Gaussian Approximation and Concentration of Constant Learning-Rate Stochastic Gradient Descent

Ziyang Wei
Department of Statistics
University of Chicago
Chicago, IL 60637
ziyangw@uchicago.edu

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

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

Abstract

We establish a comprehensive finite-sample and asymptotic theory for stochastic gradient descent (SGD) with constant learning rates. First, we propose a novel linear approximation technique to provide a quenched central limit theorem (CLT) for SGD iterates with refined tail properties, showing that regardless of the chosen initialization, the fluctuations of the algorithm around its target point converge to a multivariate normal distribution. Our conditions are substantially milder than those required in the classical CLTs for SGD, yet offering a stronger convergence result. Furthermore, we derive the first Berry-Esseen bound – the Gaussian approximation error – for the constant learning-rate SGD, which is sharp compared to the decaying learning-rate schemes in the literature. Beyond the moment convergence, we also provide the Nagaev-type inequality for the SGD tail probabilities by adopting the autoregressive approximation techniques, which entails non-asymptotic large-deviation guarantees. These results are verified via numerical simulations, paving the way for theoretically grounded uncertainty quantification, especially with non-asymptotic validity.

1 Introduction

In large-scale optimization and streaming-data applications, online learning has played a crucial role, where stochastic approximation serves as a fundamental ingredient to improve computational efficiency and save memory. However, to facilitate trustworthy AI and reliable decision-making based on stochastic approximation, it is important to understand its inherent variability, especially in finite-sample settings.

This paper considers a popular recursive algorithm in online learning – Stochastic Gradient Descent (SGD), also known as the Robbins-Monro algorithm [Robbins and Monro, 1951], which is widely used due to its memory efficiency, computational simplicity, and algorithmic stability. The convergence and distributional theory of SGD and its variants have been extensively studied [Fabian, 1968; Woodroofe, 1972; Pflug, 1986; Polyak and Juditsky, 1992; Kushner and Yin, 1997; Shamir and Zhang, 2013]. Nevertheless, the uncertainty quantification of constant learning-rate SGD remains partially understood. Though central limit theorems (CLT) have been explored in literature [Pflug, 1986; Dieuleveut et al., 2020], refined theoretical properties such as non-asymptotic Gaussian approx-

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

imation rate and sharp concentration inequalities are lacking. To fill in this gap, we provide the first Berry-Esseen bound and the first Nagaev-type tail probability for constant learning-rate SGD.

Specifically, we are interested in the following optimization problem:

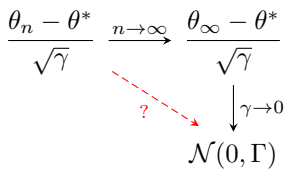
$$\theta^* = \arg\min_{\theta\in\mathbb{R}^d} G(\theta) \triangleq \arg\min_{\theta\in\mathbb{R}^d} \mathbb{E}_{X\sim\Pi} g(\theta, X), \tag{1}$$

where G is the objective function from $\mathbb{R}^d$ to $\mathbb{R}$, and $g(\theta, X)$ is the noise-perturbed loss function. With sequentially arriving i.i.d. random data $X_i \sim \Pi$ for some unknown distribution Π, SGD updates the estimated model parameter as

$$\theta_i = \theta_{i-1} - \gamma \nabla g(\theta_{i-1}, X_i),\ i \geq 1, \tag{2}$$

where ∇g is the gradient of $g(\theta, X)$ with respect to the first argument θ, and $\gamma > 0$ is the fixed learning rate.

Asymptotic normality of the SGD iterates $\{\theta_i\}_{i\in\mathbb{N}}$ has been investigated in many works. For example, Pflug [1986] proved the first CLT of SGD with constant learning rate, where the SGD iterates were viewed as a time-homogeneous Markov chain. This provides a traditional framework to address the non-stationarity of constant learning-rate SGD. Under certain regularity conditions, such a Markov chain has been shown to converge to a stationary probability measure π_γ geometrically fast as the iteration n grows to infinity [Dieuleveut et al., 2020; Merad and Gaïffas, 2023]. Building on this asymptotic stationarity, the more recent works focus on the scaling of this stationary distribution in regard to the learning rate γ, and establish the CLT as $\gamma \to 0$ [Chen et al., 2022; Wei et al., 2025]. Since the constant learning-rate SGD will not converge to θ^* but oscillates around it with stochastic order of $\sqrt{\gamma}$ [Dieuleveut et al., 2020], we denote by $\theta_\infty \in \mathbb{R}^d$ the random vector following the stationary measure. The diagram below illustrates the key ingredients of classical analysis.



As we can see from the diagram, all the aforementioned works cannot avoid one critical problem: they require the iterations $n \to \infty$ before $\gamma \to 0$. In other words, their asymptotic normality result only holds for the stationary SGD sequence, which, however, is unrealistic in practice, since practitioners usually arbitrarily fix the initialization θ_0 that yields a non-stationary sequence θ_i, while the stationary sequence is unattainable in finite time. In contrast, the actual procedure of stochastic approximation is to first determine a small learning rate, then update the algorithm with a large sample size, which means $n \to \infty$ and $\gamma \to 0$ have to perform concurrently and dependently. Under this circumstance, it remains unclear whether CLT still holds for the SGD iterate itself and, if it does, what constraints between n and γ are required. To address the issue, we directly approximate the SGD sequence with iterative linear random functions, skipping the intermediate stationary process. Our methodology exhibits superiority in the sense that, with weaker assumptions, it reveals the simple yet essential relationship between the number of iterations and learning rate sufficient for a quenched version of CLT that holds for any initialization θ_0, and also demonstrates how standardized SGD sequences converge to the normal distribution despite the ordering of limitation on n and γ. Furthermore, with a slightly stronger condition, linear approximation leads to another powerful finite-sample Gaussian approximation – the Berry-Esseen inequality [Chen and Shao, 2001; Korolev and Shevtsova, 2010; Raič, 2019] – which characterizes the explicit order of distance between SGD and the Gaussian distribution with the number of iterations and learning rate.

Beyond the asymptotic theory, the existing literature on non-asymptotic uncertainty quantification has primarily focused on the linear model [Durmus et al., 2021; Zhu et al., 2022; Agrawalla et al., 2023; Samsonov et al., 2024] or the power-law decaying learning rate [Anastasiou et al., 2019; Shao and Zhang, 2022; Sheshukova et al., 2025], which introduces additional tuning parameters and suffers from surged sensitivity to hyperparameters and slow convergence [Nemirovski et al., 2009; Chee and Toulis, 2018]. Nevertheless, fixed learning rate has recently gained popularity for its simpler tuning requirements and rapid forgetting of the initial value. Moreover, it also enables parallelization of multiple SGD runs to accelerate convergence and one can employ extrapolation techniques for bias

correction [Dieuleveut et al., 2020]. However, establishing non-asymptotic theoretical guarantees for constant learning-rate SGD poses greater challenges due to its persistent oscillations around a stationary region induced by non-diminishing step sizes [Dieuleveut et al., 2020; Cardot et al., 2013, 2017]. To grasp the recursive nature of SGD and comprehend its nonlinear dependence structure, we present a systematic and sophisticated framework based on coupling and dependence measure theory established in Wu [2005]. The high-level idea throughout this paper is to reveal the intrinsic complexity of time series by slight modifications of input data or the iteration mechanism.

1.1 Contributions

This paper advances the theoretical understanding of constant learning-rate SGD by introducing novel approximation techniques and deriving sharp finite-sample guarantees. Our main contributions are:

- **A linear-approximation CLT for constant-step SGD.** We develop a new linearization framework that captures the drift and noise dynamics of SGD iterates. Under substantially weaker smoothness and moment conditions than those in Pflug [1986] and Dieuleveut et al. [2020], we prove a central limit theorem showing that, as the step size $\gamma \to 0$, the properly scaled SGD iterate converges in distribution to a Gaussian law.
- **Non-asymptotic p-th moment bounds for $p \geq 2$.** Going beyond weak convergence, we derive explicit finite-sample upper bounds on $\mathbb{E}[|\theta_n - \theta^*|^p]$ for any integer $p \geq 2$. The moment convergence rate quantifies how quickly all higher moments of the error decay in n, which provides a more refined theoretical guarantee compared to the weak convergence in the probability measure.
- **First Berry–Esseen bound in the Gaussian approximation.** We first provide a quenched version CLT of the last-iterate SGD. Under mild regularity conditions, if $n \to \infty$ and $\gamma \to 0$ with $n\gamma \geq \nu \log n$ for some $\nu \geq 1/(2\gamma)$, then for any initialization $\theta_0 \in \mathcal{R}^d$ the scaled iterate satisfies
$$\frac{\theta_n - \theta^*}{\sqrt{\gamma}} \xrightarrow{D} \mathcal{N}(0, \Gamma),$$
where the matrix Γ is later defined in (6). Furthermore, we also obtain the Berry–Esseen type rate (up to logarithmic factors) for the distributional distance between the scaled SGD iterate and its Gaussian limit. We quantify how rapidly the convergence to normality occurs in finite samples and fixed γ by providing the rate for
$$\sup_{\mathcal{D} \in \mathcal{V}} |\mathbb{P}(\frac{\theta_n - \theta^*}{\sqrt{\gamma}} \in \mathcal{D}) - \mathbb{P}(\mathcal{N}(0, \Gamma) \in \mathcal{D})|$$
where $\mathcal{V}$ is the collection of all convex sets in $\mathbb{R}^d$. This is the first Gaussian approximation bound for constant-learning-rate SGD in general settings, which allows one to assess the accuracy of statistical inference with finite samples.
- **Nagaev-type large-deviation inequalities.** By approximating the SGD recursion with an autoregressive process, we derive sharp Nagaev-style tail bounds that control the probability of large deviations beyond the CLT regime. These results yield explicit sub-Gaussian and polynomial terms in the convergence rate for $\mathbb{P}(|\theta_n - \theta^*| > \epsilon)$ for any $\epsilon > 0$, which relates to tight sample complexity bounds significant for statistical learning theory and applications [Valiant, 1984].

1.2 Other related works

Stochastic Gradient Descent. The large-sample behavior of SGD and its extensions dates back to the foundational analyses by Blum [1954]; Sacks [1958], who first investigated its asymptotics and has since been elaborated upon by a succession of studies [Ljung, 1977; Lai, 2003; Wang and Gao, 2010; Gandikota et al., 2022; Zhong et al., 2024; Li et al., 2024]. In particular, Fabian [1968] characterized the limiting law of the final iterate. Robbins and Siegmund [1971] then leveraged martingale arguments to establish almost-sure convergence of the procedure. Subsequent research has quantified the algorithm's convergence rates under a variety of assumptions [Toulis and Airoldi, 2017; Pillaud-Vivien et al., 2018; Muecke et al., 2019; Duchi and Ruan, 2021]. See more recent works by Zhu et al. [2022]; Hu and Fu [2024]; Lauand and Meyn [2024] among the others.

Convergence with different learning rates. The learning-rate choice in SGD, either fixed or decaying, critically shapes convergence. Pflug [1986] first studied the stationary behavior under a constant step size via a Markov-chain framework; Dieuleveut et al. [2020]; Huo et al. [2023]; Merad and Gaïffas [2023] later refined this with Wasserstein-distance analyses and high-confidence bounds. For diminishing steps, Rakhlin et al. [2012] showed the linear decaying case attains the optimal rate, and Ge et al. [2019] extended the results to polynomial decays. Burn-in strategies were proposed by Gower et al. [2019] and Nguyen et al. [2019]. Adaptive schemes include Polyak's method for over-parameterized models [Loizou et al., 2021] and a bandwidth-based family studied by Wang and Yuan [2023]. See Jiang and Stich [2024] for a comprehensive overview. However, in practice, constant learning rates are favored due to the simplicity of training, which is also the focus of this study.

(Non)-asymptotic normality. In addition to convergence guarantees, performing real-time inference with SGD-style estimators is vital for uncertainty quantification. Classical bootstrap-based M-estimation methods [Fang et al., 2018; Fang, 2019; Zhong et al., 2024] are computationally prohibitive in streaming contexts. Instead, Polyak–Ruppert averaging [Ruppert, 1988; Polyak and Juditsky, 1992] offers statistical efficiency and enables inference: the averaged SGD (ASGD) sequence [Györfi and Walk, 1996; Defossez and Bach, 2015] admits an asymptotic normality result at the optimal rate [Moulines and Bach, 2011; Dieuleveut and Bach, 2016; Dieuleveut et al., 2017; Jain et al., 2018]. By contrast, inference on the final iterate under a constant step size has seen little treatment. We close this gap by proving a quenched central limit theorem for the SGD estimator as the learning rate $\gamma \to 0$, valid from any initialization [Dahlhaus and Rao, 2006; Dahlhaus et al., 2019]. Additionally, blocking-based variance estimators [Chen et al., 2020; Zhu et al., 2023] and recursive kernel approaches [Huang et al., 2014] achieve optimal mean-squared-error rates under dependence, yielding practical, theoretically sound online inference for SGD.

