Electronic Journal of Statistics
Vol. 11 (2017) 1160–1179
ISSN: 1935-7524
DOI: 10.1214/17-EJS1264

Optimal-order uniform and nonuniform bounds on the rate of convergence to normality for maximum likelihood estimators

Iosif Pinelis

Department of Mathematical Sciences
Michigan Technological University
Houghton, Michigan 49931
e-mail: ipinelis@mtu.edu

Abstract: It is well known that, under general regularity conditions, the distribution of the maximum likelihood estimator (MLE) is asymptotically normal. Very recently, bounds of the optimal order $O(1/\sqrt{n})$ on the closeness of the distribution of the MLE to normality in the so-called bounded Wasserstein distance were obtained [2, 1], where n is the sample size. However, the corresponding bounds on the Kolmogorov distance were only of the order $O(1/n^{1/4})$. In this paper, bounds of the optimal order $O(1/\sqrt{n})$ on the closeness of the distribution of the MLE to normality in the Kolmogorov distance are given, as well as their nonuniform counterparts, which work better in tail zones of the distribution of the MLE. These results are based in part on previously obtained general optimal-order bounds on the rate of convergence to normality in the multivariate delta method. The crucial observation is that, under natural conditions, the MLE can be tightly enough bracketed between two smooth enough functions of the sum of independent random vectors, which makes the delta method applicable. It appears that the nonuniform bounds for MLEs in general have no precedents in the existing literature; a special case was recently treated by Pinelis and Molzon [20]. The results can be extended to M-estimators.

MSC 2010 subject classifications: 62F10, 62F12, 60F05, 60E15.
Keywords and phrases: Maximum likelihood estimators, Berry–Esseen bounds, delta method, rates of convergence.

Received July 2016.

Contents

1. Introduction

Let us begin with the following quote from Kiefer [9] of 1968:

> a second area of what seem to me important problems to work on has to do with the fact that we do have, in many settings, quite a good large sample theory, but we don't know how large the sample sizes have to be for that theory to take hold. Now, I'm sure most of you are familiar with the error estimate one can give for the classical central-limit theorem, which goes by the name of the Berry-Esseen estimate, and which tells you that under certain assumptions one can actually give an explicit bound on the departure from the normal distribution of the sample mean for a given sample size, the error term being of order $1/\sqrt{n}$. For most other statistical problems, in fact for almost anything other than the use of the sample mean, we have nothing. The most obvious example of this (and this is not original with me; many people have been concerned with this), is the maximum likelihood estimator in the case of regular estimation. We all know what the asymptotic distribution is. Can you give explicitly some useful bound on the departure from the asymptotic normal distribution as a function of the sample size n? It seems to be a terrifically difficult problem.

Since then, there has been some significant progress in this direction, especially rather recently. For instance, Berry–Esseen-type bounds of order $1/\sqrt{n}$ were obtained for U-statistics – see e.g. [10]; for the Student statistic [4, 3]; and, even more recently, for rather broad classes of other statistics that depend on the observations in a nonlinear fashion [6, 20].

As Kiefer pointed out, it is well known that, under general regularity conditions, the distribution of the maximum likelihood estimator (MLE) is asymptotically normal. In this paper, we shall consider Berry–Esseen-type bounds of order $1/\sqrt{n}$ for the MLE. First such bounds were apparently obtained in the paper [14], followed by [16, 17]. Very recently, bounds on the closeness of the distribution of the MLE to normality in the so-called bounded Wasserstein distance, d_{bW}, were obtained in [2]. In the rather common special case when the MLE $\hat{\theta}$ is expressible as a smooth enough function of a linear statistic of independent identically distributed (i.i.d.) observations, the bounds obtained in [2] were sharpened and simplified in [1] by using a version of the delta method. More specifically, it was assumed in [1] that

$$q(\hat{\theta}) = \frac{1}{n}\sum_{i=1}^{n} g(X_i), \tag{1.1}$$

where $q\colon \Theta \to \mathbb{R}$ is a twice continuously differentiable one-to-one mapping, $g\colon \mathbb{R} \to \mathbb{R}$ is a Borel-measurable function, and the X_i's are i.i.d. real-valued r.v.'s.

It was noted in [2, Proposition 2.1] that for any r.v. Y and a standard normal r.v. Z one has $d_{\mathsf{Ko}}(Y,Z) \leqslant 2\sqrt{d_{\mathsf{bW}}(Y,Z)}$, where d_{Ko} denotes the Kolmogorov distance. This bound on d_{Ko} in terms of d_{bW} is the best possible one, up to a constant factor, as shown in [20]. Therefore, even though the bounds on the bounded Wasserstein distance d_{bW} obtained in [2, 1] are of the optimal order $O(1/\sqrt{n})$, the resulting bounds on the Kolmogorov distance are only of the order $O(1/n^{1/4})$. (That the order $O(1/\sqrt{n})$ is optimal for MLEs is well known; for instance, see the example of the Bernoulli family of distributions given in [14].)

