K. Ramanan, *Distributed parameter estimation in sensor networks: Nonlinear observation models and imperfect communication*, IEEE Transactions on Information Theory, 58 (2012), pp. 3575–3605.

[21] 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 (2011), pp. 674–690.

[22] U. A. Khan, S. Kar, and J. M. Moura, *Distributed average consensus: Beyond the realm of linearity*, in 2009 Conference Record of the Forty-Third Asilomar Conference on Signals, Systems and Computers, IEEE, 2009, pp. 1337–1342.

[23] L. Lei and M. I. Jordan, *On the adaptivity of stochastic gradient-based optimization*, SIAM Journal on Optimization, 30 (2020), pp. 1473–1500.

[24] H. Mania, X. Pan, D. Papailiopoulos, B. Recht, K. Ramchandran, and M. I. Jordan,

Perturbed iterate analysis for asynchronous stochastic optimization, SIAM Journal on Optimization, 27 (2017), pp. 2202–2229.
[25] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, *Robust stochastic approximation approach to stochastic programming*, SIAM Journal on optimization, 19 (2009), pp. 1574–1609.
[26] M. B. Nevelson and R. Z. Khasminskiĭ, *Stochastic approximation and recursive estimation*, vol. 47, American Mathematical Soc., 1976.
[27] F. Niu, B. Recht, C. Ré, and S. J. Wright, *Hogwild!: A lock-free approach to parallelizing stochastic gradient descent*, arXiv preprint arXiv:1106.5730, (2011).
[28] R. Pascanu, T. Mikolov, and Y. Bengio, *On the difficulty of training recurrent neural networks*, in International Conference on Machine Learning, PMLR, 2013, pp. 1310–1318.
[29] V. Pichapati, A. T. Suresh, F. X. Yu, S. J. Reddi, and S. Kumar, *Adaclip: Adaptive clipping for private sgd*, arXiv preprint arXiv:1908.07643, (2019).
[30] B. T. Polyak and Y. Z. Tsypkin, *Adaptive estimation algorithms: convergence, optimality, stability*, Avtomatika i Telemekhanika, (1979), pp. 71–84.
[31] A. Shapiro, D. Dentcheva, and A. Ruszczynski, *Lectures on stochastic programming: modeling and theory*, SIAM, 2021.
[32] U. Simsekli, M. Gürbüzbalaban, T. H. Nguyen, G. Richard, and L. Sagun, *On the heavy-tailed theory of stochastic gradient descent for deep neural networks*, arXiv preprint arXiv:1912.00018, (2019).
[33] S. S. Stanković, M. Beko, and M. S. Stanković, *A robust consensus seeking algorithm*, in IEEE EUROCON 2019-18th International Conference on Smart Technologies, IEEE, 2019, pp. 1–6.
[34] S. Sundaram and B. Gharesifard, *Consensus-based distributed optimization with malicious nodes*, in 2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton), IEEE, 2015, pp. 244–249.
[35] F. Yousefian, A. Nedić, and U. V. Shanbhag, *On stochastic gradient and subgradient methods with adaptive steplength sequences*, Automatica, 48 (2012), pp. 56–67.
[36] J. Zhang, T. He, S. Sra, and A. Jadbabaie, *Why gradient clipping accelerates training: A theoretical justification for adaptivity*, arXiv preprint arXiv:1905.11881, (2019).
[37] J. Zhang, S. P. Karimireddy, A. Veit, S. Kim, S. J. Reddi, S. Kumar, and S. Sra, *Why are adaptive methods good for attention models?*, arXiv preprint arXiv:1912.03194, (2019).

Appendix.