1.3 Notation

For a vector $v = (v_1, \ldots, v_d)^\top \in \mathbb{R}^d$ and $q > 0$, we denote $|v|_q = (\sum_{i=1}^d |v_i|^q)^{1/q}$ and $|v| = |v|_2$. For any $s > 0$ and a random vector X, we say $X \in \mathcal{L}^s$ if $\|X\|_s = (\mathbb{E}|X|_2^s)^{1/s} < \infty$. For two positive number sequences (a_n) and (b_n), we say $a_n = O(b_n)$ or $a_n \lesssim b_n$ (resp. $a_n \asymp b_n$) if there exists $C > 0$ such that $a_n/b_n \le C$ (resp. $1/C \le a_n/b_n \le C$) for all large n, and write $a_n = o(b_n)$ if $a_n/b_n \to 0$ as $n \to \infty$. Let (X_n) and (Y_n) be two sequences of random variables. Write $X_n = o_{\mathbb{P}}(Y_n)$ if $X_n/Y_n \to 0$ in probability as $n \to \infty$. Let $\langle \cdot, \cdot \rangle$ denote the canonical inner product in the finite dimensional Euclidean space $\mathbb{R}^d$.

2 Moment convergence

We first introduce the following assumptions on the objective function $G(\theta)$ and the stochastic gradients $\nabla g(\theta, X)$.

Assumption 2.1 (μ-strong convexity)**.** *The function G is twice differentiable and μ-strongly convex, i.e. for a $\mu > 0$ and for all $\theta, \theta' \in \mathbb{R}^d$, it holds that*

$$\langle \nabla G(\theta) - \nabla G(\theta'), \theta - \theta' \rangle \ge \mu |\theta - \theta'|^2.$$

Since Assumption 2.1 requires that G is twice differentiable, the Hessian matrix $\nabla^2 G(\theta^*)$ exists. Note that for the L^p convergence in Theorem 2.3, we only need G to be continuously differentiable. The existence of Hessian is only necessary to derive the CLT and Gaussian approximation in Theorem 3.4 and 3.5.

Assumption 2.2 (Stochastic Lipschitz continuity)**.** *The function $g(\theta, x)$ is continuously differentiable w.r.t. θ for any x. Moreover, for some $p \ge 2$, assume $\|\nabla g(\theta^*, X)\|_p =: M_p < \infty$, $\|\theta_0 - \theta^*\|_p =: \rho_p < \infty$, and the stochastic Lipschitz continuity,*

$$\|\nabla g(\theta_1, X) - \nabla g(\theta_2, X)\|_p \le L_p |\theta_1 - \theta_2|, \quad \textit{for all } \theta_1, \theta_2 \in \mathbb{R}^d.$$

Here, the condition $\|\theta_0 - \theta^*\|_p < \infty$ trivially holds in the fixed-initialization setting. Both strong convexity and Lipschitz continuity conditions are commonly adopted in the literature; see for example, Dieuleveut et al. [2020]; Zhu et al. [2022]; Merad and Gaïffas [2023]. Notably, Assumption 2.2 also implies the L-smoothness of the gradient of objective function G, that is

$$|\nabla G(\theta_1) - \nabla G(\theta_2)| \le L_p |\theta_1 - \theta_2|, \quad \text{for all } \theta_1, \theta_2 \in \mathbb{R}^d.$$

We refer to Li et al. [2024] for a detailed discussion. Now we are ready to present the L^p moment convergence of constant learning-rate SGD.

Theorem 2.3 (L^p Convergence)**.** *Suppose that Assumptions 2.1 and 2.2 hold. Let α_0 be some constant satisfying*

$$0 < \alpha_0 \leq \min\Big\{\frac{1}{\gamma}, 2\mu - (6p-5)L_p^2\gamma\Big\}. \tag{3}$$

Then, for any $n \geq 1$ and γ satisfying

$$0 < \gamma < \frac{2\mu}{(6p-5)L_p^2}, \tag{4}$$

we have,

$$\|\theta_n - \theta^*\|_p^2 \leq (1-\alpha_0\gamma)^n\rho_p^2 + 3(p-1)M_p^2\alpha_0^{-1}\gamma. \tag{5}$$

Remark (Rate of moment convergence)**.** *The right-hand side of* (5) *demonstrates that SGD with a constant learning rate forgets its initial condition with an exponential pace. Besides, if we determine the learning rate γ based on the knowledge of the total sample size or number of iterations n, a constraint $n\gamma \to \infty$ should be satisfied to ensure convergence, i.e., the total length of steps needs to be sufficiently large such that SGD can move on. This relationship is quite natural and required by most of the literature on SGD [Polyak and Juditsky, 1992; Kushner and Yin, 1997; Sheshukova et al., 2025]. For example, consider the decaying learning rate schedule $\eta_n \asymp n^{-\alpha}$, it is well-known that α can not exceed 1 for the sake of $\sum_{i=1}^n \eta_i \to \infty$.*

3 CLT and Berry-Esseen theorem

In this section, we introduce how to approximate SGD sequences with recursive linear random functions. Consequently, we establish asymptotic normality and finite-sample Gaussian approximation directly on the SGD iterations.

3.1 Dependency of SGD iterates

In recursive algorithms, each new updated estimator depends on the last update and the new-coming random sample, which involves an intricate dependency structure that poses a challenge to normality analysis. To address this issue, we leverage the theory of functional dependence measure introduced in Wu [2005] to quantify the dependence structure of the SGD algorithm. Define $\theta_n = \tau_n(X_1, X_2, ..., X_n)$ for some measurable function τ_n that can vary for different n and $\theta_n^{(t)} = \tau_n(X_1, X_2, ..., X_{t-1}, X_t', X_{t+1}, ..., X_n)$, where X_t' is an i.i.d. copy of X_t. The functional dependence measure is defined as

$$\psi(n,t,p) = \|\theta_n - \theta_n^{(t)}\|_p.$$

We denote $\psi(n,t) = \psi(n,t,2)$ for simplicity. In particular, $\psi(n,t,p)$ quantifies the effect of the random sample X_t on the n-th SGD iterate θ_n. By utilizing this tool, we can derive exact Gaussian approximation and tail probability rates. We first provide an essential bound of this dependence measure, which will be heavily used in the proofs.

Theorem 3.1 (Functional dependence measure)**.** *Suppose the same conditions of Theorem 2.3 hold. The functional dependence measure satisfies*

$$\psi(n,t,p) \leq 2\sqrt{2}\gamma(1-\alpha_0\gamma)^{(n-t)/2}\sqrt{M_p^2 + L_p^2[(1-\alpha_0\gamma)^{t-1}\rho_p^2 + 3(p-1)M_p^2\alpha_0^{-1}\gamma]}.$$

Remark (Rate of dependence measure)**.** *Theorem 3.1 encapsulates the temporal dependence of constant learning rate SGD, which decays at an exponential rate as the time lag increases.*

3.2 Refined linear approximation and asymptotic normality

We first introduce another assumption only required for the rest of this section.

Assumption 3.2 (Local Smoothness)**.** *There exists some constants $L \geq 0$ and $\kappa > 0$ such that for all $|\theta - \theta^*| \leq \kappa$,*

$$|\nabla^2 G(\theta) - \nabla^2 G(\theta^*)| \leq L|\theta - \theta^*|.$$

The local smoothness of the Hessian is standard in the literature on statistical inference and normal approximation of online learning algorithms [Anastasiou et al., 2019; Shao and Zhang, 2022; Li et al., 2022; Sheshukova et al., 2025]. Assumption 1 in [Anastasiou et al., 2019] requires a stronger global Lipschitz smoothness. We only need this condition to ensure the validity of Taylor's expansion of the gradient, i.e.,

$$|\nabla G(\theta) - \nabla^2 G(\theta^*)(\theta - \theta^*)| \lesssim |\theta - \theta^*|^2,$$

which is also commonly imposed [Ruppert, 1988; Polyak and Juditsky, 1992] and necessary for Assumption H6 in [Moulines and Bach, 2011].

Define $A := \nabla^2 G(\theta^*)$ as the Hessian matrix of the objective function $G(\theta)$ at $\theta = \theta^*$, and $S = \mathbb{E}[\nabla g(\theta^*, X_n) \nabla g(\theta^*, X_n)^\top]$ as the covariance of the stochastic gradients, also at the true parameter θ^*. Denote the estimation error by $\Delta_n = \theta_n - \theta^*$. We have

$$\Delta_n = \Delta_{n-1} - \gamma A \Delta_{n-1} + \gamma R_n + \gamma D_n - \gamma \nabla g(\theta^*, X_n),$$

where we define the Taylor expansion remainder and the martingale difference noise term respectively by

$$\begin{aligned} R_n &= A\Delta_{n-1} - \nabla G(\theta_{n-1}), \\ D_n &= \nabla G(\theta_{n-1}) - \nabla g(\theta_{n-1}, X_n) + \nabla g(\theta^*, X_n). \end{aligned}$$

Recursively updating the formula, we get

$$\Delta_n = (\mathbf{I}_d - \gamma A)^n \Delta_0 - L_n + I_{1,n} + I_{2,n},$$

where

$$L_n = \gamma \sum_{k=1}^{n} (\mathbf{I}_d - \gamma A)^{n-k} \nabla g(\theta^*, X_k),$$

$I_{1,n} = \gamma \sum_{k=1}^{n} (\mathbf{I}_d - \gamma A)^{n-k} R_k$ and $I_{2,n} = \gamma \sum_{k=1}^{n} (\mathbf{I}_d - \gamma A)^{n-k} D_k$. Due to stochastic Lipschitz continuity and Theorem 5, the terms $I_{1,n}$ and $I_{2,n}$ are infinitesimal of higher order, and the estimation error of SGD can be well approximated by the linear sequence L_n. Let $\lambda^* > \lambda_* > 0$ denote the largest and smallest eigenvalue of A, and $\lambda = \min\{\lambda_*, \alpha_0\}$. We have the following result,

Lemma 3.3 (Linear approximation)**.** *Under Assumption 3.2 and same conditions of Theorem 2.3, for*

$$\gamma < \min\{\frac{1}{\lambda^*}, \frac{\mu}{(6p-5)L_p^2}\},$$

we have

$$\|\Delta_n - L_n\|_1 \leq \Big(\frac{3C_0 M_2^2}{\lambda \alpha_0} + \frac{2\sqrt{3} L_2 M_2}{\sqrt{\lambda \alpha_0}}\Big)\gamma + \frac{2L_2 \rho_2}{\sqrt{\lambda}} \sqrt{\gamma} (1 - \lambda\gamma)^{\frac{n-1}{2}} + C_0 \rho_2^2 n \gamma (1 - \lambda\gamma)^{n-1},$$

where $C_0 = \max\{L, 2L_2 \kappa^{-1}\}$.

We will show in the appendix that the scaled limiting covariance of the linear sequence L_n can be obtained from the Lyapunov equation. Specifically, let Γ be the unique solution of

$$A\Gamma + \Gamma A = S,$$

which can also be written as

$$\Gamma = \int_{\mathbb{R}^+} e^{-At} S e^{-At} dt. \tag{6}$$

Then the following limitation holds

$$\Gamma = \lim_{\substack{n\gamma \to \infty \\ \gamma \to 0}} \frac{\mathrm{Cov}(L_n)}{\gamma}.$$

This asymptotic covariance is in accordance with results in [Pflug, 1986; Chen et al., 2022; Wei et al., 2025]. In the theorem below, we present a refined CLT result via the linear approximation technique.

Theorem 3.4 (Quenched Central Limit Theorem)**.** *Under same conditions of Lemma 3.3, let $n \to \infty$ and $\gamma \to 0$ such that $n\gamma \geq \nu \log n$ for some constant $\nu > 1/2\lambda$. Then, for an SGD sequence $\{\theta_n\}_{n\in\mathbb{N}}$ with arbitrarily initialization $\theta_0 \in \mathbb{R}^d$, we have*

$$\frac{\theta_n - \theta^*}{\sqrt{\gamma}} \xrightarrow{D} \mathcal{N}(0, \Gamma).$$

Remark (Quenched version of asymptotic normality)**.** *In most classical CLT-type results for SGD, one assumes that the iterate sequence $\{\theta_n\}_{n\in\mathbb{N}}$ is stationary, which means choosing the initialization θ_0 exactly following the limiting stationary distribution of θ_n [Pflug, 1986; Dieuleveut et al., 2020; Chen et al., 2022]. Instead, our quenched CLT yields a stronger result which guarantees the asymptotic normality of SGD sequences with any arbitrarily initial point $\theta_0 \in \mathbb{R}^d$.*

Notably, the relationship $n\gamma \gtrsim \log n$ is a minimal condition for SGD in most settings as discussed before. Consider the decaying learning rate schedule $\eta_n \asymp n^{-\alpha}$, taking $\alpha = 1$, the minimal rate of the total step size $\sum_{i=1}^{n} \eta_i$ is also $\mathcal{O}(\log n)$.

3.3 Non-asymptotic Gaussian approximation

The central limit theorem 3.4 establishes a weak convergence of SGD to the Gaussian distribution. However, it does not tell how close the SGD iterates are to Gaussian, and how fast it converges when the number of iterations grows or the learning rate is turned down. For example, it is natural to ask how small the following distance could be

$$|\mathbb{P}((\theta_n - \theta^*)/\sqrt{\gamma} \in \hat{\mathcal{C}}) - \mathbb{P}(\mathcal{N}(0, \Gamma) \in \hat{\mathcal{C}})|$$

for some confidence interval $\hat{\mathcal{C}}$ of concern, but the CLT can not help due to its asymptotic essence. Consequently, the effectiveness of statistical inference is still questionable with a finite sample and fixed learning rate.

