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 Science, 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: Berry–Esseen bound; Multivariate normal approximation; Randomized concentration inequality; Stein’s method; M-estimators; Averaged stochastic gradient descent algorithms

1. Introduction

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

$$T = W + D,$$

(1.1)

where

$$W = \sum_{i=1}^{n} \xi_i, \quad D := D(X_1, \ldots, X_n) = T - W,$$

(1.2)

$$\xi_i := h_i(X_i) \in \mathbb{R}^d$$

and $h_i : \mathcal{X} \mapsto \mathbb{R}^d$ is a Borel measurable function. Assume that

$$\mathbb{E}\xi_i = 0 \text{ for each } 1 \leq i \leq n \text{ and } \sum_{i=1}^{n} \mathbb{E}\xi_i \xi_i^T = I_d.$$ 

(1.3)
Let
\[\gamma := \gamma_n = \sum_{i=1}^{n} \mathbb{E}[\|\xi_i\|^3].\] 

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 \(d^{1/4} n\), 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:
\[\sup_{z \in \mathbb{R}} |\mathbb{P}(T \leq z) - \Phi(z)| \leq 6.1 \gamma + \mathbb{E}|WD| + \sum_{i=1}^{n} \mathbb{E}|\xi_i(D - D^{(i)})|,\] 

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 \(d \geq 1\) and \(x = (x_1, \ldots, x_d)\) be a vector in \(\mathbb{R}^d\). For \(x, y \in \mathbb{R}^d\), denote \(\langle x, y \rangle\) 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 \(A^T\) the transpose of \(A\) and by \(\|A\|\) the spectral norm, i.e., \(\|A\| := (\lambda_{\max}(A^T A))^{1/2}\). Let \(I_d\) be the \(d\)-dimensional identity matrix. For \(X \in \mathbb{R}\) (resp. \(\mathbb{R}^d\) and \(p \geq 1\), 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 \((X_1, \ldots, X_n), (\xi_1, \ldots, \xi_n), W, T\) and \(D\) be defined as in (1.1) and (1.2). Let \(A\) 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,

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

(2.1)

for any random variables \(\Delta\) and \((\Delta^{(i)})_{1 \leq i \leq n}\) such that \(\Delta \geq \|D\|\) 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 \((X'_1, \ldots, X'_n)\) be an independent copy of \((X_1, \ldots, X_n)\), one may choose \(\Delta = \|D\|\) and \(\Delta^{(i)} = \|D^{(i)}\|\), where \(D^{(i)} = D(X_1, \ldots, X_{i-1}, X'_i, X_{i+1}, \ldots, X_n)\). One can also choose \(D^{(i)} = D(X_1, \ldots, X_{i-1}, 0, X_{i+1}, \ldots, X_n)\). 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

\[
259\gamma + 2E\{\|W\|\Delta\} + 2\sum_{i=1}^{n} E|\xi_i(\Delta - \Delta^{(i)})|,
\]

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

\[
\sup_{A \in \mathcal{A}} |\mathbb{P}(W \in A) - \mathbb{P}(Z \in A)| \leq Cd^{1/4}\gamma,
\]

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 Račič [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 \(\Delta \geq \|D\|\mathbb{1}(O)\). Under the conditions of Theorem 2.1, we have

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

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 $(\xi_1, \ldots, \xi_n)$ be defined as in (1.1) and (1.2). Assume that $(\xi_1, \ldots, \xi_n)$ satisfies:

$$E\{\xi_i\} = 0 \text{ for } 1 \leq i \leq n \text{ and } \sum_{i=1}^{n} E\{\xi_i \xi_i^T\} = \Sigma,$$

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

$$\sup_{A \in A} |P(T \in A) - P(S^{1/2} Z \in A)| \leq 259 \sigma^{-3/2} d^{1/2} \gamma + 2\sigma^{-1} E\{\|W\|\Delta\} + 2\sigma^{-1} \sum_{i=1}^{n} E\{\|\xi_i\|\Delta - \Delta^{(i)}\},$$

for any random variables $\Delta$ and $(\Delta^{(i)})_{1 \leq i \leq n}$ such that $\Delta \geq \|D\|$ 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 $X, X_1, \ldots, X_n$ be i.i.d. random variables with common probability distribution $P$ that take values in a measurable space $(\mathcal{X}, B(\mathcal{X}))$. For any function $f : \mathcal{X} \mapsto \mathbb{R}$, let

$$P_n f = \frac{1}{n} \sum_{i=1}^{n} f(X_i), \quad P f = \int_{\mathcal{X}} f(x) P(dx), \quad G_n f = \sqrt{n}(P_n - P)f.$$  \hfill (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

$$M_n(\theta) = P_n m_\theta, \quad M(\theta) = P m_\theta.$$  \hfill (3.2)

Following the notations in Van der Vaart [32], we briefly write

$$\hat{m}_\theta(x) = \nabla_\theta m_\theta(x), \quad \check{m}_\theta(x) = \nabla^2_\theta m_\theta(x),$$  \hfill (3.3)

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

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

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

$$\hat{\theta}_n = \arg\min_{\theta \in \Theta} M_n(\theta).$$  \hfill (3.5)

For any $p \geq 1$ and $Y \in \mathbb{R}^d$, let $\|Y\|_p = (E\{|Y|^p\})^{1/p}$ 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 [33], Van der Vaart [32] and the references therein for a thorough reference. Under some regularity conditions, one has $\hat{\theta}_n \overset{P}{\to} \theta^*$, 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 $A \preceq (\text{resp.} \succeq) B$ 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 $\|m_1(X)\|_9 \leq c_1$ and $\|m_2(X)\|_4 \leq c_2$, such that for any $\theta \in \Theta$,

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

and

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

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

$$m_{\theta^*}(x) \leq m_3(x)I_d \quad \text{and} \quad \|m_3(X)\|_4 \leq c_3. \tag{3.9}$$

(M2) Let $\xi_i = m_{\theta^*}(X_i) := (\xi_{i,1}, \ldots, \xi_{i,d})^T$, $\Sigma = \mathbb{E}\{\xi\xi^T\}$ and $V = \mathbb{E}\{m_{\theta^*}(X)\}$. Assume that there exist constants $\lambda_1 > 0$ and $\lambda_2 > 0$ such that $\lambda_{\min}(\Sigma) \geq \lambda_1$ and $\lambda_{\min}(V) \geq \lambda_2$. Moreover, assume that there exists a constant $c_4 > 0$ such that

$$\|\xi_i\|_4 \leq c_4d^{1/2}. \tag{3.10}$$

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

$$\sup_{A \in \mathcal{A}} \left| \mathbb{P}(n^{1/2}\Sigma^{-1/2}V(\hat{\theta}_n - \theta^*) \in A) - \mathbb{P}(Z \in A) \right| \leq C d^{\theta/4} n^{-1/2},$$

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 $|m_3(X_i)| \leq c_2$ for each $1 \leq i \leq n$ almost surely, then the assumption for $m_3(x)$ can be replaced by $\|m_1(X)\|_5 \leq c_1$. The condition (3.10) is satisfied if $\|\xi_{ij}\|_4 \leq c_4$ for all $1 \leq i \leq n$ and $1 \leq j \leq d$. 

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

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

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 \( h_\theta : \mathcal{X} \rightarrow \mathbb{R}_+ \), 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

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

The estimator \( \hat{\theta}_n \) in (3.11) is often called a Z-estimator of \( \theta^* \), see e.g., Van der Vaart [32]. 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

\[
\langle \Psi(\theta) - \Psi(\theta^*), \dot{\Psi}_0(\theta - \theta^*) \rangle = o(\|\theta - \theta^*\|) \quad \text{as} \quad \theta \rightarrow \theta^*.
\]

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 \( \dot{\Psi}^{-1}_0 Z \), where \( Z \sim N(0, \mathbb{E}\{h_{\theta^*}(X_i)h_{\theta^*}(X_i)\} \) \). Bentkus, Bloznelis and Götze [6] proved a Berry–Esseen bound of order \( O(n^{-1/2}) \) 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 \( p \geq 3 \) 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

\[
\langle \Psi(\theta_1) - \Psi(\theta_2), \theta_1 - \theta_2 \rangle \geq \mu \|\theta_1 - \theta_2\|^2,
\] (3.12)

and

\[
\|\Psi(\theta) - \Psi(\theta^*) - \dot{\Psi}_0(\theta - \theta^*)\| \leq c_1 \|\theta - \theta^*\|^2, \quad \lambda_{\min}(\dot{\Psi}_0) \geq \lambda_1.
\] (3.13)

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

\[
|h_{\theta,j}(X) - h_{\theta',j}(X)| \leq h_0(X)\|\theta - \theta'\|.
\] (3.14)

and

\[
\|h_0(X)\|_p \leq c_2.
\] (3.15)
Berry–Esseen bounds for multivariate nonlinear statistics

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

$$\lambda_{\text{min}}(\Sigma) \geq \lambda_2,$$

and

$$\|\xi_1\|_p \leq c_3 d^{1/2}. \quad (3.17)$$

**Remark 3.4.** Following notations in Theorem 3.1, we can choose $h_{\theta}(x) = \tilde{m}_{\theta}(x)$. 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 $p \geq 3$ and $D_\Theta := \sup_{\theta_1, \theta_2 \in \Theta} \|\theta_1 - \theta_2\|$, the diameter of the parameter space $\Theta$. Assume that conditions (B1)–(B3) are satisfied. Then,

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

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 [33, Section 2.2] for more details. Let $\psi : [0, \infty) \rightarrow [0, \infty)$ 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

$$\|Y\|_\psi = \inf \left\{ C > 0 : \mathbb{E}\left\{ \psi\left( \frac{\|Y\|}{C} \right) \right\} \leq 1 \right\}. \quad (3.19)$$

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

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

$$\|h_0(X)\|_{\psi_1} \leq c_4. \quad (3.20)$$

where $c_4 > 0$ is a constant.

