# Berry–Esseen Bounds for Multivariate Nonlinear Statistics with Applications to M-estimators and Stochastic Gradient Descent Algorithms

**Qi-Man Shao**[*1,2] **and Zhuo-Song Zhang**[†2,3]

[1] *Department of Statistics and Data Scinece, Southern University of Science and Technology, Shenzhen, Guangdong, P.R. China. e-mail:* shaoqm@sustech.edu.cn

[2] *Department of Statistics, The Chinese University of Hong Kong, Shatin, N.T. Hong Kong*

[3] *Department of Statistics and Applied Probability, National University of Singapore, Singapore 117546. e-mail:* zszhang.stat@gmail.com

**Abstract:** We establish a Berry–Esseen bound for general multivariate nonlinear statistics by developing a new multivariate-type randomized concentration inequality. The bound is the best possible for many known statistics. As applications, Berry–Esseen bounds for M-estimators and averaged stochastic gradient descent algorithms are obtained.

**MSC2020 subject classifications:** Primary 60F05, 62E20; secondary 62F12.
**Keywords and phrases:** Berry–Esseen bound, Multivariate normal approximation, Randomized concentration inequality, Stein's method, M-estimators, Averaged stochastic gradient descent algorithms

.

## 1. Introduction

Let $X_1, \dots, X_n$ be independent random variables taking values on $\mathcal{X}$ and let $T \coloneqq T(X_1, \dots, X_n)$ be a general $d$-dimensional nonlinear statistic. In many cases the nonlinear statistic can be written as a linear statistic plus an error term:

$$T = W + D, \tag{1.1}$$

where

$$W = \sum_{i=1}^n \xi_i, \quad D \coloneqq D(X_1, \dots, X_n) = T - W, \tag{1.2}$$

$\xi_i \coloneqq h_i(X_i) \in \mathbb{R}^d$ and $h_i \colon \mathcal{X} \mapsto \mathbb{R}^d$ is a Borel measurable function. Assume that

$$\mathbb{E}\xi_i = 0 \text{ for each } 1 \leqslant i \leqslant n \text{ and } \sum_{i=1}^n \mathbb{E}\{\xi_i \xi_i^{\intercal}\} = I_d. \tag{1.3}$$

*Research partially supported by NSFC12031005 and Shenzhen Outstanding Talents Training Fund and also by Hong Kong RGC GRF 14302515 and 14304917.

†Corresponding author. Research partially supported by Singapore Ministry of Education Academic Research Fund MOE 2018-T2-076.

Let

$$\gamma := \gamma_n = \sum_{i=1}^{n} \mathbb{E}\|\xi_i\|^3. \tag{1.4}$$

Since $\xi_i$ is standardized, we remark that $h_i = h_{n,i}$ and $\xi_i = \xi_{n,i}$. If $\|D\| \xrightarrow{p} 0$ and $\gamma \to 0$ as $n \to \infty$, then, clearly, $T$ converges in distribution to a $d$-dimensional standard normal distribution $N(0, I_d)$.

The aim of this paper is to provide a Berry–Esseen bound of the multivariate normal approximation for the nonlinear statistic $T$. The Berry–Esseen bound for multivariate normal approximation has been well studied in the past decades. For the linear statistic $W$, Bentkus [4, 5] used induction and Taylor's expansion to prove a Berry–Esseen bound of order $d^{1/4}\gamma$, which is the best known result for the dependence on the dimension $d$. We refer to Nagaev [16], Senatov [28], Götze [14], Bhattacharya and Holmes [7] and Raič [25] for other results for independent random vectors.

In the case where $d = 1$, Chen and Shao [9] proved a Berry–Esseen bound for $T$ using the Berry–Esseen bound for $W$ and a randomized-type concentration inequality approach:

$$\sup_{z\in\mathbb{R}} |\,\mathbb{P}(T \leqslant z) - \Phi(z)| \leqslant 6.1\gamma + \mathbb{E}|WD| + \sum_{i=1}^{n} \mathbb{E}|\xi_i(D - D^{(i)})|, \tag{1.5}$$

where $D^{(i)}$ is any random variable such that $\xi_i$ is independent of $D^{(i)}$ and $\Phi$ is the standard normal distribution function. For the Berry–Esseen bound for multivariate normal approximation, Chen and Fang [10] proved a concentration inequality for $d$-dimensional exchangeable pairs. We also refer to Barbour [3], Götze [14], Goldstein and Rinott [13], Chatterjee and Meckes [8], Reinert and Röllin [26], Bhattacharya and Holmes [7], Chen et al. [11], Chen and Fang [10] and Raič [25] for the development of Stein's method for multivariate normal approximations.

The main purpose of this paper is to prove a Berry–Esseen bound for nonlinear multivariate statistics by developing a new randomized multivariate concentration inequality which generalizes the results of Chen and Shao [9] and Chen and Fang [10]. Our main result can be applied to a large class of non-linear statistics, including M-estimators and averaged stochastic gradient descent estimators.

Throughout this paper, we use the following notations. Let $d \geqslant 1$ and $x = (x_1, \dots, x_d)$ be a vector in $\mathbb{R}^d$. For $x, y \in \mathbb{R}^d$, denote by $\langle x, y\rangle$ the inner product of $x$ and $y$. Let $\|x\| = \sqrt{\langle x, x\rangle}$ be the $l_2$-norm of $x$. For a $d \times d$ matrix $A$, and let $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ be the minimal and maximal eigenvalue of $A$, respectively. Denote by $A^\intercal$ the transpose of $A$ and by $\|A\|$ the spectral norm, i.e., $\|A\| := (\lambda_{\max}(A^\intercal A))^{1/2}$. Let $I_d$ be the $d$-dimensional identity matrix. For $X \in \mathbb{R}$ (resp. $\mathbb{R}^d$) and $p \geqslant 1$, let $\|X\|_p = (\mathbb{E}\{|X|^p\})^{1/p}$ (resp. $(\mathbb{E}\{\|X\|^p\})^{1/p}$) be the $L_p$-norm of $X$.

The rest of this paper is organized as follows. In Section 2, we present the Berry–Esseen bound of the multivariate normal approximation for $T$. In Section 3, we apply our main result to M-estimators and averaged stochastic gradient descent algorithms. In Section 4, we present a randomized concentration inequality for multivariate linear statistics and give the proof of the main result. The proofs of theorems in Section 3 are postponed to Section 5.

## 2. Main results

Let $(X_1, \dots, X_n), (\xi_1, \dots, \xi_n), W, T$ and $D$ be defined as in (1.1) and (1.2). Let $\mathcal{A}$ be the collection of all convex sets in $\mathbb{R}^d$. Let $Z \sim N(0, I_d)$. The following theorem provides a Berry–Esseen bound for $T$.

**Theorem 2.1.** *Assume that* (1.3) *is satisfied. Then,*

$$\sup_{A\in\mathcal{A}}\bigl|\mathbb{P}(T \in A) - \mathbb{P}(Z \in A)\bigr| \leqslant 259 d^{1/2}\gamma + 2\,\mathbb{E}\bigl\{\|W\|\Delta\bigr\} + 2\sum_{i=1}^{n}\mathbb{E}\bigl\{\|\xi_i\||\Delta - \Delta^{(i)}|\bigr\}, \tag{2.1}$$

*for any random variables $\Delta$ and $(\Delta^{(i)})_{1\leqslant i\leqslant n}$ such that $\Delta \geqslant \|D\|$ and $\Delta^{(i)}$ is independent of $X_i$, where $\gamma$ is as defined in* (1.4).

**Remark 2.1.** The choices of $\Delta$ and $\Delta^{(i)}$ are flexible. For example, let $(X_1', \dots, X_n')$ be an independent copy of $(X_1, \dots, X_n)$, one may choose $\Delta = \|D\|$ and $\Delta^{(i)} = \|D^{(i)}\|$, where $D^{(i)} = D(X_1, \dots, X_{i-1}, X_i', X_{i+1}, \dots, X_n)$. One can also choose $D^{(i)} = D(X_1, \dots, X_{i-1}, 0, X_{i+1}, \dots, X_n)$. Moreover, the last term in (2.1) cannot be removed, and we refer to Chen and Shao [9, Section 4] for a counterexample.

**Remark 2.2.** For $d = 1$, the right hand side of (2.1) reduces to

$$259\gamma + 2\,\mathbb{E}\{|W|\Delta\} + 2\sum_{i=1}^{n}\mathbb{E}\bigl|\xi_i(\Delta - \Delta^{(i)})\bigr|,$$

which differs from (1.5) up to a constant factor.

The Berry–Esseen bound (2.1) provides an optimal order in terms of $n$ for many applications. However, the order in $d$ may not be optimal in (2.1). For a linear statistic $W$, Bentkus [5] proved that

$$\sup_{A\in\mathcal{A}}\bigl|\mathbb{P}(W \in A) - \mathbb{P}(Z \in A)\bigr| \leqslant C d^{1/4}\gamma,$$

where $C > 0$ is an absolute constant and $d^{1/4}$ is believed to be the best possible. Here, $C > 0$ is an absolute constant, and Raič [25] recently obtained a bound with an explicit constant $42d^{1/4} + 16$ by using Stein's method. However, it is not clear how to obtain the order $d^{1/4}$ in our result.

Using the technique of truncation, we obtain the following corollary, which may be useful for applications.

**Corollary 2.2.** *Let $O$ be a measurable set and $\Delta$ be a random variable such that $\Delta \geqslant \|D\|\mathbb{1}(O)$. Under the conditions of Theorem 2.1, we have*

$$\sup_{A\in\mathcal{A}}\bigl|\mathbb{P}(T\in A)-\mathbb{P}(Z\in A)\bigr| \leqslant 259d^{1/2}\gamma + 2\,\mathbb{E}\bigl\{\|W\|\Delta\bigr\} + 2\sum_{i=1}^n \mathbb{E}\bigl\{\|\xi_i\||\Delta-\Delta^{(i)}|\bigr\} + \mathbb{P}(O^c),$$

*where $\Delta^{(i)}$ is any measurable random variable that is independent of $X_i$.*

Condition (1.3) can be extended to a general case. We have the following corollary.

**Corollary 2.3.** *Let $T, W, D$ and $(\xi_1,\dots,\xi_n)$ be defined as in (1.1) and (1.2). Assume that $(\xi_1,\dots,\xi_n)$ satisfies:*

$$\mathbb{E}\{\xi_i\} = 0 \text{ for } 1\leqslant i\leqslant n \text{ and } \sum_{i=1}^n \mathbb{E}\{\xi_i\xi_i^{\intercal}\} = \Sigma,$$

*where $\Sigma$ is a positive definite matrix with $\lambda_{\min}(\Sigma)\geqslant \sigma > 0$. Then*

$$\begin{aligned}\sup_{A\in\mathcal{A}}\bigl|\mathbb{P}(T\in A)-\mathbb{P}(\Sigma^{1/2}Z\in A)\bigr| &\leqslant 259\sigma^{-3/2}d^{1/2}\gamma + 2\sigma^{-1}\,\mathbb{E}\bigl\{\|W\|\Delta\bigr\}\\ &\quad + 2\sigma^{-1}\sum_{i=1}^n \mathbb{E}\bigl\{\|\xi_i\||\Delta-\Delta^{(i)}|\bigr\},\end{aligned}$$

*for any random variables $\Delta$ and $(\Delta^{(i)})_{1\leqslant i\leqslant n}$ such that $\Delta\geqslant\|D\|$ and $\Delta^{(i)}$ is independent of $X_i$, where $\gamma$ is as defined in (1.4).*

## 3. Applications

In this section, we apply Theorem 2.1 to M-estimators and stochastic gradient descent algorithms.

### *3.1. M-estimators*

Let $X, X_1, \dots, X_n$ be i.i.d. random variables with common probability distribution $P$ that take values in a measurable space $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$. For any function $f:\mathcal{X}\mapsto\mathbb{R}$, let

$$\mathbb{P}_n f = \frac{1}{n}\sum_{i=1}^n f(X_i), \quad Pf = \int_{\mathcal{X}} f(x)P(dx), \quad \mathbb{G}_n f = \sqrt{n}(\mathbb{P}_n - P)f. \tag{3.1}$$

Let $\Theta\subset\mathbb{R}^d$ be a parameter space. For each $\theta\in\Theta$, let $m_\theta(\cdot):\mathcal{X}\mapsto\mathbb{R}$ be twice differentiable with respect to $\theta$, and write

$$\mathbb{M}_n(\theta) = \mathbb{P}_n m_\theta, \quad M(\theta) = Pm_\theta. \tag{3.2}$$

Following the notations in Van der Vaart [31], we briefly write

$$\dot{m}_\theta(x) = \nabla_\theta m_\theta(x), \quad \ddot{m}_\theta(x) = \nabla_\theta^2 m_\theta(x), \tag{3.3}$$

where $\nabla_\theta m_\theta(x)$ is the gradient with respect to $\theta$. Let

$$\theta^* = \arg\min_{\theta\in\Theta} M(\theta) \tag{3.4}$$

and we say $\hat{\theta}_n$ is an *M-estimator* of $\theta^*$ if

$$\hat{\theta}_n = \arg\min_{\theta\in\Theta} \mathbb{M}_n(\theta). \tag{3.5}$$

For any $p \geqslant 1$ and $Y \in \mathbb{R}^d$, let $\|Y\|_p = (\mathbb{E}\{\|Y\|^p\})^{1/p}$ be the $L_p$-norm of $Y$.

The asymptotic properties for M-estimators have been well studied in the literature, and we refer to Van der Vaart and Wellner [32], Van der Vaart [31] and the references therein for a thorough reference. Under some regularity conditions, one has $\hat{\theta}_n \xrightarrow{p} \theta^*$, and Pollard [22] showed that $\sqrt{n}(\hat{\theta}_n - \theta^*)$ converges weakly to a $d$-dimensional normal distribution. The convergence rate was also studied by many authors, for instance, Pfanzagl [19, 20, 21] proved a Berry–Esseen bound of order $O(n^{-1/2})$ for the minimum contrast estimates under some regularity conditions.

In this subsection, we provide a Berry–Esseen bound for $\sqrt{n}(\hat{\theta}_n - \theta^*)$ under some convexity conditions, which are different from those in Pfanzagl [20]. For symmetric matrices $A$ and $B$, denote by $A \preccurlyeq$ (resp. $\succcurlyeq$)$B$ if $A-B$ is non-positive (resp. non-negative) definite. We first propose the following two assumptions.

(M1) The function $m_\theta(\cdot)$ is twice differentiable with respect to $\theta$ and there exist constants $\mu > 0, c_1 > 0, c_2 > 0$ and two nonnegative functions $m_1, m_2 : \mathcal{X} \mapsto \mathbb{R}$ with $\|m_1(X)\|_9 \leqslant c_1$ and $\|m_2(X)\|_4 \leqslant c_2$, such that for any $\theta \in \Theta$,

$$M(\theta) - M(\theta^*) \geqslant \mu\|\theta - \theta^*\|^2, \tag{3.6}$$

$$|m_\theta(x) - m_{\theta^*}(x)| \leqslant m_1(x)\|\theta - \theta^*\|, \quad \forall\, x \in \mathcal{X}, \tag{3.7}$$

and

$$\|\ddot{m}_\theta(x) - \ddot{m}_{\theta^*}(x)\| \leqslant m_2(x)\|\theta - \theta^*\|, \quad \forall\, x \in \mathcal{X}. \tag{3.8}$$

Moreover, there exists a constant $c_3 \geqslant 0$ and a nonnegative function $m_3 \colon \mathcal{X} \mapsto \mathbb{R}$ such that for any $x \in \mathcal{X}$,

$$\ddot{m}_{\theta^*}(x) \preccurlyeq m_3(x) I_d \text{ and } \|m_3(X)\|_4 \leqslant c_3. \tag{3.9}$$

(M2) Let $\xi_i = \dot{m}_{\theta^*}(X_i) \coloneqq (\xi_{i,1}, \ldots, \xi_{i,d})^\intercal$, $\Sigma = \mathbb{E}\{\xi_i\xi_i^\intercal\}$ and $V = \mathbb{E}\{\ddot{m}_{\theta^*}(X)\}$. Assume that there exist constants $\lambda_1 > 0$ and $\lambda_2 > 0$ such that $\lambda_{\min}(\Sigma) \geqslant \lambda_1$ and $\lambda_{\min}(V) \geqslant \lambda_2$. Moreover, assume that there exists a constant $c_4 > 0$ such that

$$\|\xi_1\|_4 \leqslant c_4 d^{1/2}. \tag{3.10}$$

The following theorem provides a Berry–Esseen bound for the M-estimators.

**Theorem 3.1.** *Let $\theta^*$ and $\hat{\theta}_n$ be defined as in (3.4) and (3.5). Under the conditions (M1) and (M2), we have*

$$\sup_{A\in\mathcal{A}}\Big|\mathbb{P}\big(n^{1/2}\Sigma^{-1/2}V(\hat{\theta}_n-\theta^*)\in A\big)-\mathbb{P}(Z\in A)\Big|\leqslant Cd^{9/4}n^{-1/2},$$

*where $C>0$ is a constant depending only on $c_1, c_2, c_3, c_4, \mu, \lambda_1$ and $\lambda_2$.*

**Remark 3.1.** The assumptions (M1) and (M2) are neater than those in Pfanzagl [20]. Moreover, Theorem 3.1 provides a Berry–Esseen bound with the dependence on the dimension.

**Remark 3.2.** Based on the proof of Theorem 3.1, if we further assume that $|m_2(X_i)|\leqslant c_2$ for each $1\leqslant i\leqslant n$ almost surely, then the assumption for $m_1(x)$ can be replaced by $\|m_1(X)\|_5\leqslant c_1$. The condition (3.10) is satisfied if $\|\xi_{ij}\|_4\leqslant c_4$ for all $1\leqslant i\leqslant n$ and $1\leqslant j\leqslant d$.

**Remark 3.3.** The twice differentiability of $m_\theta(x)$ holds for many applications. However, in general, $\ddot{m}_\theta(x)$ does not necessarily exist. We will discuss this case in the next subsection.

When $m_\theta(\cdot)$ is smooth in $\theta$, one can compute $\hat{\theta}_n$ by solving the score equation

$$\mathbb{P}_n\dot{m}_\theta=\frac{1}{n}\sum_{i=1}^n\dot{m}_\theta(X_i)=0.$$

More generally, we can consider the estimating equations of the following type. Let $\Theta\subset\mathbb{R}^d$ be the parameter space and for each $\theta\in\Theta$, let $h_\theta:\mathcal{X}\mapsto\mathbb{R}^d$, and let

$$\Psi_n(\theta)=\frac{1}{n}\sum_{i=1}^n h_\theta(X_i),\quad \Psi(\theta)=\mathbb{E}\{h_\theta(X)\}.$$

Let $\hat{\theta}_n$ and $\theta^*$ satisfy

$$\Psi_n(\hat{\theta}_n)=0,\quad \Psi(\theta^*)=0. \tag{3.11}$$

The estimator $\hat{\theta}_n$ in (3.11) is often called a *Z-estimator* of $\theta^*$, see e.g., Van der Vaart [31]. However, although there is no maximization in (3.11), the estimator $\hat{\theta}_n$ is also called an M-estimator of $\theta^*$. Assume that $\Psi(\theta)$ is differentiable at $\theta^*$ and there exists a $d\times d$ matrix $\dot{\Psi}_0$ satisfying

$$\Psi(\theta)-\Psi(\theta^*)-\dot{\Psi}_0(\theta-\theta^*)=\mathrm{o}(\|\theta-\theta^*\|)\quad\text{as }\theta\to\theta^*.$$

Under some regularity conditions and the so called "asymptotic equi-continuity" condition, Huber [15] proved that $\sqrt{n}(\hat{\theta}_n-\theta^*)$ converges in distribution to $\dot{\Psi}_0^{-1}Z$, where $Z\sim N(0,\mathbb{E}\{h_{\theta^*}(X_i)h_{\theta^*}(X_i)^\intercal\})$. Bentkus, Bloznelis and Götze [6] proved a Berry–Esseen bound of order $\mathrm{O}(n^{-1/2})$ for the 1-dimensional case under some convexity conditions, and Paulauskas [18] proved a convergence rate result for the $d$-dimensional case under some *smooth stochastic differentiability* conditions, which are different from the conditions (B1)–(B5) below.

Let $p \geqslant 3$ be a fixed number, and we make the following assumptions.

(B1) There exist positive constants $\mu, c_1$ and $\lambda_1$ and a positive definite matrix $\dot{\Psi}_0$ such that

$$\big\langle \Psi(\theta_1) - \Psi(\theta_2), \theta_1 - \theta_2 \big\rangle \geqslant \mu \|\theta_1 - \theta_2\|^2, \tag{3.12}$$

and

$$\|\Psi(\theta) - \Psi(\theta^*) - \dot{\Psi}_0(\theta - \theta^*)\| \leqslant c_1 \|\theta - \theta^*\|^2, \quad \lambda_{\min}(\dot{\Psi}_0) \geqslant \lambda_1. \tag{3.13}$$

(B2) Let $h_{\theta,j}$ be the $j$-th element of $h_\theta$. There exists a function $h_0 \colon \mathcal{X} \mapsto \mathbb{R}_+$ and a constant $c_2 > 0$ such that for any $\theta, \theta' \in \Theta$,

$$\big| h_{\theta,j}(X) - h_{\theta',j}(X) \big| \leqslant h_0(X) \|\theta - \theta'\|. \tag{3.14}$$

and

$$\|h_0(X)\|_p \leqslant c_2. \tag{3.15}$$

(B3) Let $\xi_i = h_{\theta^*}(X_i)$ and $\Sigma = \mathbb{E}\{\xi_i \xi_i^\intercal\}$. Assume that there exist positive constants $c_3$ and $\lambda_2$ such that

$$\lambda_{\min}(\Sigma) \geqslant \lambda_2, \tag{3.16}$$

and

$$\|\xi_1\|_p \leqslant c_3 d^{1/2}. \tag{3.17}$$

**Remark 3.4.** Following notations in Theorem 3.1, we can choose $h_\theta(x) = \dot{m}_\theta(x)$. Note that the assumption (B1) is weaker than (M1) in the sense of the differentiability of $h_\theta$, because we assume that the differentiability only holds for $\Psi(\theta)$ rather than $h_\theta(x)$.

**Theorem 3.2.** *Let $\hat{\theta}_n$ and $\theta^*$ be defined as in (3.11). Let $p \geqslant 3$ and $D_\Theta \coloneqq \sup_{\theta_1,\theta_2 \in \Theta} \|\theta_1 - \theta_2\|$, the diameter of the parameter space $\Theta$. Assume that conditions (B1)–(B3) are satisfied. Then,*

$$\sup_{A \in \mathcal{A}} \big| \mathbb{P}\big(\sqrt{n}\Sigma^{-1/2}\dot{\Psi}_0(\hat{\theta}_n - \theta^*) \in A\big) - \mathbb{P}\big(Z \in A\big) \big| \leqslant C (D_\Theta + 1)^2 d^{7/2} n^{-1/2 + \varepsilon_p}. \tag{3.18}$$

*where $\varepsilon_p = 1/(2p-2)$ and $C > 0$ is a constant depending on $p, c_1, c_2, c_3, \lambda_1, \lambda_2$ and $\mu$.*

**Remark 3.5.** Under some different conditions and assuming that $\mathbb{E}\|\xi_i\|^3$ is bounded, Paulauskas [18, Theorem 9] proved a bound of order $n^{-1/4}(\log n)^{3/4}$. In Theorem 3.2 with $p = 3$, the result (3.18) reduces to $D_\Theta^2 d^{7/2} n^{-1/4}$, which is of a sharper order than Paulauskas [18]. Moreover, Theorem 3.2 provides a result with the dependence on the dimension $d$.

The order $n^{-1/2+\varepsilon_p}$ can be improved to $n^{-1/2}\log n$ under some stronger conditions. Let us introduce the so-called *Orlicz norm*, one may refer to Van der Vaart and Wellner [32, Section 2.2] for more details. Let $\psi : [0,\infty) \mapsto [0,\infty)$ be a nondecreasing, convex function with $\psi(0) = 0$. Let $Y$ be a $\mathbb{R}^d$-valued random variable, and define the Orlicz norm of $Y$ with respect to $\psi$ to be

$$\|Y\|_\psi = \inf\Big\{C > 0 : \mathbb{E}\Big\{\psi\Big(\frac{\|Y\|}{C}\Big)\Big\} \leqslant 1\Big\}. \tag{3.19}$$

Specially, if we choose $\psi(x) = x^p$ for $p \geqslant 1$, then the corresponding Orlicz norm is simply the $L_p$-norm. Let $\psi_1(x) \coloneqq e^x - 1$. Now we propose the following assumptions.

(B4) The condition (3.15) in (B2) is replaced by

$$\|h_0(X)\|_{\psi_1} \leqslant c_4. \tag{3.20}$$

where $c_4 > 0$ is a constant.

(B5) The condition (3.17) in (B3) is replaced by

$$\|\xi_1\|_{\psi_1} \leqslant c_5, \tag{3.21}$$

where $c_5 > 0$ is a constant.

**Remark 3.6.** Let $Y$ be a random variable. It can be shown (see Vershynin [33, (5.14)–(5.16)] for example) that, there exist positive constants $K_1, K_2, K_3$ that differ from each other by at most an absolute constant factor such that the following are equivalent:

(a) $\|Y\|_{\psi_1} \leqslant K_1$;
(b) $\mathbb{P}(|Y| \geqslant t) \leqslant \exp\{1 - t/K_2\}$ for all $t \geqslant 0$;
(c) $\|Y\|_p \leqslant K_3 p$ for all $p \geqslant 1$.

We have the following theorem.

