

Assessing the multivariate normal approximation of the maximum likelihood estimator from high-dimensional, heterogeneous data

Andreas Anastasiou*

Department of Statistics, London School of Economics, Columbia House, Houghton Street, London, WC2A 2AE
e-mail: a.anastasiou@lse.ac.uk

Abstract: The asymptotic normality of the maximum likelihood estimator (MLE) under regularity conditions is a cornerstone of statistical theory. In this paper, we give explicit upper bounds on the distributional distance between the distribution of the MLE of a vector parameter, and the multivariate normal distribution. We work with possibly high-dimensional, independent but not necessarily identically distributed random vectors. In addition, we obtain upper bounds in cases where the MLE cannot be expressed analytically.

MSC 2010 subject classifications: Primary 62F12; secondary 62E17.

Keywords and phrases: Multi-parameter maximum likelihood estimation, multivariate normal approximation, Stein’s method.

Received August 2017.

Contents

1	Introduction	3795
2	Bounds for multi-parameter distributions	3799
	2.1 A general bound	3799
	2.2 Linear regression	3802
	2.3 Special case: Identically distributed random vectors	3807
3	Bounds when the MLE is not known explicitly	3811
4	Proof of Theorem 2.1	3816
	Appendix: Proofs of Lemma 4.1 and of Corollaries 2.3 and 3.1	3820
	References	3827

*The author would like to thank Gesine Reinert, Tobias Kley, and Christophe Ley for insightful comments. In addition, the author sincerely thanks the Editor, Associate Editor and one anonymous referee for suggestions that lead to an improvement of the paper.

1. Introduction

The assessment of the quality of various normal approximations has attracted the interest of statisticians for many years. In general this is not an easy task and as Kiefer (1968) points out, *to give explicitly useful bounds on the departure from the asymptotic normal distribution as a function of the sample size seems to be a terrifically difficult problem*. Since then, Berry-Esseen type bounds have been derived for general (mainly linear) statistics; see for example Koroljuk & Borovskich (1994) for the case of U -statistics.

Due to the fact that the Maximum Likelihood Estimator (MLE) is not in general a linear function of the random variables, it was only recently that the assessment of its asymptotic normality has started getting significant attention in statistical research. Obtaining a quantitative statement related to the normal approximation of the MLE can be helpful to assess whether using the limiting distribution is an acceptable approximation or not. In addition, such results can save both money and time by giving a good indication on whether a larger sample size is indeed necessary.

The case of a scalar MLE for observations from single-parameter distributions is the first that has been covered in a series of papers. The existing approaches are mainly split into two categories based on whether a technique called Stein's method (as first introduced in Stein (1972)) was employed in order to get distributional bounds, or not. In the former category, where Stein's method was used, one can measure the MLE-related normal approximation error in a wide range of metrics, such as Zolotarev-type distances (for example the Wasserstein distance) and the Kolmogorov metric. Anastasiou & Reinert (2017) provide the most general approach, where bounds on the distributional distance between the distribution of the MLE and the normal distribution are given and no restrictions are imposed on the form of the MLE. Anastasiou & Ley (2017) give a different approach to the problem based on a combination of Stein's method with the Delta method for situations where the MLE can be expressed as a function of the sum of independent terms. Their strategy consists in benefiting from this special form of the MLE, which allows the direct usage of Stein's method on a sum of random elements. The bounds given in Anastasiou & Ley (2017) are simpler than those obtained in Anastasiou & Reinert (2017). We note however, that an obvious advantage of the methodology developed in Anastasiou & Reinert (2017) is its wider applicability as it works for all MLE settings (not requiring the MLE to be of a special form) and even for cases where an analytic expression of the MLE is not known. In the recent contribution of Anastasiou (2017) the independence assumption is relaxed and the normal approximation of the MLE is assessed under the presence of a local dependence structure between the random variables. The resulting Zolotarev-type bounds are of the optimal $\mathcal{O}(n^{-1/2})$ order, while the obtained bounds on the Kolmogorov distance are $\mathcal{O}(n^{-1/4})$.

In the second category, where Stein's method is not used, bounds are given in

the Kolmogorov distance. Using the Delta method and under the requirement that the MLE can be expressed as a function of the sum of independent random elements, Pinelis & Molzon (2016) provide uniform and non-uniform Berry-Esseen bounds on the rate of convergence to normality for various statistics, among which is the MLE. The conditions used are partly different than those in Anastasiou & Ley (2017), where the Delta method was also employed. The bounds achieve the optimal $\mathcal{O}(n^{-1/2})$ order. Pinelis (2017) extends the results of Pinelis & Molzon (2016) in cases where the MLE is not necessarily a function of the sum of independent random terms. Under conditions, he shows that the MLE can be tightly enough bracketed between two smooth enough functions, which makes the Delta method applicable. With regards to the Kolmogorov distance, the obtained bounds are again of the optimal order, which is an advantage over the Stein's method related approaches of the previous paragraph, where the order of the bound on the Kolmogorov distance is only $\mathcal{O}(n^{-1/4})$. However, the results given in Anastasiou & Reinert (2017) and in the current paper are more general in the sense that firstly, depending on the choice of the test function h , they can give results for metrics different than the Kolmogorov distance and secondly, under assumptions, are applicable when the MLE is not known analytically.

In this paper, we give upper bounds on the distributional distance between the distribution of a vector MLE and the multivariate normal, which under specific regularity conditions (given at a later stage) is the MLE's limiting distribution. We partly employ multivariate Stein's method for Gaussian approximation (for multivariate Stein's method for other target distributions, see Gorham et al. (2016)) and our focus is on independent but not necessarily identically distributed random vectors. The bounds obtained are explicit in terms of the sample size and the parameter. We are the first to give results for situations where the vector MLE can not be expressed in a closed form. The wide applicability of the maximum likelihood estimation method adds to the importance of our results. Among others, an MLE is used in ordinary and generalised linear models, time series analysis and a large number of other situations related to hypothesis testing and confidence intervals; see Section 2.2 for bounds related to linear regression models.

The notation which is used throughout the paper is as follows. The parameter space is $\Theta \subset \mathbb{R}^d$ equipped with the Euclidean norm. Let $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_d)^\top$ denote a parameter from the parameter space, while $\boldsymbol{\theta}_0 = (\theta_{0,1}, \theta_{0,2}, \dots, \theta_{0,d})^\top$ denotes the true, but unknown, value of the parameter. The probability density (or probability mass) function is denoted by $f(\mathbf{x}|\boldsymbol{\theta})$, where $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)$. The likelihood function is $L(\boldsymbol{\theta}; \mathbf{x}) = f(\mathbf{x}|\boldsymbol{\theta})$. Its natural logarithm, called the log-likelihood function is denoted by $\ell(\boldsymbol{\theta}; \mathbf{x})$. A maximum likelihood estimate (not seen as a random vector) is a value of the parameter which maximises the likelihood function. For many models the MLE as a random vector exists and is also unique, in which case it is denoted by $\hat{\boldsymbol{\theta}}_n(\mathbf{X})$; see Mäkeläinen et al. (1981) for a set of assumptions that ensure existence and uniqueness. This is known as the 'regular' case. However, existence and uniqueness of the MLE can not be

taken for granted, see e.g. Billingsley (1961) for an example of non-uniqueness.

For $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$ being independent but not necessarily identically distributed (i.n.i.d.) random vectors, we denote by $f_i(\mathbf{x}, \boldsymbol{\theta})$ the probability density (or mass) function of \mathbf{X}_i . The likelihood function is $L(\boldsymbol{\theta}; \mathbf{x}) = \prod_{i=1}^n f_i(\mathbf{x}_i|\boldsymbol{\theta})$. With the parameter space Θ being an open subset of \mathbb{R}^d , the asymptotic normality of the MLE holds under the following regularity conditions as expressed in Hoadley (1971):

- (N1) $\hat{\boldsymbol{\theta}}_n(\mathbf{X}) \xrightarrow{P} \boldsymbol{\theta}_0$, as $n \rightarrow \infty$, where $\boldsymbol{\theta}_0$ is the true parameter value;
- (N2) the Hessian matrix $J_k(\mathbf{X}_k, \boldsymbol{\theta}) = \left\{ \frac{\partial^2}{\partial \theta_i \partial \theta_j} \log(f_k(\mathbf{X}_k|\boldsymbol{\theta})) \right\}_{i,j=1,2,\dots,d} \in \mathbb{R}^{d \times d}$ and the gradient vector $\nabla(\log(f_k(\mathbf{X}_k|\boldsymbol{\theta}))) \in \mathbb{R}^{d \times 1}$ exist almost surely $\forall k \in \{1, 2, \dots, n\}$ with respect to the probability measure \mathbb{P} ;
- (N3) $J_k(\mathbf{X}_k, \boldsymbol{\theta})$ is a continuous function of $\boldsymbol{\theta}$, $\forall k = 1, 2, \dots, n$, almost surely with respect to \mathbb{P} and is a measurable function of \mathbf{X}_k ;
- (N4) $\mathbb{E}_{\boldsymbol{\theta}} [\nabla(\log(f_k(\mathbf{X}_k|\boldsymbol{\theta})))] = \mathbf{0}$, $k = 1, 2, \dots, n$;
- (N5) with \mathbf{y}^\top denoting the transpose of a vector \mathbf{y} ,

$$\mathbb{E}_{\boldsymbol{\theta}} [\nabla(\log(f_k(\mathbf{X}_k|\boldsymbol{\theta}))) [\nabla(\log(f_k(\mathbf{X}_k|\boldsymbol{\theta})))]^\top] = -\mathbb{E} [J_k(\mathbf{X}_k, \boldsymbol{\theta})] =: I_k(\boldsymbol{\theta});$$

(N6) for

$$\bar{I}_n(\boldsymbol{\theta}) = \frac{1}{n} \sum_{j=1}^n I_j(\boldsymbol{\theta}), \tag{1.1}$$

there exists a matrix $\bar{I}(\boldsymbol{\theta}) \in \mathbb{R}^{d \times d}$ such that $\bar{I}_n(\boldsymbol{\theta}) \xrightarrow[n \rightarrow \infty]{} \bar{I}(\boldsymbol{\theta})$. In addition, $\bar{I}_n(\boldsymbol{\theta}), \bar{I}(\boldsymbol{\theta})$ are symmetric matrices for all $\boldsymbol{\theta}$ and $\bar{I}(\boldsymbol{\theta})$ is positive definite;

- (N7) for some $\delta > 0$, $\frac{\sum_k \mathbb{E}_{\boldsymbol{\theta}_0} |\boldsymbol{\lambda}^\top \nabla(\log(f_k(\mathbf{X}_k)))|^{2+\delta}}{n^{\frac{2+\delta}{2}}} \xrightarrow[n \rightarrow \infty]{} 0$ for all $\boldsymbol{\lambda} \in \mathbb{R}^d$;
- (N8) with $\|\cdot\|$ the ordinary Euclidean norm on \mathbb{R}^d , then for $k, i, j \in \{1, 2, \dots, d\}$ there exist $\epsilon > 0, K > 0, \delta > 0$ and random variables $B_{k,ij}(\mathbf{X}_k)$ such that
 - (i) $\sup \left\{ \left| \frac{\partial^2}{\partial \theta_i \partial \theta_j} \log(f_k(\mathbf{X}_k|\mathbf{t})) \right| : \|\mathbf{t} - \boldsymbol{\theta}_0\| \leq \epsilon \right\} \leq B_{k,ij}(\mathbf{X}_k)$;
 - (ii) $\mathbb{E} |B_{k,ij}(\mathbf{X}_k)|^{1+\delta} \leq K$.

Assuming that $\hat{\boldsymbol{\theta}}_n(\mathbf{X})$ exists and is unique, the following theorem gives the result for the asymptotic normality of the MLE in the case of i.n.i.d. random vectors in a slightly different way than Hoadley (1971).

Theorem 1.1. *Let $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$ be independent random vectors with probability density (or mass) functions $f_i(\mathbf{x}_i|\boldsymbol{\theta})$, where $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^d$. Assume that the MLE exists and is unique and that the regularity conditions (N1)-(N8) hold.*

Also let $\mathbf{Z} \sim N_d(\mathbf{0}, I_{d \times d})$, where $\mathbf{0}$ is the $d \times 1$ zero vector and $I_{d \times d}$ is the $d \times d$ identity matrix. Then, for $\bar{I}_n(\boldsymbol{\theta})$ as in (1.1)

$$\sqrt{n} [\bar{I}_n(\boldsymbol{\theta}_0)]^{\frac{1}{2}} \left(\hat{\boldsymbol{\theta}}_n(\mathbf{X}) - \boldsymbol{\theta}_0 \right) \xrightarrow[n \rightarrow \infty]{d} \mathbf{Z}. \quad (1.2)$$

Proof. Hoadley (1971) proves in Theorem 2, p.1983 that under the regularity conditions (N1)-(N8),

$$\sqrt{n} \left(\hat{\boldsymbol{\theta}}_n(\mathbf{X}) - \boldsymbol{\theta}_0 \right) \xrightarrow[n \rightarrow \infty]{d} [\bar{I}(\boldsymbol{\theta}_0)]^{-\frac{1}{2}} \mathbf{Z}.$$

Using this result and (N6) we obtain that

$$[\bar{I}_n(\boldsymbol{\theta}_0)]^{\frac{1}{2}} \sqrt{n} \left(\hat{\boldsymbol{\theta}}_n(\mathbf{X}) - \boldsymbol{\theta}_0 \right) \xrightarrow[n \rightarrow \infty]{d} [\bar{I}(\boldsymbol{\theta}_0)]^{\frac{1}{2}} [\bar{I}(\boldsymbol{\theta}_0)]^{-\frac{1}{2}} \mathbf{Z} = \mathbf{Z},$$

which is the result of the theorem. ■

The interest is on assessing the quality of the asymptotic normality of the MLE in (1.2). For any three times differentiable function $h: \mathbb{R}^d \rightarrow \mathbb{R}$, we abbreviate $\|h\| := \sup |h|$, $\|h\|_1 := \sup_i \left| \frac{\partial}{\partial x_i} h \right|$, $\|h\|_2 := \sup_{i,j} \left| \frac{\partial^2}{\partial x_i \partial x_j} h \right|$, and $\|h\|_3 := \sup_{i,j,k} \left| \frac{\partial^3}{\partial x_i \partial x_j \partial x_k} h \right|$. For $j \in \{1, 2, 3\}$, let

$$H = \{h: \mathbb{R}^d \rightarrow \mathbb{R}: h \text{ is three times differentiable with bounded } \|h\|, \|h\|_j\} \quad (1.3)$$

be the class of test functions used in the paper. We will give upper bounds on

$$\left| \mathbb{E} \left[h \left(\sqrt{n} [\bar{I}_n(\boldsymbol{\theta}_0)]^{\frac{1}{2}} \left(\hat{\boldsymbol{\theta}}_n(\mathbf{X}) - \boldsymbol{\theta}_0 \right) \right) \right] - \mathbb{E}[h(\mathbf{Z})] \right|, \quad (1.4)$$

where $\mathbf{Z} \sim N_d(\mathbf{0}, I_{d \times d})$. The bounds are explicit in terms of the sample size and $\boldsymbol{\theta}_0$. The main result of the paper is given in Theorem 2.1, where we obtain a general upper bound on (1.4) which holds under slightly weaker assumptions than the usual, sufficient regularity conditions (N1)-(N8) used for the asymptotic normality of the MLE. The generality of the bound adds to its importance as it can be applied in various different occasions. Furthermore, Theorem 3.1 is also substantial since we give an upper bound related to the asymptotic normality of the MLE, even when the MLE is not known analytically, but it is assumed to be within an ϵ -neighbourhood of $\boldsymbol{\theta}_0$, for $\epsilon > 0$.

The paper is organised as follows. Section 2 first treats the case of independent but not necessarily identically distributed (i.n.i.d.) random vectors. The upper bound on the distributional distance between the distribution of the vector MLE and the multivariate normal distribution is presented. Special attention is given to linear regression models with an application to the simplest case of the straight-line model, where apart from the upper bound, we also give results from a simulation study. Furthermore, under weaker regularity conditions, we

explain how the bound can be simplified for the case of i.i.d. random vectors. Specific theoretical and empirical results for independent random variables from the normal distribution under canonical parametrisation are given. In Section 3 we explain how the results can be expanded when no analytic expression of the vector MLE is available. We briefly illustrate the results for the Beta distribution with both shape parameters unknown. In order to make the paper easily readable, we only provide an outline of the proof of our main Theorem 2.1, with the complete proof being given in Section 4. In addition, some technical results and proofs of corollaries that are not essential for the smooth understanding of the paper are confined in the Appendix.

2. Bounds for multi-parameter distributions

In this section we examine the case of i.n.i.d. t -dimensional random vectors, for $t \in \mathbb{Z}^+$. We give an upper bound on the distributional distance between the distribution of the MLE and the multivariate normal. An example from linear models then follows. The last subsection covers, under weaker regularity conditions, the case of i.i.d. random vectors and an example from the normal distribution under canonical parametrisation serves as illustration of our results. It is worth mentioning that the MLE in this example is not a sum of random variables and classical Stein method approaches cannot be applied directly.

