This article has been accepted for publication in IEEE Transactions on Information Theory. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIT.2026.3702077

Sharp Concentration Analysis of Online Stochastic Approximation Algorithms

Danna Zhang, *Member, IEEE,* Zhipeng Lou, and Wei Biao Wu

***Abstract*—This paper introduces a comprehensive framework for analyzing online stochastic approximation algorithms from the perspective of nonlinear time series analysis. We introduce functional dependence measure to capture the underlying complex dependence structure of these algorithms, which can significantly facilitate the theoretical studies. In the more realistic setting where the gradient noise exhibits only a finite q-th moment ($q > 1$), we develop a new Burkholder-type inequality as a powerful tool for analyzing general q-th moments of online stochastic approximation algorithms, and systematically quantifying the proposed functional dependence measures. Equipped with these quantified dependence measures, we propose insightful and elegant approaches to derive sharp q-th moment and tail probability inequalities for both the iterates of the stochastic gradient descent algorithm and the Polyak-Ruppert averaging estimators. In particular, to address the non-stationary nature of the outputs of these algorithms, we introduce a previously underexplored blocking technique with time-varying block sizes, which can be of independent interest when dealing with non-stationary processes.**

***Index Terms*—Concentration inequalities, Martingale, Nonlinear time series, Online learning, Stochastic gradient descent.**

I. Introduction

Advancements in data acquisition techniques have led to an explosion in data generation over the past several decades. One commonly encountered type of dataset is large-scale streaming data, which arises across a variety of fields, including signal processing, financial econometrics, industrial monitoring, e-commerce, the Internet of Things (IoT), meteorology, agriculture, and many others. Real-time analysis of sequentially collected data streams has become increasingly crucial across a wide range of domains. However, the sequential nature, high volume and high velocity of large-scale streaming data make traditional batch-based methods (e.g., M-estimation) computationally infeasible, thereby necessitating the development of new online learning methodology. In modern statistics and machine learning, stochastic approximation algorithms, particularly the Stochastic Gradient Descent (SGD) algorithm introduced by [1] and its variants ([2, 3, 4, 5, 6, 7, 8, 9, 10]), have become fundamental tools. These algorithms are widely adopted due to their computational efficiency and scalability, especially in the context of online optimization involving large-scale streaming data. In particular, the following optimization problem is investigated:

$$\omega^* = \underset{\omega\in\mathbb{R}^d}{\arg\min}\ F(\omega), \tag{1}$$

where $F(\omega) = \mathbb{E}_{X\sim\Xi}\{f(\omega, X)\}$ is the objective function of interest, and the minimizer $\omega^* \in \mathbb{R}^d$ is the population parameter vector. This paper focuses on the statistical inference of ω^*, a fundamental problem in both statistics and machine learning.

Given a sequence of sequential random samples $X_1, X_2, \ldots \sim \Xi$ and an initial value $\omega_0 \in \mathbb{R}^d$, the SGD algorithm updates the estimate of ω^* via the iteration

$$\omega_t = \omega_{t-1} - \gamma_t \nabla f(\omega_{t-1}, X_t), \tag{2}$$

where $\nabla f(\omega, X) \in \mathbb{R}^d$ denotes the gradient of $f(\omega, X)$ with respect to ω, and $(\gamma_t)_{t\geq 1}$ is the step size sequence defined by $\gamma_t = \gamma_1 t^{-\beta}$, with $\beta \in (0, 1)$. For efficient statistical inference of the population parameter, [11] and [12] proposed the averaged SGD estimator

$$\bar{\omega}_t = \frac{\omega_1 + \cdots + \omega_t}{t}.$$

Since the introduction of the SGD algorithm, its convergence properties and distribution theory have been extensively studied ([13, 14, 15, 16, 17, 18, 11, 19, 12, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31]). However, several fundamental challenges are still unresolved. In particular, deriving sharp non-asymptotic concentration inequalities and moment bounds remains an open problem in realistic settings where the gradient noise $\nabla f(\omega^*, X)$ possesses only a finite q-th moment with $q > 1$.

The primary objective of this paper is to propose a comprehensive framework for analyzing online stochastic approximation algorithms through a new lens: nonlinear time series analysis. Specifically, for each $t \geq 1$, the iteration (2) yields the following causal representation of the SGD iterate ω_t as a function of the inputs $X_1, \ldots, X_t$ up to time t,

$$\omega_t = h_t(X_t, X_{t-1}, \ldots, X_1), \tag{3}$$

where $h_t(\cdot) = (h_{t,1}(\cdot), \ldots, h_{t,d}(\cdot))^\top$ is an $\mathbb{R}^d$-valued measurable function. Given this representation, it is evident that the sequence $(\omega_t)_{t\geq 1}$ exhibits temporal dependence and constitutes a nonlinear causal process. Notably, this dependent process is non-stationary due to the time-varying step sizes. Analyzing such a non-stationary sequence is highly nontrivial and poses significant mathematical challenges. To address these difficulties, we propose studying the sequence $(\omega_t)_{t\geq 1}$

Danna Zhang is with the Department of Mathematics, University of California, San Diego, CA 92093, USA (email: daz076@ucsd.edu).

Zhipeng Lou is with the Department of Mathematics, University of California, San Diego, CA 92093, USA (email: zlou@ucsd.edu).

Wei Biao Wu is with the Department of Statistics, The University of Chicago, IL 60637, USA (email: wbwu@uchicago.edu).

from the new perspective of dynamic systems, employing the functional dependence measure introduced by [32] to capture the intrinsic complexity of the dependence structure. These measures substantially facilitate the theoretical development of concentration properties and distribution theory for temporally dependent sequences ([32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46]).

Non-asymptotic moment inequalities provide an effective tool for evaluating the finite-sample convergence properties of online stochastic approximation algorithms. The first type of fundamental inequalities we derive are upper bounds on $\mathbb{E}|\widehat{\omega}-\omega^*|^q$ for $\widehat{\omega} \in \{\omega_t, \bar{\omega}_t\}$, where $q>1$. The special case of the last iterate $\widehat{\omega}=\omega_t$ has been studied in the literature ([14, 15, 24, 28, 47, 48]). In this work, we establish a new Burkholder-type inequality that is both more computationally tractable and applicable to a broader class of cases. For the Polyak-Ruppert averaging estimator $\widehat{\omega}=\bar{\omega}_t$, the general q-th moment inequality with $q>1$ remains largely unexplored, except for the special case $q=2$ ([24]). Leveraging the quantified functional dependence measures for the SGD sequence $(\omega_t)_{t\geq 0}$, we aim to close this gap by proposing an elegant method to derive general q-th moment inequalities for functionals of SGD iterates. Remarkably, our approach only requires the standard convexity and stochastic Lipschitz conditions of the random gradient function $\nabla f(\cdot, X)$.

Another important class of concentration inequalities concerns tail probabilities, which play a central role in modern machine learning theory ([49, 50, 51, 52]). For online stochastic approximation algorithms, concentration inequalities of the form $\mathbb{P}(|\widehat{\omega}-\omega^*|\geq z)$ for $\widehat{\omega} \in \{\omega_t, \bar{\omega}_t\}$ enable the derivation of finite-sample high-probability bounds and facilitate the study of sample complexity ([49]). Existing works primarily focus on *light-tailed* settings, where the gradient noise $\nabla f(\omega^*, X)$ is either strictly bounded or exhibits exponential-type moments ([53, 54, 25, 55, 56, 27, 57]). However, in many realistic applications, the gradient noise typically possesses only a finite q-th moment. Consequently, the development of sharp tail probability inequalities for $\widehat{\omega} \in \{\omega_t, \bar{\omega}_t\}$ remains insufficient. In this paper, we significantly relax the moment conditions to allow for heavy-tailed distributions and establish new Nagaev-type concentration inequalities for SGD iterates. The celebrated Nagaev inequality, firstly introduced by [58], provides tight control over tail probabilities of the sum of independent random variables. To illustrate the idea, let $Y_1,\dots,Y_n \in \mathbb{R}$ be independent centered random variables with $\max_{t\geq 1}\mathbb{E}|Y_t|^q<\infty$ for some constant $q>2$. Then, for any $z>0$, [58] established the following tail probability inequality:

$$\mathbb{P}\left(\left|\sum_{t=1}^n Y_t\right|>z\right) \leq \frac{K_{q,1}\sum_{t=1}^n \mathbb{E}|Y_t|^q}{z^q} + 2\exp\left(-\frac{2z^2}{K_{q,2}\sum_{t=1}^n \mathbb{E}|Y_t|^2}\right), \quad (4)$$

where $K_{q,1}=(1+2/q)^q$ and $K_{q,2}=e^q(q+2)^2$. It should be noted that the polynomial term involving $\sum_{t=1}^n \mathbb{E}|Y_t|^q/z^q$ is known to be tight in general ([58]) and is significantly sharper than $\mathbb{E}|\sum_{t=1}^n Y_t|^q/z^q$, the order that can be obtained by applying the widely used Markov inequality. Specifically, for i.i.d. centered random variables $Y_1,\dots,Y_n$, we have $\sum_{t=1}^n \mathbb{E}|Y_t|^q/z^q \asymp n/z^q$, while $\mathbb{E}|\sum_{t=1}^n Y_t|^q/z^q \asymp n^{q/2}/z^q$. The latter can be much larger than n/z^q in view of $q>2$. Extensions of the Nagaev inequality for independent random variables to weakly dependent stationary processes have been studied in [59], [42] and [43]. In the context of online stochastic learning, [30] were among the first to derive the sharp Nagaev-type concentration inequality for the Polyak-Ruppert averaging estimator of the mini-batch SGD algorithm for linear regression model estimation. However, their argument relies on the linearity of the gradient function $\nabla F(\cdot)$ and is restricted to linear regression models. Developing Nagaev-type concentration inequalities for general SGD algorithms applied to optimization problems like (1), where the gradient function is nonlinear, remains an open problem. Moreover, when the gradient noise has only finite moments, the martingale decomposition of (2), which is commonly used to derive exponential-type concentration inequalities ([53, 54, 25, 55, 56, 27, 57]) fails to yield tight tail probability bounds. This poses significant mathematical challenges. As a major theoretical contribution, we propose a new method to derive sharp tail probability inequalities for $\widehat{\omega} \in \{\omega_t, \bar{\omega}_t\}$. Our technical novelty is twofold. First, we construct a nonlinear autoregressive coupling and develop a refined analysis that captures the nonlinear dynamics underlying the SGD iterates. Second, we introduce a previously underexplored blocking technique with time-varying block sizes to address the non-stationarity of the SGD sequence. Both techniques are of independent interest for the broader study of nonlinear and non-stationary stochastic processes.

The rest of this paper is organized as follows. In Section II, we establish a sharp Burkholder-type inequality and systematically quantify the functional dependence measures for the SGD sequence. We then introduce a new approach for deriving general q-th moment inequalities for both the SGD iterates and the Polyak-Ruppert averaging estimator. In Section III, we present a nonlinear autoregressive coupling of the SGD sequence and develop sharp concentration inequalities that bound the differences between the coupling processes and the original SGD sequence. In Section IV, we derive tight q-th moment and tail probability inequalities in the heavy-tailed setting, where the gradient noise exhibits infinite variance. In Section V, we conduct a finite-sample simulation study to illustrate the dichotomous phenomenon revealed by the established inequalities. All technical proofs are provided in Section VI.

We begin by introducing some frequently used notation. For a vector $\nu=(\nu_1,\dots,\nu_m)^\top \in \mathbb{R}^m$, its Euclidean norm is denoted by $|\nu|=(\nu_1^2+\cdots+\nu_m^2)^{1/2}$. The unit sphere in $\mathbb{R}^m$ is denoted by $\mathbb{S}^m=\{\nu\in\mathbb{R}^m : |\nu|=1\}$. For any random vector $Y\in\mathbb{R}^m$ with $\mathbb{E}|Y|^q<\infty$ for some constant $q>0$, we define $\|Y\|_q=(\mathbb{E}|Y|^q)^{1/q}$. Throughout this paper, we use $\mathcal{C},\mathcal{C}_0,\mathcal{C}_1,\mathcal{C}_2,\dots$ to denote positive constants, whose values may vary across different contexts. For two sequences of positive numbers $(a_n)_{n\geq 1}$ and $(b_n)_{n\geq 1}$, we write $a_n=O(b_n)$ if there exists a positive constant $\mathcal{C}$ such that $a_n\leq \mathcal{C}b_n$ for all sufficiently large n.