**Theorem 3.3.** *Let $\hat\theta_n, \theta^*$ and $D_\Theta$ be defined as in Theorem 3.2. Under the assumptions (B1), (B4) and (B5),*

$$\sup_{A\in\mathcal{A}}\big|\mathbb{P}\big(\sqrt{n}\Sigma^{-1/2}\dot\Psi_0(\hat\theta_n - \theta^*) \in A\big) - \mathbb{P}\big(Z \in A\big)\big| \leqslant C(D_\Theta + 1)^2 d^4 n^{-1/2}\log n,$$

*where $C > 0$ is a constant depending on $c_1, c_4, c_5, \lambda_1, \lambda_2$ and $\mu$.*

### *3.2. Averaged stochastic gradient descent algorithms*

Consider the problem of searching for the minimum point $\theta^*$ of a smooth function $f(\theta), \theta \in \Theta \subset \mathbb{R}^d$. The stochastic gradient descent method provides a direct way to solve the minimization problem. In this subsection, we consider the averaged stochastic gradient descent algorithm, which is proposed by Polyak [23] and Ruppert [27]. The algorithm is given as follows: Let $\theta_0 \in \mathbb{R}^d$ be the initial value (might be random), and for $n \geqslant 1$, we update $\theta_n$ by

$$\begin{aligned} \theta_n &= \theta_{n-1} - \ell_n\big(\nabla f(\theta_{n-1}) + \zeta_n\big), \\ \bar{\theta}_n &= \frac{1}{n}\sum_{i=0}^{n-1} \theta_i. \end{aligned} \tag{3.22}$$

where $\ell_n > 0$ is the so called *learning rate* and $(\zeta_1, \zeta_2, \dots)$ is a sequence of $\mathbb{R}^d$-valued martingale differences. The convergence rate of $\mathbb{E}\|\theta_n - \theta^*\|^2$ and $\mathbb{E}\|\bar{\theta}_n - \theta^*\|^2$ was thoroughly studied in the literature, see Polyak [23] and Bach and Moulines [2]. The normality of $\sqrt{n}(\bar{\theta}_n - \theta^*)$ is also well-known, see Polyak and Juditsky [24]. Suppose that the learning rate $\ell_n = \ell_0 n^{-\alpha}$ where $\alpha \in (1/2, 1)$, under some regularity conditions, Polyak and Juditsky [24] proved that $\sqrt{n}(\bar{\theta}_n - \theta^*)$ converges weakly to a multivariate normal distribution. Recently, Anastasiou, Balasubramanian and Erdogdu [1] used Stein's method and the techniques of martingales to prove a convergence rate for a class of smooth test functions, see Anastasiou, Balasubramanian and Erdogdu [1, Theorem 4] for more details.

In this subsection, we provide a Berry–Esseen bound for the normal approximation for $\sqrt{n}(\bar{\theta}_n - \theta^*)$.

We make the following assumptions:

(C0) There exists a constant $\tau_0 > 0$ such that $\|\theta_0 - \theta^*\|_4 \leqslant \tau_0$.

(C1) The sequence $(\zeta_1, \zeta_2, \dots)$ is independent of $\theta_0$, and for each $n \geqslant 1$, $\zeta_n$ admits the decomposition

$$\zeta_n = \xi_n + \eta_n,$$

where

(i). $(\xi_1, \xi_2, \dots)$ is a sequence of independent random variables and $\mathbb{E}\{\xi_i\} = 0$ and $\mathbb{E}\big\{\xi_i \xi_i^{\intercal}\big\} = \Sigma_i$; there exist positive numbers $\lambda_1$ and $\lambda_2$ such that for any $i \geqslant 1$, $\lambda_1 \leqslant \lambda_{\min}(\Sigma_i) \leqslant \lambda_{\max}(\Sigma_i) \leqslant \lambda_2$; moreover, there exists a positive number $\tau$ such that

$$\max_{1 \leqslant i \leqslant n} \|\xi_i\|_4 \leqslant \tau;$$

(ii). let $\mathcal{F}_0 = \sigma\{\theta_0\}$, and for each $n \geqslant 0$, $\mathcal{F}_n = \sigma\{\theta_0, \xi_1, \dots, \xi_n\}$; let $g(\cdot,\cdot) : \mathbb{R}^d \times \mathbb{R}^d \mapsto \mathbb{R}^d$, and the random variable $\eta_n := g(\theta_{n-1}, \xi_n)$ satisfies $\mathbb{E}\{\eta_n | \mathcal{F}_{n-1}\} = 0$ and for any $\theta$ and $\theta'$, there exists a nonnegative number $c_1 \geqslant 0$ such that

$$\|g(\theta, \xi) - g(\theta', \xi)\| \leqslant c_1 \|\theta - \theta'\| \quad \text{and} \quad g(\theta^*, \xi) = 0 \quad \text{for } \xi \in \mathbb{R}^d. \tag{3.23}$$

(C2) The function $f$ is $L$-*smooth* and *strongly convex* with convexity constant $\mu > 0$, i.e., $f$ is twice differentiable and there exist two constants $\mu > 0$ and $L > 0$ such that

$$\mu I_d \preccurlyeq \nabla^2 f(\theta) \preccurlyeq L I_d, \text{ for all } \theta \in \Theta. \tag{3.24}$$

(C3) There exist positive constants $c_2$ and $\beta$ such that for all $\theta$ with $\|\theta - \theta^*\| \leqslant \beta$,

$$\big\|\nabla^2 f(\theta) - \nabla^2 f(\theta^*)\big\| \leqslant c_2 \|\theta - \theta^*\|. \tag{3.25}$$

Let $G := \nabla^2 f(\theta^*)$. Recall that $(\ell_n)_{n\geqslant 1}$ is the learning rate sequence in (3.22), and let

$$Q_i = \ell_i \prod_{j=i}^{n-1} \prod_{k=i+1}^{j} (I_d - \ell_k G).$$

Here, for any $n \geqslant 0$, set $\prod_{i=n+1}^n A_i = I_d, \prod_{i=n+1}^n a_i = 1$, where $(A_i)_{i\geqslant 1}$ is a $\mathbb{R}^{d\times d}$-valued sequence and $(a_i)_{i\geqslant 1}$ is a $\mathbb{R}$-valued sequence. Let

$$\Sigma_n = \frac{1}{n} \sum_{i=1}^{n-1} Q_i \Sigma_i Q_i^{\intercal}.$$

We have the following theorem.

**Theorem 3.4.** *Let $\ell_n = \ell_0 n^{-\alpha}$ where $\ell_0 > 0$ and $1/2 < \alpha \leqslant 1$. Under the assumptions (C0)–(C3), we have*

(1) *if $\alpha \in (1/2, 1)$,*

$$\begin{aligned} &\sup_{A\in\mathcal{A}} \Big| \mathbb{P}\big(\sqrt{n}\Sigma_n^{-1/2}(\bar{\theta}_n - \theta^*) \in A\big) - \mathbb{P}(Z \in A) \Big| \\ &\qquad \leqslant C\big(d^{3/2} + \tau^3 + \tau_0^3\big)(d^{1/2} n^{-1/2} + n^{-\alpha+1/2}); \end{aligned} \tag{3.26}$$

(2) *if $\ell_n = \ell_0 n^{-1}$ with $\ell_0\mu \geqslant 1$, we have*

$$\begin{aligned} &\sup_{A\in\mathcal{A}} \Big| \mathbb{P}\big(\sqrt{n}\Sigma_n^{-1/2}(\bar{\theta}_n - \theta^*) \in A\big) - \mathbb{P}(Z \in A) \Big| \\ &\qquad \leqslant C n^{-1/2} (d^{3/2} + \tau^3 + \tau_0^3) \times \begin{cases} d^{1/2} + \log n, & \ell_0\mu > 1; \\ d^{1/2} (\log n)^3, & \ell_0 \mu = 1. \end{cases} \end{aligned} \tag{3.27}$$

*Here, $C > 0$ is a constant depending only on $\ell_0, \lambda_1, \lambda_2, c_1, c_2, \alpha, \beta, L$ and $\mu$ and independent of $d, \tau$ and $\tau_0$.*

**Remark 3.7.** Typically, $\tau \sim \tau_0 \sim d^{1/2}$. Specially, if $\alpha = 1 - \varepsilon$ with an arbitrary $0 < \varepsilon < 1/2$, then the RHS of (3.26) reduces to $C(d^2 n^{-1/2} + d^{3/2} n^{-1/2+\varepsilon})$. If $\alpha = 1$ with $\ell_0 \mu \geqslant 1$, the Berry–Esseen bound (3.27) is of an optimal order up to a polynomial of a $(\log n)^3$ factor.

**Remark 3.8.** For $\alpha = 1$, it has been proved (see Bach and Moulines [2, Theorem 2] and also Lemma 5.12) that

$$\mathbb{E}\|\theta_n - \theta^*\|^2 \leqslant \begin{cases} n^{-1}, & \ell_0\mu > 1; \\ n^{-1}(\log n), & \ell_0\mu = 1; \\ n^{-\ell_0\mu/2}, & 0 < \ell_0\mu < 1. \end{cases}$$

Therefore, for $\alpha = 1$, the choice of $\ell_0$ is critical, but the problem is: a small $\ell_0$ leads to a very slow convergence rate of order $n^{-\ell_0\mu/2}$ while a large $\ell_0$ might lead to explosion due to the initial condition (see, e.g., Bach and Moulines [2] and Nemirovski, Juditsky, Lan and Shapiro [17] for more details). In practice, one prefers to use a learning rate of order $n^{-\alpha}$ with $0 < \alpha < 1$.

**Theorem 3.5.** *Consider the model* (3.2)*. Let*

$$\theta^* = \min_{\theta\in\mathbb{R}^d} M(\theta),$$

*and the algorithm*

$$\theta_n = \theta_{n-1} - \ell_n \dot{m}_{\theta_{n-1}}(X_n),$$

*where* $\dot{m}_\theta$ *is as in* (3.3)*,* $\ell_n = \ell_0 n^{-\alpha}$ *is the learning rate,* $\ell_0 > 0$*,* $1/2 < \alpha \leqslant 1$ *and* $\theta_0$ *is the initial value that is independent of* $(X_1, \dots, X_n)$*. Let*

$$\begin{aligned} \xi_n &= \dot{m}_{\theta^*}(X_n) - \nabla M(\theta^*), \\ \eta_n &= \dot{m}_{\theta_{n-1}}(X_n) - \dot{m}_{\theta^*}(X_n) - \nabla M(\theta_{n-1}) + \nabla M(\theta^*). \end{aligned}$$

*Assume that* (C1(i)) *is satisfied for* $(\xi_1, \dots, \xi_n)$ *and for any* $\theta_1, \theta_2 \in \mathbb{R}^d$*,*

$$\sup_{z\in\mathcal{X}} \|\dot{m}_{\theta_1}(z) - \dot{m}_{\theta_2}(z)\| \leqslant L_F \|\theta_1 - \theta_2\|. \tag{3.28}$$

*Assume further that* (C0)*,* (C2) *and* (C3) *are satisfied with* $f(\theta) = M(\theta)$*, and let* $\bar{\theta}_n$ *be as defined in* (3.22)*. Then, we have* (3.26) *and* (3.27) *hold with* $c_1 = 2L_F$*.*

*Proof.* We only need to check the condition (C1(ii)) is satisfied. Note that for each $n \geqslant 1$,

$$\begin{aligned} \dot{m}_{\theta_{n-1}}(X_n) &= \nabla M(\theta_{n-1}) + \big(\dot{m}_{\theta_{n-1}}(X_n) - \nabla M(\theta_{n-1})\big) \\ &= \nabla M(\theta_{n-1}) + \big(\dot{m}_{\theta^*}(X_n) - \nabla M(\theta^*)\big) \\ &\quad + \big(\dot{m}_{\theta_{n-1}}(X_n) - \dot{m}_{\theta^*}(X_n) - \nabla M(\theta_{n-1}) + \nabla M(\theta^*)\big) \\ &= \nabla M(\theta_{n-1}) + \xi_n + \eta_n. \end{aligned}$$

For $n \geqslant 0$, let $\mathcal{F}_n = \sigma(\theta_0, X_1, \dots, X_n)$. Then we have $\mathbb{E}\{\xi_n\} = 0$ and $\mathbb{E}\big\{\eta_n \mid \mathcal{F}_{n-1}\big\} = 0$. By (3.28), it follows that the condition (C1(ii)) in (3.6) holds with $c_1 = 2L_F$. Hence, Theorem 3.4 implies the desired result. □

## 4. Proofs of main results

### *4.1. A randomized concentration inequality*

To prove (2.1), we need to develop a randomized concentration inequality for sums of multivariate independent random vectors. We use the following notation. For a subset $A$ of $\mathbb{R}^d$, let $d(x, A) = \inf\{\|x - y\| : y \in A\}$. For a given number $\varepsilon > 0$, define $A^\varepsilon = \big\{x \in \mathbb{R}^d : d(x, A) \leqslant \varepsilon\big\}$, and $A^{-\varepsilon} = \{x \in A : B(x, \varepsilon) \subset A\}$, where $B(x, \varepsilon)$ is the $d$-dimensional ball centered in $x$ with radius $\varepsilon$. Specially, for $\varepsilon = 0$, let $A^\varepsilon = A$. Let $\bar{A}$ be the closure of $A$ and let $r(\bar{A}) = \max\{y : B(x, y) \subset \bar{A}$ for some $x \in \mathbb{R}^n\}$ be the *inradius* of $\bar{A}$. For $a, b \in \mathbb{R}$, write $a \wedge b = \min(a, b)$ and $a \vee b = \max(a, b)$. Let $\gamma = \sum_{i=1}^n \mathbb{E}\{\|\xi_i\|^3\}$ be as in (1.4). We have the following proposition.

**Proposition 4.1.** *Let $W = \sum_{i=1}^n \xi_i$, where $(\xi_i)_{i=1}^n$ is a sequence of $\mathbb{R}^d$-valued independent random vectors satisfying that $\mathbb{E}\{\xi_i\} = 0$ for $1 \leqslant i \leqslant n$ and $\sum_{i=1}^n \mathbb{E}\{\xi_i \xi_i^\intercal\} = I_d$. Let $\Delta_1$ and $\Delta_2$ be nonnegative random variables. Then we have for all $A \in \mathcal{A}$ such that $r(\bar{A}) > \gamma$,*

$$\mathbb{P}\big(W \in A^{4\gamma+\Delta_1} \setminus A^{4\gamma-\bar{\Delta}_2}\big) \leqslant 19d^{1/2}\gamma + 2\,\mathbb{E}\big\{\|W\|(\Delta_1 + \Delta_2)\big\} + 2\sum_{i=1}^n \sum_{j=1}^2 \mathbb{E}\big\{\|\xi_i\||\Delta_j - \Delta_j^{(i)}|\big\}, \tag{4.1}$$

*where $\bar{\Delta}_2 = \Delta_2 \wedge (r(\bar{A}) - \gamma)$ and $\Delta^{(i)}$ is a random variable independent of $\xi_i$.*

The proof of this proposition is postponed in Subsection 4.3.

**Remark 4.1.** Specially, if $\Delta_1 = \varepsilon$ and $\Delta_2 = 0$ where $\varepsilon > 0$ is a constant, then (4.1) reduces to

$$P(W \in A^{4\gamma+\varepsilon} \setminus A^{4\gamma}) \leqslant 2d^{1/2}\varepsilon + 19d^{1/2}\gamma,$$

which is equivalent to the result in Chen and Fang [10] up to a constant factor.

**Remark 4.2.** When $d = 1$, the right hand side of (4.1) reduces to

$$19\gamma + 2\,\mathbb{E}|W(\Delta_1 + \Delta_2)| + 2\sum_{i=1}^n \sum_{j=1}^2 \mathbb{E}|\xi_i(\Delta_j - \Delta_j^{(i)})|,$$

which is equivalent to Chen and Shao [9]'s concentration inequality result. Recently, Shao and Zhou [29] proved that the term $\mathbb{E}|W\Delta|$ can be improved to be $\mathbb{E}|\Delta|$ in (1.5). However, due to some technical difficulty, we are not able to remove the $W$ term in our result. Nevertheless, the order in $n$ is optimal in many applications.

### *4.2. Proofs of Theorem 2.1 and Corollaries 2.2 and 2.3*

We first give the proof of Theorem 2.1.

*Proof of Theorem 2.1.* Without loss of generality, let $A$ be an arbitrary nonempty convex subset of $\mathbb{R}^d$. Let $Z \sim N(0, I_d)$ be independent of all others. It has been shown in Chen and Fang [10, Proposition 2.5 and Theorem 3.5] that for $\varepsilon_1, \varepsilon_2 \geqslant 0$,

$$\sup_{A \in \mathcal{A}} \big|\mathbb{P}(W \in A) - \mathbb{P}(Z \in A)\big| \leqslant 115d^{1/2}\gamma, \tag{4.2}$$

$$\mathbb{P}(Z \in A^{\varepsilon_1} \setminus A^{-\varepsilon_2}) \leqslant d^{1/2}(\varepsilon_1 + \varepsilon_2). \tag{4.3}$$

For each $1 \leqslant i \leqslant n$, let $\Delta^{(i)}$ be any random variable that is independent of $\xi_i$. Note that $\|T - W\| \leqslant \Delta$ and that $r(\bar{A}^{2\gamma}) > \gamma$. Applying Proposition 4.1 to $A^{2\gamma}$ with $\Delta_1 = \Delta$ and $\Delta_2 = 0$, and by (4.2) and (4.3), we have

$$\begin{aligned}
&\mathbb{P}(T \in A) - \mathbb{P}(Z \in A) \\
&\quad \leqslant \mathbb{P}(T \in A^{6\gamma}) - \mathbb{P}(W \in A^{6\gamma}) + \mathbb{P}(W \in A^{6\gamma}) - \mathbb{P}(Z \in A^{6\gamma}) + \mathbb{P}(Z \in A^{6\gamma} \setminus A) \\
&\quad \leqslant \mathbb{P}(W \in (A^{2\gamma})^{\Delta + 4\gamma} \setminus (A^{2\gamma})^{4\gamma}) + 121 d^{1/2}\gamma \\
&\quad \leqslant 140 d^{1/2}\gamma + 2\,\mathbb{E}\big\{\|W\|\Delta\big\} + 2\sum_{i=1}^{n} \mathbb{E}\big\{\|X_i\||\Delta - \Delta^{(i)}|\big\}.
\end{aligned} \tag{4.4}$$

This proves the upper bound of $\mathbb{P}(T \in A) - \mathbb{P}(Z \in A)$. For the upper bound of $\mathbb{P}(Z \in A) - \mathbb{P}(T \in A)$, we introduce the following notation. Recall that $\bar{A}$ is the closure of $A$ and $r := r(\bar{A})$ is the inradius of $\bar{A}$. We consider the following two cases.

If $r < 9\gamma$, then $A^{-9\gamma} = \emptyset$. By (4.3),

$$\mathbb{P}(Z \in A) - \mathbb{P}(T \in A) \leqslant \mathbb{P}(Z \in A \setminus A^{-9\gamma}) \leqslant 9\, d^{1/2}\gamma. \tag{4.5}$$

Now we consider the case where $r \geqslant 9\gamma$. Let $A_0 = A^{-4\gamma}$ and it follows that $A_0 \neq \emptyset$ and $r(\bar{A}_0) = r - 4\gamma$. Let $\Delta_0 = \Delta \wedge (r - 5\gamma) = \Delta \wedge (r(\bar{A}_0) - \gamma)$. Since $A_0^{4\gamma} = (A^{-4\gamma})^{4\gamma} \subset A$, we have

$$\mathbb{P}\big(Z \in A\big) - \mathbb{P}\big(T \in A\big) \leqslant \mathbb{P}\big(Z \in A\big) - \mathbb{P}\big(T \in A_0^{4\gamma}\big) = Q_1 + Q_2 + Q_3$$

where

$$\begin{aligned}
Q_1 &= \mathbb{P}\big(Z \in A\big) - \mathbb{P}\big(Z \in A_0^{4\gamma}\big), \\
Q_2 &= \mathbb{P}\big(Z \in A_0^{4\gamma}\big) - \mathbb{P}\big(W \in A_0^{4\gamma}\big), \\
Q_3 &= \mathbb{P}\big(W \in A_0^{4\gamma}\big) - \mathbb{P}\big(T \in A_0^{4\gamma}\big).
\end{aligned}$$

For $Q_1$, by (4.3), we have

$$|Q_1| \leqslant \mathbb{P}\big(Z \in A \setminus A_0\big) \leqslant 4 d^{1/2}\gamma.$$

For $Q_2$, noting that $A_0^{4\gamma}$ is also convex, by (4.2), we have

$$|Q_2| \leqslant 115 d^{1/2}\gamma.$$

We now move to give an upper bound of $Q_3$. If $0 \leqslant \Delta \leqslant r - 5\gamma$,

$$\mathbb{1}\{w \in A_0^{4\gamma}\} - \mathbb{1}\{w + D \in A_0^{4\gamma}\} \leqslant \mathbb{1}\{w \in A_0^{4\gamma} \setminus A_0^{4\gamma - \Delta}\}. \tag{4.6}$$

If $\Delta > r - 5\gamma$, then

$$\begin{aligned}
&\mathbb{1}\{w \in A_0^{4\gamma}\} - \mathbb{1}\{w + D \in A_0^{4\gamma}\} \\
&\leqslant \mathbb{1}\{w \in A_0^{4\gamma}\} \\
&\leqslant \mathbb{1}\{w \in A_0^{4\gamma} \setminus A_0^{9\gamma - r}\} + \mathbb{1}\{w \in A_0^{9\gamma - r}\} \\
&\leqslant \mathbb{1}\{w \in A_0^{4\gamma} \setminus A_0^{4\gamma - (r - 5\gamma)}\} + \mathbb{1}\{w \in A^{5\gamma - r}\},
\end{aligned} \tag{4.7}$$

where the last line follows from the fact that $(A^{-4\gamma})^{9\gamma-r} \subset A^{5\gamma-r}$. Equations (4.6) and (4.7) yield

$$\begin{aligned} Q_3 &= \mathbb{P}\big(W \in A_0^{4\gamma}\big) - \mathbb{P}\big(W + D \in A_0^{4\gamma}\big) \\ &\leqslant \mathbb{P}\big(W \in A_0^{4\gamma} \setminus A_0^{4\gamma-\Delta_0}\big) + \mathbb{P}\big(W \in A^{5\gamma-r}\big). \end{aligned} \tag{4.8}$$

For each $1 \leqslant i \leqslant n$, let $\Delta_0^{(i)} = \Delta^{(i)} \wedge (r(\bar{A}_0) - \gamma)$. For the first term of (4.8), by Proposition 4.1, we have

$$\begin{aligned} \mathbb{P}\big(W \in A_0^{4\gamma} \setminus A_0^{4\gamma-\Delta_0}\big) &\leqslant 19d^{1/2}\gamma + 2\,\mathbb{E}\{\|W\|\Delta_0\} + 2\sum_{i=1}^n \mathbb{E}\{\|\xi_i\||\Delta_0 - \Delta_0^{(i)}|\} \\ &\leqslant 19d^{1/2}\gamma + 2\,\mathbb{E}\{\|W\|\Delta\} + 2\sum_{i=1}^n \mathbb{E}\{\|\xi_i\||\Delta - \Delta^{(i)}|\}. \end{aligned}$$

For the second term of (4.8), since $A^{-r-\gamma} = \emptyset$ and $A^{5\gamma-r}$ is convex and nonempty, by (4.2) and (4.3), we have

$$\begin{aligned} \mathbb{P}(W \in A^{5\gamma-r}) &\leqslant |\mathbb{P}(W \in A^{5\gamma-r}) - \mathbb{P}(Z \in A^{5\gamma-r})| + \mathbb{P}(Z \in A^{5\gamma-r} \setminus A^{-\gamma-r}) \\ &\leqslant 115d^{1/2}\gamma + 6d^{1/2}\gamma \leqslant 121d^{1/2}\gamma. \end{aligned}$$

Then it follows that

$$Q_3 \leqslant 140d^{1/2}\gamma + 2\,\mathbb{E}\{\|W\|\Delta\} + 2\sum_{i=1}^n \mathbb{E}\{\|\xi_i\||\Delta - \Delta^{(i)}|\}.$$

Combining the upper bounds of $Q_1, Q_2$ and $Q_3$, we have

$$\mathbb{P}(Z \in A) - \mathbb{P}(T \in A) \leqslant 259d^{1/2}\gamma + 2\,\mathbb{E}\big\{\|W\|\Delta\big\} + 2\sum_{i=1}^n \mathbb{E}\big\{\|X_i\||\Delta - \Delta^{(i)}|\big\}. \tag{4.9}$$

By (4.4), (4.5) and (4.9), we have

$$\sup_{A\in\mathcal{A}} |\mathbb{P}(T \in A) - \mathbb{P}(W \in A)| \leqslant 259d^{1/2}\gamma + 2\,\mathbb{E}\{\|W\|\Delta\} + 2\sum_{i=1}^n \mathbb{E}\{\|X_i\||\Delta - \Delta^{(i)}|\},$$

as desired. □

*Proof of Corollary 2.2.* Let $\widetilde{T} = W + D\mathbb{1}(O)$. For any $A \in \mathcal{A}$,

$$|\mathbb{P}(T \in A) - \mathbb{P}(\widetilde{T} \in A)| \leqslant \mathbb{P}(O^c).$$

Applying Theorem 2.1 to $\widetilde{T}$ yields

$$\sup_{A\in\mathcal{A}} \big|\mathbb{P}(\widetilde{T} \in A) - \mathbb{P}(Z \in A)\big| \leqslant 259d^{1/2}\gamma + 2\,\mathbb{E}\big\{\|W\|\Delta\big\} + 2\sum_{i=1}^n \mathbb{E}\big\{\|\xi_i\||\Delta - \Delta^{(i)}|\big\}.$$

Combining the foregoing inequalities we obtain the desired result. □

*Proof of Corollary 2.3.* For any convex set $A \subset \mathbb{R}^d$, we have $\Sigma^{-1/2}A := \{y \in \mathbb{R}^d : y = \Sigma^{-1/2}x, x \in A\}$ is also a convex subset of $\mathbb{R}^d$. To see this, it suffices to show that for any $y_1, y_2 \in \Sigma^{-1/2}A$ and for any $0 \leqslant t \leqslant 1$,

$$ty_1 + (1-t)y_2 \in \Sigma^{-1/2}A. \tag{4.10}$$

Since $y_1, y_2 \in \Sigma^{-1/2}A$, it follows that there exist $x_1, x_2 \in A$ such that

$$y_1 = \Sigma^{-1/2}x_1, \quad y_2 = \Sigma^{-1/2}x_2.$$

Moreover, as $A$ is convex, we have for any $0 \leqslant t \leqslant 1$,

$$tx_1 + (1-t)x_2 \in A,$$

and thus

$$\begin{aligned} ty_1 + (1-t)y_2 &= t\Sigma^{-1/2}x_1 + (1-t)\Sigma^{-1/2}x_2 \\ &= \Sigma^{-1/2}(tx_1 + (1-t)x_2) \in \Sigma^{-1/2}A. \end{aligned}$$

This proves (4.10) and hence $\Sigma^{-1/2}A$ is convex. Note that

$$\mathbb{P}(T \in A) - P(\Sigma^{1/2}Z \in A) = \mathbb{P}(\Sigma^{-1/2}T \in \Sigma^{-1/2}A) - \mathbb{P}(Z \in \Sigma^{-1/2}A),$$

and we have

$$\sup_{A \in \mathcal{A}} \big|\mathbb{P}(T \in A) - \mathbb{P}(\Sigma^{-1/2}A \in A)\big| = \sup_{A \in \mathcal{A}} \big|\mathbb{P}(\Sigma^{-1/2}T \in A) - \mathbb{P}(Z \in A)\big|.$$

Applying Theorem 2.1 yields the desired result. □

### 4.3. Proof of Proposition 4.1

We apply the ideas in Chen and Shao [9] and Chen and Fang [10] to prove Proposition 4.1 in this subsection. Before the proof, we first introduce some definitions and lemmas.