A. Proof of Theorem 5.2. We first state and prove the following Lemma.

Lemma 8.1. *Consider (deterministic) sequence*

$$v^{t+1} = \left(1 - \frac{a_3}{(t+1)^\delta}\right) v^t + \frac{a_4}{(t+1)^\delta},\ t \geq t_0,$$

with $a_3, a_4 > 0$ *and* $0 < \delta \leq 1$, $t_0 > 0$, *and* $v^{t_0} \geq 0$. *Further, assume that* t_0 *is such that* $\frac{a_3}{(t+1)^\delta} \leq 1$, *for all* $t \geq t_0$. *Then,* $\lim_{t\to\infty} v^t = \frac{a_4}{a_3}$.

Proof. Let $e^t = v^t - \frac{a_4}{a_3}$. It is easy to verify that:

$$e^{t+1} = \left(1 - \frac{a_3}{(t+1)^\delta}\right) e^t,\ t \geq t_0.$$

Then, for all $t \geq t_0$, there holds:

$$|e^{t+1}| = \left(1 - \frac{a_3}{(t+1)^\delta}\right) |e^t| \leq \exp\left(-a_3 \sum_{s=t_0}^{t} \frac{1}{(s+1)^\delta}\right) |e^{t_0}| \tag{8.1}$$

where in (8.1) we used the inequality $1 + a \leq \exp(a)$, $a > 0$. Letting $t \to \infty$ and the fact that $\delta \leq 1$ so that the sequence $\frac{1}{(s+1)^\delta}$, $s \geq t_0$, is non-summable, we obtain that $e^t \to 0$, which in turn implies the claim of the Lemma. ∎

We now continue with proving Theorem 5.2. First, let us prove claim (1). Note that:

$$z^{t+1} \leq \left(1 - \frac{a_1}{(t+1)^{\delta_1}}\right) z^t + \frac{a_2}{(t+1)^{\delta_2}},\ t \geq t'. \tag{8.2}$$

Multiplying the above inequality with $(t+1)^{\delta_2-\delta_1}$, defining $\widehat{z}^t = t^{\delta_2-\delta_1}\, z^t$, we get:

$$\widehat{z}^{t+1} \leq \left(1 - \frac{a_1}{(t+1)^{\delta_1}}\right) (1+1/t)^{\delta_2-\delta_1}\, \widehat{z}^t + \frac{a_2}{(t+1)^{\delta_1}}.$$

Next, using, e.g., a Taylor expansion of function $a \mapsto (1+a)^{\delta_2-\delta_1}$, it can be shown that $(1+1/t)^{\delta_2-\delta_1} \leq 1 + \frac{2(\delta_2-\delta_1)}{t}$, for any $t \geq t_\delta$, for appropriately chosen $t_\delta > 0$. Therefore,

$$\begin{aligned}
&\left(1 - \frac{a_1}{(t+1)^{\delta_1}}\right) (1+1/t)^{\delta_2-\delta_1} \\
&\leq 1 - \frac{a_1}{(t+1)^{\delta_1}} + \frac{2(\delta_2-\delta_1)}{t} - \frac{2a_1(\delta_2-\delta_1)}{t(t+1)^{\delta_1}} \leq 1 - \frac{a_1}{2(t+1)^{\delta_1}},
\end{aligned}$$

for any $t \geq t_1$, for appropriately taken $t_1 > 0$. Using the latter bound, we obtain: $\widehat{z}^{t+1} \leq \left(1 - \frac{a_1}{2(t+1)^{\delta_1}}\right) \widehat{z}^t + \frac{a_2}{(t+1)^{\delta_1}}$, $t \geq t_1$. Now, applying Lemma 8.1, we obtain that $\widehat{z}^t = O(1)$, and therefore $z^t = O(1/t^{\delta_2-\delta_1})$. This proves claim (1) in Theorem 5.2.