This article has been accepted for publication in IEEE Transactions on Information Theory. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIT.2026.3702077

II. Moment concentration inequalities

A. Dependence measure

Recall from (3) the causal representation $\omega_t = h_t(X_t, X_{t-1}, \dots, X_1)$ for the SGD iteration, which can be viewed as a dynamic system with $\{X_t, \dots, X_1\}$ as inputs, $h_t(\cdot)$ as the data-generating filter, and ω_t as the output. The temporal dependence of the process $(\omega_t)_{t\geq 1}$ follows directly from its causal structure. To quantify the underlying dependence structure, we employ the functional dependence measure introduced by [32] within the framework of nonlinear time series analysis, which is grounded in the concept of coupling. We first introduce some preliminary definitions. Let $\mathcal{F}_0 = \varnothing$, and for $h \geq 1$, define $\mathcal{F}_h = (X_h, \dots, X_1)$ and the projection operator

$$\mathcal{P}_h(\cdot) = \mathbb{E}(\cdot|\mathcal{F}_h) - \mathbb{E}(\cdot|\mathcal{F}_{h-1}).$$

Then, $\{\mathcal{P}_h(\cdot)\}_{h\geq 1}$ forms a sequence of martingale differences adapted to the filtration $\{\mathcal{F}_h\}_{h\geq 1}$. Martingale decomposition is a key tool in theoretical investigations of temporally dependent processes (e.g., [32]).

Definition 1 (Functional dependence measure)**.** Let $(X'_t)_{t\in\mathbb{N}}$ be independent and identically distributed (i.i.d.) copies of $(X_t)_{t\geq 1}$. For each $h \in \{1, \dots, t\}$, define the coupled random vector of ω_t as

$$\omega_{t,\{h\}} = h_t(X_t, \dots, X_{h+1}, X'_h, X_{h-1}, \dots, X_1).$$

Assume that $\max_{t\geq 1} \mathbb{E}|\omega_t|^q < \infty$ for some constant $q \geq 1$. The functional dependence measure of the temporally dependent sequence $(\omega_t)_{t\geq 1}$ is defined as

$$\varphi_{q,t,h} = \|\omega_t - \omega_{t,\{h\}}\|_q.$$

The functional dependence measure $\varphi_{q,t,h}$ defined above admits an important interpretation: *it quantifies the contribution of training sample X_h to the t-th SGD iterate ω_t.* Moreover, for each $h \in \{1, \dots, t\}$, $\varphi_{q,t,h}$ also provides an upper bound for the q-th moment of the martingale difference $\mathcal{P}_h(\omega_t)$, in view of Jensen's inequality. The framework of functional dependence measures provides the foundation for the subsequent analysis of functionals of the non-stationary sequence $(\omega_t)_{t\geq 0}$. For example, sharp concentration bounds for $|\bar{\omega}_t - \omega^*|$ can be derived by leveraging functional dependence measures together with the martingale decomposition $\omega_t = \mathbb{E}(\omega_t) + \sum_{h=1}^t \mathcal{P}_h(\omega_t)$ for each $t \geq 1$.

In Section II-B, we derive sharp upper bounds for the functional dependence measure $\varphi_{q,t,h}$. To this end, we impose the following regularity conditions.

Assumption 1. There exists a positive constant $\mu > 0$ such that for $\omega, \omega' \in \mathbb{R}^d$,

$$(\omega - \omega')^\top\{\nabla F(\omega) - \nabla F(\omega')\} \geq \mu|\omega - \omega'|^2.$$

Assumption 2. There exist positive constants $M_q < \infty$ and $L_q < \infty$ such that $\|\nabla f(\omega^*, X)\|_q \leq M_q$, and for $\omega, \omega' \in \mathbb{R}^d$,

$$\|\nabla f(\omega, X) - \nabla f(\omega', X)\|_q \leq L_q|\omega - \omega'|. \tag{5}$$

Assumption 1 guarantees that the objective function $F(\cdot)$ is strongly convex, which is a standard condition in the literature (e.g., [12, 24, 6, 60]). Assumption 2 imposes a natural Lipschitz continuity condition on the random gradient function $\nabla f(\cdot, X)$, measured in terms of the q-th moment. This generalizes Assumption H2 in [24], which corresponds the special case $q = 2$. Moreover, Assumption 2 ensures the Lipschitz continuity of the population gradient $\nabla F(\cdot)$. In particular, combining (5) with Jensen's inequality yields $|\nabla F(\omega) - \nabla F(\omega')| \leq L_1|\omega - \omega'|$ for $\omega, \omega' \in \mathbb{R}^d$.

B. Concentration inequality

We begin by establishing a Burkholder-type moment inequality, which plays an essential role in deriving sharp concentration inequalities. This result applies to a broad class of random vectors and may be of independent interest in probability theory. For instance, when applied to the sum of martingale differences, it recovers the classical Burkholder moment inequality for martingales (Lemma 13).

Proposition 1. *Let $\xi \in \mathbb{R}^d$ and $\zeta \in \mathbb{R}^d$ be two random vectors such that $\mathbb{E}|\xi|^q < \infty$ and $\mathbb{E}|\zeta|^q < \infty$ for some constant $q \geq 2$. Then the following inequality holds:*

$$\begin{aligned}&\|\xi + \zeta\|_q^2 \\ \leq &\|\xi\|_q^2 + 2\|\xi\|_q^{2-q}\mathbb{E}\{|\xi|^{q-2}\xi^\top\mathbb{E}(\zeta|\xi)\} + (q-1)\|\zeta\|_q^2.\end{aligned} \tag{6}$$

Remark 1. Proposition 1 can be viewed as a stochastic generalization of Lemma 2.1 in [61] and Proposition 7 in [62], both of which concern deterministic (non-random) vectors. When $\mathbb{E}(\zeta|\xi) = 0$, (6) reduces to

$$\|\xi + \zeta\|_q^2 \leq \|\xi\|_q^2 + (q-1)\|\zeta\|_q^2. \tag{7}$$

A special case of inequality (7), where $\xi, \zeta \in \mathbb{R}$ are scalar random variables, was established in [63]. It is also worth noting that the equality in (6) holds when $q = 2$. For further discussion on the tightness of the constant $q - 1$, see Remark 2.1 in [63]. For the infinite-variance case $q \in (1, 2)$, an analogous Burkholder-type inequality was established in [64] and is provided in the Appendix (Lemma 12) for reference. □

Equipped with Proposition 1, we establish an upper bound for the functional dependence measure of $(\omega_t)_{t\geq 1}$ in the following lemma. In particular, the bound holds uniformly for arbitrary step-size schemes $(\gamma_t)_{t\geq 1}$ and is applicable to a wide range of scenarios.

Lemma 2. *Under Assumptions 1 and 2, it holds that for all $t \geq 1$ and $h \in \{1, \dots, t\}$,*

$$\begin{aligned}\|\omega_t - \omega_{t,\{h\}}\|_q^2 \leq &8(L_q^2\|\omega_{h-1} - \omega^*\|_q^2 + M_q^2)\gamma_h^2 \\ &\cdot \prod_{m=h+1}^t \{1 - 2\mu\gamma_m + (q-1)L_q^2\gamma_m^2\}.\end{aligned} \tag{8}$$

Remark 2. It is worth noting that the upper bound in (8) depends explicitly on the q-th moment of the SGD iterates, as well as on both t and h. This feature highlights the ability of the bound to capture the complex temporal dependence structure of the SGD sequence in a fully non-asymptotic manner. Throughout this paper, we focus on the time-varying step-size scheme $\gamma_t = \gamma_1 t^{-\beta}$. By substituting γ_t, together with the non-asymptotic moment bound established in Lemma 3

This article has been accepted for publication in IEEE Transactions on Information Theory. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIT.2026.3702077

and Theorem 4 below, into (8), we obtain an explicit upper bound for the functional dependence measure, which plays a central role in the subsequent theoretical analysis. Recently, [65] derived an upper bound on the functional dependence measure for SGD with a constant step size, which can be viewed as a special case of the more general upper bound provided by (8). □

To quantify the effect of the initial estimate ω_0 on the SGD iterates $(\omega_t)_{t\geq 0}$ and to streamline the exposition of the subsequent theoretical analysis, we introduce an auxiliary coupled process $(\omega_t^\circ)_{t\geq 0}$. This process is initialized at the true parameter, $\omega_0^\circ = \omega^*$, and for $t \geq 1$, updated recursively by

$$\omega_t^\circ = \omega_{t-1}^\circ - \gamma_t \nabla f(\omega_{t-1}^\circ, X_t). \tag{9}$$

The process $(\omega_t^\circ)_{t\geq 0}$ shares the same temporal dependence structure as the original $(\omega_t)_{t\geq 0}$ defined in (2), differing only in the initialization. Hence (8) continues to hold for $(\omega_t^\circ)_{t\geq 0}$.

In the following lemma, we establish non-asymptotic q-th moment bounds for the differences $(\omega_t - \omega_t^\circ)$ and $(\bar{\omega}_t - \bar{\omega}_t^\circ)$.

Lemma 3. *Under Assumptions 1 and 2, for all $t \geq 1$, we have*

$$\|\omega_t - \omega_t^\circ\|_q \leq \mathcal{C}_1 \exp(-\mathcal{C}_0 t^{1-\beta})|\omega_0 - \omega^*|,$$
$$\|\bar{\omega}_t - \bar{\omega}_t^\circ\|_q \leq \frac{\mathcal{C}_2|\omega_0 - \omega^*|}{t},$$

where $\mathcal{C}_0$, $\mathcal{C}_1$ and $\mathcal{C}_2$ are positive constant independent of t and ω_0.

Building on Lemma 3, we focus on the coupling process $(\omega_t^\circ)_{t\geq 0}$ in what follows. We begin by deriving a non-asymptotic bound for the q-th moment of the SGD iterates $(\omega_t^\circ)_{t\geq 1}$.

Theorem 4. *Under Assumptions 1 and 2, for all $t \geq 1$, we have*

$$\|\omega_t^\circ - \omega^*\|_q^2 \leq \mathcal{C} M_q^2 \gamma_t, \tag{10}$$

where $\mathcal{C}$ is a positive constant independent of t.

Remark 3. It is remarkable that the rate of convergence $\|\omega_t^\circ - \omega^*\|_q = O(\sqrt{\gamma_t})$ is tight for any $\beta \in (0,1)$, given the asymptotic normality of $(\omega_t^\circ - \omega^*)/\sqrt{\gamma_t}$ as $t \to \infty$. Moreover, for all $t \geq 1$ and $h \in \{1,\ldots,t\}$, combining the non-asymptotic upper bound (10) with (8) yields the following bound for the functional dependence measure,

$$\|\omega_t^\circ - \omega_{t,\{h\}}^\circ\|_q^2 \leq \mathcal{C} M_q^2 \gamma_h^2 \prod_{m=h+1}^{t} \{1 - 2\mu\gamma_m + (q-1)L_q^2\gamma_m^2\}, \tag{11}$$

where $\mathcal{C}$ is a positive constant independent of t and h. In the special case of a constant step size $\gamma_t \equiv \gamma_1$, a non-asymptotic upper bound of a similar nature was established in the recent work [65] This result can be viewed as a counterpart of Theorem 4 corresponding to the case $\beta = 0$. □

We now turn to deriving the moment bound for the Polyak-Ruppert averaging estimator $\bar{\omega}_t^\circ$. Unlike the individual SGD iterates, deriving tight concentration inequalities for the general q-th moment of $\bar{\omega}_t^\circ$ remains relatively unexplored in the literature and presents significant technical challenges. In Theorem 5 below, we present a new result providing an upper bound for the general q-th moment of $\bar{\omega}_t^\circ$, leveraging the martingale decomposition and quantified functional dependence measures established in (11). Importantly, our analysis requires only the existence of the q-th moment of the stochastic noise (Assumption 2). Our approach decomposes the deviation of $\bar{\omega}_t^\circ$ into stochastic and bias component as follows,

$$\|\bar{\omega}_t^\circ - \omega^*\|_q \leq \|\bar{\omega}_t^\circ - \mathbb{E}(\bar{\omega}_t^\circ)\|_q + |\mathbb{E}(\bar{\omega}_t^\circ) - \omega^*|.$$

We propose a new approach to upper bound $\|\bar{\omega}_t^\circ - \mathbb{E}(\bar{\omega}_t^\circ)\|_q$ based on a martingale decomposition. Specifically, using the sequence of projection operators $\{\mathcal{P}_h(\cdot)\}_{h\geq 1}$, $\bar{\omega}_t^\circ$ admits the following martingale decomposition,

$$\bar{\omega}_t^\circ - \mathbb{E}(\bar{\omega}_t^\circ) = \frac{1}{t}\sum_{s=1}^{t}\sum_{h=0}^{s-1} \mathcal{P}_{s-h}(\omega_s^\circ) = \frac{1}{t}\sum_{h=0}^{t-1}\sum_{s=h+1}^{t} \mathcal{P}_{s-h}(\omega_s^\circ). \tag{12}$$

For each $h \in \{0,1,\ldots,s-1\}$, $\{\mathcal{P}_{s-h}(\omega_s^\circ)\}_{s\geq h+1}$ is a sequence of martingale differences adapted to the filtration $\{\mathcal{F}_{s-h}\}_{s\geq h+1}$, satisfying $\|\mathcal{P}_{s-h}(\omega_s^\circ)\|_q \leq \|\omega_s^\circ - \omega_{s,\{s-h\}}^\circ\|_q$. Consequently, the martingale moment inequality (Lemma 13 in the Appendix) can be applied.