In [20], optimal-order bounds of the form $O(1/\sqrt{n})$ on the rate of convergence to normality in the general multivariate delta method were given. Those results are applicable when the statistic of interest can be expressed as a smooth enough function of the sum of independent random vectors. Accordingly, various kinds of applications were presented in [20]. In particular, uniform and nonuniform bounds of the optimal order on the closeness of the distribution of the MLE to normality were obtained in [20] under conditions similar to the mentioned conditions assumed in [1].

In this paper we present a way to extend those results in [20] to the general case, without an assumption of the form (1.1), made in [1, 20]. Of course, in general the MLE cannot be represented as a function of the sum of independent random vectors (see the Appendix in the arXiv version [19] of this paper for details). However, the crucial observation here is that, under natural conditions, the MLE can be tightly enough bracketed between two such smooth enough functions, which makes the delta method applicable. Thus, the present paper is methodologically different from the preceding work on Berry–Esseen-type bounds for the MLE, in that it relies on the general result developed in [20], rather than on methods specially designed to deal with the MLE.

Perhaps more importantly, the new method yields not only uniform bounds (that is, in the Kolmogorov metric) of the optimal order $O(1/\sqrt{n})$ on the closeness of the distribution of the MLE to normality but also their so-called nonuniform counterparts, which work much better for large deviations, that is, in tail zones of the distribution of the MLE – which are usually of foremost interest in statistical tests. Such nonuniform bounds for MLEs in general appear to have no precedents in the existing literature (except that, as stated above, a special case of nonuniform bounds for MLEs was recently treated in [20]).

The paper is organized as follows. The general setting of the problem is described in Section 2. The key step of tight enough bracketing of the MLE between two functions of the sum of independent random vectors is made in Section 3. General uniform and nonuniform optimal-order bounds from [20] on the convergence rate in the multivariate delta method are presented in Section 4. In Section 5, we make the bracketing work by applying the general bounds in the multivariate delta method. Yet, this leaves out the problem of bounding a remainder, which is a probability of large deviations of the MLE from the true value of the parameter. It is shown in Section 6 that under natural conditions this remainder is exponentially fast decreasing (in n) and thus asymptotically negligible as compared to the main term on the order of $1/\sqrt{n}$. All these findings are summarized in Section 7, where the main result of this paper is presented, along with corresponding discussion.

2. General setting

Let $X, X_1, X_2, \dots$ be random variables (r.v.'s) mapping a measurable space $(\Omega, \mathcal{A})$ to another measurable space $(\mathcal{X}, \mathcal{B})$ and let $(\mathsf{P}_\theta)_{\theta\in\Theta}$ be a parametric family of probability measures on $(\Omega, \mathcal{A})$ such that the r.v.'s $X, X_1, X_2, \dots$ are i.i.d. with respect to each of the probability measures P_θ with $\theta \in \Theta$; here the parameter space Θ is assumed to be a subset of the real line $\mathbb{R}$. As usual, let E_θ denote the expectation with respect to the probability measure P_θ. Suppose that for each $\theta \in \Theta$ the distribution $\mathsf{P}_\theta\, X^{-1}$ of X has a density p_θ with respect to a measure μ on $\mathcal{B}$. Because the extended real line $[-\infty, \infty]$ is compact, for each $n \in \mathbb{N}$ and each point $\mathbf{x} = \mathbf{x}_n = (x_1, \dots, x_n) \in \mathcal{X}^n$ the likelihood function $\Theta \ni \theta \mapsto L_{\mathbf{x}}(\theta) := \prod_{i=1}^n p_\theta(x_i)$ has at least one generalized maximizer $\hat\theta_n(\mathbf{x})$ in the closure of the set Θ in $[-\infty, \infty]$, in the sense that $\sup_{\theta\in\Theta} L_{\mathbf{x}}(\theta) = \limsup_{\theta\to\hat\theta_n(\mathbf{x})} L_{\mathbf{x}}(\theta)$. Picking, for each $\mathbf{x} = (x_1, \dots, x_n) \in \mathcal{X}^n$, any one of such generalized maximizers $\hat\theta_n(\mathbf{x})$, one obtains a map $\Omega \ni \omega \mapsto \hat\theta_n(\mathbf{X}(\omega))$, where $\mathbf{X} := \mathbf{X}_n := (X_1, \dots, X_n)$; any such map will be denoted here by $\hat\theta_n(\mathbf{X})$ (or simply by $\hat\theta_n$ or $\hat\theta$) and referred to as a maximum likelihood estimator (MLE) of θ. This is a somewhat more general definition of the MLE than usual, and in general an MLE $\hat\theta$ will not have to be a r.v.; that is, it can be non-measurable with respect to the sigma-algebra $\mathcal{A}$. However, to simplify the presentation, we shall still refer to sets of the form $\{\hat\theta \in J\} := \{\omega \in \Omega\colon \hat\theta_n(\mathbf{X}(\omega)) \in J\}$ for Borel sets $J \subseteq \Theta$ as events and write $\mathsf{P}_\theta(\hat\theta \in J)$ implying that the latter expression may and should be understood as either one of the expressions $(\mathsf{P}_\theta)^*(\hat\theta \in J)$ or $(\mathsf{P}_\theta)_*(\hat\theta \in J)$, where * and $_*$ stand for the corresponding outer and inner measures. Of course, when the map $\hat\theta$ is measurable, then one can use the bona fide expressions of the mentioned form $\mathsf{P}_\theta(\hat\theta \in J)$.