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

$$\|\xi_1\|_{\psi_1} \leq c_5, \quad (3.21)$$

where $c_5 > 0$ is a constant.

**Remark 3.6.** Let $Y$ be a random variable. It can be shown (see Vershynin [34, (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:
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),

$$
\sup_{A \in A} \left| \mathbb{P}\left( \sqrt{n} \Sigma^{-1/2} \psi_0(\hat{\theta}_n - \theta^*) \in A \right) - \mathbb{P}(Z \in A) \right| \leq C(D_\Theta + 1)^2 d^4 n^{-1/2} \log n,
$$

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 $n \geq 1$, we update $\theta_n$ by

$$
\begin{align*}
\theta_n &= \theta_{n-1} - \ell_n (\nabla f(\theta_{n-1}) + \zeta_n), \\
\hat{\theta}_n &= \frac{1}{n} \sum_{i=0}^{n-1} \theta_i, \\
\ell_n &= \text{learning rate, and } (\zeta_1, \zeta_2, \ldots) \text{ is a sequence of } \mathbb{R}^d \text{-valued martingale differences.}
\end{align*}
$$

where $\ell_n > 0$ is the so called learning rate and $(\zeta_1, \zeta_2, \ldots)$ is a sequence of $\mathbb{R}^d$-valued martingale differences. The convergence rate of $\mathbb{E}\|\theta_n - \theta^*\|^2$ and $\mathbb{E}\|\hat{\theta}_n - \theta^*\|^2$ was thoroughly studied in the literature, see Polyak [23] and Bach and Moulines [2]. The normality of $\sqrt{n}(\hat{\theta}_n - \theta^*)$ 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 $\sqrt{n}(\hat{\theta}_n - \theta^*)$ 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 $\sqrt{n}(\hat{\theta}_n - \theta^*)$.

We make the following assumptions:

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

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

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

where

(i) $(\xi_1, \xi_2, \ldots)$ is a sequence of independent random variables and $\mathbb{E}\{\xi_i\} = 0$ and $\mathbb{E}\{\xi_i \xi_j^\top\} = \Sigma_i$; there exist positive numbers $\lambda_1$ and $\lambda_2$ such that for any $i \geq 1$, $\lambda_1 \xi_i \leq \lambda_{\max}(\Sigma_i) \leq \lambda_2$; moreover, there exists a positive number $\tau$ such that

$$
\max_{1 \leq i \leq n} \|\xi_i\|_4 \leq \tau;
$$
Here, for any \( n \geq 0 \) and \( (\xi) \) let \( \eta_n := g(\theta_{n-1}, \xi_n) \) satisfies \( \mathbb{E}\{\eta_n|\mathcal{F}_{n-1}\} = 0 \) and for any \( \theta \) and \( \theta' \), there exists a nonnegative number \( c_1 \geq 0 \) such that
\[
\|g(\theta, \xi) - g(\theta', \xi)\| \leq c_1 \|\theta - \theta'\| \quad \text{and} \quad g(\theta^*, \xi) = 0 \quad \text{for} \ \xi \in \mathbb{R}^d. \tag{3.23}
\]

(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
\[
\mu I_d \preceq \nabla^2 f(\theta) \preceq LI_d, \quad \text{for all} \ \theta \in \Theta. \tag{3.24}
\]

(C3) There exist positive constants \( c_2 \) and \( \beta \) such that for all \( \theta \) with \( \|\theta - \theta^*\| \leq \beta \),
\[
\|\nabla^2 f(\theta) - \nabla^2 f(\theta^*)\| \leq c_2 \|\theta - \theta^*\|. \tag{3.25}
\]

Let \( G := \nabla^2 f(\theta^*) \). Recall that \( (\ell_n)_{n \geq 1} \) 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 \( n \geq 0 \), set \( \prod_{i=n+1}^n A_i = I_d, \prod_{i=n+1}^n a_i = 1 \), where \( (A_i)_{i \geq 1} \) is a \( \mathbb{R}^{d \times d} \)-valued sequence and \( (a_i)_{i \geq 1} \) is a \( \mathbb{R} \)-valued sequence. Let
\[
\Sigma_n = \frac{1}{n} \sum_{i=1}^{n-1} Q_i \Sigma_i Q_i^\top.
\]

We have the following theorem.

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

1. if \( \alpha \in (1/2, 1) \),
\[
\sup_{A \in \mathcal{A}} \left| \mathbb{P}(\sqrt{n} \Sigma_n^{-1/2}(\theta_n - \theta^*) \in A) - \mathbb{P}(Z \in A) \right| 
\leq C (d^{3/2} + \tau^3 + \tau_0^3) (d^{1/2} n^{-1/2} + n^{-\alpha+1/2}); \tag{3.26}
\]
2. if \( \ell_n = \ell_0 n^{-1} \) with \( \ell_0 \mu \geq 1 \), we have
\[
\sup_{A \in \mathcal{A}} \left| \mathbb{P}(\sqrt{n} \Sigma_n^{-1/2}(\theta_n - \theta^*) \in A) - \mathbb{P}(Z \in A) \right| 
\leq C n^{-1/2} (d^{3/2} + \tau^3 + \tau_0^3) \times \begin{cases} 
d^{1/2} + \log n, \quad \ell_0 \mu > 1; \\
d^{1/2}(\log n)^3, \quad \ell_0 \mu = 1.
\end{cases} \tag{3.27}
\]

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

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

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

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 

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

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

$$
\xi_n = \bar{m}_{\theta^*}(X_n) - \nabla M(\theta^*),
$$

$$
\eta_n = \bar{m}_{\theta_{n-1}}(X_n) - \bar{m}_{\theta^*}(X_n) - \nabla M(\theta_{n-1}) + \nabla M(\theta^*).
$$

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

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

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 $n \geq 1$, 

$$
\bar{m}_{\theta_{n-1}}(X_n) = \nabla M(\theta_{n-1}) + (\bar{m}_{\theta_{n-1}}(X_n) - \nabla M(\theta_{n-1}))
$$

$$
= \nabla M(\theta_{n-1}) + (\bar{m}_{\theta^*}(X_n) - \nabla M(\theta^*))
$$

$$
+ (\bar{m}_{\theta_{n-1}}(X_n) - \bar{m}_{\theta^*}(X_n) - \nabla M(\theta_{n-1}) + \nabla M(\theta^*))
$$

$$
= \nabla M(\theta_{n-1}) + \xi_n + \eta_n.
$$

For $n \geq 0$, let $\mathcal{F}_n = \sigma(\theta_0, X_1, \ldots, X_n)$. Then we have $\mathbb{E}\{\xi_n\} = 0$ and $\mathbb{E}\{\eta_n \mid \mathcal{F}_{n-1}\} = 0$. 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. \qed
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 $A^\varepsilon = \{x \in \mathbb{R}^d : d(x,A) \leq \varepsilon\}$, and $A^{-\varepsilon} = \{x \in A : B(x,\varepsilon) \subset A\}$, where $B(x,\varepsilon)$ is the $d$-dimensional ball 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 $\bar{\gamma} = \sum_{i=1}^{n} \mathbb{E}\{\|\xi_i\|^2\}$ 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 $1 \leq i \leq n$ and $\sum_{i=1}^{n} \mathbb{E}\{\xi_i \xi_j^\top\} = I_d$. Let $\Delta_1$ and $\Delta_2$ be nonnegative random variables. Then we have for all $A \in \mathcal{A}$ such that $r(\bar{A}) > \gamma$,

$$\mathbb{P}(W \in A^{\gamma+\Delta_1} \setminus A^{\gamma-\Delta_2}) \leq 19d^{1/2}\gamma + 2\mathbb{E}\{\|W\|(|\Delta_1 + \Delta_2|)\} + 2\sum_{i=1}^{n} \sum_{j=1}^{2} \mathbb{E}\{\|\xi_i\|\Delta_j - \Delta_j^{(i)}\},$$

