Convergence Rates of Stochastic Gradient Descent under Infinite Noise Variance

Hongjian Wang*
Carnegie Mellon University
hjnwang@cmu.edu

Mert Gürbüzbalaban
Rutgers University
mg1366@rutgers.edu

Lingjiong Zhu
Florida State University
zhu@math.fsu.edu

Umut Şimşekli
INRIA & ENS – PSL Research University
umut.simsekli@inria.fr

Murat A. Erdogdu
University of Toronto & Vector Institute
erdogdu@cs.toronto.edu

Abstract

Recent studies have provided both empirical and theoretical evidence illustrating that heavy tails can emerge in stochastic gradient descent (SGD) in various scenarios. Such heavy tails potentially result in iterates with diverging variance, which hinders the use of conventional convergence analysis techniques that rely on the existence of the second-order moments. In this paper, we provide convergence guarantees for SGD under a state-dependent and heavy-tailed noise with a potentially infinite variance, for a class of strongly convex objectives. In the case where the p-th moment of the noise exists for some $p \in [1, 2)$, we first identify a condition on the Hessian, coined 'p-positive (semi-)definiteness', that leads to an interesting interpolation between the positive semi-definite cone ($p = 2$) and the cone of diagonally dominant matrices with non-negative diagonal entries ($p = 1$). Under this condition, we provide a convergence rate for the distance to the global optimum in L^p. Furthermore, we provide a generalized central limit theorem, which shows that the properly scaled Polyak-Ruppert averaging converges weakly to a multivariate α-stable random vector. Our results indicate that even under heavy-tailed noise with infinite variance, SGD can converge to the global optimum without necessitating any modification neither to the loss function nor to the algorithm itself, as typically required in robust statistics. We demonstrate the implications of our results over misspecified models, in the presence of heavy-tailed data.

1 Introduction

We consider the unconstrained minimization problem

$$\underset{\boldsymbol{x}\in\mathbb{R}^n}{\text{minimize}}\ f(\boldsymbol{x}), \tag{1.1}$$

using the stochastic gradient descent (SGD) algorithm. Initialized at $\boldsymbol{x}_0 \in \mathbb{R}^n$, the SGD algorithm is given by the iterations,

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \gamma_{t+1}\big(\boldsymbol{\nabla} f(\boldsymbol{x}_t) + \boldsymbol{\xi}_{t+1}(\boldsymbol{x}_t)\big), \quad t = 0, 1, 2, ... \tag{1.2}$$

where $\{\gamma_t\}_{t\in\mathbb{N}^+}$ denotes the step-size sequence, and $\{\boldsymbol{\xi}_t\}_{t\in\mathbb{N}^+}$ is a martingale difference sequence adapted to a filtration $\{\mathcal{F}_t\}_{t\in\mathbb{N}}$, characterizing the noise in the gradient (the sequence $\{\boldsymbol{x}_t\}_{t\in\mathbb{N}}$ is also adapted to the same filtration, if we assume $\boldsymbol{x}_0$ is $\mathcal{F}_0$-measurable). Our focus is on the case where the noise is state dependent, and its variance is infinite, i.e., $\mathbb{E}\big[\|\boldsymbol{\xi}_t\|_2^2\big] = \infty$.

*Work partially conducted while affiliated with the Vector Institute.

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

Many problems in modern statistical learning can be written in the form (1.1), where $f(\boldsymbol{x})$ typically corresponds to the population risk, that is, $f(\boldsymbol{x}) := \mathbb{E}_{z\sim\nu}[\ell(\boldsymbol{x}, z)]$ for a given loss function ℓ and an unknown data distribution ν. In practice, one observes independent and identically distributed (i.i.d.) samples $z_i \sim \nu$ for $i \in [n]$, and estimates the population gradient $\boldsymbol{\nabla} f(\boldsymbol{x})$ with a noisy gradient at each iteration, which is based on an empirical average over a subset of the samples $\{z_i\}_{i\in[n]}$. Due to its simplicity, superior generalization performance, and well-understood theoretical guarantees, SGD has been the method of choice for minimization problems arising in statistical machine learning.

Starting from the pioneering works of Robbins and Monro [1951], Chung [1954], Sacks [1958], Fabian [1968], Ruppert [1988], Shapiro [1989], Polyak and Juditsky [1992], theoretical properties of the SGD algorithm and its variants have been receiving a growing attention under different scenarios. Recent works, for example Tripuraneni et al. [2018], Su and Zhu [2018], Duchi and Ruan [2021], Toulis and Airoldi [2017], Fang et al. [2018], Anastasiou et al. [2019], Yu et al. [2020] established convergence rates for SGD in various settings. By building on the analysis of Polyak and Juditsky [1992] to prove a central limit theorem (CLT) for the Polyak-Ruppert averaging, these works led to novel methodologies to compute confidence intervals using SGD. However, a recurring assumption in this line of work is the finite noise variance, which may be violated frequently in modern frameworks.

Heavy-tailed behavior in statistical methodology may naturally arise from the underlying model, or through the iterative optimization algorithm used during model training. In robust statistics, one often encounters heavy-tailed noise behavior in data, which in conjunction with standard loss functions leads to infinite noise variance in SGD. Very recently, heavy-tailed behavior is shown to emerge from the multiplicative noise in SGD, when the step-size is large and/or the batch-size is small [Hodgkinson and Mahoney, 2021, Gürbüzbalaban et al., 2021]. On the other hand, there is strong empirical evidence in modern machine learning that the gradient noise often exhibits a heavy-tailed behavior, which indicates an infinite variance. For example, this is observed in fully connected and convolutional neural networks [Şimşekli et al., 2019, Gürbüzbalaban and Hu, 2021] as well as recurrent neural networks [Zhang et al., 2020]. Thus, understanding the behavior of SGD under infinite noise variance becomes extremely important for at least two reasons. A *computational complexity reason:* modern machine learning and robust statistics frameworks lead to heavy-tailed behavior in SGD; thus, understanding the performance of this algorithm in terms of precise convergence rates as well as the required conditions on the step-size sequence as a function of the 'heaviness' of the tail become crucial in this setup. A *statistical reason:* many inference methods that rely on Polyak-Ruppert averaging utilize a CLT which holds under finite noise variance (see e.g. online bootstrap and variance estimation approaches [Fang et al., 2018, Su and Zhu, 2018, Chen et al., 2020]). Using the same methodology in the aforementioned modern frameworks (under heavy-tailed noise) will ultimately result in incorrect confidence intervals, jeopardizing the statistical procedure. Thus, establishing the limit distribution in this setting is of great importance.

In this work, we study the behavior of the SGD algorithm with diminishing step-sizes for a class of strongly convex problems when the noise variance is infinite. We establish the convergence rates of the SGD iterates towards the global minimum, and identify a sufficient condition on the Hessian of f, which interpolates between the positive semi-definite cone and the cone of diagonally dominant matrices (with non-negative diagonal entries). We further study the Polyak-Ruppert averaging of the SGD iterates, and show that the limit distribution is a multivariate α-stable distribution. We illustrate our theory on linear regression and generalized linear models, demonstrating how to verify the conditions of our theorems. Perhaps surprisingly, our results show that even under heavy-tailed noise with infinite variance, SGD with diminishing step-sizes can converge to the global optimum without requiring any modification neither to the loss function nor to the algorithm itself, as opposed to the conventional techniques used in robust statistics [Huber, 2004]. Finally, we argue that our work has potential implications in constructing confidence intervals in the infinite noise variance setting.

2 Preliminaries and Technical Background

Notational Conventions. By $\mathbb{N}$, $\mathbb{N}^+$ and $\mathbb{R}$ we denote the set of non-negative integers, positive integers, and real numbers, respectively. For $m \in \mathbb{N}^+$, we define $[m] = \{1, \dots, m\}$. We use italic letters (e.g. x, ξ) to denote scalars and scalar-valued functions, $\operatorname{sign}(x)$ to denote the sign of x, bold face italic letters (e.g. $\boldsymbol{x}, \boldsymbol{\xi}$) to denote vectors and vector-valued functions, and bold face upper case letters (e.g. $\mathbf{A}$) to denote matrices. We use $|\boldsymbol{x}|$ and $\|\boldsymbol{x}\|_p$ to denote the 2-norm and p-norm of a vector $\boldsymbol{x}$; $\|\mathbf{A}\|$ and $\|\mathbf{A}\|_p$ the operator 2-norm and operator p-norm of a matrix $\mathbf{A}$. The transpose

of a matrix $\mathbf{A}$ and a vector $\boldsymbol{x}$ (viewed as a matrix with 1 column) are denoted by $\mathbf{A}^\mathsf{T}$ and $\boldsymbol{x}^\mathsf{T}$. If $\{\mathbf{A}_i\}_{i\in\mathbb{N}}$ is a sequence of matrices and $k > \ell$, the empty product $\prod_{i=k}^{\ell} \mathbf{A}_i$ is understood to be the identity matrix $\mathbf{I}$. For two sequences of real numbers $\{a_t\}_{t\in\mathbb{N}}$, $\{b_t\}_{t\in\mathbb{N}}$, we write $a_t = \mathcal{O}(b_t)$ if $\limsup_{t\to\infty} |a_t|/|b_t| < \infty$, $a_t = o(b_t)$ if $\limsup_{t\to\infty} |a_t|/|b_t| = 0$, $a_t = \Theta(b_t)$ if both $a_t = \mathcal{O}(b_t)$ and $b_t = \mathcal{O}(a_t)$ hold, and $a_t \asymp b_t$ if $\lim_{t\to\infty} |a_t|/|b_t|$ exists and is in $(0,\infty)$. If $a_t = \mathcal{O}(b_t t^\varepsilon)$ for any $\varepsilon > 0$, we say $a_t = \tilde{\mathcal{O}}(b_t)$. Sufficiently large or sufficiently small positive constants whose values do not matter are written as $C, C_0, C_1, \ldots$, sometimes without prior introduction. If $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$ is a sequence of random vectors taking value in $\mathbb{R}^n$ and μ is a probability measure on $\mathbb{R}^n$, we write $\boldsymbol{X}_t \xrightarrow[t\to\infty]{\mathcal{D}} \mu$ if $\{\boldsymbol{X}_t\}_{t\in\mathbb{N}^+}$ converges in distribution (also called 'weak convergence') to μ.

Stochastic Approximation. In the SGD recursion (1.2), we can replace $\boldsymbol{\nabla} f$ with an arbitrary continuous function $\boldsymbol{R} : \mathbb{R}^n \to \mathbb{R}^n$, and consider the same iterations that stochastically approximate the zero $\boldsymbol{x}^*$ of $\boldsymbol{R}$,

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \gamma_{t+1}\big(\boldsymbol{R}(\boldsymbol{x}_t) + \boldsymbol{\xi}_{t+1}(\boldsymbol{x}_t)\big). \tag{2.1}$$

This is called the *stochastic approximation* process [Robbins and Monro, 1951], which is a predecessor of stochastic gradient descent (SGD) and describes a larger family of iterative algorithms (see e.g. [Kushner and Yin, 2003, Chapters 2 and 3]). Theoretical investigation of the recursion (2.1) has been active ever since its invention, especially under finite noise variance assumption: Robbins and Monro [1951] prove that the recursion (2.1) can lead to the L^2 convergence $\lim_{t\to\infty} \mathbb{E}[|\boldsymbol{x}_t - \boldsymbol{x}^*|^2] = 0$; Chung [1954] further calculates an exact convergence rate (see (3.6) in Section 3); Blum [1954] presents an elegant proof that the convergence of $\boldsymbol{x}_t$ to $\boldsymbol{x}^*$ can hold almost surely. The asymptotic distribution of (2.1) can be attributed to [Chung, 1954, Theorem 6], which states that the expression $\gamma_t^{-1/2}(\boldsymbol{x}_t - \boldsymbol{x}^*)$ converges weakly to a normal distribution. In their seminal works, Polyak and Juditsky [1992] and Ruppert [1988] independently introduce the concept of 'averaging the iterates',

$$\overline{\boldsymbol{x}}_t = \frac{\boldsymbol{x}_0 + \ldots + \boldsymbol{x}_{t-1}}{t},$$

showing the striking result that $\sqrt{t}(\overline{\boldsymbol{x}}_t - \boldsymbol{x}^*)$ converges weakly to a fixed normal distribution *regardless of the choice of the step-size* $\{\gamma_t\}_{t\in\mathbb{N}^+}$ as long as it satisfies mild conditions. Recently, optimization algorithms that can handle heavy-tailed noise sequence $\{\boldsymbol{\xi}_t\}_{t\in\mathbb{N}^+}$ have been proposed [Davis et al., 2019, Nazin et al., 2019, Gorbunov et al., 2020]; however, they still rely on a *uniformly* bounded variance assumption, hence do not cover our setting.

Compared with the copious collection of theoretical studies on stochastic approximation algorithms with finite variance as mentioned above, papers that study stochastic approximation under *infinite* noise variance are extremely scarce; we shall summarize only a few papers known to us. Krasulina [1969] is the first to consider such problems, proving almost sure and L^p convergence for the one-dimensional stochastic approximation process without variance. The weak convergence of the iterates (without averaging) $t^{1/\alpha}(\boldsymbol{x}_t - \boldsymbol{x}^*)$ is also considered by Krasulina [1969], but only for the fastest-decaying step-size $\gamma_t = 1/t$. Goodsell and Hanson [1976] discuss how $\boldsymbol{x}_t \to \boldsymbol{x}^*$ in probability can imply $\boldsymbol{x}_t \to \boldsymbol{x}^*$ almost surely, when no finite variance is assumed, and Li [1994] provides a necessary and sufficient condition for almost sure convergence of $\boldsymbol{x}_t \to \boldsymbol{x}^*$, stating that faster-decaying step-size $\gamma_t = o(t^{-1/p})$ is required when moments of lower orders $\mathbb{E}[|\boldsymbol{\xi}_t|^p]$ are not in place. Anantharam and Borkar [2012] show that although step-size that decays slower than $t^{-1/p}$ cannot yield almost sure convergence, L^p convergence can still hold under what they call the 'stability assumption', but their analysis technique provides no convergence rate. Recently, Şimşekli et al. [2019] and Zhang et al. [2020] considered SGD with heavy-tailed noise $\boldsymbol{\xi}_t$ having *uniformly bounded* p-th order moments. Besides not being able to handle state-dependent noise due to this uniform moment condition, Şimşekli et al. [2019] imposed further conditions on $\boldsymbol{R} = \boldsymbol{\nabla} f$ such as global Hölder continuity for a non-convex f, whereas Zhang et al. [2020] modified SGD with 'gradient clipping', in order to be able to compensate the effects of the heavy-tailed noise.

Finally, we shall mention that a class of stochastic recursions similar to (2.1) have been considered in the dynamical systems theory [Mirek, 2011, Buraczewski et al., 2012, 2016], for which generalized central limit theorems with α-stable limits have been established. However, such techniques typically require $\boldsymbol{R}$ to be (asymptotically) linear and the step-sizes to be constant as they heavily rely on the theory of time-homogeneous Markov processes. Hence, their approach does not readily generalize to the setting of our interest, i.e., non-linear $\boldsymbol{R}$ and diminishing step-sizes, where the latter is crucial for ensuring convergence towards the global optimum.