This restriction motivates us to investigate the non-asymptotic convergence rate to normality. The Berry-Esseen theorem is a powerful tool to quantify the maximum approximation error. In the original groundbreaking work [Berry, 1941; Esseen, 1942], the Kolmogorov–Smirnov distance between the empirical distribution and the Gaussian is specified as

$$\sup_{x\in\mathbb{R}} |F_n(x) - \Phi(x)| \leq \frac{0.33554\mathbb{E}|X_1^3| + 0.4748(\mathbb{E}X_1^2)^{2/3}}{(\mathbb{E}X_1^2)^{2/3}\sqrt{n}},$$

where F_n is the cumulative distribution function (cdf) of i.i.d. average of $X_1, ..., X_n$ with a finite third moment, and Φ is the standard normal cdf.

Taking advantage of the linear recursion and functional dependence measure, we develop the following optimal Gaussian approximation result for constant learning rate SGD.

Theorem 3.5 (Berry-Esseen bound)**.** *Suppose that Assumptions 2.1 and 2.2 hold with $p \geq 4$, and the same conditions of Lemma 3.3 hold. Let $\mathcal{V} = \{\mathcal{D} \in \mathbb{R}^d : \mathcal{D} \text{ is convex.}\}$ and Y_Γ be a mean zero normal vector in $\mathbb{R}^d$ with covariance Γ. Then, we have the following Berry-Esseen inequality:*

$$\begin{aligned}&\sup_{\mathcal{D}\in\mathcal{V}} |\mathbb{P}(\frac{\Delta_n}{\sqrt{\gamma}} \in \mathcal{D}) - \mathbb{P}(Y_\Gamma \in \mathcal{D})| \\ \leq& C\big[\sqrt{\gamma} + (1-\lambda\gamma)^{\frac{n-1}{2}} + \sqrt{\gamma}n(1-\lambda\gamma)^{n-1} + \gamma + (1-\lambda\gamma)^{2n}\big],\end{aligned}$$

where C is a constant independent of n, γ and θ_0.

Remark (Rate of Gaussian approximation)**.** *The first term comes from the third moment of linear approximation sequence. The second and third terms come from the dependence structure of SGD and the error of linear approximation. The last two terms are due to the difference between finite-sample and asymptotic covariance. To the best of our knowledge, Theorem 3.5 is the first Gaussian approximation result for constant learning-rate SGD that explicitly bounds the distance between distributions with a specific order of γ and n. The dominant term $\sqrt{\gamma}$ can not be improved, since the bias of SGD is $\mathcal{O}(\gamma)$ which can not be eliminated [Dieuleveut et al., 2020]. As a result, $\Delta_n/\sqrt{\gamma}$ is at least $\mathcal{O}(\sqrt{\gamma})$ away from the centered Gaussian vector.*

Suppose we know the number of iterations n a priori. To choose an appropriate γ scaling with n, the right-hand side can be nearly optimized by setting $\gamma \propto \nu \log n/n$ for some $\nu > 1/\lambda$. Then the order of Gaussian approximation becomes

$$\sup_{\mathcal{D}\in\mathcal{V}} |\mathbb{P}(\frac{\Delta_n}{\sqrt{\gamma}} \in \mathcal{D}) - \mathbb{P}(Y_\Gamma \in \mathcal{D})| \lesssim \frac{\sqrt{\log n}}{\sqrt{n}}.$$

A similar result of ASGD with decaying learning rate schedule $\eta_n \asymp n^{-\alpha}$ is developed by [Shao and Zhang, 2022; Samsonov et al., 2024], where the optimal approximation rate is $\mathcal{O}(n^{-1/4})$. In comparison, the normal approximation rate is sharper with a constant learning rate SGD.

4 Sharp concentration of tail probability

In addition to L_p convergence, CLT, and Gaussian approximation, another vital concern is the tail behavior of the estimation error, as the expected error rate can not be achieved by a one-time implementation. Practitioners want to ensure an ideal performance of the algorithm in a single trial, especially under non-linear and heavy-tailed noisy settings. To this end, we generalize the method developed in Nagaev [1979] to provide a high probability guarantee of the estimation accuracy without the requirement of bounded or sub-Gaussian gradient noise.

4.1 Non-linear autoregressive approximation

We use the following non-linear autoregressive sequence to approximate the SGD sequence. Let $\beta_0 = \theta^*$ and

$$\beta_n = \beta_{n-1} - \gamma\nabla G(\beta_{n-1}) - \gamma\nabla g(\theta^*, X_n).$$

The sequence β_n can be viewed as the gradient descent iteration with i.i.d. noise. The following lemma characterizes the convergence of β_n and the approximation error rate.

Lemma 4.1. *Under same conditions of Theorem 2.3,*

$$\|\beta_n - \theta^*\|_p^2 \le \frac{3(p-1)M_p^2}{\alpha_0}\gamma,$$

and for some constant α_1 such that

$$\alpha_1 \le \min\{2\mu - \gamma L_p^2, \frac{1}{\gamma}\},$$

we have

$$\|\theta_n - \beta_n\|_p^2 \le [(1-\alpha_1\gamma)^n + 4(p-1)L_p^2 n\gamma^2(1-\alpha_0\gamma)^n]\rho_p^2 + \frac{3(p-1)M_p^2}{\alpha_0\alpha_1}\gamma^2.$$

Remark. *Compared to Lemma 3.3, the non-linear autoregression sequence can approximate the SGD in L_p-space for $p \ge 2$, which is more useful for a precise derivation of tail probability.*

4.2 Nagaev-type inequality of tail probability

To analyze the concentration property of SGD with fixed learning rates, we first focus on the linear functional of the estimation error, i.e., $v^\top(\theta_n - \theta^*)$ for all $v \in \mathbb{S}^d$, the unit sphere in $\mathbb{R}^d$. The tail probability bounds directly from the L_p convergence or Gaussian approximation rate are too conservative. Simply applying the Berry-Esseen inequality 3.5, for instance, will result in a sub-Gaussian tail plus an inevitable $\sqrt{\gamma}$ term. Another naive way is to use the Markov inequality on Theorem 2.3. If we set the degree of tolerance as ϵ, it yields

$$\mathbb{P}(|v^\top(\theta_n - \theta^*)| > \epsilon) \lesssim \frac{1}{\epsilon^p}\big[(1-\alpha_0\gamma)^{np/2}\rho_p^p + M_p^p\alpha_0^{-p/2}\gamma^{p/2}\big], \tag{7}$$

where the polynomial term $\gamma^{p/2}$ is far from optimal. As a result, for some credible level $0 < \delta < 1$, with probability $1-\delta$, one can only have

$$|v^\top(\theta_n - \theta^*)| = \mathcal{O}(\delta^{-1/p}\sqrt{\gamma}).$$

According to $n \gtrsim \gamma^{-1}$, the resulting estimate of sample complexity $N(\epsilon, \delta)$ (i.e., the minimum number of iterations such that the MSE satisfies the given credible level δ and degree of tolerance ϵ),

$$N(\epsilon, \delta) = \mathcal{O}(\frac{\delta^{-2/p}}{\epsilon}),$$

grows rapidly for any tolerance level $\delta \in (0, 1)$.

The next theorem provides a Nagaev-type high-confidence bound that is substantially tighter than existing results by incorporating functional dependence measures, non-linear autoregressive approximation and the moment-generating function.

Theorem 4.2 (Nagaev inequality)**.** *Under same conditions of Theorem 2.3, for any $\epsilon > 0$, we have*

$$\sup_{v \in \mathbb{S}^d} \mathbb{P}(|\langle v, \Delta_n \rangle| > \epsilon) \leq \frac{C_1 \gamma^{p-1}}{\epsilon^p} + 2\exp\{-\frac{C_2 \epsilon^2}{\gamma}\} + \frac{2^{p-1} \rho_p^p (1 - \alpha_0 \gamma)^{np/2}}{\epsilon^p} \Big(1 + 4(p-1) L_p^2 n \gamma^2\Big)^{p/2},$$

where C_1 and C_2 are some positive constants independent of ϵ, n, γ and θ_0.

Remark (Rate of tail probability)**.** *The polynomial term in Theorem 4.2 is γ^{p-1}, much sharper than those obtained from Gaussian approximation or Markov inequality when $p > 2$. The sub-Gaussian term $\exp\{-C_2\epsilon^2/\gamma\}$ is optimal in the sense that $\Delta_n/\sqrt{\gamma}$ is asymptotically normal. With a high degree of tolerance ϵ, the polynomial term dominates, and the estimate of sample complexity can be greatly improved to $\mathcal{O}(\delta^{-1/(p-1)}\epsilon^{-p/2(p-1)})$ in this case. With a low degree of tolerance, the sub-Gaussian term dominates.*

We can directly obtain the same result of the uniform distance between θ_n and θ^* by the union bound. Taking v over the standard basis $(0, ..., 0, 1, 0, ..., 0)^\top$ yields

$$\begin{aligned}\mathbb{P}(|\|\Delta_n\|_\infty > \epsilon) \leq & \frac{C_1 d \gamma^{p-1}}{\epsilon^p} + 2d\exp\{-\frac{C_2\epsilon^2}{\gamma}\} \\ & + \frac{2^{p-1} d \rho_p^p (1 - \alpha_0\gamma)^{np/2}}{\epsilon^p}[1 + 4(p-1)L_p^2 n\gamma^2]^{p/2}.\end{aligned}$$

5 Numerical studies

5.1 Simulation setting

We conduct a simulation to demonstrate that our Nagaev-type inequality in Theorem 4.2 is indeed valid and tight. Consider the following data generating mechanism for the logistic regression model: $X_i = (a_i, b_i), i = 1, 2, ...$ are i.i.d. random vectors where a_i are generated from a 5-dimensional independent t distribution with degrees of freedom df $= 3$. $b_i \in \{1, -1\}$ follows a Bernoulli distribution with the probability given by $\mathbb{P}(b_i|a_i) = 1/(1 + \exp(-b_i a_i^\top \theta^*))$. The loss function is defined as the negative log-likelihood,

$$g(\theta, X_i) = \log(1 + \exp(-b_i a_i^\top \theta)).$$

We investigate logistic regression for its non-linearity. The t distribution with df $= \nu$ only has finite p-th moments with $p < \nu$. This property enables us to study the performance of different tail probability bounds with specific values of p. Elementary calculation shows that

$$\nabla g(\theta, X) = \nabla g(\theta, a, b) = \frac{-ba}{1 + \exp(ba^\top \theta)},$$

and therefore $\nabla g(\theta^*, X)$ only has finite p-th moment with $p < 3$ in our setting. Theoretically, we can choose p arbitrarily close to 3 and apply the inequalities discussed in Section 4.2. For simplicity, we take $p = 3$ since the results we report in the simulation are continuous with respect to p.

We run 1000 independent trials with $n = 500000$ and $\gamma = 0.005, 0.001, 0.0002$. Since the number of iterations is large enough, the main contribution in the tail probability bounds are the polynomial terms and the sub-Gaussian term, i.e.,

$$II_1 = \frac{\gamma^{p/2}}{\epsilon^p}$$

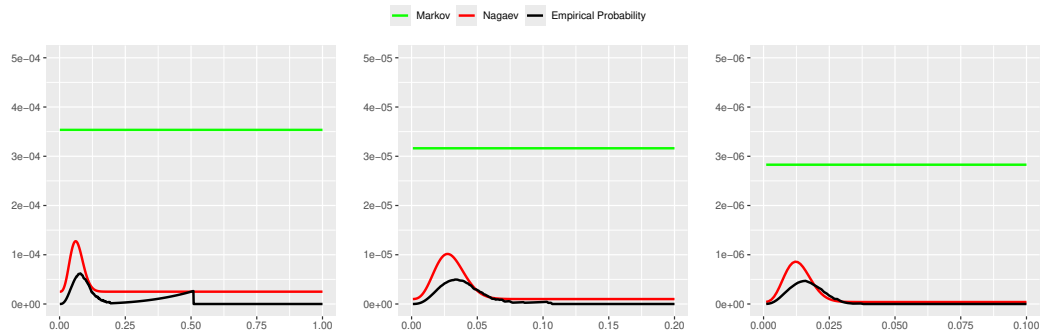


Figure 1: Comparison of different tail probability bounds against the empirical probability. The learning rate $\gamma = 0.005, 0.001, 0.0002$ from left to right.

in the Markov-type bound, and

$$II_2 = \frac{\gamma^{p-1}}{\epsilon^p} + 2\exp\{-\frac{C_2\epsilon^2}{\gamma}\}$$

in the Nagaev-type bound. In our simulation, we set $C_2 = 2$, because Theorem 3.4 provides the equation that the asymptotic covariance Γ satisfies. Elementary calculation Chen et al. [2020] shows that for the logistic model, $A = S$, and the unique solution to $A\Gamma + \Gamma A = S$ becomes $\Gamma = \mathbf{I}_d/2$, so we accordingly scale the sub-Gaussian term.

