
Statistical Guarantees for High-Dimensional Stochastic Gradient Descent

Jiaqi Li

Department of Statistics
University of Chicago
Chicago, IL 60637
jqli@uchicago.edu

Zhipeng Lou

Department of Mathematics
University of California, San Diego
La Jolla, CA 92093
zlou@ucsd.edu

Johannes Schmidt-Hieber

Department of Applied Mathematics
University of Twente
Enschede, Netherlands
a.j.schmidt-hieber@utwente.nl

Wei Biao Wu

Department of Statistics
University of Chicago
Chicago, IL 60637
wbwu@uchicago.edu

Abstract

Stochastic Gradient Descent (SGD) and its Ruppert–Polyak averaged variant (ASGD) lie at the heart of modern large-scale learning, yet their theoretical properties in high-dimensional settings are rarely understood. In this paper, we provide rigorous statistical guarantees for constant learning-rate SGD and ASGD in high-dimensional regimes. Our key innovation is to transfer powerful tools from high-dimensional time series to online learning. Specifically, by viewing SGD as a nonlinear autoregressive process and adapting existing coupling techniques, we prove the geometric-moment contraction of high-dimensional SGD for constant learning rates, thereby establishing asymptotic stationarity of the iterates. Building on this, we derive the q -th moment convergence of SGD and ASGD for any $q \geq 2$ in general ℓ^s -norms, and, in particular, the ℓ^∞ -norm that is frequently adopted in high-dimensional sparse or structured models. Furthermore, we provide sharp high-probability concentration analysis which entails the probabilistic bound of high-dimensional ASGD. Beyond closing a critical gap in SGD theory, our proposed framework offers a novel toolkit for analyzing a broad class of high-dimensional learning algorithms.

1 Introduction

Stochastic gradient descent (SGD) has been a cornerstone in large-scale machine learning since the seminal work by Robbins and Monro [1951]. It is especially efficient in high-dimensional and overparameterized settings where the number of unknown parameters can exceed the number of training samples [Arpit et al., 2017, Zhang et al., 2017, He et al., 2016]. SGD can also be combined with regularization techniques such as dropout to prevent overfitting in large networks [Krizhevsky et al., 2012, Srivastava et al., 2014]. Despite the vast amount of theoretical work on SGD, generalization bounds of SGD in high-dimensional regimes remain limited [Garrigos and Gower, 2023]. Considering a strongly convex objective function, we provide statistical guarantees for constant learning-rate SGD and its Ruppert–Polyak averaged variant (ASGD) [Ruppert, 1988, Polyak and Juditsky, 1992] in high-dimensional settings.

Specifically, we consider a general optimization problem

$$\beta^* \in \arg \min_{\beta \in \mathbb{R}^d} G(\beta), \text{ where } \beta \mapsto G(\beta) := \mathbb{E}_{\xi \sim \Pi} g(\beta, \xi), \quad (1)$$

$g(\cdot)$ is the noise-perturbed measurement of $G(\cdot)$, and ξ denotes a random element sampled from some unknown distribution Π . Given i.i.d. random samples $\xi_1, \xi_2, \dots$ and some initialization $\beta_0 \in \mathbb{R}^d$, the k -th SGD iteration is

$$\beta_k = \beta_{k-1} - \alpha \nabla g(\beta_{k-1}, \xi_k), \quad k = 1, 2, \dots, \quad (2)$$

for some constant learning rate $\alpha > 0$, and $\nabla g(\beta, \xi) = \nabla_{\beta} g(\beta, \xi)$ the stochastic gradient with respect to β . For $k \geq 1$, the ASGD variant is defined by

$$\bar{\beta}_k = \frac{1}{k} \sum_{i=1}^k \beta_i. \quad (3)$$

We are interested in the high-dimensional setting where the parameter dimension d can be very large. Here, a notable divide between empirical success and theoretical understanding is that practitioners often employ a large constant learning rate α in (2) to accelerate convergence in high-dimensional problems [Wu et al., 2018, Cohen et al., 2021, Cai et al., 2024]. However, such choices can induce pronounced non-stationarity in the SGD iterates $\{\beta_k\}_{k \in \mathbb{N}}$ which will not converge to a point but oscillates around the mean of a stationary distribution. In other words, β_k is non-stationary but asymptotically stationary, which converges only in distribution as $k \rightarrow \infty$, while the mean of this distribution differs from the exact minimizer β^* due to the non-diminishing bias of order $O(\alpha)$ [Dieuleveut et al., 2020, Merad and Gaïffas, 2023]. Classical theory mostly relies on decaying learning rates [Zhang, 2004, Nemirovski et al., 2009, Jentzen and von Wurstemberger, 2020, Shi et al., 2023]. To address the non-stationarity issue, we apply powerful tools from nonlinear time series analysis [Wu and Shao, 2004] to online learning, particularly by adapting the coupling techniques to show the geometric-moment contraction of SGD for constant learning rates. Specifically, for any two SGD sequences $\{\beta_k\}_{k \in \mathbb{N}}$ and $\{\beta'_k\}_{k \in \mathbb{N}}$ that share the same random samples but have different initial vectors β_0 and β'_0 , we show in Theorem 1 that for all sufficiently small constant learning rates α , the initialization is forgotten exponentially fast in the sense that

$$(\mathbb{E}|\beta_k - \beta'_k|_s^q)^{1/q} \leq r_{\alpha, s, q}^k |\beta_0 - \beta'_0|_s \quad \text{holds for all } k \in \mathbb{N}, \quad (4)$$

for contraction speed $0 \leq r_{\alpha, s, q} < 1$, and $|\cdot|_s$ the ℓ^s -norm, that is,

$$|(v_1, \dots, v_d)^\top|_s = \left(\sum_{i=1}^d |v_i|^s \right)^{1/s}, \quad s \geq 1.$$

This asserts the existence of a limiting stationary distribution of β_k as $k \rightarrow \infty$, thereby facilitating a systematic convergence theory of SGD even in nonlinear, overparameterized models.

Building on this new framework, we provide non-asymptotic bounds for higher-order moments of the SGD error in general ℓ^s -norms for any finite $s \geq 2$ beyond the usual ℓ^2 -norm, extendable to max-norm ℓ^∞ by choosing $s \approx \log(d)$. Notably, the ℓ^∞ -norm is frequently adopted in high-dimensional sparse or structured estimation [Wainwright, 2019]. See for instance, the max-norm convergence of the Lasso and Dantzig selector [Lounici, 2008]; the pivotal method for sup-norm bounds of the square-root Lasso [Belloni et al., 2011]; and the max-norm error control for confidence intervals in high-dimensional regression problems [Javanmard and Montanari, 2013]. In stochastic-approximation (SA), Wainwright [2019] derived ℓ^∞ -norm bounds for Q-learning with decaying learning rates; Chen et al. [2023] derived maximal concentration bounds for SA under arbitrary norms with decaying learning rates and with contraction as an assumption; Agarwal et al. [2012] considered high-dimensional SA for strongly convex objectives with a sparse optimum, but using decaying learning rates and restricting the tails of stochastic gradients to be sub-Gaussian. To date, all the existing results are restricted to low-dimensional settings or decaying learning rates and do not carry over to overparameterized models with constant learning rates. To address this gap, we derive a sharp *high-dimensional moment inequality* (see Lemma 2) valid for a broad class of learning problems, delivering explicit non-asymptotic bounds of $\mathbb{E}|\beta_k - \beta^*|_s^q$ and its ASGD variant for any $q, s \geq 2$ with mild conditions, together with matching complexity guarantees, i.e., given some target error $\varepsilon > 0$ (see Proposition 2), the required number of iterations k such that

$$\mathbb{E}|\bar{\beta}_k - \beta^*|_s^q \leq \varepsilon.$$

Although moment bounds capture average-case performance, a single execution of (A)SGD in practice demands high-probability guarantees [Valiant, 1984, Vapnik, 2000, Bach and Moulines, 2013, Durmus et al., 2021, Zhong et al., 2024]. Recent advances include a generic high-probability framework for both convex and nonconvex SGD with sub-Gaussian gradient noises [Liu et al., 2023], high-probability rates for clipped-SGD with heavy-tailed noises [Nguyen et al., 2023], and high-probability guarantees for nonconvex stochastic approximation via robust gradient clipping [Li and Liu, 2022]. However, these established high-probability bounds focus again on decaying learning rates and low dimension. Moreover, early work primarily addressed light-tailed noises where the gradients are bounded or have exponential-type moments [Nemirovski et al., 2009, Rakhlin et al., 2012, Ghadimi and Lan, 2013, Cardot et al., 2017, Harvey et al., 2019, Mou et al., 2020, Chen et al., 2023]. For the cases that only admit a polynomial tail with finite q -th moment, Lou et al. [2022] were the first to derive a Nagaev-type inequality [Nagaev, 1979] for low-dimensional ASGD. The rate was shown to be optimal but their bound heavily relies on the linearity of gradients and is only suitable for decaying learning rates. By leveraging a dependency-adjusted *functional dependence measure* in high-dimensional time series [Zhang and Wu, 2017], we derive a high-probability concentration bounds for high-dimensional ASGD with constant learning rates. Given a tolerance level $\delta \in (0, 1)$ and a target error $\varepsilon > 0$, we provide bounds for the required number of iterations k to guarantee that

$$\mathbb{P}(|\bar{\beta}_k - \beta^*|_s \leq \varepsilon) \geq 1 - \delta.$$

This tail-decay result (see Eq. (10)) is proved via a new Fuk-Nagaev-type inequality (see Theorem 4) and complements our moment and complexity characterizations of large-step stochastic optimization.

1.1 Our Contributions

This paper contributes to theoretical advancements for understanding constant learning-rate SGD and its averaged variant (ASGD) in the challenging high-dimensional regime. Our main technical innovations and results include:

(1) Handling Constant Learning Rates in High Dimensions. In practice, large-scale machine learning models commonly deploy fixed, large learning rates to speed up optimization in high-dimensional settings. To address this, we introduce novel coupling techniques inspired by high-dimensional nonlinear time series and establish the asymptotic stationarity of the SGD iterates with arbitrary initialization (Section 2).

(2) Generalized Moment Convergence in ℓ^s - and ℓ^∞ -Norms. By deriving a sharp high-dimensional moment inequality, we establish explicit, non-asymptotic q -th moment bounds for arbitrary ℓ^s -norms of (A)SGD iterates for any $q \geq 2$ and even integers s , generalizing previous theory primarily focusing on mean squared error (MSE) convergence with $q = s = 2$. Our results extend naturally to the max-norm case (i.e., ℓ^∞) by selecting $s \approx \log(d)$, that is essential for modern sparse and structured estimation in high-dimensional data (Section 3).

(3) High-Probability Tail Bounds. While average-case (moment) bounds are informative, single runs require tail guarantees. We derive the first high-probability concentration bounds for ASGD in high-dimensional settings with constant learning rates. By developing a tight Fuk-Nagaev-type inequality using the coupling techniques in nonlinear time series, we control the algorithmic complexity required to achieve targeted accuracy with high confidence (Section 4).

1.2 Related Works

Stochastic Gradient Descent and its Variants. The SGD algorithm can be traced back to Robbins and Monro [1951], Kiefer and Wolfowitz [1952]. Popular SGD variants include Nesterov’s accelerated gradient [Nesterov, 1983], AdaGrad [Duchi et al., 2011], AdaDelta [Zeiler, 2012], Adam [Kingma and Ba, 2014], AMSGrad [Reddi et al., 2018], AdamW [Loshchilov and Hutter, 2018], SAG [Schmidt et al., 2017], SVRG [Johnson and Zhang, 2013], SARAH [Nguyen et al., 2017], SPIDER [Fang et al., 2018] and Katyusha [Allen-Zhu, 2017]. The theoretical foundations of SGD under decaying learning rates were established in the early studies by [Blum, 1954, Dvoretzky, 1956, Sacks, 1958], with stronger almost-sure guarantees by Fabian [1968], Robbins and Siegmund [1971], Ljung [1977], Lai [2003], Wang and Gao [2010]. Existing works for smooth, strongly-convex objectives with decaying step sizes include Ruppert [1988], Polyak and Juditsky [1992], Nemirovski et al. [2009], Bach and Moulines [2013], Rakhlin et al. [2012], Mertikopoulos et al. [2020] among others. Despite the rich literature on SGD, the theoretical understanding in high-dimensional settings remains limited.

Exceptions are Paquette et al. [2021, 2022] who study high-dimensional SGD for the least-squares loss.

Constant Learning Rate. In high-dimensional scenarios, constant learning rates prevail due to simpler tuning procedures and faster convergence [Wang et al., 2022]. More recent theoretical and empirical studies of large-step SGD include Wu et al. [2018], Cohen et al. [2021] and the very recent Cai et al. [2024], which formalize the resurgence of constant-step methods in modern machine learning. A useful way to analyze constant-step SGD is to treat its iterates as a time-homogeneous Markov chain [Pflug, 1986], which makes it possible to characterize its long-run behavior and stationary law. However, previous works only derived convergence in Wasserstein distance [Dieuleveut et al., 2020, Merad and Gaïffas, 2023]. Such convergence in probability measures can hardly provide refined (non)-asymptotics such as higher-moment convergence and concentration inequalities, and seems nontrivial to extend to high-dimensional regimes.

High-Dimensional Nonlinear Time Series. An alternative approach for constant learning-rate SGD is to view it as an iterated random function [Dubins and Freedman, 1966, Barnsley and Demko, 1985, Diaconis and Freedman, 1999, Diaconis and Duflo, 2000], or a nonlinear autoregressive (AR) process. This interpretation facilitates the theory of online learning with non-stationarity and complex dependency structures; see, for example, the recent work by Li et al. [2024c] on SGD with dropout regularization building on the GMC framework [Wu and Shao, 2004]. To extend this systematic theory to high-dimensional settings, we adapt the coupling techniques in time series [Wu, 2005, 2007, 2009, 2011, Xiao and Wu, 2012, Berkes et al., 2014, Wu and Wu, 2016, Karmakar and Wu, 2020], especially the ones for high-dimensional regimes [Zhang and Wu, 2017, 2021, Li et al., 2024a] to online learning algorithms.

1.3 Notation

Denote column vectors in $\mathbb{R}^d$ by lowercase bold letters $\mathbf{x} = (x_1, \dots, x_d)^\top$ and the ℓ^s -norm of $\mathbf{x}$ by $|\mathbf{x}|_s = (\sum_{i=1}^d |x_i|^s)^{1/s}$, $s \geq 1$. Write $\mathbf{x}^{\odot s} = (x_1^s, \dots, x_d^s)^\top$. The expectation and covariance of random vectors are respectively denoted by $\mathbb{E}[\cdot]$ and $\text{Cov}(\cdot)$. For $q > 0$ and a random variable X , we write $X \in \mathcal{L}^q$ iff $\|X\|_q = [\mathbb{E}(|X|^q)]^{1/q} < \infty$. We denote matrices by uppercase letters. Given matrices A and B of compatible dimension, their matrix product is denoted by juxtaposition. Write $A^\top$ for the transpose of A and I_d for $d \times d$ identity matrix. For two positive number sequences (a_n) and (b_n) , we say $a_n = O(b_n)$ (resp. $a_n \asymp b_n$) if there exists $c > 0$ such that $a_n/b_n \leq c$ (resp. $1/c \leq a_n/b_n \leq c$) for all large n . Let (x_n) and (y_n) be two sequences of random variables. Write $x_n = O_{\mathbb{P}}(y_n)$ if for $\forall \epsilon > 0$, there exists $c > 0$ such that $\mathbb{P}(|x_n/y_n| \leq c) > 1 - \epsilon$ for all large n .

Notation	Definition	Reference	Index Range
β^*	minimizer of the loss function $G(\beta)$	Eq. (1)	/
β_k	SGD iterates	Eq. (2)	$k \in \mathbb{N}$
β_k°	stationary SGD iterates	Thm. (2)	$k \in \mathbb{Z}$
$\bar{\beta}_k$	ASGD iterates	Eq. (3)	$k \in \mathbb{N}$
$\bar{\beta}_k^\circ$	stationary ASGD iterates	Eq. (9)	$k \in \mathbb{Z}$

Table 1: List of the sequences defined in the paper.

2 Convergence of SGD to a Stationary Distribution

In this section, we establish the GMC property of high-dimensional SGD with constant learning rates. Our technique is to construct a smooth surrogate for the non-differentiable ℓ^∞ -norm via the ℓ^s -norm, so that standard gradient-based tools become available. We defer the technical details to Section 6.1. Furthermore, we provide a novel high-dimensional moment inequality (see Section 6.2) and use it to derive the dimension-dependent range of the constant learning rate that guarantees the contraction.

We first impose the following assumptions on the objective function and the stochastic gradients.