Stable Distributions. In probability theory, a random variable X is *stable* if its distribution is non-degenerate and satisfies the following property: Let X_1 and X_2 be independent copies of X. Then, for any constants $a, b > 0$, the random variable $aX_1 + bX_2$ has the same distribution as $cX + d$ for some constants $c > 0$ and d (see e.g. [Samorodnitsky and Taqqu, 1994]). The stable distribution is also referred to as the α-stable distribution, first proposed by Lévy [1937], where $\alpha \in (0, 2]$ denoting the stability parameter. The case $\alpha = 2$ corresponds to the normal distribution, and the variance under this distribution is undefined for any $\alpha < 2$. The multivariate α-stable distribution dates back to Feldheim [1937], which is a multivariate generalization of the univariate α-stable distribution, which is also uniquely characterized by its characteristic function. In particular, an $\mathbb{R}^n$-valued random vector X has a multivariate α-stable distribution, denoted as $\boldsymbol{X} \sim \mathcal{S}(\alpha, \Lambda, \delta)$ if the joint characteristic function of $\boldsymbol{X}$ is given by

$$\mathbb{E}\big[\exp\big(i\boldsymbol{u}^\mathsf{T}\boldsymbol{X}\big)\big] = \exp\Big\{ - \int_{\boldsymbol{s}\in S_2} (|\boldsymbol{u}^\mathsf{T}\boldsymbol{s}|^\alpha + i\nu(\boldsymbol{u}^\mathsf{T}\boldsymbol{s}, \alpha))\Lambda(\mathrm{d}\boldsymbol{s}) + i\boldsymbol{u}^\mathsf{T}\delta\Big\},$$

for any $\boldsymbol{u} \in \mathbb{R}^n$, and $0 < \alpha \leqslant 2$. Here, α is the tail-index, Λ is a finite measure on S_2 known as the spectral measure, $\boldsymbol{\delta} \in \mathbb{R}^n$ is a shift vector, and $\nu(y, \alpha) := -\operatorname{sign}(y)\tan(\pi\alpha/2)|y|^\alpha$ for $\alpha \neq 1$ and $\nu(y, \alpha) := (2/\pi)y\log|y|$ for $\alpha = 1$ for any $y \in \mathbb{R}$, and S_2 denotes the unit sphere in $\mathbb{R}^n$; i.e. $S_2 = \{\boldsymbol{s} \in \mathbb{R}^n : |\boldsymbol{s}| = 1\}$. Stable distribution also emerges as the limit distribution in the Generalized Central Limit Theorem (GCLT) [Gnedenko and Kolmogorov, 1954], which states that for a sequence of i.i.d. random variables whose distributions have a power-law tail with index $0 < \alpha < 2$, the normalized sum converges to an α-stable distribution as the number of summands go to ∞.

Domains of Normal Attraction of Stable Distributions. Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_t$ be an i.i.d. sequence of random vectors in $\mathbb{R}^n$ with a common distribution function $F(\boldsymbol{x})$. If there exists some constant $a > 0$ and a sequence $b_t \in \mathbb{R}^n$ such that

$$\frac{\boldsymbol{X}_1 + \cdots + \boldsymbol{X}_t}{at^{1/\alpha}} - b_t \xrightarrow[t\to\infty]{\mathcal{D}} \mu, \tag{2.2}$$

then $F(\boldsymbol{x})$ is said to belong to the *domain of normal attraction* of the law μ, and α is the characteristic exponent of μ [Gnedenko and Kolmogorov, 1954, page 181]. If μ is an α-stable distribution, then we say $F(\boldsymbol{x})$ belongs to the domain of normal attraction of an α-stable distribution. For example, the Pareto distribution belongs to the domain of normal attraction of an α-stable law. In Section C in the supplementary document, we provide more details as well as a sufficient and necessary conditions for being in the domain of normal attraction of an α-stable law.

3 Convergence of SGD under Heavy-tailed Gradient Noise

In this section, we identify sufficient conditions for the convergence of SGD under heavy-tailed gradient noise, and derive explicit rate estimates. In the standard setting when the noise variance is finite, it is sufficient to assume Hessian is uniformly positive definite in order to achieve contraction in the subsequent SGD iterations (see for example Polyak and Juditsky [1992], Tripuraneni et al. [2018], Su and Zhu [2018], Duchi and Ruan [2021], Toulis and Airoldi [2017], Fang et al. [2018], Anastasiou et al. [2019]). When the noise variance is infinite with a finite p-th moment for $p \in [1, 2)$, a stronger notion of positive definiteness is required in our analysis to achieve such a contraction, which leads to an interesting interpolation between the positive semi-definite cone (as $p \to 2$), and the cone of diagonally dominant matrices with non-negative diagonal entries ($p = 1$).

3.1 p-Positive Definiteness

First, we introduce the signed power of vectors which will be used to define a family of matrices.

For a vector $\boldsymbol{v} = (v^1, \ldots, v^n)^\mathsf{T} \in \mathbb{R}^n$ and $q \geqslant 0$, the signed power of $\boldsymbol{v}$ is defined as

$$\boldsymbol{v}^{\langle q\rangle} := \Big(\operatorname{sign}\big(v^1\big)\big|v^1\big|^q, \ldots, \operatorname{sign}(v^n)|v^n|^q\Big)^\mathsf{T}. \tag{3.1}$$

Denoting the n-dimensional ℓ_p unit sphere with $S_p = \{\boldsymbol{v} \in \mathbb{R}^n : \|\boldsymbol{v}\|_p = 1\}$, and the set of $n \times n$ symmetric matrices with $\mathbb{S}$, we now define the following subset of $\mathbb{S}$.

Definition 1 (p-positive definiteness)**.** *Let $p \geqslant 1$ and $\mathbf{Q}$ be a symmetric matrix. We say that $\mathbf{Q}$ is p-positive definite if for all $\boldsymbol{v} \in S_p$, we have $\boldsymbol{v}^\mathsf{T}\mathbf{Q}\boldsymbol{v}^{\langle p-1\rangle} > 0$. Similarly, we say that $\mathbf{Q}$ is p-positive semi-definite if for all $\boldsymbol{v} \in S_p$, we have $\boldsymbol{v}^\mathsf{T}\mathbf{Q}\boldsymbol{v}^{\langle p-1\rangle} \geqslant 0$.*

It is not hard to see that the set of p-positive semi-definite matrices (p-PSD) defines a closed pointed cone, which we denote by $\mathbb{S}^p_+$, with interior as the set of p-positive definite matrices (p-PD), denoted by $\mathbb{S}^p_{++}$. We are mainly interested in the case $1 \leqslant p < 2$. Note that $\mathbb{S}^2_+$ coincides with the standard PSD cone, and we show in Section A.2 that $\mathbb{S}^1_+$ is exactly the cone of diagonally dominant matrices with non-negative diagonal entries, denoted by $\mathbb{D}_+$. For any $p \in [1, 2]$, these cones satisfy the following

$$\mathbb{D}_+ = \mathbb{S}^1_+ \subseteq \mathbb{S}^p_+ \subseteq \mathbb{S}^2_+.$$

Figure 1 is a hypothetical illustration of the inclusion relationship between these cones.

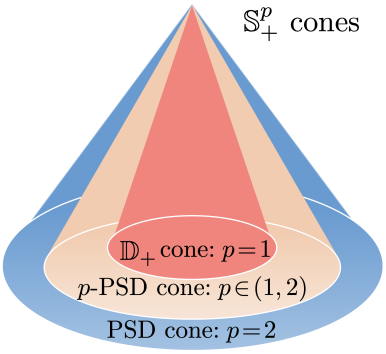

Figure 1: Geometry of p-PSD matrices. $\mathbb{D}_+$ cone refers to the cone of diagonally dominant matrices with non-negative diagonal entries. Their inclusion relationship is given in Propositions 13 and 14.

Similar to the uniform PD condition on the Hessian, which is commonly used in classical analysis (i.e. strong convexity), we also define a uniform version of Definition 1. We recall that every operator norm $\|\cdot\|_p$ induces the same topology on the set of n-dimensional matrices, which is just the usual topology on $\mathbb{R}^{n\times n}$. Further, the set of symmetric matrices $\mathbb{S}$, as the set of zeros of the continuous function $\mathbf{X} \mapsto \mathbf{X} - \mathbf{X}^\mathsf{T}$, is a closed set. Hence for a set $\mathcal{M} \subseteq \mathbb{S}$, denoting its topological closure with $\overline{\mathcal{M}}$, we also have $\overline{\mathcal{M}} \subseteq \mathbb{S}$. We are interested in the case where $\mathcal{M}$ is a bounded set.

Definition 2 (uniform p-PD)**.** *Let $p \geqslant 1$ and $\mathcal{M} \subset \mathbb{S}$ be a non-empty set of symmetric matrices. We say that $\mathcal{M}$ is uniformly p-PD if for all $\mathbf{Q} \in \overline{\mathcal{M}}$, we have $\mathbf{Q} \in \mathbb{S}^p_{++}$.*

Notice that $\mathcal{M}$ is uniformly 2-PD if and only if the eigenvalues of the symmetric matrices in the set $\mathcal{M}$ are all lower bounded by a positive real number. Notice also that a finite subset of symmetric matrices is uniformly p-PD if and only if each element of the set is p-PD.

p-PSD cone emerges naturally when analyzing SGD algorithm in the heavy-tailed setting, interpolating between the standard PSD cone to the cone of diagonally dominant matrices with non-negative diagonal entries. To the best of our knowledge, we are the first to study such families of matrices and their application in stochastic optimization. For further details about these cones, we refer interested reader to Section A.2 in the supplementary document.

We make the following uniform smoothness and curvature assumptions on the objective function.

Assumption 1. *The set of matrices $\{\boldsymbol{\nabla}^2 f(\boldsymbol{x}) : \boldsymbol{x} \in \mathbb{R}^n\}$ is bounded and uniformly p-PD.*

3.2 Rate of Convergence in L^p

We fix a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with filtration $\{\mathcal{F}_t\}_{t\in\mathbb{N}}$, and we let $\boldsymbol{x}_0$ be $\mathcal{F}_0$-measurable. We make the following assumption on the gradient noise sequence.

Assumption 2. *The gradient noise sequence $\{\boldsymbol{\xi}_t\}_{t\in\mathbb{N}^+}$ admits the following decomposition*

$$\boldsymbol{\xi}_{t+1}(\boldsymbol{x}_t) = \boldsymbol{m}_{t+1}(\boldsymbol{x}_t) + \boldsymbol{\zeta}_{t+1}, \tag{3.2}$$

where $\{\boldsymbol{\zeta}_t\}_{t\in\mathbb{N}^+}$ is an i.i.d. sequence with $\mathbb{E}[\boldsymbol{\zeta}_t] = 0$, and $\mathbb{E}[|\boldsymbol{\zeta}_t|^p] < \infty$ for some p, and $\{\boldsymbol{m}_t\}_{t\in\mathbb{N}^+}$ is a martingale difference sequence, and both sequences are adapted to the filtration $\{\mathcal{F}_t\}_{t\in\mathbb{N}}$.

Further, the state dependent component of the noise satisfies, for some $K > 0$,

$$\mathbb{E}\Big[|\boldsymbol{m}_{t+1}(\boldsymbol{x}_t)|^2 \mid \mathcal{F}_t\Big] \leqslant K\big(1 + |\boldsymbol{x}_t|^2\big). \tag{3.3}$$

We call $\boldsymbol{m}_t$ the *state-dependent component* of the gradient noise, which naturally has a state-dependent conditional second moment. The variance of this component can be arbitrarily large depending on the state $\boldsymbol{x}_t$. The *heavy-tailed* noise behavior is due to $\boldsymbol{\zeta}_t$, which may have an infinite variance for $p < 2$ (i.e., the second moment is undefined). Compared to recent works on SGD with heavy-tailed noise, our noise model in Assumption 2 is significantly more general. For instance, the noise model in the recent work Zhang et al. [2020] assumes $\mathbb{E}\big[|\boldsymbol{\xi}_{t+1}(\boldsymbol{x})|^p\big] \leq \sigma^p$ for all $\boldsymbol{x}$, where σ does not depend on $\boldsymbol{x}$. Therefore, this noise model cannot handle state-dependent noise, and does not even hold in the

linear regression example given in Section 5 (as the moments of the noise must scale with the norm of $\boldsymbol{x}$). On the contrary, in many instances of stochastic approximation methods subject to heavy-tailed noise with long-range dependencies, one can verify that the noise admits the decomposition (3.2). We shall show in Section 5 that the noise model (3.2) arises in practical applications such as linear regression and generalized linear models subject to heavy-tailed data (see also Anantharam and Borkar [2012] for a detailed discussion on this noise model).

In our first result, assuming that the objective function f has a uniformly p-PD Hessian and the noise sequence $\{\boldsymbol{\xi}_t\}_{t\in\mathbb{N}^+}$ has infinite variance but satisfies Assumption 2, we establish an asymptotic convergence rate in L^p for the SGD algorithm to the unique minimizer $\boldsymbol{x}^*$.

Theorem 3. *Suppose Assumptions 1 and 2 hold for some $1 < p \leqslant 2$. For step-size satisfying $\gamma_t \asymp t^{-\rho}$ with $\rho \in (0,1)$, the error of the SGD iterates $\{\boldsymbol{x}_t\}_{t\in\mathbb{N}}$ from the minimizer $\boldsymbol{x}^*$ satisfies*

$$\mathbb{E}[|\boldsymbol{x}_t - \boldsymbol{x}^*|^p] = \mathcal{O}\Big(t^{-\rho(p-1)}\Big). \tag{3.4}$$

Consequently, we have $\sup_{t\in\mathbb{N}^+} \mathbb{E}[|\boldsymbol{\xi}_t|^p] < \infty$.

The proof of Theorem 3 is provided in Section B in the supplementary document. We observe that the convergence rate of SGD depends on the highest order finite moment p of the noise sequence, and faster rates are achieved for larger values of p. The fastest convergence rate implied by our result is near $\mathcal{O}\big(t^{-p+1}\big)$, which is achieved for $\rho \approx 1$. However, SGD converges even for very slowly decaying step-size sequences as ρ gets closer to 0.

If the noise sequence has further integrability properties with a finite p-th moment for all $p \in [q, \alpha)$ for some $1 < q < \alpha$ and if uniform p-PD condition (i.e. Assumption 1) holds, then faster rates are achievable. In particular, the following result is an interesting consequence of Theorem 3, and its proof is provided in Section B in the supplementary document.

Corollary 4. *For constants q, α satisfying $1 < q < \alpha \leqslant 2$, suppose that Assumptions 1 and 2 hold for every $p \in [q, \alpha)$. For step-size satisfying $\gamma_t \asymp t^{-\rho}$ with $\rho \in (0,1)$, the error of the SGD iterates $\{\boldsymbol{x}_t\}_{t\in\mathbb{N}}$ from the minimizer $\boldsymbol{x}^*$ satisfies*

$$\mathbb{E}[|\boldsymbol{x}_t - \boldsymbol{x}^*|^q] = \tilde{\mathcal{O}}\Big(t^{-\rho q \frac{\alpha-1}{\alpha}}\Big). \tag{3.5}$$

Remark. The additional integrability assumption yields faster rates for any feasible step-size sequence since $p(\alpha-1)/\alpha \geqslant p-1$ for $p \in (1,2]$.

Let us briefly compare our results stated above to those in the setting where the noise sequence has a finite variance. A classical convergence result that goes back to Chung [1954, Theorem 5][2] states that

$$\mathbb{E}[|\boldsymbol{x}_t - \boldsymbol{x}^*|^r] = \Theta\Big(t^{-\rho r/2}\Big), \tag{3.6}$$

where $r \geqslant 2$ is an integer such that the r-th moment exists for the stochastic approximation process, and this is achieved for strongly convex objective functions in one dimension (whose second derivative $\{f''(\boldsymbol{x}) : \boldsymbol{x} \in \mathbb{R}\}$ satisfies the uniformly 2-PD property) with a step-size choice $\gamma_t \asymp t^{-\rho}$ for some $\rho \in (1/2, 1)$. We point out that our rate (3.5) recovers the rate implied by (3.6) when $r = 2$, and extends it further to the case $1 \leqslant r < 2$.

In the presence of heavy-tailed noise, the folklore is to modify SGD (e.g., clipped gradients) in order to tame the heavy-tails, which considerably simplifies the problem and makes it amenable to classical analysis tools. For instance, to motivate modifying SGD in this regime, in [Zhang et al., 2020, Remark 1] authors prove that $\mathbb{E}\big[|\boldsymbol{\nabla} f(\boldsymbol{x}_t)|^2\big] = \infty$ for vanilla SGD and argue that SGD diverges in this setting. On the contrary, our results show that, without any modifications, SGD can still converge to the optimum in L^p, even when it does not converge in L^2 since the second moment is not defined.