(4.1)

where $\Delta_2 = \Delta_2 \wedge (r(\bar{A}) - \gamma)$ and $\Delta_j^{(i)}$ 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

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

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 [30] 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 $\varepsilon_1, \varepsilon_2 \geq 0$,

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

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

For each $1 \leq i \leq n$, let $\Delta^{(i)}$ be any random variable that is independent of $\xi_i$. Note that $\|T - W\| \leq \Delta$ and that $r(A^{2\gamma}) > \gamma$. 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

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

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),

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

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

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

where

$$Q_1 = \mathbb{P}(Z \in A) - \mathbb{P}(Z \in A_0^{4\gamma}),$$

$$Q_2 = \mathbb{P}(Z \in A_0^{4\gamma}) - \mathbb{P}(W \in A_0^{4\gamma}),$$

$$Q_3 = \mathbb{P}(W \in A_0^{4\gamma}) - \mathbb{P}(T \in A_0^{4\gamma}).$$

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

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

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

$$|Q_2| \leq 115d^{1/2}\gamma.$$
Berry–Esseen bounds for multivariate nonlinear statistics

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

$$
\mathbb{I}\{w \in A_0^{4\gamma}\} - \mathbb{I}\{w + D \in A_0^{4\gamma}\} \leq \mathbb{I}\{w \in A_0^{4\gamma} \setminus A_0^{4\gamma - \Delta}\}. \quad (4.6)
$$

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

$$
\mathbb{I}\{w \in A_0^{4\gamma}\} - \mathbb{I}\{w + D \in A_0^{4\gamma}\} \\
\leq \mathbb{I}\{w \in A_0^{4\gamma}\} \\
\leq \mathbb{I}\{w \in A_0^{4\gamma} \setminus A_0^{4\gamma - \Delta}\} + \mathbb{I}\{w \in A_0^{4\gamma - (r - 5\gamma)}\} + \mathbb{I}\{w \in A_0^{4\gamma - (r - 5\gamma)}\}, \\
\leq \mathbb{I}\{w \in A_0^{4\gamma} \setminus A_0^{4\gamma - \Delta}\} + \mathbb{I}\{w \in A_0^{4\gamma - (r - 5\gamma)}\} + \mathbb{I}\{w \in A_0^{4\gamma - (r - 5\gamma)}\}, \\
\text{where the last line follows from the fact that } (A^{-4\gamma})^{9\gamma - r} \subset A_0^{5\gamma - r}. \text{ Equations (4.6) and (4.7) yield}
$$

$$
Q_3 = \mathbb{P}(W \in A_0^{4\gamma}) - \mathbb{P}(W + D \in A_0^{4\gamma}) \\
\leq \mathbb{P}(W \in A_0^{4\gamma} \setminus A_0^{4\gamma - \Delta_0}) + \mathbb{P}(W \in A_0^{5\gamma - r}). \quad (4.8)
$$

For each $1 \leq i \leq n$, let $\Delta_0^{(i)} = \Delta(i) \land (r(A_0) - \gamma)$. For the first term of (4.8), by Proposition 4.1, we have

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

For the second term of (4.8), since $A^{-\gamma - r} = \emptyset$ and $A_0^{5\gamma - r}$ is convex and nonempty, by (4.2) and (4.3), we have

$$
\mathbb{P}(W \in A_0^{5\gamma - r}) \leq |\mathbb{P}(W \in A_0^{5\gamma - r}) - \mathbb{P}(Z \in A_0^{5\gamma - r})| + \mathbb{P}(Z \in A_0^{5\gamma - r} \setminus A^{-\gamma - r}) \\
\leq 115d^{1/2}\gamma + 6d^{1/2}\gamma \leq 121d^{1/2}\gamma.
$$

Then it follows that

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

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

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

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

$$
\sup_{A \in \mathcal{A}} |\mathbb{P}(T \in A) - \mathbb{P}(W \in A)| \leq 259d^{1/2}\gamma + 2\mathbb{E}\{\|W\|\Delta\} + 2\sum_{i=1}^{n}\mathbb{E}\{\|X_i\|\Delta - \Delta^{(i)}\}|,
$$
Proof of Corollary 2.2. Let $\tilde{T} = W + D(1 (O).$ For any $A \in \mathcal{A}$,
\begin{equation*}
\left| \mathbb{P}(T \in A) - \mathbb{P}(\tilde{T} \in A) \right| \leq \mathbb{P}(O^c).
\end{equation*}
Applying Theorem 2.1 to $\tilde{T}$ yields
\begin{equation*}
\sup_{A \in \mathcal{A}} \left| \mathbb{P}(\tilde{T} \in A) - \mathbb{P}(Z \in A) \right| \leq 259d^{1/2}\gamma + 2\mathbb{E}\{\|W\|\Delta\} + 2n \sum_{i=1}^{n} \mathbb{E}\{\|\xi_i\|\Delta \Delta(i)}\}.\end{equation*}
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 $0 \leq t \leq 1$,
\begin{equation}
(4.10) \quad ty_1 + (1-t)y_2 \in \Sigma^{-1/2}A.
\end{equation}
Since $y_1, y_2 \in \Sigma^{-1/2}A$, it follows that there exist $x_1, x_2 \in A$ such that
\begin{equation*}
y_1 = \Sigma^{-1/2}x_1, \quad y_2 = \Sigma^{-1/2}x_2.
\end{equation*}
Moreover, as $A$ is convex, we have for any $0 \leq t \leq 1$,
\begin{equation*}
tx_1 + (1-t)x_2 \in A,
\end{equation*}
and thus
\begin{equation*}
(4.10) \quad ty_1 + (1-t)y_2 = t\Sigma^{-1/2}x_1 + (1-t)\Sigma^{-1/2}x_2 \in \Sigma^{-1/2}A.
\end{equation*}
This proves (4.10) and hence $\Sigma^{-1/2}A$ is convex. Note that
\begin{equation*}
\mathbb{P}(T \in A) - \mathbb{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),
\end{equation*}
and we have
\begin{equation*}
\sup_{A \in \mathcal{A}} \left| \mathbb{P}(T \in A) - \mathbb{P}(\Sigma^{-1/2}A \in A) \right| = \sup_{A \in \mathcal{A}} \left| \mathbb{P}(\Sigma^{-1/2}T \in A) - \mathbb{P}(Z \in A) \right|.
\end{equation*}
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 \( \varepsilon \geq 0 \), we construct \( f_{A,\varepsilon} : \mathbb{R}^d \rightarrow \mathbb{R}^d \) as follows. Let \( P_A \) be the projection operator on \( A \), that is, for any \( x \in \mathbb{R}^d \), let
\[
P_A(x) := \arg\min_{y \in A} \|x - y\|.
\]
Therefore, \( P_A(x) \) is the nearest point of \( x \) in the set \( A \).

Let \( \bar{A} \) be the closure of \( A \), and
\[
f_{A,\varepsilon}(x) = \begin{cases} 
0, & x \in \bar{A}, \\
x - P_{\bar{A}}(x), & x \in A^\varepsilon \setminus \bar{A}, \\
P_{(A^\varepsilon \setminus \bar{A})}(x) - P_A(x), & x \in \mathbb{R}^d \setminus A^\varepsilon. 
\end{cases}
\] (4.11)

Let \( r(\bar{A}) = \max\{y : B(x, y) \subset \bar{A} \text{ for some } x \in \mathbb{R}^d\} \) 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) \( \|f\| \leq \varepsilon + 8\gamma \); 
(ii) for all \( \xi, \eta \in \mathbb{R}^d \), \( \langle \xi, f(\eta + \xi) - f(\eta) \rangle \geq 0 \); 
(iii) for \( w \in A^{4\gamma+\varepsilon} \setminus A^{4\gamma} \) and \( \|x\| \leq 4\gamma \), we have
\[
\langle x, f(w) - f(w - x) \rangle \geq \frac{3}{4} (x \cdot h_1)^2,
\]
where \( h_1 = (w_0 - w)/\|w_0 - w\| \) and \( w_0 = P_{\bar{A}}(w) \).

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 \( \Delta_2 = \Delta_2 \wedge (r - \gamma) \).