2.1. A general bound

The normal approximation in (1.2) is an asymptotic result and our motivation is to assess the quality of this normal approximation through explicit, for finite sample size, upper bounds on the distributional distance of interest. From now on, $\bar{I}_n(\boldsymbol{\theta})$ is as in (1.1). Let the subscript (m) denote an index for which the quantity $|\hat{\theta}_n(\mathbf{x})_{(m)} - \theta_{0,(m)}|$ is the largest among the d components;

$$(m) \in \{1, \dots, d\} \text{ is such that } |\hat{\theta}_n(\mathbf{x})_{(m)} - \theta_{0,(m)}| \geq |\hat{\theta}_n(\mathbf{x})_j - \theta_{0,j}|, \forall j \in \{1, \dots, d\}.$$

For ease of presentation, let us introduce the following notation:

$$\begin{aligned} Q_{(m)} &= Q_{(m)}(\mathbf{X}, \boldsymbol{\theta}_0) := \hat{\theta}_n(\mathbf{X})_{(m)} - \theta_{0,(m)} \\ Q_j &= Q_j(\mathbf{X}, \boldsymbol{\theta}_0) := \hat{\theta}_n(\mathbf{X})_j - \theta_{0,j}, \quad \forall j \in \{1, 2, \dots, d\} \\ T_{lj} &= T_{lj}(\boldsymbol{\theta}_0, \mathbf{X}) = \frac{\partial^2}{\partial \theta_l \partial \theta_j} \ell(\boldsymbol{\theta}_0; \mathbf{X}) + n[\bar{I}_n(\boldsymbol{\theta}_0)]_{lj}, \quad j, l \in \{1, 2, \dots, d\} \\ \tilde{V} &= \tilde{V}(n, \boldsymbol{\theta}_0) := [\bar{I}_n(\boldsymbol{\theta}_0)]^{-\frac{1}{2}} \\ \xi_{ij} &= \frac{1}{\sqrt{n}} \sum_{k=1}^d \tilde{V}_{jk} \frac{\partial}{\partial \theta_k} \log(f_i(\mathbf{X}_i | \boldsymbol{\theta}_0)), \quad i \in \{1, 2, \dots, n\}, \quad j \in \{1, 2, \dots, d\}. \end{aligned} \tag{2.5}$$

Notice that, using conditions (N5) and (N6), $\mathbb{E}[T_{lj}] = 0$ and in general, we expect T_{lj} to be small. The main result of the paper is as follows.

Theorem 2.1. *Let $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$ be i.n.i.d. \mathbb{R}^t -valued, $t \in \mathbb{Z}^+$, random vectors with probability density (or mass) function $f_i(\mathbf{x}_i|\boldsymbol{\theta})$, for which the parameter space Θ is an open subset of \mathbb{R}^d . Assume that the MLE exists and is unique and that (N1)-(N6) are satisfied. In addition, assume that for any $\boldsymbol{\theta}_0 \in \Theta$ there exists $0 < \epsilon = \epsilon(\boldsymbol{\theta}_0)$ and functions $M_{kjl}(\mathbf{x})$, $\forall k, j, l \in \{1, 2, \dots, d\}$ such that $\left| \frac{\partial^3}{\partial \theta_k \partial \theta_j \partial \theta_l} \ell(\boldsymbol{\theta}, \mathbf{x}) \right| \leq M_{kjl}(\mathbf{x})$ for all $\boldsymbol{\theta} \in \Theta$ with $|\theta_j - \theta_{0,j}| < \epsilon \forall j \in \{1, 2, \dots, d\}$. Also, for $Q_{(m)}$ as in (2.5), assume that $\mathbb{E} \left[(M_{kjl}(\mathbf{X}))^2 \mid Q_{(m)} < \epsilon \right] < \infty$. Let $\{\mathbf{X}'_i, i = 1, 2, \dots, n\}$ be an independent copy of $\{\mathbf{X}_i, i = 1, 2, \dots, n\}$. For $\mathbf{Z} \sim N_d(\mathbf{0}, I_{d \times d})$, $h \in H$, where H is as in (1.3), and with Q_j, T_{lj} , and ξ_{ik} as in (2.5), it holds that*

$$\begin{aligned} & \left| \mathbb{E} \left[h \left(\sqrt{n} [\bar{I}_n(\boldsymbol{\theta}_0)]^{\frac{1}{2}} \left(\hat{\boldsymbol{\theta}}_n(\mathbf{X}) - \boldsymbol{\theta}_0 \right) \right) \right] - \mathbb{E}[h(\mathbf{Z})] \right| \\ & \leq \frac{1}{\sqrt{n}} (\|h\|_1 K_1(\boldsymbol{\theta}_0) + \|h\|_2 K_2(\boldsymbol{\theta}_0) + \|h\|_3 K_3(\boldsymbol{\theta}_0)) + \frac{2\|h\|}{\epsilon^2} \mathbb{E} \left[\sum_{j=1}^d Q_j^2 \right], \end{aligned} \quad (2.6)$$

where,

$$\begin{aligned} K_1(\boldsymbol{\theta}_0) &= \sum_{k=1}^d \sum_{l=1}^d |\tilde{V}_{lk}| \sum_{j=1}^d \sqrt{\mathbb{E}[Q_j^2] \mathbb{E}[T_{kj}^2]} \\ &+ \frac{1}{2} \sum_{k=1}^d \sum_{l=1}^d |\tilde{V}_{lk}| \sum_{j=1}^d \sum_{v=1}^d \sqrt{\mathbb{E}[Q_j^2 Q_v^2]} \sqrt{\mathbb{E}[(M_{kjl}(\mathbf{X}))^2 \mid Q_{(m)} < \epsilon]} \end{aligned} \quad (2.7)$$

$$K_2(\boldsymbol{\theta}_0) = \frac{1}{4\sqrt{n}} \left\{ \sum_{j=1}^d \sqrt{\sum_{i=1}^n \text{Var}[n\xi_{ij}^2]} + 2 \sum_{k=1}^{d-1} \sum_{j=k+1}^d \sqrt{\sum_{i=1}^n \text{Var}[n\xi_{ij}\xi_{ik}]} \right\} \quad (2.8)$$

$$K_3(\boldsymbol{\theta}_0) = \frac{1}{12n} \sum_{i=1}^n \mathbb{E} \left[\sum_{m=1}^d \left| \sum_{l=1}^d \tilde{V}_{ml} \left(\frac{\partial}{\partial \theta_l} \{ \log(f_i(\mathbf{X}'_i|\boldsymbol{\theta}_0)) - \log(f_i(\mathbf{X}_i|\boldsymbol{\theta}_0)) \} \right) \right|^2 \right] \quad (2.9)$$

Remark 2.1. (1) At first glance, the bound seems complicated. However, the examples that follow show that the terms are easily calculated giving an expression for the bound, which is of the optimal $n^{-1/2}$ order.

(2) Assuming that $\bar{I}_n(\boldsymbol{\theta}_0) = \mathcal{O}(1)$ in (1.1) yields, for fixed d , $\mathbb{E} \left[\sum_{j=1}^d Q_j^2 \right] = \mathcal{O}(n^{-1})$. To see this, first use that from the asymptotic normality of the MLE as expressed in Theorem 1.1, $\sqrt{n} \mathbb{E} \left[\hat{\boldsymbol{\theta}}_n(\mathbf{X}) - \boldsymbol{\theta}_0 \right] \xrightarrow{n \rightarrow \infty} \mathbf{0}$ and thus

$$\mathbb{E}[Q_j] = o\left(\frac{1}{\sqrt{n}}\right), \quad \forall j \in \{1, 2, \dots, d\}.$$

Secondly, from Theorem 1.1 we also get that

$$n [\bar{I}_n(\boldsymbol{\theta}_0)]^{\frac{1}{2}} \text{Cov} [\hat{\boldsymbol{\theta}}_n(\mathbf{X})] [\bar{I}_n(\boldsymbol{\theta}_0)]^{\frac{1}{2}} \xrightarrow{n \rightarrow \infty} I_{d \times d}. \tag{2.10}$$

Assuming that the matrix $\bar{I}_n(\boldsymbol{\theta}_0)$ is $\mathcal{O}(1)$, it follows from (2.10) that $\text{Var} [\hat{\boldsymbol{\theta}}_n(\mathbf{X})_j] = \mathcal{O}(n^{-1})$, $\forall j \in \{1, 2, \dots, d\}$ and therefore,

$$\mathbb{E} [Q_j^2] = \text{Var} [\hat{\boldsymbol{\theta}}_n(\mathbf{X})_j] + [\mathbb{E} [Q_j]]^2 = \mathcal{O}(n^{-1}). \tag{2.11}$$

(3) With T_{lj} as in (2.5), using (N5), (N6), and the fact that $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$ are independent yields

$$\mathbb{E} [T_{lj}^2] = \sum_{i=1}^n \text{Var} \left[\frac{\partial^2}{\partial \theta_l \partial \theta_j} \log (f_i(\mathbf{X}_i | \boldsymbol{\theta}_0)) \right], \tag{2.12}$$

meaning that $\mathbb{E} [T_{lj}^2]$ is $\mathcal{O}(n)$.

(4) Using (2.11) and (2.12), then if $\bar{I}_n(\boldsymbol{\theta}_0) = \mathcal{O}(1)$ it can be deduced that

$$K_1(\boldsymbol{\theta}_0) = \mathcal{O}(1), \quad K_2(\boldsymbol{\theta}_0) = \mathcal{O}(1), \quad K_3(\boldsymbol{\theta}_0) = \mathcal{O}(1),$$

where $K_1(\boldsymbol{\theta}_0), K_2(\boldsymbol{\theta}_0), K_3(\boldsymbol{\theta}_0)$ are as in (2.7), (2.8), (2.9), respectively. Hence, the upper bound in Theorem 2.1 is $\mathcal{O}(n^{-1/2})$.

(5) In terms of the dimensionality d of the parameter, having that $\xi_{ij} = \mathcal{O}(d)$, then $K_1(\boldsymbol{\theta}_0) = \mathcal{O}(d^4)$, $K_2(\boldsymbol{\theta}_0) = \mathcal{O}(d^4)$ and $K_3(\boldsymbol{\theta}_0) = \mathcal{O}(d^6)$ as can be deduced from (2.7), (2.8) and (2.9), respectively. The last term of the bound in (2.6) is of order d in terms of the dimensionality of the parameter. Thus, for $d \gg n$ the bound does not behave well, but d could grow moderately with n . For example $d = o(n^\alpha)$, $0 < \alpha < \frac{1}{12}$ would still yield a bound which goes to zero as n goes to infinity.

Outline of the proof. From the definition of the MLE, $\frac{\partial}{\partial \theta_k} \ell(\hat{\boldsymbol{\theta}}_n(\mathbf{x}); \mathbf{x}) = 0 \forall k \in \{1, 2, \dots, d\}$. A second-order Taylor expansion of $\frac{\partial}{\partial \theta_k} \ell(\hat{\boldsymbol{\theta}}_n(\mathbf{x}); \mathbf{x})$ about $\boldsymbol{\theta}_0$ yields for Q_j as in (2.5)

$$\begin{aligned} \sum_{j=1}^d Q_j \left(\frac{\partial^2}{\partial \theta_k \partial \theta_j} \ell(\boldsymbol{\theta}_0; \mathbf{x}) \right) &= - \frac{\partial}{\partial \theta_k} \ell(\boldsymbol{\theta}_0; \mathbf{x}) \\ &\quad - \frac{1}{2} \sum_{j=1}^d \sum_{q=1}^d Q_j Q_q \left(\frac{\partial^3}{\partial \theta_k \partial \theta_j \partial \theta_q} \ell(\boldsymbol{\theta}; \mathbf{x}) \Big|_{\boldsymbol{\theta} = \boldsymbol{\theta}_0^*} \right), \end{aligned}$$

with $\boldsymbol{\theta}_0^*$ between $\hat{\boldsymbol{\theta}}_n(\mathbf{x})$ and $\boldsymbol{\theta}_0$. Adding $\sum_{j=1}^d n [\bar{I}_n(\boldsymbol{\theta}_0)]_{kj} Q_j$ on both sides of the above equation gives, for T_{kj} as in (2.5), that

$$\begin{aligned} \sum_{j=1}^d n[\bar{I}_n(\boldsymbol{\theta}_0)]_{kj} Q_j &= \frac{\partial}{\partial \theta_k} \ell(\boldsymbol{\theta}_0; \mathbf{x}) + \sum_{j=1}^d Q_j T_{kj} \\ &+ \frac{1}{2} \sum_{j=1}^d \sum_{q=1}^d Q_j Q_q \left(\frac{\partial^3}{\partial \theta_k \partial \theta_j \partial \theta_q} \ell(\boldsymbol{\theta}; \mathbf{x}) \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0^*} \right). \end{aligned} \quad (2.13)$$

Using (2.13), which holds $\forall k \in \{1, 2, \dots, d\}$, and with \tilde{V} as in (2.5),

$$\begin{aligned} &\sqrt{n}[\bar{I}_n(\boldsymbol{\theta}_0)]^{\frac{1}{2}}(\hat{\boldsymbol{\theta}}_n(\mathbf{x}) - \boldsymbol{\theta}_0) \\ &= \frac{\tilde{V}}{\sqrt{n}} \left\{ \nabla(\ell(\boldsymbol{\theta}_0; \mathbf{x})) + \sum_{j=1}^d Q_j \left(\nabla \left(\frac{\partial}{\partial \theta_j} \ell(\boldsymbol{\theta}_0; \mathbf{x}) \right) + n[\bar{I}_n(\boldsymbol{\theta}_0)]_{[j]} \right) \right. \\ &\quad \left. + \frac{1}{2} \sum_{j=1}^d \sum_{q=1}^d Q_j Q_q \left(\nabla \left(\frac{\partial^2}{\partial \theta_j \partial \theta_q} \ell(\boldsymbol{\theta}; \mathbf{x}) \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0^*} \right) \right) \right\}, \end{aligned}$$

where $[\bar{I}_n(\boldsymbol{\theta}_0)]_{[j]}$ is the j^{th} column of the matrix $\bar{I}_n(\boldsymbol{\theta}_0)$. The triangle inequality gives that

$$\begin{aligned} &\left| \mathbb{E} \left[h \left(\sqrt{n} [\bar{I}_n(\boldsymbol{\theta}_0)]^{\frac{1}{2}} (\hat{\boldsymbol{\theta}}_n(\mathbf{X}) - \boldsymbol{\theta}_0) \right) \right] - \mathbb{E}[h(\mathbf{Z})] \right| \\ &\leq \left| \mathbb{E} \left[h \left(\frac{\tilde{V}}{\sqrt{n}} \nabla(\ell(\boldsymbol{\theta}_0; \mathbf{X})) \right) \right] - \mathbb{E}[h(\mathbf{Z})] \right| \end{aligned} \quad (2.14)$$

$$+ \left| \mathbb{E} \left[h \left(\sqrt{n} [\bar{I}_n(\boldsymbol{\theta}_0)]^{\frac{1}{2}} (\hat{\boldsymbol{\theta}}_n(\mathbf{X}) - \boldsymbol{\theta}_0) \right) - h \left(\frac{\tilde{V}}{\sqrt{n}} \nabla(\ell(\boldsymbol{\theta}_0; \mathbf{X})) \right) \right] \right|. \quad (2.15)$$

Now, (2.14) is based on $\nabla(\ell(\boldsymbol{\theta}_0; \mathbf{x})) = \sum_{i=1}^n \nabla(\log(f_i(\mathbf{x}_i|\boldsymbol{\theta}_0)))$ which is a sum of independent random vectors. For this expression, a bound using Stein's method for multivariate normal approximation will be derived. In contrast, (2.15) will be bounded using multivariate Taylor expansions. Technical difficulties arise as the third-order partial derivatives of the log-likelihood function may not be uniformly bounded in $\boldsymbol{\theta}$. Therefore, for $0 < \epsilon = \epsilon(\boldsymbol{\theta}_0)$ we will condition on whether $|Q_{(m)}|$ as defined in (2.5) is greater or less than the positive constant ϵ and each case will be treated separately by bounding conditional expectations. Known probability inequalities, such as the Cauchy-Schwarz and Markov's inequality, will be employed in order to derive the upper bounds in each case.

2.2. Linear regression

This subsection calculates the bound in (2.6) for linear regression models. The asymptotic normality of the MLE in linear regression models has been proven in Fahrmeir & Kaufmann (1985). We give the example of a straight-line regression and the bound turns out to be, as expected, of order $\mathcal{O}(n^{-1/2})$, where n is

the sample size. The following notation is used throughout this subsection. The vector $\mathbf{Y} = (Y_1, Y_2, \dots, Y_n)^\top \in \mathbb{R}^{n \times 1}$ denotes the response variable for the linear regression, while $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_d)^\top \in \mathbb{R}^{d \times 1}$ is the vector of the d parameters and $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \dots, \epsilon_n)^\top \in \mathbb{R}^{n \times 1}$ is the vector of the error terms, which are i.i.d. random variables with $\epsilon_i \sim N(0, \sigma^2), \forall i \in \{1, 2, \dots, n\}$. The true value of the unknown parameter $\boldsymbol{\beta}$ is denoted by $\boldsymbol{\beta}_0 = (\beta_{0,1}, \beta_{0,2}, \dots, \beta_{0,d})^\top \in \mathbb{R}^{d \times 1}$. The design matrix is

$$X = \begin{pmatrix} 1 & x_{1,2} & \dots & x_{1,d} \\ 1 & x_{2,2} & \dots & x_{2,d} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,2} & \dots & x_{n,d} \end{pmatrix}.$$

For the model

$$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

the aim is to find upper bounds on the distributional distance between the distribution of the MLE, $\hat{\boldsymbol{\beta}}$, and the normal distribution. The probability density function for Y_i is

$$f_i(y_i|\boldsymbol{\beta}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{1}{2\sigma^2} (y_i - X_{[i]}\boldsymbol{\beta})^2 \right\}, \tag{2.16}$$

where $X_{[i]}$ denotes the i^{th} row of the design matrix. The parameter space $\Theta = \mathbb{R}^d$ is open and if $X^\top X$ is of full rank, the matrix $X^\top X$ is invertible and

$$\hat{\boldsymbol{\beta}} = (X^\top X)^{-1} X^\top \mathbf{Y}. \tag{2.17}$$

We now bound the corresponding distributional distance.