To bound the estimation bias $|\mathbb{E}(\bar{\omega}_t^\circ) - \omega^*|$, we impose the following Lipschitz continuity assumption on the Hessian $\nabla^2 F(\cdot)$.

Assumption 3. There exists a positive constant $\varrho < \infty$ such that for $\omega \in \mathbb{R}^d$,

$$|\nabla F(\omega) - \nabla^2 F(\omega^*)(\omega - \omega^*)| \leq \varrho|\omega - \omega^*|^2.$$

Assumption 3 is standard and widely adopted in the literature on concentration inequalities and asymptotic distributions of Polyak-Ruppert averaging estimators in stochastic approximation algorithms ([11, 12, 24]).

Theorem 5. *Under Assumptions 1, 2 and 3, for all $t \geq 1$, we have*

$$\|\bar{\omega}_t^\circ - \omega^*\|_q \leq \frac{\mathcal{C}_1 M_q}{t^{1/2}} + \frac{\mathcal{C}_2 \max\{M_2^2, 1\}}{t^\beta}, \tag{13}$$

where $\mathcal{C}_1$ and $\mathcal{C}_2$ are positive constants independent of t.

Remark 4. Most existing results in the literature focus on the special case $q = 2$ and typically require a higher-order moment assumptions on the stochastic noise $\nabla f(\omega^*, X)$. For example, [24] and [47] assume the existence of fourth-order moments. In contrast, Theorem 5 only requires the existence of the q-th moment of the stochastic noise (Assumption 2), substantially weakening the moment conditions. As a result, our approach provides new insights even in the classical setting $q = 2$.

The second term on the right-hand side of (13), $\mathcal{C}_2 \max\{M_2^2, 1\} t^{-\beta}$, quantifies the estimation bias $|\mathbb{E}(\bar{\omega}_t^\circ) - \omega^*|$, which arises from the nonlinearity of the gradient function $\nabla F(\cdot)$. This bias term becomes asymptotically negligible when $\beta > 1/2$. According to the central limit theorem for the Polyak-Ruppert averaging estimator ([11, 12]), the convergence rate $t^{-1/2}$ in the first polynomial term is known to be tight. However, the numerator $\mathcal{C}_1 M_q$ in this term is generally suboptimal and can be improved under stronger moment assumptions. More specifically, for any $q \geq 2$, suppose that Assumption 2 holds with $2q$. Then, by leveraging the linear time-varying autoregressive coupling in (22), we obtain a refined upper bound for $\|\bar{\omega}_t^\circ - \omega^*\|_q$, as stated in the following proposition. □

Proposition 6. *Suppose that Assumptions 1 and 3 hold, and that Assumption 2 is satisfied with $2q$. Then, for all $t \geq 1$, we have*

$$\|\bar{\omega}_t^\circ - \omega^*\|_q \leq \Big\| \frac{1}{t} \sum_{s=1}^{t} H^{-1} \nabla f(\omega^*, X_s) \Big\|_q + O\Big(t^{\max\{-\beta, \beta/2-1\}}\Big), \tag{14}$$

where $H = \nabla^2 F(\omega^*) \in \mathbb{R}^{d \times d}$ denotes the Hessian matrix.

Remark 5. Proposition 6 extends Theorem 3 in [24], which considers only the case $q = 2$. Notably, the first term on the right hand side of (14) is independent of β and, when squared in the case $q = 2$, matches the Cramer-Rao lower bound $\mathrm{Tr}(\Sigma)/t$ in certain statistical models ([47]), where

$$\Sigma = H^{-1} \mathrm{Cov}\{\nabla f(\omega^*, X)\} H^{-1} \in \mathbb{R}^{d \times d}.$$

Since this bound is improvable for such models, (14) provides more reasonable guidance than (13) for selecting β in the step size $\gamma_t = \gamma_1 t^{-\beta}$. In particular, the convergence rate of the second term on the right hand side of (14) is minimized by choosing $\beta = 2/3$. When $q = 2$, (14) was further refined by [47] to

$$\|\bar{\omega}_t^\circ - \omega^*\|_2^2 \leq \frac{\mathrm{Tr}(\Sigma)}{t} + O\Big(t^{-\min\{(\beta+1/2),(2-\beta)\}}\Big),$$

under the additional assumption that the covariance matrix of the random gradient function $\nabla f(\omega, X)$ is Lipschitz continuous with respect to ω. It would be of interest to investigate whether analogous refinements can be established for $\|\bar{\omega}_t^\circ - \omega^*\|_q^2$ with $q > 2$, which we leave for future work. □

III. Nonlinear autoregressive coupling

The primary goal of this section is to develop tight concentration inequalities for the tail probabilities of SGD iterates, which are essential for analyzing finite-sample high-probability bounds and the sample complexity of stochastic algorithms. While concentration inequalities have been extensively studied in the context of modern machine learning (e.g., [66, 52, 67] and references therein), most existing results focus on independent data and cannot be extended to the sequence of SGD iterates. The non-stationary nature of the sequence $(\omega_t^\circ)_{t\geq 0}$ poses significant mathematical challenges for deriving sharp tail probability inequalities. Consequently, even the existing generalizations of concentration inequalities to stationary distributions (e.g., [59, 42, 68]) fail to apply here. A straightforward method is to apply Markov's inequality along with the moment inequalities in Theorems 4 and 5; however, this typically does not yield tight tail probability bounds when $q > 2$. In this work, we address this open problem and advance the probabilistic toolkit available for analyzing SGD sequence.

We propose a new approach to capture the nonlinear stochastic dynamics underlying the SGD iterates $(\omega_t^\circ)_{t\geq 0}$. Our analysis begins by introducing a nonlinear autoregressive coupling for $(\omega_t^\circ)_{t\geq 0}$. Observe that (9) can be equivalently expressed as

$$\omega_t^\circ = \omega_{t-1}^\circ - \gamma_t \nabla F(\omega_{t-1}^\circ) - \gamma_t\{\nabla f(\omega_{t-1}^\circ, X_t) - \nabla F(\omega_{t-1}^\circ)\}.$$

Under the Lipschitz continuity assumption on the gradient function (Assumption 2), it follows that $\|\{\nabla f(\omega_{t-1}^\circ, X_t) - \nabla F(\omega_{t-1}^\circ)\} - \nabla f(\omega^*, X_t)\|_q \leq 2L_q \|\omega_{t-1}^\circ - \omega^*\|_q$. This observation motivates the construction of an auxiliary nonlinear autoregressive coupling for the sequence $\{(\omega_t^\circ - \omega^*)\}_{t\geq 0}$. Specifically, we define the process $(\nu_t)_{t\geq 0}$, initialized with $\nu_0 = \omega_0^\circ - \omega^*$, and recursively updated according to

$$\nu_t = \nu_{t-1} - \gamma_t \nabla F(\omega^* + \nu_{t-1}) - \gamma_t \nabla f(\omega^*, X_t).$$

For notational convenience, we define $\varepsilon_t = -\nabla f(\omega^*, X_t)$ for $t \in \mathbb{N}$ in what follows. Since $\varepsilon_t \in \mathbb{R}^d$, $t \in \mathbb{N}$, are i.i.d. random vectors, the process $(\nu_t)_{t\geq 0}$ can be interpreted as a time-varying nonlinear autoregressive process. In the following lemma, we provide a non-asymptotic quantification of the coupling approximation error.

Lemma 7. *Under the conditions of Theorem 4, for all $t \geq 1$, we have*

$$\|(\omega_t^\circ - \omega^*) - \nu_t\|_q^2 \leq \mathcal{C}_1 M_q^2 \gamma_t^2, \tag{15}$$

and, for any vector $\alpha \in \mathbb{S}^d$,

$$\mathbb{P}(|\alpha^\top \mathbb{E}_0(\bar{\omega}_t^\circ - \bar{\nu}_t)| > z) \leq \frac{\mathcal{C}_2 t^{1-\beta} (\log t)^{q-1} M_q^q}{(tz)^q} + 4\exp\left(-\frac{\mathcal{C}_3 z^2 t^{1+\beta}}{M_2^2 \log t}\right), \tag{16}$$

where $\mathbb{E}_0(\cdot) = \cdot - \mathbb{E}(\cdot)$, and $\mathcal{C}_1$, $\mathcal{C}_2$, and $\mathcal{C}_3$ are positive constants independent of t and α.

Remark 6. The tail probability upper bound in (16) is essential for deriving a sharp Nagaev inequality for $\bar{\omega}_t^\circ$ in Section III-B. Specifically, as $\beta > 0$, it suffices to establish the concentration inequality for the nonlinear autoregressive coupling process $(\nu_t)_{t\geq 0}$, in view of Theorem 10. Unlike the q-th moment bound in (15), deriving the tail probability bound in (16) is highly nontrivial and mathematically challenging due to the complex dependence structure in the difference process $\{(\omega_t^\circ - \omega^* - \nu_t)\}_{t\geq 1}$. In particular, applying Markov's inequality yields the upper bound $\mathbb{E}|\alpha^\top \mathbb{E}_0(\bar{\omega}_t^\circ - \bar{\nu}_t)|^q / z^q$ for the tail probability $\mathbb{P}(|\alpha^\top \mathbb{E}_0(\bar{\omega}_t^\circ - \bar{\nu}_t)| > z)$, which is in general not sharp, especially when q is large. In classical time series analysis, the m-approximation method combined with the blocking technique is useful for handling temporal dependence ([36, 59, 40, 42, 43, 45]). However, the traditional blocking technique, which employs equal block sizes, is primarily designed for stationary sequences and is not well suited to SGD algorithms with time-varying step sizes $\gamma_t = \gamma_1 t^{-\beta}$, as the SGD sequence $(\omega_t^\circ)_{t\geq 0}$ is intrinsically non-stationary. To address this challenge, we propose a new Δ_k-approximation method, where $(\Delta_k)_{k\geq 1}$ is a sequence of non-decreasing integers, and modify the blocking technique using time-varying block sizes, adapting it to the non-stationary nature of the SGD sequence.

We briefly outline the construction below. For notational simplicity, define $\vartheta_t = \omega_t^\circ - \omega^* - \nu_t$. Let $\psi_0 = 0$ and let $(\psi_k)_{k\ge 1}$ be a sequence of positive integers such that $\psi_k - \psi_{k-1}$ is even and

$$\psi_k - \psi_{k-1} \asymp k^{\beta/(1-\beta)}(\log t)^{1/(1-\beta)}, \quad k \ge 1. \tag{17}$$

Without loss of generality, assume that $\psi_K = t$ for some integer $K \ge 1$. Define the consecutive block sums

$$\mathcal{B}_k = \sum_{s=\psi_{k-1}+1}^{\psi_{k-1}+\Delta_k} \vartheta_s \quad \text{and} \quad \mathcal{B}_k^\diamond = \sum_{s=\psi_{k-1}+\Delta_k+1}^{\psi_k} \vartheta_s$$

for $k = 1, \ldots, K$, where $\Delta_k = (\psi_k - \psi_{k-1})/2$. By construction, $\sum_{s=1}^t \vartheta_s = \sum_{k=1}^K (\mathcal{B}_k + \mathcal{B}_k^\diamond)$. Since $\mathcal{B}_k$ and $\mathcal{B}_k^\diamond$ exhibit different dependence structure across k, we introduce Δ_k-approximation for $\mathcal{B}_k$ and $\mathcal{B}_k^\diamond$ to be

$$\widetilde{\mathcal{B}}_k = \sum_{s=\psi_{k-1}+1}^{\psi_{k-1}+\Delta_k} \mathbb{E}(\vartheta_s | X_s, \ldots, X_{s-\Delta_k}),$$
$$\widetilde{\mathcal{B}}_k^\diamond = \sum_{s=\psi_{k-1}+\Delta_k+1}^{\psi_k} \mathbb{E}(\vartheta_s | X_s, \ldots, X_{s-\Delta_k}).$$

Under this construction, both the sequences $\{\widetilde{\mathcal{B}}_k\}_{k\ge1}$ and $\{\widetilde{\mathcal{B}}_k^\diamond\}_{k\ge1}$ consist of independent random vectors. Consequently, the Nagaev inequality in (4) can be applied to bound the tail probabilities of both $\sum_{k=1}^K \widetilde{\mathcal{B}}_k$ and $\sum_{k=1}^K \widetilde{\mathcal{B}}_k^\diamond$. It is worth mentioning that the sequence $(\psi_k)_{k\ge1}$ in (17) is chosen to balance these tail probabilities and the approximation error $|\sum_{s=1}^t \vartheta_s - \sum_{k=1}^K (\widetilde{\mathcal{B}}_k + \widetilde{\mathcal{B}}_k^\diamond)|$ introduced by the Δ_k-approximation. □

We now introduce another time-varying autoregressive process to quantify the estimation bias $|\mathbb{E}(\nu_t)|$ for $t \ge 1$. Under Assumptions 2 and 3, for any constant p satisfying $q/2 \le p \le q$, we have

$$\|\nabla F(\omega^* + \nu_{t-1}) - \nabla^2 F(\omega^*)\nu_{t-1}\|_p \le (2L_1 + \varrho)\|\nu_{t-1}\|_q^{q/p}. \tag{18}$$

This bound motivates the definition of an auxiliary linear time-varying autoregressive process $(\upsilon_t)_{t\ge0}$, initialized as $\upsilon_0 = \nu_0$, and updated for $t \ge 1$ via

$$\upsilon_t = \upsilon_{t-1} - \gamma_t \nabla^2 F(\omega^*)\upsilon_{t-1} + \gamma_t \varepsilon_t. \tag{19}$$

The process $(\upsilon_t)_{t\ge0}$ is a linear time-varying autoregressive process driven by i.i.d. innovations $(\varepsilon_t)_{t\ge1}$, with deterministic matrices $I_d - \gamma_t \nabla^2 F(\omega^*) =: A_t, t \ge 1$, as coefficient matrices. Consequently, $(\upsilon_t)_{t\ge1}$ admits the explicit linear representation

$$\upsilon_t = \sum_{h=1}^t \gamma_h \left(\prod_{m=h+1}^t A_m \right) \varepsilon_h, \tag{20}$$

and satisfies $\mathbb{E}(\upsilon_t) = 0$ for all $t \ge 1$. In the following lemma, we establish non-asymptotic bound on the p-th moment of the difference between the two coupling processes $(\nu_t)_{t\ge1}$ and $(\upsilon_t)_{t\ge1}$.