Let \( \Delta_1^{(i)} \) and \( \Delta_2^{(i)} \) be any random variables that are independent of \( \xi_i \) and let \( \Delta_2^{(i)} = \Delta_2 \wedge (r - \gamma) \).
For any \( a \geq 0 \) and \( 0 \leq b \leq r - \gamma \), define \( g_{a,b} = f_{A^{a-b},8\gamma+a+b} \). Noting that \( \mathbb{E}\xi_i = 0 \) and observing that \( \Delta_1^{(i)} \) and \( \Delta_2^{(i)} \) are independent of \( \xi_i \), we have
\[
\mathbb{E}\{\langle \xi_i, g_{\Delta_1^{(i)},\Delta_2^{(i)}}(W - \xi_i) \rangle\} = 0,
\]
and thus,
\[
\mathbb{E}\{\langle W, g_{\Delta_1^{(i)},\Delta_2^{(i)}}(W) \rangle\} = \sum_{i=1}^{n} \left( \mathbb{E}\{\langle \xi_i, g_{\Delta_1^{(i)},\Delta_2^{(i)}}(W) \rangle\} - \mathbb{E}\{\langle \xi_i, g_{\Delta_1^{(i)},\Delta_2^{(i)}}(W - \xi_i) \rangle\} \right) = H_1 + H_2,
\] (4.12)
where
\[
H_1 = \sum_{i=1}^{n} \mathbb{E}\{\langle \xi_i, g_{\Delta_1^{(i)},\Delta_2^{(i)}}(W - \xi_i) \rangle\},
\]
and
\[
H_2 = \sum_{i=1}^{n} \mathbb{E}\{\langle \xi_i, g_{\Delta_1^{(i)},\Delta_2^{(i)}}(W - \xi_i) - g_{\Delta_1^{(i)},\Delta_2^{(i)}}(W - \xi_i) \rangle\}.
\]
For the upper bound of $H_2$, by the definition of $f$, we have
\[
\|g_{\Delta_2,\Delta_2}(w) - g_{\Delta_1,\Delta_1}(w)\| \leq \|g_{\Delta_2,\Delta_2}(w) - g_{\Delta_1,\Delta_1}(w)\| + \|g_{\Delta_1,\Delta_1}(w) - g_{\Delta_1,\Delta_1}(w)\|.
\] (4.13)
Without loss of generality, assume that $\Delta_1 = \Delta_2$. Let $A_2 = (A^{\Delta_2}) = A_3 = A^{\Delta_1 + \Delta_1}$, and $w_i = P_{A_j}(w)$ for $j = 2, 3, 4$.

If $w \in A_3 \subset A_4$, then
\[
g_{\Delta_1,\Delta_2}(w) = g_{\Delta_1,\Delta_2}(w);
\]
if $w \in A_4 \setminus A_3$, then
\[
g_{\Delta_1,\Delta_2}(w) = w - w_2, \quad g_{\Delta_1,\Delta_2}(w) = w_3 - w_2,
\]
and
\[
\|w - w_3\| = \|w_4 - w_3\|, \quad \text{for } w \in A_4 \setminus A_3;
\]
if $w \in A_1$, then
\[
g_{\Delta_1,\Delta_2}(w) = w_4 - w_2 \quad \text{and} \quad g_{\Delta_1,\Delta_2}(w) = w_3 - w_2.
\]
By the definition of $w_3$ and $w_4$, it follows that $\|w_4 - w_3\| \leq |\Delta_1 - \Delta_1(i)|$. Hence,
\[
\|g_{\Delta_1,\Delta_2}(w) - g_{\Delta_1,\Delta_2}(w)\| \leq |\Delta_1 - \Delta_1(i)|.
\] (4.14)
Similarly,
\[
\|g_{\Delta_1,\Delta_2}(w) - g_{\Delta_1,\Delta_2}(w)\| \leq |\Delta_2 - \Delta_2(i)| \leq |\Delta_2 - \Delta_2(i)|.
\] (4.15)
By (4.13)–(4.15),
\[
H_2 \leq \sum_{i=1}^{n} \mathbb{E}\{|\Delta_1 - \Delta_1(i)| + |\Delta_2 - \Delta_2(i)|\}. \quad (4.16)
\]
We next estimate the lower bound of $H_1$. By Lemma 4.2, we have
\[
H_1 = \sum_{i=1}^{n} \mathbb{E}\{\langle \xi_i, g_{\Delta_1,\Delta_2}(W) - g_{\Delta_1,\Delta_2}(W - \xi_i)\rangle\}
\] \[
\geq \sum_{i=1}^{n} \mathbb{E}\{\langle \xi_i, g_{\Delta_1,\Delta_2}(W) - g_{\Delta_1,\Delta_2}(W - \xi_i)\rangle \mathbb{I}(|\xi_i| \leq 4\gamma) \mathbb{I}(W \in A^{4\gamma + \Delta_1 \setminus A^{4\gamma - \Delta_2}})\}
\] \[
\geq \frac{3}{4} \sum_{i=1}^{n} \mathbb{E}\{\langle \xi_i, U \rangle^2 \mathbb{I}(|\xi_i| \leq 4\gamma) \mathbb{I}(W \in A^{4\gamma + \Delta_1 \setminus A^{4\gamma - \Delta_2}})\} := \frac{3}{4} R \quad (4.17)
\]
where $U := (W_0 - W) / \|W_0 - W\| = (U_1, \ldots, U_d)$ and $W_0 = P_{A}(W)$. Observe that by (4.17),
\[
R = \sum_{i=1}^{n} \sum_{j=1}^{d} \mathbb{E}\{\xi_i^2 U_j^2 \mathbb{I}(|\xi_i| \leq 4\gamma) \mathbb{I}(W \in A^{4\gamma + \Delta_1 \setminus A^{4\gamma - \Delta_2}})\}$
we have

\[ \text{For } R_1, \text{ rearranging the summations yields} \]

\[ R_1 = \sum_{j=1}^{d} \mathbb{E}\left\{ \mathbb{I}(W \in A^{4\gamma+\Delta_1} \setminus A^{4\gamma-\Delta_2}) U_j^2 \sum_{i=1}^{n} \xi_{ij}^2 \mathbb{I}(\|\xi_i\| \leq 4\gamma) \right\} \]

\[ = \sum_{j=1}^{d} \mathbb{E}\left\{ \mathbb{I}(W \in A^{4\gamma+\Delta_1} \setminus A^{4\gamma-\Delta_2}) U_j^2 \left( \sum_{i=1}^{n} \left( \xi_{ij}^2 \mathbb{I}(\|\xi_i\| \leq 4\gamma) - \mathbb{E}\xi_{ij}^2 \mathbb{I}(\|\xi_i\| \leq 4\gamma) \right) \right) \right\} \]

\[ + \sum_{j=1}^{d} \mathbb{E}\left\{ \mathbb{I}(W \in A^{4\gamma+\Delta_1} \setminus A^{4\gamma-\Delta_2}) U_j^2 \right\} \sum_{i=1}^{n} \mathbb{E}\left\{ \xi_{ij}^2 \mathbb{I}(\|\xi_i\| \leq 4\gamma) \right\} \]

\[ := R_{11} + R_{12}. \]

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

\[ a = U_j^2 \quad \text{and} \quad b = \left| \sum_{i=1}^{n} \left( \xi_{ij}^2 \mathbb{I}(\|\xi_i\| \leq 4\gamma) - \mathbb{E}\xi_{ij}^2 \mathbb{I}(\|\xi_i\| \leq 4\gamma) \right) \right|, \]

we have

\[ |R_{11}| \leq \sum_{j=1}^{d} \mathbb{E}\left\{ U_j^2 \left| \sum_{i=1}^{n} \left( \xi_{ij}^2 \mathbb{I}(\|\xi_i\| \leq 4\gamma) - \mathbb{E}\xi_{ij}^2 \mathbb{I}(\|\xi_i\| \leq 4\gamma) \right) \right| \right\} \]

\[ \leq \gamma \sum_{j=1}^{d} \mathbb{E}\left\{ U_j^4 \right\} + \frac{1}{4\gamma} \sum_{j=1}^{d} \text{Var} \left( \sum_{i=1}^{n} \xi_{ij}^2 \mathbb{I}(\|\xi_i\| \leq 4\gamma) \right) \]

\[ \leq \gamma \sum_{j=1}^{d} \mathbb{E}\left\{ U_j^4 \right\} + \frac{1}{4\gamma} \sum_{j=1}^{d} \sum_{i=1}^{n} \mathbb{E}\left\{ \xi_{ij}^4 \mathbb{I}(\|\xi_i\| \leq 4\gamma) \right\}. \]