4 Stable Limits for the Polyak-Ruppert Averaging

In this section, we establish the limit distribution of the Polyak-Ruppert averaging under infinite noise variance, extending the asymptotic normality result given by Polyak and Juditsky [1992] to α-stable distributions. Let us fix an $\alpha \in (1,2]$ and assume the following throughout this subsection.

[2] This result, like many other similar studies in the 1950s, concerns only the one-dimensional case. But they generalize easily to higher dimensions.

Assumption 3. *The gradient noise sequence $\{\boldsymbol{\xi}_t\}_{t\in\mathbb{N}^+}$ admits the following decomposition*

$$\boldsymbol{\xi}_{t+1}(\boldsymbol{x}_t) = \boldsymbol{m}_{t+1}(\boldsymbol{x}_t) + \boldsymbol{\zeta}_{t+1}, \tag{4.1}$$

where $\{\boldsymbol{\zeta}_t\}_{t\in\mathbb{N}^+}$ is an i.i.d. sequence with $\mathbb{E}[\boldsymbol{\zeta}_t] = 0$, which is in the domain of normal attraction of an n-dimensional symmetric α-stable distribution μ, i.e.,

$$\frac{\boldsymbol{\zeta}_1 + \ldots + \boldsymbol{\zeta}_t}{t^{1/\alpha}} \xrightarrow[t\to\infty]{\mathcal{D}} \mu. \tag{4.2}$$

The state dependent component $\{\boldsymbol{m}_t\}_{t\in\mathbb{N}^+}$ is a martingale difference sequence with a second-moment satisfying (3.3), and both sequences are adapted to the filtration $\{\mathcal{F}_t\}_{t\in\mathbb{N}}$.

The above assumption is arguably more stringent than Assumption 2, and it implies that $\mathbb{E}[|\boldsymbol{\zeta}_t|^p] < \infty$ for all $p \in [1, \alpha)$; thus, Assumption 2 is satisfied for any $p \in [1, \alpha)$. The condition (4.2) is a special case of (2.2) which holds if, for example, $\boldsymbol{\zeta}_t$'s have power-law tails (i.e. Paretian tail) with index α.

Denoting the Polyak-Ruppert averaging by $\overline{\boldsymbol{x}}_t := \frac{1}{t}(\boldsymbol{x}_0 + ... + \boldsymbol{x}_{t-1})$, we are interested in the asymptotic behavior of

$$t^{1-1/\alpha}(\overline{\boldsymbol{x}}_t - \boldsymbol{x}^*) = \frac{(\boldsymbol{x}_0 + \ldots + \boldsymbol{x}_{t-1}) - t\boldsymbol{x}^*}{t^{1/\alpha}},$$

for $\alpha \in (1, 2]$. In the special case when $\alpha = 2$, it is known that this limit converges to a multivariate normal distribution (which is a 2-stable distribution), a result proven in the seminal work by Polyak and Juditsky [1992]. Similarly, we begin with a result that considers a quadratic objective where the function $\boldsymbol{\nabla} f(\boldsymbol{x})$ is linear in $\boldsymbol{x}$, and then building on this result, we establish the limit distribution of Polyak-Ruppert averaging also in the more general non-linear case.

Theorem 5 (linear case)**.** *Suppose the function $\boldsymbol{\nabla} f(\boldsymbol{x})$ is affine, i.e. $\boldsymbol{\nabla} f(\boldsymbol{x}) = \mathbf{A}\boldsymbol{x} - \boldsymbol{b}$ for a real matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ and a real vector $\boldsymbol{b} \in \mathbb{R}^n$ and there exist scalars p, ρ satisfying*

$$\max\left(\frac{\alpha + \alpha\rho}{1 + \alpha\rho}, \alpha\rho\right) \leqslant p \leqslant \alpha,$$

such that $\mathbf{A}$ is p-PD and $\rho \in (0, 1)$. If the noise sequence satisfies Assumption 3 for the same parameter α, then for the step-size satisfying $\gamma_t \asymp t^{-\rho}$, the normalized average $t^{1-1/\alpha}(\overline{\boldsymbol{x}}_t - \boldsymbol{x}^*)$ converges weakly to an n-dimensional α-stable distribution.

The above theorem states that Polyak-Ruppert averaging for any step-size sequence with index $\rho \in (0, 1]$ converges weakly to an α-stable limit. Thus, in the linear case, the size of this feasible interval is the same in both heavy- and light-tailed noise settings (see e.g. Polyak and Juditsky [1992] and Ruppert [1988]). Notably, α-stable limit of the averaged iterates does not depend on the index ρ, i.e., limit distribution does not depend on how fast the step-size decays as long as $\rho \in (0, 1]$.

The next result generalizes Theorem 5 to the setting where $\boldsymbol{\nabla} f(\boldsymbol{x})$ is non-linear.

Theorem 6 (non-linear case)**.** *Let $1 < 1/\rho < q < \alpha$ and suppose Assumption 1 holds for every $p \in [q, \alpha)$. Assume further that the gradient $\boldsymbol{\nabla} f(\boldsymbol{x})$ can be approximated using the Hessian matrix $\boldsymbol{\nabla}^2 f(\boldsymbol{x}^*)$ around the minimizer $\boldsymbol{x}^*$ as*

$$\left|\boldsymbol{\nabla} f(\boldsymbol{x}) - \boldsymbol{\nabla}^2 f(\boldsymbol{x}^*)(\boldsymbol{x} - \boldsymbol{x}^*)\right| \leqslant K|\boldsymbol{x} - \boldsymbol{x}^*|^q. \tag{4.3}$$

If the noise sequence satisfies Assumption 3, for the step-size satisfying $\gamma_t \asymp t^{-\rho}$, the normalized average $t^{1-1/\alpha}(\overline{\boldsymbol{x}}_t - \boldsymbol{x}^*)$ converges weakly to an n-dimensional α-stable distribution.

The condition (4.3) is standard (see e.g. Polyak and Juditsky [1992, Assumption 3.2]), which simply imposes a local linearity condition on the gradient of the objective function f, with an order-q polynomial error term. This assumption holds, for example, whenever the Hessian is Lipschitz continuous. We notice that the size of the feasible interval is $\rho \in (1/\alpha, 1)$, which is smaller this time compared to the light tailed case; Polyak and Juditsky [1992, Theorem 2] allows $\rho \in (1/2, 1)$.

The above theorem establishes that, when the noise has diverging variance, the Polyak-Ruppert averaging admits an α-stable limit rather than a standard CLT. This result has potential implications in statistical inference in the presence of heavy-tailed data. Inference procedures that take into account the computational part of the training procedure (instead of drawing conclusions for the minimizer

of the empirical risk) rely typically on variations of Polyak-Ruppert averaging and the CLT they admit [Fang et al., 2018, Su and Zhu, 2018, Chen et al., 2020]. The above theorem simply states this CLT does not hold under heavy-tailed gradient noise. Therefore, many of these procedures require further adaptation, if the gradient has undefined variance. Finally, it is well-known that Polyak-Ruppert averaging achieves the Cramér-Rao lower bound [Polyak and Juditsky, 1992, Gadat and Panloup, 2017], which is a lower bound on the variance of an unbiased estimator. However, it is not clear what this type of optimality means when the variance is not defined. These are important directions that require thorough investigations, and they will be studied elsewhere.

5 Examples in the Presence of Heavy-tailed Noise

In this section, we demonstrate how the stochastic approximation framework discussed in our paper covers several interesting examples. More specifically, we verify the assumptions required for Theorems 3, 5, and 6, for linear regression and generalized linear models (GLMs), where the heavy-tailed noise behavior may naturally arise due to heavy-tailed data.

5.1 Ordinary Least Squares

Let us first consider the following linear model,

$$y = \boldsymbol{z}^{\mathsf{T}}\boldsymbol{\beta}_0 + \epsilon,$$

where $\boldsymbol{\beta}_0 \in \mathbb{R}^n$ is the true coefficients, $y \in \mathbb{R}$ is the response, the random vector $\boldsymbol{z} \in \mathbb{R}^n$ denotes the covariates with a positive-definite covariance $0 \prec \mathbb{E}[\boldsymbol{z}\boldsymbol{z}^{\mathsf{T}}]$ and a finite fourth moment $\mathbb{E}[|\boldsymbol{z}|^4] < \infty$, and ϵ is the noise with zero conditional mean $\mathbb{E}[\epsilon|\boldsymbol{z}] = 0$. In the classical setting, the noise ϵ is assumed to be Gaussian whose variance is well-defined. In this case, the population version of the maximum likelihood estimation (MLE) problem corresponds to minimizing

$$f(\boldsymbol{x}) = \frac{1}{2}\mathbb{E}\Big[\big(y - \boldsymbol{z}^{\mathsf{T}}\boldsymbol{x}\big)^2\Big], \tag{5.1}$$

(where the expectation is over the $(y, \boldsymbol{z})$ pair), or equivalently solving the following normal equations

$$\boldsymbol{\nabla} f(\boldsymbol{x}) \coloneqq \mathbb{E}\big[\boldsymbol{z}\boldsymbol{z}^{\mathsf{T}}\big]\boldsymbol{x} - \mathbb{E}[\boldsymbol{z}y] = 0. \tag{5.2}$$

We easily observe that the true coefficients $\boldsymbol{\beta}_0$ is the unique zero of the above equation, i.e., $\boldsymbol{x}^* = \boldsymbol{\beta}_0$.

Now, suppose we are given access to a stream of i.i.d. drawn instances of the pair $(y, \boldsymbol{z})$, denoted by $\{y_t, \boldsymbol{z}_t\}_{t\in\mathbb{N}^+}$. In large-scale settings, one generally runs the following stochastic approximation process, which is simply online SGD on the population MLE objective $f(\boldsymbol{x})$:

$$\boldsymbol{x}_t = \boldsymbol{x}_{t-1} - \gamma_t\big(\boldsymbol{z}_t\boldsymbol{z}_t^{\mathsf{T}}\boldsymbol{x}_{t-1} - \boldsymbol{z}_t y_t\big). \tag{5.3}$$

Manifestly, (5.3) is a special case of (2.1), where the gradient noise admitting the decomposition $\boldsymbol{\xi}_t = \boldsymbol{\zeta}_t + \boldsymbol{m}_t$, for an i.i.d. component $\boldsymbol{\zeta}_t$ and a state-dependent component $\boldsymbol{m}_t$ (cf. (4.1)),

$$\begin{cases} \boldsymbol{\zeta}_t \coloneqq \mathbb{E}[\boldsymbol{z}y] - \boldsymbol{z}_t y_t, \\ \boldsymbol{m}_t \coloneqq \big(\boldsymbol{z}_t\boldsymbol{z}_t^{\mathsf{T}} - \mathbb{E}\big[\boldsymbol{z}\boldsymbol{z}^{\mathsf{T}}\big]\big)\boldsymbol{x}_{t-1}. \end{cases} \tag{5.4}$$

In the presence of heavy-tailed noise, for example when the noise ϵ has infinite variance, the population MLE objective $f(\boldsymbol{x})$ may not be even finite, and one should resort to methods from M-estimation and choose a suitable loss function within robust statistics framework [Huber, 2004, Van der Vaart, 2000]. However, the iterations (5.3) may still be employed to estimate the true coefficients $\boldsymbol{\beta}_0$ (potentially due to model misspecification), as we demonstrate below. Note that in this case, iterations (5.3) should be seen as solving the root-finding problem (5.2) via stochastic approximation (2.1), rather than a minimization problem stated in (5.1).

First, we verify Assumption 2. We observe from the decomposition given in (5.4) that the i.i.d. component $\{\boldsymbol{\zeta}_t\}_{t\in\mathbb{N}}$ exhibits the heavy-tailed behavior since it contains $y_t = \boldsymbol{z}_t^{\mathsf{T}}\boldsymbol{\beta}_0 + \epsilon_t$. Assume that this component has the highest order finite moment p satisfying $1 \leqslant p < 2$, i.e., $\mathbb{E}[|\boldsymbol{\zeta}_t|^p] < \infty$. Further, the state dependent component $\boldsymbol{m}_t$ defines a martingale difference sequence, and the condition (3.3) is met since the covariates $\boldsymbol{z}$ have finite fourth moment, i.e.,

$$\mathbb{E}\big[|\boldsymbol{m}_t|^2 \mid \boldsymbol{x}_{t-1}\big] \leqslant C|\boldsymbol{x}_{t-1}|^2.$$

Hence, Assumption 2 is satisfied. Next, assuming that the second moment of the covariates $\nabla^2 f(\boldsymbol{x}) = \mathbb{E}[\boldsymbol{z}\boldsymbol{z}^\mathsf{T}]$ is p-PD, one can guarantee that Assumption 1 is satisfied. Therefore, the convergence results of our theorems hold. Finally, we note that p-PD assumption is always satisfied if $\mathbb{E}[\boldsymbol{z}\boldsymbol{z}^\mathsf{T}]$ is diagonally dominant, but the condition for $p > 1$ is weaker.

5.2 Generalized Linear Models

Generalized linear models (GLMs) play a crucial role in numerous problems in statistics, and provide a miscellaneous framework for many regression and classification tasks, with many applications [McCullagh and Nelder, 1989, Nelder and Wedderburn, 1972]. In this section, we consider minimizing the objective function arising from GLMs, for which there are many methods available (see e.g. Erdogdu [2015, 2016] and the references therein). However, we restrict ourselves to the misspecified and online setting. That is, the minimization problem corresponds to a GLM, but the model is misspecified so that the response can be heavy-tailed.

For a response $y \in \mathbb{R}$ and random covariates $\boldsymbol{z} \in \mathbb{R}^n$, the population version of an ℓ_2-regularized MLE problem in the canonical GLM framework reads

$$\underset{\boldsymbol{x}}{\text{minimize}}\ f(\boldsymbol{x}) := \mathbb{E}\Big[\psi\big(\boldsymbol{x}^\mathsf{T}\boldsymbol{z}\big) - y\boldsymbol{x}^\mathsf{T}\boldsymbol{z}\Big] + \frac{\lambda}{2}|\boldsymbol{x}|^2 \qquad \text{for} \qquad \lambda > 0. \tag{5.5}$$

Here, $\psi : \mathbb{R} \to \mathbb{R}$ is referred to as the cumulant generating function (CGF), and it is convex. Notable examples include $\psi(x) = x^2/2$ yielding linear regression, $\psi(x) = \log(1 + e^x)$ yielding logistic regression, and $\psi(x) = e^x$ yielding Poisson regression. Gradient of the objective (5.5) is given by

$$\nabla f(\boldsymbol{x}) = \mathbb{E}\Big[\boldsymbol{z}\psi'\big(\boldsymbol{z}^\mathsf{T}\boldsymbol{x}\big)\Big] - \mathbb{E}[\boldsymbol{z}y] + \lambda\boldsymbol{x}. \tag{5.6}$$

We define the unique solution of the population GLM problem as the unique zero of (5.6), which we denote by $\boldsymbol{x}^*$. To reiterate, we do not assume a model on data, allowing for model misspecification; we simply consider the resulting optimization problem similar to Erdogdu et al. [2016, 2019]. As in the previous section, we assume that the covariates have finite fourth moment and the response y_t is contaminated with heavy-tailed noise with infinite variance. In this setting, the objective function is always well-defined, even if the response has infinite variance.