Corollary 2.1. *Let $Y_i, i \in \{1, 2, \dots, n\}$ be independent normal random variables with*

$$Y_i \sim N(X_{[i]}\boldsymbol{\beta}_0, \sigma^2),$$

where σ^2 is known. Assume that the $d \times d$ matrix $X^\top X$ is of full rank. Let $\{Y'_i, i = 1, 2, \dots, n\}$ be an independent copy of $\{Y_i, i = 1, 2, \dots, n\}$ and $\mathbf{Z} \sim N_d(\mathbf{0}, I_{d \times d})$ and $\bar{I}_n(\boldsymbol{\beta})$ is as in (1.1). Then for $h \in H$ as in (1.3),

$$\begin{aligned} & \left| \mathbb{E} \left[h \left(\sqrt{n} [\bar{I}_n(\boldsymbol{\beta}_0)]^{\frac{1}{2}} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \right) \right] - \mathbb{E}[h(\mathbf{Z})] \right| \\ & \leq \frac{\|h\|_2}{4} \sum_{j=1}^d \left[\sum_{i=1}^n \text{Var} \left[\left(\sum_{k=1}^d \frac{X_{ik}}{\sigma} [X^\top X]^{-\frac{1}{2}} \right)_{jk} \left(Y_i - \sum_{m=1}^d X_{im}\beta_{0,m} \right) \right]^2 \right]^{\frac{1}{2}} \\ & + \frac{\|h\|_2}{2} \sum_{k=1}^{d-1} \sum_{j=k+1}^d \left[\sum_{i=1}^n \text{Var} \left[\sum_{q=1}^d \sum_{v=1}^d \frac{X_{iq}X_{iv}}{\sigma^2} [X^\top X]^{-\frac{1}{2}} \right]_{jq} [X^\top X]^{-\frac{1}{2}} \right]_{kv} \\ & \quad \times \left(Y_i - \sum_{m=1}^d X_{im}\beta_{0,m} \right)^2 \right]^{\frac{1}{2}} \end{aligned}$$

$$+ \frac{\|h\|_3}{12} \sum_{i=1}^n \mathbb{E} \left[\sum_{m=1}^d \left| \sum_{l=1}^d \frac{X_{il}}{\sigma} \left[[X^\top X]^{-\frac{1}{2}} \right]_{ml} (Y_i - Y'_i) \right| \right]^3. \quad (2.18)$$

Proof. Using (2.16), we can see that the Hessian matrix for the log-likelihood function does not depend on \mathbf{y} and $\bar{I}_n(\boldsymbol{\beta}_0) = \frac{1}{n\sigma^2} X^\top X$. The result in (2.17) yields

$$\begin{aligned} \sqrt{n} [\bar{I}_n(\boldsymbol{\beta}_0)]^{\frac{1}{2}} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) &= \frac{1}{\sigma} \left\{ [X^\top X]^{-\frac{1}{2}} X^\top \mathbf{Y} - [X^\top X]^{\frac{1}{2}} \boldsymbol{\beta}_0 \right\} \\ &= \frac{1}{\sqrt{n}} \left[\sigma \sqrt{n} [X^\top X]^{-\frac{1}{2}} \right] \frac{1}{\sigma^2} (X^\top \mathbf{Y} - X^\top X \boldsymbol{\beta}_0) \\ &= \frac{1}{\sqrt{n}} [I_n(\boldsymbol{\beta}_0)]^{-\frac{1}{2}} \frac{d}{d\boldsymbol{\beta}} \ell(\boldsymbol{\beta}; \mathbf{y}) \Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}. \end{aligned} \quad (2.19)$$

Having a closer look at the expression in (2.19), we notice that actually the quantity of interest $\left| \mathbb{E} \left[h \left(\sqrt{n} [\bar{I}_n(\boldsymbol{\beta})]^{\frac{1}{2}} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \right) \right] - \mathbb{E}[h(\mathbf{Z})] \right|$ is equal to (2.14), with (2.15) being equal to zero for this specific case of the linear regression model. Thus, using (4.50) and

$$\frac{\partial}{\partial \beta_k} \log(f_i(Y_i|\boldsymbol{\beta}_0)) = \frac{X_{ik}}{\sigma^2} \left(Y_i - \sum_{m=1}^d X_{im} \beta_{0,m} \right)$$

in Theorem 2.1 yields the result of the corollary. ■

Example: The simple linear model ($d=2$)

Here, we apply the results of (2.18) to the case of a straight-line regression with two unknown parameters. The model is

$$Y_i = \beta_1 + \beta_2(x_i - \bar{x}) + \epsilon_i, \quad \forall i \in \{1, 2, \dots, n\}.$$

The unknown parameters β_1 and β_2 are the *intercept* and *slope* of the regression, respectively. As before, the i.i.d. random variables $\epsilon_i \sim N(0, \sigma^2)$, $\forall i \in \{1, 2, \dots, n\}$. The MLE exists, it is unique and $\hat{\boldsymbol{\beta}} = \left(\bar{Y}, \frac{\sum_{i=1}^n (x_i - \bar{x}) Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} \right)^\top$.

Corollary 2.2. *Let Y_1, Y_2, \dots, Y_n be independent random variables with $Y_i \sim N(\beta_1 + \beta_2(x_i - \bar{x}), \sigma^2)$. The case of $x_i = x_j$, $\forall i, j \in \{1, 2, \dots, n\}$ with $i \neq j$ is excluded and for $\mathbf{Z} \sim N_2(\mathbf{0}, I_{2 \times 2})$ and $h \in H$ as in (1.3),*

$$\begin{aligned} & \left| \mathbb{E} \left[h \left(\sqrt{n} [\bar{I}_n(\boldsymbol{\beta}_0)]^{\frac{1}{2}} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \right) \right] - \mathbb{E}[h(\mathbf{Z})] \right| \\ & \leq \frac{\|h\|_2}{4} \left(\frac{3\sqrt{2}}{\sqrt{n}} + \frac{\sqrt{2 \sum_{i=1}^n (x_i - \bar{x})^4}}{\sum_{i=1}^n (x_i - \bar{x})^2} \right) + \frac{8\|h\|_3}{3\sqrt{\pi}} \left(\frac{1}{\sqrt{n}} + \frac{\sum_{i=1}^n |x_i - \bar{x}|^3}{[\sum_{i=1}^n (x_i - \bar{x})^2]^{\frac{3}{2}}} \right). \end{aligned} \quad (2.20)$$

Remark 2.2. (1) The calculation of the bound is easy and relies only on simple sums. As expected, the order of the bound is $\mathcal{O}(n^{-1/2})$, which is optimal.

(2) We exclude the case of $x_i = x_j, \forall i, j \in \{1, 2, \dots, n\}$ with $i \neq j$, only to ensure that $X^\top X$ is invertible.

Proof. We have that

$$X = \begin{pmatrix} 1 & x_1 - \bar{x} \\ 1 & x_2 - \bar{x} \\ \vdots & \vdots \\ 1 & x_n - \bar{x} \end{pmatrix}, \quad X^\top X = \begin{pmatrix} n & 0 \\ 0 & \sum_{i=1}^n (x_i - \bar{x})^2 \end{pmatrix}. \quad (2.21)$$

The result in (2.21) shows that $X^\top X$ is invertible if and only if $\sum_{i=1}^n (x_i - \bar{x})^2 \neq 0$, which holds if x_i 's are not all identical. The quantities of the bound in (2.18) are calculated for this specific case. We use that $Y_i - \beta_1 - (x_i - \bar{x})\beta_2 \stackrel{d}{=} \sigma Z_i$, where $Z_i \sim N(0, 1)$. For the first term in (2.18) we obtain that

$$\begin{aligned} & \sum_{j=1}^2 \left[\sum_{i=1}^n \text{Var} \left[\left(\sum_{k=1}^2 \frac{X_{ik}}{\sigma} [X^\top X]^{-\frac{1}{2}} \right)_{jk} \left(Y_i - \sum_{m=1}^2 X_{im} \beta_m \right) \right]^2 \right]^{\frac{1}{2}} \\ &= \sum_{j=1}^2 \left[\sum_{i=1}^n \text{Var} \left[\left(\left(\frac{X_{i1}}{\sigma} [X^\top X]^{-\frac{1}{2}} \right)_{j1} + \frac{X_{i2}}{\sigma} [X^\top X]^{-\frac{1}{2}} \right)_{j2} \sigma Z_i \right]^2 \right]^{\frac{1}{2}} \\ &= \frac{1}{n} \left[\sum_{i=1}^n \text{Var} [Z_i^2] \right]^{\frac{1}{2}} + \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} \left[\sum_{i=1}^n (x_i - \bar{x})^4 \text{Var} [Z_i^2] \right]^{\frac{1}{2}} \\ &= \sqrt{\frac{2}{n}} + \frac{\sqrt{2 \sum_{i=1}^n (x_i - \bar{x})^4}}{\sum_{i=1}^n (x_i - \bar{x})^2}. \end{aligned} \quad (2.22)$$

For the second term of (2.18), since $d = 2$ then $k = 1, j = 2$ leading to

$$\begin{aligned} & \left[\sum_{i=1}^n \text{Var} \left[\sum_{q=1}^2 \sum_{v=1}^2 \frac{X_{iq} X_{iv}}{\sigma^2} [X^\top X]^{-\frac{1}{2}} \right]_{2q} [X^\top X]^{-\frac{1}{2}} \right]_{1v} \left(Y_i - \sum_{m=1}^2 X_{im} \beta_m \right)^2 \right]^{\frac{1}{2}} \\ &= \left[\sum_{i=1}^n \text{Var} \left[\frac{X_{i2} X_{i1}}{\sigma^2} [X^\top X]^{-\frac{1}{2}} \right]_{22} [X^\top X]^{-\frac{1}{2}} \right]_{11} (\sigma Z_i)^2 \right]^{\frac{1}{2}} \\ &= \frac{1}{\sqrt{n \sum_{i=1}^n (x_i - \bar{x})^2}} \left[\sum_{i=1}^n (x_i - \bar{x})^2 \text{Var} [Z_i^2] \right]^{\frac{1}{2}} = \sqrt{\frac{2}{n}}. \end{aligned} \quad (2.23)$$

For the final term of (2.18), because Y'_i is an independent copy of Y_i , then $Y'_i - Y_i \sim N(0, 2\sigma^2)$, with $\mathbb{E} |Y'_i - Y_i|^3 = 8 \frac{\sigma^3}{\sqrt{\pi}}$. Using that

$$(|a| + |b|)^3 \leq 4(|a|^3 + |b|^3), \quad a, b \in \mathbb{R} \quad (2.24)$$

yields

$$\begin{aligned}
& \sum_{i=1}^n \mathbb{E} \left[\left| \sum_{m=1}^2 \left| \sum_{l=1}^2 \frac{X_{il}}{\sigma} [X^\top X]^{-\frac{1}{2}} \right|_{ml} (Y_i - Y'_i) \right| \right]^3 \\
&= \sum_{i=1}^n \mathbb{E} \left[\left| \left(\frac{X_{i1}}{\sigma} [X^\top X]^{-\frac{1}{2}} \right)_{11} + \frac{X_{i2}}{\sigma} [X^\top X]^{-\frac{1}{2}} \right|_{22} (Y_i - Y'_i) \right|^3 \\
&\leq \sum_{i=1}^n \mathbb{E} \left[\left(\frac{1}{\sigma\sqrt{n}} + \frac{|x_i - \bar{x}|}{\sigma\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}} \right) |Y'_i - Y_i| \right]^3 \\
&\leq 4 \sum_{i=1}^n \left(\frac{8}{n^{\frac{3}{2}}\sqrt{\pi}} + \frac{8|x_i - \bar{x}|^3}{[\sum_{i=1}^n (x_i - \bar{x})^2]^{\frac{3}{2}}\sqrt{\pi}} \right) = \frac{32}{\sqrt{\pi}} \left(\frac{1}{\sqrt{n}} + \frac{\sum_{i=1}^n |x_i - \bar{x}|^3}{[\sum_{i=1}^n (x_i - \bar{x})^2]^{\frac{3}{2}}} \right). \tag{2.25}
\end{aligned}$$

Summarizing, in the case of Y_1, Y_2, \dots, Y_n being independent random variables with $Y_i \sim N(\beta_1 + \beta_2(x_i - \bar{x}), \sigma^2)$, we apply to (2.18) the results of (2.22), (2.23) and (2.25) to obtain the assertion of the corollary. ■

Empirical results

Here, we study the accuracy of our bounds by simulations. For $n = 10^j$, $j = 4, 5, 6, 7$, we start by generating 10^3 trials of n random independent observations, y , which follow $N(\beta_1 + \beta_2(x_i - \bar{x}), \sigma^2)$, where $\beta_1 = 1, \beta_2 = 2, \sigma^2 = 1$ and each x_i is sampled from the discrete uniform distribution in the set $\{1, 2, \dots, 100\}$. Then $\sqrt{n} [\bar{I}_n(\beta_0)]^{\frac{1}{2}} (\hat{\beta} - \beta_0)$ is evaluated in each trial, which in turn gives a vector of 10^3 values. We apply to these values the function $h(x, y) = (x^2 + y^2 + 1)^{-1}$ and we calculate their sample mean, denoted by $\hat{\mathbb{E}} \left[h \left(\sqrt{n} [\bar{I}_n(\beta_0)]^{\frac{1}{2}} (\hat{\beta} - \beta_0) \right) \right]$. The function h is a member of the class H as in (1.3) with

$$\|h\| = 1, \quad \|h\|_1 = \frac{3\sqrt{3}}{8}, \quad \|h\|_2 = 2, \quad \|h\|_3 < 4.7. \tag{2.26}$$

We use these values to calculate the bound in (2.20). We define

$$Q_h(\beta_0) := \left| \hat{\mathbb{E}} \left[h \left(\sqrt{n} [\bar{I}_n(\beta_0)]^{\frac{1}{2}} (\hat{\beta} - \beta_0) \right) \right] - \tilde{\mathbb{E}}[h(\mathbf{Z})] \right|,$$

where $\tilde{\mathbb{E}}[h(\mathbf{Z})] = 0.461$ is the approximation of $\mathbb{E}[h(\mathbf{Z})]$ up to three decimal places. We compare $Q_h(\beta_0)$ with the bound in (2.20), using the difference between their values as a measure of the error. The results are presented in Table 1 and are based on this particular function h , while the theoretical bounds that we have already given hold for any test function that belongs in the class H defined in (1.3).

TABLE 1
Simulation results for the simple linear model

n	$Q_n(\beta_0)$	Upper bound	Error
10^4	5×10^{-3}	0.207	0.202
10^5	3×10^{-3}	0.065	0.062
10^6	1×10^{-3}	0.022	0.021
10^7	3×10^{-4}	0.007	0.0067

The table indicates that the bound and the error decrease as the sample size gets larger. When at each step we increase the sample size by a factor of ten, then the value of the upper bound drops by approximately a $\sqrt{10}$ factor, which is expected as the expression in (2.20) is $\mathcal{O}(n^{-1/2})$.

2.3. Special case: Identically distributed random vectors

In this subsection we use weaker regularity conditions than (N1)-(N6) which were used in Theorem 2.1, in order to find an upper bound in the case of independent and identically distributed random vectors. Following Davison (2008) and for $\ell_{\mathbf{x}_i}(\boldsymbol{\theta}) := \log(f(\mathbf{x}_i|\boldsymbol{\theta}))$, we make the following assumptions:

(R.C.1) The densities defined by any two different values of $\boldsymbol{\theta}$ are distinct;

(R.C.2) $\ell(\boldsymbol{\theta}; \mathbf{x})$ is three times differentiable with respect to the unknown vector parameter, $\boldsymbol{\theta}$, and the third partial derivatives are continuous in $\boldsymbol{\theta}$;

(R.C.3) for any $\boldsymbol{\theta}_0 \in \Theta$ and for \mathbb{X} denoting the support of the data, there exists $\epsilon_0 > 0$ and functions $M_{rst}(\mathbf{x})$ (they can depend on $\boldsymbol{\theta}_0$), such that for $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_d)$ and $r, s, t, j = 1, 2, \dots, d$,

$$\left| \frac{\partial^3}{\partial \theta_r \partial \theta_s \partial \theta_t} \ell(\boldsymbol{\theta}; \mathbf{x}) \right| \leq M_{rst}(\mathbf{x}), \quad \forall \mathbf{x} \in \mathbb{X}, \quad |\theta_j - \theta_{0,j}| < \epsilon_0,$$

with $\mathbb{E}[M_{rst}(\mathbf{X})] < \infty$;

(R.C.4) for all $\boldsymbol{\theta} \in \Theta$, $\mathbb{E}_{\boldsymbol{\theta}}[\ell_{\mathbf{x}_i}(\boldsymbol{\theta})] = 0$;

(R.C.5) the expected Fisher information matrix for one random vector $I(\boldsymbol{\theta})$ is finite, symmetric and positive definite. For $r, s = 1, 2, \dots, d$, its elements satisfy

$$n[I(\boldsymbol{\theta})]_{rs} = \mathbb{E} \left\{ \frac{\partial}{\partial \theta_r} \ell(\boldsymbol{\theta}; \mathbf{X}) \frac{\partial}{\partial \theta_s} \ell(\boldsymbol{\theta}; \mathbf{X}) \right\} = \mathbb{E} \left\{ - \frac{\partial^2}{\partial \theta_r \partial \theta_s} \ell(\boldsymbol{\theta}; \mathbf{X}) \right\}.$$

This condition implies that $nI(\boldsymbol{\theta})$ is the covariance matrix of $\nabla(\ell(\boldsymbol{\theta}; \mathbf{x}))$.