As for \( R_{12} \), recalling that \( \sum_{j=1}^{d} U_j^2 = 1 \) and \( \sum_{i=1}^{n} \mathbb{E}\{\xi_i^T\} = I_d \), we have \( \sum_{i=1}^{n} \mathbb{E}\{\xi_{ij}^2\} = 1 \) for each \( 1 \leq j \leq d \), and

\[ R_{12} = \sum_{j=1}^{d} \mathbb{E}\left\{ \mathbb{I}(W \in A^{4\gamma+\Delta_1} \setminus A^{4\gamma-\Delta_2}) U_j^2 \right\} \left( \sum_{i=1}^{n} \mathbb{E}\left\{ \xi_{ij}^2 \mathbb{I}(\|\xi_i\| > 4\gamma) \right\} \right) \]

\[ = \mathbb{P}(W \in A^{4\gamma+\Delta_1} \setminus A^{4\gamma-\Delta_2}) \]

\[ - \mathbb{E}\left\{ \mathbb{I}(W \in A^{4\gamma+\Delta_1} \setminus A^{4\gamma-\Delta_2}) \sum_{j=1}^{d} U_j^2 \left( \sum_{i=1}^{n} \mathbb{E}\left\{ \xi_{ij}^2 \mathbb{I}(\|\xi_i\| > 4\gamma) \right\} \right) \right\}. \]
By (4.18) and (4.19), it follows that

\[ R_1 \geq \mathbb{P}(W \in A^{4\gamma + \Delta_1} \setminus A^{4\gamma - \Delta_2}) - \gamma \sum_{j=1}^{d} \mathbb{E}\{U_j^4\} - \frac{1}{4\gamma} \sum_{i=1}^{n} \sum_{j=1}^{d} \mathbb{E}\{\xi_{ij}^4 \mathbb{1}(\|\xi_i\| \leq 4\gamma)\} \]

\[ - \mathbb{E}\left\{ \mathbb{1}(W \in A^{4\gamma + \Delta_1} \setminus A^{4\gamma - \Delta_2}) \sum_{j=1}^{d} U_j^2 \left( \sum_{i=1}^{n} \mathbb{E}\{\xi_{ij}^2 \mathbb{1}(\|\xi_i\| > 4\gamma)\} \right) \right\}. \]

Similarly, noting that \( \sum_{i=1}^{n} \mathbb{E}\{\xi_i \xi_{ij}^2\} = 0 \) for \( j \neq j' \), we have

\[ R_2 \geq -\gamma \sum_{j \neq j'} \mathbb{E}\{U_j^2 U_{j'}^2\} - \frac{1}{4\gamma} \sum_{i=1}^{n} \sum_{1 \leq j \neq j' \leq d} \mathbb{E}\{(\xi_{ij} \xi_{ij'})^2 \mathbb{1}(\|\xi_i\| \leq 4\gamma)\} \]

\[ - \sum_{i=1}^{n} \sum_{1 \leq j \neq j' \leq d} \mathbb{E}\left\{ \mathbb{1}(W \in A^{4\gamma + \Delta_1} \setminus A^{4\gamma - \Delta_2}) U_j U_{j'} \mathbb{E}\{\xi_{ij} \xi_{ij'} \mathbb{1}(\|\xi_i\| \leq 4\gamma)\} \right\}. \]

\[ = -\gamma \sum_{j \neq j'} \mathbb{E}\{U_j^2 U_{j'}^2\} - \frac{1}{4\gamma} \sum_{i=1}^{n} \sum_{1 \leq j \neq j' \leq d} \mathbb{E}\{(\xi_{ij} \xi_{ij'})^2 \mathbb{1}(\|\xi_i\| \leq 4\gamma)\} \]

\[ - \sum_{i=1}^{n} \sum_{1 \leq j \neq j' \leq d} \mathbb{E}\left\{ \mathbb{1}(W \in A^{4\gamma + \Delta_1} \setminus A^{4\gamma - \Delta_2}) U_j U_{j'} \mathbb{E}\{\xi_{ij} \xi_{ij'} \mathbb{1}(\|\xi_i\| > 4\gamma)\} \right\}. \]

\[ = -\gamma \sum_{j \neq j'} \mathbb{E}\{U_j^2 U_{j'}^2\} - \frac{1}{4\gamma} \sum_{i=1}^{n} \sum_{1 \leq j \neq j' \leq d} \mathbb{E}\{(\xi_{ij} \xi_{ij'})^2 \mathbb{1}(\|\xi_i\| \leq 4\gamma)\} \]

\[ - \sum_{i=1}^{n} \sum_{1 \leq j \neq j' \leq d} \mathbb{E}\left\{ \mathbb{1}(W \in A^{4\gamma + \Delta_1} \setminus A^{4\gamma - \Delta_2}) U_j U_{j'} \mathbb{E}\{\xi_{ij} \xi_{ij'} \mathbb{1}(\|\xi_i\| > 4\gamma)\} \right\}. \]

Observe that

\[ \sum_{i=1}^{n} \mathbb{E}\{\|\xi_i\|^4 \mathbb{1}(\|\xi_i\| \leq 4\gamma)\} \leq 4\gamma \sum_{i=1}^{n} \mathbb{E}\|\xi_i\|^3 \leq 4\gamma^2 \]

and by the Markov inequality,

\[ \sum_{i=1}^{n} \mathbb{E}\{\|\xi_i\|^2 \mathbb{1}(\|\xi_i\| > 4\gamma)\} \leq \frac{1}{4\gamma} \sum_{i=1}^{n} \mathbb{E}\|\xi_i\|^3 = \frac{1}{4}. \]
Recall that $\|U\| = 1$ and $\gamma = \sum_{i=1}^{n} E\{|\xi_i|^3\}$, and thus (4.20)–(4.23) yield

\[
R \geq \mathbb{P}(W \in A^{4\gamma+\bar{\Delta}_1} \setminus A^{4\gamma-\bar{\Delta}_2}) - \gamma \mathbb{E}\{\|U\|^4\} - \frac{1}{4\gamma} \sum_{i=1}^{n} \mathbb{E}\{\|\xi_i\|^4 1(\|\xi_i\| \leq 4\gamma)\}
\]

\[
- \sum_{i=1}^{n} \mathbb{E}\{1(W \in A^{4\gamma+\bar{\Delta}_1} \setminus A^{4\gamma-\bar{\Delta}_2}) \mathbb{E}\{\left(\sum_{j=1}^{d} U_j \xi_{ij}\right)^2 1(\|\xi_i\| > 4\gamma)\}\}
\]

\[
\geq \mathbb{P}(W \in A^{4\gamma+\bar{\Delta}_1} \setminus A^{4\gamma-\bar{\Delta}_2}) - \gamma \mathbb{E}\{\|U\|^4\} - \frac{1}{4\gamma} \sum_{i=1}^{n} \mathbb{E}\{\|\xi_i\|^4 1(\|\xi_i\| \leq 4\gamma)\}
\]

\[
- \sum_{i=1}^{n} \mathbb{E}\{1(W \in A^{4\gamma+\bar{\Delta}_1} \setminus A^{4\gamma-\bar{\Delta}_2}) \mathbb{E}\{\|\xi_i\|^2 1(\|\xi_i\| > 4\gamma)\}\}
\]

\[
\geq \mathbb{P}(W \in A^{4\gamma+\bar{\Delta}_1} \setminus A^{4\gamma-\bar{\Delta}_2}) - 2\gamma - \frac{1}{4} \mathbb{P}(W \in A^{4\gamma+\bar{\Delta}_1} \setminus A^{4\gamma-\bar{\Delta}_2})
\]

\[
= \frac{3}{4} \mathbb{P}(W \in A^{4\gamma+\bar{\Delta}_1} \setminus A^{4\gamma-\bar{\Delta}_2}) - 2\gamma.
\]

By (4.17) and (4.24), we have

\[
H_1 \geq \frac{9}{16} \mathbb{P}(W \in A^{4\gamma+\bar{\Delta}_1} \setminus A^{4\gamma-\bar{\Delta}_2}) - \frac{3\gamma}{2}.
\]

On the other hand, note that $\mathbb{E}\|W\|^2 = d$ and by Lemma 4.2, $\|g_{\bar{\Delta}_1, \bar{\Delta}_2}(W)\| \leq (\Delta_1 + \Delta_2 + 8\gamma)$. Thus,

\[
|\mathbb{E}\langle W, g_{\bar{\Delta}_1, \bar{\Delta}_2}(W) \rangle| \leq \mathbb{E}\|W\|(\Delta_1 + \Delta_2) + 8d^{1/2}\gamma.
\]

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

\[
\mathbb{P}(W \in A^{4\gamma+\bar{\Delta}_1} \setminus A^{4\gamma-\bar{\Delta}_2})
\]

\[
\leq 2\mathbb{E}\|W\|(\Delta_1 + \Delta_2) + 16d^{1/2}\gamma + 3\gamma + 2 \sum_{i=1}^{n} E|\xi_i| \sum_{j=1}^{2} |\Delta_j - \Delta_j^{(i)}|
\]

\[
\leq 2\mathbb{E}\|W\|(\Delta_1 + \Delta_2) + 16d^{1/2}\gamma + \sum_{i=1}^{n} E|\xi_i| \sum_{j=1}^{2} |\Delta_j - \Delta_j^{(i)}|.
\]