Lemma 8. *Under the conditions of Theorem 5, for any constant $q/2 \le p \le q$, and for all $t \ge 1$, we have*

$$\|\nu_t - \upsilon_t\|_p \le \mathcal{C}(2L_1 + \varrho) M_q^{q/p} \gamma_t^{q/(2p)}, \tag{21}$$

where $\mathcal{C}$ is a positive constant independent of t.

Remark 7. As shown in (21), the approximation error bound between $(\nu_t)_{t\ge1}$ and $(\upsilon_t)_{t\ge1}$ becomes tighter as the constant ϱ in Assumption 3 decreases. Moreover, since $\mathbb{E}(\upsilon_t) = 0$ for all $t \ge 1$, when $q \ge 2$, inequality (21) implies that the bias term satisfies $|\mathbb{E}(\nu_t)| \le \|\nu_t - \upsilon_t\|_{q/2} = O(\gamma_t)$. Beyond bounding the estimation bias, Lemma 8 also quantifies the linear approximation error for both the SGD iterate ω_t° and the Polyak-Ruppert averaging estimator $\bar{\omega}_t^\circ$. This property is particularly useful for establishing the corresponding asymptotic distributions. Specifically, combining Lemma 8 with Lemma 7 yields

$$\|(\omega_t^\circ - \omega^*) - \upsilon_t\|_p^2 = O(t^{-q\beta/p}),$$
$$\|(\bar{\omega}_t^\circ - \omega^*) - \bar{\upsilon}_t\|_p^2 = O(t^{-q\beta/p}), \tag{22}$$

where $\bar{\upsilon}_t = (\upsilon_1 + \cdots + \upsilon_t)/t$. In view of (20), both υ_t and $\bar{\upsilon}_t$ are linear combination of i.i.d. centered random vectors $\varepsilon_1, \ldots, \varepsilon_t$. Consequently, to characterize the asymptotic behavior of $\omega_t^\circ - \omega^*$ and $\bar{\omega}_t^\circ - \omega^*$, it suffices to derive the asymptotic distributions of υ_t and $\bar{\upsilon}_t$, under the conditions $q > p$ and $q\beta > p$, respectively. □

A. Concentration of the SGD iterates

For any non-random vector $\alpha \in \mathbb{S}^d$ and $z > 0$, combining the q-th moment bound established in Theorem 9 with the Markov inequality yields the following tail probability bound

$$\mathbb{P}(|\alpha^\top(\omega_t^\circ - \omega^*)| > z) \le \frac{\mathcal{C} M_q^q t^{-q\beta/2}}{z^q}. \tag{23}$$

However, the order $-q\beta/2$ in the polynomial term $\mathcal{C} M_q^q t^{-q\beta/2} z^{-q}$ is not tight when $q > 2$, resulting in a less sharp analysis of the sample complexity. In the following theorem, we develop a tighter Nagaev-type inequality for the SGD iterates by leveraging the functional dependence measures in Lemma 2 and the nonlinear autoregressive coupling introduced above.

Theorem 9. *Under the conditions of Theorem 4, for any vector $\alpha \in \mathbb{S}^d$ and any $z > 0$, we have*

$$\mathbb{P}(|\alpha^\top(\omega_t^\circ - \omega^*)| > z) \le \frac{\mathcal{C} M_q^q t^{(1-q)\beta}}{z^q} + 2\exp\left(-\frac{\mathcal{C}_0 z^2 t^\beta}{M_2^2}\right), \tag{24}$$

where $\mathcal{C}$ and $\mathcal{C}_0$ are positive constants independent of t and α.

Remark 8. It is worth mentioning that Theorem 9 is valid for arbitrary $\beta \in (0, 1)$. In particular, since $(1-q)\beta < -q\beta/2$ for any constant $q > 2$ and $\beta > 0$, the polynomial term $\mathcal{C} M_q^q t^{(1-q)\beta} z^{-q}$ is tighter than the corresponding term in (23). In fact, following the arguments in Section 4 of [30], it can be shown that this polynomial order $\mathcal{C} M_q^q t^{(1-q)\beta} z^{-q}$ cannot be improved when the stochastic gradient noise $\nabla f(\omega^*, X)$

only exhibits a finite q-th moment. The additional exponential term in (24) provides a sub-Gaussian-type concentration result. Thus, Theorem 9 reveals an interesting dichotomous phenomenon: for small z, the sub-Gaussian tail dominates, while for large z, the polynomial term $\mathcal{C}M_q^q t^{(1-q)\beta} z^{-q}$ dominates. The transition point between these two regimes satisfies $z \sim c_0 t^{-\beta/2}\sqrt{\log t}$ as $t \to \infty$, where $c_0 = M_2\sqrt{\beta(q/2-1)/\mathcal{C}_0}$. When the step size is constant, that is, $\gamma_t \equiv \gamma_1$ corresponding to $\beta = 0$, a tail bound of a similar type was recently obtained in [65]. That result addresses the constant step size regime and complements (24) in Theorem 9. □

B. Concentration of the Polyak-Ruppert averaging estimators

Theorem 10. *Assume that $\beta \in (1/2, 1)$. Then, under the conditions of Theorem 5, for any vector $\alpha \in \mathbb{S}^d$ and any $z > 0$, we have*

$$\mathbb{P}(|\alpha^\top(\bar{\omega}_t^\circ - \omega^*)| > z) \le \frac{\mathcal{C}M_q^q t}{(tz)^q} + 2\exp\left(-\frac{\mathcal{C}_0 z^2 t}{M_2^2}\right), \quad (25)$$

where $\mathcal{C}$ and $\mathcal{C}_0$ are positive constants independent of t and α.

Remark 9. Theorem 10 provides a Nagaev-type inequality for the Polyak-Ruppert averaging estimator. It aligns with the result for the special case of online linear regression in [30], which employs the mini-batch SGD algorithm. Similar to the discussions in Section III-A, the polynomial term $\mathcal{C}M_q^q t^{1-q} z^{-q}$ in (25) is significantly tighter than the tail probability bound obtained by applying Markov's inequality with the upper bound on the q-th moment $\|\bar{\omega}_t - \omega^*\|_q$ in (13). Moreover, as illustrated in [30], this polynomial order is necessary and unimprovable. □

IV. Infinite variance

A prominent characteristic of large-scale streaming data is the presence of outliers and heavy-tailed distributions. Recently, there has been growing interest in the convergence properties and distribution theory of the SGD algorithm and its variants under heavy-tailed settings ([69, 70, 62, 71, 72, 73]). To advance the current understanding of stochastic dynamics in online stochastic approximation algorithms operating in heavy-tailed environments, in this section, we relax the moment condition by allowing the gradient noise $\nabla f(\omega^*, X)$ to have only a finite q-th moment for some constant $q \in (1, 2)$ and derive new concentration inequalities for both moments and tail probabilities, applied to both the SGD iterates and the Polyak-Ruppert averaging estimators.

Similar to Lemma 3, the non-asymptotic q-th moment bounds that characterize the effect of initial estimate ω_0 remain valid for $q \in (1, 2)$. Consequently, it suffices to focus on the coupled process $(\omega_t^\circ)_{t\ge 0}$. In the following theorem, we establish non-asymptotic concentration inequalities for both ω_t° and $\bar{\omega}_t^\circ$, quantified in terms of their q-th moments.

Theorem 11. *Under Assumptions 1 and 2, for all $t \ge 1$, we have*

$$\|\omega_t^\circ - \omega^*\|_q \le \mathcal{C}_0 M_q \gamma_t^{1-1/q}. \quad (26)$$

Furthermore, suppose that Assumption 3 holds. Then, we have

$$\|\bar{\omega}_t^\circ - \omega^*\|_q \le \mathcal{C}_1 M_q t^{(1-q)/q} + \mathcal{C}_2 M_q^q t^{(1-q)\beta}. \quad (27)$$

Here, $\mathcal{C}_0$, $\mathcal{C}_1$, and $\mathcal{C}_2$ are positive constants independent of t.

Remark 10. When combined with Markov's inequality, Theorem 11 provides upper bounds on the tail probabilities of both ω_t° and $\bar{\omega}_t^\circ$, which in turn directly imply the following high-probability bounds

$$|\omega_t^\circ - \omega^*| = O(t^{\beta/q-\beta}),$$
$$|\bar{\omega}_t^\circ - \omega^*| = O(t^{(1-q)\min\{1/q,\beta\}}).$$

Notably, the last term $\mathcal{C}_2 M_q^q t^{(1-q)\beta}$ in (27) quantifies the estimation bias $|\mathbb{E}(\bar{\omega}_t^\circ) - \omega^*|$ arising from the nonlinearity of the population gradient $\nabla F(\cdot)$. To ensure valid statistical inference based on $\bar{\omega}_t^\circ$, Theorem 11 suggests selecting step sizes with $\beta > 1/q$, so that the estimation bias becomes asymptotically negligible. Compared to the finite variance setting discussed in Section II-B, smaller step sizes are required to study the asymptotic distribution of $\bar{\omega}_t^\circ$ when $q \in (1, 2)$. □

V. Simulation study

In this section, we conduct a finite-sample simulation study to illustrate the dichotomous phenomenon revealed by the Nagaev-type inequalities presented in Theorems 9 and 10. We consider the model $X = \omega^* + \varepsilon$, where $\omega^* = \mathbb{E}(X)$ and ε denotes the random noise. The parameter ω^* is estimated using SGD with step sizes $\gamma_s = 0.5 \times s^{-0.505}$ and initial estimate $\omega_0 = \omega^*$. The algorithm is run for $t = 5000$ iterations, with updates given by

$$\omega_s = \omega_{s-1} - \gamma_s(\omega_{s-1} - X_s), \quad s = 1, \ldots, t.$$

Our primary objective is to identify the two distinct decay regimes of the tail probabilities as functions of $z > 0$. According to Theorem 9, when z is small, the sub-Gaussian-type term $2\exp(-\mathcal{C}_0 z^2 t^\beta / M_2^2)$ in (24) dominates, whereas for large z, the polynomial-type term $\mathcal{C}M_q^q t^{(1-q)\beta} z^{-q}$ becomes dominant. An analogous dichotomy for the Polyak-Ruppert averaging estimator follows from (25) in Theorem 10.

To facilitate visualization, we focus on the transformed tail probabilities

$$P_z^* = \sqrt{-\log \mathbb{P}(|\omega_t - \omega^*| > z)},$$
$$\bar{P}_z^* = \sqrt{-\log \mathbb{P}(|\bar{\omega}_t - \omega^*| > z)}.$$

Under this transformation, sub-Gaussian-type tails correspond to a linear scaling of P_z^* with respect to z, while polynomial-type tails yield a dominant growth rate $\sqrt{q \log z}$ for large z. A similar pattern is expected for $\bar{P}_z^*$ in accordance with Theorem 10. We note that both P_z^* and $\bar{P}_z^*$ depend on t, which is suppressed for notational simplicity.

Without loss of generality, we set $\omega^* = 0$. For the random noise ε, we consider scaled Student's t-distributions, i.e., $\varepsilon \sim t_k / \sqrt{\frac{k}{k-2}}$, with degrees of freedom $k \in \{2.5, 3, 3.5\}$. The scaling ensures that $M_2^2 = \mathrm{Var}(\varepsilon) = 1$ for all considered values of k. The empirical transformed tail probabilities are computed based on $N = 10^7$ independent replications.

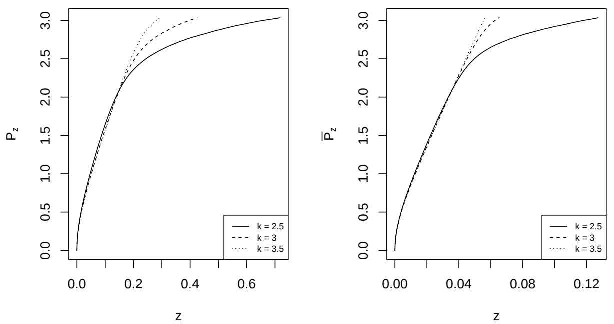

Fig. 1. Empirical transformed tail probabilities, computed via (28), for the SGD iterates and the Polyak-Ruppert averaging estimators.