These regularity conditions in the multi-parameter case resemble those in Anastasiou & Reinert (2017) where the parameter is scalar. Under (R.C.1)-(R.C.5), Davison (2008) shows that $\sqrt{n}[I(\boldsymbol{\theta}_0)]^{\frac{1}{2}} \left(\hat{\boldsymbol{\theta}}_n(\mathbf{X}) - \boldsymbol{\theta}_0 \right) \xrightarrow[n \rightarrow \infty]{d} \mathbf{Z}$. The upper

bound on the distributional distance between the distribution of a vector MLE and the multivariate normal in the case of i.i.d. random vectors is the same as the bound in Theorem 2.1 and thus it is not given again. The bound can be simplified because in the i.i.d. case $\bar{I}_n(\boldsymbol{\theta}_0) = I(\boldsymbol{\theta}_0)$ and $f_i(\mathbf{x}_i) = f(\mathbf{x}_i)$, $\forall i \in \{1, 2, \dots, n\}$. In the next example of independent random variables from the normal distribution under canonical parametrisation with both natural parameters unknown, the bound can be easily calculated and it is, as expected, of the order $\mathcal{O}(n^{-1/2})$.

Example: The normal distribution under canonical parametrisation

Many popular distributions which have the same underlying structure based on simple properties are exponential families, such as the normal, Gamma and Beta distributions; generalisations of exponential families can be found in Lauritzen (1988) and Berk (1972). Most of the times, the interest is on working under the canonical parametrisation; the distribution of a random variable, X , is said to be a canonical *multi-parameter exponential family distribution* if, for $\boldsymbol{\eta} \in \mathbb{R}^d$, the probability density (or mass) function is of the form

$$f(x|\boldsymbol{\eta}) = \exp \left\{ \sum_{j=1}^d \eta_j T_j(x) - A(\boldsymbol{\eta}) + S(x) \right\} \mathbb{1}_{\{x \in B\}},$$

where the set $B = \{x : f(x|\theta) > 0\}$ is the support of X and does not depend on $\boldsymbol{\eta}$; $A(\boldsymbol{\eta})$ is a function of the parameter; $T_j(x)$ and $S(x)$ are functions only of the data. The vectors $\boldsymbol{\eta} = (\eta_1, \eta_2, \dots, \eta_d)$ and $\mathbf{T}(\mathbf{x}) = (T_1(x), T_2(x), \dots, T_d(x))$ are called the *natural parameter vector* and *natural sufficient statistic*, respectively. There is a number of reasons why the canonical parametrisation is more convenient. To start with, written in its canonical form, the probability density (or mass) function of an exponential family distribution has some convexity properties, which are then useful in dealing with moments and other functions of the natural sufficient statistic $\mathbf{T}(\mathbf{x})$. Furthermore, for each $j \in \{1, 2, \dots, d\}$, if X follows a canonical exponential family distribution, then $T_j(X)$ also follows an exponential family distribution and also

$$\mathbb{E}[T_j(X)] = \frac{\partial}{\partial \eta_j} A(\boldsymbol{\eta}), \quad \text{Cov}[T_k(X), T_j(X)] = \frac{\partial^2}{\partial \eta_k \partial \eta_j} A(\boldsymbol{\eta}), \quad 1 \leq k, j \leq d.$$

Apart from simplifying the theory and computational complexity in generalised linear models, there are other application areas, where natural exponential family distributions play a significant role. An example is the area of Gaussian graphical models (see Lauritzen (1996) for more information) and the precision matrix estimation (Massam et al., 2018).

Here, we apply Theorem 2.1 in the case of X_1, X_2, \dots, X_n independent and identically distributed random variables from $N(\mu, \sigma^2)$, which is an exponential family distribution. Due to the importance, as explained above, of the natural

parametrisation in exponential families, we are interested in

$$\boldsymbol{\eta}_0 = (\eta_1, \eta_2) = \left(\frac{1}{2\sigma^2}, \frac{\mu}{\sigma^2} \right), \tag{2.27}$$

which is the natural parameter vector. The MLE for $\boldsymbol{\eta}_0$ exists, it is unique and equal to $\hat{\boldsymbol{\eta}}(\mathbf{X}) = (\hat{\eta}_1, \hat{\eta}_2)^\top = \frac{n}{\sum_{i=1}^n (X_i - \bar{X})^2} \left(\frac{1}{2}, \bar{X} \right)^\top$; to see this, use the invariance property of the MLE and the result of Davison (2008), p.116, where the MLEs for μ and σ^2 are given. In contrast to Corollary 2.2, the MLE in the current example of the Gaussian distribution under canonical parametrisation is not a sum of random variables; therefore, classical Stein’s method approaches, which require that the quantity of interest is a sum, cannot be employed. It appears that our results are the first that can be applied for such cases where the vector MLE has a general form in order to get upper bounds on the absolute value of the difference of expectations on the class of functions H in (1.3). The results of Pinelis & Molzon (2016) and Pinelis (2017) can also be applied through the Delta method to give upper bounds only on the Kolmogorov distance though. The conditions (R.C.1)-(R.C.5) are satisfied. Corollary 2.3 provides a bound on the distributional distance of interest and the proof is in the Appendix.

Corollary 2.3. *Let X_1, X_2, \dots, X_n be i.i.d. random variables that follow the $N(\mu, \sigma^2)$ distribution. Let $\boldsymbol{\eta}_0$ be as in (2.27) and for ease of presentation, we denote $\alpha := \alpha(\eta_1, \eta_2) = \eta_1(1 + \sqrt{\eta_1})^2 + \eta_2^2$ and $\beta := \beta(\eta_1, \eta_2) = \eta_1(1 + \sqrt{\eta_1}) + \eta_2^2$. For $\mathbf{Z} \sim N_2(\mathbf{0}, I_{2 \times 2})$ and $h \in H$ as defined in (1.3), we have that for $n > 9$*

$$\begin{aligned} & \left| \mathbb{E} \left[h \left(\sqrt{n} [I(\boldsymbol{\eta}_0)]^{\frac{1}{2}} (\hat{\boldsymbol{\eta}}(\mathbf{X}) - \boldsymbol{\eta}_0) \right) \right] - \mathbb{E} [h(\mathbf{Z})] \right| \leq \frac{8 \|h\| \left((\eta_1^2 + \eta_2^2) (2n + 15) + 2n\eta_1 \right)}{\eta_1^2 (n-3)(n-5)} \\ & + \frac{\sqrt{2n}^{\frac{3}{2}} \|h\|_1}{\sqrt{\alpha} (n-5)(n-9)} \left\{ 2 \sqrt{\frac{130}{\eta_1} + \frac{1473\eta_2^2}{\eta_1^2}} \left((\eta_1 + |\eta_2|)(\eta_1 + 3|\eta_2| + 2\sqrt{\eta_1}) + \eta_1 \right) \right. \\ & \quad + \left(\frac{39\eta_2^2}{\eta_1^3} + \frac{10}{\eta_1^2} \right) \left(4|\eta_2^3| + \eta_1(2|\eta_2| + \eta_1)(3|\eta_2| + 2 + 2\sqrt{\eta_1}) + \eta_1^{\frac{5}{2}} + \eta_1^3 \right) \\ & \quad \left. + 156(\sqrt{\eta_1} + |\eta_2| + \eta_1) \left(1 + \frac{3(|\eta_2| + \frac{\eta_1}{2})^2}{\eta_1} \right) \right\} \\ & + \frac{\|h\|_2}{2\sqrt{2n}\alpha} \left\{ \sqrt{7} \left(\frac{\eta_2^2}{\eta_1} + 1 \right) \alpha + \eta_1 \eta_2^2 + \frac{\beta^2}{\eta_1} \right. \\ & \quad \left. + 3\sqrt{2}\eta_2 \left[\left(\alpha - \eta_2^2 \right) \left(5 + \frac{\eta_2^2}{\eta_1} \right)^2 + \beta^2 + \left(2\sqrt{\eta_1}\eta_2 + \frac{\alpha}{\eta_2} \right)^2 \left(\frac{5}{3} + \frac{\eta_2^2}{\eta_1} \right) \right]^{\frac{1}{2}} \right\} \\ & + \frac{64\sqrt{2}\|h\|_3}{3\sqrt{n}\alpha^{\frac{3}{2}}} \left\{ 18 \left(1 + \frac{\eta_2^3}{2\eta_1^{\frac{3}{2}}\sqrt{\pi}} \right) \left(\eta_1^{\frac{3}{2}} (1 + \sqrt{\eta_1})^3 + |\eta_2|^3 \right) + \frac{\eta_1^3 |\eta_2|^3 + \beta^3}{\sqrt{\pi}\eta_1^{\frac{3}{2}}} \right\}. \tag{2.28} \end{aligned}$$

Remark 2.3. **(1)** The rate of convergence of the upper bound in (2.28) is $\frac{1}{\sqrt{n}}$. Although the bound might seem complicated, the proof of the corollary shows that what is required for the derivation of the bound is basic calculation of expectations.

(2) As already mentioned, this example consists an indication of the advantages of our method in comparison to classical multivariate Stein's method results, which require that the quantity of interest is a sum of random variables. This is not the case in Corollary 2.3 because $\hat{\eta}_1 = \frac{n}{2 \sum_{i=1}^n (X_i - \bar{X})^2}$.

Empirical results

We carry out a large-scale simulation study to investigate the accuracy of the bound in (2.28). The procedure is similar to the one followed previously when we obtained empirical results related to the example of the simple linear model in Corollary 2.2. Therefore, we start by generating 10^3 trials of $n = 10^j, j = 5, 6, 7, 8$, random independent observations, y , following $N(\mu, \sigma^2)$, and the vector parameter of interest is $\boldsymbol{\eta}_0 = (\eta_1, \eta_2)$ as in (2.27). We take $\mu = 1, \sigma^2 = 1$ for our simulations. Then $\sqrt{n} [I(\boldsymbol{\eta}_0)]^{\frac{1}{2}} (\hat{\boldsymbol{\eta}}(\mathbf{X}) - \boldsymbol{\eta}_0)$ is evaluated in each trial, which in turn gives a vector of 10^3 values. The function $h(x, y) = (x^2 + y^2 + 1)^{-1}$, which belongs to the class H as in (1.3), is then applied to these values in order to get the sample mean, denoted by $\hat{\mathbb{E}} \left[h \left(\sqrt{n} [I(\boldsymbol{\eta}_0)]^{\frac{1}{2}} (\hat{\boldsymbol{\eta}}(\mathbf{X}) - \boldsymbol{\eta}_0) \right) \right]$. Using (2.26), we calculate the bound in (2.28). We define

$$Q_h(\boldsymbol{\eta}_0) := \left| \hat{\mathbb{E}} \left[h \left(\sqrt{n} [I(\boldsymbol{\eta}_0)]^{\frac{1}{2}} (\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0) \right) \right] - \tilde{\mathbb{E}}[h(\mathbf{Z})] \right|,$$

where $\tilde{\mathbb{E}}[h(\mathbf{Z})] = 0.461$ is the approximation of $\mathbb{E}[h(\mathbf{Z})]$ up to three decimal places. We compare $Q_h(\boldsymbol{\eta}_0)$ with the bound in (2.28), using the difference between their values as a measure of the error. The results from the simulations are shown in Table 2 below.

TABLE 2
Simulation results for the $N(1, 1)$ distribution under a canonical parametrisation

n	$Q_h(\boldsymbol{\eta}_0)$	Upper bound	Error
10^5	0.009	28.616	28.607
10^6	0.005	9.048	9.042
10^7	0.004	2.861	2.857
10^8	0.001	0.905	0.904

As in the results of Table 1, we also see here that the bound and the error decrease as the sample size gets larger. To be more precise, when at each step we increase the sample size by a factor of ten, the value of the upper bound drops by a factor close to $\sqrt{10}$, which is expected since the order of the bound is $\mathcal{O}(n^{-1/2})$, as can be seen from (2.28). In this example, the bounds are not as small as in Table 1, with the reason being that the expression for the bound in (2.28) is the result of a series of simplifications in order to obtain a relatively

compact representation; see the proof of Corollary 2.3 in the Appendix for the exact steps that lead to the expression in (2.28). The bound has more of a conceptual character and better constants are possible at the cost of a more complicated expression.

3. Bounds when the MLE is not known explicitly

The calculation of some of the terms in the general bound of Theorem 2.1 requires an analytic expression for the MLE. This fact generates problems in models where no closed-form solution to the maximization problem is known or available; in these cases, a numerical method, such as the Newton-Raphson algorithm, can often be used to approximate the MLE and a normal approximation is still of interest. In this section, we will first explain how, under some further assumptions, we can put the dependence of the bound on the MLE only through the MSE, $\mathbb{E} \left[\sum_{j=1}^d Q_j^2 \right]$. Then, the MSE will get bounded by a quantity which is independent of $\hat{\boldsymbol{\theta}}_n(\mathbf{X})$ and it can therefore be used to get upper bounds on the distributional distance of interest that can be applied when the vector MLE is not expressed in a closed-form. To the best of our knowledge, such bounds have not appeared before in the literature for the case of a vector MLE that can not be expressed in a closed form. The extra assumptions are

(Con.1) For an $\epsilon = \epsilon(\boldsymbol{\theta}_0) > 0$, the MLE is within an ϵ -neighbourhood of $\boldsymbol{\theta}_0$, in the sense that $\forall j \in \{1, 2, \dots, d\}$, $|\hat{\boldsymbol{\theta}}_n(\mathbf{X})_j - \theta_{0,j}| < \epsilon$;

(Con.2) for all $\boldsymbol{\theta}_0 \in \Theta$, where Θ is the open parameter space,

$$\sup_{\substack{\boldsymbol{\theta}: |\theta_q - \theta_{0,q}| < \epsilon \\ \forall q \in \{1, 2, \dots, d\}}} \left| \frac{\partial^3}{\partial \theta_k \partial \theta_j \partial \theta_i} \log f(\mathbf{x}_1 | \boldsymbol{\theta}) \right| \leq M_{kji},$$

where $M_{kji} = M_{kji}(\boldsymbol{\theta}_0)$ is a constant that may depend on $\boldsymbol{\theta}_0$;

(Con.3) the Hessian matrix of the second-order partial derivatives of the log-likelihood function is symmetric and invertible.

Section 2 gave an upper bound for the distributional distance between the distribution of the MLE and the multivariate normal distribution. As explained in the outline of the proof of Theorem 2.1, this bound in (2.6) can be split into terms coming from Stein's method, and terms due to Taylor expansions and conditional expectations. With \tilde{V} as in (2.5), for ease of presentation we abbreviate the terms coming from Stein's method by

$$D = D(\boldsymbol{\theta}_0, h, \mathbf{X}) := \frac{\|h\|_2}{4\sqrt{n}} \sum_{j=1}^d \left[\text{Var} \left[\left(\sum_{k=1}^d \tilde{V}_{jk} \frac{\partial}{\partial \theta_k} \log f(\mathbf{X}_1 | \boldsymbol{\theta}_0) \right)^2 \right] \right]^{\frac{1}{2}}$$

$$\begin{aligned}
& + \frac{\|h\|_2}{2\sqrt{n}} \sum_{k=1}^{d-1} \sum_{j=k+1}^d \left[\text{Var} \left[\sum_{q=1}^d \sum_{v=1}^d \tilde{V}_{jq} \frac{\partial}{\partial \theta_q} \log f(\mathbf{X}_1 | \boldsymbol{\theta}_0) \tilde{V}_{kv} \frac{\partial}{\partial \theta_v} \log f(\mathbf{X}_1 | \boldsymbol{\theta}_0) \right] \right]^{\frac{1}{2}} \\
& + \frac{\|h\|_3}{12\sqrt{n}} \mathbb{E} \left[\sum_{i=1}^d \left| \sum_{l=1}^d \tilde{V}_{il} \left(\frac{\partial}{\partial \theta_l} \log f(\mathbf{X}'_1 | \boldsymbol{\theta}_0) - \frac{\partial}{\partial \theta_l} \log f(\mathbf{X}_1 | \boldsymbol{\theta}_0) \right) \right|^3 \right]. \quad (3.29)
\end{aligned}$$

We will now first explain how we can put the dependence of the general bound in (2.6) on MLE only through the quantity $\mathbb{E} \left[\sum_{j=1}^d Q_j^2 \right]$ with Q_j as in (2.5). After that, we will give an upper bound for $\mathbb{E} \left[\sum_{j=1}^d Q_j^2 \right]$.