Given $A \in \mathcal{A}$ and $\varepsilon \geqslant 0$, we construct $f_{A,\varepsilon} : \mathbb{R}^d \to \mathbb{R}^d$ as follows. Let $\mathcal{P}_A$ be the projection operator on $A$, that is, for any $x \in \mathbb{R}^d$, let

$$\mathcal{P}_A(x) := \operatorname*{arg\,min}_{y \in A} \|x - y\|.$$

Therefore, $\mathcal{P}_A(x)$ is the nearest point of $x$ in the set $A$.

Let $\bar{A}$ be the closure of $A$, and

$$f_{A,\varepsilon}(x) = \begin{cases} 0, & x \in \bar{A}, \\ x - \mathcal{P}_{\bar{A}}(x), & x \in A^\varepsilon \setminus \bar{A}, \\ \mathcal{P}_{(\bar{A})^\varepsilon}(x) - \mathcal{P}_{\bar{A}}(x), & x \in \mathbb{R}^d \setminus A^\varepsilon. \end{cases} \tag{4.11}$$

Let $r(\bar{A}) = \max\{y : B(x,y) \subset \bar{A}$ for some $x \in \mathbb{R}^d\}$ be the inradius of $\bar{A}$. We introduce the following lemma, whose proof can be found in Chen and Fang [10, Lemmas 2.1, 2.2 and Proposition 2.7].

**Lemma 4.2.** *Let $\varepsilon > 0$ and $\gamma > 0$ and $f := f_{A,\varepsilon+8\gamma}$ be as in (4.11). We have*

(i) $\|f\| \leqslant \varepsilon + 8\gamma$;
(ii) *for all* $\xi, \eta \in \mathbb{R}^d$, $\langle \xi, f(\eta+\xi) - f(\eta)\rangle \geqslant 0$;
(iii) *for* $w \in A^{4\gamma+\varepsilon} \setminus A^{4\gamma}$ *and* $\|x\| \leqslant 4\gamma$, *we have*

$$\langle x, f(w) - f(w-x)\rangle \geqslant \frac{3}{4}(x \cdot h_1)^2,$$

*where* $h_1 = (w_0 - w)/\|w_0 - w\|$ *and* $w_0 = \mathcal{P}_{\bar{A}}(w)$.

Now we are ready to give the proof of Proposition 4.1.

*Proof of Proposition 4.1.* Let $A \in \mathcal{A}$ be nonempty such that $r := r(\bar{A}) > \gamma$. Set $\bar{\Delta}_2 = \Delta_2 \wedge (r-\gamma)$. Let $\Delta_1^{(i)}$ and $\Delta_2^{(i)}$ be any random variables that are independent of $\xi_i$ and let $\bar{\Delta}_2^{(i)} = \Delta_2^{(i)} \wedge (r-\gamma)$. For any $a \geqslant 0$ and $0 \leqslant b \leqslant r-\gamma$, define $g_{a,b} = f_{A^{-b}, 8\gamma+a+b}$. Noting that $\mathbb{E}\xi_i = 0$ and observing that $\Delta_1^{(i)}$ and $\bar{\Delta}_2^{(i)}$ are independent of $\xi_i$, we have

$$\mathbb{E}\big\{\langle \xi_i, g_{\Delta_1^{(i)}, \bar{\Delta}_2^{(i)}}(W - \xi_i)\rangle\big\} = 0,$$

and thus,

$$\begin{aligned}\mathbb{E}\{\langle W, g_{\Delta_1,\bar{\Delta}_2}(W)\rangle\} &= \sum_{i=1}^n \Big(\mathbb{E}\{\langle \xi_i, g_{\Delta_1,\bar{\Delta}_2}(W)\rangle\} - \mathbb{E}\big\{\langle \xi_i, g_{\Delta_1^{(i)},\bar{\Delta}_2^{(i)}}(W-\xi_i)\rangle\big\}\Big)\\ &= H_1 + H_2,\end{aligned} \tag{4.12}$$

where

$$\begin{aligned} H_1 &= \sum_{i=1}^n \mathbb{E}\big\{\langle \xi_i, g_{\Delta_1,\bar{\Delta}_2}(W) - g_{\Delta_1,\bar{\Delta}_2}(W-\xi_i)\rangle\big\},\\ H_2 &= \sum_{i=1}^n \mathbb{E}\big\{\langle \xi_i, g_{\Delta_1,\bar{\Delta}_2}(W-\xi_i) - g_{\Delta_1^{(i)},\bar{\Delta}_2^{(i)}}(W-\xi_i)\rangle\big\}.\end{aligned}$$

For the upper bound of $H_2$, by the definition of $f$, we have

$$\|g_{\Delta_1,\bar{\Delta}_2}(w) - g_{\Delta_1^{(i)},\bar{\Delta}_2^{(i)}}(w)\| \leqslant \big\|g_{\Delta_1,\bar{\Delta}_2}(w) - g_{\Delta_1^{(i)},\bar{\Delta}_2}(w)\big\| + \big\|g_{\Delta_1^{(i)},\bar{\Delta}_2}(w) - g_{\Delta_1^{(i)},\bar{\Delta}_2^{(i)}}(w)\big\|. \tag{4.13}$$

Without loss of generality, assume that $\Delta_1^{(i)} \leqslant \Delta_1$. Let $A_2 = (\bar{A})^{-\bar{\Delta}_2}$, $A_3 = A^{8\gamma+\Delta_1^{(i)}}$, $A_4 = A^{8\gamma+\Delta_1}$ and $w_j = \mathcal{P}_{A_j}(w)$ for $j = 2,3,4$.

If $w \in A_3 \subset A_4$, then

$$g_{\Delta_1,\bar{\Delta}_2}(w) = g_{\Delta_1^{(i)},\bar{\Delta}_2}(w);$$

if $w \in A_4 \setminus A_3$, then

$$g_{\Delta_1,\bar{\Delta}_2}(w) = w - w_2, \quad g_{\Delta_1^{(i)},\bar{\Delta}_2}(w) = w_3 - w_2,$$

and

$$\|w - w_3\| = \|w_4 - w_3\|, \quad \text{for } w \in A_4 \setminus A_3;$$

if $w \in A_4^c$, then

$$g_{\Delta_1, \bar{\Delta}_2}(w) = w_4 - w_2 \text{ and } g_{\Delta_1^{(i)}, \bar{\Delta}_2}(w) = w_3 - w_2.$$

By the definition of $w_3$ and $w_4$, it follows that $\|w_4 - w_3\| \leqslant |\Delta_1 - \Delta_1^{(i)}|$. Hence,

$$\big\| g_{\Delta_1, \bar{\Delta}_2}(w) - g_{\Delta_1^{(i)}, \bar{\Delta}_2}(w) \big\| \leqslant |\Delta_1 - \Delta_1^{(i)}|. \tag{4.14}$$

Similarly,

$$\big\| g_{\Delta_1^{(i)}, \bar{\Delta}_2}(w) - g_{\Delta_1^{(i)}, \bar{\Delta}_2^{(i)}}(w) \big\| \leqslant |\bar{\Delta}_2 - \bar{\Delta}_2^{(i)}| \leqslant |\Delta_2 - \Delta_2^{(i)}|. \tag{4.15}$$

By (4.13)–(4.15),

$$H_2 \leqslant \sum_{i=1}^n \mathbb{E}\{\|\xi_i\|(|\Delta_1 - \Delta_1^{(i)}| + |\Delta_2 - \Delta_2^{(i)}|)\}. \tag{4.16}$$

We next estimate the lower bound of $H_1$. By Lemma 4.2, we have

$$\begin{aligned} H_1 &= \sum_{i=1}^n \mathbb{E}\big\{ \big\langle \xi_i, g_{\Delta_1, \bar{\Delta}_2}(W) - g_{\Delta_1, \bar{\Delta}_2}(W - \xi_i) \big\rangle \big\} \\ &\geqslant \sum_{i=1}^n \mathbb{E}\Big\{ \big\langle \xi_i, g_{\Delta_1, \bar{\Delta}_2}(W) - g_{\Delta_1, \bar{\Delta}_2}(W - \xi_i) \big\rangle \mathbb{1}(|\xi_i| \leqslant 4\gamma) \mathbb{1}(W \in A^{4\gamma + \Delta_1} \setminus A^{4\gamma - \bar{\Delta}_2}) \Big\} \\ &\geqslant \frac{3}{4} \sum_{i=1}^n \mathbb{E}\Big\{ \langle \xi_i, U \rangle^2 \mathbb{1}(\|\xi_i\| \leqslant 4\gamma) \mathbb{1}(W \in A^{4\gamma + \Delta_1} \setminus A^{4\gamma - \bar{\Delta}_2}) \Big\} \coloneqq \frac{3}{4} R \end{aligned} \tag{4.17}$$

where $U \coloneqq (W_0 - W)/\|W_0 - W\| = (U_1, \ldots, U_d)$ and $W_0 = \mathcal{P}_{\bar{A}}(W)$. Observe that by (4.17),

$$\begin{aligned} R &= \sum_{i=1}^n \sum_{j=1}^d \mathbb{E}\big\{ \xi_{ij}^2 U_j^2 \mathbb{1}(\|\xi_i\| \leqslant 4\gamma) \mathbb{1}(W \in A^{4\gamma + \Delta_1} \setminus A^{4\gamma - \bar{\Delta}_2}) \big\} \\ &\quad + \sum_{i=1}^n \sum_{j \neq j'} \mathbb{E}\big\{ \xi_{ij} \xi_{ij'} U_j U_{j'} \mathbb{1}(\|\xi_i\| \leqslant 4\gamma) \mathbb{1}(W \in A^{4\gamma + \Delta_1} \setminus A^{4\gamma - \bar{\Delta}_2}) \big\} \\ &\coloneqq R_1 + R_2. \end{aligned}$$

For $R_1$, rearranging the summations yields

$$\begin{aligned}
R_1 = {} & \sum_{j=1}^{d} \mathbb{E}\Big\{ \mathbb{1}(W \in A^{4\gamma+\Delta_1} \setminus A^{4\gamma-\bar{\Delta}_2}) U_j^2 \sum_{i=1}^{n} \xi_{ij}^2 \mathbb{1}(\|\xi_i\| \leqslant 4\gamma) \Big\} \\
= {} & \sum_{j=1}^{d} \mathbb{E}\Big\{ \mathbb{1}(W \in A^{4\gamma+\Delta_1} \setminus A^{4\gamma-\bar{\Delta}_2}) U_j^2 \Big( \sum_{i=1}^{n} \Big( \xi_{ij}^2 \mathbb{1}(\|\xi_i\| \leqslant 4\gamma) - \mathbb{E}\xi_{ij}^2 \mathbb{1}(\|\xi_i\| \leqslant 4\gamma) \Big) \Big) \Big\} \\
& + \sum_{j=1}^{d} \Big( \mathbb{E}\{ \mathbb{1}(W \in A^{4\gamma+\Delta_1} \setminus A^{4\gamma-\bar{\Delta}_2}) U_j^2 \} \Big) \Big( \sum_{i=1}^{n} \mathbb{E}\{ \xi_{ij}^2 \mathbb{1}(\|\xi_i\| \leqslant 4\gamma) \} \Big) \\
:= {} & R_{11} + R_{12}.
\end{aligned}$$

By the basic inequality that $ab \leqslant \gamma a^2 + (1/4\gamma) b^2$ for $a, b \geqslant 0$, it follows that with

$$a = U_j^2 \quad \text{and} \quad b = \Big| \sum_{i=1}^{n} \Big( \xi_{ij}^2 \mathbb{1}(\|\xi_i\| \leqslant 4\gamma) - \mathbb{E}\xi_{ij}^2 \mathbb{1}(\|\xi_i\| \leqslant 4\gamma) \Big) \Big|,$$

we have

$$\begin{aligned}
|R_{11}| & \leqslant \sum_{j=1}^{d} \mathbb{E}\Big\{ U_j^2 \Big| \sum_{i=1}^{n} \Big( \xi_{ij}^2 \mathbb{1}(\|\xi_i\| \leqslant 4\gamma) - \mathbb{E}\xi_{ij}^2 \mathbb{1}(\|\xi_i\| \leqslant 4\gamma) \Big) \Big| \Big\} \\
& \leqslant \gamma \sum_{j=1}^{d} \mathbb{E}\{U_j^4\} + \frac{1}{4\gamma} \sum_{j=1}^{d} \mathrm{Var}\Big( \sum_{i=1}^{n} \xi_{ij}^2 \mathbb{1}(\|\xi_i\| \leqslant 4\gamma) \Big) \qquad (4.18) \\
& \leqslant \gamma \sum_{j=1}^{d} \mathbb{E}\{U_j^4\} + \frac{1}{4\gamma} \sum_{j=1}^{d} \sum_{i=1}^{n} \mathbb{E}\{ \xi_{ij}^4 \mathbb{1}(\|\xi_i\| \leqslant 4\gamma) \}.
\end{aligned}$$

As for $R_{12}$, recalling that $\sum_{j=1}^{d} U_i^2 = 1$ and $\sum_{i=1}^{n} \mathbb{E}\{\xi_i \xi_i^{\intercal}\} = I_d$, we have $\sum_{i=1}^{n} \mathbb{E}\{\xi_{ij}^2\} = 1$ for each $1 \leqslant j \leqslant d$, and

$$\begin{aligned}
R_{12} = {} & \sum_{j=1}^{d} \mathbb{E}\Big\{ \mathbb{1}(W \in A^{4\gamma+\Delta_1} \setminus A^{4\gamma-\bar{\Delta}_2}) U_j^2 \Big\} \Big( \sum_{i=1}^{n} \mathbb{E}\{\xi_{ij}^2\} - \sum_{i=1}^{n} \mathbb{E}\{ \xi_{ij}^2 \mathbb{1}(\|\xi_i\| > 4\gamma) \} \Big) \\
= {} & \mathbb{P}(W \in A^{4\gamma+\Delta_1} \setminus A^{4\gamma-\bar{\Delta}_2}) \qquad (4.19) \\
& - \mathbb{E}\Big\{ \mathbb{1}(W \in A^{4\gamma+\Delta_1} \setminus A^{4\gamma-\bar{\Delta}_2}) \sum_{j=1}^{d} U_j^2 \Big( \sum_{i=1}^{n} \mathbb{E}\{ \xi_{ij}^2 \mathbb{1}(\|\xi_i\| > 4\gamma) \} \Big) \Big\}.
\end{aligned}$$

By (4.18) and (4.19), it follows that

$$R_1 \geqslant \mathbb{P}(W \in A^{4\gamma+\Delta_1} \setminus A^{4\gamma-\bar{\Delta}_2}) - \gamma \sum_{j=1}^{d} \mathbb{E}\{U_j^4\} - \frac{1}{4\gamma} \sum_{i=1}^{n} \sum_{j=1}^{d} \mathbb{E}\big\{\xi_{ij}^4 \mathbb{1}(\|\xi_i\| \leqslant 4\gamma)\big\}$$
$$- \mathbb{E}\bigg\{\mathbb{1}(W \in A^{4\gamma+\Delta_1} \setminus A^{4\gamma-\bar{\Delta}_2}) \sum_{j=1}^{d} U_j^2 \bigg(\sum_{i=1}^{n} \mathbb{E}\{\xi_{ij}^2 \mathbb{1}(\|\xi_i\| > 4\gamma)\}\bigg)\bigg\}. \tag{4.20}$$

Similarly, noting that $\sum_{i=1}^{n} \mathbb{E}\{\xi_{ij}\xi_{ij'}\} = 0$ for $j \neq j'$, we have

$$\begin{aligned}
R_2 &\geqslant -\gamma \sum_{j\neq j'} \mathbb{E}\big\{U_j^2 U_{j'}^2\big\} - \frac{1}{4\gamma} \sum_{i=1}^{n} \sum_{1\leqslant j\neq j'\leqslant d} \mathbb{E}\big\{(\xi_{ij}\xi_{ij'})^2 \mathbb{1}(\|\xi_i\| \leqslant 4\gamma)\big\} \\
&\quad - \sum_{i=1}^{n} \sum_{1\leqslant j\neq j'\leqslant d} \mathbb{E}\Big\{\mathbb{1}(W \in A^{4\gamma+\Delta_1} \setminus A^{4\gamma-\bar{\Delta}_2}) U_j U_{j'}\, \mathbb{E}\{\xi_{ij}\xi_{ij'} \mathbb{1}(\|\xi_i\| \leqslant 4\gamma)\}\Big\} \\
&= -\gamma \sum_{j\neq j'} \mathbb{E}\big\{U_j^2 U_{j'}^2\big\} - \frac{1}{4\gamma} \sum_{i=1}^{n} \sum_{1\leqslant j\neq j'\leqslant d} \mathbb{E}\big\{(\xi_{ij}\xi_{ij'})^2 \mathbb{1}(\|\xi_i\| \leqslant 4\gamma)\big\} \\
&\quad - \sum_{i=1}^{n} \sum_{1\leqslant j\neq j'\leqslant d} \mathbb{E}\Big\{\mathbb{1}(W \in A^{4\gamma+\Delta_1} \setminus A^{4\gamma-\bar{\Delta}_2}) U_j U_{j'}\, \mathbb{E}\{\xi_{ij}\xi_{ij'}\}\Big\} \\
&\quad - \sum_{i=1}^{n} \sum_{1\leqslant j\neq j'\leqslant d} \mathbb{E}\Big\{\mathbb{1}(W \in A^{4\gamma+\Delta_1} \setminus A^{4\gamma-\bar{\Delta}_2}) U_j U_{j'}\, \mathbb{E}\{\xi_{ij}\xi_{ij'} \mathbb{1}(\|\xi_i\| > 4\gamma)\}\Big\} \\
&= -\gamma \sum_{j\neq j'} \mathbb{E}\big\{U_j^2 U_{j'}^2\big\} - \frac{1}{4\gamma} \sum_{i=1}^{n} \sum_{1\leqslant j\neq j'\leqslant d} \mathbb{E}\big\{(\xi_{ij}\xi_{ij'})^2 \mathbb{1}(\|\xi_i\| \leqslant 4\gamma)\big\} \\
&\quad - \sum_{i=1}^{n} \sum_{1\leqslant j\neq j'\leqslant d} \mathbb{E}\Big\{\mathbb{1}(W \in A^{4\gamma+\Delta_1} \setminus A^{4\gamma-\bar{\Delta}_2}) U_j U_{j'}\, \mathbb{E}\{\xi_{ij}\xi_{ij'} \mathbb{1}(\|\xi_i\| > 4\gamma)\}\Big\}.
\end{aligned} \tag{4.21}$$

Observe that

$$\sum_{i=1}^{n} \mathbb{E}\{\|\xi_i\|^4 \mathbb{1}(\|\xi_i\| \leqslant 4\gamma)\} \leqslant 4\gamma \sum_{i=1}^{n} \mathbb{E}\|\xi_i\|^3 \leqslant 4\gamma^2 \tag{4.22}$$

and by the Markov inequality,

$$\sum_{i=1}^{n} \mathbb{E}\{\|\xi_i\|^2 \mathbb{1}(\|\xi_i\| > 4\gamma)\} \leqslant \frac{1}{4\gamma} \sum_{i=1}^{n} \mathbb{E}\|\xi_i\|^3 = \frac{1}{4}. \tag{4.23}$$

Recall that $\|U\| = 1$ and $\gamma = \sum_{i=1}^n \mathbb{E}\{\|\xi_i\|^3\}$, and thus (4.20)–(4.23) yield

$$\begin{aligned} R &\geqslant \mathbb{P}(W \in A^{4\gamma+\Delta_1} \setminus A^{4\gamma-\bar{\Delta}_2}) - \gamma\,\mathbb{E}\{\|U\|^4\} - \frac{1}{4\gamma}\sum_{i=1}^n \mathbb{E}\{\|\xi_i\|^4 \mathbb{1}(\|\xi_i\| \leqslant 4\gamma)\} \\ &\quad - \sum_{i=1}^n \mathbb{E}\Big\{\mathbb{1}(W \in A^{4\gamma+\Delta_1} \setminus A^{4\gamma-\bar{\Delta}_2})\,\mathbb{E}\Big\{\Big(\sum_{j=1}^d U_j \xi_{ij}\Big)^2 \mathbb{1}(\|\xi_i\| > 4\gamma)\Big\}\Big\} \\ &\geqslant \mathbb{P}(W \in A^{4\gamma+\Delta_1} \setminus A^{4\gamma-\bar{\Delta}_2}) - \gamma\,\mathbb{E}\{\|U\|^4\} - \frac{1}{4\gamma}\sum_{i=1}^n \mathbb{E}\{\|\xi_i\|^4 \mathbb{1}(\|\xi_i\| \leqslant 4\gamma)\} \\ &\quad - \sum_{i=1}^n \mathbb{E}\Big\{\mathbb{1}(W \in A^{4\gamma+\Delta_1} \setminus A^{4\gamma-\bar{\Delta}_2})\,\mathbb{E}\{\|U\|^2\|\xi_i\|^2 \mathbb{1}(\|\xi_i\| > 4\gamma)\}\Big\} \\ &= \mathbb{P}(W \in A^{4\gamma+\Delta_1} \setminus A^{4\gamma-\bar{\Delta}_2}) - \gamma\,\mathbb{E}\{\|U\|^4\} - \frac{1}{4\gamma}\sum_{i=1}^n \mathbb{E}\{\|\xi_i\|^4 \mathbb{1}(\|\xi_i\| \leqslant 4\gamma)\} \\ &\quad - \sum_{i=1}^n \mathbb{E}\Big\{\mathbb{1}(W \in A^{4\gamma+\Delta_1} \setminus A^{4\gamma-\bar{\Delta}_2})\,\mathbb{E}\{\|\xi_i\|^2 \mathbb{1}(\|\xi_i\| > 4\gamma)\}\Big\} \\ &\geqslant \mathbb{P}(W \in A^{4\gamma+\Delta_1} \setminus A^{4\gamma-\bar{\Delta}_2}) - 2\gamma - \frac{1}{4}\mathbb{P}(W \in A^{4\gamma+\Delta_1} \setminus A^{4\gamma-\bar{\Delta}_2}) \\ &= \frac{3}{4}\mathbb{P}(W \in A^{4\gamma+\Delta_1} \setminus A^{4\gamma-\bar{\Delta}_2}) - 2\gamma. \end{aligned} \tag{4.24}$$

By (4.17) and (4.24), we have

$$H_1 \geqslant \frac{9}{16}\mathbb{P}(W \in A^{4\gamma+\Delta_1} \setminus A^{4\gamma-\bar{\Delta}_2}) - \frac{3}{2}\gamma. \tag{4.25}$$

On the other hand, note that $\mathbb{E}\|W\|^2 = d$ and by Lemma 4.2, $\|g_{\Delta_1,\bar{\Delta}_2}(W)\| \leqslant (\Delta_1 + \Delta_2 + 8\gamma)$. Thus,

$$|\,\mathbb{E}\langle W, g_{\Delta_1,\bar{\Delta}_2}(W)\rangle| \leqslant \mathbb{E}\|W\|(\Delta_1 + \Delta_2) + 8d^{1/2}\gamma. \tag{4.26}$$

Combining inequalities (4.12), (4.16), (4.25) and (4.26) yields

$$\begin{aligned} &\mathbb{P}(W \in A^{4\gamma+\Delta_1} \setminus A^{4\gamma-\bar{\Delta}_2}) \\ &\quad \leqslant 2\Big|\mathbb{E}\langle W, g_{\Delta_1,\Delta_2}(W)\rangle\Big| + 2H_2 + 3\gamma \\ &\quad \leqslant 2\,\mathbb{E}\big\{\|W\|(\Delta_1 + \Delta_2)\big\} + 16d^{1/2}\gamma + 3\gamma + 2\sum_{i=1}^n\sum_{j=1}^2 \mathbb{E}\|\xi_i\||\Delta_j - \Delta_j^{(i)}| \\ &\quad \leqslant 2\,\mathbb{E}\|W\|(\Delta_1 + \Delta_2) + 19d^{1/2}\gamma + 2\sum_{i=1}^n\sum_{j=1}^2 \mathbb{E}\|\xi_i\|\big|\Delta_j - \Delta_j^{(i)}\big|. \end{aligned}$$

This proves (4.1). ◻

## 5. Proofs of other results

In this section, we give the proofs of the theorems in Section 3.