Specifically, let $\omega_{t,\ell}$ and $\bar{\omega}_{t,\ell}$, $\ell = 1, \ldots, N$, denote the SGD iterate and the Polyak-Ruppert averaging estimator obtained from the ℓ-th replication, respectively. For any $z > 0$, the empirical counterparts of P_z^* and $\bar{P}_z^*$ are computed as

$$P_z = \sqrt{-\log\left(\frac{1}{N}\sum_{\ell=1}^{N}\mathbb{I}\{|\omega_{t,\ell} - \omega^*| > z\}\right)},$$
$$\bar{P}_z = \sqrt{-\log\left(\frac{1}{N}\sum_{\ell=1}^{N}\mathbb{I}\{|\bar{\omega}_{t,\ell} - \omega^*| > z\}\right)}. \quad (28)$$

Figure 1 summarizes the empirical transformed tail probabilities P_z and $\bar{P}_z$ over a range of values of z. When z is small, both P_z and $\bar{P}_z$ exhibit an approximately linear dependence on z, consistent with the dominance of the sub-Gaussian-type tails. In particular, these linear curves are nearly identical across different values of $k \in \{2.5, 3, 3.5\}$, indicating the tightness of the sub-Gaussian-type tails. As z increases, the linear trend disappears and discrepancies across different values of k emerge, as $\sqrt{q \log z}$ becomes the leading term. Overall, Figure 1 aligns well with the established theory, clearly illustrating the dichotomous phenomenon for both the tail probabilities of SGD iterates and the Polyak–Ruppert averaging estimators.

VI. Proofs

In this section, we provide the detailed proofs of all theoretical results stated in the main text.

A. Lemmas

The following lemma, adapted from Lemma 1 in [64], provides a Burkholder-type inequality for the infinite-variance case $q \in (1, 2)$.

Lemma 12. *Let $\xi \in \mathbb{R}^d$ and $\zeta \in \mathbb{R}^d$ be two random vectors such that $\mathbb{E}|\xi|^q < \infty$ and $\mathbb{E}|\zeta|^q < \infty$ for some constant $q \in (1, 2)$. Then, we have*

$$\mathbb{E}|\xi + \zeta|^q \le \mathbb{E}|\xi|^q + q\mathbb{E}\left\{|\xi|^{q-2}\xi^\top \mathbb{E}(\zeta|\xi)\right\} + 2^{2-q}\mathbb{E}|\zeta|^q.$$

Lemma 13 below presents the Burkholder inequality ([74, 63, 64]) for the q-th moment of a martingale, valid for $q > 1$.

Lemma 13. *Let $(D_t)_{t\ge 1}$ be a sequence of martingale differences and let $\mathcal{D}_n = \sum_{t=1}^n D_t$. Assume that $\max_{t\ge1}\mathbb{E}|D_t|^q < \infty$ for some constant $q > 1$. Then, we have*

$$\|\mathcal{D}_n\|_q^{\min\{q,2\}} \le C_q \sum_{t=1}^n \|D_t\|_q^{\min\{q,2\}},$$

where

$$C_q = \begin{cases} (q-1), & q > 2, \\ 2^{2-q}, & 1 < q \le 2. \end{cases}$$

B. Proof of Proposition 1

Proof of Proposition 1. Let $\varphi(t) = \|\xi + t\zeta\|_q^2$, $0 \le t \le 1$. If $\varphi(t) = 0$ for some $t \in [0,1]$, then (6) trivially holds. Now we assume $\varphi(t) \ne 0$ for all $t \in [0,1]$. Elementary calculations lead to

$$\varphi'(t) = 2\|\xi + t\zeta\|_q^{2-q}\mathbb{E}\left\{|\xi + t\zeta|^{q-2}\left(t|\zeta|^2 + \xi^\top\zeta\right)\right\},$$

and

$$\begin{aligned}\varphi''(t) &\le 2(q-1)\|\xi + t\zeta\|_q^{2-q}\mathbb{E}\left(|\xi + t\zeta|^{q-2}|\zeta|^2\right)\\ &\le 2(q-1)\|\xi + t\zeta\|_q^{2-q}\|\xi + t\zeta\|_q^{q-2}\|\zeta\|_q^2\\ &= 2(q-1)\|\zeta\|_q^2,\end{aligned}$$

where the first inequality follows from Cauchy's inequality and the fact that $2-q<0$, and the second inequality follows from Hölder's inequality. Consequently, applying Taylor's theorem yields

$$\begin{aligned}&\|\xi + \zeta\|_q^2 = \varphi(1) = \varphi(0) + \varphi'(0) + \int_0^1 (1-t)\varphi''(t)dt\\ &\le \|\xi\|_q^2 + 2\|\xi\|_q^{2-q}\mathbb{E}\left(|\xi|^{q-2}\xi^\top\zeta\right) + 2(q-1)\|\zeta\|_q^2\int_0^1(1-t)dt\\ &= \|\xi\|_q^2 + 2\|\xi\|_q^{2-q}\mathbb{E}\left(|\xi|^{q-2}\xi^\top\zeta\right) + (q-1)\|\zeta\|_q^2.\end{aligned}$$

□

C. Proof of Theorem 4

Proof of Theorem 4. For simplicity of notation, we denote $\delta_t = \omega_t - \omega^*$. Observe that $\nabla F(\omega^*) = 0$ and

$$\mathbb{E}\{\nabla f(\omega_{t-1}, X_t)|\mathcal{F}_{t-1}\} = \nabla F(\omega_{t-1}).$$

Hence, by Assumption 2, we have

$$\begin{aligned}&\mathbb{E}\left\{|\delta_{t-1}|^{q-2}\delta_{t-1}^\top\nabla f(\omega_{t-1}, X_t)\right\}\\ =&\mathbb{E}\left\{|\delta_{t-1}|^{q-2}\delta_{t-1}^\top(\nabla F(\omega_{t-1}) - \nabla F(\omega^*))\right\} \ge \mu\mathbb{E}|\delta_{t-1}|^q,\end{aligned}$$

and $\|\nabla f(\omega_{t-1}, X_t)\|_q^2 \le 2M_q^2 + 2L_q^2\|\delta_{t-1}\|_q^2$. Then, it follows from Proposition 1 that

$$\begin{aligned}\|\delta_t\|_q^2 &\le \|\delta_{t-1}\|_q^2 + (q-1)\gamma_t^2\|\nabla f(\omega_{t-1}, X_t)\|_q^2\\ &\quad - 2\gamma_t\|\delta_{t-1}\|_q^{2-q}\mathbb{E}\left\{|\delta_{t-1}|^{q-2}\delta_{t-1}^\top\nabla f(\omega_{t-1}, X_t)\right\}\\ &\le \|\delta_{t-1}\|_q^2 - 2\mu\gamma_t\|\delta_{t-1}\|_q^{2-q}\|\delta_{t-1}\|_q^q\\ &\quad + 2(q-1)\gamma_t^2L_q^2\|\delta_{t-1}\|_q^2 + 2(q-1)\gamma_t^2M_q^2\\ &= \{1 - 2\mu\gamma_t + 2(q-1)L_q^2\gamma_t^2\}\|\delta_{t-1}\|_q^2 + 2(q-1)\gamma_t^2M_q^2. \quad (29)\end{aligned}$$

Consequently, an argument analogous to the proof of Theorem 1 in [24] yields (10). □

D. Proof of Lemma 2

Proof of Lemma 2. By Theorem 4 and Assumption 2, for each $t \geq 1$, we have

$$\begin{aligned}\|\omega_t - \omega_{t,\{t\}}\|_q^2 &= \|\nabla f(\omega_{t-1}, X_t') - \nabla f(\omega_{t-1}, X_t)\|_q^2 \gamma_t^2 \\ &\leq 8L_q^2 \|\omega_{t-1} - \omega^*\|_q^2 \gamma_t^2 + 8M_q^2 \gamma_t^2.\end{aligned}$$

Then, for each $h \in \{1, \dots, t-1\}$, similar to (29), it follows by Proposition 1 that

$$\begin{aligned}\|\omega_t - \omega_{t,\{h\}}\|_q^2 \leq& \|\omega_h - \omega_{h,\{h\}}\|_q^2 \prod_{m=h+1}^{t} V_m \\ \leq& 8(L_q^2 \|\omega_{h-1} - \omega^*\|_q^2 + M_q^2)\gamma_h^2 \prod_{m=h+1}^{t} V_m.\end{aligned}$$

where $V_m = 1 - 2\mu\gamma_m + (q-1)L_q^2\gamma_m^2$ □

E. Proof of Theorem 5

Proof of Theorem 5. In view of (19), we have $\mathbb{E}(\upsilon_t) = 0$ for all $t \geq 0$. It then follows from Lemmas 7 and 8 that

$$\begin{aligned}|\mathbb{E}(\bar{\omega}_t^\circ - \omega^*)| &= |\mathbb{E}\{(\bar{\omega}_t^\circ - \omega^*) - \bar{\nu}_t\} + \mathbb{E}(\bar{\nu}_t - \bar{\upsilon}_t)| \\ &\leq \|(\bar{\omega}_t^\circ - \omega^*) - \bar{\nu}_t\|_2 + \|\bar{\nu}_t - \bar{\upsilon}_t\|_1 \\ &\leq \frac{\mathcal{C}\max\{1, M_2^2\}}{t^\beta}.\end{aligned} \tag{30}$$

Then it remains to upper bound $\|\mathbb{E}_0(\bar{\omega}_t^\circ)\|_q$. Recall that $V_m = 1 - 2\mu\gamma_m + (q-1)L_q^2\gamma_m^2$. By (11), Lemma 13 and the martingale decomposition in (12), we have

$$\begin{aligned}\|\mathbb{E}_0(\bar{\omega}_t^\circ)\|_q &\leq \frac{\mathcal{C}}{t}\sum_{h=0}^{t-1}\sqrt{\sum_{s=h+1}^{t}\|\omega_s^\circ - \omega_{s,\{s-h\}}^\circ\|_q^2} \\ &\leq \frac{\mathcal{C}M_q}{t}\sum_{h=0}^{t-1}\sqrt{\sum_{s=h+1}^{t}\gamma_{s-h}^2\prod_{m=s-h+1}^{s}V_m} \leq \frac{\mathcal{C}M_q}{t^{1/2}}.\end{aligned}$$

Combining this with (30) yields (13). □

F. Proof of Proposition 6

Proof of Proposition 6. By Lemma 7 and Lemma 8, we have

$$\begin{aligned}\|(\omega_t^\circ - \omega^*) - \upsilon_t\|_q &\leq \|(\omega_t^\circ - \omega^*) - \nu_t\|_q + \|\nu_t - \upsilon_t\|_q \\ &\leq \mathcal{C}_1 M_q \gamma_t + \mathcal{C}_2 M_{2q}^2 \gamma_t.\end{aligned}$$

Then, applying the triangle inequality yields

$$\|(\bar{\omega}_t^\circ - \omega^*) - \bar{\upsilon}_t\|_q = O(t^{-\beta}). \tag{31}$$

We next derive an upper bound for $\|\bar{\upsilon}_t\|_q$. By the definition of the process $(\upsilon_t)_{t\geq 1}$ in (19), for each $s \geq 1$,

$$\frac{\upsilon_{s-1} - \upsilon_s}{\gamma_s} = H\upsilon_{s-1} - \varepsilon_s.$$

Recall that $\varepsilon_s = -\nabla f(\omega^*, X_s)$ for each $s \in \mathbb{N}$. Hence, using the upper bound $\|\upsilon_s\|_q \leq \mathcal{C}M_q\gamma_s^{1/2}$, we obtain

$$\begin{aligned}&\Big\|\bar{\upsilon}_t + \frac{1}{t}\sum_{s=2}^{t+1} H^{-1}\nabla f(\omega^*, X_s)\Big\|_q \\ =&\Big\|\frac{1}{t}\sum_{s=2}^{t+1}\frac{H^{-1}\upsilon_{s-1} - H^{-1}\upsilon_s}{\gamma_s}\Big\|_q \\ \leq&\Big\|\frac{1}{t}\sum_{s=2}^{t} H^{-1}\upsilon_s\big(\frac{1}{\gamma_{s+1}} - \frac{1}{\gamma_s}\big) + \frac{H^{-1}\upsilon_1}{t\gamma_2} - \frac{H^{-1}\upsilon_{t+1}}{t\gamma_{t+1}}\Big\|_q \\ \leq&\frac{1}{t}\sum_{s=2}^{t}\|H^{-1}\upsilon_s\|_q\big(\frac{1}{\gamma_{s+1}} - \frac{1}{\gamma_s}\big) + \frac{\|H^{-1}\upsilon_1\|_q}{t\gamma_2} + \frac{\|H^{-1}\upsilon_{t+1}\|_q}{t\gamma_{t+1}} \\ =&O\big(t^{\beta/2-1}\big).\end{aligned}$$