5.2 Numerical results

We compare the concentration inequality (7) and our Nagaev-type bound, Theorem 4.2, both with the empirical probability: $II_3 = \text{Avg}_v(\mathbb{P}(|v^\top(\theta_n - \theta^*)| > \epsilon))$, where v ranges over the standard basis vectors $(0, \ldots, 0, 1, 0, \ldots, 0)^\top$. For more transparent visualization, we multiply these quantities by ϵ^p and plot the results against the degree of tolerance ϵ. In Figure 1, the x-axis is the degree of tolerance. The green, red, and black curves are $\epsilon^p II_1$, $\epsilon^p II_2$, and $\epsilon^p II_3$, respectively. They represent the ϵ^p-scaled Markov bound, Nagaev bound, and empirical probability.

Figure 1 clearly indicates that the tail probability bound from Markov inequality is excessively conservative. In contrast, our Nagaev-type inequality from Theorem 4.2 yields a much sharper upper bound for SGD across different learning-rate scales. The shape of empirical probability closely matches the dichotomous phenomenon in theory: the dominance transits from the polynomial term to the sub-Gaussian term as ϵ decreases. The experiment results confirm that our Nagaev-type bound is both valid and tight, precisely describing the tail behavior of constant learning rate SGD.

Acknowledgments and Disclosure of Funding

We sincerely thank the program chair, senior area chair, area chair, and the five 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). Wei Biao Wu's research is partially supported by the NSF (Grant NSF/DMS-2311249). We would like to thank Johannes Schmidt-Hieber for helpful discussions.

References

Agrawalla, B., K. Balasubramanian, and P. Ghosal (2023). High-dimensional central limit theorems for linear functionals of online least-squares sgd. *arXiv e-prints*, arXiv–2302.

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

Berry, A. C. (1941). The accuracy of the gaussian approximation to the sum of independent variates. *Transactions of the American Mathematical Society 49*(1), 122–136.

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

Cardot, H., P. Cénac, and A. Godichon-Baggioni (2017). Online estimation of the geometric median in Hilbert spaces: Nonasymptotic confidence balls. *The Annals of Statistics 45*(2), 591 – 614.

Cardot, H., P. Cénac, and P.-A. Zitt (2013). Efficient and fast estimation of the geometric median in Hilbert spaces with an averaged stochastic gradient algorithm. *Bernoulli 19*(1), 18 – 43.

Chee, J. and P. Toulis (2018). Convergence diagnostics for stochastic gradient descent with constant learning rate. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, pp. 1476–1485. PMLR.

Chen, L. H. and Q.-M. Shao (2001). A non-uniform berry–esseen bound via stein's method. *Probability Theory and Related Fields 120*, 236–254.

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

Chen, Z., S. Mou, and S. T. Maguluri (2022). Stationary behavior of constant stepsize sgd type algorithms: An asymptotic characterization. *Proceedings of the ACM on Measurement and Analysis of Computing Systems 6*(1), 1–24.

Dahlhaus, R. and S. S. Rao (2006). Statistical inference for time-varying ARCH processes. *The Annals of Statistics 34*(3), 1075–1114.

Dahlhaus, R., S. Richter, and W. B. Wu (2019). Towards a general theory for nonlinear locally stationary processes. *Bernoulli 25*(2), 1013–1044.

Defossez, A. and F. Bach (2015). Averaged least-mean-squares: Bias-variance trade-offs and optimal sampling distributions. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*, pp. 205–213.

Devroye, L., A. Mehrabian, and T. Reddad (2018). The total variation distance between high-dimensional gaussians with the same mean. *arXiv preprint arXiv:1810.08693*.

Dieuleveut, A. and F. Bach (2016). Nonparametric stochastic approximation with large step-sizes. *The Annals of Statistics 44*(4), 1363–1399.

Dieuleveut, A., A. Durmus, and F. Bach (2020). Bridging the gap between constant step size stochastic gradient descent and Markov chains. *The Annals of Statistics 48*(3), 1348–1382.

Dieuleveut, A., N. Flammarion, and F. Bach (2017). Harder, better, faster, stronger convergence rates for least-squares regression. *The Journal of Machine Learning Research 18*(1), 3520–3570.

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

Durmus, A., E. Moulines, A. Naumov, S. Samsonov, K. Scaman, and H.-T. Wai (2021). Tight high probability bounds for linear stochastic approximation with fixed stepsize. *Advances in Neural Information Processing Systems 34*, 30063–30074.

Esseen, C.-G. (1942). On the liapunov limit error in the theory of probability. *Ark. Mat. Astr. Fys. 28*, 1–19.

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

Fang, Y. (2019). Scalable statistical inference for averaged implicit stochastic gradient descent. *Scandinavian Journal of Statistics 46*(4), 987–1002.

Fang, Y., J. Xu, and L. Yang (2018). Online bootstrap confidence intervals for the stochastic gradient descent estimator. *Journal of Machine Learning Research 19*, 1–21.

Gandikota, V., D. Kane, R. K. Maity, and A. Mazumdar (2022). vqsgd: Vector quantized stochastic gradient descent. *IEEE Transactions on Information Theory 68*(7), 4573–4587.

Ge, R., S. M. Kakade, R. Kidambi, and P. Netrapalli (2019). The step decay schedule: A near optimal, geometrically decaying learning rate procedure for least squares. *arXiv preprint*. arXiv:1904.12838.

Gower, R. M., N. Loizou, X. Qian, A. Sailanbayev, E. Shulgin, and P. Richtarik (2019). SGD: General analysis and improved rates. *Proceedings of the 36th International Conference on Machine Learning 97*, 5200–5209.

Györfi, L. and H. Walk (1996). On the averaged stochastic approximation for linear regression. *SIAM Journal on Control and Optimization 34*(1), 31–61.

Hu, J. and M. C. Fu (2024). Technical note—on the convergence rate of stochastic approximation for gradient-based stochastic optimization. *Operations Research*. To appear.

Huang, Y., X. Chen, and W. B. Wu (2014). Recursive nonparametric estimation for time series. *IEEE Transactions on Information Theory 60*(2), 1301–1312.

Huo, D., Y. Chen, and Q. Xie (2023). Bias and extrapolation in markovian linear stochastic approximation with constant stepsizes. *arXiv preprint*. arXiv:2210.00953.

Jain, P., S. M. Kakade, R. Kidambi, P. Netrapalli, and A. Sidford (2018). Parallelizing stochastic gradient descent for least squares regression: Mini-batching, averaging, and model misspecification. *Journal of Machine Learning Research 18*(223), 1–42.

Jiang, X. and S. U. Stich (2024). Adaptive SGD with polyak stepsize and line-search: robust convergence and variance reduction. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, pp. 26396–26424. Curran Associates Inc.

Korolev, V. Y. and I. G. Shevtsova (2010). On the upper bound for the absolute constant in the berry–esseen inequality. *Theory of Probability & Its Applications 54*(4), 638–658.

Kushner, H. J. and G. G. Yin (1997). Stochastic approximation and recursive algorithm and applications. *Application of Mathematics 35*(10).

Lai, T. L. (2003). Stochastic approximation: invited paper. *The Annals of Statistics 31*(2), 391 – 406.

Lauand, C. K. and S. Meyn (2024). Revisiting step-size assumptions in stochastic approximation. *arXiv preprint*. arXiv:2405.17834.

Li, C. J., W. Mou, M. Wainwright, and M. Jordan (2022). Root-sgd: Sharp nonasymptotics and asymptotic efficiency in a single algorithm. In *Conference on Learning Theory*, pp. 909–981. PMLR.

Li, J., Z. Lou, S. Richter, and W. B. Wu (2024). The stochastic gradient descent from a nonlinear time series persective. Manuscript.

Li, J., J. Schmidt-Hieber, and W. B. Wu (2024). Asymptotics of stochastic gradient descent with dropout regularization in linear models. *arXiv preprint*. arXiv:2409.07434.

Ljung, L. (1977). Analysis of recursive stochastic algorithms. *IEEE Transactions on Automatic Control 22*(4), 551–575.

Loizou, N., S. Vaswani, I. H. Laradji, and S. Lacoste-Julien (2021). Stochastic polyak step-size for SGD: An adaptive learning rate for fast convergence. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, pp. 1306–1314. PMLR.

Merad, I. and S. Gaïffas (2023). Convergence and concentration properties of constant step-size SGD through Markov chains. *arXiv preprint*. arXiv:2306.11497.

Moulines, E. and F. Bach (2011). Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems*, pp. 856–864.

Muecke, N., G. Neu, and L. Rosasco (2019). Beating sgd saturation with tail-averaging and minibatching. In *Advances in Neural Information Processing Systems*, Volume 32. Curran Associates, Inc.

Nagaev, S. V. (1979). Large deviations of sums of independent random variables. *The Annals of Probability*, 745–789.

Nemirovski, A., A. Juditsky, G. Lan, and A. Shapiro (2009). Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization 19*(4), 1574–1609.

Nguyen, P. H., L. Nguyen, and M. van Dijk (2019). Tight Dimension Independent Lower Bound on the Expected Convergence Rate for Diminishing Step Sizes in SGD. In *Advances in Neural Information Processing Systems*, Volume 32. Curran Associates, Inc.

Pflug, G. C. (1986). Stochastic minimization with constant step-size: asymptotic laws. *SIAM Journal on Control and Optimization 24*(4), 655–666.

Pillaud-Vivien, L., A. Rudi, and F. Bach (2018). Statistical optimality of stochastic gradient descent on hard learning problems through multiple passes. In *Advances in Neural Information Processing Systems*, Volume 31. Curran Associates, Inc.

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

Raič, M. (2019). A multivariate berry–esseen theorem with explicit constants. *Bernoulli 25*(4A), 2824–2853.

Rakhlin, A., O. Shamir, and K. Sridharan (2012). Making gradient descent optimal for strongly convex stochastic optimization. *arXiv preprint*. arXiv:1109.5647.

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

Robbins, H. and S. Monro (1951). A stochastic approximation method. *The Annals of Mathematical Statistics 22*(4), 400–407.

Robbins, H. and D. Siegmund (1971). A convergence theorem for non-negative almost supermartingales and some applications. In J. S. Rustagi (Ed.), *Optimizing Methods in Statistics*, pp. 233–257. Academic Press.

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

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

Samsonov, S., E. Moulines, Q.-M. Shao, Z.-S. Zhang, and A. Naumov (2024). Gaussian approximation and multiplier bootstrap for polyak-ruppert averaged linear stochastic approximation with applications to td learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Shamir, O. and T. Zhang (2013). Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In *International Conference on Machine Learning*, pp. 71–79. PMLR.

Shao, Q.-M. and Z.-S. Zhang (2022). Berry–Esseen bounds for multivariate nonlinear statistics with applications to M-estimators and stochastic gradient descent algorithms. *Bernoulli 28*(3), 1548 – 1576.

Sheshukova, M., S. Samsonov, D. Belomestny, E. Moulines, Q.-M. Shao, Z.-S. Zhang, and A. Naumov (2025). Gaussian approximation and multiplier bootstrap for stochastic gradient descent. *arXiv preprint arXiv:2502.06719*.

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

Valiant, L. G. (1984). A theory of the learnable. *Communications of the ACM 27*(11), 1134–1142.

Wang, X. and N. Gao (2010). Stochastic resource allocation over fading multiple access and broadcast channels. *IEEE Transactions on Information Theory 56*(5), 2382–2391.

Wang, X. and Y.-x. Yuan (2023). On the convergence of stochastic gradient descent with bandwidth-based step size. *Journal of Machine Learning Research 24*(48), 1–49.

Wei, Z., J. Li, L. Chen, and W. B. Wu (2025). Online inference for quantiles by constant learning-rate stochastic gradient descent. *arXiv preprint arXiv:2503.02178*.

Woodroofe, M. (1972). Normal approximation and large deviations for the robbins-monro process. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 329–338.

Wu, W. B. (2005). Nonlinear system theory: Another look at dependence. *Proceedings of the National Academy of Sciences 102*(40), 14150–14154.

Zhong, Y., J. Li, and S. Lahiri (2024). Probabilistic guarantees of stochastic recursive gradient in non-convex finite sum problems. In *Advances in Knowledge Discovery and Data Mining*, Singapore, pp. 142–154. Springer Nature Singapore.

Zhu, W., X. Chen, and W. B. Wu (2023). Online covariance matrix estimation in stochastic gradient descent. *Journal of the American Statistical Association 118*(541), 393–404.

Zhu, W., Z. Lou, and W. B. Wu (2022). Beyond sub-gaussian noises: Sharp concentration analysis for stochastic gradient descent. *Journal of Machine Learning Research 23*(46), 1–22.

A Technical Appendices and Supplementary Material

All theoretical results are proved in the appendix. We first introduce some technical lemma.