### 5.1. Proof of Theorem 3.1

Note that $\hat{\theta}_n$ minimizes $\mathbb{M}_n(\theta)$, and $m_\theta$ is smooth for $\theta$. By the Taylor expansion, it follows that

$$0 = \frac{1}{n}\sum_{i=1}^{n} \dot{m}_{\hat{\theta}_n}(X_i) = \frac{1}{n}\sum_{i=1}^{n} \dot{m}_{\theta^*}(X_i) + \frac{1}{n}\sum_{i=1}^{n}\int_0^1 \big(\ddot{m}_{\theta_t}(X_i)\big)\big(\hat{\theta}_n - \theta^*\big)dt,$$

where $\theta_t = \theta^* + t(\hat{\theta}_n - \theta^*)$. Therefore, recalling that $V = \mathbb{E}\{\ddot{m}_{\theta^*}(X)\}$ and $\xi_i = \dot{m}_{\theta^*}(X_i)$,

$$\begin{aligned} V\big(\hat{\theta}_n - \theta^*\big) = {} & -\frac{1}{n}\sum_{i=1}^{n}\xi_i - \Big(\frac{1}{n}\sum_{i=1}^{n}\ddot{m}_{\theta^*}(X_i) - V\Big)\big(\hat{\theta}_n - \theta^*\big) \\ & - \frac{1}{n}\sum_{i=1}^{n}\int_0^1 \big(\ddot{m}_{\theta_t}(X_i) - \ddot{m}_{\theta^*}(X_i)\big)\big(\hat{\theta}_n - \theta^*\big)dt, \end{aligned}$$

Let

$$W = -\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Sigma^{-1/2}\xi_i,$$

and

$$\begin{aligned} D = {} & -\sqrt{n}\Sigma^{-1/2}\Big(\frac{1}{n}\sum_{i=1}^{n}\ddot{m}_{\theta^*}(X_i) - V\Big)\big(\hat{\theta}_n - \theta^*\big) \\ & - \sqrt{n}\Sigma^{-1/2}\Big(\frac{1}{n}\sum_{i=1}^{n}\int_0^1\big(\ddot{m}_{\theta(t)}(X_i) - \ddot{m}_{\theta^*}(X_i)\big)dt\Big)\big(\hat{\theta}_n - \theta^*\big). \end{aligned}$$

Then, we have

$$T \coloneqq \sqrt{n}\Sigma^{-1/2}V(\hat{\theta}_n - \theta^*) = W + D. \tag{5.1}$$

By (M1) and (M2), we have

$$\|D\| \leqslant n^{1/2}{\lambda_1}^{-1/2}\big(H_1\|\hat{\theta}_n - \theta^*\| + H_2\|\hat{\theta}_n - \theta^*\|^2\big), \tag{5.2}$$

where

$$H_1 = \Big\|\frac{1}{n}\sum_{i=1}^{n}\big(\ddot{m}_{\theta^*}(X_i) - \mathbb{E}\{\ddot{m}_{\theta^*}(X_i)\}\big)\Big\|, \qquad H_2 = \frac{1}{n}\sum_{i=1}^{n}\|m_2(X_i)\|.$$

Let $\Delta = n^{1/2}\lambda_1{}^{-1/2}\big(H_1\|\hat{\theta}_n-\theta^*\|+H_2\|\hat{\theta}_n-\theta^*\|^2\big)$, and it follows that $\|T-W\| \leqslant \Delta$. Let $(X_1', \dots, X_n')$ be an independent copy of $(X_1, \dots, X_n)$, and define

$$X_j^{(i)} = \begin{cases} X_j, & j \neq i; \\ X_i', & j = i. \end{cases}$$

Moreover, let

$$H_1^{(i)} = \Big\|\frac{1}{n}\sum_{j=1}^{n}\big(\ddot{m}_{\theta^*}(X_j^{(i)}) - \mathbb{E}\{\ddot{m}_{\theta^*}(X_j)\}\big)\Big\|, \quad H_2^{(i)} = \frac{1}{n}\sum_{j=1}^{n}\|m_2(X_j^{(i)})\|$$

$$\hat{\theta}_n^{(i)} = \operatorname*{arg\,min}_{\theta\in\Theta} \frac{1}{n}\sum_{j=1}^{n} m_\theta(X_j^{(i)}), \quad \Delta^{(i)} = n^{1/2}\lambda_1{}^{-1/2}\big(H_1^{(i)}\|\hat{\theta}_n^{(i)} - \theta^*\| + H_2^{(i)}\|\hat{\theta}_n^{(i)} - \theta^*\|^2\big).$$

Then, $\Delta^{(i)}$ is independent of $X_i$ and $\xi_i$. Theorem 3.1 follows directly from Theorem 2.1 and the following proposition.

**Proposition 5.1.** *Assume that conditions (M1) and (M2) are satisfied. Then, we have*

$$\sum_{i=1}^{n}\mathbb{E}\|\Sigma^{-1/2}\xi_i\|^3 \leqslant Cd^{3/2}n, \tag{5.3}$$

$$\mathbb{E}\{\|W\|\Delta\} \leqslant Cd^{13/8}n^{-1/2}, \tag{5.4}$$

$$\sum_{i=1}^{n}\mathbb{E}\|\Sigma^{-1/2}\xi_i\||\Delta - \Delta^{(i)}| \leqslant Cd^{9/4}, \tag{5.5}$$

*where $C > 0$ is a constant depending only on $\lambda_1, \lambda_2, c_1, c_2, c_3, c_4$ and $\mu$.*

In order to prove Proposition 5.1, we first need to prove three useful lemmas, whose proofs are postponed to Appendix A.

The following lemma provides an upper bound for the $p$-th moment of $\|\hat{\theta}_n - \theta^*\|$, whose proof (see Appendix A.2) is based on the ideas in Van der Vaart and Wellner [32, Theorem 3.2.5].

**Lemma 5.2.** *Assume that there exist $p \geqslant 2$ and $a_6 > 0$ such that (3.6) and (3.7) are satisfied with $\|m_1(X)\|_{p+1} \leqslant a_6$. Then, we have*

$$\mathbb{E}\|\hat{\theta}_n - \theta^*\|^p \leqslant C\mu^{-p-1}a_6^{p+1}d^{\frac{p+1}{2}}n^{-\frac{p}{2}},$$

*where $C > 0$ is a constant depending only on $p$.*

The next lemma gives upper bounds for the fourth moments of $H_1$ and $H_2$. The proof can be found in Appendix A.3, where we use the Rosenthal-type inequality for random matrices (see, e.g, Chen et al. [12, Theorem A.1]).

**Lemma 5.3.** *Under the assumption (M1), we have*

$$\mathbb{E}\big\{H_1^4\big\} \leqslant Cc_3^4n^{-2}, \tag{5.6}$$

$$\mathbb{E}\big\{H_2^4\big\} \leqslant Cc_2^4, \tag{5.7}$$

*where $C > 0$ is an absolute constant.*

The following lemma gives an upper bound of $\mathbb{E}\|\hat{\theta}_n - \hat{\theta}_n^{(i)}\|^2$, whose proof is given in Appendix A.4.

**Lemma 5.4.** *Under the assumptions (M1) and (M2), we have*

$$\mathbb{E}\|\hat{\theta}_n - \hat{\theta}_n^{(i)}\|^2 \leqslant C\lambda_2^{-2}n^{-2}\Big(c_4^2 d + \mu^{-9/4}c_1^{9/4}c_3^2 d^{9/8} + \mu^{-9/2}c_1^{9/2}c_2^2 d^{9/4}\Big) \tag{5.8}$$

*where $C > 0$ is an absolute constant.*

With Lemmas 5.2–5.4, we are ready to give the proof of Proposition 5.1.

*Proof of Proposition 5.1.* In this proof, we denote by $C$ a general positive constant depending on $\lambda_1$, $\lambda_2$, $c_1$,
$c_2, c_3, c_4$ and $\mu$, and the value of $C$ might be different in different places. The bound (5.3) follows from (M2). For (5.4), since

$$\Delta = n^{1/2}\lambda_1{}^{-1/2}(H_1\|\hat{\theta}_n - \theta^*\| + H_2\|\hat{\theta}_n - \theta^*\|^2),$$

it follows from Lemmas 5.2 and 5.3 and the Hölder inequality that

$$\begin{aligned}\mathbb{E}\big\{\Delta\|W\|\big\} &\leqslant Cn^{1/2}\lambda_1{}^{-1/2}\Big(\mathbb{E}\big\{H_1\|W\|\|\hat{\theta}_n - \theta^*\| + H_2\|W\|\|\hat{\theta}_n - \theta^*\|^2\big\}\Big)\\ &\leqslant Cn^{1/2}\lambda_1{}^{-1/2}\Big(\|H_1\|_2\|W\|_4\|\hat{\theta}_n - \theta^*\|_2 + \|H_2\|_2\|W\|_4\|\hat{\theta}_n - \theta^*\|_4\Big)\\ &\leqslant Cd^{\frac{13}{8}}n^{-1/2},\end{aligned} \tag{5.9}$$

and

$$\mathbb{E}\Delta \leqslant Cd^{9/8}n^{-1/2}. \tag{5.10}$$

Therefore (5.9) and (5.10) yield (5.4).

For (5.5), we have

$$\begin{aligned}|\Delta - \Delta^{(i)}| \leqslant n^{1/2}\lambda_1{}^{-1/2}\Big(&H_1\|\hat{\theta}_n - \hat{\theta}_n^{(i)}\| + |H_1 - H_1^{(i)}|\|\hat{\theta}_n - \theta_n^{(i)}\| + |H_2 - H_2^{(i)}|\|\hat{\theta}_n^{(i)} - \theta^*\|^2\\ &+ H_2(\|\hat{\theta}_n - \theta^*\| + \|\hat{\theta}_n^{(i)} - \theta^*\|)\|\hat{\theta}_n - \hat{\theta}_n^{(i)}\|\Big).\end{aligned}$$

By the assumption (M1),

$$\|H_1 - H_1^{(i)}\|_4 \leqslant \frac{1}{n}\|\ddot{m}_{\theta^*}(X_i) - \ddot{m}_{\theta^*}(X_i^{(i)})\|_4 \leqslant Cn^{-1}$$

and

$$\|H_2 - H_2^{(i)}\|_4 \leqslant \frac{1}{n}\Big(\|m_2(X_i)\|_4 + \|m_2(X_i^{(i)})\|_4\Big) \leqslant Cn^{-1}.$$

By Lemmas 5.2–5.4 and the Hölder inequality, we have

$$\mathbb{E}\{\|\Sigma^{-1/2}\xi_i\||\Delta - \Delta^{(i)}|\} \leqslant Cd^{13/8}n^{-1},$$

which yields (5.5). □

### 5.2. Proof of Theorem 3.2

Let $\delta_n = (D_\Theta + 1) d n^{-(p-2)/(2p-2)}$, where $D_\Theta$ is the diameter of the parameter space $\Theta$. As $p \geqslant 3$, it follows that $\delta_n \geqslant n^{-1/2}$. In this subsection, we denote by $C > 0$ a constant depending only on $p, c_1, c_2, c_3, \lambda_1, \lambda_2$ and $\mu$, which might be different in different places. The main idea is to rewrite $\sqrt{n}\Sigma^{-1/2}\dot{\Psi}_0(\hat{\theta}_n - \theta^*)$ as a summation of a linear statistic plus an error term, and then apply Corollary 2.2 to prove (5.16). To this end, by (3.11),

$$
\begin{aligned}
&\sqrt{n}\big(\Psi(\hat{\theta}_n) - \Psi(\theta^*)\big)\\
&= \sqrt{n}\Big(\Psi(\hat{\theta}_n) - \Psi_n(\hat{\theta}_n)\Big)\\
&= -\sqrt{n}\big(\Psi_n(\theta^*) - \Psi(\theta^*)\big) - \Big(\sqrt{n}(\Psi_n(\hat{\theta}_n) - \Psi(\hat{\theta}_n)) - \sqrt{n}(\Psi_n(\theta^*) - \Psi(\theta^*))\Big).
\end{aligned} \tag{5.11}
$$

By (5.11), we obtain

$$
T \coloneqq \sqrt{n}\Sigma^{-1/2}\dot{\Psi}_0(\hat{\theta}_n - \theta^*) = W + D,
$$

where

$$
\begin{aligned}
W &= -\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Sigma^{-1/2}\xi_i,\\
D &= -\Sigma^{-1/2}\Big(\sqrt{n}(\Psi_n(\hat{\theta}_n) - \Psi(\hat{\theta}_n)) - \sqrt{n}(\Psi_n(\theta^*) - \Psi(\theta^*))\Big)\\
&\quad - \sqrt{n}\Sigma^{-1/2}\Big(\Psi(\hat{\theta}_n) - \Psi(\theta^*) - \dot{\Psi}_0(\hat{\theta}_n - \theta^*)\Big).
\end{aligned}
$$

By (3.13) and (3.16),

$$
\|D\|\mathbb{1}(\|\hat{\theta}_n - \theta^*\| \leqslant \delta_n) \leqslant \Delta_1 + \Delta_2,
$$

where

$$
\begin{aligned}
\Delta_1 &= \lambda_2^{-1/2}\sqrt{n}\sup_{\theta:\|\theta-\theta^*\|\leqslant\delta_n}\Big\|\big(\Psi_n(\theta) - \Psi(\theta)\big) - \big(\Psi_n(\theta^*) - \Psi(\theta^*)\big)\Big\|\\
\Delta_2 &= c_1\lambda_2^{-1/2}\sqrt{n}\|\hat{\theta}_n - \theta^*\|^2\mathbb{1}\big(\|\hat{\theta}_n - \theta^*\| \leqslant \delta_n\big).
\end{aligned}
$$

Now we construct random variables $\Delta_1^{(i)}$ and $\Delta_2^{(i)}$ that are independent of $\xi_i$. Let $(X_1'$,
$\dots, X_n')$ be an independent copy of $(X_1, \dots, X_n)$ and let

$$
\Psi_n^{(i)}(\theta) = \Psi_n(\theta) - \frac{1}{n}\big(h_\theta(X_i) - h_\theta(X_i')\big), \quad 1 \leqslant i \leqslant n.
$$

Let $\hat{\theta}_n^{(i)}$ be the minimizer of $\Psi_n^{(i)}$, and let

$$
\begin{aligned}
\Delta_1^{(i)} &= \lambda_2^{-1/2}\sqrt{n}\sup_{\theta:\|\theta-\theta^*\|\leqslant\delta_n}\Big\|\big(\Psi_n^{(i)}(\theta) - \Psi(\theta)\big) - \big(\Psi_n^{(i)}(\theta^*) - \Psi(\theta^*)\big)\Big\|\\
\Delta_2^{(i)} &= c_1\lambda_2^{-1/2}\sqrt{n}\|\hat{\theta}_n^{(i)} - \theta^*\|\mathbb{1}\big(\|\hat{\theta}_n^{(i)} - \theta^*\| \leqslant \delta_n\big).
\end{aligned}
$$

To apply Corollary 2.2, we need to develop the following proposition.

**Proposition 5.5.** *Let $B_\delta = \{\theta \in \Theta : \|\theta - \theta^*\| \leqslant \delta\}$. Under the conditions (B1)–(B3),*

$$\mathbb{P}\big(\hat{\theta}_n \in B^c_{\delta_n}\big) \leqslant C(D_\Theta + 1)^p d^p \delta_n^{-p} n^{-p/2}, \tag{5.12}$$

$$\sum_{i=1}^n \mathbb{E}\|\Sigma^{-1/2}\xi_i\|^3 \leqslant Cd^{3/2}n, \tag{5.13}$$

$$\mathbb{E}\big\{\|W\|(\Delta_1 + \Delta_2)\big\} \leqslant Cd^{3/2}\delta_n + C(D_\Theta + 1)^2 d^{5/2} n^{-1/2}, \tag{5.14}$$

$$\begin{aligned}\sum_{i=1}^n \sum_{j=1}^2 \mathbb{E}\big\{\|\xi_i\|\big|\Delta_j - \Delta_j^{(i)}\big|\big\} \leqslant C\Big(&(D_\Theta + 1)^p d^{p+1/2} \delta_n^{-p+2} n^{-\frac{p-3}{2}} \\ &+ (D_\Theta + 1)d^2 + (D_\Theta + 1)d^{5/2} n^{1/2}\delta_n\Big).\end{aligned} \tag{5.15}$$

By Corollary 2.2 with $O = \{\|\hat{\theta}_n - \theta^*\| \leqslant \delta_n\}$ and Proposition 5.5,

$$\begin{aligned}&\sup_{A\in\mathcal{A}}\big|\mathbb{P}\big(\sqrt{n}\Sigma^{-1/2}\dot{\Psi}_0(\hat{\theta}_n - \theta^*) \in A\big) - \mathbb{P}\big(Z \in A\big)\big| \\ &\quad\leqslant Cd^{1/2}n^{-3/2}\sum_{i=1}^n \mathbb{E}\|\Sigma^{-1/2}\xi_i\|^3 + C\,\mathbb{E}\{\|W\|(\Delta_1 + \Delta_2)\} \\ &\qquad + Cn^{-1/2}\sum_{i=1}^n \mathbb{E}\big\{\|\xi_i\|(\Delta_1 + \Delta_2 - \Delta_1^{(i)} - \Delta_2^{(i)})\big\} + \mathbb{P}\big(\|\hat{\theta}_n - \theta^*\| > \delta_n\big) \\ &\quad\leqslant Cn^{-1/2}\big(d^2 + (D_\Theta + 1)^2 d^{5/2}\big) + C(D_\Theta + 1)^2 d^{5/2}\delta_n \\ &\qquad + C(D_\Theta + 1)^p d^{p+1/2}\delta_n^{-p+2} n^{-(p-2)/2} + C(D_\Theta + 1)^p d^p \delta_n^{-p} n^{-p/2}.\end{aligned} \tag{5.16}$$

Recall that $p \geqslant 3$, and then $\delta_n^2 n \geqslant 1$. Therefore,

$$\begin{aligned}\text{RHS of (5.16)} &\leqslant Cn^{-1/2}(D_\Theta + 1)^2 d^{5/2} + C(D_\Theta + 1)^2 d^{5/2}\delta_n \\ &\quad + C(D_\Theta + 1)^p d^{p+1/2}\delta_n^{-p+2} n^{-(p-2)/2} \\ &\leqslant C(D_\Theta + 1)^2\Big(d^{5/2}n^{-1/2} + d^{7/2}n^{-\frac{p-2}{2(p-1)}}\Big) \\ &\leqslant C(D_\Theta + 1)^2 d^{7/2} n^{-1/2+\varepsilon_p},\end{aligned}$$

where $\varepsilon_p = 1/(2p - 2)$. This proves Theorem 3.2.

It suffices to prove Proposition 5.5, and we need to apply some preliminary lemmas, whose proofs are put in Appendix A.

**Lemma 5.6.** *Let $B_\delta$ be as in Proposition 5.5. Under the assumptions (B1)–(B3), we have*

$$\|W\|_2 \leqslant \lambda_2^{-1/2} c_3 d^{1/2}, \quad \|W\|_3 \leqslant C\lambda_2^{-1/2} c_3 d^{1/2}, \tag{5.17}$$

$$\mathbb{E}\Delta_1^2 \leqslant C\lambda_2^{-1} c_2^2 d^2 \delta_n^2, \tag{5.18}$$

*where $C > 0$ is an absolute constant. Moreover,*

$$\mathbb{E}\big\{\|\hat{\theta}_n - \theta^*\|^p\big\} \leqslant C(D_\Theta + 1)^p d^p n^{-p/2}, \tag{5.19}$$

$$\mathbb{E}\Big\{\|\hat{\theta}_n - \hat{\theta}_n^{(i)}\|^p \mathbb{1}\big(\hat{\theta}_n \in B_\delta, \hat{\theta}_n^{(i)} \in B_\delta\big)\Big\} \leqslant C\big(d^{p/2} n^{-p} + d^p n^{-p/2}\delta^p\big), \tag{5.20}$$

*where $C > 0$ is a constant depending only on $c_2, c_3, \mu$ and $p$.*

**Lemma 5.7.** *We have*

$$\mathbb{E}\big\{\|\xi_i\|\|\hat{\theta}_n - \theta^*\|^2 \mathbb{1}\big(\hat{\theta}_n \in B_\delta, \hat{\theta}_n^{(i)} \in B_\delta^c\big)\big\} \leqslant C(D_\Theta + 1)^p d^{p+1/2}\delta^{-p+2} n^{-p/2}, \tag{5.21}$$

$$\mathbb{E}\big\{\|\xi_i\|\|\hat{\theta}_n^{(i)} - \theta^*\|^2 \mathbb{1}\big(\hat{\theta}_n \in B_\delta^c, \hat{\theta}_n^{(i)} \in B_\delta\big)\big\} \leqslant C(D_\Theta + 1)^p d^{p+1/2}\delta^{-p+2} n^{-p/2}, \tag{5.22}$$

*and*

$$\begin{aligned} &\mathbb{E}\Big\{\|\xi_i\|(\|\hat{\theta}_n - \theta^*\| + \|\hat{\theta}_n^{(i)} - \theta^*\|)\|\hat{\theta}_n - \hat{\theta}_n^{(i)}\| \mathbb{1}\big(\hat{\theta}_n \in B_\delta, \hat{\theta}_n^{(i)} \in B_\delta\big)\Big\} \\ &\quad \leqslant C(D_\Theta + 1)\Big(d^2 n^{-3/2} + d^{5/2} n^{-1}\delta\Big). \end{aligned} \tag{5.23}$$

Now we are ready to give the proof of Proposition 5.5.

*Proof of Proposition 5.5.* The inequality (5.12) follows directly from (5.19) and the Chebyshev inequality.

For (5.13), by (B3), we have

$$\mathbb{E}\|\Sigma^{-1/2}\xi_i\|^3 \leqslant \lambda_2^{-1/2} c_3 d^{3/2},$$

and thus (5.13) holds.

For (5.14), it suffices to give the bounds for the moments of $\|W\|, \Delta_1$ and $\Delta_2$. By (5.17) and (5.18) and the Cauchy inequality, it follows that

$$\mathbb{E}\{\|W\|\Delta_1\} \leqslant C d^{3/2}\delta_n. \tag{5.24}$$

Recall that $p \geqslant 3$, $\Delta_2 \leqslant c_1 \lambda_2^{-1/2} n^{1/2}\|\hat{\theta}_n - \theta_0\|^2$ and by (5.19),

$$\mathbb{E}\|\hat{\theta}_n - \theta^*\|^3 \leqslant C(D_\Theta + 1)^3 d^3 n^{-3/2}, \tag{5.25}$$

and then by (5.17), (5.25) and the Hölder inequality,

$$\mathbb{E}\{\|W\|\Delta_2\} \leqslant C n^{1/2}\ \{\|W\|_3 \|\hat{\theta}_n - \theta^*\|_3^2\} \leqslant C(D_\Theta + 1)^2 d^{5/2} n^{-1/2}. \tag{5.26}$$

Combining (5.24) and (5.26) yields (5.14).

It suffices to prove (5.15). By the definition of $\Delta_1$ and $\Delta_1^{(i)}$,

$$\big|\Delta_1 - \Delta_1^{(i)}\big| \leqslant 2\lambda_2^{-1/2} n^{-1/2} \sup_{\theta:\|\theta-\theta^*\|\leqslant\delta_n} \big\|h_\theta(X_i) - h_\theta(X_i')\big\| \leqslant 2d^{1/2}\lambda_2^{-1/2} n^{-1/2}\delta_n |h_0(X)|,$$

and by (3.15),

$$\mathbb{E}\big|\Delta_1 - \Delta_1^{(i)}\big|^2 \leqslant Cdn^{-1}\delta_n^2.$$

Thus, by (3.17) and the Cauchy inequality,

$$\sum_{i=1}^{n} \mathbb{E}\Big\{\|\xi_i\|\big(|\Delta_1 - \Delta_1^{(i)}|\big)\Big\} \leqslant Cdn^{1/2}\delta_n. \tag{5.27}$$

For $\Delta_2 - \Delta_2^{(i)}$, we have

$$\begin{aligned}
&\big|\Delta_2 - \Delta_2^{(i)}\big| \\
&\leqslant c_2\lambda_2^{-1/2}\sqrt{n}\Big(\|\hat{\theta}_n - \theta^*\|^2 \mathbb{1}\big(\|\hat{\theta}_n - \theta^*\| \leqslant \delta_n, \|\hat{\theta}_n^{(i)} - \theta^*\| > \delta_n\big) \\
&\qquad + \|\hat{\theta}_n^{(i)} - \theta^*\|^2 \mathbb{1}\big(\|\hat{\theta}_n - \theta^*\| > \delta_n, \|\hat{\theta}_n^{(i)} - \theta^*\| \leqslant \delta_n\big) \\
&\qquad + (\|\hat{\theta}_n - \theta^*\| + \|\hat{\theta}_n^{(i)} - \theta^*\|)\|\hat{\theta}_n - \hat{\theta}_n^{(i)}\| \mathbb{1}\big(\|\hat{\theta}_n - \theta^*\| \leqslant \delta_n, \|\hat{\theta}_n^{(i)} - \theta^*\| \leqslant \delta_n\big)\Big).
\end{aligned}$$

By Lemmas 5.6 and 5.7, it follows that

$$\begin{aligned}
\sum_{i=1}^{n} \mathbb{E}\Big\{\|\xi_i\|\big(\big|\Delta_2 - \Delta_2^{(i)}\big|\big)\Big\} &\leqslant C(D_\Theta + 1)^p d^{p+1/2}\delta^{-p+2} n^{-(p-3)/2} \\
&\quad + C(D_\Theta + 1)d^2 + C(D_\Theta + 1)d^{5/2}n^{1/2}\delta_n.
\end{aligned} \tag{5.28}$$

Together with (5.27) and (5.28), we obtain (5.15). □

### 5.3. *Proof of Theorem 3.3*

Theorem 3.3 follows from the proof of Theorem 3.2 and the following proposition.