A bound depending on the mean squared error: Under (Con.1) and with \tilde{V} , $Q_{(m)}$, Q_j and T_{kj} as in (2.5), and for D in (3.29), we obtain, using (2.6), that

$$\begin{aligned}
& \left| \mathbb{E} \left[h \left(\sqrt{n} [I(\boldsymbol{\theta}_0)]^{\frac{1}{2}} (\hat{\boldsymbol{\theta}}_n(\mathbf{X}) - \boldsymbol{\theta}_0) \right) \right] - \mathbb{E}[h(\mathbf{Z})] \right| \leq D \\
& + \frac{\|h\|_1}{\sqrt{n}} \sum_{k=1}^d \sum_{l=1}^d |\tilde{V}_{lk}| \sum_{j=1}^d \left[\mathbb{E} [Q_j^2] \mathbb{E} [T_{kj}^2] \right]^{\frac{1}{2}} \quad (3.30)
\end{aligned}$$

$$+ \frac{\|h\|_1}{2\sqrt{n}} \left\{ \sum_{k=1}^d \sum_{l=1}^d |\tilde{V}_{lk}| \mathbb{E} \left[\sum_{j=1}^d \sum_{i=1}^d Q_j Q_i \frac{\partial^3}{\partial \theta_k \partial \theta_j \partial \theta_i} \ell(\boldsymbol{\theta}_0^*; \mathbf{X}) \right] \right\}. \quad (3.31)$$

Step 1: Upper bound for (3.30). Since $\mathbb{E} [T_{kj}] = 0$, $\forall j, k \in \{1, 2, \dots, d\}$,

$$\begin{aligned}
(3.30) & = \|h\|_1 \sum_{k=1}^d \sum_{l=1}^d |\tilde{V}_{lk}| \sum_{j=1}^d \sqrt{\mathbb{E} [Q_j^2]} \sqrt{\text{Var} \left[\frac{\partial^2}{\partial \theta_k \partial \theta_j} \log f(\mathbf{X}_1 | \boldsymbol{\theta}_0) \right]} \\
& \leq \|h\|_1 \sum_{k=1}^d \sum_{l=1}^d |\tilde{V}_{lk}| \sum_{j=1}^d \sqrt{\mathbb{E} [Q_j^2]} \sqrt{\sum_{i=1}^d \text{Var} \left[\frac{\partial^2}{\partial \theta_k \partial \theta_i} \log f(\mathbf{X}_1 | \boldsymbol{\theta}_0) \right]}, \quad (3.32)
\end{aligned}$$

where the inequality comes from the trivial bound

$$\text{Var} \left[\frac{\partial^2}{\partial \theta_k \partial \theta_j} \log f(\mathbf{X}_1 | \boldsymbol{\theta}_0) \right] \leq \sum_{i=1}^d \text{Var} \left[\frac{\partial^2}{\partial \theta_k \partial \theta_i} \log f(\mathbf{X}_1 | \boldsymbol{\theta}_0) \right]$$

since the variance of a random variable is always non-negative. Now, using that $\left(\sum_{j=1}^d \alpha_j \right)^2 \leq d \left(\sum_{j=1}^d \alpha_j^2 \right)$ for $\alpha_j \in \mathbb{R}$, yields

$$\left(\sum_{j=1}^d \sqrt{\mathbb{E} [Q_j^2]} \right)^2 \leq d \sum_{j=1}^d \mathbb{E} [Q_j^2].$$

Taking square roots in both sides of the above inequality and applying this result to (3.32) gives

$$(3.30) \leq \|h\|_1 \sqrt{d} \sum_{k=1}^d \sum_{l=1}^d |\tilde{V}_{lk}| \sqrt{\sum_{i=1}^d \text{Var} \left[\frac{\partial^2}{\partial \theta_k \partial \theta_i} \log f(\mathbf{X}_1 | \boldsymbol{\theta}_0) \right]} \sqrt{\mathbb{E} \left[\sum_{j=1}^d Q_j^2 \right]}. \quad (3.33)$$

Step 2: Upper bound for (3.31). Notice that from (Con.2), $\left| \frac{\partial^3}{\partial \theta_k \partial \theta_j \partial \theta_i} \ell(\boldsymbol{\theta}_0^*; \mathbf{x}) \right| = \left| \sum_{l=1}^n \frac{\partial^3}{\partial \theta_k \partial \theta_j \partial \theta_i} \log f(\mathbf{x}_l | \boldsymbol{\theta}_0^*) \right| \leq n M_{kji}$. Also,

$$\sum_{j=1}^d \sum_{i=1}^d |Q_j Q_i| M_{kji} = \sum_{j=1}^d Q_j^2 M_{kjj} + 2 \sum_{i=1}^{d-1} \sum_{j=i+1}^d |Q_j| |Q_i| M_{kij}.$$

Using now that $2\alpha\beta \leq \alpha^2 + \beta^2, \forall \alpha, \beta \in \mathbb{R}$,

$$\begin{aligned} \sum_{j=1}^d \sum_{i=1}^d |Q_j Q_i| M_{kji} &\leq \sum_{j=1}^d Q_j^2 M_{kjj} + \sum_{i=1}^{d-1} \sum_{j=i+1}^d [Q_j^2 + Q_i^2] M_{kji} = \sum_{j=1}^d Q_j^2 \sum_{i=1}^d M_{kji} \\ &\leq \sum_{j=1}^d Q_j^2 \sum_{m=1}^d \sum_{i=1}^d M_{kmi}. \end{aligned} \quad (3.34)$$

Using (3.34) yields

$$(3.31) \leq \frac{\|h\|_1 \sqrt{n}}{2} \sum_{k=1}^d \sum_{l=1}^d |\tilde{V}_{lk}| \sum_{m=1}^d \sum_{i=1}^d M_{kmi} \mathbb{E} \left[\sum_{j=1}^d Q_j^2 \right]. \quad (3.35)$$

Hence, from (3.33) and (3.35),

$$\begin{aligned} &\left| \mathbb{E} \left[h \left(\sqrt{n} [I(\boldsymbol{\theta}_0)]^{\frac{1}{2}} (\hat{\boldsymbol{\theta}}_n(\mathbf{X}) - \boldsymbol{\theta}_0) \right) \right] - \mathbb{E}[h(\mathbf{Z})] \right| \leq D \\ &+ \|h\|_1 \sqrt{d} \sum_{k=1}^d \sum_{l=1}^d |\tilde{V}_{lk}| \sqrt{\sum_{i=1}^d \text{Var} \left[\frac{\partial^2}{\partial \theta_k \partial \theta_i} \log f(\mathbf{X}_1 | \boldsymbol{\theta}_0) \right]} \sqrt{\mathbb{E} \left[\sum_{j=1}^d Q_j^2 \right]} \\ &+ \frac{\|h\|_1 \sqrt{n}}{2} \sum_{k=1}^d \sum_{l=1}^d |\tilde{V}_{lk}| \sum_{m=1}^d \sum_{i=1}^d M_{kmi} \mathbb{E} \left[\sum_{j=1}^d Q_j^2 \right]. \end{aligned} \quad (3.36)$$

Since D as defined in (3.29), is not related to the MLE, the upper bound in (3.36) depends on $\hat{\boldsymbol{\theta}}_n(\mathbf{X})$ only through $\mathbb{E} \left[\sum_{j=1}^d Q_j^2 \right]$. Our purpose now is to find a bound for $\mathbb{E} \left[\sum_{j=1}^d Q_j^2 \right]$ that does not contain any terms related to $\hat{\boldsymbol{\theta}}_n(\mathbf{X})$.

A bound on the mean squared error: In order to give an upper bound

when $\hat{\boldsymbol{\theta}}_n(\mathbf{X})$ is not known explicitly but (Con.1)-(Con.3) are satisfied, we bound $\mathbb{E} \left[\sum_{j=1}^d Q_j^2 \right]$, for Q_j as in (2.5), by a quantity which does not require knowledge of the MLE. The result is given in Theorem 3.1 below, followed by the proof.

Theorem 3.1. *Let $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$ be i.i.d. \mathbb{R}^t -valued random elements, for $t \in \mathbb{N}$, with probability density (or mass) function $f(\mathbf{x}_i|\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is the d -valued vector parameter. Assume that (Con.1), (Con.3) are satisfied. We assume existence and uniqueness of $\hat{\boldsymbol{\theta}}_n(\mathbf{X})$. For $J(\mathbf{x}, \boldsymbol{\theta}) = \left\{ \frac{\partial^2}{\partial \theta_i \partial \theta_j} \ell(\boldsymbol{\theta}; \mathbf{x}) \right\}_{i,j=1,2,\dots,d}$, the Hessian matrix, it holds that*

$$\begin{aligned} \mathbb{E} \left[\sum_{j=1}^d Q_j^2 \right] &\leq \mathbb{E} \left[\sum_{k=1}^d \sum_{q=1}^d \frac{\partial}{\partial \theta_k} \ell(\boldsymbol{\theta}_0; \mathbf{X}) \frac{\partial}{\partial \theta_q} \ell(\boldsymbol{\theta}_0; \mathbf{X}) \sup_{\substack{\boldsymbol{\theta}: |\theta_j - \theta_{0,j}| < \epsilon \\ \forall j \in \{1,2,\dots,d\}}} \left\{ [J^{-2}(\mathbf{X}, \boldsymbol{\theta})]_{kq} \right\} \right] \\ &=: U_1. \end{aligned} \quad (3.37)$$

Proof. From the definition of the MLE, we have that $\frac{\partial}{\partial \theta_k} \ell(\hat{\boldsymbol{\theta}}_n(\mathbf{x}); \mathbf{x}) = 0$, $\forall k \in \{1, 2, \dots, d\}$. A first-order Taylor expansion of $\frac{\partial}{\partial \theta_k} \ell(\hat{\boldsymbol{\theta}}_n(\mathbf{x}); \mathbf{x})$ about $\boldsymbol{\theta}_0$ leads to

$$\sum_{j=1}^d (\hat{\theta}_n(\mathbf{x})_j - \theta_{0,j}) \frac{\partial^2}{\partial \theta_k \partial \theta_j} \ell(\tilde{\boldsymbol{\theta}}; \mathbf{x}) = - \frac{\partial}{\partial \theta_k} \ell(\boldsymbol{\theta}_0; \mathbf{x}), \quad (3.38)$$

where $\tilde{\boldsymbol{\theta}}$ is between $\boldsymbol{\theta}_0$ and $\hat{\boldsymbol{\theta}}_n(\mathbf{x})$. Since the result in (3.38) holds $\forall k \in \{1, 2, \dots, d\}$, we deduce that

$$\hat{\boldsymbol{\theta}}_n(\mathbf{x}) - \boldsymbol{\theta}_0 = - \left[J(\tilde{\boldsymbol{\theta}}; \mathbf{x}) \right]^{-1} \nabla(\ell(\boldsymbol{\theta}_0; \mathbf{x}))$$

and therefore

$$\left(\hat{\boldsymbol{\theta}}_n(\mathbf{x}) - \boldsymbol{\theta}_0 \right)^\top \left(\hat{\boldsymbol{\theta}}_n(\mathbf{x}) - \boldsymbol{\theta}_0 \right) = \left[\nabla(\ell(\boldsymbol{\theta}_0; \mathbf{x})) \right]^\top \left[J(\tilde{\boldsymbol{\theta}}; \mathbf{x}) \right]^{-2} (\ell(\boldsymbol{\theta}_0; \mathbf{x})).$$

Going a step further and using (Con.1), we get that

$$\begin{aligned} &\mathbb{E} \left[\left(\hat{\boldsymbol{\theta}}_n(\mathbf{X}) - \boldsymbol{\theta}_0 \right)^\top \left(\hat{\boldsymbol{\theta}}_n(\mathbf{X}) - \boldsymbol{\theta}_0 \right) \right] \\ &\leq \mathbb{E} \left[\left[\nabla(\ell(\boldsymbol{\theta}_0; \mathbf{X})) \right]^\top \sup_{\substack{\boldsymbol{\theta}: |\theta_j - \theta_{0,j}| < \epsilon \\ \forall j \in \{1,2,\dots,d\}}} \left\{ [J(\boldsymbol{\theta}; \mathbf{X})]^{-2} \right\} (\ell(\boldsymbol{\theta}_0; \mathbf{X})) \right], \end{aligned}$$

which finishes the proof. ■

Remark 3.1. **(1)** As the bound (3.37) does not include $\hat{\boldsymbol{\theta}}_n(\mathbf{X})$, in cases where a closed-form expression for the vector MLE is not available, we can still get an upper bound on the distributional distance between the distribution of the MLE and the d -variate standard normal, under the assumptions (R.C.1)-(R.C.5) and (Con.1)-(Con.3). Combining the results in (3.36) and (3.37) and for D as in (3.29) and U_1 as in (3.37), we obtain that

$$\begin{aligned} & \left| \mathbb{E} \left[h \left(\sqrt{n} [I(\boldsymbol{\theta}_0)]^{\frac{1}{2}} (\hat{\boldsymbol{\theta}}_n(\mathbf{X}) - \boldsymbol{\theta}_0) \right) \right] - \mathbb{E}[h(\mathbf{Z})] \right| \leq D \\ & + \|h\|_1 \sqrt{d} U_1 \sum_{k=1}^d \sum_{l=1}^d \left| \tilde{V}_{lk} \right| \sqrt{\sum_{i=1}^d \text{Var} \left[\frac{\partial^2}{\partial \theta_k \partial \theta_i} \log f(\mathbf{X}_1 | \boldsymbol{\theta}_0) \right]} \\ & + \frac{\|h\|_1 \sqrt{n}}{2} U_1 \sum_{k=1}^d \sum_{l=1}^d \left| \tilde{V}_{lk} \right| \sum_{m=1}^d \sum_{i=1}^d M_{kmi}. \end{aligned} \tag{3.39}$$

(2) In the special case where the second-order partial derivatives of the log-likelihood function do not depend on \mathbf{x} , then the result can be simplified, since in such scenarios $J(\boldsymbol{\theta}; \mathbf{X}) = -n[I(\boldsymbol{\theta})]$ and $[J(\boldsymbol{\theta}; \mathbf{X})]^{-2} = \frac{1}{n^2} [I(\boldsymbol{\theta})]^{-2}$. Applying this on (3.37), leads to

$$U_1 = \frac{1}{n^2} \sum_{k=1}^d \sum_{q=1}^d \sup_{\substack{\boldsymbol{\theta}: |\theta_j - \theta_{0,j}| < \epsilon \\ \forall j \in \{1, 2, \dots, d\}}} \left\{ [I^{-2}(\boldsymbol{\theta})]_{kq} \right\} \mathbb{E} \left[\frac{\partial}{\partial \theta_k} \ell(\boldsymbol{\theta}_0; \mathbf{X}) \frac{\partial}{\partial \theta_q} \ell(\boldsymbol{\theta}_0; \mathbf{X}) \right]. \tag{3.40}$$

Example: The Beta distribution

Here, we briefly explain how we can calculate U_1 in (3.37) for the specific example of i.i.d. random variables from the Beta distribution with both shape parameters unknown. An analytic expression for the MLE is not available. Let $\Psi_j(\cdot)$ to be the j^{th} derivative of the digamma function Ψ , with $\Psi(z) = \frac{\Gamma'(z)}{\Gamma(z)}$, $z > 0$. The function $\Psi_j(z)$ can be defined through a sum, with

$$\Psi_m(z) = (-1)^{m+1} m! \sum_{k=0}^{\infty} \frac{1}{(z+k)^{m+1}}, \text{ for } z \in \mathbb{C} \setminus \{\mathbb{Z}_0^-\} \text{ and } m > 0. \tag{3.41}$$

Corollary 3.1 gives the bound U_1 for the MSE in the case of the Beta distribution. The proof is given in the Appendix. For ease of presentation, and for $x, y > 0$, allow us from now on to denote by

$$\begin{aligned} \delta_I & := \delta_I(\alpha, \beta) = \Psi_1(\alpha)\Psi_1(\beta) - \Psi_1(\alpha + \beta) (\Psi_1(\alpha) + \Psi_1(\beta)) \\ C_1(x, y) & := \Psi_1(x) - \Psi_1(x + y). \end{aligned} \tag{3.42}$$

Corollary 3.1. *Let X_1, X_2, \dots, X_n be i.i.d. random variables from the Beta(α, β) distribution with $\boldsymbol{\theta}_0 = (\alpha, \beta)$. Under (Con.1)-(Con.3) and with U_1 as in (3.37)*

and $\delta_I, C_1(x, y)$ as in (3.42), we get that

$$U_1 = \frac{1}{n[\delta_I(\alpha + \epsilon, \beta + \epsilon)]^2} \left\{ \begin{aligned} &C_1(\alpha, \beta) [(\alpha + \epsilon)^2 \Psi_2^2(\beta - \epsilon) + \Psi_1^2(\alpha + \beta - 2\epsilon)] \\ &+ C_1(\beta, \alpha) [(\beta + \epsilon)^2 \Psi_2^2(\alpha - \epsilon) + \Psi_1^2(\alpha + \beta - 2\epsilon)] \\ &+ 2\Psi_1(\alpha + \beta) [(\beta + \epsilon) \Psi_2(\alpha - \epsilon) + (\alpha + \epsilon) \Psi_2(\beta - \epsilon)] \end{aligned} \right\}. \quad (3.43)$$

Remark 3.2. This bound basically relies on the calculation of the expressions defined in (3.37). It can be easily seen that it is of order $\mathcal{O}(n^{-1})$. We deduce that, if we use the result in (3.43) in order to calculate the bound in (3.39) for the specific case of the Beta distribution, then the obtained bound will be, as expected, of order $\mathcal{O}(n^{-1/2})$.