Lemma A.1 (Rio's inequality [Rio, 2009])**.** *Let $X \in \mathbb{R}^d$ and $Y \in \mathbb{R}^d$ be two random vectors such that $\mathbb{E}|X|^p < \infty$ and $\mathbb{E}|Y|^p < \infty$ for some $p \geq 2$. Then we have*

$$\|X + Y\|_p^2 \leq \|X\|_p^2 + (p-1)\|Y\|_p^2. \tag{8}$$

Lemma A.2. *Under Assumption 2.1 and 2.2, we have*

$$|\nabla G(\theta_1) - \nabla G(\theta_2)| \leq L_p|\theta_1 - \theta_2|.$$

Proof. By Assumption 2.2,

$$\|\nabla g(\theta_1, X) - \nabla g(\theta_2, X)\|_2^2 \leq L_p^2|\theta_1 - \theta_2|^2, \quad \text{for all } \theta_1, \theta_2 \in \mathbb{R}^d.$$

By the convexity of $|\cdot|^2$ and Jensen's inequality,

$$\begin{aligned}|\nabla G(\theta_1) - \nabla G(\theta_2)|^2 &= |\mathbb{E}_X(\nabla g(\theta_1, X) - \nabla g(\theta_2, X))|^2 \\ &\leq \mathbb{E}_X|\nabla g(\theta_1, X) - \nabla g(\theta_2, X)|^2 \\ &\leq L_p^2|\theta_1 - \theta_2|^2.\end{aligned}$$

□

A.1 Proof of Moments Convergence of SGD

Lemma A.3. *Consider the SGD iterates $\{\theta_t\}_{t\geq 1}$ in* (2)*. Under the same conditions of Theorem 2.3, for some universal constant α_0 such that*

$$0 < \alpha_0 \leq \min\Big\{\frac{1}{\gamma}, 2\mu - (6p-5)L_p^2\gamma\Big\}, \tag{9}$$

we have, for all $n \geq 1$,

$$\|\Delta_n\|_p^2 \leq (1 - \alpha_0\gamma)\|\Delta_{n-1}\|_p^2 + 3(p-1)\gamma^2 M_p^2. \tag{10}$$

Proof of Lemma A.3. Since ξ_t, for $t \geq 1$, are i.i.d. random samples, it follows from the tower rule that

$$\mathbb{E}[\nabla g(\theta_{n-1}, \xi_n) - \nabla G(\theta_{n-1}) \mid \theta_{n-1}] = 0. \tag{11}$$

Therefore, by applying Rio's inequality in Lemma A.1, for $p \geq 2$, we have

$$\begin{aligned}\|\Delta_n\|_p^2 &\leq \|\theta_{n-1} - \theta^* - \gamma\nabla G(\theta_{n-1})\|_p^2 + (p-1)\gamma^2\|\nabla g(\theta_{n-1}, \xi_n) - \nabla G(\theta_{n-1})\|_p^2 \\ &=: \mathbb{I}_1 + \mathbb{I}_2.\end{aligned} \tag{12}$$

We shall bound the two parts $\mathbb{I}_1$ and $\mathbb{I}_2$ separately. For the first part $\mathbb{I}_1$, note that $\nabla G(\theta^*) = 0$ and by the triangle inequality, we have

$$\begin{aligned}\mathbb{I}_1 &= \|\theta_{n-1} - \theta^* - \gamma\nabla G(\theta_{n-1})\|_p^2 \\ &= \Big\|\big\langle\theta_{n-1} - \theta^*, \theta_{n-1} - \theta^*\big\rangle - 2\gamma\big\langle\theta_{n-1} - \theta^*, \nabla G(\theta_{n-1}) - \nabla G(\theta^*)\big\rangle \\ &\quad + \gamma^2\big\langle\nabla G(\theta_{n-1}) - \nabla G(\theta^*), \nabla G(\theta_{n-1}) - \nabla G(\theta^*)\big\rangle\Big\|_{p/2} \\ &\leq \Big\|\big\langle\theta_{n-1} - \theta^*, \theta_{n-1} - \theta^*\big\rangle - 2\gamma\big\langle\theta_{n-1} - \theta^*, \nabla G(\theta_{n-1}) - \nabla G(\theta^*)\big\rangle\Big\|_{p/2} \\ &\quad + \gamma^2\big\|\nabla G(\theta_{n-1}) - \nabla G(\theta^*)\big\|_p^2.\end{aligned} \tag{13}$$

By applying Assumption 2.1 to the first term and Assumption 2.2 to the second term, we can obtain

$$\mathbb{I}_1 \leq (1 - 2\gamma\mu + \gamma^2 L_p^2)\|\theta_{n-1} - \theta^*\|_p^2. \tag{14}$$

Regarding the second part $\mathbb{I}_2$, since $\nabla G(\theta^*) = 0$, we have

$$\begin{aligned}&\|\nabla g(\theta_{n-1}, \xi_n) - \nabla G(\theta_{n-1})\|_p \\ &\le \|\nabla g(\theta_{n-1}, \xi_n) - \nabla g(\theta^*, \xi_n)\|_p + \|\nabla G(\theta_{n-1}) - \nabla G(\theta^*)\|_p + \|\nabla g(\theta^*, \xi_n)\|_p. \end{aligned} \tag{15}$$

Hence, by Assumption 2.2, we can achieve

$$\|\nabla g(\theta_{n-1}, \xi_n) - \nabla G(\theta_{n-1})\|_p^2 \le 6L_p^2 \|\theta_{n-1} - \theta^*\|_p^2 + 3\|\nabla g(\theta^*, \xi_n)\|_p^2. \tag{16}$$

Combining results from $\mathbb{I}_1$ and $\mathbb{I}_2$, we can obtain

$$\|\Delta_n\|_p^2 \le (1 - 2\gamma\mu + (6p-5)\gamma^2 L_p^2)\|\theta_{n-1} - \theta^*\|_p^2 + 3(p-1)\gamma^2 \|\nabla g(\theta^*, \xi_n)\|_p^2.$$

This can directly lead to the desired inequality. □

Proof of Theorem 2.3. By recursively applying Lemma A.3, we have

$$\|\Delta_n\|_p^2 \le \prod_{k=1}^{n}(1 - \alpha_0\gamma)\|\Delta_0\|_p^2 + 3(p-1)M_p^2 \sum_{j=1}^{n}\gamma^2 \prod_{k=j+1}^{n}(1 - \alpha_0\gamma). \tag{17}$$

By elementary calculations,

$$\|\Delta_n\|_p^2 \le (1 - \alpha_0\gamma)^n \rho_p^2 + 3(p-1)M_p^2 \gamma \alpha_0^{-1}. \tag{18}$$

□

A.2 Proof of the Bound for Functional Dependence Measure

Recall that the SGD sequence θ_n can be represented by $\theta_n = \tau_n(X_1, ..., X_n)$ and $\theta_n^{(t)} = \tau_n(X_1, ..., X_{t-1}, X_t', X_{t+1}, ..., X_n)$ for some measurable function τ_n that can vary for different n, where X_t' is an i.i.d. copy of X_t. The functional dependence measure was defined as $\psi(n, t, p) = \|\theta_n - \theta_n^{(t)}\|_p$. We prove the bound for $\psi(n, t, p)$ as stated in Theorem 3.1, which is fundamental to the proofs of CLT, Berry-Esseen inequality and Nagaev-type inequality.

Proof. By applying Rio's inequality, for each $t \le n-1$, we have

$$\begin{aligned}\|\theta_n - \theta_n^{(t)}\|_p^2 &\le (1 - 2\gamma\mu + (6p-5)\gamma^2 L_p^2)^{n-t}\|\theta_t - \theta_t^{(t)}\|_p^2 \\ &\le (1 - \alpha_0\gamma)^{n-t}\|\theta_t - \theta_t^{(t)}\|_p^2. \end{aligned} \tag{19}$$

It follows from Assumption 2.2 that for all $t \ge 1$,

$$\begin{aligned}\|\nabla g(\theta_{t-1}, X_t)\|_p^2 &\le 2\|\nabla g(\theta_{t-1}, X_t) - \nabla g(\theta^*, X_t)\|_p^2 + 2\|\nabla g(\theta^*, X_t)\|_p^2 \\ &\le 2L_p^2\|\theta_{t-1} - \theta^*\|_p^2 + 2M_p^2. \end{aligned} \tag{20}$$

As a direct consequence, we can achieve

$$\begin{aligned}\|\theta_t - \theta_t^{(t)}\|_p^2 &= \gamma^2\|\nabla g(\theta_{t-1}, X_t) - \nabla g(\theta_{t-1}, X_t')\|_p^2 \\ &\le \gamma^2\big(2\|\nabla g(\theta_{t-1}, X_t)\|_p^2 + 2\|\nabla g(\theta_{t-1}, X_t')\|_p^2\big) \\ &\le 4\gamma^2\big(L_p^2\|\theta_{t-1} - \theta^*\|_p^2 + M_p^2\big). \end{aligned} \tag{21}$$

This along with expression (19) and Theorem 2.3 provides the desired result. □

A.3 Proof of Quenched CLT

Here and in the sequel, we will repeatedly use a basic property that $|\mathbf{I}_d - \gamma A| \le 1 - \gamma\lambda_* \le 1 - \gamma\lambda$. We denote $Z_i = \nabla g(\theta^*, X_i)$ as the gradient noise.

Proof of Lemma 3.3. We first show that $I_{1,n}$ and $I_{2,n}$ vanish. Define $\mathcal{F}_0 = \emptyset$ and $\mathcal{F}_t = \sigma(X_1, ..., X_t)$ as the filtration generated by the data. It is clear that $\{D_n\}$ is a martingale difference sequence w.r.t. the filtration $\{\mathcal{F}_t\}$, and we have

$$\begin{aligned}
\|I_{2,n}\|^2 &\leq \gamma^2 \sum_{k=1}^{n} \|(\mathbf{I}_d - \gamma A)\|^{2n-2k} \|D_k\|^2 \\
&\leq 4\gamma^2 L_2^2 \sum_{k=1}^{n} \|(\mathbf{I}_d - \gamma A)\|^{2n-2k} \|\theta_{k-1} - \theta^*\|^2 \\
&\leq 4\gamma^2 L_2^2 \sum_{k=1}^{n} \|(\mathbf{I}_d - \gamma A)\|^{2n-2k} (1-\alpha_0\gamma)^{k-1} \rho_2^2 + 12 L_2^2 \gamma^3 M_2^2 \alpha_0^{-1} \sum_{k=1}^{n} \|(\mathbf{I}_d - \gamma A)\|^{2n-2k} \\
&\leq 4\gamma^2 L_2^2 \rho_2^2 \frac{(1-\gamma\lambda)^{n-1} - (1-\gamma\lambda)^{2n-1}}{\gamma\lambda} + 12 L_2^2 \gamma^3 M_2^2 \alpha_0^{-1} \frac{1-(1-\gamma\lambda)^{2n}}{2\gamma\lambda - \gamma^2\lambda^2} \\
&\leq 4 L_2^2 \rho_2^2 \gamma \frac{(1-\lambda\gamma)^{n-1}}{\lambda} + \frac{12 L_2^2 \gamma^2 M_2^2}{\alpha_0(2\lambda - \gamma\lambda^2)} \\
&\leq 4 L_2^2 \rho_2^2 \gamma \frac{(1-\lambda\gamma)^{n-1}}{\lambda} + \frac{12 L_2^2 \gamma^2 M_2^2}{\alpha_0 \lambda}.
\end{aligned}$$

Here we use the fact that $|D_n| \leq 2L_2|\Delta_{n-1}|$ due to Assumption 2.2, Lemma A.2, and the triangular inequality. The last inequality comes from $\gamma \leq 1/\alpha_0$ and $\gamma \leq 1/\lambda_*$. By Taylor expansion around θ^*, since $\nabla G(\theta^*) = 0$, we have

$$\begin{aligned}
R_n &= A\Delta_{n-1} - (\nabla G(\theta_{n-1}) - \nabla G(\theta^*)) \\
&= -\int_0^1 [\nabla^2 G(\theta^* + t(\theta_{n-1} - \theta^*)) - \nabla^2 G(\theta^*)](\theta_{n-1} - \theta^*) dt.
\end{aligned}$$

By Assumption 3.2, when $|\theta_{n-1} - \theta^*| \leq \kappa$, we have $|R_n| \leq L|\theta_{n-1} - \theta^*|^2$. For $|\theta_{n-1} - \theta^*| > \kappa$, the Lipschitz continuity of the gradient implies $|R_n| \leq 2L_2|\theta_{n-1} - \theta^*| \leq 2L_2\kappa^{-1}|\theta_{n-1} - \theta^*|^2$. So we finally have $|R_n| \leq C_0|\Delta_{n-1}|^2$ where $C_0 = \max\{L, 2L_2\kappa^{-1}\}$. As a result,

$$\begin{aligned}
\mathbb{E}|I_{1,n}| &\leq C_0\gamma \sum_{k=1}^{n} \|\mathbf{I}_d - \gamma A\|^{n-k} \|\theta_{k-1} - \theta^*\|^2 \\
&\leq n\gamma C_0 \rho_2^2 (1-\gamma\lambda)^{n-1} + 3C_0 M_2^2 \alpha_0^{-1} \gamma^2 \frac{1-(1-\gamma\lambda)^n}{\gamma\lambda} \\
&\leq n\gamma C_0 \rho_2^2 (1-\gamma\lambda)^{n-1} + \frac{3C_0 M_2^2}{\lambda\alpha_0}\gamma. \qquad (22)
\end{aligned}$$

We also have $\|\mathbf{I}_d - \gamma A)^n \Delta_0\| \leq (1-\gamma_*\lambda)^n \rho_1$. Combining these inequalities will lead to the bound in Lemma 3.3. □

Proof of Theorem 3.4. As $\gamma \to 0$ and $n\gamma \geq \nu \log n$ for some constant $\nu > 1/2\lambda$, elementary calculation shows that the difference between $\Delta_n/\sqrt{\gamma}$ and $L_n/\sqrt{\gamma}$ goes to 0, i.e., by Lemma 3.3 we have $\|\Delta_n/\sqrt{\gamma} - L_n/\sqrt{\gamma}\|_1 \to 0$. Then it suffices to prove that

$$\frac{L_n}{\sqrt{\gamma}} \xrightarrow{D} \mathcal{N}(0, \Gamma). \qquad (23)$$

Notice that $L_n/\sqrt{\gamma}$ is a linear combination of i.i.d. random vectors, the covariance matrix of which is

$$\Gamma_n(\gamma) = \gamma \sum_{k=1}^{n} (\mathbf{I}_d - \gamma A)^{n-k} S (\mathbf{I}_d - \gamma A)^{n-k}.$$

We introduce the following auxiliary lemma.