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 $M_n(\theta)$, and $m_\theta$ is smooth for $\theta$. By the Taylor expansion, it follows that

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

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

$$V(\hat{\theta}_n - \theta^*) = -\frac{1}{n} \sum_{i=1}^{n} \xi_i - \left(\frac{1}{n} \sum_{i=1}^{n} \tilde{m}_{\theta^*}(X_i) - V\right)(\hat{\theta}_n - \theta^*)$$

$$- \frac{1}{n} \sum_{i=1}^{n} \int_{0}^{1} (\tilde{m}_{\theta_t}(X_i) - \tilde{m}_{\theta^*}(X_i))(\hat{\theta}_n - \theta^*) dt,$$

Let

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

and

$$D = -\sqrt{n} \Sigma^{-1/2} \left(\frac{1}{n} \sum_{i=1}^{n} \tilde{m}_{\theta^*}(X_i) - V\right)(\hat{\theta}_n - \theta^*)$$

$$- \sqrt{n} \Sigma^{-1/2} \left(\frac{1}{n} \sum_{i=1}^{n} \int_{0}^{1} (\tilde{m}_{\theta_t}(X_i) - \tilde{m}_{\theta^*}(X_i)) dt\right)(\hat{\theta}_n - \theta^*).$$

Then, we have

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

By (M1) and (M2), we have

$$\|D\| \leq n^{1/2} \lambda_1^{-1/2} \left(\{H_1\|\hat{\theta}_n - \theta^\ast\| + H_2\|\hat{\theta}_n - \theta^\ast\|^2\right), \quad (5.2)$$

where

$$H_1 = \left\|\frac{1}{n} \sum_{i=1}^{n} \left(\tilde{m}_{\theta^*}(X_i) - \mathbb{E}\{\tilde{m}_{\theta^*}(X_i)\}\right)\right\|, \quad H_2 = \frac{1}{n} \sum_{i=1}^{n} \|m_2(X_i)\|.$$

Let $\Delta = n^{1/2} \lambda_1^{-1/2} \left(H_1\|\hat{\theta}_n - \theta^\ast\| + H_2\|\hat{\theta}_n - \theta^\ast\|^2\right)$, and it follows that $\|T - W\| \leq \Delta$. Let $(X'_1, \ldots, X'_n)$ be an independent copy of $(X_1, \ldots, X_n)$, and define

$$X'_{j(i)} = \begin{cases} X_j, & j \neq i; \\ X'_i, & j = i. \end{cases}$$
Moreover, let
\[
H_1^{(i)} = \left\| \frac{1}{n} \sum_{j=1}^{n} (\bar{m}_{\theta^*}(X_j^{(i)}) - E\{\bar{m}_{\theta^*}(X_j)\}) \right\|, \quad H_2^{(i)} = \frac{1}{n} \sum_{j=1}^{n} \|m_{2}(X_j^{(i)})\|
\]
\[
\hat{\theta}_n^{(i)} = \arg\min_{\theta \in \Theta} \frac{1}{n} \sum_{j=1}^{n} m_{\theta}(X_j^{(i)}), \quad \Delta^{(i)} = n^{1/2} \lambda_1^{-1} n^{-1/2} \left( H_1^{(i)} \|\hat{\theta}_n^{(i)} - \theta^\ast\| + H_2^{(i)} \|\hat{\theta}_n^{(i)} - \theta^\ast\| \right).
\]

Then, \( \Delta^{(i)} \) is independent of \( X_i \) and \( \xi_i \). Theorem 3.1 now follows directly from Theorem 2.1 and the following proposition, whose proof can be found in Section A of the Supplementary Material [29].

**Proposition 5.1.** Assume that conditions (M1) and (M2) are satisfied. Then, we have
\[
\sum_{i=1}^{n} E\|\Sigma^{-1/2} \xi_i\|^3 \leq C \delta^{3/2} n, \quad (5.3)
\]
\[
E\{\|W\|\Delta\} \leq C \delta^{13/8} n^{-1/2}, \quad (5.4)
\]
\[
\sum_{i=1}^{n} E\|\Sigma^{-1/2} \xi_i\|\|\Delta - \Delta^{(i)}\| \leq C \delta^{3/4}, \quad (5.5)
\]
where \( C > 0 \) is a constant depending only on \( \lambda_1, \lambda_2, c_1, c_2, c_3, c_4 \) and \( \mu \).

### 5.2. Proof of Theorem 3.2

Let \( \delta_n = (D_\Theta + 1)dn^{-(p-2)/(2p-2)} \), where \( D_\Theta \) is the diameter of the parameter space \( \Theta \). As \( p \geq 3 \), it follows that \( \delta_n \geq n^{-1/2} \). 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 \( \sqrt{n} \Sigma^{-1/2} \hat{\Psi}_0(\hat{\theta}_n - \theta^\ast) \) as a summation of a linear statistic plus an error term, and then apply Corollary 2.2 to prove (5.11). To this end, by (3.11),
\[
\sqrt{n}(\hat{\Psi}(\hat{\theta}_n) - \Psi(\theta^\ast)) = \sqrt{n}(\hat{\Psi}(\hat{\theta}_n) - \Psi_{n}(\hat{\theta}_n)) = -\sqrt{n}(\Psi_{n}(\theta^\ast) - \Psi(\theta^\ast)) - \left( \sqrt{n}(\Psi_{n}(\hat{\theta}_n) - \hat{\Psi}(\hat{\theta}_n)) - \sqrt{n}(\Psi_{n}(\theta^\ast) - \Psi(\theta^\ast)) \right).
\]

By (5.6), we obtain
\[
T := \sqrt{n} \Sigma^{-1/2} \hat{\Psi}_0(\hat{\theta}_n - \theta^\ast) = W + D,
\]
where
\[
W = -\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Sigma^{-1/2} \xi_i, \quad D = -\Sigma^{-1/2} \left( \sqrt{n}(\Psi_{n}(\hat{\theta}_n) - \hat{\Psi}(\hat{\theta}_n)) - \sqrt{n}(\Psi_{n}(\theta^\ast) - \Psi(\theta^\ast)) \right).
\]
Let Proposition 5.2.

By (3.13) and (3.16),

$$\|D\|_1(\|\hat{\theta}_n - \theta^*\| \leq \delta_n) \leq \Delta_1 + \Delta_2,$$

where

$$\Delta_1 = c_2 \lambda_2^{-1/2} \sqrt{n} \sup_{\theta: \|\theta - \theta^*\| \leq \delta_n} \|\Psi_n(\theta) - \Psi(\theta) - (\Psi_n(\theta^*) - \Psi(\theta^*))\|,$$

$$\Delta_2 = c_1 \lambda_2^{-1/2} \sqrt{n} \|\hat{\theta}_n - \theta^*\| \Psi(\hat{\theta}_n - \theta^*) \leq \delta_n).$$

Now we construct random variables $\Delta_1^{(i)}$ and $\Delta_2^{(i)}$ that are independent of $\xi_i$. Let $(X'_1, \ldots, X'_n)$ be an independent copy of $(X_1, \ldots, X_n)$ and let

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

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

$$\Delta_1^{(i)} = c_2 \lambda_2^{-1/2} \sqrt{n} \sup_{\theta: \|\theta - \theta^*\| \leq \delta_n} \|\Psi_n^{(i)}(\theta) - \Psi(\theta) - (\Psi_n^{(i)}(\theta^*) - \Psi(\theta^*))\|,$$

$$\Delta_2^{(i)} = c_1 \lambda_2^{-1/2} \sqrt{n} \|\hat{\theta}_n^{(i)} - \theta^*\| \Psi(\hat{\theta}_n^{(i)} - \theta^*) \leq \delta_n).$$

To apply Corollary 2.2, we need to develop the following proposition, whose proof is given in Subection B.1 of the Supplementary Material [29].