We now prove claim (2). Multiplying (8.2) by $(t+1)^{\delta_2-1}$, and defining $\widehat{z}^t = t^{\delta_2-1}\, z^t$, we obtain:

$$\begin{aligned}
\widehat{z}^{t+1} &\leq \left(1 - \frac{a_1}{(t+1)}\right) (1+1/t)^{\delta_2-1} \widehat{z}^t + \frac{a_2}{t+1} \\
&\leq \left(1 - \frac{a_1 - (\delta_2 - 1)}{t} + \frac{C_{22}}{t^2}\right) \widehat{z}^t + \frac{a_2}{t+1} \qquad (8.3) \\
&\leq \left(1 - \frac{a_1 - (\delta_2 - 1)}{2\,(t+1)}\right) \widehat{z}^t + \frac{a_2}{t+1},\ t \geq t_2, \qquad (8.4)
\end{aligned}$$

for appropriately chosen $t_2 > 0$ and $C_{22} > 0$. In (8.3), we used the fact that $(1 + 1/t)^{\delta_2-1} \leq 1 + \frac{\delta_2-1}{t} + \frac{C_{23}}{t^2}$, for all $t \geq 1$ and some $C_{23} > 0$ (the inequality can be obtained, e.g., via a Taylor approximation). The claim (2) of Theorem 5.2 now follows by applying Lemma 8.1 to (8.4).

We now prove claim (3). Let $a_1 < \delta_2 - 1$, and fix an arbitrary positive number ζ, $\zeta < a_1$. Then, we have, for $\widehat{z}^t = t^\zeta z^t$:

$$\begin{aligned}
\widehat{z}^{t+1} &\leq \left(1 - \frac{a_1}{(t+1)}\right) (1+1/t)^{\zeta} \widehat{z}^t + \frac{a_2}{(t+1)^{\delta_2-\zeta}} \\
&\leq \left(1 - \frac{a_1 - \zeta}{t} + \frac{C_{24}}{t^2}\right) \widehat{z}^t + \frac{a_2}{(t+1)^{\delta_2-\zeta}} \\
&\leq \left(1 - \frac{a_1 - \zeta}{2\,(t+1)}\right) \widehat{z}^t + \frac{a_2}{t+1},\ t \geq t_3,
\end{aligned}$$

for appropriately chosen $t_3 > 0$ and $C_{24} > 0$. In the last inequality, we used the fact that $\zeta < a_1 \leq \delta_2 - 1$, and so $\delta_2 - \zeta > 1$. Finally, applying Lemma 8.1, claim (3) follows. □

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 Assumption 1, Assumption 3, and Assumption 5, 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:

$$x^{t+1} = (1 - \alpha_t)\, x^t - \alpha_t\, \nu^t,\ t = 0, 1, ..., \tag{8.5}$$

with arbitrary deterministic initialization $x^0 \in \mathbb{R}$. Then, squaring (8.5), 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 $t \geq 1$.

C. Extension of Theorem 3.2 for gradient noise vector with mutually dependent entries. We show that Theorem 3.2 continues to hold when we have an i.i.d. zero mean noise vector sequence $\{\boldsymbol{\nu}^t\}$ 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 $\mathbf{y} \neq 0$:

$$\int \psi(\mathbf{y} + \mathbf{u})^\top \mathbf{y}\, p(\mathbf{u})\, d\mathbf{u}. \tag{8.6}$$

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:

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

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

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 5.3. In particular, ϕ_i is also odd, and hence:

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

$$= \sum_{i=1}^d \phi_i(y_i) y_i = \sum_{i=1}^d |\phi_i(y_i)|\, |y_i|.$$

The last inequality holds because, for any $i = 1, ..., d$, quantities $\phi_i(y_i)$ and y_i have equal sign. The proof now proceeds analogously to that of Theorem 3.2.