4. Proof of Theorem 2.1

In this section, the complete steps of the proof of the main theorem of our paper are given. The following lemma (special case of Chebyshev's 'other' inequality) is useful for bounding conditional expectations, which sometimes can be difficult to derive. The proof is given in the Appendix.

Lemma 4.1. *Let $\mathbf{M} \in \mathbb{R}^d$ be a random vector with $M_i > 0 \forall i = 1, 2, \dots, d$ and $\epsilon > 0$. For every continuous function $f: \mathbb{R}^d \rightarrow \mathbb{R}$ such that $f(\mathbf{m})$ is increasing and $f(\mathbf{m}) \geq 0$, for $m_i > 0 \forall i \in \{1, 2, \dots, d\}$, where $\mathbf{m} = (m_1, m_2, \dots, m_d)$,*

$$\mathbb{E}[f(\mathbf{M}) | M_i < \epsilon \forall i = 1, 2, \dots, d] \leq \mathbb{E}[f(\mathbf{M})].$$

Proof of Theorem 2.1. It has already been shown in the outline of the proof that the triangle inequality yields

$$\left| \mathbb{E} \left[h \left(\sqrt{n} [\bar{I}_n(\boldsymbol{\theta}_0)]^{\frac{1}{2}} \left(\hat{\boldsymbol{\theta}}_n(\mathbf{X}) - \boldsymbol{\theta}_0 \right) \right) \right] - \mathbb{E}[h(\mathbf{Z})] \right| \leq (2.14) + (2.15).$$

Step 1: Upper bound for (2.14). First, $\nabla(\ell(\boldsymbol{\theta}_0; \mathbf{x})) = \sum_{i=1}^n \nabla(\log(f_i(\mathbf{x}_i | \boldsymbol{\theta}_0)))$ due to independence. With \tilde{V} as in (2.5), the results of Theorem 2.1 of Reinert & Röllin (2009) will be used for

$$\mathbf{W} = \frac{1}{\sqrt{n}} \tilde{V} \sum_{i=1}^n \nabla(\log(f_i(\mathbf{X}_i | \boldsymbol{\theta}_0))) = (W_1, W_2, \dots, W_d)^\top \in \mathbb{R}^{d \times 1}. \quad (4.44)$$

From (4.44) we have that for all $k \in \{1, 2, \dots, d\}$, $W_k = \sum_{i=1}^n \xi_{ik}$, with ξ_{ik} as in (2.5). From the regularity conditions, $\mathbb{E}[\nabla(\ell(\boldsymbol{\theta}_0; \mathbf{X}))] = \mathbf{0}$ and thus $\mathbb{E}[\mathbf{W}] = \mathbf{0}$.

Also, $\bar{I}_n(\boldsymbol{\theta}_0)$ is symmetric. Therefore, \tilde{V} is also symmetric. Using the regularity conditions we know that $\sum_{i=1}^n \text{Cov}[\nabla(\log(f_i(\mathbf{X}_i|\boldsymbol{\theta}_0)))] = n\bar{I}_n(\boldsymbol{\theta}_0)$ and basic calculations show that $\text{Cov}[\mathbf{W}] = I_{d \times d}$. Since $\mathbb{E}[\mathbf{W}] = \mathbf{0}$ and $\mathbb{E}[\mathbf{W}\mathbf{W}^\top] = I_{d \times d}$, the first assumption of Theorem 2.1 from Reinert & Röllin (2009) is satisfied. This theorem also assumes that $\exists \mathbf{W}'$ such that $(\mathbf{W}, \mathbf{W}')$ is an exchangeable pair meaning that $(\mathbf{W}, \mathbf{W}') \stackrel{d}{=} (\mathbf{W}', \mathbf{W})$, where $\stackrel{d}{=}$ denotes equality in distribution. In addition, it is assumed that

$$\mathbb{E}[\mathbf{W}' - \mathbf{W} | \mathbf{W}] = -\Lambda \mathbf{W} + \mathbf{R} \tag{4.45}$$

for an invertible $d \times d$ matrix Λ and a $\sigma(\mathbf{W})$ -measurable random vector \mathbf{R} . To define \mathbf{W}' in our case such that (4.45) is satisfied, let $\{\mathbf{X}'_i, i = 1, 2, \dots, n\}$ be an independent copy of $\{\mathbf{X}_i, i = 1, 2, \dots, n\}$ and let the index $I \in \{1, 2, \dots, n\}$ follow the uniform distribution on $\{1, 2, \dots, n\}$, independently of the set $\{\mathbf{X}_i, \mathbf{X}'_i, i = 1, 2, \dots, n\}$. Let

$$\xi'_{ik} = \frac{1}{\sqrt{n}} \sum_{j=1}^d \tilde{V}_{kj} \frac{\partial}{\partial \theta_j} \log(f_i(\mathbf{X}'_i|\boldsymbol{\theta}_0))$$

and

$$W'_k = W_k - \xi_{Ik} + \xi'_{Ik}, \quad \forall k \in \{1, 2, \dots, d\},$$

with $\mathbb{E}[W'_k - W_k | \mathbf{W}] = \mathbb{E}[\xi'_{Ik} - \xi_{Ik} | \mathbf{W}] = -\mathbb{E}[\xi_{Ik} | \mathbf{W}] = -\frac{1}{n} \sum_{i=1}^n \mathbb{E}[\xi_{ik} | \mathbf{W}] = -\frac{W_k}{n}$. Hence (4.45) is satisfied with $\Lambda = \frac{1}{n} I_{d \times d}$ and $\mathbf{R} = \mathbf{0}$. Therefore, Theorem 2.1 from Reinert & Röllin (2009) gives in our case that

$$\begin{aligned} & |\mathbb{E}[h(\mathbf{W})] - \mathbb{E}[h(\mathbf{Z})]| \\ & \leq n \left(\frac{\|h\|_2}{4} \sum_{i=1}^d \sum_{j=1}^d [\text{Var}[\mathbb{E}[(W'_i - W_i)(W'_j - W_j) | \mathbf{W}]]] \right)^{\frac{1}{2}} \end{aligned} \tag{4.46}$$

$$+ n \left(\frac{\|h\|_3}{12} \sum_{i=1}^d \sum_{j=1}^d \sum_{k=1}^d \mathbb{E}|(W'_i - W_i)(W'_j - W_j)(W'_k - W_k)| \right). \tag{4.47}$$

To bound the variance of the conditional expectations in (4.46), let $\mathcal{A} = \sigma(\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n)$. Since $\sigma(\mathbf{W}) \subset \mathcal{A}$, for any random variable Y , we have that $\text{Var}[\mathbb{E}[Y | \mathbf{W}]] \leq \text{Var}[\mathbb{E}[Y | \mathcal{A}]]$. Then,

$$\begin{aligned} (4.46) & \leq n \frac{\|h\|_2}{4} \left\{ \sum_{j=1}^d \sqrt{\text{Var}[\mathbb{E}[(\xi'_{Ij} - \xi_{Ij})^2 | \mathcal{A}]]} \right. \\ & \quad \left. + 2 \sum_{k=1}^{d-1} \sum_{j=k+1}^d \sqrt{\text{Var}[\mathbb{E}[(\xi'_{Ik} - \xi_{Ik})(\xi'_{Ij} - \xi_{Ij}) | \mathcal{A}]]} \right\}. \end{aligned} \tag{4.48}$$

Since $\{\mathbf{X}'_i, i = 1, 2, \dots, n\}$ is an independent copy of $\{\mathbf{X}_i, i = 1, 2, \dots, n\}$ and ξ'_{ik} is independent of \mathcal{A} ,

$$\begin{aligned}
 (4.48) &= n \frac{\|h\|_2}{4} \left\{ \sum_{j=1}^d [\text{Var} [\mathbb{E} [(\xi'_{I_j})^2] - 2\mathbb{E}[\xi'_{I_j}]\mathbb{E}[\xi_{I_j}|\mathcal{A}] + \mathbb{E} [\xi_{I_j}^2|\mathcal{A}]]] \right\}^{\frac{1}{2}} \\
 &+ 2 \sum_{k=1}^{d-1} \sum_{j=k+1}^d [\text{Var} [\mathbb{E} [\xi'_{Ik}\xi'_{I_j}] - \mathbb{E} [\xi'_{I_j}] \mathbb{E} [\xi_{Ik}|\mathcal{A}] - \mathbb{E} [\xi'_{Ik}] \mathbb{E} [\xi_{I_j}|\mathcal{A}] \\
 &+ \mathbb{E} [\xi_{Ik}\xi_{I_j}|\mathcal{A}]]] \right\}^{\frac{1}{2}}. \tag{4.49}
 \end{aligned}$$

Using that $\mathbb{E} [\xi'_{ik}] = 0$,

$$\begin{aligned}
 (4.49) &= n \frac{\|h\|_2}{4} \left\{ \sum_{j=1}^d \left[\frac{1}{n^2} \text{Var} \left[\sum_{i=1}^n \mathbb{E} [\xi_{ij}^2|\mathcal{A}] \right] \right]^{\frac{1}{2}} \right. \\
 &\quad \left. + 2 \sum_{k=1}^{d-1} \sum_{j=k+1}^d \left[\frac{1}{n^2} \text{Var} \left[\sum_{i=1}^n \mathbb{E} [\xi_{ik}\xi_{ij}|\mathcal{A}] \right] \right]^{\frac{1}{2}} \right\} \\
 &= \frac{\|h\|_2}{4} \left\{ \sum_{j=1}^d \left[\text{Var} \left[\sum_{i=1}^n \xi_{ij}^2 \right] \right]^{\frac{1}{2}} + 2 \sum_{k=1}^{d-1} \sum_{j=k+1}^d \left[\text{Var} \left[\sum_{i=1}^n \xi_{ik}\xi_{ij} \right] \right]^{\frac{1}{2}} \right\} \\
 &= \frac{\|h\|_2}{\sqrt{n}} K_2(\boldsymbol{\theta}_0),
 \end{aligned}$$

with $K_2(\boldsymbol{\theta}_0)$ defined in (2.8). For (4.47), using again the definition of ξ_{ik} in (2.5), after basic calculations we obtain that

$$(4.47) \leq \frac{\|h\|_3}{\sqrt{n}} K_3(\boldsymbol{\theta}_0),$$

with $K_3(\boldsymbol{\theta}_0)$ as in (2.9). Therefore,

$$(2.14) \leq \frac{\|h\|_2}{\sqrt{n}} K_2(\boldsymbol{\theta}_0) + \frac{\|h\|_3}{\sqrt{n}} K_3(\boldsymbol{\theta}_0). \tag{4.50}$$

Step 2: Upper bound for (2.15). With \tilde{V} as in (2.5), for ease of presentation let us denote by

$$\begin{aligned}
 \mathbf{R}_1(\boldsymbol{\theta}_0; \mathbf{x}) &= \frac{1}{2\sqrt{n}} \tilde{V} \sum_{j=1}^d \sum_{q=1}^d Q_j Q_q \left(\nabla \left(\frac{\partial^2}{\partial \theta_j \partial \theta_q} \ell(\boldsymbol{\theta}; \mathbf{x}) \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0^*} \right) \right) \\
 T_1 &= T_1(\boldsymbol{\theta}_0; \mathbf{X}, h) := h \left(\sqrt{n} [\bar{I}_n(\boldsymbol{\theta}_0)]^{\frac{1}{2}} \left(\hat{\boldsymbol{\theta}}_n(\mathbf{X}) - \boldsymbol{\theta}_0 \right) \right) \\
 &\quad - h \left(\frac{1}{\sqrt{n}} \tilde{V} (\nabla(\ell(\boldsymbol{\theta}_0; \mathbf{x}))) + \mathbf{R}_1(\boldsymbol{\theta}_0; \mathbf{X}) \right)
 \end{aligned}$$

$$T_2 = T_2(\boldsymbol{\theta}_0; \mathbf{X}, h) := h \left(\frac{1}{\sqrt{n}} \tilde{V}(\nabla(\ell(\boldsymbol{\theta}_0; \mathbf{x}))) + \mathbf{R}_1(\boldsymbol{\theta}_0; \mathbf{x}) \right) - h \left(\frac{1}{\sqrt{n}} \tilde{V}(\nabla(\ell(\boldsymbol{\theta}_0; \mathbf{X}))) \right). \tag{4.51}$$

Using the above notation and the triangle inequality

$$(2.15) = |\mathbb{E}[T_1 + T_2]| \leq \mathbb{E}|T_1| + \mathbb{E}|T_2|.$$

With $A_{[j]}$ the j^{th} row of a matrix A , a first order multivariate Taylor expansion gives that

$$|T_1| \leq \|h\|_1 \left| \sum_{j=1}^d \left(\sqrt{n} [\bar{I}_n(\boldsymbol{\theta}_0)]^{\frac{1}{2}} \right)_{[j]} (\hat{\boldsymbol{\theta}}_n(\mathbf{X}) - \boldsymbol{\theta}_0) - \frac{1}{\sqrt{n}} \tilde{V}_{[j]} \nabla(\ell(\boldsymbol{\theta}_0; \mathbf{X})) - \frac{1}{2\sqrt{n}} \tilde{V}_{[j]} \left\{ \sum_{k=1}^d \sum_{q=1}^d Q_k Q_q \left(\nabla \left(\frac{\partial^2}{\partial \theta_k \partial \theta_q} \ell(\boldsymbol{\theta}; \mathbf{x}) \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0^*} \right) \right\} \right) \right|.$$

Using (2.13) component-wise and the Cauchy-Schwarz inequality, we have that

$$\mathbb{E}|T_1| \leq \frac{\|h\|_1}{\sqrt{n}} \sum_{k=1}^d \sum_{l=1}^d |\tilde{V}_{lk}| \sum_{j=1}^d \sqrt{\mathbb{E}[Q_j^2] \mathbb{E} \left[\left(\frac{\partial^2}{\partial \theta_j \partial \theta_k} \ell(\boldsymbol{\theta}_0; \mathbf{X}) + n[\bar{I}_n(\boldsymbol{\theta}_0)]_{kj} \right)^2 \right]}. \tag{4.52}$$

To bound now $\mathbb{E}|T_2|$, with T_2 as in (4.51), we need to take into account that $\frac{\partial^3}{\partial \theta_k \partial \theta_q \partial \theta_j} \ell(\boldsymbol{\theta}; \mathbf{x}) \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0^*}$ is in general not uniformly bounded. For $\epsilon > 0$, the law of total expectation and Markov's inequality yield

$$\begin{aligned} \mathbb{E}|T_2| &\leq 2\|h\| \mathbb{P}(|Q_{(m)}| \geq \epsilon) + \mathbb{E}[|T_2| | |Q_{(m)}| < \epsilon] \\ &\leq \frac{2\|h\|}{\epsilon^2} \mathbb{E} \left[\sum_{j=1}^d Q_j^2 \right] + \mathbb{E}[|T_2| | |Q_{(m)}| < \epsilon], \end{aligned} \tag{4.53}$$

with $Q_{(m)}$ as in (2.5). To bound $\mathbb{E}[|T_2| | |Q_{(m)}| < \epsilon]$, a first-order Taylor expansion and (2.13) yield

$$|T_2| \leq \frac{\|h\|_1}{2\sqrt{n}} \sum_{k=1}^d \sum_{l=1}^d |\tilde{V}_{lk}| \sum_{j=1}^d \sum_{v=1}^d \left| Q_j Q_v \frac{\partial^3}{\partial \theta_k \partial \theta_j \partial \theta_v} \ell(\boldsymbol{\theta}; \mathbf{X}) \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0^*} \right|. \tag{4.54}$$

Therefore, from (4.53) and (4.54) we have that

$$\mathbb{E}|T_2| \leq \frac{2\|h\|}{\epsilon^2} \mathbb{E} \left[\sum_{j=1}^d Q_j^2 \right]$$

$$+ \frac{\|h\|_1}{2\sqrt{n}} \sum_{k=1}^d \sum_{l=1}^d |\tilde{V}_{lk}| \mathbb{E} \left[\sum_{j=1}^d \sum_{v=1}^d \left| Q_j Q_v \frac{\partial^3}{\partial \theta_k \partial \theta_j \partial \theta_v} \ell(\boldsymbol{\theta}; \mathbf{X}) \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0^*} \right| \left| Q_{(m)} \right| < \epsilon \right].$$

The Cauchy-Schwarz inequality and Lemma 4.1 yield

$$\begin{aligned} \mathbb{E}|T_2| &\leq \frac{2\|h\|}{\epsilon^2} \mathbb{E} \left[\sum_{j=1}^d Q_j^2 \right] \\ &+ \frac{\|h\|_1}{2\sqrt{n}} \left\{ \sum_{k=1}^d \sum_{l=1}^d |\tilde{V}_{lk}| \sum_{j=1}^d \sum_{v=1}^d [\mathbb{E} [Q_j^2 Q_v^2]]^{\frac{1}{2}} \left[\mathbb{E} \left[(M_{kjv}(\mathbf{X}))^2 \mid |Q_{(m)}| < \epsilon \right] \right]^{\frac{1}{2}} \right\}. \end{aligned} \quad (4.55)$$

Therefore, from (4.52) and (4.55) we obtain that

$$(2.15) \leq \frac{2\|h\|}{\epsilon^2} \mathbb{E} \left[\sum_{j=1}^d Q_j^2 \right] + \frac{\|h\|_1}{\sqrt{n}} K_1(\boldsymbol{\theta}_0), \quad (4.56)$$

where $K_1(\boldsymbol{\theta}_0)$ is as in (2.7). Using now (4.50) and (4.56) we obtain the assertion. ■