We are given access to a stream of i.i.d. drawn instances of the pair $(y, \boldsymbol{z})$, denoted by $\{y_t, \boldsymbol{z}_t\}_{t\in\mathbb{N}^+}$, and we solve the above non-linear problem using the following stochastic process,

$$\boldsymbol{x}_t = \boldsymbol{x}_{t-1} - \gamma_t\big(\boldsymbol{z}_t\psi'\big(\boldsymbol{z}_t^\mathsf{T}\boldsymbol{x}_{t-1}\big) - \boldsymbol{z}_t y_t + \lambda\boldsymbol{x}_{t-1}\big),$$

with gradient noise admitting the decomposition $\boldsymbol{\xi}_t = \boldsymbol{\zeta}_t + \boldsymbol{m}_t$ where

$$\begin{cases}\boldsymbol{\zeta}_t := \mathbb{E}[\boldsymbol{z}y] - \boldsymbol{z}_t y_t, \\ \boldsymbol{m}_t := \boldsymbol{z}_t\psi'\big(\boldsymbol{z}_t^\mathsf{T}\boldsymbol{x}_{t-1}\big) - \mathbb{E}\big[\boldsymbol{z}_t\psi'\big(\boldsymbol{z}_t^\mathsf{T}\boldsymbol{x}_{t-1}\big)\big].\end{cases}$$

In what follows, we verify our assumptions for a CGF satisfying $|\psi'(x)| \leqslant C(1 + |x|)$, $\psi''(x) \geqslant 0$, and $|\psi'''(x)| \leq L$ for all $x \in \mathbb{R}$. These assumptions can be easily verified for any second-order smooth CGF that grows at most quadratically (e.g. if the misspecified model is binomial with k number of trials, we have $\psi(x) = k\log(1 + e^x)$). $\boldsymbol{\zeta}_t$'s are i.i.d. and contain the entire heavy-tailed part of the gradient noise. Assume that this component has the highest defined moment order $1 \leqslant p < 2$, i.e., $\mathbb{E}[|\boldsymbol{\zeta}_t|^p] < \infty$. Further observe that the state dependent component defines a martingale difference sequence and satisfies the condition (3.3) since the covariates $\boldsymbol{z}$ have finite fourth moment, and $|\psi'|$ grows at most linearly. Therefore, Assumption 2 is satisfied.

We note that the Hessian of the objective f is given as

$$\nabla^2 f(\boldsymbol{x}) = \mathbb{E}\big[\boldsymbol{z}\boldsymbol{z}^\mathsf{T}\psi''\big(\boldsymbol{z}^\mathsf{T}\boldsymbol{x}\big)\big] + \lambda\mathbf{I}.$$

Since $\psi''(x) \geqslant 0$, $\nabla^2 f(\boldsymbol{x})$ is clearly PD for all $\lambda > 0$. For sufficiently large λ, this matrix can also be made diagonally dominant, which implies the p-PD condition for any $p \geqslant 1$, further implying Assumption 1. We further note that if $\nabla^2 f(\boldsymbol{x})$ is Lipshitz (e.g. for the binomail CGF), then (4.3) holds for $q = 2$ globally; thus it holds for any $q < 2$ locally. Therefore, for an appropriate step-size sequence, our convergence results on the SGD can be applied to this framework.

6 Conclusion

In this paper, we considered SGD subject to state-dependent and heavy-tailed noise with potentially infinite variance, when the objective belongs to a class of strongly convex functions (termed as p-PD condition). We provided asymptotic L^p convergence rates for vanilla SGD, demonstrating that SGD without any modifications can be still used in the presence of heavy-tailed noise. Furthermore, we provided a generalized central limit theorem for the Polyak-Ruppert averaging, i.e., we proved that the averaged iterates converge to a multivariate α-stable distribution.

We emphasize that p-PD condition is a sufficient condition, and further investigation is needed to see if this condition can be replaced with the standard strong convexity assumption. We also highlight that non-asymptotic L^p rates in the current setting should be achievable, which will be studied elsewhere. Finally, while we leave it for a future study, we emphasized the importance of adapting existing statistical inference techniques that rely on the averaged SGD iterates when the gradient noise is heavy-tailed, which arises naturally in modern statistical learning applications.

Acknowledgements

MAE is partially funded by CIFAR AI Chairs program, and CIFAR AI Catalyst grant, NSERC Grant [2019-06167]. MG's research is supported in part by the grants Office of Naval Research Award Number N00014-21-1-2244, National Science Foundation (NSF) CCF-1814888, NSF DMS-2053485, NSF DMS-1723085. LZ is grateful to the partial support from NSF DMS-2053454 and a Simons Foundation Collaboration Grant. UŞ's research is supported by the French government under management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute).

References

V. Anantharam and V. S. Borkar. Stochastic approximation with long range dependent and heavy tailed noise. *Queueing Systems*, 71(1-2):221–242, 2012.

A. Anastasiou, K. Balasubramanian, and M. A. Erdogdu. Normal approximation for stochastic gradient descent via non-asymptotic rates of martingale CLT. In *Conference on Learning Theory*, pages 115–137, 2019.

J. R. Blum. Approximation methods which converge with probability one. *The Annals of Mathematical Statistics*, 25(2):382–386, 1954.

D. Buraczewski, E. Damek, and M. Mirek. Asymptotics of stationary solutions of multivariate stochastic recursions with heavy tailed inputs and related limit theorems. *Stochastic Processes and their Applications*, 122(1):42–67, 2012.

D. Buraczewski, E. Damek, and T. Mikosch. *Stochastic Models with Power-Law Tails*. Springer, 2016.

X. Chen, J. D. Lee, X. T. Tong, and Y. Zhang. Statistical inference for model parameters in stochastic gradient descent. *The Annals of Statistics*, 48(1):251–273, 2020.

Y. Cherapanamjeri, N. Tripuraneni, P. L. Bartlett, and M. I. Jordan. Optimal mean estimation without a variance. *arXiv preprint arXiv:2011.12433*, 2020.

K. L. Chung. On a stochastic approximation method. *The Annals of Mathematical Statistics*, 25(3): 463–483, 1954.

D. Davis, D. Drusvyatskiy, L. Xiao, and J. Zhang. From low probability to high confidence in stochastic convex optimization. *arXiv preprint arXiv:1907.13307*, 2019.

J. Duchi and F. Ruan. Asymptotic optimality in stochastic optimization. *The Annals of Statistics*, 49 (1):21–48, 2021.

M. A. Erdogdu. Newton-stein method: A second order method for glms via stein's lemma. *Advances in Neural Information Processing Systems*, 28:1216–1224, 2015.

M. A. Erdogdu. Newton-stein method: An optimization method for glms via stein's lemma. *The Journal of Machine Learning Research*, 17(1):7565–7616, 2016.

M. A. Erdogdu, L. H. Dicker, and M. Bayati. Scaled least squares estimator for glms in large-scale problems. *Advances in Neural Information Processing Systems*, 29:3324–3332, 2016.

M. A. Erdogdu, M. Bayati, and L. H. Dicker. Scalable approximations for generalized linear problems. *The Journal of Machine Learning Research*, 20(1):231–275, 2019.

V. Fabian. Stochastic approximation of minima with improved asymptotic speed. *The Annals of Mathematical Statistics*, 38(1):191–200, 1967.

V. Fabian. On asymptotic normality in stochastic approximation. *The Annals of Mathematical Statistics*, 39(4):1327–1332, 1968.

Y. Fang, J. Xu, and L. Yang. Online bootstrap confidence intervals for the stochastic gradient descent estimator. *The Journal of Machine Learning Research*, 19(1):3053–3073, 2018.

N. Farsad, W. Guo, C. B. Chae, and A. Eckford. Stable distributions as noise models for molecular communication. *2015 IEEE Global Communications Conference, GLOBECOM 2015*, 2015.

E. Feldheim. *Étude de la stabilité des lois de probabilité*. PhD thesis, Faculté des Sciences de Paris, Paris, 1937.

W. Feller. *An Introduction to Probability Theory and Its Applications*. Wiley, New York, 2nd edition, 1971.

A. Fiche, J. C. Cexus, A. Martin, and A. Khenchaf. Features modeling with an α-stable distribution: Application to pattern recognition based on continuous belief functions. *Information Fusion*, 14 (4):504–520, 2013.

S. Gadat and F. Panloup. Optimal non-asymptotic bound of the Ruppert-Polyak averaging without strong convexity. *arXiv preprint arXiv:1709.03342*, 2017.

J. L. Geluk and L. de Hann. Stable probability distributions and their domains of attraction: A direct approach. *Probability and Mathematical Statistics*, 20:169–188, 2000.

B. V. Gnedenko and A. Kolmogorov. *Limit Distributions for Sums of Independent Random Variables*. Addison-Wesley, Cambridge, MA, 1954. Translated by Kai Lai Chung.

C. Goodsell and D. Hanson. Almost sure convergence for the Robbins-Monro process. *The Annals of Probability*, 4(6):890–901, 1976.

E. Gorbunov, M. Danilova, and A. Gasnikov. Stochastic optimization with heavy-tailed noise via accelerated gradient clipping. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

M. Gürbüzbalaban and Y. Hu. Fractional moment-preserving initialization schemes for training fully-connected neural networks. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130, pages 2233–2241. PMLR, 2021.

M. Gürbüzbalaban, U. Şimşekli, and L. Zhu. The heavy-tail phenomenon in SGD. In *International Conference on Machine Learning*, pages 3964–3975, 2021.

L. Hodgkinson and M. W. Mahoney. Multiplicative noise and heavy tails in stochastic optimization. In *International Conference on Machine Learning*, pages 4262–4274, 2021.

P. J. Huber. *Robust Statistics*, volume 523. John Wiley & Sons, 2004.

T. P. Krasulina. On stochastic approximation processes with infinite variance. *Theory of Probability & Its Applications*, 14(3):522–526, 1969.

H. Kushner and G. G. Yin. *Stochastic Approximation and Recursive Algorithms and Applications*, volume 35. Springer Science & Business Media, 2003.

P. Lévy. Théorie de l'addition des variables aléatoires. *Gauthiers-Villars, Paris*, 1937.

G. Li. Almost sure convergence of stochastic approximation procedures. *Statistica Sinica*, 4(1): 361–372, 1994.

P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Chapman and Hall, 2nd edition, 1989.

M. Mirek. Heavy tail phenomenon and convergence to stable laws for iterated Lipschitz maps. *Probability Theory and Related Fields*, 151(3):705–734, 2011.

A. V. Nazin, A. S. Nemirovsky, A. B. Tsybakov, and A. B. Juditsky. Algorithms of robust stochastic optimization based on mirror descent method. *Automation and Remote Control*, 80(9):1607–1627, 2019.

J. A. Nelder and R. W. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3):370–384, 1972.

J. Neveu. *Discrete-Parameter Martingales*, volume 10. North-Holland Amsterdam, 1975.

B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.

H. Robbins and S. Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951.

D. Ruppert. Efficient estimations from a slowly convergent Robbins-Monro process. Technical report, Cornell University Operations Research and Industrial Engineering, 1988.

J. Sacks. Asymptotic distribution of stochastic approximation procedures. *The Annals of Mathematical Statistics*, 29(2):373–405, 1958.

G. Samorodnitsky and M. S. Taqqu. *Stable Non-Gaussian Random Processes: Stochastic Models with Infinite Variance*. Chapman & Hall, New York, 1994.

K. Sarafrazi and M. Yazdi. Skewed alpha-stable distribution for natural texture modeling and segmentation in contourlet domain. *Eurasip Journal on Image and Video Processing*, 2019(1): 1–12, 2019.

A. Shapiro. Asymptotic properties of statistical estimators in stochastic programming. *The Annals of Statistics*, 17(2):841–858, 1989.

U. Şimşekli, 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.

U. Şimşekli, L. Sagun, and M. Gürbüzbalaban. A tail-index analysis of stochastic gradient noise in deep neural networks. In *International Conference on Machine Learning*, pages 5827–5837, 2019.

W. Su and Y. Zhu. Statistical inference for online learning and stochastic approximation via hierarchical incremental gradient descent. *arXiv preprint arXiv:1802.04876*, 2018.

P. Toulis and E. M. Airoldi. Asymptotic and finite-sample properties of estimators based on stochastic gradients. *The Annals of Statistics*, 45(4):1694–1727, 2017.

N. Tripuraneni, N. Flammarion, F. Bach, and M. I. Jordan. Averaging stochastic gradient descent on Riemannian manifolds. In *Proceedings of the 31st Conference on Learning Theory*, 2018.

A. W. Van der Vaart. *Asymptotic Statistics*, volume 3. Cambridge University Press, 2000.

L. Yu, K. Balasubramanian, S. Volgushev, and M. A. Erdogdu. An analysis of constant step size SGD in the non-convex regime: Asymptotic normality and bias. *arXiv preprint arXiv:2006.07904*, 2020.

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? In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Convergence Rates of Stochastic Gradient Descent under Infinite Noise Variance

SUPPLEMENTARY DOCUMENT

A Lemmas and Discussions

A.1 Key Lemmas

In this subsection, we present some key lemmas used in the proof of our main theorems, which are helpful when considering stochastic problems with *infinite* variance.

The concept of *uncorrelatedness* has long been used by probabilists as a trick when computing and estimating variance. For example, consider a sequence of uncorrelated random vectors $\{\boldsymbol{X}_t\}_{t\in\mathbb{N}^+}$ (e.g. square-integrable martingale difference). Then

$$\mathbb{E}\big[|\boldsymbol{X}_1+\ldots+\boldsymbol{X}_t|^2\big]=\mathbb{E}\big[|\boldsymbol{X}_1|^2\big]+\ldots+\mathbb{E}\big[|\boldsymbol{X}_t|^2\big]. \tag{A.1}$$

Indeed, this type of expansion is used in Polyak and Juditsky [1992] to show L^2 convergence in the normality analysis of stochastic approximation problems.

However, correlatedness is *only* defined when random elements have *finite* variance. The following lemma provides an infinite-variance version of expansion (A.1), stating that the p-th moment ($p<2$) of a martingale without square-integrability assumption can also be bounded *simpliciter* by the sum of the p-th moments of its differences, at the cost of a multiplicative constant that may depend only on p and the dimension n. It is a generalization of the recent study Cherapanamjeri et al. [2020, Lemma 4.2].

Lemma 7. *Suppose $p\in[0,1]$ and let $\{\boldsymbol{S}_t\}_{t\in\mathbb{N}}$ be an n-dimensional martingale adapted to the filtration $\{\mathcal{F}_t\}_{t\in\mathbb{N}}$, with $\mathbb{E}[|\boldsymbol{S}_t|^{1+p}]<\infty$ for every t and $\boldsymbol{S}_0=0$. Let $\boldsymbol{X}_i=\boldsymbol{S}_i-\boldsymbol{S}_{i-1}$. Then*

$$\mathbb{E}\Big[|\boldsymbol{S}_t|^{1+p}\Big]\leqslant 2^{1-p}n^{1-\frac{1+p}{2}}\sum_{i=1}^{t}\mathbb{E}\Big[|\boldsymbol{X}_i|^{1+p}\Big].$$

Next, we present a Taylor-expansion-type inequality for the function $\|\boldsymbol{x}\|_p^p$. Recall that we have defined the signed power of a vector in (3.1).

Lemma 8. *Let $p\in[1,2]$. For any $\boldsymbol{x},\boldsymbol{y}\in\mathbb{R}^n$, $\|\boldsymbol{x}+\boldsymbol{y}\|_p^p\leqslant\|\boldsymbol{x}\|_p^p+4\|\boldsymbol{y}\|_p^p+p\boldsymbol{y}^\top\boldsymbol{x}^{\langle p-1\rangle}$.*

This inequality traces back to Krasulina [1969], where the one-dimensional version $|x+y|^p\leqslant|x|^p+C|y|^p+pyx^{p-1}\operatorname{sign}(x)$ is used[3] to derive an L^p rate of convergence for the one-dimensional stochastic approximation process with step-size $1/t$. In our current study, this lemma is used not only to derive L^p rate of convergence for general infinite-variance process in $\mathbb{R}^n$ with variable step-size scheme (Theorem 3), but also in the proof of the equivalent definitions of p-PD (Theorem 10).

Finally, we quote Fabian [1967, Lemma 4.2], which we shall use to calculate the exact convergence rate (see also Chung [1954]).

Lemma 9 (Fabian [1967], Lemma 4.2)**.** *Let $\{b_t\}_{t\in\mathbb{N}}, A, B, \alpha, \beta$ be real numbers such that $0<\alpha<1$, $A>0$ and suppose the recursion*

$$b_{t+1}=b_t(1-At^{-\alpha})+Bt^{-\alpha-\beta}$$

holds. Then, $b_t=\Theta(t^{-\beta})$.