Combining this with (31), we conclude that

$$\begin{aligned}&\|\bar{\omega}_t^\circ - \omega^*\|_q \\ \leq&\|(\bar{\omega}_t^\circ - \omega^*) - \bar{\upsilon}_t\|_q + \Big\|\bar{\upsilon}_t + \frac{1}{t}\sum_{s=2}^{t+1} H^{-1}\nabla f(\omega^*, X_s)\Big\|_q \\ &+ \Big\|\frac{1}{t}\sum_{s=2}^{t+1} H^{-1}\nabla f(\omega^*, X_s)\Big\|_q \\ \leq&\Big\|\frac{1}{t}\sum_{s=2}^{t+1} H^{-1}\nabla f(\omega^*, X_s)\Big\|_q + O\Big(t^{\max\{-\beta, \beta/2-1\}}\Big).\end{aligned}$$

□

G. Proof of Lemma 7

Proof of Lemma 7. For notational simplicity, we denote $\delta_t^\circ = \omega_t^\circ - \omega^*$ and $\bar{\delta}_t^\circ = (\delta_1^\circ + \cdots + \delta_t^\circ)/t$. By Proposition 1 and Theorem 4, we have $\|\delta_t^\circ\|_q^2 \leq \mathcal{C}M_q^2\gamma_t^2$ for all $t \geq 1$, and

$$\begin{aligned}&\|\delta_t^\circ - \nu_t\|_q^2 \\ \leq&\|\delta_{t-1}^\circ - \nu_{t-1} - \gamma_t\{\nabla F(\omega^* + \delta_{t-1}^\circ) - \nabla F(\omega^* + \nu_{t-1})\}\|_q^2 \\ &+ (q-1)\gamma_t^2\|\nabla F(\omega^* + \delta_{t-1}^\circ) - \nabla f(\omega^* + \delta_{t-1}^\circ, X_t) - \varepsilon_t\|_q^2 \\ \leq&\{1 - 2\mu\gamma_t + (q-1)L_1^2\gamma_t^2\}\|\delta_{t-1}^\circ - \nu_{t-1}\|_q^2 \\ &+ 4(q-1)\gamma_t^2 L_q^2\|\delta_{t-1}^\circ\|_q^2.\end{aligned}$$

Consequently, an argument analogous to the proof of Theorem 1 in [24] yields (15).

Now we bound $\mathbb{P}(|\alpha^\top\mathbb{E}_0(\bar{\delta}_t^\circ - \bar{\nu}_t)| > z)$. Note that $\psi_k \asymp k^{1/(1-\beta)}(\log t)^{1/(1-\beta)}$ for each $k \geq 1$ and $K \asymp t^{1-\beta}/(\log t)$. By the triangle inequality,

$$\begin{aligned}|t\mathbb{E}_0(\bar{\delta}_t^\circ - \bar{\nu}_t)| \leq&\Big|\sum_{k=1}^{K}\mathbb{E}_0(\widetilde{\mathcal{B}}_k)\Big| + \Big|\sum_{k=1}^{K}\mathbb{E}_0(\widetilde{\mathcal{B}}_k^\diamond)\Big| \\ &+ \Big|\sum_{k=1}^{K}(\mathcal{B}_k - \widetilde{\mathcal{B}}_k)\Big| + \Big|\sum_{k=1}^{K}(\mathcal{B}_k^\diamond - \widetilde{\mathcal{B}}_k^\diamond)\Big| \\ =:&\Upsilon_1 + \Upsilon_2 + \Upsilon_3 + \Upsilon_4.\end{aligned}$$

We begin with upper bounding $\mathbb{P}(\Upsilon_1 > z)$ and $\mathbb{P}(\Upsilon_2 > z)$. For any constant $q \geq 2$, by the triangle inequality and (15),

$$\begin{aligned}\|\mathbb{E}_0(\widetilde{\mathcal{B}}_k)\|_q + \|\mathbb{E}_0(\widetilde{\mathcal{B}}_k^\diamond)\|_q &\leq 2 \sum_{s=\psi_{k-1}+1}^{\psi_k} \|\delta_s^\circ - \nu_s\|_q \\ &\leq \mathcal{C} M_q \sum_{s=\psi_{k-1}+1}^{\psi_k} \gamma_s \leq \mathcal{C} M_q (\log t),\end{aligned}$$

which implies that

$$\begin{aligned}\sum_{k=1}^K \|\mathbb{E}_0(\widetilde{\mathcal{B}}_k)\|_q^q + \sum_{k=1}^K \|\mathbb{E}_0(\widetilde{\mathcal{B}}_k^\diamond)\|_q^q &\leq \mathcal{C} K M_q^q (\log t)^q \\ &\asymp M_q^q t^{1-\beta} (\log t)^{q-1}.\end{aligned}$$

Consequently, since $\widetilde{\mathcal{B}}_1, \dots, \widetilde{\mathcal{B}}_K$ are independent, it follows by Corollary 1.8 in [58] that

$$\begin{aligned}\mathbb{P}(\Upsilon_1 > z) \leq & \frac{\mathcal{C}_q M_q^q t^{1-\beta} (\log t)^{q-1}}{z^q} \\ & + 2 \exp\Big\{ - \frac{\mathcal{C}_q z^2}{M_2^2 t^{1-\beta} (\log t)} \Big\}.\end{aligned}$$

Similarly, the inequality above also holds for $\mathbb{P}(\Upsilon_2 > z)$. It remains to upper bound $\mathbb{P}(\Upsilon_3 > z)$ and $\mathbb{P}(\Upsilon_4 > z)$. Note that

$$\begin{aligned}\sum_{k=1}^K (\mathcal{B}_k - \widetilde{\mathcal{B}}_k) &= \sum_{k=1}^K \sum_{s=\psi_{k-1}+1}^{\psi_k} \sum_{h=1}^{s-\Delta_k-1} \mathcal{P}^h (\delta_s^\circ - \nu_s) \\ &=: \sum_{m=1}^K \sum_{h=1\vee(\psi_{m-1}-\Delta_{m-1})}^{\psi_m-\Delta_m-1} \Lambda(m,h),\end{aligned}$$

where

$$\Lambda(m,h) = \mathcal{P}^h \Big(\sum_{k=m}^K \sum_{s=(\psi_{k-1}+1)\vee(h+\Delta_k+1)}^{\psi_k} (\delta_s^\circ - \nu_s) \Big)$$

and $\mathcal{P}^h(\cdot) = \mathbb{E}(\cdot | X_h, X_{h+1}, \dots) - \mathbb{E}(\cdot | X_{h+1}, X_{h+2}, \dots)$ is the projection operator. By Lemma 2,

$$\begin{aligned}&\sum_{s=(\psi_{k-1}+1)\vee(h+\Delta_k+1)}^{\psi_k} \|\mathcal{P}^h(\delta_s^\circ - \nu_s)\|_q \\ \leq & \mathcal{C}_q M_q \Delta_k \exp\left\{ - \frac{\mathcal{C}\Delta_k}{(\Delta_k \wedge h)^\beta} \right\} \gamma_h \leq \mathcal{C}_q M_q t^{-C},\end{aligned}$$

where the last inequality follows from the fact that $\Delta_k^{1-\beta} \asymp k^\beta (\log t)$ and $\Delta_k / h^\beta \geq \Delta_k / \psi_k^\beta \geq \mathcal{C}(\log t)$ for $h \leq \psi_k - \Delta_k - 1 \leq \psi_k$. Consequently, since $\{\Lambda(m,h)\}_{h \geq 1}$ is a sequence of martingale differences, it follows by Lemma 13 and the triangle inequality that

$$\begin{aligned}&\Big\| \sum_{k=1}^K (\mathcal{B}_k - \widetilde{\mathcal{B}}_k) \Big\|_q^2 \\ \leq & \mathcal{C}_q \sum_{m=1}^K \sum_{h=1\vee(\psi_{m-1}-\Delta_{m-1})}^{\psi_m-\Delta_m-1} \|\Lambda(m,h)\|_q^2 \leq \mathcal{C}_q M_q^2 t^{-C}.\end{aligned}$$

The same upper bound holds for $\| \sum_{k=1}^K (\mathcal{B}_k^\diamond - \widetilde{\mathcal{B}}_k^\diamond) \|_q^2$. Then, by the Markov inequality, we obtain $\mathbb{P}(\Upsilon_3 > z) + \mathbb{P}(\Upsilon_4 > z) \leq 2(\mathcal{C}_q M_q^2 t^{-C})^{q/2} / z^q$. Combining all these results leads to (16). □

H. Proof of Lemma 8

Proof of Lemma 8. Recall that $A_t = I_d - \gamma_t \nabla^2 F(\omega^*) \in \mathbb{R}^{d \times d}$. By (18) and the bound $\|\nu_t\|_q^2 \leq \mathcal{C} M_q^2 \gamma_t$ for all $t \geq 1$, we obtain

$$\begin{aligned}\|\nu_t - \upsilon_t\|_p \leq & \|A_t(\nu_{t-1} - \upsilon_{t-1})\|_p \\ & + \gamma_t \|\nabla F(\omega^* + \nu_{t-1}) - \nabla^2 F(\omega^*) \nu_{t-1}\|_p \\ \leq & \|A_t\|_{\mathrm{op}} \|\nu_{t-1} - \upsilon_{t-1}\|_p + \mathcal{C}\gamma_t (2L_1 + \varrho)(M_q \gamma_t^{1/2})^{q/p},\end{aligned}$$

where $\|\cdot\|_{\mathrm{op}}$ stands for the operator norm. Observe that $\|A_t\|_{\mathrm{op}} \leq 1 - \gamma_t \lambda_{\min}\{\nabla^2 F(\omega^*)\}$ when $1 - \gamma_t \lambda_{\max}\{\nabla^2 F(\omega^*)\} \geq 0$, and $\|A_t\|_{\mathrm{op}} \leq 1 + \gamma_t \lambda_{\max}\{\nabla^2 F(\omega^*)\}$ otherwise. Consequently, elementary calculations yield (21). □

I. Proof of Theorem 9

Proof of Theorem 9. Recall that $\delta_t^\circ = \omega_t^\circ - \omega^*$. By Lemma 7 and the Markov inequality,

$$\mathbb{P}(|\delta_t^\circ - \nu_t| > z) \leq \frac{\|\delta_t^\circ - \nu_t\|_q^q}{z^q} \leq \frac{\mathcal{C} M_q^q t^{-q\beta}}{z^q}.$$

By Theorem 4 and Lemma 7, we have $|\mathbb{E}(\nu_t)| \leq \|\nu_t\|_2 \leq \mathcal{C}_1 M_2 t^{-\beta/2}$, where $\mathcal{C}_1$ is a positive constant independent of t. Hence, by choosing $\mathcal{C}_0 \leq (\log 2)/(4\mathcal{C}_1^2)$, (24) holds trivially for all $z \leq 2\mathcal{C}_1 M_2 t^{-\beta/2}$. Otherwise, for $z > 2\mathcal{C}_1 M_2 t^{-\beta/2}$, we have $\mathbb{P}(|\alpha^\top \nu_t| > z) \leq \mathbb{P}(|\mathbb{E}_0(\alpha^\top \nu_t)| > z/2)$. Therefore, it suffices to derive an upper bound for $\mathbb{P}(|\mathbb{E}_0(\alpha^\top \nu_t)| > z)$ for any $z > 0$. By a similar argument as the proof of Lemma 2, it follows that for each $h \in \{0, 1, \dots, t-1\}$,

$$\begin{aligned}|\mathcal{P}_{t-h}(\alpha^\top \nu_t)| \leq & \mathbb{E}(|\varepsilon_{t-h} - \varepsilon'_{t-h}| | \mathcal{F}_{t-h}) \gamma_{t-h} \\ & \cdot \prod_{m=t-h+1}^t \{1 - 2\mu\gamma_m + (q-1) L_q^2 \gamma_m^2\} \\ =: & \mathbb{E}(|\varepsilon_{t-h} - \varepsilon'_{t-h}| | \mathcal{F}_{t-h}) \times \mathcal{A}(t,h), \qquad (32)\end{aligned}$$

where $\varepsilon'_1, \dots, \varepsilon'_t \in \mathbb{R}^d$ are i.i.d. copies of ε_1. For any constant $q \geq 2$, elementary calculations yield

$$\sum_{h=0}^{t-1} \mathcal{A}^q(t,h) \leq \mathcal{C}_q t^{(1-q)\beta}. \qquad (33)$$