Assumption 1 (Coercivity). *Assume that for any sequence $\beta_1, \beta_2, \dots$ with $|\beta_n|_s \rightarrow \infty$ the loss function $G(\cdot)$ in (1) satisfies $\lim_{n \rightarrow \infty} G(\beta_n) = \infty$.*

Assumption 2 (Strong Convexity – ℓ^s -norm). *Let $s \geq 2$ be an even integer and write $\mathbf{v}^{\odot s} := (v_1^s, \dots, v_d^s)^\top$ for a vector $\mathbf{v} = (v_1, \dots, v_d)^\top$. Assume there exists $\mu > 0$ such that*

$$\langle (\boldsymbol{\beta} - \boldsymbol{\beta}')^{\odot(s-1)}, \nabla G(\boldsymbol{\beta}) - \nabla G(\boldsymbol{\beta}') \rangle \geq \mu |\boldsymbol{\beta} - \boldsymbol{\beta}'|_s^s, \quad \text{for all } \boldsymbol{\beta}, \boldsymbol{\beta}' \in \mathbb{R}^d.$$

In Lemma 3 in the supplementary materials, we show that under Assumptions 1 and 2, a unique global minimizer $\boldsymbol{\beta}^*$ exists for the optimization problem (1). When $s = 2$, Assumption 2 reduces to the regular strong convexity frequently adopted in the literature [Polyak and Juditsky, 1992, Moulines and Bach, 2011, Dieuleveut et al., 2020, Mies and Steland, 2023]. For general s and the linear regression model, Section 8.2 in the supplementary material interprets the ℓ^s -type strong convexity assumption via the ℓ^s -norm induced matrix norm. As different norms are involved, there does not seem to be an apparent relationship between the classical strong convexity and the case $s > 2$.

Assumption 3 (Stochastic Lipschitz Continuity – ℓ^s -norm). *Let $\boldsymbol{\beta}^*$ be the global minimizer. For some $q \geq 2$ and an even integer $s \geq 2$, assume that*

$$M_{s,q} := (\mathbb{E} |\nabla g(\boldsymbol{\beta}^*, \boldsymbol{\xi})|_s^q)^{1/q} < \infty.$$

Further assume there exists a constant $L_{s,q} > 0$ such that

$$(\mathbb{E} |\nabla g(\boldsymbol{\beta}, \boldsymbol{\xi}) - \nabla g(\boldsymbol{\beta}', \boldsymbol{\xi})|_s^q)^{1/q} \leq L_{s,q} |\boldsymbol{\beta} - \boldsymbol{\beta}'|_s, \quad \text{for all } \boldsymbol{\beta}, \boldsymbol{\beta}' \in \mathbb{R}^d.$$

Later we will choose $s = O(\log(d))$ to bound the max-norm. The above defined Lipschitz constant $L_{s,q}$ and the moments $M_{s,q}$ will then grow as d increases. Taking linear regression as an example, we investigate the dimension dependence of $L_{s,q}$ and $M_{s,q}$ in Section 8.2. All bounds in this work will contain the explicit dependence on $(L_{s,q}, M_{s,q})$.

We now state the first main result of this paper, which plays a crucial role in establishing moment convergence and tail probability results in the following sections. The statement quantifies the exponential rate at which the initialization $\boldsymbol{\beta}_0$ will be forgotten and the SGD iterates $\boldsymbol{\beta}_k$ converges to a stationary distribution π_α .

Theorem 1 (Convergence of SGD to stationary distribution). *Suppose that Assumptions 1–3 hold for some $\mu > 0$, $q \geq 2$ and even integer $s \geq 2$. Given a constant learning rate*

$$0 < \alpha < \alpha_{s,q} := \frac{2\mu}{\max\{q, s\} L_{s,q}^2}, \quad (5)$$

for any two d -dimensional SGD sequences $\{\boldsymbol{\beta}_k(\alpha)\}_{k \in \mathbb{N}}$ and $\{\boldsymbol{\beta}'_k(\alpha)\}_{k \in \mathbb{N}}$ sharing the same i.i.d. noise injections $\{\boldsymbol{\xi}_k\}_{k \geq 1}$ but possibly different initializations $\boldsymbol{\beta}_0, \boldsymbol{\beta}'_0 \in \mathbb{R}^d$, the geometric-moment contraction (GMC)

$$\|\boldsymbol{\beta}_k - \boldsymbol{\beta}'_k\|_s \leq r_{\alpha,s,q}^k \|\boldsymbol{\beta}_0 - \boldsymbol{\beta}'_0\|_s, \quad \text{for all } k = 0, 1, \dots$$

holds with contraction constant

$$r_{\alpha,s,q} = 1 - 2\mu\alpha + \max\{q, s\} L_{s,q}^2 \alpha^2 < 1. \quad (6)$$

Moreover, there exists a unique stationary distribution π_α with a finite q -th moment, that is, $\int |\mathbf{u}|_s^q \pi_\alpha(d\mathbf{u}) < \infty$, such that

$$\boldsymbol{\beta}_k \Rightarrow \pi_\alpha, \quad \text{as } k \rightarrow \infty.$$

Equivalently, for any continuous function $f \in \mathcal{C}(\mathbb{R}^d)$ with $|f|_\infty < \infty$,

$$\mathbb{E}[f(\boldsymbol{\beta}_k)] \rightarrow \int f(\mathbf{u}) \pi_\alpha(d\mathbf{u}), \quad \text{as } k \rightarrow \infty.$$

The result generalizes Li et al. [2024b] to large dimension d and extends the ℓ^2 -type GMC based on Lemma 9 to general ℓ^s -norms. Moreover, choosing $s = s_d$ with

$$s_d := 2 \min\{\ell \in \mathbb{N} : 2\ell > \log(d)\}, \quad (7)$$

and using the inequality

$$|\mathbf{x}|_\infty \leq |\mathbf{x}|_{s_d} \leq d^{1/s_d} |\mathbf{x}|_\infty \leq e |\mathbf{x}|_\infty, \quad (8)$$

shows the equivalence of the ℓ^{s_d} - and ℓ^∞ -norms. Consequently, by choosing $s = s_d$, the previous theorem can also be used to derive the GMC property with respect to the ℓ^∞ -norm.

3 Convergence of High-Dimensional SGD and ASGD

In this section, we derive convergence rates for the moments of the last iterate $\mathbb{E}|\beta_k - \beta^*|_\infty^q$ and the moments of the averaged SGD.

3.1 Convergence of SGD

Proposition 1. *If Assumptions 1–3 hold for some $q \geq 2$, an even integer $s \geq 2$, and a constant $M_{s,q}$, then,*

$$\|\beta_k - \beta^*\|_s^2 \leq (1 - 2\alpha\mu + 7 \max\{q, s\}\alpha^2 L_{s,q}^2) \|\beta_{k-1} - \beta^*\|_s^2 + 3 \max\{q, s\}\alpha^2 M_{s,q}^2,$$

for all $k \geq 1$. The same inequality holds if β_k is replaced by the stationary SGD iterates $\beta_k^\circ \sim \pi_\alpha$, $k \geq 1$.

Theorem 2 (Moment convergence of SGD). *Let $0 < \alpha < \alpha_{s,q}/7$ with $\alpha_{s,q}$ as defined in (5). Suppose that Assumptions 1–3 hold for $q \geq 2$ and even integer $s \geq 2$. Then for the stationary SGD iterates $\beta_k^\circ \sim \pi_\alpha$,*

$$\|\beta_k^\circ - \beta^*\|_q = O\left(M_{s,q} \sqrt{\max\{q, s\}\alpha}\right) \quad \text{for all } k \geq 1$$

and for the SGD iterate β_k with arbitrary initialization β_0 ,

$$\|\beta_k - \beta^*\|_q = O\left(M_{s,q} \sqrt{\max\{q, s\}\alpha} + r_{\alpha,s,q}^k \|\beta_0 - \beta_0^\circ\|_q\right) \quad \text{for all } k \geq 1.$$

Choosing $s = s_d$ in (7) yields a bound with respect to the ℓ^∞ -norm.

3.2 Convergence of Ruppert-Polyak Averaged SGD

Consider now the Ruppert-Polyak Averaged SGD (ASGD) $\bar{\beta}_k = \frac{1}{k} \sum_{i=1}^k \beta_i$ as defined in (3). For the initialization $\beta_0^\circ \sim \pi_\alpha$, define the stationary ASGD sequence

$$\bar{\beta}_k^\circ = \frac{1}{k} \sum_{i=1}^k \beta_i^\circ, \quad k \geq 1. \quad (9)$$

Theorem 3. *Consider the ASGD sequence $\{\bar{\beta}_k\}_{k \geq 1}$. Suppose that Assumptions 1–3 hold with some $q \geq 2$ and even integer $s = s_d$ in (7), the conditions of Theorem 8 hold and the learning rate satisfies $\alpha \in (0, \alpha_{s_d,q})$ with $\alpha_{s_d,q}$ defined in (5). For any $k \geq 1$ and some universal constants $C_1, C_2, C_3 > 0$,*

$$\begin{aligned} \|\bar{\beta}_k - \beta^*\|_q &\leq C_1 \underbrace{\left\{ \sqrt{\frac{c_q s_d}{k}} M_{s_d,q} \left(L_{s_d,q} \sqrt{\alpha \max\{q, s_d\}} + 1 \right) \right\}}_{\text{stochastic variance}} \\ &\quad + C_2 \underbrace{\left\{ \frac{1}{k(1 - r_{\alpha,s_d,q})} \|\beta_0 - \beta_0^\circ\|_q \right\}}_{\text{initialization bias}} + C_3 \underbrace{\left\{ M_{s_d,q}^2 \max\{q, s_d\} \alpha d^{\frac{q}{q-1} \cdot (1 - \frac{2}{s_d})} \right\}}_{\text{bias of constant learning rate}}. \end{aligned}$$

Proposition 2 (Complexity bound). *Under the assumptions of Theorem 3, let $\Delta_0 = \|\beta_0 - \beta_0^\circ\|_\infty$,*

$$V = L_{s_d,q} M_{s_d,q} \sqrt{\max\{q, s_d\}} + M_{s_d,q}, \quad B = M_{s_d,q}^2 \max\{q, s_d\} d^{\frac{q}{q-1} (1 - \frac{2}{s_d})}.$$

Given a tolerance $\varepsilon > 0$,

$$\alpha \leq \min \left\{ \frac{\varepsilon}{3C_3 B}, \frac{\alpha_{s_d,q}}{7} \right\}, \quad \text{and } k \geq \max \left\{ \frac{9C_1^2 c_q s_d V^2 \alpha}{\varepsilon^2}, \frac{3C_2 \Delta_0}{\alpha \varepsilon} \right\},$$

we have $\|\bar{\beta}_k - \beta^*\|_q \leq \varepsilon$.

A proof outline is given in Section 6.3 and the full proof is deferred to the supplementary material. The sharpest complexity bound of SA for ℓ^∞ -norm known to date was derived by Wainwright [2019] proving that the number of iterations required to obtain an ε -accurate solution of Q-learning scales as $(1 - \gamma)^{-4} \cdot \varepsilon^{-2}$ with the discount factor γ . In Proposition 2, our complexity bound for SGD is also of the order of $O(1/\varepsilon^2)$ if the dimension d is fixed, which is consistent with the degenerate Q-learning case in Wainwright [2019]. The derived result allows to determine the dependence on the dimension d .

4 Sharp Concentration and Gaussian Approximation

Via the following tail probability inequality for the averaged SGD estimator $\bar{\beta}_k$, one can further derive high-probability concentration bound of $|\bar{\beta}_k - \beta^*|_\infty$. Recall that $s_d = 2 \min\{\ell \in \mathbb{N} : 2\ell > \log(d)\}$.

Theorem 4 (Fuk-Nagaev inequality). *Under the conditions of Theorem 3, for any $z > 0$, we have*

$$\mathbb{P}(|\bar{\beta}_k - \beta^*|_\infty > z) \lesssim \frac{\|\beta_0 - \beta_0^\circ\|_q^q}{(k\alpha z)^q} + \frac{(\log d)^{\frac{3q}{2}} (\log k)^{1+2q} M_{s_d,q}^q}{z^q k^{q-1} \alpha^{q/2-1}} + \exp\left(-\frac{C_k z^2 \alpha^{1-2/q}}{M_{s_d,q}^2 \log d}\right),$$

where the constants in $\lesssim$ are independent of k, d, s and α .

As an immediate consequence of Theorem 4, we obtain a sharp high-probability upper bound for $|\bar{\beta}_k - \beta^*|_\infty$, that is, for any given tolerance rate $\delta \in (0, 1)$, with at least probability $1 - \delta$, we have

$$|\bar{\beta}_k - \beta^*|_\infty = O\left(\frac{\|\beta_0 - \beta_0^\circ\|_q^q}{k\alpha\delta^{1/q}} + \frac{(\log d)^{3/2} (\log k)^{1/q+2} M_{s_d,q}}{k^{1-1/q} \alpha^{1/2-1/q} \delta^{1/q}} + \sqrt{\frac{M_{s_d,q}^2 \log d \log(1/\delta)}{k\alpha^{1-2/q}}}\right). \quad (10)$$

Notably, if the q -th moment of the gradient noise is finite ($M_{s_d,q} < \infty$), the second term of the right hand side, involving k^{1-q} , is generally unimprovable [Nagaev, 1979, Lou et al., 2022].

The distribution convergence for the high-dimensional ASGD relies on the following result. Let $M_{2,q}$ be as defined in Assumption 3.

Theorem 5 (Gaussian approximation). *Consider stationary SGD iterates $\beta_k^\circ \sim \pi_\alpha$ with π_α as defined in Theorem 1, initialization $\beta_0^\circ \sim \pi_\alpha$, and learning rate $\alpha \in (0, \alpha_{s_d,q})$. Suppose that Assumptions 1–3 hold for some $q > 2$. Then, on a potentially different probability space, and for a number of iterations T satisfying $d \leq cT$, where $c > 0$ is some constant, there exist random vectors $\{\tilde{\beta}_k\}_{k=1}^T \stackrel{\mathcal{D}}{=} \{\beta_k^\circ\}_{k=1}^T$ and independent Gaussian random vectors $\{z_k\}_{k=1}^T$ with mean zero and covariance matrix*

$$\Xi = \sum_{k=-\infty}^{\infty} \text{Cov}(\beta_0^\circ, \beta_k^\circ), \quad (11)$$

such that

$$\left(\mathbb{E} \max_{k \leq T} \left| \frac{1}{\sqrt{k}} \sum_{i=1}^k [(\tilde{\beta}_i - \mathbb{E}[\beta_1^\circ]) - z_i] \right|_2^2 \right)^{1/2} \leq C_{\alpha,q}^* M_{2,q} \sqrt{d \log(T)} \left(\frac{d}{T}\right)^{\frac{q-2}{6q-4}}, \quad (12)$$

with $C_{\alpha,q}^*$ a constant that only depends on c , the learning rate α , and the moment index q .

For diverging moment index $q \rightarrow \infty$, the Gaussian approximation rate in (12) approaches the rate $O(\sqrt{\log(T)}(d^4/T)^{1/6})$. Thus, to obtain a nontrivial Gaussian approximation bound within T iterations, we need dimension dependence $d = o(T^{1/4-\zeta})$ with $\zeta > 0$.

5 Constant Learning Rate for Large Dimension

Recall that $L_{s,q}$ is the Lipschitz constant introduced in Assumption 3. We established asymptotic stationarity and non-asymptotic convergence if $\alpha < \alpha_{s,q}/7$ with $\alpha_{s,q}$ defined in (5), leading to the upper bound

$$\alpha < \frac{\alpha_{s,q}}{7} = \frac{2\mu}{7 \max\{q, s\} L_{s,q}^2} \asymp \frac{1}{d^2 \log(d)},$$

if we choose $s = s_d$ in (7) and if $L_{s_d,q} \asymp d$. We refer to Section 8.2 for the derivation of the dimension dependence of $L_{s,q}$ in the linear regression model.

Alternatively, the upper bound for the learning rate α can also be derived by a linear approximation technique (see Lemma 1), defined as the nontrivial solution to the following equation

$$1 - q\mu\alpha + \frac{q[|q-s| + (s-1)] L_{s,q}^2 \alpha^2 (1 + \alpha L_{s,q})^{q-2}}{2} = 1. \quad (13)$$

A derivation of this equation is provided in Section 6.1. The existence of a solution of (13) is shown below the proof of Lemma 1 in the supplementary materials. When $q = 2$, the range of α simplifies to

$$\alpha < \frac{2\mu}{7[|s-2| + (s-1)]L_{s,2}^2},$$

which is also proportional to $1/[d^2 \log(d)]$ if we choose $s = s_d$ in (7) and if $L_{s_d,2} \asymp d$, matching the rate of $\alpha_{s,q}$ in (5) derived by Lemma 2, though with a slightly more conservative constant for general s . In the special case with $s = 2$, both bounds reduce to the classical $\alpha < 2\mu/L_{2,2}^2$. If $L_{2,2} \asymp d$ for large dimension d , which is shown to be true for the linear regression model in Section 8.2 in the supplementary materials, the ℓ^∞ - and the ℓ^2 -norm yield similar upper bounds for the learning rate α .

6 Proof Sketches

6.1 Bridge between ℓ^s - and ℓ^∞ - Norms

In high-dimensional regimes, convergence rates of constant-learning-rate SGD (2) with respect to the ℓ^∞ -norm are of particular interest [Wainwright, 2019, Chen et al., 2023]. However, it is extremely challenging to directly study the convergence of $|\beta_k - \beta^*|_\infty$ since the ℓ^∞ -norm is not differentiable thereby ruling out standard gradient-based tools for proving convergence rates or concentration. To address this issue, we instead study $|\cdot|_{s_d}$ with s_d defined in (7). By the equivalence between ℓ^{s_d} - and ℓ^∞ -norms shown in (8), contraction in ℓ^∞ -norm follows from ℓ^{s_d} -norm contraction.