D. Proof of Lemma 5.3. The proof can be found in [30]; we include similar arguments for completeness. For claim 1., note that

$$\phi(a) = \int_{-\infty}^{+\infty} \Psi(a + u) p(u) du = -\int_{-\infty}^{+\infty} \Psi(-a - u) p(u) du$$

$$= -\int_{-\infty}^{+\infty} \Psi(-a + w) p(w) dw = -\phi(-a),$$

for any $a \in \mathbb{R}$, where we use the fact that Ψ is odd. For claim 2., note that $|\phi(a)| \leq \int_{-\infty}^{+\infty} |\Psi(a+u)|p(u)du \leq C_1 \int_{-\infty}^{+\infty} p(u)du = C_1$, where we used Assumption 7. Proof of claim 3. is similar to that of claim 2. For claim 4., note that $\phi(a) = \int_0^{+\infty} (\Psi(u+a) - \Psi(u-a))\, p(u)du$, and so, for $a' > a$, we have

$$\phi(a') - \phi(a) = \int_0^{+\infty} [(\Psi(u+a') - \Psi(u+a)) + \\ + (\Psi(u-a) - \Psi(u-a'))]\, p(u)du \geq 0,$$

because Ψ is non-decreasing. Finally, for claim 5., to show that $\phi'(0)$ is given by (5.4), see the proof of Lemma 6 in [30]. To verify that $\phi'(0)$ is strictly positive, consider first the case that Ψ has a discontinuity at zero. Then, because $p(0) > 0$ by Assumption 3, it follows from (5.4) that $\phi'(0) \geq (\Psi(0+) - \Psi(0-))p(0) > 0$. Otherwise, if Ψ is continuous at zero, we have: $\phi'(0) \geq \int_{-c}^{c} \Psi'(u)p(u)du > 0$, where $c > 0$ is taken such that $\Psi(u)$ is continuous and strictly increasing and $p(u)$ is strictly positive for $|u| < c$.[8] Such c exists in view of Assumptions 3 and 5.

E. Derivations for Example 3.3. We calculate the rate ζ in Theorem 3.2 for the component-wise clipping nonlinearity with saturation value m, $m > 1$. Here, it can be shown, by doing direct calculations, that

$$\phi(w) = 2\, w \int_0^{m-w} p(u)du + \int_{m-w}^{m+w} (m + w - u)p(u)du,\ w \in [0, m]. \tag{8.7}$$

Furthermore, it can be shown that (see Appendix F): $\phi'(0) = 2\int_0^m p(u)du$. Noting that the second integral in (8.7) is nonnegative, and using the form $p(u)$ in (3.2), we obtain:

$$\phi(w) \geq 2\, w \int_0^{m-w} p(u)du = w\left(1 - \frac{1}{(m-w+1)^{\alpha-1}}\right),\ w \in [0, m]. \tag{8.8}$$

Also, we have: $\phi'(0) = 1 - \frac{1}{(m+1)^{\alpha}}$. From the latter equation and (8.8), we estimate that ξ can be taken as: $\xi = m + 1 - \left(\frac{2}{1+(m+1)^{-\alpha}}\right)^{1/(\alpha-1)} \geq m - 1$, for any $\alpha > 2$, for any $m > 1$. Hence, we can also take $\xi = m - 1$. Substituting the obtained estimates for $\phi'(0)$ and ξ into the rate ζ, we obtain the rate estimate in (3.3).

F. Derivation of $\phi'(0)$ for Example 3.5. Consider the coordinate-wise clipping nonlinearity Ψ with floor level $m > 0$. The function Ψ here is piece-wise differentiable, with the derivative $\Psi'(a) = 1$, for $a \in (-m, m)$, and $\Psi'(a) = 0$, for $|a| > m$. We now apply claim 5. in Lemma 5.3 and use formula (5.4) for evaluating $\phi'(0)$. As the coordinate-wise clipping function does not have discontinuity points, (5.4) simplifies to the following:

$$\phi'(0) = \int_{u\in\mathbb{R},\, u\neq -m,\, u\neq m} \Psi'(u)p(u)\, du = \int_{-m}^{+m} p(u)du = 2\int_0^m p(u)du,$$

where the last equality uses symmetry of function $p(u)$.

[8] If there are some (at most countably many) points inside interval $(-c, c)$ where Ψ is continuous but not differentiable, these points are excluded from the integration set in $\int_{-c}^{c} \Psi'(u)p(u)du$ without change in the integration result.