**Proposition 5.8.** *Under the conditions (B1), (B4) and (B5), we have*

$$\begin{aligned}
\mathbb{P}\big(\hat{\theta}_n \in B_\delta^c\big) &\leqslant C\exp\Big(-\frac{C'\sqrt{n}\delta}{(D_\Theta + 1)d^{3/2}}\Big), \\
\sum_{i=1}^{n} \mathbb{E}\|\Sigma^{-1/2}\xi_i\|^3 &\leqslant Cd^{3/2}n \\
\mathbb{E}\big\{\|W\|(\Delta_1 + \Delta_2)\big\} &\leqslant Cd^{3/2}\delta_n + C(D_\Theta + 1)^2 d^{5/2} n^{-1/2}, \\
\sum_{i=1}^{n} \mathbb{E}\big\{\|\xi_i\|\big|\Delta_1 + \Delta_2 - \Delta_1^{(i)} - \Delta_2^{(i)}\big|\big\} &\leqslant C(D_\Theta + 1)^2 d^{5/2}\exp\Big(-\frac{C'\sqrt{n}\delta_n}{4(D_\Theta + 1)d^{3/2}}\Big) \\
&\quad + C(D_\Theta + 1)^2 d^2 + C(D_\Theta + 1)d^{5/2}n^{1/2}\delta_n,
\end{aligned}$$

*where $C' > 0$ is a constant depending only on $c_4, c_5$ and $\mu$ and $C > 0$ is a constant depending only on $c_1, c_4, c_5, \mu, \lambda_1$ and $\lambda_2$.*

Similar to the proof of Theorem 3.2, and by Proposition 5.8, we have

$$\sup_{A\in\mathcal{A}}\big|\mathbb{P}\big(\sqrt{n}\Sigma^{-1/2}\dot{\Psi}_0(\hat{\theta}_n-\theta^*)\in A\big)-\mathbb{P}\big(Z\in A\big)\big|$$
$$\leqslant C(D_\Theta+1)^2d^{5/2}n^{-1/2}+C(D_\Theta+1)d^{5/2}\delta_n+C(D_\Theta+1)^2d^{5/2}\exp\Big(-\frac{C'\sqrt{n}\delta_n}{4(D_\Theta+1)d^{3/2}}\Big).$$

Choosing $\delta_n=(C')^{-1}(D_\Theta+1)d^{3/2}n^{-1/2}\log n$, we completes the proof of Theorem 3.3. It suffices to prove Proposition 5.8.

The following lemma is a modification of Lemmas 5.6 and 5.7, whose proof is given in Appendix A.

**Lemma 5.9.** *Let $B_\delta$ be as in Proposition 5.5. Under the assumptions (B1), (B4) and (B5), we have (5.19) and (5.20) hold for each $p\geqslant 1$ with a positive constant $C$ depending on $c_1,c_4,c_5,\mu,\lambda_1,\lambda_2$ and $p$. Moreover, we have there exists a constant $C'>0$ depending only on $c_4,c_5$ and $\mu$ such that*

$$\mathbb{P}\big(\|\hat{\theta}_n-\theta^*\|>t\big)\leqslant 2\exp\Big(-\frac{C'n^{1/2}t}{(D_\Theta+1)d^{3/2}}\Big),\qquad \textit{for } t>0, \tag{5.29}$$

*and*

$$\mathbb{E}\Big\{\|\xi_i\|\|\hat{\theta}_n-\theta^*\|^2\mathbb{1}\big(\hat{\theta}_n\in B_\delta,\hat{\theta}_n^{(i)}\in B_\delta^c\big)\Big\}\leqslant C(D_\Theta+1)^2d^{5/2}n^{-1}\exp\Big(-\frac{C'n^{1/2}\delta}{4(D_\Theta+1)d^{3/2}}\Big), \tag{5.30}$$

$$\mathbb{E}\Big\{\|\xi_i\|\|\hat{\theta}_n^{(i)}-\theta^*\|^2\mathbb{1}\big(\hat{\theta}_n\in B_\delta^c,\hat{\theta}_n^{(i)}\in B_\delta\big)\Big\}\leqslant C(D_\Theta+1)^2d^{5/2}n^{-1}\exp\Big(-\frac{C'n^{1/2}\delta}{4(D_\Theta+1)d^{3/2}}\Big), \tag{5.31}$$

*where $C>0$ depending only on $c_1,c_4,c_5,\mu,\lambda_1$ and $\lambda_2$.*

*Proof of Proposition 5.8.* The first inequality follows directly from (5.29), and the last three inequalities follow from (5.30) and (5.31) and from the proof of Proposition 5.5. □

### *5.4. Proof of Theorem 3.4*

Without loss of generality, we assume that $n\geqslant 4\{(2L\ell_0)^\alpha+1\}$; otherwise, the bound (3.26) is trivial.

In this subsection, we denote by $C,C_1,C_2,\dots$ a sequence of general positive constants depending only on $\ell_0,\lambda_1,\lambda_2,c_1,c_2,\alpha,\beta,L$ and $\mu$ and independent of $\tau$ and $\tau_0$. Let $L_1:=\max\{c_2,2L/\beta\}$ and $L_2:=c_1+L$. We introduce the following family of functions: Let $\varphi_\beta\colon\mathbb{R}_+\setminus\{0\}\mapsto\mathbb{R}$ be given by

$$\varphi_\beta(t)=\begin{cases}\frac{t^\beta-1}{\beta}, & \text{if } \beta\neq 0,\\ \log t, & \text{if } \beta=0.\end{cases} \tag{5.32}$$

*Proof of Theorem 3.4.* Note that $\theta^*$ is the minimum point of $f$ and by the differentiability and convexity of $f$, we have $\nabla f(\theta^*) = 0$. By (3.25),

$$\nabla f(\theta) = \nabla^2 f(\theta^*)(\theta - \theta^*) + H(\theta), \tag{5.33}$$

where

$$\begin{aligned} H(\theta) &= \nabla f(\theta) - \nabla^2 f(\theta^*)(\theta - \theta^*) \\ &= \nabla f(\theta) - \nabla f(\theta^*) - \nabla^2 f(\theta^*)(\theta - \theta^*) \\ &= \int_0^1 \{\nabla^2 f(\theta^* + t(\theta - \theta^*)) - \nabla^2 f(\theta^*)\}(\theta - \theta^*) dt. \end{aligned}$$

By (C2) and (C3), it follows that

$$\|H(\theta)\| \mathbb{1}(\|\theta - \theta^*\| \leqslant \beta) \leqslant c_2 \|\theta - \theta^*\|^2,$$

and

$$\|H(\theta)\| \mathbb{1}(\|\theta - \theta^*\| > \beta) \leqslant 2L\|\theta - \theta^*\| \mathbb{1}(\|\theta - \theta^*\| > \beta) \leqslant \frac{2L}{\beta}\|\theta - \theta^*\|^2.$$

Hence, with $L_1 := \max\{c_2, 2L/\beta\}$, we have

$$\|H(\theta)\| \leqslant L_1 \|\theta - \theta^*\|^2. \tag{5.34}$$

Recall that $G := \nabla^2 f(\theta^*)$, and it follows from (3.22) and (5.33) that for any $n \geqslant 1$,

$$\begin{aligned} \theta_n &= \theta_{n-1} - \ell_n\big(\nabla f(\theta_{n-1}) + \zeta_n\big) \\ &= \theta_{n-1} - \ell_n\big(G(\theta_{n-1} - \theta^*) + \xi_n + \eta_n + H(\theta_{n-1})\big). \end{aligned} \tag{5.35}$$

By definition, $(\bar\theta_n - \theta^*) = n^{-1}\sum_{i=0}^{n-1}(\theta_i - \theta^*)$. Solving the recursive system (5.35) yields

$$\sqrt{n}(\bar\theta_n - \theta^*) = \frac{1}{\sqrt{n}\ell_0} Q_0(\theta_0 - \theta^*) + \frac{1}{\sqrt{n}}\sum_{i=1}^{n-1} Q_i\big(\xi_i + \eta_i + H(\theta_{i-1})\big),$$

where $Q_i = \ell_i \sum_{j=i}^{n-1}\prod_{k=i+1}^{j}(I - \ell_k G)$.

Recall that $\Sigma_n := n^{-1}\sum_{i=1}^{n-1} Q_i \Sigma_i Q_i^\intercal$. Let

$$T_n = n^{-1/2}\Sigma_n^{-1/2}\sum_{i=0}^{n-1}(\theta_n - \theta^*), \quad \zeta_i = \frac{1}{\sqrt{n}}\Sigma_n^{-1/2} Q_i \xi_i, \quad W_n = \sum_{i=1}^{n-1}\zeta_i$$

and

$$\begin{aligned} D_n &= \frac{1}{\sqrt{n}\ell_0}\Sigma_n^{-1/2} Q_0(\theta_0 - \theta^*) + \frac{1}{\sqrt{n}}\Sigma_n^{-1/2}\sum_{i=1}^{n-1} Q_i\eta_i + \frac{1}{\sqrt{n}}\Sigma_n^{-1/2}\sum_{i=1}^{n-1} Q_i H(\theta_{i-1}) \\ &:= D_{1,n} + D_{2,n} + D_{3,n}. \end{aligned}$$

It is easy to show that

$$\mathbb{E}W_n = 0, \quad \mathrm{Var}(W_n) = I_d.$$

and

$$T_n = W_n + D_n.$$

Also,

$$\begin{aligned}\|D_n\| &\leqslant n^{-1/2}\ell_0^{-1}\|\Sigma_n^{-1/2}\| \cdot \|Q_0\| \cdot \|\theta_0 - \theta^*\| \\ &\quad + n^{-1/2}\|\Sigma_n^{-1/2}\|\left\|\sum_{i=1}^{n-1} Q_i\eta_i\right\| + n^{-1/2}\|\Sigma_n^{-1/2}\|\sum_{i=1}^{n-1}\|Q_iH(\theta_{i-1})\|^2\end{aligned}$$

The following proposition provides the bounds of $Q_j$ and $\Sigma_n^{-1}$.

**Proposition 5.10.** *Suppose that $n \geqslant 4\{(2L\ell_0)^\alpha + 1\}$. If $\ell_i = \ell_0 i^{-\alpha}$ with $1/2 < \alpha \leqslant 1$, then there exists a sequence $(p_i)_{i\geqslant 1}$, and two positive constants $C_1$ and $C_2$ depending on $\ell_0, \lambda_1, \lambda_2, c_1, c_2, \alpha, \beta, L$ and $\mu$ such that for each $0 \leqslant i \leqslant n-1$,*

$$\begin{aligned}\Sigma_n^{-1} &\preccurlyeq C_1 I_d, \\ -p_i I_d &\preccurlyeq Q_i \preccurlyeq p_i I_d,\end{aligned} \tag{5.36}$$

*where*

$$p_i \leqslant \begin{cases} C_2, & \text{if } (\alpha = 1, \ell_0\mu > 1) \text{ or } (\alpha \in (1/2, 1)); \\ C_2 \log n, & \text{if } (\alpha = 1, \ell_0\mu = 1). \end{cases}$$

Let $(\xi_1', \ldots, \xi_n')$ be an independent copy of $(\xi_1, \ldots, \xi_n)$. For each $1 \leqslant i \leqslant n-1$, we now construct $D_{2,n}^{(i)}$ and $D_{3,n}^{(i)}$ which are independent of $\xi_i$. Firstly, for each $i$, we construct $\theta_1^{(i)}, \ldots, \theta_n^{(i)}$ as follows:

(a) If $j < i$, let $\theta_j^{(i)} = \theta_j$.
(b) If $j = i$, let $\theta_j^{(i)} = \theta_{j-1}^{(i)} - \ell_j(\nabla f(\theta_{j-1}^{(i)}) + \xi_j' + \eta_j^{(i)})$, where $\eta_j^{(i)} = g(\theta_{j-1}^{(i)}, \xi_i')$.
(c) If $j > i$, let $\theta_j^{(i)} = \theta_{j-1}^{(i)} - \ell_j(\nabla f(\theta_{j-1}^{(i)}) + \xi_j + \eta_j^{(i)})$, where $\eta_j^{(i)} = g(\theta_{j-1}^{(i)}, \xi_j)$.

Secondly, let

$$\begin{aligned}D_{2,n}^{(i)} &= n^{-1/2}\Sigma_n^{-1/2}\sum_{j=1}^{n-1} Q_i\eta_j^{(i)}, \\ D_{3,n}^{(i)} &= n^{-1/2}\Sigma_n^{-1/2}\sum_{j=1}^{n-1} Q_iH(\theta_{j-1}^{(i)}).\end{aligned}$$

Then, we have for each $1 \leqslant i \leqslant n-1$, $D_{2,n}^{(i)}$ and $D_{3,n}^{(i)}$ is independent of $\xi_i$. Let

$$\Delta = \Delta_1 + \Delta_2 + \Delta_3,$$

where $\Delta_1 = \|D_{1,n}\|, \Delta_2 = \|D_{2,n}\|$ and $\Delta_3 = C_1 L_1 n^{-1/2} \sum_{i=1}^{n-1} p_i \big\|\theta_{i-1} - \theta^*\big\|^2$, where $C_1$ is given as in (5.38). By (5.34), it follows that $\|D_{3,n}\| \leqslant \Delta_3$. Also, for each $1 \leqslant i \leqslant n-1$, define

$$\begin{aligned}
\Delta_1^{(i)} &= \|D_{1,n}\|,\\
\Delta_2^{(i)} &= \|D_{2,n}^{(i)}\|,\\
\Delta_3^{(i)} &= C_1 L_1 n^{-1/2} \sum_{j=1}^{n-1} p_j \big\|\theta_{j-1}^{(i)} - \theta^*\big\|^2.
\end{aligned}$$

Clearly, $\Delta_1^{(i)}$, $\Delta_2^{(i)}$ and $\Delta_3^{(i)}$ are independent of $\xi_i$ for each $1 \leqslant i \leqslant n-1$. The following proposition provides the bounds of the moments for $\Delta_j$ and $\Delta_j - \Delta_j^{(i)}$, $j = 1, 2, 3$.

**Proposition 5.11.** *We have $\Delta_1$ is independent of $(\xi_1, \ldots, \xi_n)$ and*

$$\mathbb{E}\{\Delta_1 \|W\|\} \leqslant C(\tau^2 + \tau_0^2) n^{-1/2}.$$

1. *For $\alpha \in (1/2, 1)$,*

$$\begin{aligned}
\mathbb{E}\{\Delta_2 \|W\|\} &\leqslant C d^{1/2} (\tau + \tau_0) n^{-\alpha/2},\\
\mathbb{E}\{\Delta_3 \|W\|\} &\leqslant C d^{1/2} (\tau^2 + \tau_0^2) n^{-\alpha + 1/2}.
\end{aligned}$$

*and*

$$\begin{aligned}
\sum_{i=1}^{n-1} \mathbb{E}\{|\Delta_2 - \Delta_2^{(i)}| \|\zeta_i\|\} &\leqslant C(\tau^2 + \tau_0^2) n^{-\alpha + 1/2},\\
\sum_{i=1}^{n-1} \mathbb{E}\{|\Delta_3 - \Delta_3^{(i)}| \|\zeta_i\|\} &\leqslant C(\tau^3 + \tau_0^3) n^{-\alpha/2}.
\end{aligned}$$

2. *For $\alpha = 1$,*

$$\begin{aligned}
\mathbb{E}\{\Delta_2 \|W\|\} &\leqslant \begin{cases} C d^{1/2} (\tau + \tau_0) n^{-1/2} (\log n)^{1/2}, & \ell_0 \mu > 1;\\ C d^{1/2} (\tau + \tau_0) n^{-1/2} (\log n)^2, & \ell_0 \mu = 1. \end{cases}\\
\mathbb{E}\{\Delta_3 \|W\|\} &\leqslant \begin{cases} C d^{1/2} (\tau^2 + \tau_0^2) n^{-1/2} (\log n), & \ell_0 \mu > 1;\\ C d^{1/2} (\tau^2 + \tau_0^2) n^{-1/2} (\log n)^{5/2}, & \ell_0 \mu = 1, \end{cases}\\
\sum_{i=1}^{n-1} \mathbb{E}\big\{\big|\Delta_2 - \Delta_2^{(i)}\big| \|\zeta_i\|\big\} &\leqslant C(\tau^2 + \tau_0^2) \times \begin{cases} n^{-1/2}, & \ell_0 \mu > 1;\\ n^{-1/2} (\log n)^{5/2}, & \ell_0 \mu = 1. \end{cases}
\end{aligned}$$

*and*

$$\sum_{i=1}^{n-1} \mathbb{E}\big\{\big|\Delta_3 - \Delta_3^{(i)}\big| \|\zeta_i\|\big\} \leqslant C(\tau^3 + \tau_0^3) \times \begin{cases} n^{-1/2}, & \mu \ell_0 > 1;\\ n^{-1/2} (\log n)^{5/2}, & \mu \ell_0 = 1. \end{cases}$$

We apply Theorem 2.1 to prove the Berry–Esseen bound for $\sqrt{n}\Sigma_n^{-1/2}(\bar{\theta}_n - \theta^*)$.

**(1). For** $1/2 < \alpha < 1$**.** Firstly, by Proposition 5.10 and (C1), we have

$$\sum_{i=1}^{n-1} \mathbb{E}\|\zeta_i\|^3 \leqslant Cn^{-3/2} \sum_{i=1}^{n-1} \mathbb{E}\|\xi_i\|^3 \leqslant Cn^{-1/2}\tau^3. \tag{5.37}$$

By Proposition 5.11, we have

$$\begin{aligned} &\mathbb{E}\{\|W\|\Delta\} \leqslant C(d^{3/2} + \tau^3 + \tau_0^3) n^{-\alpha+1/2}, \\ &\sum_{i=1}^{n-1} \mathbb{E}\{\|\zeta_i\| \cdot |\Delta - \Delta^{(i)}|\} \leqslant C(d^{3/2} + \tau^3 + \tau_0^3) n^{-\alpha/2}. \end{aligned} \tag{5.38}$$

Substituting (5.37) and (5.38) to Theorem 2.1 yields (3.26).

**(2). For** $\alpha = 1$**.** By the definition of $\zeta_i$ and by (5.36),

$$\begin{aligned} \sum_{i=1}^{n-1} \mathbb{E}\|\zeta_i\|^3 &\leqslant Cn^{-3/2} \sum_{i=1}^{n-1} p_i^3\, \mathbb{E}\|\xi_i\|^3 \\ &\leqslant \begin{cases} C\tau^3 n^{-1/2}, & \text{if } \ell_0\mu > 1, \\ C\tau^3 n^{-1/2}(\log n)^3, & \text{if } \ell_0\mu = 1. \end{cases} \end{aligned} \tag{5.39}$$

By Proposition 5.11, we have

$$\begin{aligned} \mathbb{E}\{\Delta\|W\|\} \leqslant{}& C(d + \tau^2 + \tau_0^2) n^{-1/2} \\ &+ C(d^{3/2} + \tau^3 + \tau_0^3) \times \begin{cases} n^{-1/2}(\log n), & \ell_0\mu > 1; \\ n^{-1/2}(\log n)^{5/2}, & \ell_0\mu = 1. \end{cases} \end{aligned} \tag{5.40}$$

Define $\Delta^{(i)} = \Delta_1 + \Delta_2^{(i)} + \Delta_3^{(i)}$, then we have $\Delta^{(i)}$ is independent of $\zeta_i$. Also, $\Delta - \Delta^{(i)} = \Delta_2 - \Delta_2^{(i)} + \Delta_3 - \Delta_3^{(i)}$. By Proposition 5.11, we have

$$\begin{aligned} &\sum_{i=1}^{n-1} \mathbb{E}\{|\Delta - \Delta^{(i)}|\|\zeta_i\|\} \\ &\leqslant C(d^{3/2} + \tau^3 + \tau_0^3) \times \begin{cases} n^{-1/2}, & \ell_0\mu > 1; \\ n^{-1/2}(\log n)^{5/2}, & \ell_0\mu = 1. \end{cases} \end{aligned} \tag{5.41}$$

Then the bound (3.27) follows from Theorem 2.1 and (5.39)–(5.41). □

Now we are ready to give the proofs of Propositions 5.10 and 5.11. We first prove Proposition 5.11. To prove Proposition 5.11, we need to apply some preliminary lemmas, which provide the bounds for $\mathbb{E}\|\theta_n - \theta^*\|^2, \mathbb{E}\|\theta_n - \theta_n^{(i)}\|^2$ and $\mathbb{E}\|\theta_n - \theta_0\|^4$. The proofs of the lemmas can be found in Appendix A.

**Lemma 5.12.** *For $\alpha \in (1/2, 1)$, we have*

$$\mathbb{E}\|\theta_n - \theta^*\|^2 \leqslant Cn^{-\alpha}(\tau^2 + \tau_0^2), \quad \text{for } n \geqslant 1. \tag{5.42}$$

*For $\alpha = 1$, we have*

$$\mathbb{E}\|\theta_n - \theta^*\|^2 \leqslant \begin{cases} Cn^{-1}(\tau^2 + \tau_0^2), & \mu\ell_0 > 1, \\ Cn^{-1}(\log n)(\tau^2 + \tau_0^2), & \mu\ell_0 = 1, \\ Cn^{-\mu\ell_0}(\tau^2 + \tau_0^2), & 0 < \mu\ell_0 < 1. \end{cases} \tag{5.43}$$

**Lemma 5.13.** *For $\alpha \in (1/2, 1)$, we have*

$$\begin{aligned} \mathbb{E}\|\theta_j - \theta_j^{(i)}\|^2 &\leqslant C(\tau^2 + \tau_0^2) i^{-2\alpha} \exp\Big\{-\mu\big(\varphi_{1-\alpha}(j) - \varphi_{1-\alpha}(i)\big)\Big\} & (5.44)\\ &\leqslant C(\tau^2 + \tau_0^2) j^{-2\alpha}. & (5.45) \end{aligned}$$

*For $\alpha = 1$, we have*

$$\mathbb{E}\|\theta_j - \theta_j^{(i)}\|^2 \leqslant C(\tau^2 + \tau_0^2) i^{-2} \Big(\frac{i}{j}\Big)^{2\mu\ell_0}. \tag{5.46}$$

*Here, $\varphi_{1-\alpha}$ is as given in (5.32).*

**Lemma 5.14.** *For $\alpha \in (0, 1)$,*

$$\mathbb{E}\|\theta_j - \theta^*\|^4 \leqslant Cj^{-2\alpha}(\tau^4 + \tau_0^4). \tag{5.47}$$

*For $\alpha = 1$,*

$$\mathbb{E}\|\theta_j - \theta^*\|^4 \leqslant \begin{cases} Cj^{-2}, & \ell_0\mu > 1, \\ Cj^{-2}\log j, & \ell_0\mu = 1. \end{cases} \tag{5.48}$$

*Proof of Proposition 5.11.* Recall that we assume that $n \geqslant 4\{(2L\ell_0)^\alpha + 1\}$. Now we consider the following two cases.