To prove the GMC property of SGD as introduced in (4), it suffices to show that for any two d -dimensional SGD sequences $\{\beta_k\}_{k \in \mathbb{N}}$ and $\{\beta'_k\}_{k \in \mathbb{N}}$ sharing the same i.i.d. observations $\{\xi_k\}_{k \geq 1}$ but possibly different initializations $\beta_0, \beta'_0 \in \mathbb{R}^d$, the contraction holds for $|\beta_k - \beta'_k|_{s_d}$ for all $k \geq 1$. To this end, we need to determine a range of constant learning rates α such that for any $q \geq 2$ and $\beta, \beta' \in \mathbb{R}^d$, the GMC in Theorem 1 holds, i.e.,

$$(\mathbb{E}|\beta - \alpha \nabla g(\beta, \xi) - (\beta' - \alpha \nabla g(\beta', \xi))|_s^{1/q} \leq r|\beta - \beta'|_s, \quad \text{for some } r = r_{\alpha,s,q} < 1. \quad (14)$$

To derive the inequality, we first provide a lemma based on linear approximation by considering the scalar function

$$\alpha \mapsto |\mathbf{x} - \alpha \mathbf{z}|_s^q, \quad \text{where } \mathbf{x} = \beta - \beta', \quad \mathbf{z} = \nabla g(\beta, \xi) - \nabla g(\beta', \xi),$$

and linearizing it around $\alpha = 0$. Then, one only needs to prove that $\mathbb{E}|\mathbf{x} - \alpha \mathbf{z}|_s^q \leq r|\mathbf{x}|_s^q$. By the second-order Taylor expansion of $|\mathbf{x} - \alpha \mathbf{z}|_s^q$ in α , we have the linear approximation

$$|\mathbf{x} - \alpha \mathbf{z}|_s^q \approx |\mathbf{x}|_s^q - q\alpha|\mathbf{x}|_s^{q-s}\langle \mathbf{x}^{s-1}, \mathbf{z} \rangle, \quad (15)$$

with remainder term of order α^2 , see Section 2 in the supplementary materials for details. Since a simple triangle inequality argument $\|\mathbf{x} - \alpha \mathbf{z}\|_q \leq \|\mathbf{x}\|_q + \alpha\|\mathbf{z}\|_q$ fails to control this remainder sufficiently to yield a contraction constant $r < 1$, we establish a more precise bound.

Lemma 1. *Recall that $\mathbf{v}^{\otimes s} = (v_1^s, \dots, v_d^s)^\top$ for a vector $\mathbf{v} = (v_1, \dots, v_d)^\top$. For any $q \geq 2$, any even integer $s \geq 2$, any two vectors $\mathbf{x}, \mathbf{z} \in \mathbb{R}^d$, and any $\alpha > 0$,*

$$\left| |\mathbf{x} - \alpha \mathbf{z}|_s^q - |\mathbf{x}|_s^q + q\alpha|\mathbf{x}|_s^{q-s}\langle \mathbf{x}^{s-1}, \mathbf{z} \rangle \right| \leq \frac{q\alpha^2}{2} [q-s + (s-1)] (|\mathbf{x}|_s + \alpha|\mathbf{z}|_s)^{q-2} |\mathbf{z}|_s^2.$$

If $s = 2, q = 2$, the right-hand side is $\alpha^2|\mathbf{z}|_2^2$. This is consistent with the Taylor remainder of the right-hand side in Lemma 9 derived by Li et al. [2024c]. Using this inequality to prove the contraction in (14) is remarkably different from the approaches relying on the martingale decomposition (MD) that is frequently adopted in the literature [Dieuleveut et al., 2020, Mertikopoulos et al., 2020, Mies and Steland, 2023]. Our proposed method requires mild moment conditions on the stochastic gradients and yields simpler proofs that can be generalized to a broad class of online learning problems. We refer to Li et al. [2024b] for detailed discussion. Nevertheless, we remark in advance that a Rio-type inequality (Lemma 2) with slightly sharper constants will be used directly in our main contraction proof, while we retain Lemma 1 here for its intuitive appeal. Finally, by choosing $s = s_d$ as in (7) for (14), we can expect the ℓ^∞ -norm type GMC to hold for high-dimensional SGD iterates.

6.2 High-Dimensional Moment Inequality

To prove Theorem 1, we derive a high-dimensional version of Rio's inequality [Rio, 2009], adapted to the q -th moment of ℓ^s -norm. This result provides a slightly sharper constant than Lemma 1 and is used directly in our moment-contraction analysis.

Lemma 2 (High-dimensional moment inequality). *For any $q \geq 2$, any even integer $s \geq 2$, and any two d -dimensional random vectors $\mathbf{x}, \mathbf{y}$, we have*

$$\|\mathbf{x} + \mathbf{y}\|_q^2 \leq \|\mathbf{x}\|_q^2 + 2\|\mathbf{x}\|_q^{2-q} \mathbb{E} \left(|\mathbf{x}|_s^{q-s} \sum_{j=1}^d x_j^{s-1} y_j \right) + (\max\{q, s\} - 1) \|\mathbf{y}\|_q^2.$$

Moreover, if $\mathbb{E}[\mathbf{y} \mid \mathbf{x}] = 0$ almost surely, then

$$\|\mathbf{x} + \mathbf{y}\|_q^2 \leq \|\mathbf{x}\|_q^2 + (\max\{q, s\} - 1) \|\mathbf{y}\|_q^2. \quad (16)$$

Repeatedly applying Lemma 2 leads to the high-dimensional maximal moment inequality in Lemma 8 in the supplementary materials, which is of independent interest.

6.3 Stationarity, Variation and Bias of ASGD

We prove the moment bound $\|\bar{\beta}_k - \beta^*\|_\infty$ via the decomposition

$$\|\bar{\beta}_k - \beta^*\|_\infty \leq \|\bar{\beta}_k - \bar{\beta}_k^\circ\|_\infty + \|\bar{\beta}_k^\circ - \mathbb{E}[\bar{\beta}_k^\circ]\|_\infty + \|\mathbb{E}[\bar{\beta}_k^\circ] - \beta^*\|_\infty.$$

The first term accounts for the deviation due to the non-stationarity of $\bar{\beta}_k$ as it is initialized from an arbitrarily fixed β_0 ; this can be bounded using the GMC property of β_k shown in Theorem 1. The second term captures the stochastic variance of the stationary ASGD sequence. Bounding this term is more delicate because of the intricate dependency structure of $\bar{\beta}_k^\circ$. To address this, we deploy another powerful tool in time series – the *functional dependence measure* [Wu, 2005] in Section 8.6 of the supplementary materials, which can effectively quantify the contribution of the random sample ξ_i to the k -th SGD iterate β_k° for all $i \leq k$. As such, by controlling the cumulative dependence measures, we can bound this variance. Lastly, we handle the third term, which represents the non-diminishing bias of $\bar{\beta}_k^\circ$ induced by the constant learning rate α [Dieuleveut et al., 2020, Huo et al., 2023]. This can be dealt with by extending the approach in Li et al. [2024b] to high-dimensional settings.

Theorem 6 (Asymptotic stationarity). *Consider the ASGD iterates $\bar{\beta}_k$ and the stationary version $\bar{\beta}_k^\circ$. Suppose that Assumptions 1–3 are satisfied for some $q \geq 2$ and some even integer $s \geq 2$. Then, for the learning rate $\alpha \in (0, \alpha_{s,q})$ with $\alpha_{s,q}$ defined in (5),*

$$\|\bar{\beta}_k - \bar{\beta}_k^\circ\|_q \leq \frac{1}{k} \cdot \frac{1}{1 - r_{\alpha,s,q}} |\beta_0 - \beta_0^\circ|_s.$$

As a direct consequence of Theorem 6, we have $\|\bar{\beta}_k - \bar{\beta}_k^\circ\|_q \lesssim |\beta_0 - \beta_0^\circ|_s / (k\alpha)$, which indicates the asymptotic stationarity of high-dimensional ASGD sequences. When the bias induced by the initialization is controlled, i.e., $|\beta_0 - \beta_0^\circ|_s < \infty$, as $k\alpha \rightarrow \infty$, the ASGD iterate $\bar{\beta}_k$ approaches the stationary solution $\bar{\beta}_k^\circ$ in the sense that $\|\bar{\beta}_k - \bar{\beta}_k^\circ\|_q \rightarrow 0$. By Theorem 6, we only need to show the convergence for stationary ASGD.

Theorem 7 (Stochasticity of stationary ASGD). *Consider the stationary SGD sequence $\{\beta_k^\circ\}_{k \geq 1}$. Suppose that Assumptions 1–3 hold with some $q \geq 2$ and some even integer $s \geq 2$. Then there exists a constant $c_q > 0$ only depending on q , such that, for all $k \geq 1$,*

$$\|\bar{\beta}_k^\circ - \mathbb{E}[\bar{\beta}_k^\circ]\|_q \leq \sqrt{\frac{c_q s}{k}} M_{s,q} \left(L_{s,q} \sqrt{\alpha \max\{q, s\}} + 1 \right).$$

In the low-dimensional case, we take $s = 2$ as a special example. Then, $L_{s,q} \sqrt{\alpha \max\{q, s\}} = O(1)$ such that the bound is $\|\bar{\beta}_k^\circ - \mathbb{E}[\bar{\beta}_k^\circ]\|_q = O\{1/\sqrt{k}\}$. This rate is optimal considering the central limit theorem of the stationary ASGD.

Next, we consider the bias induced by the constant learning rate. We first introduce some necessary notation. Recall $G(\beta) = \mathbb{E}[\nabla g(\beta, \xi)]$ and $\nabla G(\beta) = (\partial_1 G(\beta), \dots, \partial_d G(\beta))^\top$, where $\beta = (\beta_1, \dots, \beta_d)^\top$. Denote $\partial_i G(\beta) = \partial G(\beta) / \partial \beta_i$, $1 \leq i \leq d$,

$$\nabla^2 G(\beta) = [\partial_i \partial_j G(\beta)]_{1 \leq i, j \leq d}, \quad \nabla^3 G_i(\beta) = [\partial_i \partial_l \partial_r G(\beta)]_{1 \leq l, r \leq d}. \quad (17)$$

We provide the non-asymptotic bound for the bias of stationary ASGD in the following lemma.

Theorem 8 (Bias of stationary ASGD). *Under Assumptions 1–3, consider the stationary ASGD $\bar{\beta}_k^\circ$. Assume that $g(\beta, \xi)$ is twice differentiable with respect to β with positive definite Hessian matrix $\nabla^2 G(\beta^*)$, and uniformly bounded derivatives $\max_{1 \leq i \leq d} \|\nabla^3 G_i(\beta)\|_\infty < \infty$, where*

$$\|\nabla^3 G_i(\beta)\|_\infty := \max_{1 \leq l \leq d} \sum_{r=1}^d \left| (\nabla^3 G_i(\beta))_{l,r} \right|.$$

Then, we have

$$\|\mathbb{E}[\bar{\beta}_k^\circ - \beta^*]\|_\infty = O\left(M_{s_d,q}^2 \max\{q, s_d\} \alpha d^{\frac{q}{q-1} \cdot (1 - \frac{2}{s_d})}\right).$$

7 Conclusions and Discussion

This work advances the theoretical understanding of the constant learning-rate SGD algorithms in high-dimensional settings. By introducing novel coupling techniques in nonlinear time series, we establish asymptotic stationarity of SGD with any initialization. We then derive non-asymptotic q -th moment bounds in general ℓ^s - and ℓ^∞ -norms, and develop the first Fuk-Nagaev high-probability tail bound for ASGD. While this paper assumes strong convexity and smoothness of the objective, the nonlinear time series perspective offers a principled framework applicable to a broad class of over-parameterized optimization tasks and can be extended to non-convex regimes, providing fundamental insights into the stability, convergence, and reliability of large-scale learning algorithms.

Acknowledgments and Disclosure of Funding

We sincerely thank the program chair, senior area chair, area chair, and the four reviewers for their constructive feedback and involved discussion, which has greatly improved the clarity of our paper. Jiaqi Li’s research is partially supported by the NSF (Grant NSF/DMS-2515926). Johannes Schmidt-Hieber has received funding from the Dutch Research Council (NWO) via the Vidi grant VI.Vidi.192.021. Wei Biao Wu’s research is partially supported by the NSF (Grants NSF/DMS-2311249, NSF/DMS-2027723). We would like to thank Insung Kong for helpful discussions.

References

- A. Agarwal, S. Negahban, and M. J. Wainwright. Stochastic optimization and sparse statistical recovery: Optimal algorithms for high dimensions. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.
- Z. Allen-Zhu. Katyusha: the first direct acceleration of stochastic gradient methods. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1200–1205, Montreal Canada, 2017. ACM.
- D. Arpit, S. Jastrzyski, N. Ballas, D. Krueger, E. Bengio, M. S. Kanwal, T. Maharaj, A. Fischer, A. Courville, Y. Bengio, and S. Lacoste-Julien. A closer look at memorization in deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 233–242. PMLR, 2017.
- F. Bach and E. Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate $O(1/n)$. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.
- M. F. Barnsley and S. Demko. Iterated function systems and the global construction of fractals. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 399 (1817):243–275, 1985.
- A. Belloni, V. Chernozhukov, and L. Wang. Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806, 2011.
- I. Berkes, W. Liu, and W. B. Wu. Komlós–Major–Tusnády approximation under dependence. *The Annals of Probability*, 42(2):794–817, 2014.

- J. R. Blum. Approximation methods which converge with probability one. *The Annals of Mathematical Statistics*, 25(2):382–386, 1954.
- Y. Cai, J. Wu, S. Mei, M. Lindsey, and P. Bartlett. Large stepsize gradient descent for non-homogeneous two-layer networks: Margin improvement and fast optimization. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.
- H. Cardot, P. Cénac, and A. Godichon-Baggioni. Online estimation of the geometric median in Hilbert spaces: Nonasymptotic confidence balls. *The Annals of Statistics*, 45(2):591–614, 2017.
- L. Chen, G. Keilbar, and W. B. Wu. Recursive quantile estimation: Non-asymptotic confidence bounds. *Journal of Machine Learning Research*, 24(91):1–25, 2023.
- Z. Chen, S. Theja Maguluri, and M. Zubeldia. Concentration of contractive stochastic approximation: Additive and multiplicative noise. *arXiv e-prints*, art. arXiv:2303.15740, 2023.
- V. Chernozhukov, D. Chetverikov, and K. Kato. Comparison and anti-concentration bounds for maxima of Gaussian random vectors. *Probability Theory and Related Fields*, pages 47–70, 2015.
- J. Cohen, S. Kaur, Y. Li, J. Z. Kolter, and A. Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. In *International Conference on Learning Representations*, 2021.
- P. Diaconis and M. Duflo. Random iterative models. In *Journal of the American Statistical Association*, volume 95, page 342, 2000.
- P. Diaconis and D. Freedman. Iterated random functions. *SIAM Review*, 41(1):45–76, 1999.
- A. Dieuleveut, A. Durmus, and F. Bach. Bridging the gap between constant step size stochastic gradient descent and Markov chains. *The Annals of Statistics*, 48(3):1348–1382, 2020.
- L. E. Dubins and D. A. Freedman. Invariant probabilities for certain Markov processes. *The Annals of Mathematical Statistics*, 37(4):837–848, 1966.
- J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159, 2011.
- A. Durmus, E. Moulines, A. Naumov, S. Samsonov, K. Scaman, and H. T. Wai. Tight high probability bounds for linear stochastic approximation with fixed stepsize. In *Advances in Neural Information Processing Systems*, 2021.
- A. Dvoretzky. On stochastic approximation. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, volume 3.1, pages 39–56. University of California Press, 1956.
- V. Fabian. On asymptotic normality in stochastic approximation. *The Annals of Mathematical Statistics*, 39(4):1327–1332, 1968.
- C. Fang, C. J. Li, Z. Lin, and T. Zhang. SPIDER: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. *Advances in Neural Information Processing Systems*, 31, 2018.
- G. Garrigos and R. M. Gower. Handbook of convergence theorems for (stochastic) gradient methods. *arXiv e-prints*, art. arXiv:2301.11235, 2023. doi: 10.48550/arXiv.2301.11235.
- S. Ghadimi and G. Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.
- N. J. A. Harvey, C. Liaw, Y. Plan, and S. Randhawa. Tight analyses for non-smooth stochastic gradient descent. In *Proceedings of the Thirty-Second Conference on Learning Theory*, pages 1579–1613. PMLR, 2019.
- K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.
- R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, Cambridge, 1985.