A.2 Discussions on p-Positive Definiteness and Uniform p-Positive Definiteness

Let us now focus on p-PD and uniform p-PD conditions which are defined in Definition 1, Definition 2 (also see Assumption 1). The next theorem provides several equivalent characterizations of p-PD condition, which will be used in the proof of L^p convergence.

[3] The paper Krasulina [1969] contains a minor error in ignoring the signum function $\operatorname{sign}(x)$ in this inequality. Our proof of Theorem 3 can be thought of its correction as well as extension.

Theorem 10 (Equivalent definitions of p-PD)**.** *Let $\mathbf{Q}$ be a symmetric matrix. The following are equivalent when $p \in [1, 2]$.*

i) There exist $\delta, L > 0$, such that $\|\mathbf{I} - t\mathbf{Q}\|_p^p \leqslant 1 - Lt$ for all $t \in [0, \delta)$.

ii) There exists $\lambda > 0$ such that for all $\boldsymbol{v} \in \mathbb{R}^n$, $\boldsymbol{v}^\mathsf{T}\mathbf{Q}\boldsymbol{v}^{\langle p-1\rangle} \geqslant \lambda\|\boldsymbol{v}\|_p^p$.

iii) For all $\boldsymbol{v} \in S_p$, $\boldsymbol{v}^\mathsf{T}\mathbf{Q}\boldsymbol{v}^{\langle p-1\rangle} > 0$.

iv) For all $\boldsymbol{v} \in S_p$, there exists $t_0 > 0$ such that $\|\boldsymbol{v} - t_0\mathbf{Q}\boldsymbol{v}\|_p < 1$.

Next, we provide several equivalent characterizations of uniform p-PD.

Theorem 11 (Equivalent definitions of uniform p-PD)**.** *Let $\mathcal{M}$ be a bounded set of symmetric matrices. The following are equivalent when $p \in [1, 2]$.*

i) There exist $\delta, L > 0$, such that $\|\mathbf{I} - t\mathbf{Q}\|_p^p \leqslant 1 - Lt$ for all $t \in [0, \delta)$ and $\mathbf{Q} \in \mathcal{M}$.

ii) There exists $\lambda > 0$ such that for all $\boldsymbol{v} \in \mathbb{R}^n$ and $\mathbf{Q} \in \mathcal{M}$, $\boldsymbol{v}^\mathsf{T}\mathbf{Q}\boldsymbol{v}^{\langle p-1\rangle} \geqslant \lambda\|\boldsymbol{v}\|_p^p$.

iii) For all $\boldsymbol{v} \in S_p$ and $\mathbf{Q} \in \overline{\mathcal{M}}$, $\boldsymbol{v}^\mathsf{T}\mathbf{Q}\boldsymbol{v}^{\langle p-1\rangle} > 0$.

iv) For all $\boldsymbol{v} \in S_p$ and $\mathbf{Q} \in \overline{\mathcal{M}}$, there exists $t_0 > 0$ such that $\|\boldsymbol{v} - t_0\mathbf{Q}\boldsymbol{v}\|_p < 1$.

We notice that some mild assumptions can indeed imply p-PD. For example, we will show that diagonal dominance implies p-PD. Recall that a symmetric matrix $\mathbf{Q} = (q_{ij})_{n\times n}$ is called diagonally dominant (with non-negative diagonal) if for every $i \in [n]$,

$$q_{ii} - \sum_{j\in[n]\setminus\{i\}} |q_{ij}| > 0.$$

Further, we say that a non-empty set $\mathcal{M}$ of symmetric matrices is *uniformly diagonally dominant* (with non-negative diagonal) if

$$\inf_{(q_{ij})_{n\times n}\in\mathcal{M}} \min_{i\in[n]} \left(q_{ii} - \sum_{j\in[n]\setminus\{i\}} |q_{ij}| \right) > 0.$$

We have the following observations which we shall prove in Section B. First, we observe that the uniform p-PD assumption is weaker than the notion of uniform diagonally dominance (with non-negative diagonal).

Proposition 12. *A uniformly diagonally dominant (with non-negative diagonal) set of symmetric matrices is uniformly p-PD for every $p \in [1, 2]$.*

Next, we notice that the result in Proposition 12 is tight for $p = 1$.

Proposition 13. *Uniform 1-PD is equivalent to uniform diagonal dominance (with non-negative diagonal).*

Finally, we observe that the notion of uniform 2-PD is weaker than uniform p-PD for any $p \in [1, 2]$.

Proposition 14. *Let $p \in [1, 2]$. Uniform p-PD implies uniform 2-PD.*

B Omitted Proofs

In this section, we first prove the lemmas, theorems, and propositions in Section A, then prove the theorems in Sections 3 and 4. Throughout this section, we denote by $\boldsymbol{\delta}_t$ the error of the approximation $\boldsymbol{x}_t - \boldsymbol{x}^*$, and by $\overline{\boldsymbol{\delta}}_t$ the averaged error $(\boldsymbol{\delta}_0 + \ldots + \boldsymbol{\delta}_{t-1})/t$. The gradient $\boldsymbol{\nabla} f(\boldsymbol{x})$ and the Hessian $\boldsymbol{\nabla}^2 f(\boldsymbol{x})$ will be written as $\boldsymbol{R}(\boldsymbol{x})$ and $\boldsymbol{\nabla}\boldsymbol{R}(\boldsymbol{x})$ respectively, not only for notational simplicity, but also to stress the fact that our results can be applied to any instance of stochastic approximation (2.1) including SGD.

Proof of Lemma 7 We first prove the $n = 1$ case. Suppose $\{S_t\}$ is a one-dimensional martingale and $X_i = S_i - S_{i-1}$. Notice that the function $g(x) = |x|^{1+p}$ satisfies the inequality (see e.g. Cherapanamjeri et al. [2020, Lemma A.3]):

$$|g'(x) - g'(y)| \leqslant 2^{1-p} g'(|x-y|),$$

where the weak derivative $g'(x) = \operatorname{sign}(x)$ is used in the inequality above in the case of $p = 0$, where

$$\operatorname{sign}(x) := \begin{cases} 1 & \text{if} \quad x > 0, \\ -1 & \text{if} \quad x < 0, \\ 0 & \text{if} \quad x = 0. \end{cases}$$

Furthermore, by $\mathbb{E}[X_i g'(S_{i-1}) \mid \mathcal{F}_{i-1}] = g'(S_{i-1})\mathbb{E}[X_i \mid \mathcal{F}_{i-1}] = 0$, we have

$$\begin{aligned}
\mathbb{E}[g(S_t)] &= \sum_{i=1}^{t} \mathbb{E}\left[\int_{S_{i-1}}^{S_i} g'(x)\mathrm{d}x\right] \\
&= \sum_{i=1}^{t} \mathbb{E}\left[X_i g'(S_{i-1}) + \int_{S_{i-1}}^{S_i} [g'(x) - g'(S_{i-1})]\mathrm{d}x\right] \\
&= \sum_{i=1}^{t} \mathbb{E}\left[\int_{S_{i-1}}^{S_i} [g'(x) - g'(S_{i-1})]\mathrm{d}x\right] \\
&= \sum_{i=1}^{t} \mathbb{E}\left[\int_{0}^{X_i} [g'(S_{i-1} + \tau) - g'(S_{i-1})]\mathrm{d}\tau\right] \\
&= \sum_{i=1}^{t} \mathbb{E}\left[\int_{0}^{|X_i|} |g'(S_{i-1} + \operatorname{sign}(X_i)\tau) - g'(S_{i-1})|\mathrm{d}\tau\right] \\
&\leqslant 2^{1-p} \sum_{i=1}^{t} \mathbb{E}\left[\int_{0}^{|X_i|} g'(\tau)\mathrm{d}\tau\right] \\
&= 2^{1-p} \sum_{i=1}^{t} \mathbb{E}[g(|X_i|)]. \qquad \text{(B.1)}
\end{aligned}$$

Next, for the higher dimension $n > 1$, we denote by S_i^j (resp. X_i^j) the j-th entry of the vector $\boldsymbol{S}_i$ (resp. $\boldsymbol{X}_i$). We can apply the inequality (B.1) obtained above to S_t^j by taking a $(1+p)$-norm,

$$\begin{aligned}
\mathbb{E}\Big[\|\boldsymbol{S}_t\|_{1+p}^{1+p}\Big] &= \sum_{j=1}^{n} \mathbb{E}\left[\left|S_t^j\right|^{1+p}\right] \\
&\leqslant \sum_{j=1}^{n} 2^{1-p} \sum_{i=1}^{t} \mathbb{E}\left[\left|X_i^j\right|^{1+p}\right] \\
&= 2^{1-p} \sum_{i=1}^{t} \sum_{j=1}^{n} \mathbb{E}\left[\left|X_i^j\right|^{1+p}\right] \\
&= 2^{1-p} \sum_{i=1}^{t} \mathbb{E}\Big[\|\boldsymbol{X}_i\|_{1+p}^{1+p}\Big].
\end{aligned}$$

Finally, the inequalities

$$|\boldsymbol{x}| \leqslant \|\boldsymbol{x}\|_{1+p} \leqslant n^{\frac{1}{1+p} - \frac{1}{2}} |\boldsymbol{x}|$$

give our desired result:

$$\mathbb{E}\big[|\boldsymbol{S}_t|^{1+p}\big] \leqslant 2^{1-p} n^{1 - \frac{1+p}{2}} \sum_{i=1}^{t} \mathbb{E}\big[|\boldsymbol{X}_i|^{1+p}\big].$$

The proof is complete. □

Proof of Lemma 8 By the inequality that $|1+a|^p \leqslant 1 + ap + 4|a|^p$ for any $p \in [1,2]$ and $a \in \mathbb{R}$, we have that for any $p \in [1,2]$ and $x, y \in \mathbb{R}$,

$$|x+y|^p \leqslant |x|^p + py|x|^{p-1}\operatorname{sign}(x) + 4|y|^p. \tag{B.2}$$

Next, for any $\boldsymbol{x} = (x^1, \ldots, x^n)^\mathsf{T}, \boldsymbol{y} = (y^1, \ldots, y^n)^\mathsf{T} \in \mathbb{R}^n$, by taking the p-norm and applying the inequality (B.2), we obtain

$$\begin{aligned}
\|\boldsymbol{x}+\boldsymbol{y}\|_p^p &= \sum_{i=1}^n \left|x^i + y^i\right|^p \\
&\leqslant \sum_{i=1}^n \left(\left|x^i\right|^p + py^i\left|x^i\right|^{p-1}\operatorname{sign}(x^i) + 4\left|y^i\right|^p\right) \\
&= \|\boldsymbol{x}\|_p^p + 4\|\boldsymbol{y}\|_p^p + p\sum_{i=1}^n y^i\left|x^i\right|^{p-1}\operatorname{sign}(x^i) \\
&= \|\boldsymbol{x}\|_p^p + 4\|\boldsymbol{y}\|_p^p + p\boldsymbol{y}^\mathsf{T}\boldsymbol{x}^{\langle p-1\rangle},
\end{aligned}$$

which completes the proof. □

Since Theorem 10 is just a special case of Theorem 11, we will only prove the latter. Before we proceed, let us first state a useful technical lemma.

Lemma 15. *Let $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^n$ and consider the function $\varphi(t) = \|\boldsymbol{u} + t\boldsymbol{v}\|_p^p = \sum_{i=1}^n |u^i + v^i t|^p$. The function φ is convex and has the following derivative (when $1 < p \leqslant 2$) or subderivative (when $p = 1$):*

$$\varphi'(t) = \sum_{i=1}^n p\left|u^i + v^i t\right|^{p-1}\operatorname{sign}\left(u^i + v^i t\right)v^i = p\boldsymbol{v}^\mathsf{T}(\boldsymbol{u} + t\boldsymbol{v})^{\langle p-1\rangle}.$$

The proof of Lemma 15 is straightforward and is hence omitted here.

Now we are ready to prove Theorem 11.

Proof of Theorem 11 We shall show that i) $\Longrightarrow$ iv) $\Longrightarrow$ iii) $\Longrightarrow$ ii) $\Longrightarrow$ i).

i) $\Longrightarrow$ iv) Take a sequence $\{\mathbf{Q}_1, \mathbf{Q}_2, \ldots\} \subseteq \mathcal{M}$ such that $\lim_{m\to\infty} \mathbf{Q}_m = \mathbf{Q}$. iv) follows from $\|\mathbf{I} - (\delta/2)\mathbf{Q}_m\|_p^p \leqslant 1 - L\delta/2$.

iv) $\Longrightarrow$ iii) For all $\boldsymbol{v} \in S_p$ and $\mathbf{Q} \in \overline{\mathcal{M}}$, consider the function $\varphi(t) = \|\boldsymbol{v} - t\mathbf{Q}\boldsymbol{v}\|_p^p$. According to Lemma 15, $\varphi(t)$ is convex. Furthermore, $\varphi(t_0) < 1 = \varphi(0)$. Hence it follows that $\varphi'(0) < 0$; that is, $\boldsymbol{v}^\mathsf{T}\mathbf{Q}\boldsymbol{v}^{\langle p-1\rangle} > 0$.

iii) $\Longrightarrow$ ii) Since the function $(\boldsymbol{v}, \mathbf{Q}) \mapsto \boldsymbol{v}^\mathsf{T}\mathbf{Q}\boldsymbol{v}^{\langle p-1\rangle}$ is continuous, it maps the compact set $S_p \times \overline{\mathcal{M}}$ to a compact set. Hence there exists some $\lambda > 0$ such that for all $\boldsymbol{v} \in S_p$ and $\mathbf{Q} \in \overline{\mathcal{M}}$, $\boldsymbol{v}^\mathsf{T}\mathbf{Q}\boldsymbol{v}^{\langle p-1\rangle} \geqslant \lambda$. Now, for every $\boldsymbol{u} \in \mathbb{R}^n \setminus \{0\}$, by setting $\boldsymbol{v} = \boldsymbol{u}/\|\boldsymbol{u}\|_p$, we get $\boldsymbol{u}^\mathsf{T}\mathbf{Q}\boldsymbol{u}^{\langle p-1\rangle} \geqslant \lambda\|\boldsymbol{u}\|_p^p$.

ii) $\Longrightarrow$ i) For arbitrary $\boldsymbol{v} \in \mathbb{R}^n$ and $\mathbf{Q} \in \mathcal{M}$, by Lemma 8 we have $\|(\mathbf{I} - t\mathbf{Q})\boldsymbol{v}\|_p^p = \|\boldsymbol{v} - t\mathbf{Q}\boldsymbol{v}\|_p^p \leqslant \|\boldsymbol{v}\|_p^p + 4t^p\|\mathbf{Q}\boldsymbol{v}\|_p^p - pt(\boldsymbol{v}^\mathsf{T}\mathbf{Q}\boldsymbol{v}^{\langle p-1\rangle}) \leqslant \|\boldsymbol{v}\|_p^p + 4t^p\|\mathbf{Q}\|_p^p\|\boldsymbol{v}\|_p^p - pt\lambda\|\boldsymbol{v}\|_p^p$. This implies i).