For any $z > 0$, let $y = qz/(q+2)$ and define another sequence $(\nu_s^\diamond)_{s=0}^t$ with $\nu_0^\diamond = 0$ and

$$\begin{aligned}\nu_s^\diamond = & \nu_{s-1}^\diamond - \gamma_s \nabla F(\omega^* + \nu_{s-1}^\diamond) \\ & + \gamma_s \varepsilon_s \times \min\left\{1, \frac{y}{2\mathcal{A}(t, t-s)|\varepsilon_s|}\right\}, \qquad (34)\end{aligned}$$

for $s = 1, \dots, t$. Then, we have

$$\begin{aligned}\mathbb{P}(|\mathbb{E}_0(\alpha^\top \nu_t)| > z) \leq & \sum_{s=1}^t \mathbb{P}\left\{ |\varepsilon_s| > \frac{y}{2\mathcal{A}(t, t-s)} \right\} \\ & + \mathbb{P}(|\mathbb{E}_0(\alpha^\top \nu_t^\diamond)| > z) \\ \leq & \frac{\mathcal{C}_q M_q^q t^{(1-q)\beta}}{z^q} + \mathbb{P}(|\mathbb{E}_0(\alpha^\top \nu_t^\diamond)| > z),\end{aligned}$$

and it suffices to establish an upper bound for $\mathbb{P}(|\mathbb{E}_0(\alpha^\top \nu_t^\diamond)| > z)$. To this end, we shall first upper bound the following moment generating function

$$M_t^\diamond(\phi) = \mathbb{E}\exp\{\phi\mathbb{E}_0(\alpha^\top \nu_t^\diamond)\} = \mathbb{E}\exp\Big(\phi\sum_{h=0}^{t-1}\mathcal{P}_{t-h}(\alpha^\top\nu_t^\diamond)\Big),$$

for $\phi > 0$. For each $h \in \{0,1,\dots,t-1\}$, we have $\sup_{|\alpha|=1}|\mathcal{P}_{t-h}(\alpha^\top\nu_t^\diamond)| \le y$ in view of the definition in (34). Consequently, by Lemma 1.4 in [58], we have

$$\begin{aligned}
&\mathbb{E}\{\exp(\phi\mathcal{P}_{t-h}(\alpha^\top\nu_t^\diamond))|\mathcal{F}_{t-h-1}\}\\
\le&1+\frac{\exp(q)\phi^2}{2}\mathbb{E}\{|\mathcal{P}_{t-h}(\alpha^\top\nu_t^\diamond)|^2|\mathcal{F}_{t-h-1}\}\\
&+\frac{\exp(\phi y)-1-\phi y}{y^q}\mathbb{E}\{|\mathcal{P}_{t-h}(\alpha^\top\nu_t^\diamond)|^q|\mathcal{F}_{t-h-1}\}\mathbb{I}\{\phi>q/y\}\\
\le&1+\exp(q)\phi^2\mathcal{A}^2(t,h)M_2^2\\
&+\frac{\exp(\phi y)-1-\phi y}{y^q}2^q\mathcal{A}^q(t,h)M_q^q\mathbb{I}\{\phi>q/y\}. \qquad (35)
\end{aligned}$$

Observe that the upper bound above is independent of $\mathcal{F}_{t-h-1}$. Hence, by (33),

$$\begin{aligned}
M_t^\diamond(\phi)\le&\prod_{h=0}^{t-1}\Big(1+\exp(q)\phi^2\mathcal{A}^2(t,h)M_2^2\\
&+\frac{\exp(\phi y)-1-\phi y}{y^q}2^q\mathcal{A}^q(t,h)M_q^q\times\mathbb{I}\{\phi>q/y\}\Big)\\
\le&\exp\big(\frac{\mathcal{C}_q\phi^2M_2^2}{t^\beta}+\mathcal{C}_q\frac{\exp(\phi y)-1-\phi y}{y^qt^{(q-1)\beta}}M_q^q\mathbb{I}\{\phi>q/y\}\big)(36)
\end{aligned}$$

Following the proof of Theorem 1.3 in [58], we obtain

$$\begin{aligned}
&\mathbb{P}(|\mathbb{E}_0(\alpha^\top\nu_t^\diamond)|>z)\\
\le&2\left(\frac{\mathcal{C}_qz^qM_q^q}{t^{(q-1)\beta}}+1\right)^{-1}+2\exp\left(-\frac{\mathcal{C}_qz^2t^\beta}{M_2^2}\right)\\
\le&\frac{\mathcal{C}_qM_q^qt^{(1-q)\beta}}{z^q}+2\exp\left(-\frac{\mathcal{C}_qz^2t^\beta}{M_2^2}\right).
\end{aligned}$$

□

J. Proof of Theorem 10

Proof of Theorem 10. By Lemma (3) and the Markov inequality, for any $z>0$, we have

$$\mathbb{P}(|\bar\omega_t-\bar\omega_t^\circ|>z)\le\frac{C_q|\delta_0|^q}{(tz)^q}.$$

Since $|\mathbb{E}(\bar\omega_t^\circ-\omega^*)|\le\mathcal{C}_2\max\{M_2^2,1\}t^{-\beta}$ and $\beta>1/2$, it suffices to upper bound $\mathbb{P}(|\mathbb{E}_0(\alpha^\top\bar\nu_t)|>z)$ for all $z>0$ in view of (16). Note that

$$\sum_{s=1}^t\mathbb{E}_0(\alpha^\top\nu_s)=\sum_{h=1}^t\mathcal{P}_h\left(\sum_{s=h}^t\alpha^\top\nu_s\right)=:\sum_{h=1}^t\chi_h.$$

Here $(\chi_h)_{h\ge1}$ is a sequence of martingale differences adapted to the filtration $(\mathcal{F}_h)_{h\ge1}$ and

$$\begin{aligned}
|\chi_h|\le&\sum_{s=h}^t\mathcal{A}(s,s-h)\mathbb{E}(|\varepsilon_h-\varepsilon_h'||\mathcal{F}_h)\\
=:&\widetilde{\mathcal{A}}(t,h)\mathbb{E}(|\varepsilon_h-\varepsilon_h'||\mathcal{F}_h).
\end{aligned}$$

For any constant $q\ge2$, elementary calculations lead to $\sum_{h=1}^t\widetilde{\mathcal{A}}^q(t,h)\lesssim t$. Recall that $y=qz/(q+2)$. Similar to (34), we define the sequence $(\nu_s^\circ)_{s=0}^t$ with $\nu_0^\circ=0$ and

$$\begin{aligned}
\nu_s^\circ=&\nu_{s-1}^\circ-\gamma_s\nabla F(\omega^*+\nu_{s-1}^\diamond)\\
&+\gamma_s\varepsilon_s\times\min\left\{1,\frac{y}{2\widetilde{\mathcal{A}}(t,t-s)|\varepsilon_s|}\right\},\quad s=1,\dots,t,
\end{aligned}$$

and let $\chi_h^\circ=\mathcal{P}_h(\sum_{s=h}^t\alpha^\top\nu_s^\circ)$ for $h\in\{1,\dots,t\}$. Similar to (35) and (36), we have

$$\begin{aligned}
&\mathbb{E}\exp\left(\phi\sum_{h=1}^t\chi_h^\circ\right)\\
\le&\exp\big(\mathcal{C}_qt\phi^2M_2^2+\mathcal{C}_q\frac{\exp(\phi y)-1-\phi y}{y^q}tM_q^q\times\mathbb{I}\{\phi>q/y\}\big).
\end{aligned}$$

Consequently, we obtain

$$\begin{aligned}
&\mathbb{P}\left(\left|\sum_{h=1}^t\chi_h\right|>z\right)\\
\le&\sum_{s=1}^t\mathbb{P}\left\{|\varepsilon_s|>\frac{y}{2\widetilde{\mathcal{A}}(t,t-s)}\right\}+\mathbb{P}\left(\left|\sum_{h=1}^t\chi_h^\circ\right|>z\right)\\
\le&\frac{\sum_{s=1}^t\widetilde{\mathcal{A}}^q(t,t-s)M_q^q}{(y/2)^q}+\frac{\mathcal{C}_qM_q^qt}{z^q}+2\exp\left(-\frac{\mathcal{C}_qz^2}{M_2^2t}\right)\\
\le&\frac{\mathcal{C}_qM_q^qt}{z^q}+2\exp\left(-\frac{\mathcal{C}_qz^2}{M_2^2t}\right).
\end{aligned}$$

□

K. Proof of Theorem 11

Proof of Theorem 11. By Lemma 12, under Assumption 1 and Assumption 2, it follows that

$$\begin{aligned}
&\mathbb{E}|\omega_t^\circ-\omega^*|^q\\
\le&(1-q\mu\gamma_t)\mathbb{E}|\omega_{t-1}^\circ-\omega^*|^q+2^{2-q}\gamma_t^q\mathbb{E}|\nabla f(\omega_{t-1}^\circ,X_t)|^q\\
\le&(1-q\mu\gamma_t+2L_q^q\gamma_t^q)\mathbb{E}|\omega_{t-1}^\circ-\omega^*|^q+2M_q^q\gamma_t^q. \qquad (37)
\end{aligned}$$

Consequently, an argument analogous to the proof of Theorem 1 in [24] yields (26).

To establish (27), we first bound $\|\bar\omega_t^\circ-\omega^*\|_q$. By (26), Lemma 12, and an argument analogous to that for (37), it follows that, for all $t\ge1$ and $h\in\{1,\dots,t\}$,

$$\mathbb{E}|\omega_t^\circ-\omega_{t,\{h\}}^\circ|^q\le\mathcal{C}M_q^q\gamma_h^q\prod_{m=h+1}^t(1-q\mu\gamma_m+2^{2-q}L_q^q\gamma_m^q).$$

Then, by Lemma 13 and the martingale decomposition of $\bar\omega_t^\circ$ in (12), we have

$$\begin{aligned}
\|\mathbb{E}_0(\bar\omega_t^\circ)\|_q\le&\frac{\mathcal{C}}{t}\sum_{h=0}^{t-1}\Big(\sum_{s=h+1}^t\mathbb{E}|\omega_s^\circ-\omega_{s,\{s-h\}}^\circ|^q\Big)^{1/q}\\
\le&\mathcal{C}M_qt^{(1-q)/q}. \qquad (38)
\end{aligned}$$

We now bound the estimation bias $|\mathbb{E}(\bar{\omega}_t^\circ)-\omega^*|$. Recall that $\delta_t^\circ = \omega_t - \omega^*$. By Lemma 12 and an argument analogous to that for (37), we obtain

$$\begin{aligned}
&\mathbb{E}|\delta_t^\circ - \nu_t|^q \\
\le&\mathbb{E}|\delta_{t-1}^\circ - \nu_{t-1} - \gamma_t\{\nabla F(\omega^* + \delta_{t-1}^\circ) - \nabla F(\omega^* + \nu_{t-1})\}|^q \\
&+ 2^{2-q}\gamma_t^q \mathbb{E}|\nabla F(\omega^* + \delta_{t-1}^\circ) - \nabla f(\omega^* + \delta_{t-1}^\circ, X_t) - \varepsilon_t|^q \\
\le&(1 - q\mu\gamma_t + 2^{2-q}L_1^q\gamma_t^q)\mathbb{E}|\delta_{t-1}^\circ - \nu_{t-1}|^q + 4\gamma_t^q L_q^q \mathbb{E}|\delta_{t-1}^\circ|^q.
\end{aligned}$$

Combining this with (26) yields $\mathbb{E}|\delta_t^\circ - \nu_t|^q \le \mathcal{C}M_q^q\gamma_t^{2(q-1)}$ for all $t \ge 1$. Since $q \in (1,2)$, setting $p = 1$ in (18) yields

$$\mathbb{E}|\nabla F(\omega^* + \nu_{t-1}) - \nabla^2 F(\omega^*)\nu_{t-1}| \le (2L_1 + \varrho)\mathbb{E}|\nu_{t-1}|^q.$$

Combined with the triangle inequality, this implies that

$$\begin{aligned}
\mathbb{E}|\nu_t - \upsilon_t| &\le \mathbb{E}|A_t(\nu_{t-1} - \upsilon_{t-1})| + \gamma_t\mathbb{E}|\nabla F(\omega^* + \nu_{t-1}) \\
&\quad - \nabla^2 F(\omega^*)\nu_{t-1}| \\
&\le \|A_t\|_{\mathrm{op}}\mathbb{E}|\nu_{t-1} - \upsilon_{t-1}| + \gamma_t(2L_1 + \varrho)\mathbb{E}|\nu_{t-1}|^q \\
&\le \|A_t\|_{\mathrm{op}}\mathbb{E}|\nu_{t-1} - \upsilon_{t-1}| + \mathcal{C}(2L_1 + \varrho)M_q^q\gamma_t^q.
\end{aligned}$$

Then we have $\mathbb{E}|\nu_t - \upsilon_t| \le \mathcal{C}M_q^q\gamma_t^{q-1}$ for all $t \ge 1$. Consequently, for $q \in (1,2)$, the triangle inequality implies

$$\begin{aligned}
|\mathbb{E}(\bar{\omega}_t^\circ) - \omega^*| &\le \|(\bar{\omega}_t^\circ - \omega^*) - \bar{\nu}_t\|_q + \|\bar{\nu}_t - \bar{\upsilon}_t\|_1 \\
&\le \mathcal{C}M_q^q t^{(1-q)\beta},
\end{aligned}$$

and combining this with (38) yields (27). □

VII. Conclusion and discussion

In this paper, we introduce a comprehensive framework for analyzing online stochastic approximation algorithms by establishing connections with nonlinear causal time series analysis. By quantifying the associated functional dependence measures, we establish tight moment and tail probability inequalities under the assumption that the gradient noise exhibits only finite q-th moment ($q > 1$). Our proposed methods may serve as valuable tools for further theoretical studies of these algorithms, including weak convergence and the strong invariance principle.