- D. Huo, Y. Chen, and Q. Xie. Bias and extrapolation in Markovian linear stochastic approximation with constant stepsizes. In *Abstract Proceedings of the 2023 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*, pages 81–82, 2023.
- A. Javanmard and A. Montanari. Confidence intervals and hypothesis testing for high-dimensional statistical models. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.
- A. Jentzen and P. von Wurstemberger. Lower error bounds for the stochastic gradient descent optimization algorithm: Sharp convergence rates for slowly and fast decaying learning rates. *Journal of Complexity*, 57:101438, 2020.
- R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. *Advances in Neural Information Processing Systems*, 26, 2013.
- S. Karmakar and W. B. Wu. Optimal Gaussian approximation for multiple time series. *Statistica Sinica*, 30(3):1399–1417, 2020.
- J. Kiefer and J. Wolfowitz. Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics*, 23(3):462–466, 1952.
- D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations*, 2014.
- A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.
- T. L. Lai. Stochastic approximation: invited paper. *The Annals of Statistics*, 31(2):391–406, 2003.
- J. Li, L. Chen, W. Wang, and W. B. Wu. ℓ^2 inference for change points in high-dimensional time series via a two-way MOSUM. *Ann. Statist.*, 52(2):602–627, 2024a.
- J. Li, Z. Lou, S. Richter, and W. B. Wu. The stochastic gradient descent from a nonlinear time series perspective, 2024b. Manuscript.
- J. Li, J. Schmidt-Hieber, and W. B. Wu. Asymptotics of stochastic gradient descent with dropout regularization in linear models. *arXiv preprint*, 2024c. arXiv:2409.07434.
- S. Li and Y. Liu. High probability guarantees for nonconvex stochastic gradient descent with heavy tails. In *Proceedings of the 39th International Conference on Machine Learning*, pages 12931–12963. PMLR, 2022.
- Z. Liu, T. D. Nguyen, T. H. Nguyen, A. Ene, and H. L. Nguyen. High probability convergence of stochastic gradient methods. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *ICML’23*, pages 21884–21914, Honolulu, Hawaii, USA, 2023. JMLR.org.
- L. Ljung. Analysis of recursive stochastic algorithms. *IEEE Transactions on Automatic Control*, 22(4):551–575, 1977.
- I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018.
- Z. Lou, W. Zhu, and W. B. Wu. Beyond sub-Gaussian noises: Sharp concentration analysis for stochastic gradient descent. *Journal of Machine Learning Research*, 23(46), 2022.
- K. Lounici. Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators. *Electronic Journal of Statistics*, 2(none):90–102, 2008.
- I. Merad and S. Gaïffas. Convergence and concentration properties of constant step-size SGD through Markov chains. *arXiv e-prints*, art. arXiv:2306.11497, 2023.
- P. Mertikopoulos, N. Hallak, A. Kavis, and V. Cevher. On the almost sure convergence of stochastic gradient descent in non-convex problems. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1117–1128. Curran Associates, Inc., 2020.

- F. Mies and A. Steland. Sequential Gaussian approximation for nonstationary time series in high dimensions. *Bernoulli*, 29(4):3114–3140, 2023.
- W. Mou, C. J. Li, M. J. Wainwright, P. L. Bartlett, and M. I. Jordan. On linear stochastic approximation: Fine-grained Polyak-Ruppert and non-asymptotic concentration. In *Proceedings of Thirty Third Conference on Learning Theory*, pages 2947–2997. PMLR, 2020.
- E. Moulines and F. Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011.
- S. V. Nagaev. Large deviations of sums of independent random variables. *The Annals of Probability*, 7(5):745–789, 1979.
- A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.
- Y. E. Nesterov. A method for solving the convex programming problem with convergence rate $O(1/k^2)$. *Dokl. Akad. Nauk SSSR*, 269(3):543–547, 1983.
- L. M. Nguyen, J. Liu, K. Scheinberg, and M. Takáč. SARAH: A novel method for machine learning problems using stochastic recursive gradient. *arXiv:1703.00102*, 2017.
- T. D. Nguyen, T. H. Nguyen, A. Ene, and H. Nguyen. Improved convergence in high probability of clipped gradient methods with heavy tailed noise. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.
- C. Paquette, K. Lee, F. Pedregosa, and E. Paquette. SGD in the Large: Average-case Analysis, Asymptotics, and Stepsize Criticality. In *Proceedings of Thirty Fourth Conference on Learning Theory*, pages 3548–3626. PMLR, July 2021.
- C. Paquette, E. Paquette, B. Adlam, and J. Pennington. Implicit regularization or implicit conditioning? exact risk trajectories of SGD in high dimensions. In *Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS '22*, pages 35984–35999, Red Hook, NY, USA, Nov. 2022. Curran Associates Inc.
- G. C. Pflug. Stochastic minimization with constant step-size: Asymptotic laws. *SIAM Journal on Control and Optimization*, 24(4):655–666, 1986.
- B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.
- A. Rakhlin, O. Shamir, and K. Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. In *Proceedings of the 29th International Conference on Machine Learning*, pages 1571–1578, 2012.
- S. J. Reddi, S. Kale, and S. Kumar. On the convergence of Adam and beyond. In *International Conference on Learning Representations*, 2018.
- E. Rio. Moment inequalities for sums of dependent random variables under projective conditions. *Journal of Theoretical Probability*, 22(1):146–163, 2009.
- H. Robbins and S. Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951.
- H. Robbins and D. Siegmund. A convergence theorem for non negative almost supermartingales and some applications. In J. S. Rustagi, editor, *Optimizing Methods in Statistics*, pages 233–257. Academic Press, 1971.
- 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.

- M. Schmidt, N. Le Roux, and F. Bach. Minimizing finite sums with the stochastic average gradient. *Math. Program.*, 162(1-2):83–112, 2017.
- B. Shi, W. J. Su, and M. I. Jordan. On learning rates and Schrödinger operators. *Journal of Machine Learning Research*, 24:1–53, 2023.
- N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014.
- L. G. Valiant. A theory of the learnable. *Commun. ACM*, 27(11):1134–1142, 1984.
- V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, New York, NY, 2000.
- R. Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 2018.
- M. J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 2019.
- M. J. Wainwright. Stochastic approximation with cone-contractive operators: Sharp ℓ_∞ -bounds for Q -learning. *arXiv e-prints*, art. arXiv:1905.06265, 2019.
- X. Wang and N. Gao. Stochastic resource allocation over fading multiple access and broadcast channels. *IEEE Transactions on Information Theory*, 56(5):2382–2391, 2010.
- Y. Wang, M. Chen, T. Zhao, and M. Tao. Large learning rate tames homogeneity: Convergence and balancing effect. In *International Conference on Learning Representations*, 2022.
- N. Wiener. *Nonlinear problems in random theory*. MIT Press, 1958.
- L. Wu, C. Ma, and W. E. How SGD selects the global minima in over-parameterized learning: A dynamical stability perspective. In *Advances in Neural Information Processing Systems*, volume 31, 2018.
- W. B. Wu. Nonlinear system theory: Another look at dependence. *PNAS*, 102(40):14150–14154, 2005.
- W. B. Wu. Strong invariance principles for dependent random variables. *The Annals of Probability*, 35(6):2294–2320, 2007.
- W. B. Wu. Recursive estimation of time-average variance constants. *Ann. Appl. Probab.*, 19(4):1529–1552, 2009.
- W. B. Wu. Asymptotic theory for stationary processes. *Statistics and Its Interface*, 4(2):207–226, 2011.
- W. B. Wu and X. Shao. Limit theorems for iterated random functions. *Journal of Applied Probability*, 41(2):425–436, 2004.
- W. B. Wu and Y. N. Wu. Performance bounds for parameter estimates of high-dimensional linear models with correlated errors. *Electronic Journal of Statistics*, 10(1):352–379, 2016.
- H. Xiao and W. B. Wu. Covariance matrix estimation for stationary time series. *The Annals of Statistics*, 40(1), 2012.
- M. D. Zeiler. Adadelta: An adaptive learning rate method. *arXiv e-prints*, art. arXiv:1212.5701, 2012.
- C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization. In *2017 International Conference on Learning Representations (ICLR)*, 2017.
- D. Zhang and W. B. Wu. Gaussian approximation for high dimensional time series. *The Annals of Statistics*, 45(5):1895–1919, 2017.

- D. Zhang and W. B. Wu. Convergence of covariance and spectral density estimates for high-dimensional locally stationary processes. *The Annals of Statistics*, 49(1), 2021.
- T. Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32(1):56–134, 2004.
- Y. Zhong, J. Li, and S. Lahiri. Probabilistic guarantees of stochastic recursive gradient in non-convex finite sum problems. In D.-N. Yang, X. Xie, V. S. Tseng, J. Pei, J.-W. Huang, and J. C.-W. Lin, editors, *Advances in Knowledge Discovery and Data Mining*, pages 142–154, Singapore, 2024. Springer Nature.

8 Technical Appendices and Supplementary Material

8.1 Existence and Uniqueness of Global Minimum

Lemma 3. *Consider the minimization problem $\beta^* \in \arg \min_{\beta \in \mathbb{R}^d} G(\beta)$. If the function G satisfies Assumptions 1 and 2, then a global minimizer β^* exists and is unique.*

Proof of Lemma 3. We first show the existence of a global minimizer. By the coercivity condition in Assumption 1, $\lim_{|\beta|_s \rightarrow \infty} G(\beta) = \infty$, which implies that we can choose some large $\delta \in \mathbb{R}$ such that the sub-level set

$$\mathcal{S}_\delta := \{\beta \in \mathbb{R}^d : G(\beta) \leq \delta\}$$

is non-empty and bounded. Since G is continuous by Assumption 2, $\mathcal{S}_\delta$ is also closed, and hence compact in $\mathbb{R}^d$ by the Heine–Borel theorem. Finally, by applying the Weierstrass extreme value theorem, there exists $\beta^* \in \mathcal{S}_\delta$ such that $G(\beta^*) = \min_{\beta \in \mathcal{S}_\delta} G(\beta)$. Since for any $\beta \notin \mathcal{S}_\delta$, $G(\beta) > \delta \geq G(\beta^*)$, $G(\beta^*) = \min_{\beta \in \mathbb{R}^d} G(\beta)$.

Next, we show the uniqueness of the global minimum. Assume that there are two distinct minimizers $\beta_1 \neq \beta_2$. By Assumption 2, there exists $\mu > 0$ such that

$$\langle (\beta_1 - \beta_2)^{\odot(s-1)}, \nabla G(\beta_1) - \nabla G(\beta_2) \rangle \geq \mu |\beta_1 - \beta_2|_s^s > 0.$$

However, since β_1 and β_2 are both minimizers, $\nabla G(\beta_1) = \nabla G(\beta_2) = 0$, while $\mu |\beta_1 - \beta_2|_s^s > 0$. This leads to contradiction, which finishes the proof. $\square$

8.2 Example: Linear Regression

As example, we consider the SGD algorithm for the high-dimensional linear regression, observing independent and identically distributed (i.i.d.) pairs $\xi_1 := (\mathbf{x}_1, y_1), \xi_2 := (\mathbf{x}_2, y_2), \dots$ satisfying

$$y_k = \mathbf{x}_k^\top \beta + \epsilon_k, \quad \text{for } k = 1, 2, \dots, \quad (18)$$

for random noises ϵ_k that are independent of $\mathbf{x}_k$ with $\mathbb{E}[\epsilon_k] = 0$ and $\mathbb{E}|\epsilon_k|^q < \infty$ for some $q \geq 2$. We verify Assumptions 2 and 3 and derive the explicit dependency of the learning-rate, the Lipschitz constant, and the moments of the gradient noise on the dimension d .

Let $\xi = (y, \mathbf{x})$ be an independent random sample from the same distribution as the data. The least-squares loss and the stochastic gradient are respectively given by

$$g(\beta, \xi) = \frac{1}{2} (y - \mathbf{x}^\top \beta)^2, \quad \text{and } \nabla g(\beta, \xi) = -(y - \mathbf{x}^\top \beta) \mathbf{x}. \quad (19)$$

Then

$$\nabla G(\beta) = \mathbb{E}[\nabla g(\beta, (y, \mathbf{x}))] = -\mathbb{E}[(y - \mathbf{x}^\top \beta) \mathbf{x}] = \mathbb{E}[\mathbf{x} \mathbf{x}^\top] (\beta - \beta^*). \quad (20)$$

Let

$$\Sigma = \mathbb{E}[\mathbf{x} \mathbf{x}^\top], \quad \mathbf{v} = \beta - \beta'. \quad (21)$$

To verify the ℓ^s -type strong convexity

$$\langle (\beta_1 - \beta_2)^{\odot(s-1)}, \nabla G(\beta) - \nabla G(\beta') \rangle \geq \mu |\beta - \beta'|_s^s, \quad \text{for all } \beta, \beta' \in \mathbb{R}^d,$$

imposed in Assumption 2, observe that $\nabla G(\beta) - \nabla G(\beta') = \Sigma \mathbf{v}$. Thus, the condition becomes

$$0 < \lambda_{\min}^{(s)} := \inf_{\mathbf{v} \in \mathbb{R}^d, \mathbf{v} \neq 0} \frac{\langle \mathbf{v}^{s-1}, \Sigma \mathbf{v} \rangle}{|\mathbf{v}|_s^s}. \quad (22)$$

Lemma 4. *Let $s \in \{2, 4, 6, \dots\}$. Writing $\Sigma = (\Sigma_{ij})_{i,j=1,\dots,d}$, we have*

$$\lambda_{\min}^{(s)} \geq \min_{i=1,\dots,d} \Sigma_{ii} - \sum_{j:j \neq i} |\Sigma_{ij}|.$$

Proof of Lemma 4. Write $\mathbf{v} = (v_1, \dots, v_d)^\top$. Because of $(|v_i|^{s-1} - |v_j|^{s-1})(|v_i| - |v_j|) \geq 0$, we obtain $|v_i^{s-1}v_j| + |v_iv_j^{s-1}| \leq v_i^s + v_j^s$ and

$$\begin{aligned} \langle \mathbf{v}^{\odot(s-1)}, \Sigma \mathbf{v} \rangle &= \sum_{i=1}^d \Sigma_{ii} v_i^s - \sum_{i < j} \Sigma_{ij} (v_i^{s-1} v_j + v_i v_j^{s-1}) \\ &\geq \sum_{i=1}^d \Sigma_{ii} v_i^s - \sum_{i < j} |\Sigma_{ij}| (v_i^s + v_j^s) \\ &\geq \left(\min_{i=1, \dots, d} \Sigma_{ii} - \sum_{j: j \neq i} |\Sigma_{ij}| \right) \sum_{\ell} v_{\ell}^s. \end{aligned}$$

□

This shows that the rightmost “Gershgorin gap” $\min_{i=1, \dots, d} \Sigma_{ii} - \sum_{j: j \neq i} |\Sigma_{ij}|$ is a universal lower bound for every s . The lower bound is non-trivial if Σ is sufficiently diagonally dominant.

For large s , the inequality $\lambda_{\min}^{(s)} \geq \min_{i=1, \dots, d} \Sigma_{ii} - \sum_{j: j \neq i} |\Sigma_{ij}|$ is nearly sharp. To see this, let i^* be the index i that minimizes $\min_{i=1, \dots, d} \Sigma_{ii} - \sum_{j: j \neq i} |\Sigma_{ij}|$. For a small $\delta > 0$, pick $\mathbf{v} = (v_1, \dots, v_d)$ by choosing $v_{i^*} := 1$ and for $i \neq i^*$, taking $v_i := -\text{sign}(\Sigma_{i^*i})(1 - \delta)$. For large s , $\mathbf{v}^{\odot(s-1)} \approx (0, 0, \dots, 1, 0, \dots, 0)$ with the 1 at the i^* -th position. Similarly, $|\mathbf{v}|_s^s \approx 1$. The i^* -th entry of $\Sigma \mathbf{v}$ is given by $\Sigma_{i^*i^*} - \sum_{j \neq i^*} |\Sigma_{i^*j}| + O(\delta)$. Hence for suitable sequences $\delta \rightarrow 0$ and $s \rightarrow \infty$, we obtain $\langle \mathbf{v}^{\odot(s-1)}, \Sigma \mathbf{v} \rangle / |\mathbf{v}|_s^s \rightarrow \Sigma_{i^*i^*} - \sum_{j \neq i^*} |\Sigma_{i^*j}| = \min_i \Sigma_{ii} - \sum_{j \neq i} |\Sigma_{ij}|$.

Regarding Assumption 3, we investigate the dependence of the Lipschitz constant $L_{s,q}$ on the dimension d in high-dimensional linear regression models. If s^* is the dual exponent of s , satisfying $1/s + 1/s^* = 1$, we show that the condition holds with

$$L_{s,q} = \|\mathbf{x}_s | \mathbf{x}_{s^*}\|_q. \quad (23)$$

To see this, for any two vectors $\beta, \beta' \in \mathbb{R}^d$, we have

$$\nabla g(\beta, \xi) - \nabla g(\beta', \xi) = -\left[(y - \mathbf{x}^\top \beta) \mathbf{x} - (y - \mathbf{x}^\top \beta') \mathbf{x}\right] = \mathbf{x} \mathbf{x}^\top (\beta - \beta'). \quad (24)$$

Taking the ℓ^s -norm on both sides, we obtain

$$|\nabla g(\beta, \xi) - \nabla g(\beta', \xi)|_s = |\mathbf{x} \mathbf{x}^\top (\beta - \beta')|_s = |\mathbf{x}|_s |\mathbf{x}^\top (\beta - \beta')|. \quad (25)$$

By Hölder’s inequality, for the dual exponent s^* satisfying $1/s + 1/s^* = 1$, it follows that

$$|\mathbf{x}^\top (\beta - \beta')| \leq |\mathbf{x}|_{s^*} |\beta - \beta'|_s. \quad (26)$$

Therefore, for $q \geq 2$, we have the q -th moment bounded as follows,

$$\left(\mathbb{E} |\nabla g(\beta, \xi) - \nabla g(\beta', \xi)|_s^q\right)^{1/q} \leq \left(\mathbb{E} [|\mathbf{x}|_s^q |\mathbf{x}_{s^*}|^q]\right)^{1/q} |\beta - \beta'|_s,$$

proving (23).