Appendix: Proofs of Lemma 4.1 and of Corollaries 2.3 and 3.1

Proof of Lemma 4.1. Let $k \in \{1, 2, \dots, d\}$. We set $M_{d+1} = 0$. It will be shown that for $k = 1, 2, \dots, d$ we have that

$$\mathbb{E}[f(\mathbf{M}) \mid M_i < \epsilon, i = k, \dots, d] \leq \mathbb{E}[f(\mathbf{M}) \mid M_i < \epsilon, i = k + 1, \dots, d].$$

From the law of total expectation,

$$\begin{aligned} &\mathbb{E}[f(\mathbf{M}) \mid M_i < \epsilon, i = k + 1, \dots, d] \\ &= \mathbb{E}[f(\mathbf{M}) \mid M_i < \epsilon, i = k, \dots, d] \mathbb{P} [M_k < \epsilon \mid M_i < \epsilon, i = k + 1, \dots, d] \\ &+ \mathbb{E}[f(\mathbf{M}) \mid M_i < \epsilon, i = k + 1, \dots, d, M_k \geq \epsilon] \mathbb{P} [M_k \geq \epsilon \mid M_i < \epsilon, \\ & \quad i = k + 1, \dots, d]. \end{aligned}$$

Using that

$$\mathbb{P} [M_k < \epsilon \mid M_i < \epsilon, i = k + 1, \dots, d] = 1 - \mathbb{P} [M_k \geq \epsilon \mid M_i < \epsilon, i = k + 1, \dots, d]$$

yields

$$\begin{aligned} &\mathbb{E}[f(\mathbf{M}) \mid M_i < \epsilon, i = k + 1, \dots, d] = \mathbb{E}[f(\mathbf{M}) \mid M_i < \epsilon, i = k, \dots, d] \\ &+ \mathbb{P} [M_k \geq \epsilon \mid M_i < \epsilon, i = k + 1, \dots, d] \left\{ \mathbb{E}[f(\mathbf{M}) \mid M_i < \epsilon, i = k + 1, \dots, d, \right. \end{aligned}$$

$$M_k \geq \epsilon] - \mathbb{E}[f(\mathbf{M})|M_i < \epsilon, i = k, \dots, d] \Bigg\}. \tag{4.57}$$

Since $f(\mathbf{m})$ is an increasing function,

$$\mathbb{E}[f(\mathbf{M})|M_i < \epsilon, i = k + 1, \dots, d, M_k \geq \epsilon] - \mathbb{E}[f(\mathbf{M})|M_i < \epsilon, i = k, \dots, d] \geq 0.$$

Applying this to (4.57) gives that

$$\mathbb{E}[f(\mathbf{M})|M_i < \epsilon, i = k, \dots, d] \leq \mathbb{E}[f(\mathbf{M})|M_i < \epsilon, i = k + 1, \dots, d].$$

A simple iteration over k gives that

$$\mathbb{E}[f(\mathbf{M})|M_i < \epsilon \forall i = 1, 2, \dots, d] \leq \mathbb{E}[f(\mathbf{M})],$$

which is the result of the lemma. ■

Proof of Corollary 2.3. For one random variable, the first and second-order partial derivatives of the logarithm of the normal density function are

$$\begin{aligned} \frac{\partial}{\partial \eta_1} \log f(x_1|\boldsymbol{\eta}_0) &= -x_1^2 + \frac{1}{2\eta_1} + \frac{\eta_2^2}{4\eta_1^2}, & \frac{\partial}{\partial \eta_2} \log f(x_1|\boldsymbol{\eta}_0) &= x_1 - \frac{\eta_2}{2\eta_1}, \\ \frac{\partial^2}{\partial \eta_1^2} \log f(x_1|\boldsymbol{\eta}_0) &= -\left(\frac{1}{2\eta_1^2} + \frac{\eta_2^2}{2\eta_1^3}\right), & \frac{\partial^2}{\partial \eta_2^2} \log f(x_1|\boldsymbol{\eta}_0) &= -\frac{1}{2\eta_1}, \\ \frac{\partial^2}{\partial \eta_1 \partial \eta_2} \log f(x_1|\boldsymbol{\eta}_0) &= \frac{\partial^2}{\partial \eta_2 \partial \eta_1} \log f(x_1|\boldsymbol{\eta}_0) = \frac{\eta_2}{2\eta_1^2}. \end{aligned} \tag{4.58}$$

Hence, the expected Fisher Information matrix for one random variable is

$$I(\boldsymbol{\eta}_0) = \frac{1}{2\eta_1} \begin{pmatrix} \frac{1}{\eta_1} + \frac{\eta_2^2}{\eta_1^2} & -\frac{\eta_2}{\eta_1} \\ -\frac{\eta_2}{\eta_1} & 1 \end{pmatrix}, \tag{4.59}$$

and after simple calculations we obtain that

$$[I(\boldsymbol{\eta}_0)]^{-\frac{1}{2}} = \tilde{V} = \sqrt{\frac{2}{\alpha}} \begin{pmatrix} \eta_1^{\frac{3}{2}} (1 + \sqrt{\eta_1}) & \eta_1 \eta_2 \\ \eta_1 \eta_2 & \eta_1 (1 + \sqrt{\eta_1}) + \eta_2^2 \end{pmatrix},$$

where $\alpha = \eta_1 (1 + \sqrt{\eta_1})^2 + \eta_2^2$ as defined in Corollary 2.3. We bound the terms in Theorem 2.1 in order of appearance. The term $K_1(\boldsymbol{\eta}_0)$ is given in (2.7) and the first quantity of $K_1(\boldsymbol{\eta}_0)$ vanishes due to the fact that

$$\mathbb{E}[T_{kj}^2] = 0, \quad \forall k, j \in \{1, 2\}. \tag{4.60}$$

This comes from the definition of T_{kj} in (2.5) and the results of (4.58) and (4.59).

For the second term of $K_1(\boldsymbol{\eta}_0)$, we note that $\text{Cov}\left[\bar{X}, \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2\right] = 0$ (Casella & Berger, 2002)[p.218] and simple calculations lead to

$$\mathbb{E}[Q_1^2 Q_2^2] < \frac{2n\eta_1^3(2n + 63) + 3\eta_1^2\eta_2^2(4n^2 + 172n + 315)}{(n - 5)^2(n - 9)^2},$$

$$\begin{aligned}\mathbb{E}[Q_1^4] &< \frac{\eta_1^4(12n^2 + 516n + 945)}{(n-5)^2(n-9)^2} \\ \mathbb{E}[Q_2^4] &< \frac{12n^2(\eta_1 + \eta_2^2)^2 + 12n\eta_2^2(43\eta_2^2 + 63\eta_1) + 945\eta_2^4}{(n-5)^2(n-9)^2},\end{aligned}\quad (4.61)$$

where Q_1 and Q_2 are defined in (2.5). In addition, for $M_{kjl}(\mathbf{x})$ and $0 < \epsilon = \epsilon(\boldsymbol{\eta}_0)$ as in the condition (R.C.3), simple calculations and (4.58) yield for $m = 1, 2$

$$\begin{aligned}\sup_{\boldsymbol{\theta}:|\theta_m-\eta_m|<\epsilon} \left| \frac{\partial^3}{\partial \theta_1^3} \ell(\boldsymbol{\theta}; \mathbf{X}) \right| &= \sup_{\boldsymbol{\theta}:|\theta_m-\eta_m|<\epsilon} \left| \frac{n}{\theta_1^3} + \frac{3n\theta_2^2}{2\theta_1^4} \right| \\ &< \frac{n}{(\eta_1 - \epsilon)^3} \left(1 + \frac{3(\eta_2 + \epsilon)^2}{2(\eta_1 - \epsilon)} \right) =: M_{111}(\mathbf{x}), \\ \sup_{\boldsymbol{\theta}:|\theta_m-\eta_m|<\epsilon} \left| \frac{\partial^3}{\partial \theta_2^3} \ell(\boldsymbol{\theta}; \mathbf{X}) \right| &= 0 =: M_{222}(\mathbf{x}), \\ \sup_{\boldsymbol{\theta}:|\theta_m-\eta_m|<\epsilon} \left| \frac{\partial^3}{\partial \theta_1^2 \partial \theta_2} \ell(\boldsymbol{\theta}; \mathbf{X}) \right| &= \left| -\frac{n\eta_2}{\eta_1^3} \right| < \frac{n(\eta_2 + \epsilon)}{(\eta_1 - \epsilon)^3} =: M_{112}(\mathbf{x}), \\ \sup_{\boldsymbol{\theta}:|\theta_m-\eta_m|<\epsilon} \left| \frac{\partial^3}{\partial \theta_1 \partial \theta_2^2} \ell(\boldsymbol{\theta}; \mathbf{X}) \right| &= \left| \frac{n}{2\eta_1^2} \right| < \frac{n}{2(\eta_1 - \epsilon)^2} =: M_{221}(\mathbf{x}).\end{aligned}\quad (4.62)$$

For the choice of $\epsilon = \epsilon_0$ as in (R.C.3), (4.62) requires that $0 < \epsilon < \eta_1$. There is a trade-off on its choice for the fourth term of the bound in (2.6) and the results in (4.62). This is because the last term of the general bound in (2.6) is divided by ϵ^2 indicating that we should choose ϵ away from zero. However the terms in (4.62) have powers of $\eta_1 - \epsilon$ on the denominator and it would be reasonable for ϵ to be close to zero and away from η_1 . An optimisation process with respect to ϵ becomes quite tedious and therefore we choose ϵ to be the midpoint of $(0, \eta_1)$, which is sufficiently away from both zero and η_1 . Using this value of ϵ and for \tilde{V} as in (2.5), our results in (4.62) and (4.61) give, for the second term of $K_1(\boldsymbol{\eta}_0)$, that

$$\begin{aligned}&\frac{1}{2} \left\{ \sum_{k=1}^2 \sum_{l=1}^2 |\tilde{V}_{lk}| \sum_{j=1}^2 \sum_{i=1}^2 \sqrt{\mathbb{E}[Q_j^2 Q_i^2]} \sqrt{\mathbb{E}[(nM_{kji}(\mathbf{X}))^2 | Q_{(m)} < \epsilon]} \right\} \\ &< \frac{n^2}{\sqrt{2\alpha}(n-5)(n-9)} \left\{ 8(\sqrt{\eta_1} + |\eta_2| + \eta_1) \sqrt{12 + \frac{516}{n} + \frac{945}{n^2}} \right. \\ &\quad \times \left(1 + \frac{3(|\eta_2| + \frac{\eta_1}{2})^2}{\eta_1} \right) \\ &\quad + 2\sqrt{12(\eta_1 + \eta_2^2)^2 + \frac{12\eta_2^2(43\eta_2^2 + 63\eta_1)}{n} + \frac{945\eta_2^4}{n^2}} \\ &\quad \times \left(\frac{4|\eta_2^3|}{\eta_1^3} + \frac{(2|\eta_2| + \eta_1)(3|\eta_2| + 2 + 2\sqrt{\eta_1})}{\eta_1^2} + \frac{1}{\sqrt{\eta_1}} + 1 \right) \end{aligned}$$

$$\begin{aligned}
& + \frac{4}{\eta_1} \sqrt{4(\eta_1 + 3\eta_2^2) + \frac{6}{n}(21\eta_1 + 86\eta_2^2) + 945\eta_2^2} \\
& \left. \times ((\eta_1 + |\eta_2|)(\eta_1 + 3|\eta_2| + 2\sqrt{\eta_1}) + \eta_1) \right\}, \quad (4.63)
\end{aligned}$$

which is an upper bound for $K_1(\boldsymbol{\eta}_0)$. We now proceed to find an upper bound on $K_2(\boldsymbol{\eta}_0)$, which is a sum of two quantities as (2.8) shows, involving the calculation of variances of ξ_{ij} as defined in (2.5). For the first quantity, using (4.58) and (4.60), after straightforward calculation of moments (up to fourth order) of X_1 and with α and β as in the corollary, we get that

$$\begin{aligned}
& \frac{1}{4} \sum_{j=1}^2 \left[\text{Var} \left[\left(\sum_{k=1}^2 \tilde{V}_{jk} \frac{\partial}{\partial \eta_k} \log f(X_1 | \boldsymbol{\eta}_0) \right)^2 \right] \right]^{\frac{1}{2}} \\
& = \frac{1}{4} \left\{ \sqrt{\text{Var} \left[\left(X_1^2 - \frac{1}{2\eta_1} - \frac{\eta_2^2}{4\eta_1^2} \right)^2 \right]} (\tilde{V}_{11}^2 + \tilde{V}_{12}^2) \right. \\
& \quad \left. + \sqrt{\text{Var} \left[\left(X_1 - \frac{\eta_2}{2\eta_1} \right)^2 \right]} (\tilde{V}_{22}^2 + \tilde{V}_{12}^2) \right\} \\
& = \frac{1}{2\alpha} \left\{ \alpha \sqrt{\frac{7}{2} + \frac{\eta_2^4}{2\eta_1^2} + \frac{7\eta_2^2}{\eta_1} + \frac{1}{\sqrt{2}\eta_1} (\eta_1^2 \eta_2^2 + \beta^2)} \right\}. \quad (4.64)
\end{aligned}$$

For the second quantity in $K_2(\boldsymbol{\eta}_0)$, simple calculation of moments leads to

$$\begin{aligned}
& \frac{1}{2} \left[\text{Var} \left[\sum_{q=1}^2 \sum_{v=1}^2 \tilde{V}_{2q} \frac{\partial}{\partial \eta_q} \log f(X_1 | \boldsymbol{\eta}_0) \tilde{V}_{1v} \frac{\partial}{\partial \eta_v} \log f(X_1 | \boldsymbol{\eta}_0) \right] \right]^{\frac{1}{2}} \\
& \leq \frac{1}{2} \left[\mathbb{E} \left[\tilde{V}_{11} \tilde{V}_{21} \left(\frac{\partial}{\partial \eta_1} \log f(X_1 | \boldsymbol{\eta}_0) \right)^2 + \tilde{V}_{22} \tilde{V}_{12} \left(\frac{\partial}{\partial \eta_2} \log f(X_1 | \boldsymbol{\eta}_0) \right)^2 \right. \right. \\
& \quad \left. \left. + \frac{\partial}{\partial \eta_1} \log f(X_1 | \boldsymbol{\eta}_0) \frac{\partial}{\partial \eta_2} \log f(X_1 | \boldsymbol{\eta}_0) (\tilde{V}_{21}^2 + \tilde{V}_{22} \tilde{V}_{11}) \right]^2 \right]^{\frac{1}{2}} \\
& \leq \frac{\sqrt{3}}{2} \left[\tilde{V}_{11}^2 \tilde{V}_{21}^2 \mathbb{E} \left[\left(\frac{\partial}{\partial \eta_1} \log f(X_1 | \boldsymbol{\eta}_0) \right)^4 \right] + \tilde{V}_{22}^2 \tilde{V}_{12}^2 \mathbb{E} \left[\left(\frac{\partial}{\partial \eta_2} \log f(X_1 | \boldsymbol{\eta}_0) \right)^4 \right] \right. \\
& \quad \left. + (\tilde{V}_{21}^2 + \tilde{V}_{22} \tilde{V}_{11})^2 \mathbb{E} \left[\left(\frac{\partial}{\partial \eta_1} \log f(X_1 | \boldsymbol{\eta}_0) \frac{\partial}{\partial \eta_2} \log f(X_1 | \boldsymbol{\eta}_0) \right)^2 \right] \right]^{\frac{1}{2}} \\
& = \frac{3\eta_2}{2\alpha} \left[(\alpha - \eta_2^2) \left(5 + \frac{\eta_2^2}{\eta_1} \right)^2 + \beta^2 + \left(2\sqrt{\eta_1} \eta_2 + \frac{\alpha}{\eta_2} \right)^2 \left(\frac{5}{3} + \frac{\eta_2^2}{\eta_1} \right) \right]^{\frac{1}{2}}. \quad (4.65)
\end{aligned}$$

For an upper bound on $K_3(\boldsymbol{\eta}_0)$ as in (2.9), we use that X'_1 is an independent copy of X_1 and also

$$\begin{aligned}\mathbb{E}\left[\left|\frac{\partial}{\partial\eta_1}\log f(X_1|\boldsymbol{\eta}_0)\right|^3\right] &= \mathbb{E}\left[\left|-X_1^2 + \frac{1}{2\eta_1} + \frac{\eta_2^2}{4\eta_1^2}\right|^3\right] \leq \frac{18}{\eta_1^3}\left(1 + \frac{\eta_2^3}{2\eta_1^{\frac{3}{2}}\sqrt{\pi}}\right) \\ \mathbb{E}\left[\left|\frac{\partial}{\partial\eta_2}\log f(X_1|\boldsymbol{\eta}_0)\right|^3\right] &= \frac{1}{\sqrt{\pi}\eta_1^{\frac{3}{2}}}.\end{aligned}$$

Then, the triangle inequality and (2.24) yield

$$\begin{aligned}&\frac{1}{12}\mathbb{E}\left[\sum_{i=1}^2\left|\sum_{l=1}^2\tilde{V}_{il}\left(\frac{\partial}{\partial\eta_l}\log f(X'_1|\boldsymbol{\eta}_0) - \frac{\partial}{\partial\eta_l}\log f(X_1|\boldsymbol{\eta}_0)\right)\right|^3\right] \\ &\leq \frac{32}{3}\left\{\mathbb{E}\left[\left|\frac{\partial}{\partial\eta_1}\log f(X_1|\boldsymbol{\eta}_0)\right|^3\right]\left(|\tilde{V}_{11}|^3 + |\tilde{V}_{21}|^3\right)\right. \\ &\quad \left.+\mathbb{E}\left[\left|\frac{\partial}{\partial\eta_2}\log f(X_1|\boldsymbol{\eta}_0)\right|^3\right]\left(|\tilde{V}_{12}|^3 + |\tilde{V}_{22}|^3\right)\right\} \\ &= \frac{64\sqrt{2}}{3\alpha^{\frac{3}{2}}}\left\{18\left(1 + \frac{\eta_2^3}{2\eta_1^{\frac{3}{2}}\sqrt{\pi}}\right)\left(\eta_1^{\frac{3}{2}}(1 + \sqrt{\eta_1})^3 + |\eta_2|^3\right) + \frac{\eta_1^3|\eta_2|^3 + \beta^3}{\sqrt{\pi}\eta_1^{\frac{3}{2}}}\right\}.\end{aligned}\tag{4.66}$$

For the last term of (2.6), we obtain that

$$\begin{aligned}\frac{2\|h\|}{\epsilon^2}\mathbb{E}\left[\sum_{j=1}^2Q_j^2\right] &= \frac{2\|h\|}{\epsilon^2(n-3)(n-5)}\left((2n+15)\eta_1^2 + 2n(\eta_2^2 + \eta_1) + 15\eta_2^2\right) \\ &= \frac{8\|h\|}{\eta_1^2(n-3)(n-5)}\left((2n+15)(\eta_1^2 + \eta_2^2) + 2n\eta_1\right),\end{aligned}\tag{4.67}$$

where for the second equality we used that $\epsilon = \frac{2n}{2}$, with our choice explained in the paragraph after (4.62). Using the results in (4.60), (4.63), (4.64), (4.65), (4.66) and (4.67) we get the result of the corollary. ■

Proof of Corollary 3.1.