Lemma A.4. *The minimum eigenvalue of $\Gamma_n(\gamma)$, denoted as $\lambda_n(\gamma)$, satisfies*

$$\lambda_n(\gamma) \geq \frac{\lambda_S(1-(1-\gamma\lambda^*)^{2n})}{2\lambda^*},$$

where λ_S is the smallest eigenvalue of S.

This lemma is easy to prove since $(\mathbf{I}_d - \gamma A)$ is positive definite with the minimum eigenvalue $1-\gamma\lambda^*$. So we have

$$\lambda_n(\gamma) \geq \lambda \sum_{k=1}^{n} (1-\gamma\lambda^*)^{2n-2k}\lambda_S$$

and elementary calculation leads to the conclusion. The lemma implies that $|\Gamma_n(\gamma)^{-1}|$ is bounded by some constant. As $\gamma \to 0$ and $n\gamma \to \infty$, we have

$$\max_{1\leq k\leq n} |\Gamma_n(\gamma)^{-1}\gamma(\mathbf{I}_d - \gamma A)^{n-k} S (\mathbf{I}_d - \gamma A)^{n-k}| \lesssim \gamma \to 0.$$

By the multivariate Lindeberg-Feller CLT, it is clear that

$$\Gamma_n(\gamma)^{-1/2} \frac{L_n}{\sqrt{\gamma}} \xrightarrow{D} \mathcal{N}(0, \mathbf{I}_d). \tag{24}$$

To determine the closeness between $\Gamma_n(\gamma)$ and Γ, notice that

$$\Gamma_{n+1}(\gamma) = (\mathbf{I}_d - \gamma A)\Gamma_n(\gamma)(\mathbf{I}_d - \gamma A) + \gamma S.$$

Minus Γ on both sides and plug $A\Gamma + \Gamma A = S$ into the formula above, we get

$$\Gamma_{n+1}(\gamma) - \Gamma = (\mathbf{I}_d - \gamma A)(\Gamma_n(\gamma) - \Gamma)(\mathbf{I}_d - \gamma A) + \gamma^2 A\Gamma A.$$

Hence, there exists a universal constant C such that

$$|\Gamma_{n+1}(\gamma) - \Gamma| \leq (1-\gamma\lambda)^2 |\Gamma_n(\gamma) - \Gamma| + C\gamma^2.$$

Let $\Gamma_0(\gamma) = \mathbf{0}_{d\times d}$. Recursively updating the inequality we get

$$|\Gamma_n(\gamma) - \Gamma| \leq \sum_{i=1}^{n} C\gamma^2 (1-\gamma\lambda)^{2(n-i)} + (1-\gamma\lambda)^{2n} |\Gamma_0(\gamma) - \Gamma| \tag{25}$$

$$\leq \frac{C\gamma}{2\lambda - \gamma\lambda^2} + (1-\gamma\lambda)^{2n}|\Gamma| \to 0. \tag{26}$$

Since the eigenvalues of $\Gamma_n(\gamma)$ are bounded and bounded away from 0, and Γ is a fixed positive definite matrix, we have $\Gamma^{-1/2}\Gamma_n(\gamma)^{1/2} \to \mathbf{I}_d$. By Slustky's theorem,

$$\Gamma^{-1/2}\frac{L_n}{\sqrt{\gamma}} = \Gamma^{-1/2}\Gamma_n(\gamma)^{1/2}\Gamma_n(\gamma)^{-1/2}\frac{L_n}{\sqrt{\gamma}} \xrightarrow{D} \mathcal{N}(0, \mathbf{I}_d),$$

and we proved (23). □

Remark on the multivariate Lindeberg–Feller CLT

Suppose a triangular sequence $y_{n,k} \in \mathbb{R}^d$ are independent with means $\mathbb{E}y_{n,k} = 0$ and covariance matrices $V_{n,k} = \mathbb{E}(y_{n,k}y_{n,k}^\top)$. Set

$$U_n = \sum_{k=1}^{n} V_{n,k}, \quad \nu_n^2 = \lambda_{\min}(U_n).$$

If $\nu_n^2 > 0$ and for all $\varepsilon > 0$

$$\lim_{n\to\infty} \frac{1}{\nu_n^2} \sum_{k=1}^{n} \mathbb{E}\big(|y_{n,k}|^2 \cdot \mathbf{1}(|y_{n,k}|^2 > \varepsilon\nu_n^2)\big) = 0,$$

then as $n \to \infty$

$$U_n^{-1/2}\left(\sum_{k=1}^{n} y_{n,k}\right) \Rightarrow \mathcal{N}(0, I_d).$$

Above is the statement of the multivariate Lindeberg–Feller CLT. We apply the theorem to the standardized sum

$$\Gamma_n(\gamma)^{-1/2}\sqrt{\gamma}\sum_{k=1}^{n}(I_d - \gamma A)^{n-k} Z_k.$$

Denote

$$B_{n,k}(\gamma) = \Gamma_n(\gamma)^{-1/2}\sqrt{\gamma}(I_d - \gamma A)^{n-k}.$$

Then

$$y_{n,k} = B_{n,k}(\gamma)Z_k, \quad U_n = I_d, \quad \nu_n^2 = 1.$$

Let $M_n(\gamma)$ be the maximum norm of the matrix prefactor of $y_{n,k}$, i.e.

$$M_n(\gamma) = \max_{1\le k\le n} |B_{n,k}(\gamma)| = \max_{1\le k\le n} |\Gamma_n(\gamma)^{-1}\gamma(I_d - \gamma A)^{2n-2k}|.$$

We have shown that $|\Gamma_n(\gamma)^{-1}|$ is bounded by some constant. As a result, $M_n(\gamma) \lesssim \gamma \to 0$. This is sufficient for the Lindeberg condition and CLT. To this end, notice that

$$\sum_{k=1}^{n} \mathbb{E}\big(|y_{n,k}|^2 \mathbf{1}(|y_{n,k}|^2 \ge \varepsilon)\big) \le \sum_{k=1}^{n} |B_{n,k}(\gamma)|^2 \mathbb{E}\big(|Z_k|^2 \mathbf{1}(M_n(\gamma)|Z_k|^2 \ge \varepsilon)\big).$$

By the dominated convergence theorem, we have

$$\mathbb{E}\big(|Z_k|^2 \mathbf{1}(M_n(\gamma)|Z_k|^2 \ge \varepsilon)\big) \to 0$$

due to $\mathbf{1}(M_n(\gamma)|Z_k|^2 \ge \varepsilon) \to 0$. By elementary calculation, $\sum_{k=1}^{n} |B_{n,k}(\gamma)|^2$ is bounded. As a result, the Lindeberg condition is justified.

A.4 Proof of the Berry-Esseen Inequality

We begin with a refined analysis of the linear approximation.

$$\Delta_n = (\mathbf{I}_d - \gamma A)^n \Delta_0 - L_n + I_{1,n} + I_{2,n},$$

where $I_{1,n} = \gamma \sum_{k=1}^{n} (\mathbf{I}_d - \gamma A)^{n-k} R_k$, $I_{2,n} = \gamma \sum_{k=1}^{n} (\mathbf{I}_d - \gamma A)^{n-k} D_k$, $R_n = A\Delta_{n-1} - \nabla G(\theta_{n-1})$, and $D_n = \nabla G(\theta_{n-1}) - \nabla g(\theta_{n-1}, X_n) + \nabla g(\theta^*, X_n)$. By Lemma A.4, the matrix $\Gamma_n(\gamma)$ is invertible, and

$$\sqrt{\gamma}\Gamma_n(\gamma)^{-1/2} \lesssim \sqrt{\gamma}.$$

We further have the following decomposition:

$$\Gamma^{-1/2}\frac{\Delta_n}{\sqrt{\gamma}} = (\Gamma^{-1/2} - \Gamma_n(\gamma)^{-1/2})\frac{\Delta_n}{\sqrt{\gamma}} + \Gamma_n(\gamma)^{-1/2}\frac{(\mathbf{I}_d - \gamma A)^n \Delta_0 - L_n + I_{1,n} + I_{2,n}}{\sqrt{\gamma}}.$$

The Gaussian approximation error will be assessed via those terms. Define

$$\mathcal{I}_n = \sqrt{\gamma} \sum_{k=1}^{n} (1 - \gamma\lambda)^{n-k} |\Delta_{k-1}|^2,$$

as an auxiliary sequence. Define $\mathcal{I}_n^{(t)}$, $D_k^{(t)}$, and $I_{2,n}^{(t)}$ in the same way as $\theta_n^{(t)}$, and

$$\psi_D(n,t) = \|I_{2,n} - I_{2,n}^{(t)}\|_2$$

as the functional dependence measure of $I_{2,n}$. The next Lemma investigates the behavior of ψ_D.

Lemma A.5. *For $1 \le t \le n$,*

$$\psi_D(n,t)^2 \lesssim \gamma^2(1 - \gamma\lambda)^{2n-t-1} + \gamma^3(1 - \gamma\lambda)^{n-t-1}$$

Proof. Notice that

$$D_k - D_k^{(t)} = \begin{cases} 0 & k < t; \\ -\nabla g(\theta_{t-1}, X_t) + \nabla g(\theta^*, X_t) + -\nabla g(\theta_{t-1}, X_t') - \nabla g(\theta^*, X_t') & k = t; \\ \nabla G(\theta_{k-1}) - \nabla g(\theta_{k-1}, X_k) - [\nabla G(\theta_{k-1}^{(t)}) - \nabla g(\theta_{k-1}^{(t)}, X_k)] & k > t, \end{cases}$$

it is clear that that $D_k - D_k^{(t)}$ is also a martingale difference sequence. Due to stochastic Lipschitz continuity,

$$\begin{aligned}\psi_D(n,t)^2 &= \gamma^2\mathbb{E}|\sum_{k=1}^{n}(\mathbf{I}_d-\gamma A)^{n-k}(D_k-D_k^{(t)})|^2\\ &= \gamma^2\sum_{k=1}^{n}(\mathbf{I}_d-\gamma A)^{2n-2k}\mathbb{E}|D_k-D_k^{(t)}|^2\\ &\le \gamma^2(1-\gamma\lambda)^{2n-2t}\mathbb{E}|D_t-D_t^{(t)}|^2+4L_2^2\gamma^2\sum_{k=t+1}^{n}(1-\gamma\lambda)^{2n-2k}\mathbb{E}|\theta_{k-1}-\theta_{k-1}^{(t)}|^2\end{aligned}$$

By Theorem 2.3 and 3.1, we have

$$\mathbb{E}|D_t-D_t^{(t)}|^2 \le 4\|\nabla g(\theta_{t-1},X_t)-\nabla g(\theta^*,X_t)\|^2 \le 4L_2^2\|\Delta_{t-1}\|^2 \lesssim \gamma+(1-\alpha_0\gamma)^{t-1},$$

and

$$\mathbb{E}|\theta_{k-1}-\theta_{k-1}^{(t)}|^2 = \psi(k-1,t)^2 \lesssim \gamma^2(1-\alpha_0\gamma)^{k-1-t}.$$

By elementary calculation,

$$\begin{aligned}\psi_D(n,t)^2 &\lesssim \gamma^3(1-\gamma\lambda)^{2n-2t}+\gamma^2(1-\gamma\lambda)^{2n-t-1}+\gamma^3(1-\gamma\lambda)^{n-t-1}\\ &\asymp \gamma^2(1-\gamma\lambda)^{2n-t-1}+\gamma^3(1-\gamma\lambda)^{n-t-1}.\end{aligned}$$

□

Now we are ready to prove the Berry-Esseen inequality.