Recall s_d defined in (7). To bound the ℓ^∞ -norm, we set the conjugates

$$s = s_d, \quad s_d^* = \frac{s_d}{s_d - 1}. \quad (27)$$

Recall that for the ℓ^s -norm, we have $|\mathbf{x}|_\infty \leq |\mathbf{x}|_{s_d} \leq d^{1/s_d} |\mathbf{x}|_\infty \leq e |\mathbf{x}|_\infty$. Similarly, for the conjugate $\ell^{s_d^*}$ -norm, $d^{\frac{1}{s_d^*} - 1} = d^{\frac{1}{s_d}} \leq e$ implies

$$\frac{1}{e} |\mathbf{x}|_1 \leq \frac{1}{d^{\frac{1}{s_d^*} - 1}} |\mathbf{x}|_1 \leq |\mathbf{x}|_{s_d^*} \leq |\mathbf{x}|_1, \quad (28)$$

which together with (23) gives

$$L_{s_d, q} \leq e \|\mathbf{x}|_\infty | \mathbf{x}|_1\|_q. \quad (29)$$

The next two lemmas show that the tail behavior of the covariate vector $\mathbf{x}_k$ determines the behavior of the Lipschitz constant $L_{s,q}$ and the moment $M_{s,q}$ defined in Assumption 3.

Lemma 5. Consider the linear regression in (18) with i.i.d. generic random samples $(\mathbf{x}, y)$, where $\mathbf{x} = (x_1, \dots, x_d)^\top$. Let $q \geq 2$ and recall s_d in (7).

(i) (Sub-Gaussian) If there is a constant K such that for all $\mathbf{u} \in \mathbb{R}^d$, $|\mathbf{u}^\top \mathbf{x}|_{\psi_2} \leq K|\mathbf{u}|_2$, where $|v|_{\psi_2} = \inf\{t > 0 : \mathbb{E}[e^{v^2/t^2}] \leq 2\}$ denotes the sub-Gaussian norm, then

$$L_{s_d, q} = O(d\sqrt{\log(d)}).$$

(ii) (Sub-exponential) If there is a constant K such that for all $\mathbf{u} \in \mathbb{R}^d$, $|\mathbf{u}^\top \mathbf{x}|_{\psi_1} \leq K|\mathbf{u}|_2$, where $|v|_{\psi_1} = \inf\{t > 0 : \mathbb{E}[e^{|v|/t}] \leq 2\}$ denotes the sub-exponential norm, then

$$L_{s_d, q} = O(d \log(d)).$$

(iii) (Finite moment) If there is some $p \geq 2q$ and a finite constant K_p such that for each $1 \leq j \leq d$, $\mathbb{E}|x_j|^p \leq K_p$, then

$$L_{s_d, q} = O(d^{1+\frac{1}{2q}}).$$

(iv) For all three cases (i)–(iii), when $s = 2$, $L_{2, q} = O(d)$.

Proof of Lemma 5. We write $\mathbf{x} = \mathbf{x}_k$ to denote a generic covariate. By (29) and Hölder's inequality,

$$L_{s_d, q} \leq e \|\mathbf{x}\|_\infty \|\mathbf{x}\|_1 \leq e \|\mathbf{x}\|_\infty \|\mathbf{x}\|_{2q} \|\mathbf{x}\|_1 \|\mathbf{x}\|_{2q}.$$

The convexity of the function $t \mapsto t^{2q}$ and Jensen's inequality yield $|\mathbf{x}|_1^{2q} \leq d^{2q-1} \sum_{j=1}^d |x_j|^{2q}$ and

$$\mathbb{E}|\mathbf{x}|_1^{2q} \leq d^{2q-1} \sum_{j=1}^d \mathbb{E}|x_j|^{2q} \leq d^{2q} \max_{1 \leq j \leq d} \mathbb{E}|x_j|^{2q}.$$

Therefore, for all the three cases (i)–(iii),

$$\|\mathbf{x}\|_1 \|\mathbf{x}\|_{2q} = O(d).$$

Next, we study the order of $(\mathbb{E}[\|\mathbf{x}\|_\infty^{2q}])^{1/(2q)}$ for fixed $q \geq 2$.

(i) If each x_j is sub-Gaussian, then by Section 2.5 in Vershynin [2018], we have

$$(\mathbb{E}[\max_{1 \leq j \leq d} |x_j|^{2q}])^{1/(2q)} \leq K(\sqrt{\log(d)} + \sqrt{q}) = O(\sqrt{\log(d)}).$$

(ii) If each x_j is sub-exponential, then by Section 2.7 in Vershynin [2018], we obtain

$$(\mathbb{E}[\max_{1 \leq j \leq d} |x_j|^{2q}])^{1/(2q)} = O(K(\log(d) + \log(q))) = O(\log(d)).$$

(iii) If each x_j has the finite p -th moment for some $p \geq 2q$, then

$$\mathbb{E}[\max_{1 \leq j \leq d} |x_j|^{2q}] \leq \sum_{1 \leq j \leq d} \mathbb{E}[|x_j|^{2q}] \leq dK_q = O(d).$$

Finally, for case (iv) with $s_d = 2$, by (23),

$$L_{2, q} = \|\mathbf{x}\|_2^2 \|\mathbf{x}\|_{2q}^2.$$

By the convexity of the function $t \mapsto t^q$, we apply Jensen's inequality and obtain

$$|\mathbf{x}|_2^{2q} = \left(\sum_{j=1}^d x_j^2 \right)^q \leq d^{q-1} \sum_{j=1}^d |x_j|^{2q}.$$

Therefore, for $\mathbf{x}$ satisfying case (iii),

$$\mathbb{E}|\mathbf{x}|_2^{2q} \leq d^{q-1} \sum_{j=1}^d \mathbb{E}|x_j|^{2q} \leq d^q K_p, \tag{30}$$

which yields $L_{2,q} = O(d)$. For the cases (i) and (ii), by Sections 3.4 and 2.7 in Vershynin [2018], respectively, we obtain

$$\|\mathbf{x}\|_2 = O(K(\sqrt{d} + \sqrt{q})) = O(\sqrt{d}),$$

and

$$\|\mathbf{x}\|_2 = O(K(q\sqrt{d})) = O(\sqrt{d}),$$

both indicating $L_{2,q} = O(d)$. This completes the proof. $\square$

Lemma 6. *Consider the linear regression model in (18) and assume the conditions on ϵ and $\mathbf{x}$ therein are satisfied. Recall that $M_{s,q} = \|\nabla g(\beta^*, \xi)\|_q$ is defined in Assumption 3 for some $q \geq 2$. For the same four cases (i)–(iv) as in Lemma 5 and s_d defined in (7), $M_{s_d,q}$ is respectively equal to (i) $O(\sqrt{\log(d)})$, (ii) $O(\log(d))$, (iii) $O(d^{1/(2q)})$ and (iv) $O(\sqrt{d})$.*

Proof of Lemma 6. In the linear regression model, the stochastic gradient at the global minimum β^* can be rewritten into

$$\nabla g(\beta^*, \xi) = -(y - \mathbf{x}^\top \beta^*) = -\epsilon \mathbf{x}.$$

Since the noise ϵ is independent of the covariate vector $\mathbf{x}$, we obtain

$$\|\nabla g(\beta^*, \xi)\|_q = \|\epsilon\|_q \|\mathbf{x}\|_q.$$

By inequality (8), it suffices to bound $\|\mathbf{x}\|_\infty$. Since $\|\mathbf{x}\|_\infty \leq \|\mathbf{x}\|_2$, the same arguments in the proof of Lemma 5 carry over immediately. We omit the details here. $\square$

8.3 Some Useful Lemmas

Lemma 7 (Maximal inequality [Chernozhukov et al., 2015]). *Let $\mathbf{z}_1, \dots, \mathbf{z}_n$ be independent, d -dimensional random vectors. Denote the j -th element of $\mathbf{z}_i$ by z_{ij} , $1 \leq j \leq d$. Define $M := \max_{1 \leq i \leq n} \max_{1 \leq j \leq d} |z_{ij}|$ and $\sigma^2 := \max_{1 \leq j \leq d} \sum_{i=1}^n \mathbb{E}[z_{ij}^2]$. Then,*

$$\mathbb{E} \left[\max_{1 \leq j \leq d} \left| \sum_{i=1}^n (z_{ij} - \mathbb{E}[z_{ij}]) \right| \right] \lesssim \sigma \sqrt{\log(d)} + \sqrt{\mathbb{E}[M^2] \log(d)},$$

where the universal constant in $\lesssim$ is positive and independent of n and d .

Lemma 8 (L^q maximal inequality). *Let $\mathbf{x}_1, \dots, \mathbf{x}_n$ be independent, d -dimensional random vectors. Denote by x_{ij} the j -th element of $\mathbf{x}_i$, $1 \leq j \leq d$. Then,*

$$\left\| \max_{1 \leq j \leq d} \left| \sum_{i=1}^n (x_{ij} - \mathbb{E}[x_{ij}]) \right| \right\|_q^2 \leq e^2 (\max\{q, \log(d)\} - 1) \sum_{i=1}^n \left\| \max_{1 \leq j \leq d} |x_{ij} - \mathbb{E}[x_{ij}]| \right\|_q^2.$$

This moment inequality can be derived by repeatedly applying Lemma 2. It generalizes the maximal inequality for $\mathbb{E}[\max_{1 \leq j \leq d} |\sum_{i=1}^n (x_{ij} - \mathbb{E}[x_{ij}])|]$ in Chernozhukov et al. [2015], reproduced above as Lemma 7, to general q -th moments.

Proof of Lemma 8. One can assume that the independent random vectors $\mathbf{x}_1, \dots, \mathbf{x}_n$ have zero means. By repeatedly applying Lemma 2 and choosing $s = \log(d)$,

$$\begin{aligned} \left\| \mathbf{x}_1 + \dots + \mathbf{x}_n \right\|_\infty^2 &\leq \left\| \mathbf{x}_1 + \dots + \mathbf{x}_n \right\|_s^2 \\ &\leq \left\| \mathbf{x}_1 + \dots + \mathbf{x}_{n-1} \right\|_s^2 + (\max\{q, s\} - 1) \left\| \mathbf{x}_n \right\|_s^2 \\ &\leq (\max\{q, s\} - 1) \sum_{i=1}^n \left\| \mathbf{x}_i \right\|_s^2 \\ &\leq e^2 (\max\{q, \log(d)\} - 1) \sum_{i=1}^n \left\| \mathbf{x}_i \right\|_\infty^2. \end{aligned}$$

$\square$

Lemma 9 (Moment inequality [Li et al., 2024c]). *Let $q \geq 2$. For any two random vectors $\mathbf{x}$ and $\mathbf{y}$ in $\mathbb{R}^d$ with fixed $d \geq 1$, and let*

$$\Delta = \mathbb{E} \left| \|\mathbf{x} + \mathbf{y}\|_2^q - \|\mathbf{x}\|_2^q - q\|\mathbf{x}\|_2^{q-2} \mathbf{x}^\top \mathbf{y} \right|.$$

Then, the following inequalities holds:

(i)

$$\Delta \leq \mathbb{E}(\|\mathbf{x}\|_2 + \|\mathbf{y}\|_2)^q - \mathbb{E}\|\mathbf{x}\|_2^q - q\mathbb{E}(\|\mathbf{x}\|_2^{q-1} \|\mathbf{y}\|_2).$$

(ii)

$$\Delta \leq [(\mathbb{E}\|\mathbf{x}\|_2^q)^{1/q} + (\mathbb{E}\|\mathbf{y}\|_2^q)^{1/q}]^q - \mathbb{E}\|\mathbf{x}\|_2^q - q(\mathbb{E}\|\mathbf{x}\|_2^q)^{(q-1)/q} (\mathbb{E}\|\mathbf{y}\|_2^q)^{1/q}.$$

Lemma 10 (Equivalence of ℓ^s - ℓ^∞ -induced matrix norms). *For matrix $A \in \mathbb{R}^{d \times d}$, we have the equivalence of the ℓ^{s_d} -norm and ℓ^∞ -norm induced matrix norms as follows*

$$\frac{1}{e} \|A\|_\infty \leq \|A\|_{s_d} \leq e \|A\|_\infty, \quad (31)$$

where s_d is defined as (7) and $\|A\|_s = \max_{|\mathbf{x}|_s \neq 0} |A\mathbf{x}|_s / |\mathbf{x}|_s$. If in addition, A is symmetric, then

$$\frac{1}{e} \|A\|_1 = \frac{1}{e} \|A\|_\infty \leq \|A\|_{s_d} \leq e \|A\|_\infty = e \|A\|_1. \quad (32)$$

Proof of Lemma 10. By Horn and Johnson [1985], for any $1 \leq p \leq q \leq \infty$ and matrix $A \in \mathbb{R}^{d \times d}$,

$$d^{(1/q)-(1/p)} \|A\|_q \leq \|A\|_p \leq d^{(1/p)-(1/q)} \|A\|_q. \quad (33)$$

For $p = s$ and $q = \infty$, we obtain

$$d^{-1/s} \|A\|_\infty \leq \|A\|_s \leq d^{1/s} \|A\|_\infty. \quad (34)$$

Since $d^{1/s} \leq e$ by choosing $s = s_d$ in (7), we obtain (31).

For symmetric $A = (a_{ij})_{1 \leq i, j \leq d}$, $a_{ij} = a_{ji}$ for all i, j . Therefore,

$$\|A\|_1 = \max_{1 \leq j \leq d} \sum_{i=1}^d |a_{ij}| = \max_{1 \leq i \leq d} \sum_{j=1}^d |a_{ij}| = \|A\|_\infty. \quad (35)$$

This completes the proof. $\square$

8.4 Proofs for Section 2

Derivation of (15): Since s is an even integer, we can write

$$f(\alpha) := |\mathbf{x} - \alpha \mathbf{z}|_s^q = \left\{ \sum_{i=1}^d (x_i - \alpha z_i)^s \right\}^{\frac{q}{s}}. \quad (36)$$

Taking the derivative with respect to α , we obtain

$$\begin{aligned} f'(\alpha) &:= \frac{d}{d\alpha} f(\alpha) = \frac{q}{s} \left\{ \sum_{i=1}^d (x_i - \alpha z_i)^s \right\}^{\frac{q}{s}-1} \sum_{i=1}^d \frac{d}{d\alpha} (x_i - \alpha z_i)^s \\ &= \frac{q}{s} \left\{ \sum_{i=1}^d (x_i - \alpha z_i)^s \right\}^{\frac{q}{s}-1} \sum_{i=1}^d s (x_i - \alpha z_i)^{s-1} (-z_i) \\ &= -q \left\{ \sum_{i=1}^d (x_i - \alpha z_i)^s \right\}^{\frac{q}{s}-1} \sum_{i=1}^d (x_i - \alpha z_i)^{s-1} z_i. \end{aligned} \quad (37)$$

Therefore,

$$\begin{aligned} f'(0) &= -q \left\{ \sum_{i=1}^d x_i^s \right\}^{\frac{q}{s}-1} \sum_{i=1}^d x_i^{s-1} z_i \\ &= -q |\mathbf{x}|_s^{q-s} \sum_{i=1}^d x_i^{s-1} z_i. \end{aligned} \quad (38)$$

A first-order Taylor expansion yields then (15).

Proof of Lemma 1. Recall that we have defined

$$f(\alpha) = |\mathbf{x} - \alpha \mathbf{z}|_s^q = \left\{ \sum_{i=1}^d (x_i - \alpha z_i)^s \right\}^{\frac{q}{s}}.$$

A second order Taylor expansion gives $f(\alpha) = f(0) + \alpha f'(0) + \frac{1}{2} \alpha^2 f''(\eta)$ for some $\eta \in [0, \alpha]$. It suffices to bound $\sup_{u \in [0, \alpha]} |f''(u)|$. Defining

$$M(u) := \sum_{i=1}^d (x_i - u z_i)^s = |\mathbf{x} - u \mathbf{z}|_s^s, \quad (39)$$

we have $f(u) = [M(u)]^{\frac{q}{s}}$,

$$M'(u) = -s \sum_{i=1}^d (x_i - u z_i)^{s-1} z_i, \quad (40)$$

$$M''(u) = s(s-1) \sum_{i=1}^d (x_i - u z_i)^{s-2} z_i^2, \quad (41)$$

and the first two derivatives of $f(u)$ can be respectively expressed by

$$f'(u) = \frac{q}{s} [M(u)]^{\frac{q}{s}-1} M'(u), \quad (42)$$

$$f''(u) = \frac{q}{s} \left(\frac{q}{s} - 1 \right) [M(u)]^{\frac{q}{s}-2} [M'(u)]^2 + M''(u) \frac{q}{s} [M(u)]^{\frac{q}{s}-1}. \quad (43)$$

Since s is an even integer, it follows from Hölder's inequality that

$$\begin{aligned} [M'(u)]^2 &= s^2 \left(\sum_{i=1}^d (x_i - u z_i)^{s-1} z_i \right)^2 \\ &\leq s^2 \left(\left(\sum_{i=1}^d (x_i - u z_i)^s \right)^{\frac{s-1}{s}} \left(\sum_{i=1}^d z_i^s \right)^{1/s} \right)^2 \\ &= s^2 |\mathbf{x} - u \mathbf{z}|_s^{2(s-1)} |\mathbf{z}|_s^2. \end{aligned} \quad (44)$$

By applying Hölder's inequality again, we obtain

$$\begin{aligned} |M''(u)| &\leq s(s-1) \left(\sum_{i=1}^d (x_i - u z_i)^s \right)^{\frac{s-2}{s}} \left(\sum_{i=1}^d z_i^s \right)^{2/s} \\ &= s(s-1) |\mathbf{x} - u \mathbf{z}|_s^{s-2} |\mathbf{z}|_s^2. \end{aligned} \quad (45)$$

By the two results above, we have

$$\begin{aligned} |f''(u)| &= \left| \frac{q}{s} \left(\frac{q}{s} - 1 \right) |\mathbf{x} - u \mathbf{z}|_s^{q-2s} [M'(u)]^2 + M''(u) \frac{q}{s} |\mathbf{x} - u \mathbf{z}|_s^{q-s} \right| \\ &\leq q|q-s| \cdot |\mathbf{x} - u \mathbf{z}|_s^{q-2} |\mathbf{z}|_s^2 + q(s-1) |\mathbf{x} - u \mathbf{z}|_s^{q-2} |\mathbf{z}|_s^2 \\ &\leq q[|q-s| + (s-1)] (|\mathbf{x}|_s + |\mathbf{z}|_s)^{q-2} |\mathbf{z}|_s^2. \end{aligned} \quad (46)$$

Since $u \in [0, \alpha]$, it follows that

$$\sup_{u \in [0, \alpha]} |f''(u)| \leq q[|q-s| + (s-1)] (|\mathbf{x}|_s + \alpha |\mathbf{z}|_s)^{q-2} |\mathbf{z}|_s^2. \quad (47)$$

This completes the proof. $\square$

Existence of solution to (13): To see the existence of the solution $\alpha_{s,q}$ in

$$1 - q\mu\alpha + \frac{q[|q-s| + (s-1)] L_{s,q}^2}{2} \alpha^2 (1 + \alpha L_{s,q})^{q-2} = 1.$$

denote the function $\alpha \mapsto F(\alpha) = -\mu + c\alpha(1+L)^{q-2}$ for the constant $c = [|q-s| + (s-1)] L^2/2 > 0$ and $L = L_{s,q}$. For any $q \geq 2$, and any $\alpha > 0$, $F'(\alpha) = c[(1+L\alpha)^{q-2} + \alpha(q-2)L(1+L\alpha)^{q-3}] > 0$, proving that $F(\alpha)$ is strictly increasing on $\alpha > 0$. Since $F(0) = -\mu < 0$ and $F(\infty) = +\infty$, the unique root to $F(\alpha) = 0$ exists.