Part **a**). The probability density function is

$$f(x|\boldsymbol{\theta}) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1},$$

with $\alpha, \beta > 0$ and $x \in [0, 1]$. Hence, for $j, k \in \mathbb{Z}^+$

$$\frac{\partial}{\partial\alpha}\log f(x|\boldsymbol{\theta}) = \Psi(\alpha + \beta) - \Psi(\alpha) + \log(x),$$

$$\begin{aligned} \frac{\partial}{\partial \beta} \log f(x|\boldsymbol{\theta}) &= \Psi(\alpha + \beta) - \Psi(\beta) + \log(1 - x) \\ \frac{\partial^{j+1}}{\partial \alpha^{j+1}} \log f(x|\boldsymbol{\theta}) &= \Psi_j(\alpha + \beta) - \Psi_j(\alpha), \\ \frac{\partial^{j+1}}{\partial \beta^{j+1}} \log f(x|\boldsymbol{\theta}) &= \Psi_j(\alpha + \beta) - \Psi_j(\beta) \\ \frac{\partial^{k+j}}{\partial \alpha^k \partial \beta^j} \log f(x|\boldsymbol{\theta}) &= \Psi_{k+j-1}(\alpha + \beta). \end{aligned} \tag{4.68}$$

From (4.68), we see that we are under the scenario (2) of Remark 3.1 and U_1 will be calculated using (3.40). The expected Fisher Information matrix is

$$I(\boldsymbol{\theta}_0) = \begin{pmatrix} \Psi_1(\alpha) - \Psi_1(\alpha + \beta) & -\Psi_1(\alpha + \beta) \\ -\Psi_1(\alpha + \beta) & \Psi_1(\beta) - \Psi_1(\alpha + \beta) \end{pmatrix}.$$

Simple calculations show that the inverse of $I(\boldsymbol{\theta}_0)$ is

$$[I(\boldsymbol{\theta}_0)]^{-1} = \frac{1}{\delta_I} \begin{pmatrix} C_1(\beta, \alpha) & \Psi_1(\alpha + \beta) \\ \Psi_1(\alpha + \beta) & C_1(\alpha, \beta) \end{pmatrix}.$$

Therefore,

$$[I(\boldsymbol{\theta}_0)]^{-2} = \frac{1}{\delta_I^2} \begin{pmatrix} C_1^2(\beta, \alpha) + \Psi_1^2(\alpha + \beta) & \Psi_1(\alpha + \beta)(C_1(\alpha, \beta) + C_1(\beta, \alpha)) \\ \Psi_1(\alpha + \beta)(C_1(\alpha, \beta) + C_1(\beta, \alpha)) & C_1^2(\alpha, \beta) + \Psi_1^2(\alpha + \beta) \end{pmatrix}.$$

For $k, q = 1, 2$, we now proceed to calculate the quantities

$$\sup_{\substack{\boldsymbol{\theta}: |\theta_j - \theta_{0,j}| < \epsilon \\ \forall j \in \{1,2\}}} \left\{ [I^{-2}(\boldsymbol{\theta})]_{kq} \right\}.$$

Firstly, the fact that $\delta_I(\alpha, \beta)$ as in (3.42) is a positive, decreasing function of α and β , means that

$$\sup_{\substack{\boldsymbol{\theta}: |\theta_j - \theta_{0,j}| < \epsilon \\ \forall j \in \{1,2\}}} \left\{ \frac{1}{[\delta_I(\theta_1, \theta_2)]^2} \right\} = \frac{1}{[\delta_I(\alpha + \epsilon, \beta + \epsilon)]^2}. \tag{4.69}$$

In regards to $C_1^2(\theta_1, \theta_2)$ as in (3.42), we have that using a first-order Taylor expansion and for $\tilde{\theta}$ between θ_1 and $\theta_1 + \theta_2$,

$$\begin{aligned} \sup_{\substack{\boldsymbol{\theta}: |\theta_j - \theta_{0,j}| < \epsilon \\ \forall j \in \{1,2\}}} \{C_1^2(\theta_1, \theta_2)\} &= \sup_{\substack{\boldsymbol{\theta}: |\theta_j - \theta_{0,j}| < \epsilon \\ \forall j \in \{1,2\}}} \left\{ \theta_2^2 \Psi_2^2(\tilde{\theta}) \right\} \\ &= (\beta + \epsilon)^2 \sup_{\substack{\boldsymbol{\theta}: |\theta_j - \theta_{0,j}| < \epsilon \\ \forall j \in \{1,2\}}} \left\{ \Psi_2^2(\theta_1) \right\} = (\beta + \epsilon)^2 \Psi_2^2(\alpha - \epsilon), \end{aligned} \tag{4.70}$$

since $\Psi_2^2(x)$ is a decreasing function of x ; see the definition of $\Psi_2(\cdot)$ in (3.41). In the same way, we can find an upper bound for $C_1^2(\theta_2, \theta_1)$. With regards to the quantity $C_1(\theta_1, \theta_2) + C_1(\theta_2, \theta_1)$, we have that a similar first-order Taylor expansion as in (4.70) leads to

$$C_1(\theta_1, \theta_2) + C_1(\theta_2, \theta_1) = -\theta_2 \Psi_2(\tilde{\theta}) - \theta_1 \Psi_2(\tilde{\theta}), \quad (4.71)$$

where $\tilde{\theta}$ is between θ_1 and $\theta_1 + \theta_2$, while $\tilde{\theta}$ is between θ_2 and $\theta_1 + \theta_2$. It is important to highlight that $\Psi_2(x)$, as defined in (3.41) is a negative and increasing function of x . Continuing from (4.71),

$$\sup_{\substack{\boldsymbol{\theta}: |\theta_j - \theta_{0,j}| < \epsilon \\ \forall j \in \{1,2\}}} \{C_1(\theta_1, \theta_2) + C_1(\theta_2, \theta_1)\} = -[(\beta + \epsilon) \Psi_2(\alpha - \epsilon) + (\alpha + \epsilon) \Psi_2(\beta - \epsilon)]. \quad (4.72)$$

Using the results in (4.69), (4.70), as well as the fact that $\Psi_1(x)$ defined in (3.41) is a positive, decreasing function of x , we have that

$$\begin{aligned} \sup_{\substack{\boldsymbol{\theta}: |\theta_j - \theta_{0,j}| < \epsilon \\ \forall j \in \{1,2\}}} \{[I^{-2}(\boldsymbol{\theta})]_{11}\} &= \frac{(\alpha + \epsilon)^2 \Psi_2^2(\beta - \epsilon) + \Psi_1^2(\alpha + \beta - 2\epsilon)}{[\delta_I(\alpha + \epsilon, \beta + \epsilon)]^2} \\ \sup_{\substack{\boldsymbol{\theta}: |\theta_j - \theta_{0,j}| < \epsilon \\ \forall j \in \{1,2\}}} \{[I^{-2}(\boldsymbol{\theta})]_{12}\} &= \sup_{\substack{\boldsymbol{\theta}: |\theta_j - \theta_{0,j}| < \epsilon \\ \forall j \in \{1,2\}}} \{[I^{-2}(\boldsymbol{\theta})]_{21}\} \\ &= -\frac{[(\beta + \epsilon) \Psi_2(\alpha - \epsilon) + (\alpha + \epsilon) \Psi_2(\beta - \epsilon)]}{[\delta_I(\alpha + \epsilon, \beta + \epsilon)]^2} \\ \sup_{\substack{\boldsymbol{\theta}: |\theta_j - \theta_{0,j}| < \epsilon \\ \forall j \in \{1,2\}}} \{[I^{-2}(\boldsymbol{\theta})]_{22}\} &= \frac{(\beta + \epsilon)^2 \Psi_2^2(\alpha - \epsilon) + \Psi_1^2(\alpha + \beta - 2\epsilon)}{[\delta_I(\alpha + \epsilon, \beta + \epsilon)]^2}. \end{aligned} \quad (4.73)$$

To derive the expression for U_1 as in (3.40) in this special case, we need to calculate the quantities $\mathbb{E} \left[\frac{\partial}{\partial \theta_k} \ell(\boldsymbol{\theta}_0; \mathbf{X}) \frac{\partial}{\partial \theta_q} \ell(\boldsymbol{\theta}_0; \mathbf{X}) \right]$, for $k, q = 1, 2$. Using (4.68), we have that

$$\begin{aligned} \mathbb{E} \left[\left(\frac{\partial}{\partial \alpha} \ell(\boldsymbol{\theta}_0; \mathbf{X}) \right)^2 \right] &= \mathbb{E} \left[\left(n(\Psi(\alpha + \beta) - \Psi(\alpha)) + \sum_{i=1}^n \log(X_i) \right)^2 \right] \\ &= \text{Var} \left[\sum_{i=1}^n \log(X_i) \right] = nC_1(\alpha, \beta) \\ \mathbb{E} \left[\left(\frac{\partial}{\partial \beta} \ell(\boldsymbol{\theta}_0; \mathbf{X}) \right)^2 \right] &= \mathbb{E} \left[\left(n(\Psi(\alpha + \beta) - \Psi(\beta)) + \sum_{i=1}^n \log(1 - X_i) \right)^2 \right] \\ &= \text{Var} \left[\sum_{i=1}^n \log(1 - X_i) \right] = nC_1(\beta, \alpha) \end{aligned}$$

$$\begin{aligned}
\mathbb{E} \left[\frac{\partial}{\partial \alpha} \ell(\boldsymbol{\theta}_0; \mathbf{X}) \frac{\partial}{\partial \beta} \ell(\boldsymbol{\theta}_0; \mathbf{X}) \right] &= n(\Psi(\alpha + \beta) - \Psi(\beta)) \mathbb{E} \left[\sum_{i=1}^n \log(X_i) \right] \\
&\quad + \mathbb{E} \left[\sum_{i=1}^n \sum_{j=1}^n \log(X_i) \log(1 - X_j) \right] \\
&= n^2(\Psi(\alpha + \beta) - \Psi(\beta))(\Psi(\alpha) - \Psi(\alpha + \beta)) \\
&\quad + n((\Psi(\alpha) - \Psi(\alpha + \beta))(\Psi(\beta) - \Psi(\alpha + \beta)) \\
&\quad \quad - \Psi_1(\alpha + \beta)) \\
&\quad + n(n-1)(\Psi(\alpha) - \Psi(\alpha + \beta))(\Psi(\beta) - \Psi(\alpha + \beta)) \\
&= -n\Psi_1(\alpha + \beta). \tag{4.74}
\end{aligned}$$

Applying the results of (4.73) and (4.74) to (3.40), we conclude that

$$\begin{aligned}
\mathbb{E} \left[\sum_{j=1}^2 Q_j^2 \right] &\leq \frac{1}{n[\delta_I(\alpha + \epsilon, \beta + \epsilon)]^2} \left\{ C_1(\alpha, \beta) [(\alpha + \epsilon)^2 \Psi_2^2(\beta - \epsilon) + \Psi_1^2(\alpha + \beta - 2\epsilon)] \right. \\
&\quad + C_1(\beta, \alpha) [(\beta + \epsilon)^2 \Psi_2^2(\alpha - \epsilon) + \Psi_1^2(\alpha + \beta - 2\epsilon)] \\
&\quad \left. + 2\Psi_1(\alpha + \beta) [(\beta + \epsilon) \Psi_2(\alpha - \epsilon) + (\alpha + \epsilon) \Psi_2(\beta - \epsilon)] \right\},
\end{aligned}$$

which completes the proof. ■

References

- Anastasiou, A. (2017). Bounds for the normal approximation of the maximum likelihood estimator from m -dependent random variables. *Statistics & Probability Letters*, **129**, 171–181. [MR3688530](#)
- Anastasiou, A. & Ley, C. (2017). Bounds for the asymptotic normality of the maximum likelihood estimator using the Delta method. *ALEA, Lat. Am. J. Probab. Math. Stat.*, **14**, 153–171. [MR3622464](#)
- Anastasiou, A. & Reinert, G. (2017). Bounds for the normal approximation of the maximum likelihood estimator. *Bernoulli*, **23**, 191–218. [MR3556771](#)
- Berk, R. H. (1972). Consistency and asymptotic normality of MLE's for exponential models. *The Annals of Mathematical Statistics*, **43**, 193–204. [MR0298810](#)
- Billingsley, P. (1961). Statistical Methods in Markov Chains. *The Annals of Mathematical Statistics*, **32**, No.1, 12–40. [MR0123420](#)
- Casella, G. & Berger, R. L. (2002). *Statistical Inference*. Brooks/Cole, Cengage Learning, Duxbury, Pacific Grove, second edition. [MR1051420](#)
- Davison, A. C. (2008). *Statistical Models*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, first edition. [MR1998913](#)

- Fahrmeir, L. & Kaufmann, H. (1985). Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *The Annals of Statistics*, **13**, No.1, 342–368. [MR0773172](#)
- Gorham, J., Duncan, A. B., Vollmer, S. J., & Mackey, L. (2016). Measuring sample quality with diffusions. <https://arxiv.org/pdf/1611.06972.pdf>.
- Hoadley, B. (1971). Asymptotic Properties of Maximum Likelihood Estimators for the Independent Not Identically Distributed Case. *The Annals of Mathematical Statistics*, **42**, No.6, 1977–1991. [MR0297051](#)
- Kiefer, J. C. (1968). Statistical inference. In *The future of statistics. Proceedings of a Conference on the Future of Statistics held at the University of Wisconsin, Madison, Wisconsin, June 1967*, pages 139–142. Academic Press, New York-London. [MR0234539](#)
- Koroljuk, V. S. & Borovskich, Y. V. (1994). *Theory of U-statistics*. Mathematics and its Applications **273**. Kluwer Academic Publishers Group, Dordrecht. Translated from the 1989 Russian original by P.V. Malyshev and D.V. Malyshev and revised by the authors. [MR1472486](#)
- Lauritzen, S. (1988). *Extremal Families and Systems of Sufficient Statistics*. Lecture Notes in Statistics, No.49. Springer-Verlag, Berlin-Heidelberg-New York. [MR0971253](#)
- Lauritzen, S. (1996). *Graphical Models*. Oxford: Clarendon Press. [MR1419991](#)
- Mäkeläinen, T., Schmidt, K., & Styan, G. P. H. (1981). On the existence and uniqueness of the maximum likelihood estimate of a vector-valued parameter in fixed-size samples. *The Annals of Statistics*, **9**, No.4, 758–767. [MR0619279](#)
- Massam, H., Li, Q., & Gao, X. (2018). Bayesian precision and covariance matrix estimation for graphical Gaussian models with edge and vertex symmetries. *Biometrika*, **105**, 371–388. [MR3804408](#)
- Pinelis, I. (2017). Optimal-order uniform and nonuniform bounds on the rate of convergence to normality for maximum likelihood estimators. *Electronic Journal of Statistics*, **11**, 1160–1179. [MR3634332](#)
- Pinelis, I. & Molzon, R. (2016). Optimal-order bounds on the rate of convergence to normality in the multivariate delta method. *Electronic Journal of Statistics*, **10**, 1001–1063. [MR3486424](#)
- Reinert, G. & Röllin, A. (2009). Multivariate normal approximation with Stein’s method of exchangeable pairs under a general linearity condition. *The Annals of Probability*, **37**, No.6, 2150–2173. [MR2573554](#)
- Stein, C. (1972). A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, volume **2**, pages 586–602. Berkeley: University of California Press. [MR0402873](#)