The proof is complete. □

Proof of Proposition 12 Let $\mathbf{Q} \in \mathcal{M}$ and $\boldsymbol{v} \in \mathbb{R}^n$.

$$\begin{aligned}
\boldsymbol{v}^\mathsf{T}\mathbf{Q}\boldsymbol{v}^{\langle p-1\rangle} &= \sum_{i=1}^n q_{ii}|v^i|^p + \sum_{i<j} q_{ij}(v^i|v^j|^{p-1}\operatorname{sign}(v^j) + v^j|v^i|^{p-1}\operatorname{sign}(v^i)) \\
&\geqslant \sum_{i=1}^n q_{ii}|v^i|^p - \sum_{i<j}|q_{ij}|(|v^i||v^j|^{p-1} + |v^j||v^i|^{p-1}) \\
&\geqslant \sum_{i=1}^n q_{ii}|v^i|^p - \sum_{i<j}|q_{ij}|(|v^i|^p + |v^j|^p) \\
&= \sum_{i=1}^n |v^i|^p\left(q_{ii} - \sum_{j\neq i}|q_{ij}|\right),
\end{aligned}$$

where we used the inequality $x^p + y^p \geqslant x^{p-1}y + y^{p-1}x$ for any $p \geqslant 1$ and $x, y \geqslant 0$[4] to get the third line from the second line above. Hence the uniform p-PD of $\mathcal{M}$ follows from the item ii) of Theorem 11. The proof is complete. □

Proof of Proposition 13 Suppose $\mathcal{M}$ is uniform 1-PD. By the item i) of Theorem 11, there exists $\delta, L > 0$ such that $\|\mathbf{I} - t\mathbf{Q}\|_1 \leqslant 1 - Lt$ for all $t \in [0, \delta)$ and $\mathbf{Q} \in \mathcal{M}$. Let $\mathbf{Q} = (q_{ij})_{n\times n}$ and notice that

$$\|\mathbf{I} - t\mathbf{Q}\|_1 = \max_{i\in[n]}\left(|1 - tq_{ii}| + \sum_{j\in[n]\setminus\{i\}} t|q_{ij}|\right).$$

It follows that

$$\min_{i\in[n]}\left(q_{ii} - \sum_{j\in[n]\setminus\{i\}}|q_{ij}|\right) \geqslant L > 0.$$

Hence $\mathcal{M}$ is uniformly diagonally dominant (with non-negative diagonal). The proof is complete. □

Proof of Proposition 14 Suppose $\mathcal{M}$ is uniformly p-PD but not uniformly 2-PD. Then, there exists a sequence $\{\mathbf{Q}_1, \mathbf{Q}_2, \ldots\} \subseteq \mathcal{M}$ such that the smallest eigenvalues λ_m of $\mathbf{Q}_m$ satisfy

$$\lim_{m\to\infty} \lambda_m \leqslant 0. \tag{B.3}$$

For each $m \in \mathbb{N}^+$, there exists an $\boldsymbol{v}_m \in \mathbb{R}^n \setminus \{0\}$ such that $\mathbf{Q}_m\boldsymbol{v}_m = \lambda_m\boldsymbol{v}_m$. Hence

$$\boldsymbol{v}_m^\mathsf{T}\mathbf{Q}_m\boldsymbol{v}_m^{\langle p-1\rangle} = \lambda_m\boldsymbol{v}_m^\mathsf{T}\boldsymbol{v}_m^{\langle p-1\rangle} = \lambda_m\|\boldsymbol{v}_m\|_p^p.$$

But by the item ii) of Theorem 11, there exists $\lambda > 0$ such that $\lambda_m \geqslant \lambda$. This contradicts (B.3). The proof is complete. □

Proof of Theorem 3 We use a technique similar to Krasulina [1969]. Define the function

$$\boldsymbol{T}_t(\boldsymbol{x}) = \left(T_t^1(\boldsymbol{x}), \ldots, T_t^n(\boldsymbol{x})\right)^\mathsf{T} = \boldsymbol{x} - \boldsymbol{x}^* - \gamma_{t+1}\boldsymbol{R}(x).$$

An n-dimensional (and corrected) version of the first inequality in the proof of Krasulina [1969, Theorem 2] can be obtained by applying Lemma 8 to our stochastic approximation scheme,

$$\begin{aligned}
\|\boldsymbol{x}_{t+1} - \boldsymbol{x}^*\|_p^p &= \left\|\boldsymbol{T}_t(\boldsymbol{x}_t) - \gamma_{t+1}\boldsymbol{\xi}_{t+1}\right\|_p^p \\
&\leqslant \|\boldsymbol{T}_t(\boldsymbol{x}_t)\|_p^p + 4\gamma_{t+1}^p\left\|\boldsymbol{\xi}_{t+1}\right\|_p^p + p\gamma_{t+1}\sum_{i=1}^n \xi_{t+1}^i\left|T_t^i(\boldsymbol{x}_t)\right|^{p-1}\operatorname{sign}T_t^i(\boldsymbol{x}_t)
\end{aligned} \tag{B.4}$$

Since $\mathbb{E}\left[\xi_{t+1}^i|T_t^i(\boldsymbol{x}_t)|^{p-1}\operatorname{sign}T_t^i(\boldsymbol{x}_t) \mid \boldsymbol{x}_t\right] = |T_t^i(\boldsymbol{x}_t)|^{p-1}\operatorname{sign}T_t^i(\boldsymbol{x}_t)\,\mathbb{E}[\xi_{t+1}^i \mid \boldsymbol{x}_t] = 0$, by taking expectations in (B.4), we get

$$\begin{aligned}
\mathbb{E}\left[\|\boldsymbol{\delta}_{t+1}\|_p^p\right] &= \mathbb{E}\left[\|\boldsymbol{x}_{t+1} - \boldsymbol{x}^*\|_p^p\right] \\
&\leqslant \mathbb{E}\left[\|\boldsymbol{T}_t(\boldsymbol{x}_t)\|_p^p\right] + 4\gamma_{t+1}^p\mathbb{E}\left[\left\|\boldsymbol{\xi}_{t+1}\right\|_p^p\right] \\
&= \mathbb{E}\left[\|(\boldsymbol{x}_t - \boldsymbol{x}^*) - \gamma_{t+1}\boldsymbol{R}(\boldsymbol{x}_t)\|_p^p\right] + 4\gamma_{t+1}^p\mathbb{E}\left[\left\|\boldsymbol{\xi}_{t+1}\right\|_p^p\right].
\end{aligned}$$

[4] To see this, notice that for any $p \geqslant 1$ and $x, y \geqslant 0$, $x^p + y^p - x^{p-1}y - y^{p-1}x = (x^{p-1} - y^{p-1})(x - y) \geqslant 0$.

By the mean value theorem, there exists $\boldsymbol{x}_t^\flat \in \{\boldsymbol{x}^* + \tau(\boldsymbol{x}_t - \boldsymbol{x}^*) : 0 \leqslant \tau \leqslant 1\}$, such that $\boldsymbol{R}(\boldsymbol{x}_t) = \nabla \boldsymbol{R}(\boldsymbol{x}_t^\flat)(\boldsymbol{x}_t - \boldsymbol{x}^*)$, and then

$$\begin{aligned}
&\mathbb{E}\Big[\|(\boldsymbol{x}_t - \boldsymbol{x}^*) - \gamma_{t+1}\boldsymbol{R}(\boldsymbol{x}_t)\|_p^p\Big] + 4\gamma_{t+1}^p \mathbb{E}\Big[\big\|\boldsymbol{\xi}_{t+1}\big\|_p^p\Big] \\
&= \mathbb{E}\Big[\Big\|(\mathbf{I} - \gamma_{t+1}\nabla \boldsymbol{R}(\boldsymbol{x}_t^\flat))(\boldsymbol{x}_t - \boldsymbol{x}^*)\Big\|_p^p\Big] + 4\gamma_{t+1}^p \mathbb{E}\Big[\big\|\boldsymbol{\xi}_{t+1}\big\|_p^p\Big] \\
&\leqslant \mathbb{E}\Big[\Big\|\mathbf{I} - \gamma_{t+1}\nabla \boldsymbol{R}(\boldsymbol{x}_t^\flat)\Big\|_p^p \cdot \|\boldsymbol{x}_t - \boldsymbol{x}^*\|_p^p\Big] + 4\gamma_{t+1}^p \mathbb{E}\Big[\big\|\boldsymbol{\xi}_{t+1}\big\|_p^p\Big] \\
&\leqslant \mathbb{E}\Big[\Big\|\mathbf{I} - \gamma_{t+1}\nabla \boldsymbol{R}(\boldsymbol{x}_t^\flat)\Big\|_p^p \cdot \|\boldsymbol{\delta}_t\|_p^p\Big] + C_0\gamma_{t+1}^p\big(1 + \mathbb{E}\big[\|\boldsymbol{\delta}_t\|_p^p\big]\big),
\end{aligned}$$

where the last inequality follows from

$$\begin{aligned}
\mathbb{E}[|\boldsymbol{m}_{t+1}|^p \mid \mathcal{F}_t] &\leqslant \mathbb{E}\Big[|\boldsymbol{m}_{t+1}|^2 \mid \mathcal{F}_t\Big]^{p/2} \leqslant \big[K\big(1 + |\boldsymbol{x}_t|^2\big)\big]^{p/2} \\
&\leqslant K^{p/2}(1 + |\boldsymbol{x}_t|^p) \leqslant K^{p/2}\big(1 + 2^{p-1}(|\boldsymbol{\delta}_t|^p + |\boldsymbol{x}^*|^p)\big),
\end{aligned} \tag{B.5}$$

where we used the inequality $(x+y)^r \leqslant x^r + y^r$ for any $x, y \geqslant 0$, $0 \leqslant r \leqslant 1$ to obtain the first inequality in the second line above, as well as the assumption $\mathbb{E}[|\boldsymbol{\zeta}_1|^p] < \infty$.

Note that $\big\|\mathbf{I} - \gamma_{t+1}\nabla \boldsymbol{R}(\boldsymbol{x}_t^\flat)\big\|_p^p$ can be estimated by the uniform p-PD assumption (see item i) of Theorem 11) since $\gamma_t \to 0$. For t sufficiently large,

$$\Big\|\mathbf{I} - \gamma_{t+1}\nabla \boldsymbol{R}(\boldsymbol{x}_t^\flat)\Big\|_p^p \leqslant 1 - L\gamma_{t+1}.$$

And there is a positive constant C_1 such that $1 - L\gamma_{t+1} + C_0\gamma_{t+1}^p \leqslant 1 - C_1\gamma_{t+1}$ for t sufficiently large. Hence, we arrive at the following iterative bound

$$\mathbb{E}\Big[\|\boldsymbol{\delta}_{t+1}\|_p^p\Big] \leqslant (1 - \gamma_{t+1}C_1) \cdot \mathbb{E}\Big[\|\boldsymbol{\delta}_t\|_p^p\Big] + C_0\gamma_{t+1}^p \tag{B.6}$$

for t sufficiently large.

Next, let us substitute γ_{t+1} with $t^{-\rho}$ where $0 < \rho < 1$. Consider the iteration

$$\mu_{t+1} = (1 - t^{-\rho}C_1) \cdot \mu_t + C_0 t^{-\rho p}, \tag{B.7}$$

so that by (B.6), $\mathbb{E}\Big[\|\boldsymbol{\delta}_t\|_p^p\Big] = \mathcal{O}(\mu_t)$. By virtue of Lemma 9, we get

$$\mu_t = \Theta\Big(t^{-\rho(p-1)}\Big). \tag{B.8}$$

Therefore, by (B.6), (B.7), and (B.8), we obtain the following rate of convergence:

$$\mathbb{E}[\|\boldsymbol{\delta}_t\|_p^p] = \mathcal{O}\Big(t^{-\rho(p-1)}\Big).$$

Next, since p-norms on $\mathbb{R}^n$ are all equivalent, we can drop the subscript $\|\cdot\|_p$ and obtain

$$\mathbb{E}[|\boldsymbol{\delta}_t|^p] = \mathcal{O}\Big(t^{-\rho(p-1)}\Big).$$

Finally, by (B.5), we see that $\sup_{t\in\mathbb{N}^+} \mathbb{E}[|\boldsymbol{\xi}_t|^p] \leqslant \sup_{t\in\mathbb{N}^+} \mathbb{E}[2^{p-1}(|\boldsymbol{m}_t|^p + |\boldsymbol{\zeta}_t|^p)] < \infty$. The proof is complete. □ **Proof of Corollary** 4 Under the assumptions of Corollary 4, the rate $\mathbb{E}[|\boldsymbol{\delta}_t|^p] = \mathcal{O}\big(t^{-\rho(p-1)}\big)$ holds for every $p \in [q, \alpha)$. We can thus apply Jensen's inequality to strengthen it. By Jensen's inequality and (3.4), we get

$$\mathbb{E}[|\boldsymbol{\delta}_t|^q] \leqslant \mathbb{E}[|\boldsymbol{\delta}_t|^p]^{q/p} = \mathcal{O}\Big(t^{-\rho(p-1)\frac{q}{p}}\Big).$$

By letting $p \nearrow \alpha$, we conclude that have for every $\varepsilon > 0$,

$$\mathbb{E}[|\boldsymbol{\delta}_t|^q] = o\Big(t^{-\rho q\frac{\alpha-1}{\alpha}+\varepsilon}\Big).$$

The proof is complete. □

Next, we state a series of technical lemmas as well as their proofs, which will be used in the proofs of Theorems 5 and 6.

Lemma 16. *If $\gamma_t \asymp t^{-\rho}$ with $0 < \rho < \kappa \leqslant 1$, then for all $\lambda > 0$,*

$$\lim_{t\to\infty} t^{-\kappa} \sum_{j=1}^{t-1} \exp\left(-\lambda \sum_{i=j}^{t-1} \gamma_i\right) = 0.$$

Proof. Notice that there exists some constant $B > 0$ such that

$$\sum_{i=j}^{t-1} \gamma_i \geqslant \frac{B}{\lambda}\big(t^{1-\rho} - j^{1-\rho}\big).$$

It follows that

$$t^{-\kappa} \sum_{j=1}^{t-1} \exp\left(-\lambda \sum_{i=j}^{t-1} \gamma_i\right) \leqslant t^{-\kappa} \sum_{j=0}^{t-1} \exp\big(-Bt^{1-\rho} + Bj^{1-\rho}\big) = \frac{\sum_{j=0}^{t-1} \exp(Bj^{1-\rho})}{t^{\kappa}\exp(Bt^{1-\rho})}.$$

By Stolz-Cesàro theorem, we have

$$\begin{aligned}
\frac{\sum_{j=0}^{t-1} \exp(Bj^{1-\rho})}{t^{\kappa}\exp(Bt^{1-\rho})} &\asymp \frac{\exp(Bt^{1-\rho})}{(t+1)^{\kappa}\exp(B(t+1)^{1-\rho}) - t^{\kappa}\exp(Bt^{1-\rho})} \\
&= \frac{1}{(t+1)^{\kappa}\exp[B((t+1)^{1-\rho} - t^{1-\rho})] - t^{\kappa}} \\
&= \frac{1}{(t+1)^{\kappa}\exp[B(1-\rho)(t+1)^{-\rho} + o(t^{-\rho})] - t^{\kappa}} \\
&= \frac{1}{(t+1)^{\kappa}[1 + B(1-\rho)(t+1)^{-\rho} + o(t^{-\rho})] - t^{\kappa}} \\
&= \frac{1}{B(1-\rho)(t+1)^{\kappa-\rho} + o((t+1)^{\kappa-\rho})} \\
&\to 0,
\end{aligned}$$

as $t \to \infty$. The proof is complete. □

Lemma 17. *Suppose $\gamma_t \asymp t^{-\rho}$ and $0 < \rho < \kappa \leqslant 1$; let $\mathbf{A}$ be a positive definite symmetric matrix. Consider the matrix recursion in [Polyak and Juditsky, 1992, Lemma 1],*

$$\mathbf{X}_j^j = \mathbf{I}, \quad \mathbf{X}_j^{t+1} = \mathbf{X}_j^t - \gamma_t \mathbf{A}\mathbf{X}_j^t, \quad (j \in \mathbb{N}^+)$$

and define

$$\overline{\mathbf{X}}_j^t = \gamma_j \sum_{i=j}^{t-1} \mathbf{X}_j^i, \quad \mathbf{\Phi}_j^t = \mathbf{A}^{-1} - \overline{\mathbf{X}}_j^t.$$

Then the following limit holds,

$$\lim_{t\to\infty} \frac{1}{t^{\kappa}} \sum_{j=1}^{t-1} \|\mathbf{\Phi}_j^t\| = 0.$$

Remark. Lemma 17 recovers [Polyak and Juditsky, 1992, Lemma 1] as the special case $\kappa = 1$.