Acknowledgement

The authors thank the Associate Editor and two anonymous referees for their many constructive comments.

References

[1] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, pages 400–407, 1951.

[2] Jack Kiefer and Jacob Wolfowitz. Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics*, pages 462–466, 1952.

[3] Boris T Polyak. Some methods of speeding up the convergence of iteration methods. *Ussr computational mathematics and mathematical physics*, 4(5):1–17, 1964.

[4] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

[5] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. *Advances in neural information processing systems*, 26, 2013.

[6] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

[7] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[8] Lin Xiao and Tong Zhang. A proximal stochastic gradient method with progressive variance reduction. *SIAM Journal on Optimization*, 24(4):2057–2075, 2014.

[9] Kobi Cohen, Angelia Nedić, and Rayadurgam Srikant. On projected stochastic gradient descent algorithm with weighted averaging for least squares regression. *IEEE Transactions on Automatic Control*, 62(11):5974–5981, 2017.

[10] Shi Pu, Alex Olshevsky, and Ioannis Ch Paschalidis. A sharp estimate on the transient time of distributed stochastic gradient descent. *IEEE Transactions on Automatic Control*, 67(11):5900–5915, 2021.

[11] David Ruppert. Efficient estimations from a slowly convergent robbins-monro process. Technical report, Cornell University Operations Research and Industrial Engineering, 1988.

[12] Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, 30(4):838–855, 1992.

[13] Julius R Blum. Approximation methods which converge with probability one. *The Annals of Mathematical Statistics*, pages 382–386, 1954.

[14] Kai Lai Chung. On a stochastic approximation method. *The Annals of Mathematical Statistics*, pages 463–483, 1954.

[15] Donald L Burkholder. On a class of stochastic approximation processes. *The Annals of Mathematical Statistics*, pages 1044–1059, 1956.

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

[17] V. Fabian. On asymptotic normality in stochastic approximation. *Ann. Math. Stat.*, 39(4):1327–1332, 1968.

[18] M. Woodroofe. Normal approximation and large deviations for the robbins-monro process. *Z. Wahrscheinlichkeitstheorie verw Gebiete*, 21(4):329–338, 1972.

[19] Alexander Shapiro. Asymptotic properties of statistical estimators in stochastic programming. *Ann. Statist.*, 17(2):841–858, 1989.

[20] Dirk TM Slock. On the convergence behavior of the lms and the normalized lms algorithms. *IEEE Transactions on Signal processing*, 41(9):2811–2825, 1993.

[21] Lei Guo, Lennart Ljung, and Guan-Jun Wang. Necessary and sufficient conditions for stability of lms. *IEEE Transactions on Automatic Control*, 42(6):761–770, 1997.

[22] Tze Leung Lai. Stochastic approximation. *The annals of Statistics*, 31(2):391–406, 2003.

[23] Lennart Ljung. Analysis of recursive stochastic algorithms. *IEEE transactions on automatic control*, 22(4):551–575, 2003.

[24] Eric Moulines and Francis Bach. Non-Asymptotic Analysis of Stochastic Approximation Algorithms for Machine Learning. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems*, pages 856–864, 2011.

[25] Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM journal on optimization*, 23(4):2341–2368, 2013.

[26] Prateek Jain, Dheeraj Nagaraj, and Praneeth Netrapalli. Making the last iterate of sgd information theoretically optimal. In *Conference on Learning Theory*, pages 1752–1755. PMLR, 2019.

[27] Wenlong Mou, Chris Junchi Li, Martin J Wainwright, Peter L Bartlett, and Michael I Jordan. On linear stochastic approximation: Fine-grained polyak-ruppert and non-asymptotic concentration. In *Conference on Learning Theory*, pages 2947–2997. PMLR, 2020.

[28] Arnulf Jentzen, Benno Kuckuck, Ariel Neufeld, and Philippe von Wurstemberger. Strong error analysis for stochastic gradient descent optimization algorithms. *IMA Journal of Numerical Analysis*, 41(1):455–492, 2021.

[29] Jun Liu and Ye Yuan. On almost sure convergence rates of stochastic gradient methods. In *Conference on Learning Theory*, pages 2963–2983. PMLR, 2022.

[30] Wanrong Zhu, Zhipeng Lou, and Wei Biao Wu. Beyond sub-gaussian noises: Sharp concentration analysis for stochastic gradient descent. *Journal of Machine Learning Research*, 23(46):1–22, 2022.

[31] Wanrong Zhu, Xi Chen, and Wei Biao Wu. Online covariance matrix estimation in stochastic gradient descent. *Journal of the American Statistical Association*, 118(541):393–404, 2023.

[32] Wei Biao Wu. Nonlinear system theory: another look at dependence. *Proc. Natl. Acad. Sci. USA*, 102(40):14150–14154, 2005.

[33] Wei Biao Wu. Strong invariance principles for dependent random variables. *The Annals of Probability*, pages 2294–2320, 2007.

[34] Wei Biao Wu. Asymptotic theory for stationary processes. *Statistics and its Interface*, 4(2):207–226, 2011.

[35] Han Xiao and Wei Biao Wu. A single-pass algorithm for spectrum estimation with fast convergence. *IEEE transactions on information theory*, 57(7):4720–4731, 2011.

[36] Han Xiao and Wei Biao Wu. Covariance matrix estimation for stationary time series. *The Annals of Statistics*, 40(1):466–493, 2012.

[37] Xiaohui Chen, Mengyu Xu, and Wei Biao Wu. Covariance and precision matrix estimation for high-dimensional time series. *The Annals of Statistics*, 41(6):2994–3021, 2013.

[38] Yinxiao Huang, Xiaohong Chen, and Wei Biao Wu. Recursive nonparametric estimation for time series. *IEEE Transactions on Information Theory*, 60(2):1301–1312, 2013.

[39] Zhou Zhou and Xiaofeng Shao. Inference for linear models with dependent errors. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 75(2):323–343, 2013.

[40] István Berkes, Weidong Liu, and Wei Biao Wu. Komlós-Major-Tusnády approximation under dependence. *Ann. Probab.*, 42(2):794–817, 2014.

[41] Xiaohui Chen, Mengyu Xu, and Wei Biao Wu. Regularized estimation of linear functionals of precision matrices for high-dimensional time series. *IEEE Transactions on Signal Processing*, 64(24):6459–6470, 2016.

[42] Wei-Biao Wu and Ying Nian Wu. Performance bounds for parameter estimates of high-dimensional linear models with correlated errors. *Electronic Journal of Statistics*, 10:352–379, 2016.

[43] Danna Zhang and Wei Biao Wu. Gaussian approximation for high dimensional time series. *The Annals of Statistics*, 45(5):1895–1919, 2017.

[44] Danna Zhang and Wei Biao Wu. Asymptotic theory for estimators of high-order statistics of stationary processes. *IEEE Transactions on Information Theory*, 64(7):4907–4922, 2017.

[45] Danna Zhang and Wei Biao Wu. Convergence of covariance and spectral density estimates for high-dimensional locally stationary processes. *The Annals of Statistics*, 49(1):233–254, 2021.

[46] Chi Zhang and Danna Zhang. Spectral inference for high dimensional time series. *IEEE Transactions on Information Theory*, 2025.

[47] Sébastien Gadat and Fabien Panloup. Optimal non-asymptotic analysis of the ruppert–polyak averaging stochastic algorithm. *Stochastic Processes and their Applications*, 156:312–348, 2023.

[48] Wanrong Zhu, Zhipeng Lou, Ziyang Wei, and Wei Biao Wu. High confidence level inference is almost free using parallel stochastic optimization. *arXiv preprint arXiv:2401.09346*, 2024.

[49] Leslie G Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

[50] Vladimir Vapnik. *Statistical learning theory*. John Wiley & Sons, 1998.

[51] Vladimir Vapnik. *The nature of statistical learning theory*. Springer, 2000.

[52] Pascal Massart. *Concentration inequalities and model selection*. Springer, 2007.

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

[54] Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. In *29th International Conference on Machine Learning, ICML 2012*, pages 449–456, 2012.

[55] Hervé Cardot, Peggy Cénac, and Antoine Godichon-Baggioni. Online estimation of the geometric median in hilbert spaces: Nonasymptotic confidence balls. *The Annals of Statistics*, 45(2):591–614, 2017.

[56] Nicholas JA Harvey, Christopher Liaw, Yaniv Plan, and Sikander Randhawa. Tight analyses for non-smooth stochastic gradient descent. In *Conference on Learning Theory*, pages 1579–1613. PMLR, 2019.

[57] Likai Chen, Georg Keilbar, and Wei Biao Wu. Recursive quantile estimation: Non-asymptotic confidence bounds. *Journal of Machine Learning Research*, 24(91):1–25, 2023.

[58] Sergey V Nagaev. Large deviations of sums of independent random variables. *The Annals of Probability*, pages 745–789, 1979.

[59] Weidong Liu, Han Xiao, and Wei Biao Wu. Probability and moment inequalities under dependence. *Statistica Sinica*, pages 1257–1272, 2013.

[60] Guanghui Lan. *First-order and stochastic optimization methods for machine learning*, volume 1. Springer, 2020.

[61] Lutz Dümbgen, Sara A Van De Geer, Mark C Veraar, and Jon A Wellner. Nemirovski's inequalities revisited. *The American Mathematical Monthly*, 117(2):138–160, 2010.

[62] Nuri Mert Vural, Lu Yu, Krishna Balasubramanian, Stanislav Volgushev, and Murat A Erdogdu. Mirror descent strikes again: Optimal stochastic convex optimization under infinite noise variance. In *Conference on Learning Theory*, pages 65–102. PMLR, 2022.

[63] Emmanuel Rio. Moment inequalities for sums of dependent random variables under projective conditions. *Journal of Theoretical Probability*, 22(1):146–163, 2009.

[64] Jiaqi Li, Zhipeng Lou, Stefan Richter, and Wei Biao Wu. Stochastic gradient descent: a nonlinear time series perspective. *Manuscript*, 2024.

[65] Ziyang Wei, Jiaqi Li, Zhipeng Lou, and Wei Biao Wu. Gaussian approximation and concentration of constant learning-rate stochastic gradient descent. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[66] Michel Ledoux. *The concentration of measure phenomenon*. American Mathematical Soc., 2001.

[67] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

[68] Likai Chen and Wei Biao Wu. Concentration inequalities for empirical processes of linear time series. *Journal of Machine Learning Research*, 18(231):1–46, 2018.

[69] Tatiana Pavlovna Krasulina. On stochastic approximation processes with infinite variance. *Theory of Probability & Its Applications*, 14(3):522–526, 1969.

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

[71] Zijian Liu and Zhengyuan Zhou. Stochastic nonsmooth convex optimization with heavy-tailed noises: High-probability bound, in-expectation rate and initial distance adaptation. *arXiv preprint arXiv:2303.12277*, 2023.

[72] Abdurakhmon Sadiev, Marina Danilova, Eduard Gorbunov, Samuel Horváth, Gauthier Gidel, Pavel Dvurechensky, Alexander Gasnikov, and Peter Richtárik. High-probability bounds for stochastic optimization and variational inequalities: the case of unbounded variance. In *International Conference on Machine Learning*, pages 29563–29648. PMLR, 2023.

[73] Jose Blanchet, Aleksandar Mijatović, and Wenhao Yang. Limit theorems for stochastic gradient descent with infinite variance. *arXiv preprint arXiv:2410.16340*, 2024.

[74] Donald L Burkholder. Sharp inequalities for martingales and stochastic integrals. *Astérisque*, 157(158):75–94, 1988.

Danna Zhang received the Ph.D. degree in statistics from the University of Chicago, Chicago, IL. She is currently an Associate Professor in the Department of Mathematics at University of California, San Diego, CA. Her research areas include high-dimensional time series, high-dimensional statistical testing, spectral analysis, online stochastic approximation, nonlinear time series, non-stationary time series, boostrapping and subsampling, high-order statistics, and robust analysis.

Zhipeng Lou received the Ph.D. degree in statistics from the University of Chicago, Chicago, IL. He is currently an Assistant Professor in the Department of Mathematics at University of California, San Diego, CA. His research interests include online inference theory, high-dimensional statistics, robust inference, time series analysis, nonparametric testing, and functional data analysis.

Wei Biao Wu received the Ph.D. degree in statistics from the University of Michigan, Ann Arbor, MI. He is currently a Professor in the Department of Statistics at the University of Chicago, Chicago, IL. His research interests include probability theory, time series analysis, high-dimensional statistical inference, and machine learning. His work focuses on the development of asymptotic theory for complex dependence structures, functional dependence measures, and the statistical analysis of stochastic gradient descent and online learning algorithms.