**Proposition 5.2.** Let $B_\delta = \{\theta \in \Theta : \|\theta - \theta^*\| \leq \delta\}$. Under the conditions (B1)–(B3),

$$\mathbb{P}(\hat{\theta}_n \in B_\delta^c) \leq C(D_\Theta + 1)^p d^p \delta_n^{p-1/2},$$

$$\sum_{i=1}^{n} \mathbb{E}||\Sigma^{-1/2} \xi_i||^2 \leq C d^{3/2} n,$$

$$\mathbb{E}\{\|W\|_2(\Delta_1 + \Delta_2)\} \leq C d^{3/2} \delta_n + C(D_\Theta + 1)^2 d^{5/2} n^{-1/2},$$

$$\sum_{i=1}^{n} \sum_{j=1}^{2} \mathbb{E}\{||\xi_i||_2(\|\Delta_j - \Delta_j^{(i)}\|)\} \leq C\Big((D_\Theta + 1)^p d^{p+1/2} \delta_n^{-p+2} n^{-\frac{p-2}{2}}$$

$$+ (D_\Theta + 1)^2 d^{3/2} n^{1/2} \delta_n\Big).$$
By Corollary 2.2 with \( O = \{ \| \hat{\theta}_n - \theta^* \| \leq \delta_n \} \) and Proposition 5.2,
\[
\sup_{A \in A} |\mathbb{P} \left( \sqrt{n} \Sigma^{-1/2} \hat{\Psi}_0 (\hat{\theta}_n - \theta^*) \in A \right) - \mathbb{P}(Z \in A)| \\
\leq C d^{1/2} n^{-3/2} \sum_{i=1}^{n} \mathbb{E} \| \Sigma^{-1/2} \xi_i \|^3 + C \mathbb{E} \{ \| W \| (\Delta_1 + \Delta_2) \} \\
+ C n^{-1/2} \sum_{i=1}^{n} \mathbb{E} \{ \| \xi_i \| (\Delta_1 + \Delta_2 - \Delta_1^{(i)} - \Delta_2^{(i)}) \} + \mathbb{P}(\| \hat{\theta}_n - \theta^* \| > \delta_n)
\]
(5.11)

Recall that \( p \geq 3 \), and then \( \delta_n^2 n \geq 1 \). Therefore,
\[
\text{RHS of (5.11)} \leq C n^{-1/2} (D_\Theta + 1)^2 d^{5/2} + C (D_\Theta + 1)^2 d^{5/2} \delta_n \\
+ C (D_\Theta + 1)^2 d^{p+1/2} \delta_n^{-p} n^{-((p-2)/2)} \\
\leq C (D_\Theta + 1)^2 \left( d^{5/2} n^{-1/2} + d^{7/2} n^{-2(p-2)/(p-1)} \right) \\
\leq C (D_\Theta + 1)^2 d^{7/2} n^{-1/2 + \varepsilon_p},
\]
where \( \varepsilon_p = 1/(2p - 2) \). This proves Theorem 3.2.

5.3. Proof of Theorem 3.3

Theorem 3.3 follows from the proof of Theorem 3.2 and the following proposition, which is proved in Subsection B.2 of the Supplementary Material [29].

Proposition 5.3. Under the conditions (B1), (B4) and (B5), we have
\[
\mathbb{P}(\hat{\theta}_n \in B_{\delta_n}^\varepsilon) \leq C \exp \left( - \frac{C' \sqrt{n} \delta_n}{(D_\Theta + 1) d^{3/2}} \right),
\]
\[
\sum_{i=1}^{n} \mathbb{E} \| \Sigma^{-1/2} \xi_i \|^3 \leq C d^{3/2} n \\
\mathbb{E} \{ \| W \| (\Delta_1 + \Delta_2) \} \leq C d^{3/2} \delta_n + C (D_\Theta + 1)^2 d^{5/2} n^{-1/2},
\]
\[
\sum_{i=1}^{n} \mathbb{E} \{ \| \xi_i \| \Delta_1 + \Delta_2 - \Delta_1^{(i)} - \Delta_2^{(i)} \} \leq C (D_\Theta + 1)^2 d^{5/2} \exp \left( - \frac{C' \sqrt{n} \delta_n}{4(D_\Theta + 1) d^{3/2}} \right) \\
+ C (D_\Theta + 1)^2 d^2 + C (D_\Theta + 1) d^{5/2} n^{1/2} \delta_n,
\]
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 5.3, we have

\[ \sup_{A \in \mathcal{A}} |\mathbb{P}(\sqrt{n}\Sigma^{-1/2}\tilde{\Psi}_0(\hat{\theta}_n - \theta^*) \in A) - \mathbb{P}(Z \in A)| \]

\[ \leq C(D_\Theta + 1)^2 \frac{d}{2}n^{-1/2} + C(D_\Theta + 1)d^{5/2}\delta_n + C(D_\Theta + 1)^2d^{5/2} \exp\left(-\frac{C'\sqrt{n}\delta_n}{4(D_\Theta + 1)d^{3/2}}\right). \]

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 5.3.

### 5.4. Proof of Theorem 3.4

In this subsection, we denote by \( C, C_1, C_2, \ldots \) 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 \( \gamma_0 \). Without loss of generality, we assume that \( n \geq 4\{(2L\ell_0)^\alpha + 1\}; otherwise, the bound (3.26) is trivial. Let \( L_1 := \max\{c_2, 2L/\beta\} \) and \( L_2 := c_1 + L \). We introduce the following family of functions: Let \( \varphi_\beta : \mathbb{R}_+ \setminus \{0\} \mapsto \mathbb{R} \) be given by

\[ \varphi_\beta(t) = \begin{cases} \frac{t^\beta - 1}{\beta}, & \text{if } \beta \neq 0, \\ \log t, & \text{if } \beta = 0. \end{cases} \quad (5.12) \]

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),

\[ \nabla f(\theta) = \nabla^2 f(\theta^*)(\theta - \theta^*) + H(\theta), \quad (5.13) \]

where

\[ H(\theta) = \nabla f(\theta) - \nabla^2 f(\theta^*)(\theta - \theta^*) \]

\[ = \nabla f(\theta) - \nabla f(\theta^*) - \nabla^2 f(\theta^*)(\theta - \theta^*) \]

\[ = \int_0^1 \{\nabla^2 f(\theta^* + t(\theta - \theta^*)) - \nabla^2 f(\theta^*)\}(\theta - \theta^*)dt. \]

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

\[ ||H(\theta)|| \mathbb{I}(||\theta - \theta^*|| \leq \beta) \leq c_2||\theta - \theta^*||^2, \]

and

\[ ||H(\theta)|| \mathbb{I}(||\theta - \theta^*|| > \beta) \leq 2L||\theta - \theta^*|| \mathbb{I}(||\theta - \theta^*|| > \beta) \leq \frac{2L}{\beta}||\theta - \theta^*||^2. \]

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

\[ ||H(\theta)|| \leq L_1||\theta - \theta^*||^2. \quad (5.14) \]

Recall that \( G := \nabla^2 f(\theta^*) \), and it follows from (3.22) and (5.13) that for any \( n \geq 1 \),

\[ \theta_n = \theta_{n-1} - \ell_n(\nabla f(\theta_{n-1}) + \zeta_n) \]

\[ = \theta_{n-1} - \ell_n(G(\theta_{n-1} - \theta^*) + \xi_n + \eta_n + H(\theta_{n-1})). \quad (5.15) \]
By definition, \((\tilde{\theta}_n - \theta^*) = n^{-1} \sum_{i=0}^{n-1} (\theta_i - \theta^*)\). Solving the recursive system (5.15) yields
\[
\sqrt{n}(\tilde{\theta}_n - \theta^*) = \frac{1}{\sqrt{n} \ell_0} Q_0(\theta_0 - \theta^*) + \frac{1}{\sqrt{n}} \sum_{i=1}^{n-1} Q_i (\xi_i + \eta_i + H(\theta_{i-1})) ,
\]
where \(Q_i = \ell_i \sum_{j=i}^{n-1} \prod_{k=i+1}^{j} (I - \ell_k G)\).

Recall that \(\Sigma_n := n^{-1} \sum_{i=1}^{n-1} Q_i \Sigma_i Q_i^T\). Let
\[
T_n = \frac{1}{\sqrt{n} \ell_0} \sum_{i=0}^{n-1} (\theta_i - \theta^*), \quad \zeta_i = \frac{1}{\sqrt{n} \Sigma_n^{-1/2}} \xi_i, \quad W_n = \sum_{i=1}^{n-1} \zeta_i
\]
and
\[
D_n = \frac{1}{\sqrt{n} \ell_0} \sum_{i=0}^{n-1} \Sigma_n^{-1/2} Q_0 (\theta_0 - \theta^*) + \frac{1}{\sqrt{n} n} \sum_{i=1}^{n-1} \xi_i \eta_i + \frac{1}{\sqrt{n} n} \sum_{i=1}^{n-1} \xi_i H(\theta_{i-1})
\]
\[
:= D_{1,n} + D_{2,n} + D_{3,n}.
\]

It is easy to show that
\[
\mathbb{E} W_n = 0, \quad \text{Var}(W_n) = I_d.
\]

and
\[
T_n = W_n + D_n.
\]

Also,
\[
\|D_n\| \leq n^{-1/2} \ell_0^{-1} \|\Sigma_n^{-1/2}\| \cdot \|Q_0\| \cdot \|\theta_0 - \theta^*\|
\]
\[
+ n^{-1/2} \|\Sigma_n^{-1/2}\| \left\| \sum_{i=1}^{n-1} Q_i \eta_i \right\| + n^{-1/2} \|\Sigma_n^{-1/2}\| \left\| \sum_{i=1}^{n-1} \|Q_i H(\theta_{i-1})\|^2 \right\|
\]

The following proposition provides the bounds of \(Q_j\) and \(\Sigma_n^{-1}\), whose proof can be found in Section C of the Supplementary Material [29].