Proof of Lemma 17 Modeling after Polyak and Juditsky [1992]'s proof of their Lemma 1, we define $\mathbf{S}_j^t = \sum_{i=j}^{t-1}(\gamma_i - \gamma_j)\mathbf{X}_j^i$, and we have

$$\mathbf{\Phi}_j^t = \mathbf{S}_j^t + \mathbf{A}^{-1}\mathbf{X}_j^t.$$

We will split the proofs into two parts. In the first part, we will prove $t^{-\kappa}\sum_{j=1}^{t-1}\|\mathbf{S}_j^t\| \to 0$ and then in the second part we will prove $t^{-\kappa}\sum_{j=1}^{t-1}\|\mathbf{X}_j^t\| \to 0$.

Part I. We first prove that $t^{-\kappa}\sum_{j=1}^{t-1}\|\mathbf{S}_j^t\| \to 0$.

By the Part 3 of Polyak and Juditsky [1992, Lemma 1][5], there exist some $\lambda > 0$ and $K < \infty$ such that

$$\|\mathbf{X}_j^t\| \leqslant K \exp\left(-2\lambda \sum_{i=j}^{t-1} \gamma_i\right) = K e^{-2\lambda m_j^t}, \tag{B.9}$$

where m_k^ℓ stands for $\sum_{i=k}^{\ell-1} \gamma_i$. Now we have

$$\begin{aligned}
\|\mathbf{S}_j^t\| &= \left\| \sum_{i=1}^{t} (\gamma_i - \gamma_j) \mathbf{X}_j^i \right\| \\
&= \left\| \sum_{i=1}^{t} \left[\sum_{k=j}^{i-1} (\gamma_{k+1} - \gamma_k) \right] \mathbf{X}_j^i \right\| \\
&\leqslant C_0 \sum_{i=j}^{t} \sum_{k=j}^{i-1} k^{-\rho-1} \exp\left(-2\lambda m_j^i\right) \\
&\leqslant C_0 j^{-1} \sum_{i=j}^{t} \sum_{k=j}^{i-1} k^{-\rho} \exp\left(-2\lambda m_j^i\right) \\
&\leqslant C_1 j^{-1} \sum_{i=j}^{t} m_j^i \exp\left(-2\lambda m_j^i\right) \\
&= C_1 j^{-1} \sum_{i=j}^{t} \frac{m_j^i e^{-2\lambda m_j^i} (m_j^i - m_j^{i-1})}{\gamma_i},
\end{aligned} \tag{B.10}$$

where C_0, C_1 are some positive constants.

Since the function $f_w(x) = x^\rho \exp(-w x^{1-\rho})$ is bounded on $x \in [1, \infty)$ for every $w > 0$, we have

$$\frac{j^{-\rho}}{\gamma_i} \exp\left(-\lambda m_j^i\right) \leqslant C_2 i^\rho j^{-\rho} \exp(-C_3(i^{1-\rho} - j^{1-\rho})) = C_2 f_{C_3}(i)/f_{C_3}(j) \leqslant C_4,$$

for some positive constants C_2, C_3 and C_4. Hence, continuing upon (B.10),

$$\|\mathbf{S}_j^t\| \leqslant C_1 C_4 j^{\rho-1} \sum_{i=j}^{t} m_j^i e^{-\lambda m_j^i} (m_j^i - m_j^{i-1}).$$

Since the summation $\sum_{i=j}^{t} m_j^i e^{-\lambda m_j^i} (m_j^i - m_j^{i-1})$ approximates $\int_0^{m_j^t} m e^{-\lambda m} \mathrm{d}m$, it is bounded. Hence, for some positive constant C_5,

$$\|\mathbf{S}_j^t\| \leqslant C_5 j^{\rho-1},$$

which implies the desired limit

$$\lim_{t\to\infty} t^{-\kappa} \sum_{j=1}^{t-1} \|\mathbf{S}_j^t\| = 0.$$

Part II. It remains to prove that $t^{-\kappa} \sum_{j=1}^{t-1} \|\mathbf{X}_j^t\| \to 0$.

Combining (B.9) and Lemma 16, we have $t^{-\kappa} \sum_{j=1}^{t-1} \|\mathbf{X}_j^t\| \to 0$. Hence the proof of this lemma is complete. □

Lemma 18. *Given the assumption of Theorem 5 or Theorem 6,*

$$\frac{\boldsymbol{\xi}_1 + \ldots \boldsymbol{\xi}_t}{t^{1/\alpha}} \xrightarrow[t\to\infty]{\mathcal{D}} \mu.$$

[5] We can directly use this inequality since our assumption on step-size $\gamma_t \asymp t^{-\rho}$, $0 < \rho < 1$ can meet Polyak and Juditsky [1992, Assumption 2.2].

Proof. We recall the decomposition $\boldsymbol{\xi}_t = \boldsymbol{\zeta}_t + \boldsymbol{m}_t$, where $\{\boldsymbol{\zeta}_t\}$ are i.i.d. and $\boldsymbol{\zeta}_1$ is in the domain of normal attraction of an n-dimensional centered α-stable distribution so that

$$\frac{\boldsymbol{\zeta}_1 + \ldots + \boldsymbol{\zeta}_t}{t^{1/\alpha}} \xrightarrow[t\to\infty]{\mathcal{D}} \mu.$$

Hence, it suffices to show that $t^{-1/\alpha}(\boldsymbol{m}_1 + \ldots + \boldsymbol{m}_t) \to 0$ in L^r, for some $r \geqslant 1$.

By (3.3), there exists a constant $C > 0$ such that

$$\mathbb{E}\Big[|\boldsymbol{m}_{t+1}(\boldsymbol{x}_t)|^2 \mid \mathcal{F}_t\Big] \leqslant K\big(1 + |\boldsymbol{x}_t|^2\big) \leqslant K(1 + 2|\boldsymbol{x}^*|^2 + 2|\boldsymbol{\delta}_t|^2) \leqslant C(1 + |\boldsymbol{\delta}_t|^2).$$

Hence, by using the "Remark" on p.151 of Neveu [1975] (cf. inequalities (20) of Anantharam and Borkar [2012]), we get

$$\begin{aligned}
\mathbb{E}\left[\left|\frac{\boldsymbol{m}_1 + \ldots + \boldsymbol{m}_t}{t^{1/\alpha}}\right|^r\right] &\leqslant \frac{C_1}{t^{r/\alpha}}\mathbb{E}\left[\left(\sum_{i=1}^{t}\mathbb{E}\big[|\boldsymbol{m}_i|^2 \mid \mathcal{F}_{i-1}\big]\right)^{r/2}\right] \\
&\leqslant \frac{C_2}{t^{r/\alpha}}\mathbb{E}\left[\left(\sum_{i=1}^{t}\big(1 + |\boldsymbol{\delta}_{i-1}|^2\big)\right)^{r/2}\right] \\
&\leqslant \frac{C_2}{t^{r/\alpha}}\mathbb{E}\left[t^{r/2} + \sum_{i=1}^{t}|\boldsymbol{\delta}_{i-1}|^r\right], \qquad \text{(B.11)}
\end{aligned}$$

where, for the last inequality, we use the fact that $(x+y)^s \leqslant x^s + y^s$ for any $x, y \geqslant 0$, $0 \leqslant s \leqslant 1$. If the assumption of Theorem 5 holds, take $r = p > (\alpha + \alpha\rho)/(1 + \alpha\rho)$ in the inequalities (B.11) above. Then, by Theorem 3, $\mathbb{E}[|\boldsymbol{\delta}_t|^r] = \mathcal{O}(t^{-\rho(r-1)}) = o(t^{r/\alpha - 1})$.

If the assumption of Theorem 6 holds, take $r = q > 1/\rho > \alpha/(1 + \rho(\alpha - 1))$ in the inequalities (B.11) above. Then by Corollary 4, $\mathbb{E}[|\boldsymbol{\delta}_t|^r] = \tilde{\mathcal{O}}(t^{-\rho r(\alpha-1)/\alpha}) = o(t^{r/\alpha - 1})$.

In both cases, $t^{-1/\alpha}(\boldsymbol{m}_1 + \ldots + \boldsymbol{m}_t) \to 0$ in L^r. The proof is complete. □

Finally, we are ready to prove Theorems 5 and 6.

Proof of Theorem 5 By Polyak and Juditsky [1992, Lemma 2]:

$$\frac{t}{t^{1/\alpha}}\overline{\boldsymbol{\delta}}_t = \underbrace{\frac{1}{t^{1/\alpha}}\mathbf{F}_t\boldsymbol{\delta}_0}_{\boldsymbol{I}_t^{(1)}} - \underbrace{\frac{1}{t^{1/\alpha}}\sum_{j=1}^{t-1}\mathbf{A}^{-1}\boldsymbol{\xi}_j}_{\boldsymbol{I}_t^{(2)}} - \underbrace{\frac{1}{t^{1/\alpha}}\sum_{j=1}^{t-1}\mathbf{W}_j^t\boldsymbol{\xi}_j}_{\boldsymbol{I}_t^{(3)}}, \qquad \text{(B.12)}$$

where $\mathbf{F}_t$ and $\mathbf{W}_j^t$ are deterministic matrices with uniformly bounded operator 2-norms defined as

$$\mathbf{F}_t = \sum_{i=0}^{t-1}\prod_{k=1}^{i}(\mathbf{I} - \gamma_k\mathbf{A}), \qquad \text{(B.13)}$$

$$\mathbf{W}_j^t = \gamma_j\sum_{i=j}^{t-1}\prod_{k=j+1}^{i}(\mathbf{I} - \gamma_k\mathbf{A}) - \mathbf{A}^{-1}. \qquad \text{(B.14)}$$

We have $\boldsymbol{I}_t^{(1)} \to 0$ by the boundedness of $\mathbf{F}_t$. Next, take some κ such that

$$\max(\rho, 1/\alpha) < \kappa \leqslant p/\alpha. \qquad \text{(B.15)}$$

We shall prove that $\boldsymbol{I}_t^{(3)} \to 0$ in $L^{\alpha\kappa}$ (notice that $1 < \alpha\kappa \leqslant p < \alpha$; cf. Polyak and Juditsky [1992, Proof of Theorem 1] where convergence in L^2 is proven). By Theorem 3, $\sup_j \mathbb{E}[|\boldsymbol{\xi}_j|^p] < \infty$. Hence

we can compute, by virtue of Lemma 7, that

$$\begin{aligned}\mathbb{E}\Big[\Big|\boldsymbol{I}_t^{(3)}\Big|^{\alpha\kappa}\Big] &= \mathbb{E}\left[\left|\frac{1}{t^{1/\alpha}}\sum_{j=1}^{t-1}\mathbf{W}_j^t\boldsymbol{\xi}_j\right|^{\alpha\kappa}\right] \leqslant \frac{C_0}{t^\kappa}\sum_{j=1}^{t-1}\mathbb{E}\big[\big|\mathbf{W}_j^t\boldsymbol{\xi}_j\big|^{\alpha\kappa}\big] \\ &\leqslant \left(\frac{C_0}{t^\kappa}\sum_{j=1}^{t-1}\big\|\mathbf{W}_j^t\big\|^{\alpha\kappa}\right)\sup_j\mathbb{E}\big[\big|\boldsymbol{\xi}_j\big|^{\alpha\kappa}\big] \leqslant \left(\frac{C_0}{t^\kappa}\sum_{j=1}^{t-1}\big\|\mathbf{W}_j^t\big\|\right)\sup_j\mathbb{E}\big[\big|\boldsymbol{\xi}_j\big|^{\alpha\kappa}\big] \\ &\leqslant \frac{C_1}{t^\kappa}\sum_{j=1}^{t-1}\big\|\mathbf{W}_j^t\big\|.\end{aligned}$$

Notice that the matrices $\mathbf{W}_j^t$ defined above correspond to $-\boldsymbol{\Phi}_j^t$ in Lemma 17. This infers that $\mathbb{E}\Big[|\boldsymbol{I}_t^{(3)}|^{\alpha\kappa}\Big] \leqslant \frac{K_1}{t^\kappa}\sum_{j=1}^{t-1}\|\mathbf{W}_j^t\| \to 0$ as $t\to\infty$.

Finally, Lemma 18 states that $\boldsymbol{I}_t^{(2)}$ converges weakly to an α-stable distribution. Hence we conclude the proof. □

Proof of Theorem 6 Denote by $\mathbf{A}$ the Hessian matrix $\boldsymbol{\nabla}\boldsymbol{R}(\boldsymbol{x}^*) = \boldsymbol{\nabla}^2 f(\boldsymbol{x}^*)$. Consider a corresponding linear SA process with the same noise,

$$\boldsymbol{x}_{t+1}^1 = \boldsymbol{x}_t^1 - \gamma_{t+1}\big(\mathbf{A}(\boldsymbol{x}_t^1 - \boldsymbol{x}^*) + \boldsymbol{\xi}_{t+1}(\boldsymbol{x}_t)\big), \tag{B.16}$$

with $\boldsymbol{x}_0^1 = \boldsymbol{x}_0$. We further define $\boldsymbol{\delta}_t^1 = \boldsymbol{x}_t^1 - \boldsymbol{x}^*$ and the averaging process $\overline{\boldsymbol{\delta}}_t^1 = (\boldsymbol{\delta}_0^1 + \ldots + \boldsymbol{\delta}_{t-1}^1)/t$.

Part I. We first prove that $t^{1-1/\alpha}\Big(\overline{\boldsymbol{\delta}}_t^1 - \overline{\boldsymbol{\delta}}_t\Big) \to 0$ almost surely.

By (B.12), we have

$$\frac{t}{t^{1/\alpha}}\overline{\boldsymbol{\delta}}_t^1 = \frac{1}{t^{1/\alpha}}\mathbf{F}_t\boldsymbol{\delta}_0 - \frac{1}{t^{1/\alpha}}\sum_{j=1}^{t-1}\big(\mathbf{A}^{-1} + \mathbf{W}_j^t\big)\boldsymbol{\xi}_j, \tag{B.17}$$

where the matrices $\mathbf{F}_t$ and $\mathbf{W}_j^t$ are defined back in (B.13) and (B.14). For the non-linear process (2.1), it can be viewed *as if it is a linear process with the j-th noise term being $\boldsymbol{\xi}_j + \boldsymbol{R}(\boldsymbol{x}_{j-1}) - \mathbf{A}\boldsymbol{\delta}_{j-1}$*. Hence by (B.12), we have

$$\frac{t}{t^{1/\alpha}}\overline{\boldsymbol{\delta}}_t = \frac{1}{t^{1/\alpha}}\mathbf{F}_t\boldsymbol{\delta}_0 - \frac{1}{t^{1/\alpha}}\sum_{j=1}^{t-1}\big(\mathbf{A}^{-1} + \mathbf{W}_j^t\big)\big(\boldsymbol{\xi}_j + \boldsymbol{R}(\boldsymbol{x}_{j-1}) - \mathbf{A}\boldsymbol{\delta}_{j-1}\big). \tag{B.18}$$

Combining (B.17) and (B.18) yields the difference (cf. Part 4 of Polyak and Juditsky [1992, Proof of Theorem 2])

$$\frac{t}{t^{1/\alpha}}\Big(\overline{\boldsymbol{\delta}}_t^1 - \overline{\boldsymbol{\delta}}_t\Big) = \frac{1}{t^{1/\alpha}}\sum_{j=1}^{t-1}\big(\mathbf{A}^{-1} + \mathbf{W}_j^t\big)(\boldsymbol{R}(\boldsymbol{x}_{j-1}) - \mathbf{A}\boldsymbol{\delta}_{j-1}). \tag{B.19}$$

We also recall the assumption that $|\boldsymbol{R}(\boldsymbol{x}_j) - \mathbf{A}\boldsymbol{\delta}_j| \leqslant K|\boldsymbol{\delta}_j|^q$. Hence, it suffices to show the following term vanishes almost surely as $t\to\infty$:

$$J_t = \frac{1}{t^{1/\alpha}}\sum_{j=1}^{t-1}|\boldsymbol{\delta}_j|^q.$$

To show this, first by our calculation of the rate of convergence in Corollary 4,

$$\mathbb{E}\left[\sum_{j=1}^{t-1}\frac{1}{j^{1/\alpha}}|\boldsymbol{\delta}_j|^q\right] = \sum_{j=1}^{t-1}\tilde{\mathcal{O}}\Big(j^{-\rho q\frac{\alpha-1}{\alpha}-\frac{1}{\alpha}}\Big) = \mathcal{O}(1).$$

The last equality holds since $-\rho q\frac{\alpha-1}{\alpha} - \frac{1}{\alpha} < -1$. Hence, we have

$$\mathbb{P}\left[\sum_{j=1}^{t-1}\frac{1}{j^{1/\alpha}}|\boldsymbol{\delta}_j|^q < \infty\right] = 1. \tag{B.20}$$

By Kronecker's lemma, (B.20) implies that $\mathbb{P}[\lim_{t\to\infty} J_t = 0] = 1$. This further implies that the left hand side of (B.19), $t^{1-1/\alpha}\left(\overline{\boldsymbol{\delta}}_t^1 - \overline{\boldsymbol{\delta}}_t\right)$, converges to 0 almost surely.