Proof of Lemma 2. Define $\varphi(t) = \|\mathbf{x} + t\mathbf{y}\|_s^2$ for $t \in [0, 1]$. Then

$$\begin{aligned}\varphi'(t) &= \frac{2}{q} \left[\mathbb{E} \left\{ \sum_{j=1}^d (x_j + ty_j)^s \right\}^{q/s} \right]^{2/q-1} \frac{q}{s} \mathbb{E} \left[\left\{ \sum_{j=1}^d (x_j + ty_j)^s \right\}^{q/s-1} \sum_{j=1}^d s(x_j + ty_j)^{s-1} y_j \right] \\ &= 2 \left[\mathbb{E} \left\{ \sum_{j=1}^d (x_j + ty_j)^s \right\}^{q/s} \right]^{2/q-1} \mathbb{E} \left[\left\{ \sum_{j=1}^d (x_j + ty_j)^s \right\}^{q/s-1} \sum_{j=1}^d (x_j + ty_j)^{s-1} y_j \right]\end{aligned}$$

and

$$\begin{aligned}\varphi''(t) &= 2 \left(\frac{2}{q} - 1 \right) \left[\mathbb{E} \left\{ \sum_{j=1}^d (x_j + ty_j)^s \right\}^{q/s} \right]^{2/q-2} \\ &\quad \cdot \frac{q}{s} \cdot s \left| \mathbb{E} \left[\left\{ \sum_{j=1}^d (x_j + ty_j)^s \right\}^{q/s-1} \sum_{j=1}^d (x_j + ty_j)^{s-1} y_j \right] \right|^2 \\ &\quad + 2 \left[\mathbb{E} \left\{ \sum_{j=1}^d (x_j + ty_j)^s \right\}^{q/s} \right]^{2/q-1} \\ &\quad \cdot \mathbb{E} \left[\left(\frac{q}{s} - 1 \right) \left\{ \sum_{j=1}^d (x_j + ty_j)^s \right\}^{q/s-2} s \left\{ \sum_{j=1}^d (x_j + ty_j)^{s-1} y_j \right\}^2 \right] \\ &\quad + 2 \left[\mathbb{E} \left\{ \sum_{j=1}^d (x_j + ty_j)^s \right\}^{q/s} \right]^{2/q-1} \mathbb{E} \left[\left\{ \sum_{j=1}^d (x_j + ty_j)^s \right\}^{q/s-1} (s-1) \sum_{j=1}^d (x_j + ty_j)^{s-2} y_j^2 \right] \\ &=: \Delta_1(t) + \Delta_2(t) + \Delta_3(t)\end{aligned}$$

Since $q \geq 2$, $\Delta_1(t) \leq 0$.

Case I. If $q/s - 1 \leq 0$, then $\Delta_2(t) \leq 0$ and $\varphi''(t) \leq \Delta_3(t)$. By Hölder's inequality,

$$\sum_{j=1}^d (x_j + ty_j)^{s-2} y_j^2 \leq \left\{ \sum_{j=1}^d (x_j + ty_j)^s \right\}^{(s-2)/s} \left(\sum_{j=1}^d y_j^s \right)^{2/s}$$

Consequently,

$$\begin{aligned}\Delta_3(t) &\leq 2(s-1) \left[\mathbb{E} \left\{ \sum_{j=1}^d (x_j + ty_j)^s \right\}^{q/s} \right]^{2/q-1} \\ &\quad \cdot \mathbb{E} \left[\left\{ \sum_{j=1}^d (x_j + ty_j)^s \right\}^{q/s-1} \left\{ \sum_{j=1}^d (x_j + ty_j)^s \right\}^{(s-2)/s} \left(\sum_{j=1}^d y_j^s \right)^{2/s} \right] \\ &= 2(s-1) \left[\mathbb{E} \left\{ \sum_{j=1}^d (x_j + ty_j)^s \right\}^{q/s} \right]^{2/q-1} \mathbb{E} \left[\left\{ \sum_{j=1}^d (x_j + ty_j)^s \right\}^{(q-2)/s} \left(\sum_{j=1}^d y_j^s \right)^{2/s} \right] \\ &= 2(s-1) \|\mathbf{x} + t\mathbf{y}\|_s^{2-q} \mathbb{E} \left[\left\{ \sum_{j=1}^d (x_j + ty_j)^s \right\}^{(q-2)/s} \left(\sum_{j=1}^d y_j^s \right)^{2/s} \right] \\ &\leq 2(s-1) \|\mathbf{x} + t\mathbf{y}\|_s^{2-q} \|\mathbf{x} + t\mathbf{y}\|_s^{q-2} \|\mathbf{y}\|_s^2 \\ &= 2(s-1) \|\mathbf{y}\|_s^2.\end{aligned}$$

Case II. If $q/s - 1 > 0$, by Hölder's inequality,

$$\begin{aligned}\left\{ \sum_{j=1}^d (x_j + ty_j)^{s-1} y_j \right\}^2 &= \left\{ \sum_{j=1}^d (x_j + ty_j)^{s/2} (x_j + ty_j)^{s/2-1} y_j \right\}^2 \\ &\leq \sum_{j=1}^d (x_j + ty_j)^s \sum_{j=1}^d (x_j + ty_j)^{s-2} y_j^2.\end{aligned}$$

Therefore,

$$\Delta_2(t) \leq \Delta_3(t) \frac{q-s}{s-1}$$

and

$$\varphi''(t) \leq \Delta_2(t) + \Delta_3(t) \leq \Delta_3(t) \frac{q-1}{s-1} \leq 2(q-1) \|\mathbf{y}\|_s^2.$$

Then, we have

$$\begin{aligned} \|\mathbf{x} + \mathbf{y}\|_s^2 &= \varphi(1) = \varphi(0) + \varphi'(0) + \int_0^1 (1-t) \varphi''(t) dt \\ &\leq \|\mathbf{x}\|_s^2 + 2\|\mathbf{x}\|_s^{2-q} \mathbb{E} \left(|\mathbf{x}|_s^{q-s} \sum_{j=1}^d x_j^{s-1} y_j \right) + (\max\{q, s\} - 1) \|\mathbf{y}\|_s^2. \end{aligned}$$

□

Proof of Theorem 1. Consider the iterated random function

$$F : \mathbb{R}^d \times \mathbb{R} \mapsto \mathbb{R}, \quad (\boldsymbol{\beta}, \boldsymbol{\xi}) \mapsto F_{\boldsymbol{\xi}}(\boldsymbol{\beta}) = \boldsymbol{\beta} - \alpha \nabla g(\boldsymbol{\beta}, \boldsymbol{\xi}). \quad (48)$$

To prove GMC in Theorem 1, it suffices to show that, for some $q \geq 2$ and even integer $s \geq 2$, for any fixed vectors $\boldsymbol{\beta}, \boldsymbol{\beta}' \in \mathbb{R}^d$,

$$\|F_{\boldsymbol{\xi}}(\boldsymbol{\beta}) - F_{\boldsymbol{\xi}}(\boldsymbol{\beta}')\|_s \leq r_{\alpha, s, q} \|\boldsymbol{\beta} - \boldsymbol{\beta}'\|_s.$$

Recall the inequality in Lemma 2. For $\mathbf{x}$ and $\mathbf{y}$ therein, we choose them to be $\mathbf{x} = \boldsymbol{\beta} - \boldsymbol{\beta}'$ and $\mathbf{y} = -\alpha(\nabla g(\boldsymbol{\beta}, \boldsymbol{\xi}) - \nabla g(\boldsymbol{\beta}', \boldsymbol{\xi}))$ respectively. Then, it directly follows from Lemma 2 that

$$\begin{aligned} &\|F_{\boldsymbol{\xi}}(\boldsymbol{\beta}) - F_{\boldsymbol{\xi}}(\boldsymbol{\beta}')\|_s^2 \\ &\leq \|\boldsymbol{\beta} - \boldsymbol{\beta}'\|_s^2 - 2\alpha \|\boldsymbol{\beta} - \boldsymbol{\beta}'\|_s^{2-q} \mathbb{E} \left[\|\boldsymbol{\beta} - \boldsymbol{\beta}'\|_s^{q-s} \langle (\boldsymbol{\beta} - \boldsymbol{\beta}')^{s-1}, \nabla g(\boldsymbol{\beta}, \boldsymbol{\xi}) - \nabla g(\boldsymbol{\beta}', \boldsymbol{\xi}) \rangle \right] \\ &\quad + \alpha^2 (\max\{q, s\} - 1) \|\nabla g(\boldsymbol{\beta}, \boldsymbol{\xi}) - \nabla g(\boldsymbol{\beta}', \boldsymbol{\xi})\|_s^2 \\ &= \|\boldsymbol{\beta} - \boldsymbol{\beta}'\|_s^2 - 2\alpha \|\boldsymbol{\beta} - \boldsymbol{\beta}'\|_s^{2-s} \langle (\boldsymbol{\beta} - \boldsymbol{\beta}')^{s-1}, G(\boldsymbol{\beta}) - G(\boldsymbol{\beta}') \rangle \\ &\quad + \alpha^2 (\max\{q, s\} - 1) \|\nabla g(\boldsymbol{\beta}, \boldsymbol{\xi}) - \nabla g(\boldsymbol{\beta}', \boldsymbol{\xi})\|_s^2. \end{aligned}$$

This along with Assumptions 2 and 3 yields

$$\|F_{\boldsymbol{\xi}}(\boldsymbol{\beta}) - F_{\boldsymbol{\xi}}(\boldsymbol{\beta}')\|_s^2 \leq (1 - 2\alpha\mu + \alpha^2(\max\{q, s\} - 1)L_{s, q}^2) \|\boldsymbol{\beta} - \boldsymbol{\beta}'\|_s^2,$$

which completes the proof. □

8.5 Proofs for Section 3.1

Proof of Proposition 1. Recall (17) and let $\nabla G(\boldsymbol{\beta}) = (\nabla G_1(\boldsymbol{\beta}), \dots, \nabla G_d(\boldsymbol{\beta}))^\top$ with

$$\nabla G_i(\boldsymbol{\beta}) = \partial G(\boldsymbol{\beta}) / \partial \beta_i = (\mathbb{E}[\nabla g(\boldsymbol{\beta}, \boldsymbol{\xi})])_i, \quad i = 1, \dots, d. \quad (49)$$

Since the random samples $\boldsymbol{\xi}_k$, $k \geq 1$, are independent, it follows that for the k -th iteration, $\boldsymbol{\xi}_k$ is independent of $\boldsymbol{\beta}_{k-1}$. Then, by the tower rule, for all $k \geq 1$,

$$\mathbb{E}_{\boldsymbol{\xi}} [\nabla g(\boldsymbol{\beta}_{k-1}, \boldsymbol{\xi}_k) - \nabla G(\boldsymbol{\beta}_{k-1}) \mid \boldsymbol{\beta}_{k-1}] = \mathbb{E}_{\boldsymbol{\xi}} [\nabla g(\boldsymbol{\beta}_{k-1}, \boldsymbol{\xi}_k) - \nabla G(\boldsymbol{\beta}_{k-1})] = 0. \quad (50)$$

Therefore, by applying the high-dimensional moment inequality (16) in Lemma 2, we obtain

$$\begin{aligned} \|\boldsymbol{\beta}_k - \boldsymbol{\beta}^*\|_s^2 &\leq \|\boldsymbol{\beta}_{k-1} - \boldsymbol{\beta}^* - \alpha \nabla G(\boldsymbol{\beta}_{k-1})\|_s^2 \\ &\quad + (\max\{q, s\} - 1) \alpha^2 \|\nabla g(\boldsymbol{\beta}_{k-1}, \boldsymbol{\xi}_k) - \nabla G(\boldsymbol{\beta}_{k-1})\|_s^2. \end{aligned} \quad (51)$$

For the second part in (51), noting that $\nabla G(\beta^*) = 0$, by the triangle inequality, we have

$$\begin{aligned}
& \|\nabla g(\beta_{k-1}, \xi_k) - \nabla G(\beta_{k-1})\|_s^2 \\
& \leq \left(\|\nabla g(\beta_{k-1}, \xi_k) - \nabla g(\beta^*, \xi_k)\|_s + \|\nabla G(\beta_{k-1}) - \nabla G(\beta^*)\|_s + \|\nabla g(\beta^*, \xi_k)\|_s \right)^2 \\
& \leq 3\|\nabla g(\beta_{k-1}, \xi_k) - \nabla g(\beta^*, \xi_k)\|_s^2 + 3\|\nabla G(\beta_{k-1}) - \nabla G(\beta^*)\|_s^2 + 3\|\nabla g(\beta^*, \xi_k)\|_s^2.
\end{aligned} \tag{52}$$

Since $\|\cdot\|_s$ is a convex function for $s \geq 1$, we have $|\mathbb{E}[\cdot]|_s \leq \mathbb{E}[\|\cdot\|_s]$. Thus, for all $q \geq 1$, by Jensen's inequality, we can bound

$$\begin{aligned}
\|\nabla G(\beta_{k-1}) - \nabla G(\beta^*)\|_s &= |\mathbb{E}_\xi[\nabla g(\beta_{k-1}, \xi_k) - \nabla g(\beta^*, \xi_k)]|_s \\
&\leq \mathbb{E}_\xi[\|\nabla g(\beta_{k-1}, \xi_k) - \nabla g(\beta^*, \xi_k)\|_s] \\
&\leq \left(\mathbb{E}_\xi[\|\nabla g(\beta_{k-1}, \xi_k) - \nabla g(\beta^*, \xi_k)\|_s^q] \right)^{1/q}.
\end{aligned} \tag{53}$$

This along with Assumption 3 yields

$$\begin{aligned}
\|\nabla G(\beta_{k-1}) - \nabla G(\beta^*)\|_q &\leq \left(\mathbb{E}_\beta \mathbb{E}_\xi \|\nabla g(\beta_{k-1}, \xi_k) - \nabla g(\beta^*, \xi_k)\|_s^q \right)^{1/q} \\
&= \left\| \|\nabla g(\beta_{k-1}, \xi_k) - \nabla g(\beta^*, \xi_k)\|_s \right\|_q \\
&\leq L_{s,q} \|\beta_{k-1} - \beta^*\|_q.
\end{aligned} \tag{54}$$

Inserting this result back into (52), we obtain a bound for the second term in (51) using

$$\|\nabla g(\beta_{k-1}, \xi_k) - \nabla G(\beta_{k-1})\|_s^2 \leq 6L_{s,q}^2 \|\beta_{k-1} - \beta^*\|_q^2 + 3\|\nabla g(\beta^*, \xi_k)\|_s^2. \tag{55}$$

For the first term in (51), by applying Lemma 2 again, it follows from Assumptions 2 and 3 that

$$\begin{aligned}
& \|\beta_{k-1} - \beta^* - \alpha \nabla G(\beta_{k-1})\|_q^2 \\
& \leq \|\beta_{k-1} - \beta^*\|_q^2 - 2\alpha \|\beta_{k-1} - \beta^*\|_q^{2-q} \mathbb{E} \left(\|\beta_{k-1} - \beta^*\|_s^{q-s} \sum_{j=1}^d (\beta_{k-1} - \beta^*)_j^{s-1} \nabla G_j(\beta_{k-1}) \right) \\
& \quad + \alpha^2 (\max\{q, s\} - 1) \|\nabla G(\beta_{k-1}) - \nabla G(\beta^*)\|_q^2 \\
& \leq (1 - 2\alpha\mu + \alpha^2 (\max\{q, s\} - 1) L_{s,q}^2) \|\beta_{k-1} - \beta^*\|_q^2.
\end{aligned} \tag{56}$$