**1. If $\alpha \in (1/2, 1)$.** First, by Proposition 5.10, $\Sigma_n^{-1} \preccurlyeq C_1 I_d, Q_j \preccurlyeq C_2 I_d$ for each $0 \leqslant j \leqslant n-1$ and $n \geqslant 4\{(2L\ell_0)^\alpha + 1\}$. For $\Delta_1$, by (C0), we have

$$\mathbb{E}\Delta_1^2 \leqslant Cn^{-1}\, \mathbb{E}\|\theta_0 - \theta^*\|^2 \leqslant C\tau_0^2 n^{-1}.$$

By the Cauchy inequality and noting that $\mathbb{E}\{WW^\intercal\} = I_d$, we have

$$\mathbb{E}\{\Delta_1\|W\|\} \leqslant Cd^{1/2}\tau_0 n^{-1/2}.$$

Recall that by (C1), $(\eta_j)_{j\geqslant 1}$ is a martingale difference sequence and $\|\eta_j\| \leqslant$

$c_1\|\theta_{j-1} - \theta^*\|$, and then by (5.36) and (5.42), if $\alpha \in (1/2, 1)$,

$$\begin{aligned}\mathbb{E}\Delta_2^2 &\leqslant \lambda_2^{-1}\,\mathbb{E}\Big\|\frac{1}{\sqrt{n}}\sum_{i=1}^{n-1}\Sigma_n^{-1/2}Q_i\eta_i\Big\|^2 \leqslant Cn^{-1}\sum_{i=1}^{n-1}\mathbb{E}\|\eta_i\|^2\\ &\leqslant Cn^{-1}\sum_{i=1}^{n-1}\mathbb{E}\|\theta_{i-1} - \theta^*\|^2\\ &\leqslant Cn^{-1}(\tau^2 + \tau_0^2)\sum_{i=1}^{n-1} i^{-\alpha}\\ &\leqslant Cn^{-\alpha}(\tau^2 + \tau_0^2).\end{aligned}$$

Recall that $\mathbb{E}WW^\intercal = I_d$ and thus $\mathbb{E}\|W\|^2 \leqslant Cd$, then by the Cauchy inequality again,

$$\mathbb{E}\{\Delta_2\|W\|\} \leqslant Cd^{1/2}(\tau + \tau_0)n^{-\alpha/2}.$$

For $\Delta_3$, by Proposition 5.10 and Lemma 5.14,

$$\begin{aligned}\mathbb{E}\{\Delta_3\|W\|\} &\leqslant Cn^{-1/2}\sum_{i=1}^{n-1}\mathbb{E}\{\|\theta_{i-1} - \theta^*\|^2\|W\|\}\\ &\leqslant Cd^{1/2}n^{-1/2}\sum_{i=1}^{n-1}\big(\mathbb{E}\|\theta_{i-1} - \theta^*\|^4\big)^{1/2}\\ &\leqslant Cd^{1/2}n^{-1/2}\bigg(\sum_{i=1}^{n-2}\big(\mathbb{E}\|\theta_i - \theta^*\|^4\big)^{1/2} + (\mathbb{E}\|\theta_0 - \theta^*\|^4)^{1/2}\bigg)\\ &\leqslant Cd^{1/2}n^{-1/2}(\tau^2 + \tau_0^2)\sum_{i=1}^{n-1} i^{-\alpha}\\ &\leqslant Cd^{1/2}n^{-\alpha+1/2}(\tau^2 + \tau_0^2).\end{aligned} \tag{5.49}$$

Now we move to give the bounds of $\mathbb{E}\{|\Delta_2 - \Delta_2^{(i)}|\|\zeta_i\|\}$ and $\mathbb{E}\{|\Delta_2 - \Delta_2^{(i)}|\|\zeta_i\|\}$. For $\|\zeta_i\|$, by (C1) and Proposition 5.10, we have

$$\mathbb{E}\|\zeta_i\|^4 = n^{-2}\,\mathbb{E}\|\Sigma_n^{-1/2}Q_i\xi_i\|^4 \leqslant Cn^{-2}\tau^4. \tag{5.50}$$

For $\Delta_2 - \Delta_2^{(i)}$,

$$|\Delta_2 - \Delta_2^{(i)}| \leqslant n^{-1/2}\Big\|\sum_{j=1}^{n-1}\Sigma_n^{-1}Q_j(\eta_j - \eta_j^{(i)})\Big\|, \tag{5.51}$$

and

$$\eta_j - \eta_j^{(i)} = \begin{cases}0, & j < i;\\ g(\theta_{j-1}, \xi_j) - g(\theta_{j-1}, \xi_j'), & j = i;\\ g(\theta_{j-1}, \xi_j) - g(\theta_{j-1}^{(i)}, \xi_j), & j > i.\end{cases}$$

By the construction of $\eta_j^{(i)}$ and by (3.23), for each $j$, $\eta_j =_d \eta_j^{(i)}$ and $\|\eta_j\| \leqslant c_1\|\theta_{j-1} - \theta^*\|$; and for $j > i$, $\|\eta_j - \eta_j^{(i)}\| \leqslant c_1\|\theta_{j-1} - \theta_{j-1}^{(i)}\|$. Set

$$\mathcal{F}_j^{(i)} = \begin{cases} \mathcal{F}_j, & j < i; \\ \mathcal{F}_j \bigvee \sigma(\xi_i'), & j \geqslant i. \end{cases}$$

Then $(\eta_j - \eta_j^{(i)})_{j\geqslant 1}$ is a martingale difference sequence with respect to $\mathcal{F}_j^{(i)}$. Hence, by (5.36) and (5.51),

$$\begin{aligned} \mathbb{E}\big|\Delta_2 - \Delta_2^{(i)}\big|^2 &\leqslant 2n^{-1}\,\mathbb{E}\|\eta_i - \eta_i^{(i)}\|^2 + 2\,\mathbb{E}\Big\|\frac{1}{\sqrt{n}}\sum_{j=i+1}^{n-1} Q_j(\eta_j - \eta_j^{(i)})\Big\|^2 \\ &\leqslant 4n^{-1}\,\mathbb{E}\|\eta_i\|^2 + 2n^{-1}\sum_{j=i+1}^{n-1} p_i^2\,\mathbb{E}\|\eta_j - \eta_j^{(i)}\|^2 \\ &\leqslant 4c_1^2 n^{-1}\,\mathbb{E}\|\theta_{i-1} - \theta^*\|^2 + 2c_1^2 n^{-1}\sum_{j=i+1}^{n-1} p_j^2\,\mathbb{E}\|\theta_{j-1} - \theta_{j-1}^{(i)}\|^2 \\ &\leqslant C(\tau^2 + \tau_0^2)n^{-1}i^{-\alpha} + C(\tau^2 + \tau_0^2)n^{-1}\big(\varphi_{1-2\alpha}(n-1) - \varphi_{1-2\alpha}(i)\big) \\ &\leqslant C(\tau^2 + \tau_0^2)n^{-1}i^{-2\alpha+1}, \end{aligned} \tag{5.52}$$

where we used (5.42) and (5.45) in the last second line. By (5.50) and (5.52) and the Cauchy inequality, we have

$$\sum_{i=1}^{n-1}\mathbb{E}\big\{|\Delta_2 - \Delta_2^{(i)}|\|\zeta_i\|\big\} \leqslant C(\tau^2 + \tau_0^2)n^{-\alpha+1/2}.$$

For $\Delta_3 - \Delta_3^{(i)}$, by the Hölder inequality, and noting that $\theta_j \overset{d}{=} \theta_j^{(i)}$,

$$\begin{aligned} &\sum_{i=1}^{n-1}\sum_{j=1}^{n}\mathbb{E}\Big\{\|\xi_i\|\Big(\big|\|\theta_{j-1} - \theta^*\|^2 - \|\hat{\theta}_{j-1}^{(i)} - \theta^*\|^2\big|\Big)\Big\} \\ &\quad\leqslant \sum_{i=1}^{n-1}\sum_{j=0}^{n-1}\mathbb{E}\big\{\|\xi_i\|\big(\|\theta_j - \theta^*\|\cdot\|\theta_j - \hat{\theta}_j^{(i)}\|\big)\big\} \\ &\qquad + \sum_{i=1}^{n-1}\sum_{j=0}^{n-1}\mathbb{E}\big\{\|\xi_i\|\big(\|\theta_j^{(i)} - \theta^*\|\cdot\|\theta_j - \hat{\theta}_j^{(i)}\|\big)\big\} \\ &\quad\leqslant 2\sum_{i=1}^{n-1}\sum_{j=0}^{n-1}\big(\mathbb{E}\|\xi_i\|^4\big)^{1/4}\big(\mathbb{E}\|\theta_j - \theta^*\|^4\big)^{1/4}\big(\mathbb{E}\|\theta_j - \hat{\theta}_j^{(i)}\|^2\big)^{1/2} \\ &\quad\leqslant C(\tau^3 + \tau_0^3)\sum_{i=1}^{n-1}\sum_{j=i}^{n} j^{-\alpha/2}i^{-\alpha}\exp\big\{-C\big(j^{1-\alpha} - i^{1-\alpha}\big)\big\}, \end{aligned}$$

where we used (5.44), (5.47) and (5.50) in the last line. Observe that

$$\sum_{j=i}^{n} j^{-\alpha/2} \exp\bigl\{-C\bigl(j^{1-\alpha} - i^{1-\alpha}\bigr)\bigr\} \leqslant Cn^{\alpha/2},$$

and it follows that

$$\begin{aligned}\sum_{i=1}^{n-1} \mathbb{E}\{|\Delta_3 - \Delta_3^{(i)}|\|\zeta_i\|\} &\leqslant Cn^{-1} \sum_{i=1}^{n-1}\sum_{j=0}^{n-1} \mathbb{E}\bigl\{\|\xi_i\|\bigl|\|\theta_j - \theta^*\|^2 - \|\hat{\theta}_j^{(i)} - \theta^*\|^2\bigr|\bigr\} \\ &\leqslant C(\tau^3 + \tau_0^3)n^{-\alpha/2}.\end{aligned} \tag{5.53}$$

**2. If** $\alpha = 1$**.** Since the bound of $\mathbb{E}\{\Delta_1\|W\|\}$ does not depend on $\alpha$, it suffices to give the bounds of $\mathbb{E}\{\Delta_j\|W\|\}$ and $\sum_i \mathbb{E}\{|\Delta_j - \Delta_j^{(i)}|\|\zeta_i\|\}$ for $j = 2, 3$.

By Proposition 5.10, we have $\Sigma_n^{-1/2} \preccurlyeq C_2 I_d$, and for $0 \leqslant i \leqslant n-1$,

$$p_i \leqslant \begin{cases} C, & \ell_0\mu > 1; \\ C(\log n), & \ell_0\mu = 1. \end{cases}$$

For $\Delta_2$, noting that $\|\eta_i\| \leqslant c_1\|\theta_i - \theta^*\|$, by Proposition 5.10 with $\alpha = 1$, and by (5.43), we have

$$\begin{aligned}\mathbb{E}\Delta_2^2 &\leqslant \lambda_2^{-1}\,\mathbb{E}\Bigl\|\frac{1}{\sqrt{n}}\sum_{i=1}^{n-1}\Sigma_n^{-1/2}Q_i\eta_i\Bigr\|^2 \\ &\leqslant Cn^{-1}\sum_{i=1}^{n-1} p_i^2\,\mathbb{E}\|\eta_i\|^2 \\ &\leqslant \begin{cases} C(\tau^2+\tau_0^2)n^{-1}\displaystyle\sum_{i=1}^{n-1} i^{-1}, & \text{if } \ell_0\mu > 1, \\ C(\tau^2+\tau_0^2)n^{-1}\displaystyle\sum_{i=1}^{n-1}(\log n)^2 i^{-1}\log i, & \text{if } \ell_0\mu = 1 \end{cases} \\ &\leqslant \begin{cases} C(\tau^2+\tau_0^2)n^{-1}\log n, & \text{if } \ell_0\mu > 1, \\ C(\tau^2+\tau_0^2)n^{-1}(\log n)^4, & \text{if } \ell_0\mu = 1. \end{cases}\end{aligned}$$

Recalling that $\mathbb{E}WW^\intercal = I_d$, we obtain

$$\mathbb{E}\{\Delta_2\|W\|\} \leqslant \begin{cases} Cd^{1/2}(\tau+\tau_0)n^{-1/2}(\log n)^{1/2}, & \ell_0\mu > 1; \\ Cd^{1/2}(\tau+\tau_0)n^{-1/2}(\log n)^2, & \ell_0\mu = 1, \end{cases}$$

Similar to (5.49), and by Proposition 5.10 and Lemma 5.14, we have

$$\begin{aligned}
&\mathbb{E}\{\Delta_3\|W\|\} \\
&\leqslant Cn^{-1/2}\sum_{i=1}^{n-1} p_i\,\mathbb{E}\{\|\theta_{i-1}-\theta^*\|^2\|W\|\} \\
&\leqslant Cd^{1/2}n^{-1/2}\sum_{i=1}^{n-1} p_i\Big(\mathbb{E}\|\theta_{i-1}-\theta^*\|^4\Big)^{1/2} \\
&\leqslant \begin{cases} Cd^{1/2}n^{-1/2}(\tau^2+\tau_0^2)\sum_{i=1}^{n-1} i^{-1}, & \ell_0\mu>1; \\ Cd^{1/2}n^{-1/2}\log n(\tau^2+\tau_0^2)\sum_{i=1}^{n-1} i^{-1}(\log i)^{1/2}, & \ell_0\mu=1, \end{cases} \\
&\leqslant \begin{cases} Cd^{1/2}(\tau^2+\tau_0^2)n^{-1/2}(\log n), & \ell_0\mu>1; \\ Cd^{1/2}(\tau^2+\tau_0^2)n^{-1/2}(\log n)^{5/2}, & \ell_0\mu=1. \end{cases}
\end{aligned}$$

Similar to (5.52), and note that (5.36), we have

$$\mathbb{E}\big|\Delta_2-\Delta_2^{(i)}\big|^2 \leqslant C(\tau^2+\tau_0^2)\times\begin{cases} n^{-1}i^{-1}, & \ell_0\mu>1; \\ n^{-1}(\log n)^2 i^{-1}\log i, & \ell_0\mu=1. \end{cases}$$

By (C1) and Proposition 5.10,

$$\mathbb{E}\|\zeta_i\|^2 \leqslant Cn^{-1}p_i^2\,\mathbb{E}\|\xi_i\|^2 \leqslant \begin{cases} Cn^{-1}\tau^2, & \ell_0\mu>1, \\ Cn^{-1}\tau^2(\log n)^2, & \ell_0\mu=1. \end{cases}$$

By the Cauchy inequality,

$$\begin{aligned}
&\sum_{i=1}^{n-1}\mathbb{E}\big\{\big|\Delta_2-\Delta_2^{(i)}\big|\|\zeta_i\|\big\} \\
&\leqslant C(\tau^2+\tau_0^2)n^{-1}\times\begin{cases} \displaystyle\sum_{i=1}^{n-1} i^{-1/2}, & \ell_0\mu>1, \\ \displaystyle\sum_{i=1}^{n-1} i^{-1/2}(\log i)^{1/2}(\log n)^2, & \ell_0\mu=1, \end{cases} \\
&\leqslant C(\tau^2+\tau_0^2)\times\begin{cases} n^{-1/2}, & \ell_0\mu>1; \\ n^{-1/2}(\log n)^{5/2}, & \ell_0\mu=1. \end{cases}
\end{aligned}$$

Similar to (5.53), and by (5.36), we have

$$\begin{aligned}
\sum_{i=1}^{n-1}\mathbb{E}\big\{\big|\Delta_3-\Delta_3^{(i)}\big|\|\zeta_i\|\big\} &\leqslant Cn^{-1/2}\sum_{i=1}^{n-1}\sum_{j=i}^{n} p_i^2\,\mathbb{E}\big\{\|\xi_i\|\big|\|\theta_j-\theta^*\|^2-\|\hat\theta_j^{(i)}-\theta^*\|^2\big|\big\} \\
&\leqslant C(\tau^3+\tau_0^3)\times\begin{cases} n^{-1/2}, & \mu\ell_0>1; \\ n^{-1/2}(\log n)^{9/4}, & \mu\ell_0=1. \end{cases}
\end{aligned}$$

This completes the proof. □

*Proof of Proposition 5.10.* Note that $\ell_k = \ell_0 k^{-\alpha}$ is decreasing in $k$, and recall that $G = \nabla^2 f(\theta^*)$. By (3.24), we have $\mu I_d \preccurlyeq G \preccurlyeq LI_d$ and $L^{-1} I_d \preccurlyeq G^{-1} \preccurlyeq \mu^{-1} I_d$. Let $i_0 = \lfloor (2L\ell_0)^\alpha + 1 \rfloor$. For $i > i_0$, we have $\ell_0 i^{-\alpha}\mu \leqslant \ell_0 i^{-\alpha} L \leqslant 1/2$. Then,

$$\begin{aligned} Q_i &\succcurlyeq \ell_i \sum_{j=i}^{n-1} \prod_{k=i+1}^{j} (I - \ell_i G) \\ &= \ell_i \big\{ I + (I - \ell_i G) + \cdots + (I - \ell_i G)^{n-i-1} \big\} \\ &= G^{-1} - (I - \ell_i G)^{n-i} G^{-1} \\ &\succcurlyeq L^{-1} \big\{ 1 - \big(1 - \ell_i \mu\big)^{n-i} \big\} I_d \succcurlyeq 0, \end{aligned} \tag{5.54}$$

and for $i_0 + 1 \leqslant i \leqslant n/2$,

$$\begin{aligned} 1 - \big(1 - \ell_i\mu\big)^{n-i} &\geqslant 1 - \big(1 - \ell_i\mu\big)^{n/2} \geqslant 1 - \exp\Big\{ -1/2\mu n \ell_i \Big\} \\ &\geqslant 1 - \exp\Big\{ -\frac{\ell_0}{2}\mu \Big\} \coloneqq c_G, \end{aligned} \tag{5.55}$$

where we used the fact that $n\ell_i \geqslant \ell_0$ in the last inequality.

Recall that for each $1 \leqslant i \leqslant n-1$, $\lambda_{\min}(\Sigma_i) \geqslant \lambda_1$, for any $i_0 + 1 \leqslant i \leqslant n/2$, by (C1), (C2), (5.54) and (5.55),

$$Q_i \Sigma_i Q_i \succcurlyeq c_G^2 L^{-2} \lambda_1 I_d.$$

By the assumption that $n \geqslant 4\{(2L\ell_0)^\alpha + 1\}$, it follows that $n$ is large enough such that $i_0 \leqslant n/4$. Therefore,

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n-1} Q_i \Sigma_i Q_i &= \frac{1}{n}\sum_{i=1}^{i_0} Q_i \Sigma_i T_i + \frac{1}{n}\sum_{i=i_0+1}^{n} Q_i \Sigma_i Q_i \\ &\succcurlyeq \frac{1}{n}\sum_{i=i_0+1}^{n} Q_i \Sigma_i Q_i \\ &\succcurlyeq \frac{c_G^2 \lambda_1}{4L^2} I_d, \end{aligned}$$

because $Q_i \Sigma_i Q_i \succcurlyeq 0$ for each $0 \leqslant i \leqslant i_0$. Therefore,

$$\Sigma_n^{-1} = \bigg( \frac{1}{n}\sum_{i=1}^{n-1} Q_i \Sigma_i Q_i \bigg)^{-1} \preccurlyeq \frac{4L^2}{c_G^2 \lambda_1} I_d.$$

This proves the first inequality of (5.36).

Now we move to prove the second inequality of (5.36). The following proof uses the idea of Polyak and Juditsky [24, pp. 845–846]. Write

$$V_i^m = \prod_{k=i}^{m-1} \big(I - \ell_k G\big), \quad U_i^m = (V_i^m)^\intercal G^{-1} V_i^m,$$

and it follows that

$$U_i^{m+1} = U_i^m - 2\ell_m (V_i^m)^\intercal (V_i^m) + \ell_m^2 U_i^m.$$

Recall that by (C2), $\mu I_d \preccurlyeq G \preccurlyeq L I_d$, and then

$$\begin{aligned} U_i^m &\preccurlyeq \mu^{-1} (V_i^m)^\intercal (V_i^m) \\ U_i^m &\succcurlyeq L^{-1} (V_i^m)^\intercal (V_i^m) \end{aligned} \tag{5.56}$$

and

$$U_i^{m+1} \preccurlyeq \big(1 - 2\ell_m \mu + \ell_m^2\big) U_i^m.$$

Therefore, for $j \geqslant i$,

$$U_i^j \preccurlyeq \exp\biggl\{-2\mu \sum_{k=i}^{j-1} \ell_k\biggr\} \exp\biggl\{\sum_{k=i}^{j-1} \ell_k^2\biggr\} U_i^i \preccurlyeq CL \exp\biggl\{-2\mu \sum_{k=i}^{j-1} \ell_k\biggr\} I_d. \tag{5.57}$$

If $\alpha \in (1/2, 1)$,

$$\begin{aligned} U_i^j &\preccurlyeq C \exp\Bigl\{-\mu\bigl(\varphi_{1-\alpha}(j-1) - \varphi_{1-\alpha}(i-1)\bigr)\Bigr\} I_d \\ &= C \exp\Bigl\{-\frac{\mu}{(1-\alpha)}\bigl((j-1)^{1-\alpha} - (i-1)^{1-\alpha}\bigr)\Bigr\} I_d. \end{aligned}$$

By (5.56), we have

$$V_i^j \preccurlyeq L^{1/2} \bigl(U_i^j\bigr)^{1/2} \preccurlyeq C \exp\Bigl\{-\frac{\mu}{2(1-\alpha)}\bigl((j-1)^{1-\alpha} - (i-1)^{1-\alpha}\bigr)\Bigr\} I_d. \tag{5.58}$$

For $\alpha \in (0, 1)$, by a simple calculation,

$$\sum_{j=i}^{n} \exp\Bigl\{-\frac{\mu}{2(1-\alpha)}\bigl(j^{1-\alpha} - i^{1-\alpha}\bigr)\Bigr\} \leqslant C i^{\alpha},$$

and we have

$$\begin{aligned} Q_i &\preccurlyeq \ell_i \sum_{j=i}^{n-1} V_{i+1}^{j+1} \\ &\preccurlyeq C\ell_i \sum_{j=i}^{n-1} \exp\Bigl\{-\frac{\mu}{2(1-\alpha)}\bigl(j^{1-\alpha} - i^{1-\alpha}\bigr)\Bigr\} I_d \\ &\preccurlyeq C\ell_i i^{-\alpha} I_d \\ &\preccurlyeq C I_d. \end{aligned} \tag{5.59}$$

Similarly, $Q_i \succcurlyeq -C I_d$. This proves the second inequality of (5.36) for $\alpha \in (1/2, 1)$.

For $\alpha = 1$, by (5.57), and note that $\sum_{k=1}^{\infty} \ell_k^2 \leqslant 2\ell_0^2$, we have

$$U_i^j \preccurlyeq C \exp\left\{-2\mu\big(\log(j-1) - \log(i-1)\big)\right\} I_d.$$

Similar to (5.58) and (5.59), if $1 \leqslant i \leqslant n-1$, we have

$$Q_i \preccurlyeq C(i+1)^{\ell_0\mu - 1}\{\varphi_{1-\ell_0\mu}(n-1) - \varphi_{1-\ell_0\mu}(i+1)\} I_d \preccurlyeq \begin{cases} C I_d, & \ell_0\mu > 1, \\ C(\log n) I_d, & \ell_0\mu = 1. \end{cases}$$

If $i = 0$, the result (5.36) follows from the observation that $Q_0 = \ell_0 I_d + \ell_0(1 - \ell_1 G)Q_1$. This completes the proof of the upper bound. The lower bound can be shown similarly. □

## Appendix A: Proofs of some lemmas in Section 5

In the appendix, we give the proofs of some lemmas in Section 5.

### *A.1. Preliminary lemmas*

To begin with, we introduce some preliminary lemmas. The first lemma provides a moment inequality for sums of independent random vectors.

**Lemma A.1.** *Let $\zeta_1, \dots, \zeta_n \in \mathbb{R}^d$ be mean-zero independent random vectors and $S_n = \sum_{i=1}^n \zeta_i/\sqrt{n}$. Assume that $\max_{1\leqslant i\leqslant n}\|\zeta_i\|_p \leqslant a_1$ for some $p \geqslant 2$ and $a_1 > 0$. Then,*

$$\|S_n\|_p \leqslant C a_1, \tag{A.1}$$

*where $C > 0$ is a constant depending only on $p$. Let $\|\cdot\|_\psi$ be the Orlicz norm defined as in (3.19) and $\psi_1 = e^x - 1$. Assume that $\max_{1\leqslant i\leqslant n}\|\zeta_i\|_{\psi_1} \leqslant a_2$ for some constant $a_2 > 0$. Then,*

$$\|S_n\|_{\psi_1} \leqslant C a_2. \tag{A.2}$$

*Proof.* Noting that $p \geqslant 2$, by the Hölder inequality,

$$\|\zeta_i\|_2 \leqslant \|\zeta_i\|_p \leqslant a_1,$$

and then

$$\|S_n\|_1 \leqslant \|S_n\|_2 = \left(\frac{1}{n}\sum_{i=1}^n \|\zeta_i\|_2^2\right)^{1/2} \leqslant a_1.$$

By the Hoffmann-Jørgensen inequality (see Talagrand [30, Theorem 1]), we have there exists a constant $C_1 > 0$ depending only on $p$ such that

$$\|S_n\|_p \leqslant C_1\Big(\|S_n\|_1 + n^{-1/2}\Big\|\max_{1\leqslant i\leqslant n}\|\zeta_i\|\Big\|_p\Big). \tag{A.3}$$

By the formula of integration by part,

$$\begin{aligned}\Big\|\max_{1\leqslant i\leqslant n}\|\zeta_i\|\Big\|_p^p &= \int_0^\infty \mathbb{P}\Big(\max_{1\leqslant i\leqslant n}\|\zeta_i\|^p \geqslant t\Big)dt \\ &\leqslant \sum_{i=1}^n \int_0^\infty \mathbb{P}\Big(\|\zeta_i\|^p \geqslant t\Big)dt \\ &= \sum_{i=1}^n \|\zeta_i\|_p^p \\ &\leqslant n \max_{1\leqslant i\leqslant n}\|\zeta_i\|_p^p.\end{aligned} \tag{A.4}$$

Substituting (A.4) to (A.3), we have there exist two constants $C_2$ and $C_3$ depending only on $p$ such that

$$\|S_n\|_p \leqslant C_2\Big(a_1 + n^{-1/2+1/p}\max_{1\leqslant i\leqslant n}\|\zeta_i\|_p\Big) \leqslant C_3 a_1.$$

This proves (A.1). Note that $\|\zeta_i\|_2 \leqslant \|\zeta_i\|_{\psi_1}$, and then by a similar argument and the Hoffmann-Jørgensen inequality for the $\|\cdot\|_{\psi_1}$ norm (see Talagrand [30, Theorem 3]), the inequality (A.2) holds. □

We next introduce some notations of empirical process theory, following Van der Vaart and Wellner [32]. For any function class $\mathcal{F}$, write

$$\|\mathbb{G}_n\|_{\mathcal{F}} = \sup_{f\in\mathcal{F}}\big|\mathbb{G}_n f\big|, \quad \|\mathbb{P}_n - P\|_{\mathcal{F}} = \sup_{f\in\mathcal{F}}\big|\mathbb{P}_n f - Pf\big|. \tag{A.5}$$

By (3.1), it is easy to see $\|\mathbb{G}_n\|_{\mathcal{F}} = \sqrt{n}\|\mathbb{P}_n - P\|_{\mathcal{F}}$. Let

$$\mathcal{M}_\delta \coloneqq \big\{m_\theta - m_{\theta^*} \,:\, \|\theta-\theta^*\| \leqslant \delta\big\}.$$

Then,

$$\begin{aligned}\|\mathbb{P}_n - P\|_{\mathcal{M}_\delta} &= \sup_{\theta:\|\theta-\theta^*\|\leqslant\delta}\big|(\mathbb{M}_n - M)(\theta) - (\mathbb{M}_n - M)(\theta^*)\big|, \\ \|\mathbb{G}_n\|_{\mathcal{M}_\delta} &= \sup_{\theta:\|\theta-\theta^*\|\leqslant\delta}\sqrt{n}\big|(\mathbb{M}_n - M)(\theta) - (\mathbb{M}_n - M)(\theta^*)\big|.\end{aligned}$$

Note that $\|\mathbb{G}_n\|_{\mathcal{M}_\delta}$ may not be measurable, and we need to consider its *outer expectation*, see Van der Vaart and Wellner [32, Section 1.2] for a thorough reference. Let $\mathbb{E}^*$ be the outer expectation operator and for any map $Y$,

$$\|Y\|_p^* = \big(\mathbb{E}^*\{|Y|^p\}\big)^{1/p}, \quad \|Y\|_\psi^* = \inf\Big\{C>0 : \mathbb{E}^*\Big\{\psi\Big(\frac{|Y|}{C}\Big)\Big\} \leqslant 1\Big\}.$$

Let $\mathbb{P}^*$ be the outer probability operator.

The next lemma provides a bound on the bounds of $\|\|\mathbb{G}_n\|_{\mathcal{M}_\delta}\|_q^*$ and $\|\|\mathbb{G}_n\|_{\mathcal{M}_\delta}\|_{\psi_1}^*$, the proof is based on the empirical process theory.