Proof of Theorem 3.5. We apply Theorem 2.1 in Shao and Zhang [2022]. Since we use $L_n/\sqrt{\gamma} = \sqrt{\gamma}\sum_{k=1}^{n}(\mathbf{I}_d-\gamma A)^{n-k}Z_k$ to approximate $\Delta_n/\sqrt{\gamma}$, with

$$\|L_n\|_3^3 \le \sum_{k=1}^{n}\|\sqrt{\gamma}(\mathbf{I}_d-\gamma A)^{n-k}Z_k\|_3^3 \lesssim \sum_{k=1}^{n}\gamma^{3/2}(1-\lambda\gamma)^{n-k} \lesssim \sqrt{\gamma},$$

Theorem 2.1 in Shao and Zhang [2022] states that

$$\begin{aligned}&\sup_{\mathcal{D}\in\mathcal{V}}|\mathbb{P}(\Gamma_n(\gamma)^{-1/2}\frac{\Delta_n}{\sqrt{\gamma}}\in\mathcal{D})-\mathbb{P}(\mathcal{N}(0,\mathbf{I}_d)\in\mathcal{D})|\\ \le& C\Big(\sqrt{\gamma}+\mathbb{E}\{|\Gamma_n(\gamma)^{-1/2}L_n/\sqrt{\gamma}|\mathcal{J}\}+\sum_{t=1}^{n}\mathbb{E}\{|\mathcal{J}-\mathcal{J}^{(t)}||\Gamma_n(\gamma)^{-1/2}\sqrt{\gamma}(\mathbf{I}_d-\gamma A)^{n-t}Z_t|\}\Big),\end{aligned} \tag{27}$$

where $\mathcal{J}$ and $\mathcal{J}^{(t)}$ are some quantities such that

$$\mathcal{J} \ge |\Gamma_n(\gamma)^{-1/2}\frac{\Delta_n-L_n}{\sqrt{\gamma}}| \tag{28}$$

and $\mathcal{J}^{(t)}$ is independent of X_t or Z_t. We define

$$\mathcal{J} = |\Gamma_n(\gamma)^{-1/2}\sqrt{\gamma}(\mathbf{I}_d-\gamma A)^n\Delta_0|+|\gamma^{-1/2}\Gamma_n(\gamma)^{-1/2}I_{2,n}|+\tilde{C}\mathcal{I}_n,$$

with some universal constant $\tilde{C}$ such that $\tilde{C}\mathcal{I}_n \ge |\gamma^{-1/2}\Gamma_n(\gamma)^{-1/2}I_{1,n}|$. Such $\tilde{C}$ exists because of the construction of $\mathcal{I}_n$ and Taylor's expansion. Then (28) holds by the triangle inequality. We further define

$$\mathcal{J}^{(t)} = |\Gamma_n(\gamma)^{-1/2}\sqrt{\gamma}(\mathbf{I}_d-\gamma A)^n\Delta_0|+|\gamma^{-1/2}\Gamma_n(\gamma)^{-1/2}I_{2,n}^{(t)}|+\tilde{C}\mathcal{I}_n^{(t)}$$

such that $\mathcal{J}^{(t)}$ is independent of X_t and Z_t, and decompose the difference between $\mathcal{J}$ and its copula perturbation $\mathcal{J}^{(t)}$ as

$$\begin{aligned}|\mathcal{J}-\mathcal{J}^{(t)}| &\le \tilde{C}|\mathcal{I}_n-\mathcal{I}_n^{(t)}|+\Big||\gamma^{-1/2}\Gamma_n(\gamma)^{-1/2}I_{2,n}|-|\gamma^{-1/2}\Gamma_n(\gamma)^{-1/2}I_{2,n}^{(t)}|\Big|\\ &\le \tilde{C}|\mathcal{I}_n-\mathcal{I}_n^{(t)}|+|\gamma^{-1/2}\Gamma_n(\gamma)^{-1/2}(I_{2,n}-I_{2,n}^{(t)})|.\end{aligned} \tag{29}$$

We will control each error term decomposed above. We first investigate the following quantity:

$$\begin{aligned}&\mathbb{E}|\mathcal{I}_n - \mathcal{I}_n^{(t)}||\Gamma_n(\gamma)^{-1/2}\sqrt{\gamma}(\mathbf{I}_d - \gamma A)^{n-t}Z_t| \\ &\lesssim \gamma(1-\gamma\lambda)^{n-t}\mathbb{E}\{|Z_t|\sum_{k=1}^{n}|(1-\gamma\lambda)^{n-k}(|\Delta_{k-1}|^2 - |\Delta_{k-1}^{(t)}|^2)|\} := \mathcal{T}_{n,t},\end{aligned}$$

which can be further controlled by

$$\begin{aligned}\mathcal{T}_{n,t} &\le \gamma(1-\gamma\lambda)^{n-t}\mathbb{E}\{|Z_t|\sum_{k=1}^{n}|(1-\gamma\lambda)^{n-k}|(\Delta_{k-1} - \Delta_{k-1}^{(t)})(\Delta_{k-1} + \Delta_{k-1}^{(t)})|\} \\ &\le \gamma(1-\gamma\lambda)^{n-t}\sum_{k=1}^{n}(1-\gamma\lambda)^{n-k}\mathbb{E}\Big\{|Z_t||\theta_{k-1} - \theta_{k-1}^{(t)}|\big(|\theta_{k-1} - \theta^*| + |\theta_{k-1}^{(t)} - \theta^*|\big)\Big\}\end{aligned}$$

We apply the Hölder inequality, for $k \ge t+1$,

$$\begin{aligned}&\mathbb{E}\Big\{|Z_t||\theta_{k-1} - \theta_{k-1}^{(t)}|\big(|\theta_{k-1} - \theta^*| + |\theta_{k-1}^{(t)} - \theta^*|\big)\Big\} \\ &\le 2\|Z_t\|_4\|\Delta_{k-1}\|_4\psi(k-1,t) \\ &\lesssim (\sqrt{\gamma} + (1-\alpha_0\gamma)^{(k-1)/2})\gamma(1-\alpha_0\gamma)^{(k-1-t)/2}\end{aligned}$$

where the last inequality if from Theorem 2.3 and 3.1. For $k \le t$, the expectation above is 0. Hence we have

$$\begin{aligned}\mathcal{T}_{n,t} &\lesssim \gamma(1-\gamma\lambda)^{n-t}\sum_{k=t+1}^{n}(1-\gamma\lambda)^{n-k}(\sqrt{\gamma} + (1-\alpha_0\gamma)^{(k-1)/2})\gamma(1-\alpha_0\gamma)^{(k-1-t)/2}, \\ &= \gamma^2(1-\lambda\gamma)^{n-t}\sum_{k=t+1}^{n}\big[(1-\lambda\gamma)^{n-t/2-1} + \sqrt{\gamma}(1-\gamma\lambda)^{n-(k+t+1)/2}\big] \\ &\le \gamma^2(n-t)(1-\lambda\gamma)^{2n-3t/2-1} + \gamma^{5/2}\frac{(1-\lambda\gamma)^{(n-t-1)/2}}{1-\sqrt{1-\gamma\lambda}} \\ &\lesssim \gamma^2(n-t)(1-\lambda\gamma)^{2n-3t/2-1} + \gamma^{3/2}(1-\lambda\gamma)^{(n-t-1)/2}.\end{aligned}$$

Here we use the fact that $\sqrt{1-\lambda\gamma} \le 1 - \lambda\gamma/2$. By elementary calculations,

$$\sum_{t=1}^{n}\mathbb{E}\big\{|\mathcal{I}_n - \mathcal{I}_n^{(t)}||\Gamma_n(\gamma)^{-1/2}\sqrt{\gamma}(\mathbf{I}_d - \gamma A)^{n-t}Z_t|\big\} \lesssim \sum_{t=1}^{n}\mathcal{T}_{n,t} \lesssim (1-\lambda\gamma)^{(n+1)/2} + \sqrt{\gamma}. \quad (30)$$

Similarly, by Cauchy inequality,

$$\begin{aligned}&\mathbb{E}|\gamma^{-1/2}\Gamma_n(\gamma)^{-1/2}(I_{2,n} - I_{2,n}^{(t)})||\Gamma_n(\gamma)^{-1/2}\sqrt{\gamma}(\mathbf{I}_d - \gamma A)^{n-t}Z_t| \\ &\lesssim (1-\gamma\lambda)^{n-t}\mathbb{E}\{|Z_t||I_{2,n} - I_{2,n}^{(t)}|\} \\ &\le (1-\gamma\lambda)^{n-t}\|Z_t\|_2\psi_D(n,t) \\ &\lesssim \gamma(1-\lambda\gamma)^{(4n-3t-1)/2} + \gamma^{3/2}(1-\lambda\gamma)^{(3n-3t-1)/2}.\end{aligned}$$

Summing the quantity above from $t=1$ to $t=n$ yields an upper bound

$$\sum_{t=1}^{n}\mathbb{E}|\gamma^{-1/2}\Gamma_n(\gamma)^{-1/2}(I_{2,n} - I_{2,n}^{(t)})||\Gamma_n(\gamma)^{-1/2}\sqrt{\gamma}(\mathbf{I}_d - \gamma A)^{n-t}Z_t| \lesssim (1-\lambda\gamma)^{(n-1)/2} + \sqrt{\gamma}. \quad (31)$$

Then we consider the following error terms due to linear approximation:

$$\mathbb{E}\{|\Gamma_n(\gamma)^{-1/2}\sqrt{\gamma}(\mathbf{I}_d - \gamma A)^n\Delta_0||\Gamma_n(\gamma)^{-1/2}L_n/\sqrt{\gamma}|\},$$

$$\mathbb{E}\{|\Gamma_n(\gamma)^{-1/2}\sqrt{\gamma}I_{2,n}||\Gamma_n(\gamma)^{-1/2}L_n/\sqrt{\gamma}|\},$$

and

$$\mathbb{E}\{|\mathcal{I}_n||\Gamma_n(\gamma)^{-1/2}L_n/\sqrt{\gamma}|\}.$$

We use the Cauchy inequality to bound them. Notice that by definition of $\Gamma_n(\gamma)$, the sequence $\Gamma_n(\gamma)^{-1/2}L_n/\sqrt{\gamma}$ is standardized with fixed covariance matrix $\mathbf{I}_d$. For the first term,

$$\begin{aligned}&\mathbb{E}\{|\Gamma_n(\gamma)^{-1/2}\sqrt{\gamma}(\mathbf{I}_d-\gamma A)^n\Delta_0||\Gamma_n(\gamma)^{-1/2}L_n/\sqrt{\gamma}|\}\\ \lesssim\ &\sqrt{\gamma}(1-\lambda\gamma)^n\|\Delta_0\|_2\|\Gamma_n(\gamma)^{-1/2}L_n/\sqrt{\gamma}\|_2\lesssim\sqrt{\gamma}(1-\lambda\gamma)^n.\end{aligned}\tag{32}$$

For the second term,

$$\begin{aligned}&\mathbb{E}\{|\Gamma_n(\gamma)^{-1/2}\sqrt{\gamma}I_{2,n}||\Gamma_n(\gamma)^{-1/2}L_n/\sqrt{\gamma}|\}\\ \lesssim\ &\sqrt{\gamma}\|I_{2,n}\|_2\|\Gamma_n(\gamma)^{-1/2}L_n/\sqrt{\gamma}\|_2\lesssim\gamma+\sqrt{\gamma}(1-\lambda\gamma)^{(n-1)/2}.\end{aligned}\tag{33}$$

For the third term,

$$\begin{aligned}&\mathbb{E}\{|\mathcal{I}_n||\Gamma_n(\gamma)^{-1/2}L_n/\sqrt{\gamma}|\}\\ \lesssim\ &\sqrt{\gamma}\sum_{k=1}^n(1-\gamma\lambda)^{n-k}\mathbb{E}\{|\Delta_{k-1}|^2|\Gamma_n(\gamma)^{-1/2}L_n/\sqrt{\gamma}|\}\\ \le\ &\sqrt{\gamma}\sum_{k=1}^n(1-\gamma\lambda)^{n-k}\|\Delta_{k-1}\|_4^2\\ \lesssim\ &\sqrt{\gamma}\sum_{k=1}^n(1-\gamma\lambda)^{n-k}[(1-\alpha_0\gamma)^{k-1}+\gamma]\\ \lesssim\ &\sqrt{\gamma}+\sqrt{\gamma}n(1-\gamma\lambda)^{n-1}.\end{aligned}\tag{34}$$

Combining all upper bounds of (29)-(34) and plugging them into the inequality (27) yields

$$\begin{aligned}&\sup_{\mathcal{D}\in\mathcal{V}}|\mathbb{P}(\Gamma_n(\gamma)^{-1/2}\frac{\Delta_n}{\sqrt{\gamma}}\in\mathcal{D})-\mathbb{P}(\mathcal{N}(0,\mathbf{I}_d)\in\mathcal{D})|\\ &\le C(\sqrt{\gamma}+(1-\lambda\gamma)^{\frac{n-1}{2}}+\sqrt{\gamma}n(1-\lambda\gamma)^{n-1})\end{aligned}$$

for some constant C independent of n, γ and θ_0. By the discussion of Remark 1 in Samsonov et al. [2024], since $\Gamma_n(\gamma)^{-1/2}$ is non-degenerate, and an image of a convex set under a non-degenerate linear mapping is a convex set, we have

$$\sup_{\mathcal{D}\in\mathcal{V}}|\mathbb{P}(\Gamma_n(\gamma)^{-1/2}\frac{\Delta_n}{\sqrt{\gamma}}\in\mathcal{D})-\mathbb{P}(\mathcal{N}(0,\mathbf{I}_d)\in\mathcal{D})|=\sup_{\mathcal{D}\in\mathcal{V}}|\mathbb{P}(\frac{\Delta_n}{\sqrt{\gamma}}\in\mathcal{D})-\mathbb{P}(\mathcal{N}(0,\Gamma_n(\gamma))\in\mathcal{D})|.$$

To complete the proof, we use Theorem 1.1 in Devroye et al. [2018] (or Lemma 13 in Samsonov et al. [2024]), which bounds the total variation distance of two Gaussian measures by the distance between their covariance matrix. Here we only need to bound the convex distance, i.e.,

$$\sup_{\mathcal{D}\in\mathcal{V}}|\mathbb{P}(\mathcal{N}(0,\Gamma_n(\gamma))\in\mathcal{D})-\mathbb{P}(\mathcal{N}(0,\Gamma)\in\mathcal{D})|\le\frac{3}{2}\|\Gamma^{-1/2}\Gamma_n(\gamma)\Gamma^{-1/2}-\mathbf{I}_d\|_F,$$

which trivially holds by Theorem 1.1 in Devroye et al. [2018]. Since matrix norms are equivalent, by (25) we have $|\|\Gamma^{-1/2}\Gamma_n(\gamma)\Gamma^{-1/2}-\mathbf{I}_d\|_F\lesssim\gamma+(1-\gamma\lambda)^{2n}$. The last step is to use the triangular inequality of the convex distance, and the proof is completed. □

A.5 Proof of Nagaev-type Inequality

We first prove the results of non-linear auto-regressive approximation.