Inserting this inequality and (55) into (51), we obtain the inequality

$$\begin{aligned}
\|\beta_k - \beta^*\|_q^2 &\leq (1 - 2\alpha\mu + 7(\max\{q, s\} - 1)\alpha^2 L_{s,q}^2) \|\beta_{k-1} - \beta^*\|_q^2 \\
&\quad + 3(\max\{q, s\} - 1)\alpha^2 \|\nabla g(\beta^*, \xi_k)\|_s^2.
\end{aligned}$$

The desired result is achieved since $\|\nabla g(\beta^*, \xi_k)\|_s \leq M_{s,q}$ by Assumption 3. As a special case, for the stationary SGD iterates $\beta_k^\circ \sim \pi_\alpha$, $k \geq 1$, we obtain the same result. $\square$

Proof of Theorem 2. First, we denote the contraction constant in Proposition 1 as follows

$$\tilde{r}_{\alpha,s,q} := 1 - 2\alpha\mu + 7(\max\{q, s\} - 1)\alpha^2 L_{s,q}^2. \tag{57}$$

Given the range of the constant learning rate α , we have $\tilde{r}_{\alpha,s,q} < 1$. Moreover, notice that

$$3(\max\{q, s\} - 1)\alpha^2 \|\nabla g(\beta^*, \xi_k)\|_s^2 = O(\max\{q, s\}\alpha^2 M_{s,q}^2). \tag{58}$$

Therefore, for the stationary SGD iterates $\beta_k^\circ \sim \pi_\alpha$, by Proposition 1, we can obtain

$$\|\beta_k^\circ - \beta^*\|_q^2 \leq \tilde{r}_{\alpha,s,q} \|\beta_{k-1}^\circ - \beta^*\|_q^2 + O(\max\{q, s\}\alpha^2 M_{s,q}^2). \tag{59}$$

Since the SGD iterates β_k° satisfy the geometric-moment contraction in Theorem 1, following Remark 2 in Wu and Shao [2004], the recursion $\beta_k^\circ = \beta_{k-1}^\circ - \alpha \nabla g(\beta_{k-1}^\circ, \xi_k)$ also holds for $k \leq 0$. Thus, we can recursively apply the inequality above and achieve

$$\begin{aligned} \|\beta_k^\circ - \beta^*\|_s^2 &\leq O(\max\{q, s\} \alpha^2 M_{s,q}^2) \cdot \sum_{i=0}^{\infty} \tilde{r}_{\alpha,s,q}^i \\ &= \frac{1}{1 - \tilde{r}_{\alpha,s,q}} O(\max\{q, s\} \alpha^2 M_{s,q}^2) \\ &= O(\max\{q, s\} \alpha M_{s,q}^2). \end{aligned} \quad (60)$$

This finishes the proof for the stationary SGD sequence.

Furthermore, for the general SGD iterates β_k in (2) that may not have the stationary initialization, we apply the geometric-moment contraction in Theorem 1 and obtain

$$\begin{aligned} \|\beta_k - \beta^*\|_s &\leq \|\beta_k - \beta_k^\circ\|_s + \|\beta_k^\circ - \beta^*\|_s \\ &\leq r_{\alpha,s,q}^k \|\beta_0 - \beta_0^\circ\|_s + O(M_{s,q} \sqrt{\max\{q, s\} \alpha}), \end{aligned} \quad (61)$$

which completes the proof. $\square$

8.6 Functional Dependence Measure in Time Series

The functional dependence measure in time series [Wu, 2005] is a key concept in our analysis. For that we view the high-dimensional SGD iterates $\{\beta_k\}_{k \in \mathbb{N}}$ as a nonlinear autoregressive (AR) process. Recall that $\xi_k, k \in \mathbb{Z}$, are i.i.d. Define the shift process $\mathcal{F}_k = (\xi_k, \xi_{k-1}, \dots)$ and its coupled version $\mathcal{F}_{k,\{l\}} = (\xi_k, \dots, \xi_{l+1}, \xi'_l, \xi_{l-1}, \dots)$, $l \leq k$, where ξ'_l is an i.i.d. copy of ξ_l .

The stationary sequence $\{\beta_k^\circ\}_{k \in \mathbb{Z}}$ can be represented by a functional system

$$\beta_k^\circ = h_\alpha(\xi_k, \xi_{k-1}, \dots) = h_\alpha(\mathcal{F}_k), \quad k \geq 1, \quad (62)$$

where h_α is a measurable function that depends on α [Wiener, 1958, Wu, 2005]. Define the coupled version of β_k° by

$$\beta_{k,\{l\}}^\circ = h_\alpha(\xi_k, \dots, \xi_{l+1}, \xi'_l, \xi_{l-1}, \dots) = h_\alpha(\mathcal{F}_{k,\{l\}}), \quad l \leq k. \quad (63)$$

The next lemma provides a bound for the functional dependence measure $\|\beta_k^\circ - \beta_{k,\{l\}}^\circ\|_s$. It is later used to derive the moment bounds and the tail probability of the ASGD iterates.

Lemma 11. *Consider the stationary SGD sequence $\{\beta_k^\circ\}_{k \geq 1}$. Suppose that Assumptions 2 and 3 hold with some $q \geq 2$ and even integer $s \geq 2$. Then, for all $k \geq 1$ and $l \leq k$, we have*

$$\|\beta_k^\circ - \beta_{k,\{l\}}^\circ\|_s^2 \leq 4\alpha^2 (1 - 2\alpha\mu + 7(\max\{q, s\} - 1)\alpha^2 L_{s,q}^2)^{k-l} (L_{s,q}^2 \|\beta_{l-1}^\circ - \beta^*\|_s^2 + M_{s,q}^2).$$

Proof of Lemma 11. By applying Lemma 2, it follows from similar arguments as in the proof of Proposition 1 that, for each $l \leq k-1$,

$$\|\beta_k^\circ - \beta_{k,\{l\}}^\circ\|_s^2 \leq (1 - 2\alpha\mu + 7(\max\{q, s\} - 1)\alpha^2 L_{s,q}^2)^{k-l} \|\beta_l^\circ - \beta_{l,\{l\}}^\circ\|_s^2. \quad (64)$$

By Assumption 3, for all $l \geq 1$,

$$\begin{aligned} \|\nabla g(\beta_{l-1}^\circ, \xi_l)\|_s^2 &\leq 2\|\nabla g(\beta_{l-1}^\circ, \xi_l) - \nabla g(\beta^*, \xi_l)\|_s^2 + 2\|\nabla g(\beta^*, \xi_l)\|_s^2 \\ &\leq 2L_{s,q}^2 \|\beta_{l-1}^\circ - \beta^*\|_s^2 + 2M_{s,q}^2, \end{aligned} \quad (65)$$

which yields

$$\begin{aligned} \|\beta_l^\circ - \beta_{l,\{l\}}^\circ\|_s^2 &= \alpha^2 \|\nabla g(\beta_{l-1}^\circ, \xi_l) - \nabla g(\beta_{l-1}^\circ, \xi'_l)\|_s^2 \\ &\leq \alpha^2 (2\|\nabla g(\beta_{l-1}^\circ, \xi_l)\|_s^2 + 2\|\nabla g(\beta_{l-1}^\circ, \xi'_l)\|_s^2) \\ &\leq 4\alpha^2 (L_{s,q}^2 \|\beta_{l-1}^\circ - \beta^*\|_s^2 + M_{s,q}^2) \end{aligned} \quad (66)$$

Recall $\|\nabla g(\beta^*, \xi_k)\|_s \leq M_{s,q}$ by Assumption 3. Therefore,

$$\begin{aligned} \|\beta_k^\circ - \beta_{k,\{l\}}^\circ\|_s^2 &\leq 4\alpha^2 (1 - 2\alpha\mu + 7(\max\{q, s\} - 1)\alpha^2 L_{s,q}^2)^{k-l} \\ &\quad \cdot (L_{s,q}^2 \|\beta_{l-1}^\circ - \beta^*\|_s^2 + M_{s,q}^2). \end{aligned} \quad (67)$$

This completes the proof. $\square$

8.7 Proofs for Section 3.2

In this section, we provide the proofs for the convergence results of ASGD in Section 3.2, which can be decomposed into the proofs for Theorems 6 to 8 in Section 6.

Proof of Theorem 7. Recall the i.i.d. random samples $\xi_k = (y_k, \mathbf{x}_k)$, the filtration $\mathcal{F}_k = (\xi_k, \xi_{k-1}, \dots)$ and its coupled version $\mathcal{F}_{k,\{l\}} = (\xi_k, \dots, \xi_{l+1}, \xi'_l, \xi_{l-1}, \dots)$, $l \leq k$, where ξ'_l is an i.i.d. copy of ξ_l . Following Wu [2005], we introduce the projection operator

$$\mathcal{P}_l[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}_l] - \mathbb{E}[\cdot \mid \mathcal{F}_{l-1}].$$

Then, we can rewrite the centered ASGD into

$$\bar{\beta}_k^\circ - \mathbb{E}[\bar{\beta}_k^\circ] = \frac{1}{k} \sum_{i=1}^k \sum_{l=0}^{i-1} \mathcal{P}_{i-l}(\beta_i^\circ) = \frac{1}{k} \sum_{l=0}^{k-1} \sum_{i=l+1}^k \mathcal{P}_{i-l}(\beta_i^\circ). \quad (68)$$

Since $\{\mathcal{P}_{i-l}(\beta_i^\circ)\}_{i \geq l+1}$ is a sequence of martingale differences over i for each $l = 0, 1, \dots, k-1$, following Lemma D.2 in Zhang and Wu [2021] and triangle inequality, we can obtain

$$\begin{aligned} \|\bar{\beta}_k^\circ - \mathbb{E}[\bar{\beta}_k^\circ]\|_q &= \left\| \frac{1}{k} \sum_{l=0}^{k-1} \sum_{i=l+1}^k \mathcal{P}_{i-l}(\beta_i^\circ) \right\|_q \\ &\leq \frac{1}{k} \sum_{l=0}^{k-1} \left\| \sum_{i=l+1}^k \mathcal{P}_{i-l}(\beta_i^\circ) \right\|_q \\ &\leq \frac{1}{k} \sum_{l=0}^{k-1} \left(c_q \cdot s \sum_{i=l+1}^k \|\mathcal{P}_{i-l}(\beta_i^\circ)\|_q^2 \right)^{1/2}. \end{aligned} \quad (69)$$

By Theorem 1 in Wu [2005], we have

$$\|\mathcal{P}_{i-l}(\beta_i^\circ)\|_q \leq \|\beta_i^\circ - \beta_{i,\{i-l\}}^\circ\|_q. \quad (70)$$

This along with Lemma 11 and definition of $\tilde{r}_{\alpha,s,q}$ in (58) yields

$$\begin{aligned} \|\mathcal{P}_{i-l}(\beta_i^\circ)\|_q^2 &\leq 4\alpha^2 (1 - 2\alpha\mu + 7(\max\{q, s\} - 1)\alpha^2 L_{s,q}^2)^l \\ &\quad \cdot \left(L_{s,q}^2 \|\beta_{i-l-1}^\circ - \beta^*\|_q^2 + M_{s,q}^2 \right) \\ &= 4\alpha^2 \tilde{r}_{\alpha,s,q}^l \left(L_{s,q}^2 \|\beta_{i-l-1}^\circ - \beta^*\|_q^2 + M_{s,q}^2 \right). \end{aligned} \quad (71)$$

Recall $r_{\alpha,s,q}$ in (6) and $\tilde{r}_{\alpha,s,q}$ in (58). For some constant $\omega > 0$ such that

$$\omega \leq \min \left\{ \frac{1}{\alpha}, 2\mu - 7(\max\{q, s\} - 1)\alpha L_{s,q}^2 \right\}, \quad (72)$$

we have $1 - \omega\alpha \geq 0$ and

$$r_{\alpha,s,q} \leq \tilde{r}_{\alpha,s,q} \leq 1 - \omega\alpha < 1. \quad (73)$$

Consequently, we can further bound (69) by

$$\begin{aligned} &\|\bar{\beta}_k^\circ - \mathbb{E}[\bar{\beta}_k^\circ]\|_q \\ &\leq \frac{1}{k} \sum_{l=0}^{k-1} \left[4c_q s \alpha^2 (1 - \omega\alpha)^l \sum_{i=l+1}^k \left(L_{s,q}^2 \|\beta_{i-l-1}^\circ - \beta^*\|_q^2 + M_{s,q}^2 \right) \right]^{1/2} \\ &= \frac{1}{k} \sum_{l=0}^{k-1} \left[4c_q s \alpha^2 (1 - \omega\alpha)^l \left(L_{s,q}^2 \sum_{i=l+1}^k \|\beta_{i-l-1}^\circ - \beta^*\|_q^2 + (k-l)M_{s,q}^2 \right) \right]^{1/2} \\ &\leq \frac{1}{k} \sum_{l=0}^{k-1} \left[2\alpha \sqrt{c_q s} (1 - \omega\alpha)^{l/2} L_{s,q} \sqrt{\sum_{i=l+1}^k \|\beta_{i-l-1}^\circ - \beta^*\|_q^2} \right] \\ &\quad + \frac{1}{k} \sum_{l=0}^{k-1} \left[2\alpha \sqrt{c_q s} (1 - \omega\alpha)^{l/2} \sqrt{(k-l)M_{s,q}} \right] =: \mathbf{I}_1 + \mathbf{I}_2. \end{aligned} \quad (74)$$

For the term I_1 , it follows from Theorem 2 and expression (58) that

$$\begin{aligned} \sum_{i=l+1}^k \|\beta_{i-l-1}^\circ - \beta^*\|_q^2 &\leq \sum_{i=l+1}^k \left(6M_{s,q}^2 (\max\{q, s\} - 1) \alpha \right) \\ &= 6\alpha(k-l)(\max\{q, s\} - 1)M_{s,q}^2. \end{aligned} \quad (75)$$

Inserting this back into (74) gives

$$\begin{aligned} I_1 &\leq \frac{2\alpha\sqrt{c_q s}L_{s,q}}{k} \sum_{l=0}^{k-1} (1-\omega\alpha)^{l/2} M_{s,q} \sqrt{6\alpha(k-l)(\max\{q, s\} - 1)} \\ &\leq \sqrt{c_q s}L_{s,q} \cdot \frac{c_1\sqrt{\alpha}}{\sqrt{k}} M_{s,q} \sqrt{\max\{q, s\} - 1}, \end{aligned} \quad (76)$$

for some constant $c_1 > 0$, where the last inequality is due to

$$\sum_{l=0}^{k-1} (1-\omega\alpha)^{l/2} \sqrt{k-l} \leq \sqrt{k} \sum_{l=0}^{k-1} (1-\omega\alpha)^{l/2} = O\left(\frac{\sqrt{k}}{\omega\alpha}\right). \quad (77)$$

Similarly, for some constant $c_2 > 0$,

$$I_2 \leq \frac{c_2\sqrt{c_q s}}{\sqrt{k}} M_{s,q}. \quad (78)$$

Combining the results of I_1 and I_2 , we obtain the claimed inequality

$$\|\bar{\beta}_k^\circ - \mathbb{E}[\bar{\beta}_k^\circ]\|_q \leq \sqrt{\frac{c_q s}{k}} \left(c_1 L_{s,q} \sqrt{\alpha} M_{s,q} \sqrt{\max\{q, s\} - 1} + c_2 M_{s,q} \right).$$

□

Proof of Theorem 6. For the ASGD sequence $\{\bar{\beta}_k\}_{k \in \mathbb{N}}$ with arbitrarily fixed initialization $\beta_0 \in \mathbb{R}^d$ and the stationary ASGD sequence $\{\bar{\beta}_k^\circ\}_{k \in \mathbb{N}}$ with $\beta_0^\circ \sim \pi_\alpha$, we have

$$\begin{aligned} \|\bar{\beta}_k - \bar{\beta}_k^\circ\|_q &= \frac{1}{k} \left\| \sum_{i=1}^k (\beta_i - \beta_i^\circ) \right\|_q \\ &\leq \frac{1}{k} \sum_{i=1}^k \|\beta_i - \beta_i^\circ\|_q. \end{aligned} \quad (79)$$

For each $1 \leq i \leq k$, it follows from the geometric-moment contraction in Theorem 1 that

$$\|\beta_i - \beta_i^\circ\|_q \leq r_{\alpha,s,q}^i \|\beta_0 - \beta_0^\circ\|_q. \quad (80)$$

Recall that $r_{\alpha,s,q} = 1 - 2\mu\alpha + (\max\{q, s\} - 1)L_{s,q}^2\alpha^2 < 1$ in (6). Therefore,

$$\|\bar{\beta}_k - \bar{\beta}_k^\circ\|_q \leq \frac{1}{k} \cdot \frac{r_{\alpha,s,q}(1 - r_{\alpha,s,q}^k)}{1 - r_{\alpha,s,q}} \|\beta_0 - \beta_0^\circ\|_q \leq \frac{1}{k} \cdot \frac{1}{1 - r_{\alpha,s,q}} \|\beta_0 - \beta_0^\circ\|_q. \quad (81)$$

The desired result is achieved. □

Proof of Theorem 8. Without loss of generality, assume $\beta^* = 0$. We use the notation (17) for the derivatives of G . Notice that

$$\nabla G(\beta^*) = \nabla G(0) = 0. \quad (82)$$

Consider the stationary SGD recursion

$$\beta_k^\circ = \beta_{k-1}^\circ - \alpha \nabla g(\beta_{k-1}^\circ, \xi_k), \quad k \geq 1.$$

By taking the expectation on the both sides, we obtain, for all $k \geq 1$,

$$\mathbb{E}[\nabla G(\beta_{k-1}^\circ)] = 0. \quad (83)$$

Throughout the rest of the proof, we omit the iteration index k and write $\beta = \beta_{k-1}^\circ$ when no confusion is caused. For notational convenience, write $\beta = (\beta_1, \dots, \beta_d)^\top$.