**Lemma A.2.** *For $q \geqslant 2$, assume that (3.7) is satisfied with $\|m_1(X)\|_q \leqslant a_3$ for a positive constant $a_3$. Then, we have*

$$\Big\| \|\mathbb{G}_n\|_{\mathcal{M}_\delta} \Big\|_q^* \leqslant C a_3 \sqrt{d} \delta, \tag{A.6}$$

*where $C > 0$ is a constant depending only on $q$. Assume further that there exists a constant $a_4 > 0$ such that $\|m_1(X)\|_{\psi_1} \leqslant a_4$, then we have*

$$\Big\| \|\mathbb{G}_n\|_{\mathcal{M}_\delta} \Big\|_{\psi_1}^* \leqslant C a_4 \sqrt{d} \delta, \tag{A.7}$$

*where $C > 0$ is an absolute constant.*

*Proof.* Recall that by (3.7), and $\|m_1(X)\|_2 \leqslant \|m_1(X)\|_q \leqslant a_3$, and we have $\mathcal{M}_\theta$ has an envelope $F = \delta m_1$ such that $\|F(X)\|_q \leqslant a_3 \delta$. It has been shown (see, e.g., Van der Vaart [31, Chapters 5 and 19] and Wellner [34, Corollary 3.1]) that

$$\mathbb{E}^* \|\mathbb{G}_n\|_{\mathcal{M}_\delta} \leqslant C a_3 d^{1/2} \delta,$$

where $C > 0$ is an absolute constant. By the Hoffmann-Jørgensen's inequality (see Van der Vaart and Wellner [32, Theorem 2.14.5]), we have

$$\mathbb{E}^* \|\mathbb{G}_n\|_{\mathcal{M}_\delta}^q \leqslant C \Big( \big( \mathbb{E}^* \|\mathbb{G}_n\|_{\mathcal{M}_\delta} \big)^q + \mathbb{E}|F|^q \Big) \leqslant C a_3^q d^{\frac{q}{2}} \delta^q,$$

where $C > 0$ is a constant depending only on $q$. This proves (A.6). Inequality (A.7) follows from a similar argument. ☐

**Lemma A.3.** *Let $\mathbb{G}_n$ be as in (A.5) and let $\mathcal{F}_{\delta,j} = \big\{ h_{\theta,j} - h_{\theta^*,j} \colon \|\theta - \theta^*\| \leqslant \delta \big\}$ for $1 \leqslant j \leqslant n$.*

(i) *Assume that there exists a constant $a_5 > 0$ such that (3.14) is satisfied with*

$$\|h_0(X)\|_p \leqslant a_5,$$

*then*

$$\Big\| \|\mathbb{G}\|_{\mathcal{F}_{\delta,j}} \Big\|_p^* \leqslant C a_5 d^{1/2} \delta, \tag{A.8}$$

*and*

$$\Big\| \sup_{\|\theta-\theta^*\| \leqslant \delta} \|\Psi_n(\theta) - \Psi(\theta) - \Psi_n(\theta^*) + \Psi(\theta^*)\| \Big\|_p^* \leqslant C n^{-1/2} a_5 d \delta, \tag{A.9}$$

*where $C > 0$ is a constant depending only on $p$.*

(ii) *Assume that (3.14) is satisfied with $\|h_0(X)\|_{\psi_1} \leqslant a_5$, then (A.8) and (A.9) hold for all $p \geqslant 1$, and*

$$\Big\| \sup_{\|\theta-\theta^*\| \leqslant \delta} \|\Psi_n(\theta) - \Psi(\theta) - \Psi_n(\theta^*) + \Psi(\theta^*)\| \Big\|_{\psi_1}^* \leqslant C n^{-1/2} a_5 d^{3/2} \delta, \tag{A.10}$$

*where $C > 0$ is an absolute constant.*

*Proof.* For (i), note that (A.8) follows directly from Lemma A.2, (3.14) and (3.15). For (A.9), by (A.8) and the definitions of $h_\theta$ and $h_{\theta,j}$, $j = 1, 2, \ldots, n$,

$$\begin{aligned}\mathbb{E}^* \Big\| \sup_{\|\theta-\theta^*\|\leqslant\delta} \|\Psi_n(\theta) - \Psi(\theta) - \Psi_n(\theta^*) + \Psi(\theta^*)\| \Big\|^p \\ \leqslant d^{p/2-1} n^{-p/2} \sum_{j=1}^d \mathbb{E}^* \Big\| \|\mathbb{G}\|_{\mathcal{F}_{\delta,j}} \Big\|^p \\ \leqslant C n^{-p/2} a_5^p d^p \delta^p,\end{aligned}$$

which proves (A.9).

Observe that for any $Y \in \mathbb{R}^d$ and $p \geqslant 1$,

$$\|Y\|_p \leqslant C \|Y\|_{\psi_1}.$$

Hence, under the conditions (B1), (B4) and (B5), inequalities (A.8) and (A.9) follow by a similar arguments.

For (A.10), by Van der Vaart and Wellner [32, Theorem 2.14.5], it follows from (3.14), (3.20) and (A.8) that

$$\begin{aligned}\Big\| \|\mathbb{G}\|_{\mathcal{F}_{\delta,j}} \Big\|_{\psi_1}^* &\leqslant C\Big( \Big\| \|\mathbb{G}\|_{\mathcal{F}_{\delta,j}} \Big\|_1^* + \delta \|h_0(X)\|_{\psi_1} \Big) \\ &\leqslant C a_5 d^{1/2} \delta.\end{aligned}$$

By the triangle inequality,

$$\Big\| \sup_{\|\theta-\theta^*\|\leqslant\delta} \|\Psi_n(\theta) - \Psi(\theta) - \Psi_n(\theta^*) + \Psi(\theta^*)\| \Big\|_{\psi_1}^* \leqslant n^{-1/2} \sum_{j=1}^d \Big\| \|\mathbb{G}\|_{\mathcal{F}_{\delta,j}} \Big\|_{\psi_1}^* \leqslant C n^{-1/2} a_5 d^{3/2} \delta,$$

and hence (A.10) holds. □

### *A.2. Proof of Lemma 5.2*

For each $n$, let

$$A_{j,n} = \{\theta : 2^{j-1} < \sqrt{n}\|\theta - \theta^*\| \leqslant 2^j\}, \quad j \geqslant 1.$$

Recall that $\hat\theta_n$ truly minimizes $\mathbb{M}_n(\theta)$, and thus,

$$\begin{aligned}\mathbb{E}^* \big\{ \sqrt{n}\|\hat\theta_n - \theta^*\| \big\}^p &\leqslant \sum_{j\geqslant 1} 2^{jp}\, \mathbb{P}^*\big(\hat\theta_n \in A_{j,n}\big) \\ &\leqslant \sum_{j\geqslant 1} 2^{jp}\, \mathbb{P}^*\Big( \inf_{\theta\in A_{j,n}} \big(\mathbb{M}_n(\theta) - \mathbb{M}_n(\theta^*)\big) \leqslant 0 \Big).\end{aligned} \tag{A.11}$$

By (3.6), we have

$$\inf_{\theta\in A_{j,n}} \big(M(\theta) - M(\theta^*)\big) \geqslant \mu \inf_{\theta\in A_{j,n}} \|\theta - \theta^*\|^2 \geqslant \mu n^{-1} 2^{2j-2}.$$

Set $\delta_j = 2^j n^{-1/2}$. By Lemma A.2 and the Chebyshev inequality, we have

$$\begin{aligned}
\text{RHS of (A.11)} &\leqslant \sum_{j\geqslant 1} 2^{jp}\, \mathbb{P}^*\biggl( \|\mathbb{G}_n\|_{\mathcal{M}_{\delta_j}} \geqslant \frac{\mu 2^{2j-2}}{\sqrt{n}} \biggr) \\
&\leqslant \Bigl(\frac{4}{\mu}\Bigr)^{p+1} n^{\frac{p+1}{2}} \sum_{j\geqslant 1} 2^{-j(p+2)}\, \mathbb{E}^*\Bigl\{ \|\mathbb{G}_n\|_{M_{\delta_j}}^{p+1} \Bigr\} \\
&\leqslant C\Bigl(\frac{a_6 d^{1/2}}{\mu}\Bigr)^{p+1} n^{\frac{p+1}{2}} \sum_{j\geqslant 1} 2^{-j(p+2)} \delta_j^{p+1} \\
&\leqslant C\Bigl(\frac{a_6}{\mu}\Bigr)^{p+1} d^{\frac{p+1}{2}},
\end{aligned}$$

where $C > 0$ depends only on $p$. This completes the proof.

### *A.3. Proof of Lemma* **5.3**

Write $Y_i = \ddot{m}_{\theta^*}(X_i) - \mathbb{E}\{\ddot{m}_{\theta^*}(X_i)\}$. By the Rosenthal inequality for random matrices (see, e.g., Chen et al. [12, Theorem A.1]), and noting that $Y_i$'s are symmetric $(d \times d)$-matrices satisfying (3.9),

$$\begin{aligned}
\mathbb{E}\Bigl\| \sum_{i=1}^n Y_i \Bigr\|^4 &\leqslant C \Bigl\| \Bigl( \sum_{i=1}^n \mathbb{E} Y_i^2 \Bigr)^{1/2} \Bigr\|^4 + C\, \mathbb{E}\bigl\{ \max_{1\leqslant i\leqslant n} \|Y_i\|^4 \bigr\} \\
&\leqslant Cn^{-2} \Bigl\| \bigl( \mathbb{E}\bigl\{ \ddot{m}_{\theta^*}(X)^2 \bigr\} \bigr)^{1/2} \Bigr\|^4 + Cn^{-3}\, \mathbb{E}\|\ddot{m}_{\theta^*}(X)\|^4 \\
&\leqslant Cn^{-2} \|m_3(X)\|_2^4 \times \|I_d\|^4 + Cn^{-3} \|m_3(X)\|_4^4 \times \|I_d\|^4 \\
&\leqslant Cn^{-2} c_3^4,
\end{aligned}$$

where $C > 0$ is an absolute constant and we use the fact that $\|I_d\| = 1$ in the last inequality. This proves (5.6). For $H_2$, by (3.8), we have

$$\mathbb{E}\bigl\{ H_2^4 \bigr\} \leqslant \max_{1\leqslant i\leqslant n} \mathbb{E}\{m_2(X_i)^4\} \leqslant c_2^4.$$

This completes the proof of (5.7) and hence the lemma.

### *A.4. Proof of Lemma* **5.4**

By (5.1) and (5.2) and the construction of $\hat{\theta}_n^{(i)}$, we have

$$\begin{aligned}
\|V(\hat{\theta}_n - \hat{\theta}_n^{(i)})\| \leqslant{} & \frac{1}{n} \bigl\| \xi_i - \xi_i' \bigr\| + \bigl( H_1 \|\hat{\theta}_n - \theta^*\| + H_1^{(i)} \|\hat{\theta}_n^{(i)} - \theta^*\| \bigr) \\
& + \bigl( H_2 \|\hat{\theta}_n - \theta^*\|^2 + H_2^{(i)} \|\hat{\theta}_n^{(i)} - \theta^*\|^2 \bigr),
\end{aligned}$$

where $\xi_i' = \dot{m}_{\theta^*}(X_i')$ is an independent copy of $\xi_i$.

Note that $(\xi_i, \hat{\theta}_n, H_1, H_2)$ has the same distribution as $(\xi_i', \hat{\theta}_n^{(i)}, H_1^{(i)}, H_2^{(i)})$. By Lemma 5.2 with $p = 8$ and Lemma 5.3 and the Hölder inequality, we have

$$\begin{aligned}
\mathbb{E}\|V(\hat{\theta}_n - \hat{\theta}_n^{(i)})\|^2 &\leqslant 4\Big(n^{-2}\,\mathbb{E}\|\xi_i\|^2 + \mathbb{E}\{H_1^2\|\hat{\theta}_n - \theta^*\|^2\} + \mathbb{E}\{H_2^2\|\hat{\theta}_n - \theta^*\|^4\}\Big)\\
&\leqslant 4\Big(n^{-2}\|\xi_i\|_2^2 + \|H_1\|_4^2\|\hat{\theta}_n - \theta^*\|_8^2 + \|H_2\|_4^2\|\hat{\theta}_n - \theta^*\|_8^4\Big)\\
&\leqslant Cn^{-2}\Big(c_4^2 d + \mu^{-9/4}c_1^{9/4}c_3^2 d^{9/8} + \mu^{-9/2}c_1^{9/2}c_2^2 d^{9/4}\Big),
\end{aligned}$$

where $C > 0$ is an absolute constant. The result (5.8) immediately follows from the condition that $\lambda_{\min}(V) \geqslant \lambda_2$.

### *A.5. Proof of Lemma 5.6*

The inequality (5.17) follows from Lemma A.1 and (3.17), Note that by (A.9), we have

$$\mathbb{E}\Delta_1^2 \leqslant \lambda_2^{-1} n\,\mathbb{E}^*\Big\|\sup_{\theta:\|\theta-\theta^*\|\leqslant\delta_n}\|\Psi_n(\theta) - \Psi(\theta) - \Psi_n(\theta^*) + \Psi(\theta^*)\|\Big\|^2 \leqslant C\lambda_2^{-1}c_2^2 d^2\delta_n^2.$$

This proves (5.18).

By (3.12), (3.11) and the Cauchy inequality, we have

$$\begin{aligned}
\mu\|\hat{\theta}_n - \theta^*\|^2 &\leqslant \big\langle \hat{\theta}_n - \theta^*, \Psi(\hat{\theta}_n) - \Psi(\theta^*)\big\rangle\\
&= -\big\langle \hat{\theta}_n - \theta^*, \Psi_n(\theta^*) - \Psi(\theta^*)\big\rangle\\
&\quad - \big\langle \hat{\theta}_n - \theta^*, \big(\Psi_n(\hat{\theta}_n) - \Psi(\hat{\theta}_n)\big) - \big(\Psi_n(\theta^*) - \Psi(\theta^*)\big)\big\rangle\\
&\leqslant \|\hat{\theta}_n - \theta^*\|\|\Psi_n(\theta^*) - \Psi(\theta^*)\|\\
&\quad + \|\hat{\theta}_n - \theta^*\|\big\|\big(\Psi_n(\hat{\theta}_n) - \Psi(\hat{\theta}_n)\big) - \big(\Psi_n(\theta^*) - \Psi(\theta^*)\big)\big\|,
\end{aligned}$$

which implies

$$\|\hat{\theta}_n - \theta^*\| \leqslant \frac{1}{\mu}\|\Psi_n(\theta^*) - \Psi(\theta^*)\| + \frac{1}{\mu}\big\|\big(\Psi_n(\hat{\theta}_n) - \Psi(\hat{\theta}_n)\big) - \big(\Psi_n(\theta^*) - \Psi(\theta^*)\big)\big\|. \tag{A.12}$$

By (3.17) and applying Lemma A.1 to $\Psi_n(\theta^*) - \Psi(\theta^*)$, we have

$$\|\Psi_n(\theta^*) - \Psi(\theta^*)\|_p \leqslant Cc_3 d^{1/2} n^{-1/2}. \tag{A.13}$$

Taking expectations on both sides of (A.12), by (A.13) and Lemma A.3, we obtain

$$\begin{aligned}
\|\hat{\theta}_n - \theta^*\|_p &\leqslant \frac{1}{\mu}\|\Psi_n(\theta^*) - \Psi(\theta^*)\|_p\\
&\quad + \frac{1}{\mu}\Big\|\sup_{\theta\in\Theta}\big\|\big(\Psi_n(\theta) - \Psi(\theta)\big) - \big(\Psi_n(\theta^*) - \Psi(\theta^*)\big)\big\|\Big\|_p^*\\
&\leqslant C\mu^{-1}(c_3 d^{1/2} n^{-1/2} + c_2 d n^{-1/2} D_\Theta)\\
&\leqslant C(D_\Theta + 1) d n^{-1/2}.
\end{aligned}$$

This proves (5.19).

Now we move to prove (5.20). By (3.11), we have

$$\Psi(\hat{\theta}_n) - \Psi(\theta^*) = -\big(\Psi_n(\theta^*) - \Psi(\theta^*)\big) - \big(\Psi_n(\hat{\theta}_n) - \Psi(\hat{\theta}_n)\big) + \big(\Psi_n(\theta^*) - \Psi(\theta^*)\big),$$

and

$$\Psi(\hat{\theta}_n^{(i)}) - \Psi(\theta^*) = -\big(\Psi_n^{(i)}(\theta^*) - \Psi(\theta^*)\big) - \big(\Psi_n^{(i)}(\hat{\theta}_n^{(i)}) - \Psi(\hat{\theta}_n^{(i)})\big) + \big(\Psi_n^{(i)}(\theta^*) - \Psi(\theta^*)\big).$$

On the event that $\|\hat{\theta}_n - \theta^*\| \leqslant \delta$ and $\|\hat{\theta}_n^{(i)} - \theta^*\| \leqslant \delta$, taking difference of the foregoing terms, and by (3.12) again, we have

$$\begin{aligned}
&\mu\|\hat{\theta}_n - \hat{\theta}_n^{(i)}\|\mathbb{1}\big(\|\hat{\theta}_n - \theta^*\| \leqslant \delta, \|\hat{\theta}_n^{(i)} - \theta^*\| \leqslant \delta\big)\\
&\quad\leqslant \big\|\Psi(\hat{\theta}_n) - \Psi(\hat{\theta}_n^{(i)})\big\|\mathbb{1}\big(\|\hat{\theta}_n - \theta^*\| \leqslant \delta, \|\hat{\theta}_n^{(i)} - \theta^*\| \leqslant \delta\big)\\
&\quad\leqslant \frac{1}{n}\|\xi_i - \xi_i'\| + 2\sup_{\theta:\|\theta-\theta^*\|\leqslant\delta}\big\|\big(\Psi_n(\theta) - \Psi(\theta)\big) - \big(\Psi_n(\theta^*) - \Psi(\theta^*)\big)\big\|.
\end{aligned}$$

By (3.17) and Lemma A.3,

$$\begin{aligned}
&\mu\,\mathbb{E}\{\|\hat{\theta}_n - \hat{\theta}_n^{(i)}\|^p\mathbb{1}\big(\|\hat{\theta}_n - \theta^*\| \leqslant \delta, \|\hat{\theta}_n^{(i)} - \theta^*\| \leqslant \delta\big)\}\\
&\quad\leqslant Cd^{p/2}n^{-p} + Cd^p n^{-p/2}\delta^p,
\end{aligned}$$

and then we complete the proof of (5.20).

### A.6. Proof of Lemma 5.7

For (5.21), note that $\xi_i$ and $\hat{\theta}_n^{(i)}$ are independent, and $\hat{\theta}_n$ has the same distribution as $\hat{\theta}_n^{(i)}$,

$$\begin{aligned}
&\mathbb{E}\big\{\|\xi_i\|\|\hat{\theta}_n - \theta^*\|^2\mathbb{1}\big(\hat{\theta}_n \in B_\delta, \hat{\theta}_n^{(i)} \in B_\delta^c\big)\big\}\\
&\quad\leqslant \delta^2\,\mathbb{E}\big\{\|\xi_i\|\mathbb{1}\big(\hat{\theta}_n^{(i)} \in B_\delta^c\big)\big\}\\
&\quad\leqslant \delta^2\,\mathbb{E}\|\xi_i\|\,\mathbb{P}\big(\|\hat{\theta}_n^{(i)} - \theta^*\| > \delta\big)\\
&\quad\leqslant Cd^{1/2}\delta^2\,\mathbb{P}\big(\|\hat{\theta}_n - \theta^*\| > \delta\big)\\
&\quad\leqslant Cd^{1/2}\delta^{-p+2}\,\mathbb{E}\big\{\|\hat{\theta}_n - \theta^*\|^p\big\}\\
&\quad\leqslant C(D_\Theta + 1)^p d^{p+1/2}\delta^{-p+2}n^{-p/2},
\end{aligned}$$

where we used (5.19) in the last inequality.

For (5.22), by the Hölder inequality, we have

$$\begin{aligned}
&\mathbb{E}\big\{\|\xi_i\|\|\hat{\theta}_n^{(i)} - \theta^*\|^2\mathbb{1}\big(\hat{\theta}_n \in B_\delta^c, \hat{\theta}_n^{(i)} \in B_\delta\big)\big\}\\
&\leqslant \Big(\mathbb{E}\big\{\|\xi_i\|^{p/2}\|\hat{\theta}_n^{(i)} - \theta^*\|^p\big\}\Big)^{2/p}\Big(\mathbb{P}\big(\|\hat{\theta}_n - \theta^*\| > \delta\big)\Big)^{(p-2)/p}\\
&\leqslant \delta^{-p+2}\Big(\mathbb{E}\big\{\|\xi_i\|^{p/2}\big\}\,\mathbb{E}\{\|\hat{\theta}_n^{(i)} - \theta^*\|^p\}\Big)^{2/p}\Big(\mathbb{E}\big\{\|\hat{\theta}_n - \theta^*\|^p\big\}\Big)^{(p-2)/p}\\
&\leqslant C(D_\Theta + 1)^p d^{p+1/2}\delta^{-p+2}n^{-p/2}.
\end{aligned}$$

As for (5.23), recalling that $p \geqslant 3$, and by (3.17), (5.19) and (5.20) and the Hölder inequality, we have

$$
\begin{aligned}
&\mathbb{E}\Big\{ \|\xi_i\| (\|\hat{\theta}_n - \theta^*\| + \|\hat{\theta}_n^{(i)} - \theta^*\|) \|\hat{\theta}_n - \hat{\theta}_n^{(i)}\| \mathbb{1}\big(\hat{\theta}_n \in B_\delta, \hat{\theta}_n^{(i)} \in B_\delta\big) \Big\} \\
&\quad \leqslant \|\xi_i\|_p \|\hat{\theta}_n - \theta^*\|_p \Big\| \|\hat{\theta}_n - \hat{\theta}_n^{(i)}\| \mathbb{1}\big(\hat{\theta}_n \in B_\delta, \hat{\theta}_n^{(i)} \in B_\delta\big) \Big\} \Big\|_p \\
&\qquad + \|\xi_i\|_p \|\hat{\theta}_n^{(i)} - \theta^*\|_p \Big\| \|\hat{\theta}_n - \hat{\theta}_n^{(i)}\| \mathbb{1}\big(\hat{\theta}_n \in B_\delta, \hat{\theta}_n^{(i)} \in B_\delta\big) \Big\} \Big\|_p \\
&\quad = 2\|\xi_i\|_p \|\hat{\theta}_n - \theta^*\|_p \Big\| \|\hat{\theta}_n - \hat{\theta}_n^{(i)}\| \mathbb{1}\big(\hat{\theta}_n \in B_\delta, \hat{\theta}_n^{(i)} \in B_\delta\big) \Big\} \Big\|_p \\
&\quad \leqslant C(D_\Theta + 1)\Big( d^2 n^{-3/2} + d^{5/2} n^{-1} \delta \Big).
\end{aligned}
$$

This completes the proof.

### *A.7. Proof of Lemma 5.9*

In this proof, we denote by $C$ a positive constant that depends only on $c_1$, $c_4$, $c_5$, $\mu$, $\lambda_1$ and $\lambda_2$ and $C_p$ a constant that also depends on $p$, which might take different values in different places. By (ii) of Lemma A.3 and following the proof of Lemma 5.6, we have (5.19) and (5.20) also hold for a positive constant $C_p$. This proves the first argument of this lemma. Note that $\Psi_n(\theta^*) - \Psi(\theta^*) = n^{-1}\sum_{i=1}^n \xi_i$, and by Lemma A.1 and (3.21),

$$
\|\Psi_n(\theta^*) - \Psi(\theta^*)\|_{\psi_1} \leqslant C c_5 d^{1/2} n^{-1/2}, \tag{A.14}
$$

and for any $p \geqslant 1$,

$$
\|\Psi_n(\theta^*) - \Psi(\theta^*)\|_p \leqslant C_p c_5 d^{1/2} n^{-1/2}.
$$

For (5.29), by the fact that $\mathbb{P}(\|Y\| > t) \leqslant 2e^{-t/\|Y\|_{\psi_1}}$, it follows from (A.14) and Lemma A.3 that

$$
\mathbb{P}\big(\|\Psi_n(\theta^*) - \Psi(\theta^*)\| > \mu t/2\big) \leqslant 2\exp\Big(-C'' \frac{\sqrt{n}\mu t}{c_5}\Big),
$$

and

$$
\mathbb{P}^*\bigg(\sup_{\theta\in\Theta} \|\Psi_n(\theta) - \Psi(\theta) - \Psi_n(\theta^*) + \Psi(\theta^*)\| > \mu t/2\bigg) \leqslant 2\exp\Big(-\frac{C''\sqrt{n}t}{c_4 d^{3/2}(D_\Theta + 1)}\Big),
$$

where $C'' > 0$ is an absolute constant. By (A.12), for any $t > 0$,

$$
\begin{aligned}
\mathbb{P}\big(\|\hat{\theta}_n - \theta^*\| > t\big) &\leqslant \mathbb{P}\big(\|\Psi_n(\theta^*) - \Psi(\theta^*)\| > \mu t/2\big) \\
&\qquad + \mathbb{P}^*\bigg(\sup_{\theta\in\Theta} \|\Psi_n(\theta) - \Psi(\theta) - \Psi_n(\theta^*) + \Psi(\theta^*)\| > \mu t/2\bigg) \\
&\leqslant 2\exp\Big(-\frac{C''\sqrt{n}\mu t}{c_4 d^{3/2}(D_\Theta + 1) + c_5 d^{1/2}}\Big).
\end{aligned}
$$

Taking $C' = C''\mu/(c_4+c_5)$, we complete the proof of (5.29). By Lemma 5.9 and the Hölder inequality, we have

$$\begin{aligned}\mathbb{E}\Big\{\|\xi_i\|\|\hat{\theta}_n-\theta^*\|^2\mathbb{1}\big(\hat{\theta}_n\in B_\delta,\hat{\theta}_n^{(i)}\in B_\delta^c\big)\Big\} &\leqslant \big(\mathbb{E}\|\xi_i\|^4\big)^{1/4}\Big(\mathbb{E}\|\hat{\theta}_n-\theta^*\|^4\Big)^{1/2}\Big(\mathbb{P}\big(\hat{\theta}_n^{(i)}\in B_\delta^c\big)\Big)^{1/4}\\ &\leqslant C(D_\Theta+1)^2 d^{5/2}n^{-1}\exp\Big(-\frac{C'n^{1/2}\delta}{4(D_\Theta+1)d^{3/2}}\Big),\end{aligned}$$

where we used (3.21), (5.19) and (5.29) in the last line. This proves (5.30). The inequality (5.31) can be derived using a similar argument.