Proof of Lemma 4.1. Similar to the argument of the proof of Theorem 2.3,

$$\begin{aligned}\|\beta_n-\theta^*\|_p^2&\le\|\beta_{n-1}-\theta^*-\gamma\nabla G(\beta_{n-1})\|_p^2+(p-1)\gamma^2\|Z_n\|_p^2\\ &\le(1-2\gamma\mu+(6p-5)\gamma^2L_p^2)\|\beta_{n-1}-\theta^*\|_p^2+3(p-1)\gamma^2\|Z_n\|_p^2.\end{aligned}\tag{35}$$

Recursively updating it, we get

$$\|\beta_n - \theta^*\|_p^2 \le \frac{3(p-1)M_p^2}{\alpha_0}\gamma.$$

For the second result,

$$\begin{aligned}
&\|\theta_n - \beta_n\|_p^2 \\
&\le \|\theta_{n-1} - \beta_{n-1} - \gamma[\nabla G(\theta_{n-1}) - \nabla G(\beta_{n-1})]\|_p^2 && (36)\\
&+ (p-1)\gamma^2\|\nabla G(\theta_{n-1}) - \nabla g(\theta_{n-1}, X_n) + Z_n\|_p^2 \\
&\le \Big\|\big\langle \theta_{n-1} - \beta_{n-1}, \theta_{n-1} - \beta_{n-1}\big\rangle - 2\gamma\big\langle \theta_{n-1} - \beta_{n-1}, \nabla G(\theta_{n-1}) - \nabla G(\beta_{n-1})\big\rangle\Big\|_{p/2} \\
&\quad + \gamma^2\big\|\nabla G(\theta_{n-1}) - \nabla G(\beta_{n-1})\big\|_p^2 + 4(p-1)\gamma^2 L_p^2\|\theta_{n-1} - \theta^*\|_p^2 && (37)\\
&\le (1 - 2\gamma\mu + \gamma^2 L_p^2)\|\theta_{n-1} - \beta_{n-1}\|_p^2 + 4(p-1)\gamma^2 L_p^2[(1-\alpha_0\gamma)^n\|\theta_0 - \theta^*\|_p^2 \\
&\quad + 3(p-1)M_p^2\alpha_0^{-1}\gamma^3] && (38)\\
&\le (1 - \alpha_1\gamma)\|\theta_{n-1} - \beta_{n-1}\|_p^2 + 4(p-1)\gamma^2 L_p^2[(1-\alpha_0\gamma)^n\|\theta_0 - \theta^*\|_p^2 \\
&\quad + 3(p-1)M_p^2\alpha_0^{-1}\gamma^3], && (39)\\
& && (40)
\end{aligned}$$

where $\alpha_1 \le 2\mu - \gamma L_p^2$ and $\alpha_1 \le \gamma^{-1}$. Without loss of generality, we can choose $\alpha_0 \le \alpha_1$ since the upper constraint of α_0 is more stringent. Recursively updating the inequality, we get

$$\|\theta_n - \beta_n\|_p^2 \le [(1-\alpha_1\gamma)^n + 4(p-1)L_p^2 n\gamma^2(1-\alpha_0\gamma)^n]\|\Delta_0\|_p^2 + \frac{3(p-1)M_p^2}{\alpha_0\alpha_1}\gamma^2.$$

□

Now we are ready to prove the sharp concentration inequality.

Proof of Theorem 4.2. Denote $\mathbb{E}_0 X = X - \mathbb{E}X$ as the centralized random variable X. Without loss of generality, we let $\alpha_0 \le \alpha_1$ in the following proof. Applying Markov inequality on Lemma 4.1, we have

$$\begin{aligned}
&\mathbb{P}(|\Delta_n - (\beta_n - \theta^*)| > \epsilon) && (41)\\
&= \mathbb{P}(|\theta_n - \beta_n| > \epsilon) && (42)\\
&\le \frac{\|\theta_n - \beta_n\|_p^p}{\epsilon^p} && (43)\\
&\le \frac{2^{p-1}\rho_p^p(1-\alpha_0\gamma)^{np/2}}{\epsilon^p}[1 + 4(p-1)L_p^2 n\gamma^2]^{p/2} + \frac{2^{p-1}\gamma^p(3p-3)^{p/2}M_p^p}{\epsilon^p(\alpha_0\alpha_1)^{p/2}}. && (44)
\end{aligned}$$

Then we consider the tail probability of $\mathbb{E}_0\{v^\top(\beta_n - \theta^*)\}$. Define $\mathcal{P}_k(\xi) = \mathbb{E}(\xi|\mathcal{F}_k) - \mathbb{E}(\xi|\mathcal{F}_{k-1})$ as the projection operator. Let $Z_i' = \nabla g(\theta^*, X_i')$ be the i.i.d. copy of Z_i. Similar to the proof of Theorem 3.1, we can show that for $1 \le k \le n$,

$$|\mathcal{P}_k(v^\top(\beta_n - \theta^*))| \le (1-\alpha_0\gamma)^{n-k}\gamma\mathbb{E}(|Z_k - Z_k'||\mathcal{F}_k). \tag{45}$$

Let $y = p\epsilon/(p+2)$. Define the following sequence

$$\eta_i = \eta_{i-1} - \gamma\nabla G(\eta_{i-1}) - \gamma Z_i \times \min\Big\{1, \frac{y}{2\gamma(1-\alpha_0\gamma)^{n-i}|Z_i|}\Big\},\ \eta_0 = \theta^*. \tag{46}$$

Then we have

$$\begin{aligned}
&\mathbb{P}(|\mathbb{E}_0\{v^\top(\beta_n - \theta^*)\}| > \epsilon) && (47)\\
&\le \sum_{i=1}^n \mathbb{P}\Big(|Z_i| > \frac{y}{2\gamma(1-\alpha_0\gamma)^{n-i}}\Big) + \mathbb{P}(|\mathbb{E}_0\{v^\top(\eta_n - \theta^*)\}| > \epsilon), && (48)\\
&\le \frac{2^p M_p^p}{y^p}\gamma^p\sum_{i=1}^n(1-\alpha_0\gamma)^{p(n-i)} + \mathbb{P}(|\mathbb{E}_0\{v^\top(\eta_n - \theta^*)\}| > \epsilon) && (49)\\
&\le \frac{C\gamma^{p-1}}{\epsilon^p} + \mathbb{P}(|\mathbb{E}_0\{v^\top(\eta_n - \theta^*)\}| > \epsilon), && (50)
\end{aligned}$$

for some constant C independent of ϵ, n and γ. The next step is to investigate the tail behavior of $\mathbb{E}_0\{v^\top(\eta_n - \theta^*)\}$. We consider its moment generating function: for $x > 0$,

$$\mathcal{M}_n(x) := \mathbb{E}\exp(x\mathbb{E}_0\{v^\top(\eta_n - \theta^*)\}) = \mathbb{E}\exp\left\{x\sum_{k=1}^{n}\mathcal{P}_k(v^\top(\eta_n - \theta^*)\right\}.$$

Due to definition (46) and a similar argument to (45), we have $\sup_{v\in\mathbb{S}^d}|\mathcal{P}_k(v^\top(\eta_n - \theta^*)| \le y$. As a result, we can leverage Lemma 1.4 in [Nagaev, 1979] to obtain

$$\begin{aligned}
&\mathbb{E}\left\{\exp\left(x\mathcal{P}_k(v^\top(\eta_n - \theta^*))\right)\middle|\mathcal{F}_{k-1}\right\} \\
&\le 1 + \frac{\exp(p)x^2}{2}\mathbb{E}\left\{|\mathcal{P}_k(v^\top(\eta_n - \theta^*))|^2\middle|\mathcal{F}_{k-1}\right\} \\
&\quad + \frac{\exp(xy) - 1 - xy}{y^p}\mathbb{E}\left\{|\mathcal{P}_k(v^\top(\eta_n - \theta^*))|^p\middle|\mathcal{F}_{k-1}\right\} \times \mathbb{I}\left\{x > \frac{p}{y}\right\} \\
&\le 1 + \exp(p)x^2(1-\alpha_0\gamma)^{2(n-k)}\gamma^2 M_2^2 + \frac{\exp(xy) - 1 - xy}{y^p}\frac{(1-\alpha_0\gamma)^{p(n-k)}\gamma^p M_p^p}{2^p} \times \mathbb{I}\left\{x > \frac{p}{y}\right\}.
\end{aligned}$$

Since the final upper bound does not depend on $\mathcal{F}_{k-1}$, we have

$$\begin{aligned}
\mathcal{M}_n(x) &\le \prod_{t=0}^{n-1}\left(1 + e^p x^2(1-\alpha_0\gamma)^{2t}\gamma^2 M_2^2 + \frac{e^{xy} - 1 - xy}{y^p}2^p(1-\alpha_0\gamma)^{pt}\gamma^p M_p^p \times \mathbb{I}\{x > p/y\}\right) \\
&\le \exp\left(C_p x^2 M_2^2\gamma + C_p\gamma^{(p-1)}\frac{e^{xy} - 1 - xy}{y^p}M_p^p \times \mathbb{I}\{x > p/y\}\right),
\end{aligned}$$

for some constant C_p independent of ϵ, n and γ. We use the Chernoff-type bound:

$$\mathbb{P}(|\mathbb{E}_0\{v^\top(\eta_n - \theta^*)\}| > \epsilon) \le e^{-x\epsilon}\mathcal{M}_n(x),$$

and find an $x > 0$ such that

$$C_p x^2 M_2^2\gamma - \frac{2y}{q} + C_p\gamma^{(p-1)}\frac{e^{xy} - 1 - xy}{y^p}M_p^p \times \mathbb{I}\{x > p/y\} - xy$$

is small. The calculation is identical to the proof of Theorem 1.3 in Nagaev [1979], which leads to

$$\mathbb{P}(|\mathbb{E}_0\{v^\top(\eta_n - \theta^*)\}| > \epsilon) \le \frac{2}{C_p\epsilon^p M_p^p\gamma^{1-p} + 1} + 2\exp(-\frac{C_p\epsilon^2}{\gamma M_2^2}) \le \frac{C_1\gamma^{p-1}}{\epsilon^p} + 2\exp(-\frac{C_2\epsilon^2}{\gamma}) \tag{51}$$

for some constant C_1 and C_2 independent of ϵ, n and γ.

Finally, we use union bound on (44), (50), and (51) to finish the proof of the tail probability inequality in Theorem 4.2. Notice that from Lemma 4.1 we have

$$|\mathbb{E}(\beta_n - \theta^*)| \le \|\beta_n - \theta^*\|_p \le M_p\sqrt{\frac{3(p-1)}{\alpha_0}}\sqrt{\gamma}.$$

As long as $M_p\sqrt{\frac{3(p-1)}{\alpha_0}}\sqrt{\gamma} < \epsilon$, this expectation term can be ignored in the tail probability bound. If not, i.e., $\epsilon^2/\gamma \le 3M_p^2(p-1)/\alpha_0$, we can choose $C_2 \le \log 2\alpha_0/3M_p^2(p-1)$ such that the probability bound trivially holds in this case. As a result, the expectation term does not affect the validity of our result, and the proof is completed. □

A.6 Additional Experiment Details

We conducted the experiments in R version 4.3.1 (2023-06-16) on a MacBook Air with a GPU Apple M1, 4 performance and 4 efficiency cores, and 8 GB LPDDR4 memory, equipped with macOS Big Sur version 11.5.1.

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 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 and constraints in the remarks on the results.

   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: [Yes]

   Justification: We disclose all the information in the paper and supplementary material.

   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: [Yes]

Justification: We provide them in our supplementary materials.

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, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We specify all details in the final section of the paper.

   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: [Yes]

   Justification: They are reported in the supplementary material.

   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: [Yes]

   Justification: We provide sufficient information on the computer resources in the supplementary material.

   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 statistical learning, which can inform the development of uncertainty quantification and trustworthy AI. As the work mostly focuses on theoretical aspects 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: The paper has 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: We do not use such 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: We do not introduce 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]

    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.
    - 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.