A first-order Taylor expansion on $\nabla G(\beta)$ at $\beta^* = 0$ gives

$$0 = \mathbb{E}[\nabla G(\beta)] = \nabla G(0) + \nabla^2 G(0)\mathbb{E}[\beta] + \mathcal{R}(\beta), \quad (84)$$

where $\nabla^2 G(0)$ is the $d \times d$ Jacobian matrix with entries defined by

$$[\nabla^2 G(0)]_{i,j} = \frac{\partial^2}{\partial \beta_i \partial \beta_j} G(\beta) \Big|_{\beta=0}, \quad 1 \leq i, j \leq d, \quad (85)$$

and $\mathcal{R}(\beta)$ is the d -dimensional remainder defined as

$$\mathcal{R}(\beta) = \int_0^1 \mathbb{E}([\nabla^2 G(t\beta) - \nabla^2 G(0)]\beta) dt. \quad (86)$$

The i -th entry of $\mathcal{R}(\beta)$ can be rewritten into

$$\mathcal{R}_i(\beta) = \int_0^1 (1-t) \mathbb{E}(\beta^\top \nabla^3 G_i(t\beta) \beta) dt, \quad (87)$$

where $\nabla^3 G_i(\beta)$, $1 \leq i \leq d$, is a $d \times d$ matrix whose entries are

$$[\nabla^3 G_i(\beta)]_{l,r} = \frac{\partial^3}{\partial \beta_i \partial \beta_l \partial \beta_r} G(\beta), \quad 1 \leq l, r \leq d. \quad (88)$$

Since $\nabla G(0) = 0$ and $\nabla^2 G(0)$ is invertible given that $\lambda_{\min}[\nabla^2 G(0)] > 0$, it follows from equation (84) that

$$\mathbb{E}[\beta] = -[\nabla^2 G(0)]^{-1} \mathbb{E}[\mathcal{R}(\beta)]. \quad (89)$$

We only need to bound $|\mathbb{E}[\mathcal{R}(\beta)]|_s$ using Theorem 2, that is $\mathbb{E}[|\beta_k^\circ - \beta^*|_s]^2 = O(\max\{q, s\}\alpha)$ for all $k \geq 1$.

Let $\mathbf{v} = \beta/|\beta|_s$. For each $i = 1, \dots, d$,

$$\begin{aligned} \mathbb{E}[\mathcal{R}_i(\alpha)] &= \int_0^1 (1-t) \mathbb{E}[\beta^\top \nabla^3 G_i(t\beta) \beta] dt \\ &= \int_0^1 (1-t) \mathbb{E}[|\beta|_s^2 \mathbf{v}^\top \nabla^3 G_i(t\beta) \mathbf{v}] dt. \end{aligned} \quad (90)$$

By Hölder's inequality, for $1/p + 1/q = 1$,

$$\mathbb{E}[|\beta|_s^2 \mathbf{v}^\top \nabla^3 G_i(t\beta) \mathbf{v}] \leq (\mathbb{E}[|\beta|_s^{2q}])^{1/q} \cdot (\mathbb{E}[\mathbf{v}^\top \nabla^3 G_i(t\beta) \mathbf{v}]^p)^{1/p}. \quad (91)$$

Again by Hölder's inequality,

$$\mathbb{E}[(\mathbf{v}^\top \nabla^3 G_i(t\beta) \mathbf{v})^p] \leq d^{p(1-\frac{2}{s})} \sup_{|\mathbf{v}|_s=1} \mathbb{E}|\nabla^3 G_i(t\beta) \mathbf{v}|_s^p. \quad (92)$$

Therefore, by Theorem 2 and Lemma 10,

$$\mathbb{E}[\beta] \lesssim M_{s,q}^2 \max\{q, s\} \alpha d^{\frac{q}{q-1} \cdot (1-\frac{2}{s})} \max_{1 \leq i \leq d} \|\nabla^3 G_i(\beta)\|_\infty, \quad (93)$$

where the matrix norm

$$\|\nabla^3 G_i(\beta)\|_\infty := \max_{1 \leq j_1 \leq d} \sum_{j_2=1}^d \left| (\nabla^3 G_i(\beta))_{1 \leq j_1, j_2 \leq d} \right|. \quad (94)$$

Finally, given the uniform bound $\max_{1 \leq i \leq d} \|\nabla^3 G_i(\beta)\|_\infty < \infty$,

$$\mathbb{E}[\beta] = O\left(M_{s,q}^2 \max\{q, s\} \alpha d^{\frac{q}{q-1} \cdot (1-\frac{2}{s})}\right), \quad (95)$$

which finishes the proof. $\square$

8.8 Proofs for Section 4

Proof of Theorem 4. By Theorem 6, we have $\|\bar{\beta}_k - \bar{\beta}_k^\circ\|_q \lesssim 1/(k\alpha)\|\beta_0 - \beta_0^\circ\|_q$ and consequently, it follows that

$$\mathbb{P}(\|\bar{\beta}_k - \bar{\beta}_k^\circ\|_s > z) \lesssim \frac{\|\beta_0 - \beta_0^\circ\|_q^q}{(k\alpha z)^q}, \quad z > 0. \quad (96)$$

Then it suffices to upper bound $\mathbb{P}(\|\bar{\beta}_k^\circ - \beta^*\|_s > z)$. To this end, we first bound the dependence adjusted norm (Section 2 in Zhang and Wu [2017]) for $\{\beta_k^\circ\}_{k \geq 1}$. By Theorem 1, elementary calculations yield

$$\|\beta_k^\circ - \mathbb{E}[\beta_k^\circ]\|_{q, 1/2-1/q} = O\left(\frac{M_{s,q}}{\alpha^{1/2-1/q}}\right).$$

Consequently, by Theorem 6.2 in Zhang and Wu [2017] and Theorem 8, we have

$$\mathbb{P}(\|\bar{\beta}_k^\circ - \beta^*\|_s > z) \lesssim \frac{(\log d)^{3q/2}(\log k)^{1+2q}M_{s,q}^q}{z^q k^{q-1}\alpha^{q/2-1}} + \exp\left(-\frac{Ckz^2\alpha^{1-2/q}}{M_{s,q}^2 \log d}\right).$$

Combining this with (96) completes the proof. $\square$

Theorem 9 (Theorem 3.1 in [Mies and Steland, 2023]). *Let $(\epsilon_i)_{i \in \mathbb{Z}}$ be i.i.d. random variables and $\epsilon_k = (\epsilon_k, \epsilon_{k-1}, \dots)$. Assume $X_k = G_k(\epsilon_k) \in \mathbb{R}^d$ with $\mathbb{E}[X_k] = 0$ for some measurable function G_k . For any k , denote $\tilde{\epsilon}_{k,j} = (\epsilon_k, \dots, \epsilon_{j+1}, \tilde{\epsilon}_j, \epsilon_{j-1}, \dots)$ with $\tilde{\epsilon}_j$ an i.i.d. copy of ϵ_j . Assume there exist $\Theta > 0$ and $q > 2$, such that for all k ,*

$$(\mathbb{E}|G_k(\epsilon_k) - G_k(\tilde{\epsilon}_{k,k-j})|_2^q)^{1/q} \leq \frac{\Theta}{(j \vee 1)^3}, \text{ for all } j \geq 0, \text{ and } (\mathbb{E}|G_k(\epsilon_0)|_2^q)^{1/q} \leq \Theta. \quad (97)$$

Additionally, assume that for some $\Gamma \geq 1$,

$$\sum_{k=2}^n (\mathbb{E}|G_k(\epsilon_0) - G_{k-1}(\epsilon_0)|_2^2)^{1/2} \leq \Gamma \cdot \Theta. \quad (98)$$

If $d \leq cn$ for some $c > 0$, then on a potentially different probability space, there exist random vectors $(X'_k)_{k=1}^n \stackrel{\mathcal{D}}{=} (X_k)_{k=1}^n$ and independent, mean zero, Gaussian random vectors

$$Y_k^* \sim \mathcal{N}\left(0, \sum_{h=-\infty}^{\infty} \text{Cov}(G_k(\epsilon_0), G_k(\epsilon_h))\right)$$

such that

$$\left(\mathbb{E} \max_{m \leq n} \left| \frac{1}{\sqrt{n}} \sum_{k=1}^m (X'_k - Y_k^*) \right|_2^2\right)^{1/2} \leq C\Theta\Gamma^{\frac{1}{4}}\sqrt{\log(n)}\left(\frac{d}{n}\right)^{\frac{q-2}{6q-4}},$$

for some constant C depending on (q, c) .

Instead of univariate ϵ_i , we apply Theorem 3.1 with vector-valued i.i.d. inputs ξ_i . The theorem still applies as the proof depends only on the i.i.d. random elements and their L^q bounds but not on the dimension of ξ_i .

Proof of Theorem 5. To prove the Gaussian approximation we will apply Theorem 9 (Theorem 3.1 in Mies and Steland [2023]) with $G_k \equiv G = h_\alpha$ defined in (62) since β_k° is stationary. We now verify the conditions (97) and (98).

Recall the functional dependence measure $\|\beta_k^\circ - \beta_{k,\{l\}}^\circ\|_q$ introduced in Section 8.6. Throughout the proof, the q -th moment of the Euclidean norm is denoted by

$$\|\cdot\|_q := \|\|\cdot\|_2\|_q.$$

Set

$$\rho_{\alpha,q}^2 := 1 - 2\alpha\mu + 7(\max\{q, 2\} - 1)\alpha^2 L_{2,q}^2, \quad C_{\alpha,q} := 2\alpha\sqrt{cL_{2,q}^2 \max\{q, 2\}\alpha + 1} \quad (99)$$

for some constant $c > 0$. If c is chosen sufficiently large, then, by Lemma 11 and Theorem 2, for all $k \geq 1$ and $l \leq k$, we have

$$\begin{aligned}\|\beta_k^\circ - \beta_{k,\{l\}}^\circ\|_q^2 &\leq 4\alpha^2 \rho_{\alpha,q}^{2(k-l)} \left(L_{2,q}^2 \|\beta_{l-1}^\circ - \beta^*\|_q^2 + M_{2,q}^2 \right) \\ &\leq 4\alpha^2 \rho_{\alpha,q}^{2(k-l)} M_{2,q}^2 \left(cL_{2,q}^2 \max\{q, 2\} \alpha + 1 \right) \\ &= C_{\alpha,q}^2 \rho_{\alpha,q}^{2(k-l)} M_{2,q}^2.\end{aligned}$$

For $\alpha \in (0, \alpha_{s_d,q})$, it follows that $\rho_{\alpha,q} < 1$. Let $l = k - j$. Then, for a sufficiently large constant $C'_{\alpha,q}$, we have

$$\|\beta_k^\circ - \beta_{k,\{k-j\}}^\circ\|_q \leq C_{\alpha,q} M_{2,q} \rho_{\alpha,q}^j \leq C'_{\alpha,q} M_{2,q} (j+1)^{-3}. \quad (100)$$

Therefore, the condition (97) holds with $\Theta = C'_{\alpha,q} M_{2,q}$. This verifies the first part of condition 97. For the second part of condition 97, by Assumption 3 and Theorem 2, for some constant $C''_{\alpha,q} > 0$,

$$\|h_\alpha(\xi_0, \xi_{-1}, \dots)\|_q = \|\beta_0^\circ\|_q \leq \|\beta_0^\circ - \beta^*\|_q + \|\beta^*\|_2 \leq C''_{\alpha,q} M_{2,q} < \infty. \quad (101)$$

Moreover, since β_k° is stationary, $G_k = G_{k-1} = h_\alpha$ and the left hand side of (98) is zero. Thus, condition (98) is trivially satisfied with $\Gamma = 1$.

Finally, we show that the long-run covariance matrix $\Xi = \sum_{k=-\infty}^{\infty} \text{Cov}(\beta_0^\circ, \beta_k^\circ)$ is well defined in the sense that the spectral norm $\|\Xi\|_s$ is finite. Following (63), denote

$$\beta_{k,\{\leq l\}}^\circ := \beta_{k,\{\dots, l-1, l\}}^\circ = h_\alpha(\xi_k, \dots, \xi_{l+1}, \xi'_l, \xi'_{l-1}, \dots) = h_\alpha(\mathcal{F}_{k,\{\dots, l-1, l\}}), \quad l \leq k. \quad (102)$$

Since $\beta_{k,\{\leq 0\}}^\circ$ is independent of β_0° , we have

$$\begin{aligned}\text{Cov}(\beta_0^\circ, \beta_k^\circ) &= \mathbb{E}[\beta_0^\circ \beta_k^{\circ\top}] - \mathbb{E}[\beta_0^\circ] \mathbb{E}[\beta_k^{\circ\top}] \\ &= \mathbb{E}[\beta_0^\circ \beta_k^{\circ\top}] - \mathbb{E}[\beta_0^\circ] \mathbb{E}[\beta_{k,\{\leq 0\}}^{\circ\top}] \\ &= \mathbb{E}[\beta_0^\circ (\beta_k^\circ - \beta_{k,\{\leq 0\}}^\circ)^\top].\end{aligned} \quad (103)$$

We can rewrite the difference as a telescoping sum,

$$\beta_k^\circ - \beta_{k,\{\leq 0\}}^\circ = \sum_{l=0}^{\infty} (\beta_{k,\{\leq -l+1\}}^\circ - \beta_{k,\{\leq -l\}}^\circ). \quad (104)$$

By stationarity and (100), it follows that

$$\|\beta_{k,\{\leq -l+1\}}^\circ - \beta_{k,\{\leq -l\}}^\circ\|_2 = \|\beta_{k,\{-l+1\}}^\circ - \beta_{k,\{-l\}}^\circ\|_2 \leq C_{\alpha,2} M_{2,2} \rho_{2,2}^{k+l+1}. \quad (105)$$

For the spectral norm,

$$\begin{aligned}\|\text{Cov}(\beta_0^\circ, \beta_k^\circ)\|_s &= \sup_{\mathbf{u}, \mathbf{v} \in \mathbb{R}^d, \|\mathbf{u}\|_2 = \|\mathbf{v}\|_2 = 1} \mathbb{E} \mathbf{v}^\top \beta_0^\circ (\beta_k^\circ - \beta_{k,\{\leq 0\}}^\circ)^\top \mathbf{u} \\ &\leq \sup_{\mathbf{u}, \mathbf{v} \in \mathbb{R}^d, \|\mathbf{u}\|_2 = \|\mathbf{v}\|_2 = 1} [\mathbb{E}(\mathbf{v}^\top \beta_0^\circ)^2]^{1/2} [\mathbb{E}[(\beta_k^\circ - \beta_{k,\{\leq 0\}}^\circ)^\top \mathbf{u}]^2]^{1/2} \\ &\leq \|\beta_0^\circ\|_2 \|\beta_k^\circ - \beta_{k,\{\leq 0\}}^\circ\|_2,\end{aligned} \quad (106)$$

where the first inequality is by Cauchy-Schwarz and the last inequality uses $(\mathbf{u}^\top \beta_0^\circ)^2 \leq |\beta_0^\circ|^2$ with $\|\mathbf{u}\|_2 = 1$. This, along with $M_{2,2} < \infty$ (Assumption 3) yields

$$\|\text{Cov}(\beta_0^\circ, \beta_k^\circ)\|_s \leq \|\beta_0^\circ\|_2 \|\beta_k^\circ - \beta_{k,\{\leq 0\}}^\circ\|_2 \leq C'_\alpha \rho_{\alpha,2}^k, \quad (107)$$

for some constant $C'_\alpha > 0$. As a direct consequence,

$$\|\Xi\|_s \leq \|\mathbb{E}[\beta_0^\circ \beta_0^{\circ\top}]\|_s + 2 \sum_{k=1}^{\infty} C'_\alpha \rho_{\alpha,2}^k < \infty. \quad (108)$$

This completes the proof. □

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification: *We clearly stressed our contributions and scope in the abstract and included a subsection in the introduction to list our key innovations.*

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: *We talked about the limitations in the last section, especially for the assumptions of strong convexity and smoothness.*

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: *We provide rigorous proofs for all the theoretical results in the supplementary material.*

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [NA]

Justification: *This paper contributes to theoretical guarantees of high-dimensional SGD.*

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [NA]

Justification: *This paper does not include experiments requiring code.*

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [NA]

Justification: *This paper does not include experiments.*

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [NA]

Justification: *This paper does not include experiments.*

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [NA]

Justification: *This paper does not include experiments.*

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: *We followed the NeurIPS Code of Ethics as instructed.*

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: *This paper discusses potential positive societal impacts, particularly through advancing the theoretical understanding of modern machine learning, which can inform the development of more robust and efficient algorithms. As the work is purely theoretical and does not propose or evaluate any deployable systems, we do not anticipate any direct negative societal impacts.*

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: *This paper poses no such risks.*

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [NA]

Justification: *This paper does not use existing assets.*

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset’s creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: *This paper does not release new assets.*

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: *This paper does not involve crowdsourcing nor research with human subjects.*

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

This paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: *The core method development in this research does not involve LLMs as any important, original, or non-standard components.*

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.