### A.8. Proof of Lemma 5.12

We use a recursion inequality to prove the bound. By (3.22), we have

$$\begin{aligned}\|\theta_n-\theta^*\|^2 = \|\theta_{n-1}-\theta^*\|^2 - 2\ell_n\big\langle\theta_{n-1}-\theta^*,\nabla f(\theta_{n-1})+\zeta_n\big\rangle\\ +\ell_n^2\|\nabla f(\theta_{n-1})+\zeta_n\|^2.\end{aligned}\tag{A.15}$$

Recall that by (C1), $\zeta_n=\xi_n+\eta_n$ where $\xi_n$ is independent of $\theta_{n-1}$, $\eta_n$ is $\mathcal{F}_n$ measurable, $\|\eta_n\|\leqslant c_1\|\theta_{n-1}-\theta^*\|$, $\mathbb{E}\{\eta_n|\mathcal{F}_{n-1}\}=0$ and $\theta_{n-1}\in\mathcal{F}_{n-1}$. Therefore, $\mathbb{E}\big\{\big\langle\nabla f(\theta_{n-1}),\zeta_n\big\rangle\bigm|\mathcal{F}_{n-1}\big\}=0$. Moreover, with $L_2:=c_1+L$,

$$\begin{aligned}&\mathbb{E}\big\{\big\|\nabla f(\theta_{n-1})+\zeta_n\big\|^2\bigm|\mathcal{F}_{n-1}\big\}\\ &=\mathbb{E}\big\{\big\|\nabla f(\theta_{n-1})\big\|^2\bigm|\mathcal{F}_{n-1}\big\}+\mathbb{E}\big\{\big\|\zeta_n\big\|^2\bigm|\mathcal{F}_{n-1}\big\}\\ &=\mathbb{E}\big\{\big\|\nabla f(\theta_{n-1})-\nabla f(\theta^*)\big\|^2\bigm|\mathcal{F}_{n-1}\big\}+\mathbb{E}\big\{\big\|\xi_n+\eta_n\big\|^2\bigm|\mathcal{F}_{n-1}\big\}\\ &\leqslant 2L_2^2\|\theta_{n-1}-\theta^*\|^2+2\,\mathbb{E}\|\xi_n\|^2.\end{aligned}\tag{A.16}$$

For the intersection term of (A.15), under the strong convexity assumption (3.24),

$$\big\langle\nabla f(\theta_1)-\nabla f(\theta_2),\theta_1-\theta_2\big\rangle\geqslant\mu\|\theta_1-\theta_2\|^2,$$

and noting that $\mathbb{E}\{\langle\theta_{n-1}-\theta^*,\zeta_n\rangle|\mathcal{F}_{n-1}\}=0$,

$$\begin{aligned}&\mathbb{E}\big\{\big\langle\theta_{n-1}-\theta^*,\nabla f(\theta_{n-1})+\zeta_n\big\rangle\bigm|\mathcal{F}_{n-1}\big\}\\ &\quad=\mathbb{E}\big\langle\theta_{n-1}-\theta^*,\nabla f(\theta_{n-1})-\nabla f(\theta^*)\big\rangle\\ &\quad\geqslant\mu\|\theta_{n-1}-\theta^*\|^2.\end{aligned}\tag{A.17}$$

Combining (A.15)–(A.17),

$$\mathbb{E}\big\{\|\theta_n-\theta^*\|^2\bigm|\mathcal{F}_{n-1}\big\}\leqslant\big(1-2\mu\ell_n+2L_2^2\ell_n^2\big)\|\theta_{n-1}-\theta^*\|^2+2\ell_n^2\,\mathbb{E}\|\xi_n\|^2.$$

Taking expectations on both sides yields

$$\delta_n\leqslant\big(1-2\mu\ell_n+2L_2\ell_n^2\big)\delta_{n-1}+2\ell_n^2\,\mathbb{E}\|\xi_n\|^2.\tag{A.18}$$

Observe that $\mu \leqslant L_2$ and thus $2\mu\ell_n - 2L_2^2\ell_n^2 \leqslant 2L_2\ell_n - 2L_2^2\ell_n^2 \leqslant 1/2$. This ensures that the coefficient in front of $\delta_{n-1}$ is always positive. By (C0), we have

$$\delta_0 = \mathbb{E}\|\theta_0 - \theta^*\|^2 \leqslant \tau_0^2. \tag{A.19}$$

By (A.18) and (A.19) and applying the recursion $n$ times, recalling that $\max_{1\leqslant i\leqslant n}\mathbb{E}\|\xi_i\|^2 \leqslant \tau^2$, we have

$$\begin{aligned}\delta_n &\leqslant \prod_{k=1}^n \bigl(1 - 2\mu\ell_k + 2L_2^2\ell_k^2\bigr)\delta_0 + 2\tau^2\sum_{k=1}^n\prod_{i=k+1}^n \bigl(1 - 2\mu\ell_i + 2L_2^2\ell_i^2\bigr)\ell_k^2\\ &\leqslant \prod_{k=1}^n \bigl(1 - 2\mu\ell_k + 2L_2^2\ell_k^2\bigr)\tau_0^2 + 2\tau^2\sum_{k=1}^n\prod_{i=k+1}^n \bigl(1 - 2\mu\ell_i + 2L_2^2\ell_i^2\bigr)\ell_k^2.\end{aligned}$$

Following the proof of Bach and Moulines (2011, Eqs. (18), (23) and (24)), we have

$$\delta_n \leqslant \Bigl(\tau_0^2 + \frac{\tau^2}{L_2^2}\Bigr)\exp\Bigl(-\mu\sum_{k=1}^n \ell_k + 4L_2^2\sum_{k=1}^n \ell_k^2\Bigr) + 2\tau^2\sum_{k=1}^n\prod_{i=k+1}^n (1 - \mu\ell_i)\ell_k^2 \tag{A.20}$$

$$\begin{aligned}&\leqslant \Bigl(\tau_0^2 + \frac{\tau^2}{L_2^2}\Bigr)\exp\Bigl(-\mu\sum_{k=1}^n \ell_k + 4L_2^2\sum_{k=1}^n \ell_k^2\Bigr) \qquad \text{(A.21)}\\ &\quad + 2\tau^2\biggl\{\exp\Bigl(-\mu\sum_{i=n/2+1}^n \ell_i\Bigr)\sum_{k=1}^n \ell_k^2 + \frac{\ell_{n/2}}{\mu}\biggr\}.\end{aligned}$$

We next consider the cases where $\alpha \in (1/2, 1)$ and $\alpha = 1$, separately. If $\alpha \in (1/2, 1)$, by (A.19) and (A.21), we have

$$\sum_{k=1}^n \ell_k^2 \leqslant \ell_0^2\sum_{k=1}^\infty k^{-2\alpha} \leqslant C,$$

and

$$\delta_n \leqslant C\exp\bigl(-\frac{\mu\ell_0}{4}n^{1-\alpha}\bigr)\Bigl(\tau_0^2 + \frac{\tau^2}{L_2^2}\Bigr) + \frac{4\ell_0\tau^2}{\mu n^\alpha} \leqslant Cn^{-\alpha}(\tau^2 + \tau_0^2).$$

This proves (5.42). Now we move to prove (5.43). For $\alpha = 1$, i.e., $\ell_i = \ell_0 i^{-1}$, we use the following basic inequalities:

$$\log n \leqslant \sum_{k=1}^n k^{-1} \leqslant \log n + 1, \quad \sum_{k=1}^\infty k^{-2} \leqslant 2.$$

For the first term of (A.20), we have

$$\exp\biggl(-\mu\sum_{k=1}^{n}\ell_k+4L_2^2\sum_{k=1}^{n}\ell_k^2\biggr)\leqslant\exp\bigl(8\ell_0^2L_2^2\bigr)\exp\bigl(-\mu\ell_0\log n\bigr)\leqslant Cn^{-\mu\ell_0},$$

and for the second term of (A.20), we obtain

$$\begin{aligned}\sum_{k=1}^{n}\prod_{i=k+1}^{n}(1-\mu\ell_i)\ell_k^2 &\leqslant \ell_0^2\sum_{k=1}^{n}k^{-2}\exp\Bigl\{-\mu\ell_0\sum_{i=k+1}^{n}i^{-1}\Bigr\}\\ &\leqslant \ell_0^2e^{\mu\ell_0}\sum_{k=1}^{n}k^{-2}\exp\bigl\{-\mu\ell_0\log n+\mu\ell_0\log k\bigr\}\\ &\leqslant \ell_0^2e^{\mu\ell_0}n^{-\mu\ell_0}\sum_{k=1}^{n}k^{-2+\mu\ell_0}\\ &\leqslant \begin{cases}Cn^{-1}, & \mu\ell_0>1,\\ Cn^{-1}\log n, & \mu\ell_0=1,\\ Cn^{-\mu\ell_0}, & 0<\mu\ell_0<1.\end{cases}\end{aligned}$$

Therefore, for $\alpha=1$,

$$\delta_n\leqslant\begin{cases}Cn^{-1}(\tau^2+\tau_0^2), & \mu\ell_0>1,\\ Cn^{-1}(\log n)(\tau^2+\tau_0^2), & \mu\ell_0=1,\\ Cn^{-\mu\ell_0}(\tau^2+\tau_0^2), & 0<\mu\ell_0<1.\end{cases}$$

This proves (5.43).

### *A.9. Proof of Lemma 5.13*

By the construction of $(\theta_j^{(i)})_{1\leqslant j\leqslant n}$,

$$\theta_j-\theta_j^{(i)}=\begin{cases}0, & j<i;\\ -\ell_j(\xi_j-\xi_j'+\eta_j-\eta_j^{(i)}), & j=i;\\ \bigl(\theta_{j-1}-\theta_{j-1}^{(i)}\bigr)-\ell_j\bigl(\nabla f(\theta_{j-1})-\nabla f(\theta_{j-1}^{(i)})+\eta_j-\eta_j^{(i)}\bigr), & j>i.\end{cases}$$

Since $\xi_i\overset{d}{=}\xi_i'$, $\eta_i\overset{d}{=}\eta_i^{(i)}$ and $\|\eta_i\|\leqslant c_2\|\theta_{i-1}-\theta^*\|$, it follows from (C1) and Lemma 5.12 that

$$\begin{aligned}\mathbb{E}\|\theta_i-\theta_i^{(i)}\|^2 &\leqslant C\ell_i^2(\mathbb{E}\|\xi_i\|^2+\mathbb{E}\|\eta_i\|^2)\\ &\leqslant Ci^{-2\alpha}(\tau^2+c_2\,\mathbb{E}\|\theta_{i-1}-\theta^*\|^2)\\ &\leqslant Ci^{-2\alpha}(\tau^2+\tau_0^2).\end{aligned}\tag{A.22}$$

For $j > i$, using a similar argument leading to (A.18),

$$\begin{aligned}
&\mathbb{E}\|\theta_j - \theta_j^{(i)}\|^2 \\
&= \mathbb{E}\|\theta_{j-1} - \theta_{j-1}^{(i)}\|^2 - 2\ell_j\,\mathbb{E}\Big\{\theta_{j-1} - \theta_{j-1}^{(i)}, \nabla f(\theta_{j-1}) - \nabla f(\theta_{j-1}^{(i)})\Big\} \\
&\quad + \ell_j^2\,\mathbb{E}\|\nabla f(\theta_{j-1}) - \nabla f(\theta_{j-1}^{(i)}) + \eta_j - \eta_j^{(i)}\|^2 \\
&\leqslant \big(1 - 2\mu\ell_j + 2L_2^2\ell_j^2\big)\,\mathbb{E}\|\theta_{j-1} - \theta_{j-1}^{(i)}\|^2,
\end{aligned}$$

Solving the recursive system, and by (A.22), we have

$$\begin{aligned}
\mathbb{E}\|\theta_j - \theta_j^{(i)}\|^2 &\leqslant \exp\Big\{-2\mu\sum_{k=i+1}^{j}\ell_k + 2L_2^2\sum_{k=1}^{j}\ell_k^2\Big\}\,\mathbb{E}\|\theta_i - \theta_i^{(i)}\|^2 \\
&\leqslant C(\tau^2 + \tau_0^2)i^{-2\alpha}\exp\Big\{-2\mu\sum_{k=i+1}^{j}\ell_k\Big\}.
\end{aligned}$$

For $0 < \alpha < 1$, using a similar argument as in the proof of Theorem 3.4, we have for $j \geqslant i$,

$$\begin{aligned}
\mathbb{E}\|\theta_j - \theta_j^{(i)}\|^2 &\leqslant C(\tau^2 + \tau_0^2)i^{-2\alpha}\exp\Big\{-\mu\big(\varphi_{1-\alpha}(j) - \varphi_{1-\alpha}(i)\big)\Big\} \\
&\leqslant \begin{cases} C(\tau^2 + \tau_0^2)i^{-2\alpha}, & i \leqslant j \leqslant 2i, \\ C(\tau^2 + \tau_0^2)i^{-2\alpha}\exp\Big\{-\mu\big(\varphi_{1-\alpha}(j) - \varphi_{1-\alpha}(\tfrac{j}{2})\big)\Big\}, & j > 2i, \end{cases} \\
&\leqslant \begin{cases} C(\tau^2 + \tau_0^2)(j/2)^{-2\alpha}, & i \leqslant j \leqslant 2i, \\ C(\tau^2 + \tau_0^2)i^{-2\alpha}\exp\Big\{-\tfrac{\mu}{2}\varphi_{1-\alpha}(j)\Big\}, & j > 2i \end{cases} \\
&\leqslant C(\tau^2 + \tau_0^2)j^{-2\alpha}.
\end{aligned}$$

If $\alpha = 1$, (5.46) can be shown similarly.

### *A.10. Proof of Lemma 5.14*

Recall that for any $j \geqslant 1$,

$$\theta_j - \theta^* = (\theta_{j-1} - \theta^*) - \ell_j\big(\nabla f(\theta_{j-1}) + \zeta_j\big),$$

where $\zeta_j$ is a martingale difference such that $\mathbb{E}\{\zeta_j|\mathcal{F}_{j-1}\} = 0$, and $\theta_{j-1}$ is $\mathcal{F}_{j-1}$-measurable. Hence,

$$\begin{aligned}
\|\theta_j - \theta^*\|^4 &= \|\theta_{j-1} - \theta^*\|^4 + 4\ell_j^2\big\langle\theta_{j-1} - \theta^*, \nabla f(\theta_{j-1}) + \zeta_j\big\rangle^2 + \ell_j^4\|\nabla f(\theta_{j-1}) + \zeta_j\|^4 \\
&\quad - 4\ell_j\|\theta_{j-1} - \theta^*\|^2\big\langle\theta_{j-1} - \theta^*, \nabla f(\theta_{j-1}) + \zeta_j\big\rangle \\
&\quad + 2\ell_j^2\|\theta_{j-1} - \theta^*\|^2\|\nabla f(\theta_{j-1}) + \zeta_j\|^2 \\
&\quad - 4\ell_j^3\big\langle\theta_{j-1} - \theta^*, \nabla f(\theta_{j-1}) + \zeta_j\big\rangle\|\nabla f(\theta_{j-1}) + \zeta_j\|^2 \\
&\leqslant \|\theta_{j-1} - \theta^*\| + 6\ell_j^2\|\theta_{j-1} - \theta^*\|^2\|\nabla f(\theta_{j-1}) + \zeta_j\|^2 \\
&\quad + \ell_j^4\|\nabla f(\theta_{j-1}) + \zeta_j\|^4 + 4\ell_j^3\|\theta_{j-1} - \theta^*\|\|\nabla f(\theta_{j-1}) + \zeta_j\|^3 \\
&\quad - 4\ell_j\|\theta_{j-1} - \theta^*\|^2\big\langle\theta_{j-1} - \theta^*, \nabla f(\theta_{j-1}) + \zeta_j\big\rangle,
\end{aligned}$$

and then

$$\begin{aligned}
&\mathbb{E}\big\{\|\theta_j-\theta^*\|^4|\mathcal{F}_{j-1}\big\}\\
&\leqslant \|\theta_{j-1}-\theta^*\|^4+6\ell_j^2\|\theta_{j-1}-\theta^*\|^2\,\mathbb{E}\big\{\|\nabla f(\theta_{j-1})+\zeta_j\|^2 \bigm| \mathcal{F}_{j-1}\big\}\\
&\quad+\ell_j^4\,\mathbb{E}\big\{\|\nabla f(\theta_{j-1})+\zeta_j\|^4 \bigm| \mathcal{F}_{j-1}\big\}+4\ell_j^3\|\theta_{j-1}-\theta^*\|\,\mathbb{E}\big\{\|\nabla f(\theta_{j-1})+\zeta_j\|^3 \bigm| \mathcal{F}_{j-1}\big\}\\
&\quad-4\ell_j\|\theta_{j-1}-\theta^*\|^2\big\langle\theta_{j-1}-\theta^*,\nabla f(\theta_{j-1})-\nabla f(\theta^*)\big\rangle.
\end{aligned}\tag{A.23}$$

Note that by (3.24),

$$\big\langle\theta_{j-1}-\theta^*,\nabla f(\theta_{j-1})-\nabla f(\theta^*)\big\rangle\geqslant\mu\|\theta_{j-1}-\theta^*\|^2.\tag{A.24}$$

For $1\leqslant p\leqslant 4$, recall that by (C1), $\mathbb{E}\{\|\xi_j\|^p\}\leqslant\tau^p$ and $\|\eta_j\|\leqslant c_1\|\theta_{j-1}-\theta^*\|$, and it follows that

$$\begin{aligned}
\mathbb{E}\big\{\|\nabla f(\theta_{j-1})+\zeta_j\|^p \bigm| \mathcal{F}_{j-1}\big\}&\leqslant\mathbb{E}\big\{\|\nabla f(\theta_{j-1})-\nabla f(\theta^*)+\xi_j+\eta_j\|^p \bigm| \mathcal{F}_{j-1}\big\}\\
&\leqslant 2^{p-1}\Big(L_2^p\|\theta_{j-1}-\theta^*\|^p+\tau^p\Big).
\end{aligned}\tag{A.25}$$

By (A.23)–(A.25), we have

$$\begin{aligned}
\mathbb{E}\big\{\|\theta_j-\theta^*\|^4 \bigm| \mathcal{F}_{j-1}\big\}&\leqslant\|\theta_{j-1}-\theta^*\|^4\big(1-4\mu\ell_j+12\ell_j^2L_2^2+16\ell_j^3L_2^3+8\ell_j^4L_2^4\big)\\
&\quad+20\|\theta_{j-1}-\theta^*\|^2\ell_j^2\tau^2+16\ell_j^4\tau^4\\
&\leqslant\|\theta_{j-1}-\theta^*\|^4\big(1-4\mu\ell_j+16\ell_j^2L_2^2+24\ell_j^4L_2^4\big)\\
&\quad+20\|\theta_{j-1}-\theta^*\|^2\ell_j^2\tau^2+16\ell_j^4\tau^4.
\end{aligned}$$

Taking expectations on both sides, we have

$$\begin{aligned}
\mathbb{E}\|\theta_j-\theta^*\|^4&\leqslant\mathbb{E}\|\theta_{j-1}-\theta^*\|^4\big(1-4\mu\ell_j+16\ell_j^2L_2^2+24\ell_j^4L_2^4\big)\\
&\quad+C(\tau^2+\tau_0^2)\tau^2j^{-3\alpha}+16\tau^4j^{-4\alpha},
\end{aligned}$$

where we used Lemma 5.12 in the last inequality. Using the similar arguments leading to Lemma 5.12 (see also Bach and Moulines (2011, pp. 16–19)), we have for $\alpha\in(0,1)$,

$$\mathbb{E}\|\theta_j-\theta^*\|^4\leqslant Cj^{-2\alpha}(\tau^4+\tau_0^4).$$

If $\alpha=1$, we have

$$\mathbb{E}\|\theta_j-\theta^*\|^4\leqslant Cj^{-2\mu\ell_0}\big(\varphi_{2\mu\ell_0-2}(j)+1\big)(\tau^4+\tau_0^4),$$

where $\varphi$ is as defined in (5.32). This proves (5.48), and hence the lemma.

## References

[1] Anastasiou, A., Balasubramanian, K. and Erdogdu, M. A. (2019). Normal approximation for stochastic gradient descent via non-asymptotic rates of martingale CLT. In *Proceedings of the Thirty-Second Conference on Learning Theory PMLR* **99** 115-137.

[2] Bach, F. and Moulines, E. (2011). Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Neural Information Processing Systems (NIPS)*, Spain.

[3] Barbour, A. (1990). Stein's method for diffusion approximations. *Probab. Theory Relat. Fields* **84**(3) 297–322.

[4] Bentkus, V. (2003). On the dependence of the Berry–Esseen bound on dimension. *J. Stat. Plan.* **113**(2) 385–402.

[5] Bentkus, V. (2005). A Lyapunov-type bound in $\mathbb{R}^d$. *Theory Probab. Appl.* **49**(2) 311–323.

[6] Bentkus, V., Bloznelis, M. and Götze, F. (1997). A Berry–Esséen bound for M-estimators. *Scand. J. Stat.* **24**(4) 485–502.

[7] Bhattacharya, R. and Holmes, S. (2010). An exposition of Götze's estimation of the rate of convergence in the multivariate central limit theorem. *Technical report* Stanford University.

[8] Chatterjee, S. and Meckes, E. (2008). Multivariate normal approximation using exchangeable pairs. *ALEA-Latin Am. J. Probab. Math. Stat.* **4** 257–283.

[9] Chen, L. H. and Shao, Q.-M. (2007). Normal approximation for nonlinear statistics using a concentration inequality approach. *Bernoulli* **13**(2) 581–599.

[10] Chen, L. H. Y. and Fang, X. (2015). Multivariate normal approximation by Stein's method: The concentration inequality approach. Available at arXiv:1111.4073.

[11] Chen, L. H. Y., Goldstein, L. and Shao, Q.-M. (2011). *Normal approximation by Stein's method*. Probability and its Applications, New York: Springer, Heidelberg.

[12] Chen, R. Y., Gittens, A. and Tropp, J. A. (2012). The masked sample covariance estimator: An analysis using matrix concentration inequalities. *Information and Inference* **1**(1) 2–20.

[13] Goldstein, L. and Rinott, Y. (1996). Multivariate normal approximations by Stein's method and size bias couplings. *J. Appl. Probab.* **33**(01) 1–17.

[14] Götze, F. (1991). On the rate of convergence in the multivariate CLT. *Ann. Probab.* **19**(2) 724–739.

[15] Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* **1** 221–233, University of California Press.

[16] Nagaev, S. V. (1976). An estimate of the remainder term in the multidimensional central limit theorem. In *Proceedings of the Third Japan — USSR Symposium on Probability Theory* Lecture Notes in Mathematics pp. 419–438, Berlin, Heidelberg: Springer.

[17] Nemirovski, A., Juditsky, A., Lan, G. and Shapiro, A. (2009). Robust stochastic approximation approach to stochastic programming. *SIAM J. Optim.* **19**(4) 1574–1609.

[18] Paulauskas, V. (1996). Rates of convergence in the asymptotic normality for some local maximum estimators. *Lith. Math. J.* **36**(1) 68–91.

[19] Pfanzagl, J. (1971). The Berry-Esseen bound for minimum contrast estimates. *Metrika* **17**(1) 82–91.

[20] Pfanzagl, J. (1972). Further results on asymptotic normality I. *Metrika* **18**(1) 174–198.

[21] Pfanzagl, J. (1973). The accuracy of the normal approximation for estimates of

vector parameters. *Z. Wahrscheinlichkeitstheorie verw. Geb* **25**(3) 171–198.

[22] Pollard, D. (1985). New ways to prove central limit theorems. *Econ. Theory* **1**(3) 295–313.

[23] Polyak, B. T. (1990). New stochastic approximation type procedures. *Avtomatica i Telemekhanika* **7** 98–107.

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

[25] Raič, M. (2019). A multivariate Berry–Esseen theorem with explicit constants. *Bernoulli* **25**(4A) 2824–2853.

[26] Reinert, G. and Röllin, A. (2009). Multivariate normal approximation with Stein's method of exchangeable pairs under a general linearity condition. *Ann. Probab.* **37**(6) 2150–2173.

[27] Ruppert, D. (1988). Efficient estimations from a slowly convergent Robbins–Monro process. *Technical Report* Cornell University Operations Research and Industrial Engineering.

[28] Senatov, V. V. (1981). Uniform estimates of the rate of convergence in the multidimensional central limit theorem. *Theory Probab. its Appl.* **25**(4) 745–759.

[29] Shao, Q.-M. and Zhou, W.-X. (2016). Cramér type moderate deviation theorems for self-normalized processes. *Bernoulli* **22**(4), pp. 2029–2079.

[30] Talagrand, M. (1989). Isoperimetry and integrability of the sum of independent Banach-space valued random variables. *Ann. Probab.* **17**(4) 1546–1570.

[31] Van der Vaart, A. W. (1998). *Asymptotic statistics*. Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press.

[32] Van der Vaart, A. W. and Wellner, J. A. (1996). *Weak convergence and empirical processes*. Springer Series in Statistics Springer New York.

[33] Vershynin, Roman. (2010). *Introduction to the non-asymptotic analysis of random matrices*. Compressed Sensing.

[34] Wellner, J. A. (2005). Empirical processes: Theory and applications. *Notes for a course given at Delft University of Technology*.