Part II. It remains to show that $t^{1-1/\alpha}\overline{\boldsymbol{\delta}}_t^1$ converges weakly to an α-stable distribution.

Define $\overline{\boldsymbol{x}}_t^1 = (\boldsymbol{x}_0^1 + \ldots + \boldsymbol{x}_{t-1}^1)/t$. Since $t^{1-1/\alpha}\left(\overline{\boldsymbol{x}}_t^1 - \overline{\boldsymbol{x}}_t\right) = t^{1-1/\alpha}\left(\overline{\boldsymbol{\delta}}_t^1 - \overline{\boldsymbol{\delta}}_t\right) \to 0$ almost surely, it follows *a fortiori* that $\overline{\boldsymbol{x}}_t^1 - \overline{\boldsymbol{x}}_t \to 0$ almost surely. Hence $\boldsymbol{x}_t^1 - \boldsymbol{x}_t \to 0$ almost surely, due to the well-known theorem that a real-valued sequence converges to zero if and only if the average sequence converges to zero.

Therefore, for the noise decomposition $\boldsymbol{\xi}_{t+1}(\boldsymbol{x}_t) = \boldsymbol{\zeta}_{t+1} + \boldsymbol{m}_{t+1}(\boldsymbol{x}_t)$, the state-dependent component $\boldsymbol{m}_{t+1}(\boldsymbol{x}_t)$ satisfies not only (3.3), i.e.,

$$\mathbb{E}\Big[|\boldsymbol{m}_{t+1}(\boldsymbol{x}_t)|^2 \mid \mathcal{F}_t\Big] \leqslant K\big(1 + |\boldsymbol{x}_t|^2\big),$$

but also

$$\mathbb{E}\Big[|\boldsymbol{m}_{t+1}(\boldsymbol{x}_t)|^2 \mid \mathcal{F}_t\Big] \leqslant C\big(1 + |\boldsymbol{x}_t^1|^2\big).$$

Hence, combining the discussion above and Lemma 18, we know that the linear recursion (B.16) defines a process that satisfies Theorem 5. (The only difference is that κ, instead of (B.15), can be taken from the range $(\rho, 1)$ under the assumption of the current theorem, since by Theorem 3, $\sup_{t\in\mathbb{N}^+} \mathbb{E}[|\boldsymbol{\xi}_t|^p] < \infty$ for every $1 \leqslant p < \alpha$.) It then follows from Theorem 5 that $t^{1-1/\alpha}\overline{\boldsymbol{\delta}}_t^1$ converges weakly to an α-stable distribution.

The proof is complete. □

C Additional Technical Background

C.1 Properties of α-Stable Distributions

An α-stable distributed random variable X is denoted by $X \sim \mathcal{S}_\alpha(\sigma, \theta, \mu)$, where $\alpha \in (0, 2]$ is the *tail-index*, $\theta \in [-1, 1]$ is the *skewness* parameter, $\sigma \geqslant 0$ is the *scale* parameter, and $\mu \in \mathbb{R}$ is called the *location* parameter. An α-stable random variable X is uniquely characterized by its characteristic function: $\mathbb{E}[\exp(iuX)] = e^{-\sigma^\alpha |u|^\alpha (1 - i\theta \operatorname{sgn}(u) \tan(\frac{\pi\alpha}{2})) + i\mu u}$, if $\alpha \neq 1$ and $\mathbb{E}[\exp(iuX)] = e^{-\sigma|u|(1 + i\theta \frac{2}{\pi} \operatorname{sgn}(u) \log|u|) + i\mu u}$, if $\alpha = 1$, for any $u \in \mathbb{R}$. The mean of X coincides with μ if $\alpha > 1$, and otherwise the mean of X is undefined. The skewness parameter θ is a measure of asymmetry. We say that X follows a *symmetric* α-stable distribution denoted as $\mathcal{S}\alpha\mathcal{S}(\sigma) = \mathcal{S}_\alpha(\sigma, 0, 0)$ if $\theta = 0$ (and $\mu = 0$). The tail-index parameter $\alpha \in (0, 2]$ determines the tail thickness of the distribution, and $\sigma > 0$ measures the spread of X around its mode. When $\alpha < 2$, α-stable distributions have heavy tails so that their moments are finite only up to the order α. More precisely, let $X \sim \mathcal{S}_\alpha(\sigma, \theta, \mu)$ with $0 < \alpha < 2$. Then $\mathbb{E}[|X|^p] < \infty$ for any $0 < p < \alpha$ and $\mathbb{E}[|X|^p] = \infty$ for any $p \geqslant \alpha$, which implies infinite variance (see e.g. [Samorodnitsky and Taqqu, 1994, Property 1.2.16]). When $0 < \alpha < 2$, the left tail and right tail of X are described by the formulas:

$$\lim_{x\to\infty} x^\alpha \mathbb{P}(X > x) = \frac{1+\theta}{2} C_\alpha \sigma^\alpha, \qquad \lim_{x\to\infty} x^\alpha \mathbb{P}(X < -x) = \frac{1-\theta}{2} C_\alpha \sigma^\alpha,$$

where $C_\alpha := (1 - \alpha)/(\Gamma(2 - \alpha)\cos(\pi\alpha/2))$ if $\alpha \neq 1$ and $C_\alpha := 2/\pi$ if $\alpha = 1$, (see e.g. [Samorodnitsky and Taqqu, 1994, Property 1.2.15]). The family of α-stable distributions include normal, Lévy and Cauchy distributions as special cases, and can be used to model many complex stochastic phenomena [Sarafrazi and Yazdi, 2019, Fiche et al., 2013, Farsad et al., 2015].

C.2 Domains of Attraction of Stable Distributions

Let X_i be an i.i.d. sequence with a common distribution whose distribution function is denoted as F, and let $S_n := X_1 + X_2 + \cdots + X_n$. Suppose that for some normalizing constants $a_n > 0$ and b_n, the sequence $S_n/a_n - b_n$ has a non-degenerate limit distribution with distribution function G, i.e.

$$\lim_{n\to\infty} \mathbb{P}(S_n/a_n - b_n \leqslant x) = G(x), \tag{C.1}$$

for all continuity points x of G, then such limit distributions G are stable distributions and the set of distribution functions F such that $S_n/a_n - b_n$ converges to a particular stable distribution is called its *domain of attraction*.

Next, let us provide a sufficient and necessary condition for being in the domain of attraction of a stable distribution. The class of distribution functions F for which $S_n/a_n - b_n$ converges to $\mathcal{S}\alpha\mathcal{S}(\sigma)$ is called the α-stable domain of attraction, and we denote it as $F \in D_\alpha$. Before we proceed, let us recall that a positive measurable function f is *regularly varying* if there exists a constant $\gamma \in \mathbb{R}$ such that $\lim_{t\to\infty} \frac{f(tx)}{f(t)} = x^\gamma$, for every $x > 0$. In this case, we denote $f \in RV_\gamma$, and we say a function f is slowly varying if $f \in RV_0$.

Define the characteristic functions $\phi(u) := \int_{-\infty}^{\infty} e^{iux} dF(x)$ and $\psi(u) := \int_{-\infty}^{\infty} e^{iux} dG(x)$, and also define $\lambda(u) := \phi(1/u)$ and $g(u) := \psi(1/u)$ for $u \in [-\infty, \infty]\backslash\{0\}$. We also denote $U(x) := \mathrm{Re}\lambda(x)$ and $V(x) := \mathrm{Im}\lambda(x)$. By Lévy's continuity theorem for characteristic functions (see e.g. Feller [1971, Chapter XV.3]), the convergence in (C.1) is equivalent to $\lim_{n\to\infty} \exp(-ib_n/u)\lambda^n(a_n u) = g(u)$, $u \in [-\infty, \infty]\backslash\{0\}$ uniformly on neighborhoods of $\pm\infty$. Based on this, one can show that (see e.g.) if (C.1) holds, then $|g(u)|^2 = \exp(-c|u|^{-\alpha})$ for some $\alpha \in (0, 2]$ and $c > 0$ and moreover $-\log|\lambda| \in RV_{-\alpha}$, i.e. $-\log|\lambda|$ is regularly varying with index $-\alpha$. Next, we state a sufficient and necessary condition for being in the α-stable domain of attraction.

Theorem 19 (Geluk and de Hann [2000], Theorem 1)**.** *Suppose $0 < \alpha < 2$. Every α-stable random variable X has a characteristic function of the form:*

$$\mathbb{E}[\exp(iuX)] = \exp\left(-\left\{|u|^\alpha + iu(2p-1)\{(1-\alpha)\tan(\alpha\pi/2)\}\frac{|u|^{\alpha-1}-1}{\alpha-1}\right\}\right),$$

for some $0 \leqslant p \leqslant 1$ with $(1-\alpha)\tan(\pi/2)$ defined to be $2/\pi$ at $\alpha = 1$. The following statements are equivalent:

(i) $F \in D_\alpha$.

(ii) $1 - F(x) + F(-x) \in RV_{-\alpha}$ and there exists a constant $p \in [0, 1]$ such that

$$\lim_{x\to\infty} \frac{1-F(x)}{1-F(x)+F(-x)} = p.$$

(iii) $1 - U(x) \in RV_{-\alpha}$ and there exists a constant $p \in [0, 1]$ such that

$$\lim_{x\to\infty} \frac{xuV(xu) - xV(x)}{x(1-U(x))} = (2p-1)(1-\alpha)\tan\left(\frac{\alpha\pi}{2}\right)\frac{|u|^{1-\alpha}-1}{1-\alpha}, \qquad u \in \mathbb{R}\backslash\{0\}.$$

Furthermore, [Geluk and de Hann, 2000, Theorem 1] showed that if any of (i), (ii), (iii) holds, then $\lim_{x\to\infty} \frac{1-U(x)}{1-F(x)+F(-x)} = \Gamma(1-\alpha)\cos(\alpha\pi/2)$ and $\lim_{x\to\infty} \frac{V(x)-x^{-1}\int_0^x (1-F(y)-F(-y))\mathrm{d}y}{1-F(x)+F(-x)} = (2p-1)\left(\Gamma(1-\alpha)\sin(\alpha\pi/2) - \frac{1}{1-\alpha}\right)$.

Let us illustrate [Geluk and de Hann, 2000, Theorem 1] with an example of Pareto distribution, which is a power-law distribution widely applied in various fields. A random variable X is said to follow a Pareto distribution (of type I) if there exists some $c > 0$ such that $\mathbb{P}(X > x) = (x/c)^{-\alpha}$ for any $x \geqslant c$ and $\mathbb{P}(X > x) = 1$ for any $x < c$. In this case, $F(x) = 1 - (x/c)^{-\alpha}$ for any $x \geqslant c$ and $F(x) = 0$ for any $x < c$. It follows that $1 - F(x) + F(-x) \in RV_{-\alpha}$ and $\lim_{x\to\infty} \frac{1-F(x)}{1-F(x)+F(-x)} = 1$. Therefore, $F \in D_\alpha$ and the Pareto distribution is in the α-stable domain of attraction.

When the tail-index $\alpha \in (0, 2)$, the logarithm of the characteristic function (i.e. $\log \mathbb{E}\left[e^{iuX}\right]$) of an α-stable random variable X is of the form (see [Gnedenko and Kolmogorov, 1954, equation (12), page 168]):

$$i\gamma u + c_1 \int_{-\infty}^{0} \left[e^{iux} - 1 - \frac{iux}{1+x^2}\right]\frac{\mathrm{d}x}{|x|^{1+\alpha}} + c_2 \int_0^\infty \left[e^{iux} - 1 - \frac{iux}{1+x^2}\right]\frac{\mathrm{d}x}{x^{1+\alpha}}, \tag{C.2}$$

where $c_1, c_2 \geqslant 0$ and $\gamma \in \mathbb{R}$. Since the characteristic function uniquely characterizes a probability distribution, the triplet (c_1, c_2, α) uniquely determines an α-stable law up to a constant shift $\gamma \in \mathbb{R}$ when $0 < \alpha < 2$. [Gnedenko and Kolmogorov, 1954, Theorem 2, page 175] provides another

sufficient and necessary condition for being in the domain of attraction of an α-stable distribution, which complements [Geluk and de Hann, 2000, Theorem 1]. Suppose $0 < \alpha < 2$. Then, the distribution function $F(x)$ belongs to the domain of attraction of an α-stable distribution if and only if the following conditions hold: (i) $\lim_{x\to\infty} \frac{F(-x)}{1-F(x)} = \frac{c_1}{c_2}$. (ii) For every constant $\kappa > 0$, $\lim_{x\to\infty} \frac{1-F(x)+F(-x)}{1-F(\kappa x)+F(-\kappa x)} = \kappa^\alpha$. In the case of a Pareto distribution (of type I), for some $c > 0$, we have $F(x) = 1 - (x/c)^{-\alpha}$ for any $x \geqslant c$ and $F(x) = 0$ for any $x < c$. Then we can check that $\lim_{x\to\infty} \frac{F(-x)}{1-F(x)} = 0$ and for every constant $\kappa > 0$, $\lim_{x\to\infty} \frac{1-F(x)+F(-x)}{1-F(\kappa x)+F(-\kappa x)} = \lim_{x\to\infty} \frac{(x/c)^{-\alpha}}{(\kappa x/c)^{-\alpha}} = \kappa^\alpha$. Thus, the Pareto distribution belongs to the domain of attraction of an α-stable distribution.

Finally, let us provide a sufficient and necessary condition for being in the domain of normal attraction of a stable distribution.

Theorem 20 (Gnedenko and Kolmogorov [1954], Theorem 5, page 181)**.** *Suppose $0 < \alpha < 2$. The distribution function $F(x)$ belongs to the domain of attraction of an α-stable distribution characterized by* (C.2) *if and only if*

$$F(x) = (c_1 a^\alpha + \alpha_1(x))\frac{1}{|x|^\alpha}, \qquad \textit{for } x < 0, \tag{C.3}$$

$$F(x) = 1 - (c_2 a^\alpha + \alpha_2(x))\frac{1}{x^\alpha}, \qquad \textit{for } x > 0, \tag{C.4}$$

where $a > 0$ is a positive constant and $\alpha_1(x), \alpha_2(x)$ are functions satisfying $\lim_{x\to-\infty} \alpha_1(x) = \lim_{x\to\infty} \alpha_2(x) = 0$. Indeed, the constant a in (2.2)*,* (C.3) *and* (C.4) *is the same.*

In the case of a Pareto distribution (of type I), for some $c > 0$, we have $F(x) = 1 - (x/c)^{-\alpha}$ for any $x \geqslant c$ and $F(x) = 0$ for any $x < c$. Then we can check that (C.3) and (C.4) hold with $c_1 = 0$, $\alpha_1(x) \equiv 0$, $c_2 = 1$, $\alpha_2(x) \equiv 0$ and $a = c$. Thus, the Pareto distribution belongs to the domain of normal attraction of an α-stable distribution.