**Proposition 5.4.** Suppose that \(n \geq 4 \{ (2L\ell_0)^{\alpha} + 1 \} \). If \(\ell_1 = \ell_0 i^{-\alpha}\) with \(1/2 < \alpha \leq 1\), then there exists a sequence \((p_i)_{i \geq 1}\), 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 \(0 \leq i \leq n - 1\),
\[
\Sigma_n^{-1} \preceq C_1 I_d,
\]
\[
- p_i I_d \preceq Q_i \preceq p_i I_d,
\]
where
\[
p_i \leq \begin{cases} C_2, & \text{if } (\alpha = 1, \ell_0 \mu > 1) \text{ or } (\alpha \in (1/2, 1)); \\ C_2 \log n, & \text{if } (\alpha = 1, \ell_0 \mu = 1). \end{cases}
\]
Let \((\xi_1', \ldots, \xi_n')\) be an independent copy of \((\xi_1, \ldots, \xi_n)\). For each \(1 \leq i \leq n-1\), 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 \(\theta_j^{(i)} = \theta_j^{(i-1)} - \ell_j (\nabla f(\theta_j^{(i-1)}) + \xi_j^{(i)})\), where \(\eta_j^{(i)} = g(\theta_j^{(i-1)}, \xi_j^{(i)})\).

(c) If \(j > i\), let \(\theta_j^{(i)} = \theta_j^{(i-1)} - \ell_j (\nabla f(\theta_j^{(i-1)}) + \xi_j + \eta_j^{(i)})\), where \(\eta_j^{(i)} = g(\theta_j^{(i-1)}, \xi_j)\).

Secondly, let

\[
D_{2,n}^{(i)} = n^{-1/2} \sum_{j=1}^{n-1} Q_j \eta_j^{(i)},
\]

\[
D_{3,n}^{(i)} = n^{-1/2} \sum_{j=1}^{n-1} Q_j H(\theta_j^{(i)}).
\]

Then, we have for each \(1 \leq i \leq n-1\), \(D_{2,n}^{(i)}\) 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 \(\Delta_3 = C_1 L_1 n^{-1/2} \sum_{i=1}^{n-1} p_i \|\theta_i - \theta^*\|^2\), where \(C_1\) is given as in (5.18). By (5.14), it follows that \(\|D_{3,n}\| \leq \Delta_3\). Also, for each \(1 \leq i \leq n-1\), define

\[
\Delta_1^{(i)} = \|D_{1,n}^{(i)}\|,
\]

\[
\Delta_2^{(i)} = \|D_{2,n}^{(i)}\|,
\]

\[
\Delta_3^{(i)} = C_1 L_1 n^{-1/2} \sum_{j=1}^{n-1} p_j \|\theta_j^{(i)} - \theta^*\|^2.
\]

Clearly, \(\Delta_1^{(i)}, \Delta_2^{(i)}\) and \(\Delta_3^{(i)}\) are independent of \(\xi_i\) for each \(1 \leq i \leq n-1\). The following proposition, whose proof is put in supplementary material [29], provides the bounds of the moments for \(\Delta_j\) and \(\Delta_j - \Delta_1^{(i)}\), \(j = 1, 2, 3\).

**Proposition 5.5.** We have \(\Delta_1\) is independent of \((\xi_1, \ldots, \xi_n)\) and

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

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

\[
\mathbb{E}\{\Delta_2\|W\|\} \leq C d^{1/2} (\tau + \tau_0) n^{-\alpha/2},
\]

\[
\mathbb{E}\{\Delta_3\|W\|\} \leq C d^{1/2} (\tau^2 + \tau_0^2) n^{-\alpha+1/2}.
\]

and

\[
\sum_{i=1}^{n-1} \mathbb{E}\{\|\Delta_2 - \Delta_1^{(i)}\| \|\xi_i\|\} \leq C (\tau^2 + \tau_0^2) n^{-\alpha+1/2},
\]
By Proposition 5.5, we have
\[
\sum_{i=1}^{n-1} \mathbb{E}\{ |\Delta_3 - \Delta_3^{(i)}| \|\xi_i\| \} \leq C(\tau^3 + \tau_0^3)n^{-\alpha/2}.
\]

2. For \( \alpha = 1 \),
\[
\mathbb{E}\{ \Delta_2 \|W\| \} \leq \begin{cases} 
C d^{1/2}(\tau + \tau_0)n^{-1/2}(\log n)^{1/2}, & \ell_0\mu > 1; \\
C d^{1/2}(\tau + \tau_0)n^{-1/2}(\log n)^2, & \ell_0\mu = 1.
\end{cases}
\]
\[
\mathbb{E}\{ \Delta_3 \|W\| \} \leq \begin{cases} 
C d^{1/2}(\tau^2 + \tau_0^2)n^{-1/2}(\log n), & \ell_0\mu > 1; \\
C d^{1/2}(\tau^2 + \tau_0^2)n^{-1/2}(\log n)^{5/2}, & \ell_0\mu = 1,
\end{cases}
\]
\[
\sum_{i=1}^{n-1} \mathbb{E}\{ |\Delta_2 - \Delta_2^{(i)}| \|\xi_i\| \} \leq C(\tau^2 + \tau_0^2) \times \begin{cases} 
n^{-1/2}, & \ell_0\mu > 1; \\
n^{-1/2}(\log n)^{5/2}, & \ell_0\mu = 1.
\end{cases}
\]

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

Now, we apply Theorem 2.1 to prove the Berry–Esseen bound for \( \sqrt{n\Sigma_n^{-1/2}(\theta_n - \theta^*)} \).

(1). For \( 1/2 < \alpha < 1 \). Firstly, by Proposition 5.4 and condition (C1), we have
\[
\sum_{i=1}^{n-1} \mathbb{E}\|\xi_i\|^3 \leq Cn^{-3/2} \sum_{i=1}^{n-1} \mathbb{E}\|\xi_i\|^3 \leq Cn^{-1/2}\tau^3.
\]
(5.17)

By Proposition 5.5, we have
\[
\mathbb{E}\{ \|W\|\Delta \} \leq C(d^{3/2} + \tau^3 + \tau_0^3)n^{-\alpha+1/2},
\]
\[
\sum_{i=1}^{n-1} \mathbb{E}\{ \|\xi_i\| \cdot |\Delta - \Delta^{(i)}| \} \leq C(d^{3/2} + \tau^3 + \tau_0^3)n^{-\alpha/2}.
\]
(5.18)

Substituting (5.17) and (5.18) to Theorem 2.1 yields (3.26).

(2). For \( \alpha = 1 \). By the definition of \( \zeta_i \) and by (5.16),
\[
\sum_{i=1}^{n-1} \mathbb{E}\|\zeta_i\|^3 \leq Cn^{-3/2} \sum_{i=1}^{n-1} \rho_i^3 \mathbb{E}\|\xi_i\|^3
\]
\[
\leq \begin{cases} 
C\tau^3n^{-1/2}, & \text{if } \ell_0\mu > 1, \\
C\tau^3n^{-1/2}(\log n)^3, & \text{if } \ell_0\mu = 1.
\end{cases}
\]
(5.19)

By Proposition 5.5, we have
\[
\mathbb{E}\{ \Delta \|W\| \} \leq C(d + \tau^2 + \tau_0^2)n^{-1/2}
\]
\[
+ C(d^{3/2} + \tau^3 + \tau_0^3) \times \begin{cases} 
n^{-1/2}(\log n), & \ell_0\mu > 1; \\
n^{-1/2}(\log n)^{5/2}, & \ell_0\mu = 1.
\end{cases}
\]
(5.20)
Define \( \Delta^{(i)} = \Delta_1^{(i)} + \Delta_2^{(i)} + \Delta_3^{(i)} \), 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 5.5, we have

\[
\sum_{i=1}^{n-1} \mathbb{E}\{ |\Delta - \Delta^{(i)}| \| \zeta_i \| \}
\leq C \left( d^{3/2} + \tau_3^2 + \tau_0^3 \right) \times \begin{cases} n^{-1/2}, & \ell_0 \mu > 1; \\ n^{-1/2} (\log n)^{5/2}, & \ell_0 \mu = 1. \end{cases}
\tag{5.21}
\]

Then the bound (3.27) follows from Theorem 2.1 and (5.19)–(5.21).

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–Esséen 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. *Automatica 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 estimations of the rate of convergence in the multi-dimensional central limit theorem. *Theory Probab. its Appl.* **25**(4) 745–759.

[29] Shao, Q.-M. and Zhang, Z.-S. (2021). Supplement to “Berry–Esseen Bounds for Multivariate Nonlinear Statistics with Applications to M-estimators and Stochastic Gradient Descent Algorithms”.

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

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

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

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

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

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