G. Derivations for Example 3.6. We provide here details for the derivations in Example 3.6. We first calculate σ_Ψ^2; we have:

$$\sigma_\Psi^2 = \int_{-\infty}^{\infty} |\Psi(u)|^2 p(u) du = \int_{-\infty}^{\infty} p(u) du = 1.$$

Next, by direct integration, we have for $\alpha > 3$:

$$\begin{aligned}\sigma_\nu^2 &= 2\int_0^\infty p(u) u^2 du \\ &= -(\alpha-1)\frac{[\,(\alpha-1)u((\alpha-2)u+2)\,]+2}{(\alpha-3)(\alpha-2)(\alpha-1)(1+u)^{\alpha-2}}\,|_0^\infty = \frac{2}{(\alpha-3)(\alpha-2)}.\end{aligned}$$

On the other hand, for $\alpha \in (2,3]$, we clearly have $\sigma_\nu^2 = +\infty$. Finally, using claim 5. in Lemma 5.3, and using the fact that $\Psi'(u) = 0$, for all $u \neq 0$, we obtain:

$$\phi'(0) = p(0)\,(\Psi(0+) - \Psi(0-)) = 2\,p(0) = \alpha - 1.$$

H. Derivations for Example 4.1. We consider the (joint) gradient clipping nonlinearity $\mathbf{\Psi}$ with the clipping level $M > 0$, and we consider $p(\mathbf{u})$ in (4.1).

Consider rate ζ in Theorem 4.2 that, for a sufficiently large a, can be approximated as:

$$\min\left\{2\delta - 1,\ (1-\delta)\frac{4\,\mu\,(1-\kappa)\lambda(\kappa)\mathcal{N}(1)}{L\,C_2'}\right\}. \tag{8.9}$$

Here, κ is an arbitrary scalar in $(0,1)$, and, for the gradient clipping, we have that $\mathcal{N}(1) = C_2' = M$. Note that, regarding Assumption 8, quantity B_0 can be taken here to be an arbitrary positive number. Moreover, for $p(\mathbf{u})$ in (4.1), due to the radial symmetry, we have that

$$\lambda(\kappa) = \lambda(\kappa, B_0) = \frac{1}{\pi}\arccos(1-\kappa)\,\mathcal{P}(B_0),\ \ \kappa \in (0,1),$$

where $\mathcal{P}(B_0) = \int_{\mathbf{u}:\,\|\mathbf{u}\|\leq B_0} p(\mathbf{u}) d\mathbf{u} = 1 - \frac{1+(\alpha-1)B_0}{(1-B_0)^{\alpha-1}}$. We next maximize (8.10), i.e., we maximize $(1-\kappa)\lambda(\kappa, B_0)$ with respect to $\kappa \in (0,1)$, to get the largest (tightest) estimate of ζ. It is easy to see that $\max_{\kappa\in(0,1)}(1-\kappa)\lambda(\kappa, B_0) > 0.17\,\mathcal{P}(B_0)$. Substituting all the above developments into (8.10), we obtain:

$$\begin{aligned}\zeta &\approx \min\left\{2\delta - 1,\ (1-\delta)\frac{0.68\,\mu\,\mathcal{P}(B_0)}{L}\right\} \\ &= \min\left\{2\delta - 1,\ (1-\delta)\frac{0.68\,\mu}{L}\left(1 - \frac{1+(\alpha-1)B_0}{(1+B_0)^{\alpha-1}}\right)\right\}\end{aligned} \tag{8.10}$$

As B_0 can be arbitrary positive number, letting $B_0 \to +\infty$, we obtain the following rate estimate: $\min\left\{2\delta - 1,\ (1-\delta)\frac{0.68\,\mu}{L}.\right\}$. It is easy to see that the same rate estimate can be obtained for the normalized gradient nonlinearity. The only difference in the rate derivation is that therein $\mathcal{N}(1) = C_2' = 1$.