Let $\theta_0 \in \Theta$ be the "true" value of the unknown parameter θ, such that

$$[\theta_0 - \delta, \theta_0 + \delta] \subseteq \Theta^\circ \tag{2.1}$$

for some real $\delta > 0$, where Θ° denotes the interior of the subset Θ of $\mathbb{R}$. For brevity, let

$$\mathsf{P} := \mathsf{P}_{\theta_0} \quad \text{and} \quad \mathsf{E} := \mathsf{E}_{\theta_0}\,.$$

For $x \in \mathcal{X}$ and $\theta \in \Theta$, consider the log-likelihood

$$\ell_x(\theta) := \ln p_\theta(x)$$

and assume the following:

(I) The set $\mathcal{X}_{>0} := \{x \in \mathcal{X}\colon p_\theta(x) > 0\}$ is the same for all $\theta \in [\theta_0 - \delta, \theta_0 + \delta]$, and for each $x \in \mathcal{X}_{>0}$ the density $p_\theta(x)$ and hence the log-likelihood $\ell_x(\theta)$ are thrice differentiable in θ at each point $\theta \in [\theta_0 - \delta, \theta_0 + \delta]$.

(II) Standard regularity conditions hold so that $\mathsf{E}\,\ell'_X(\theta_0) = 0$ and $\mathsf{E}\,\ell'_X(\theta_0)^2 = -\mathsf{E}\,\ell''_X(\theta_0) = I(\theta_0) \in (0, \infty)$, where $I(\theta)$ is the Fisher information at θ.

(III) $\mathsf{E}\,|\ell'_X(\theta_0)|^3 + \mathsf{E}\,|\ell''_X(\theta_0)|^3 < \infty$.

(IV) $\mathsf{E} \sup\limits_{\theta\in[\theta_0-\delta,\theta_0+\delta]} |\ell'''_X(\theta)|^3 < \infty$.

Remark 2.1. The introduction of the set $\mathcal{X}_{>0}$ in condition (I) is needed even for a careful definition of the log-likelihood. The expectation $\mathsf{E}\,\ell'_X(\theta_0)$, mentioned in condition (II), may be understood as $\int_{\mathcal{X}_{>0}} p'_x(\theta_0)\mu(\mathrm{d}x)$, where $p_x(\theta) := p_\theta(x)$; similarly, for the other expectations mentioned in conditions (II)–(IV). Of course, all the derivatives at this point are with respect to θ.

Concerning the "standard regularity conditions" mentioned in condition (II), it will be enough to assume that $\mathsf{P}(\frac{\partial}{\partial\theta}p_\theta(X) \neq 0) > 0$ and for some measurable function $g\colon \mathcal{X}_{>0} \to [0, \infty)$ such that $\int_{\mathcal{X}_{>0}} g\,\mathrm{d}\mu < \infty$ and all $\theta \in [\theta_0 - \delta, \theta_0 + \delta]$ and $x \in \mathcal{X}_{>0}$ we have $|\frac{\partial}{\partial\theta}p_\theta(x)| + |\frac{\partial^2}{\partial\theta^2}p_\theta(x)| \leqslant g(x)$; see e.g. [12, Lemma 5.3, page 116] and [18, Lemma 2.4] (more general conditions can be given using [18, Lemma 2.3]). Then $I(\theta)$ will also be continuous in $\theta \in [\theta_0 - \delta, \theta_0 + \delta]$.

Conditions (I)–(IV) are rather similar to regularity conditions used in related literature; see Remark 7.4 on page 1177 for details. It appears that these conditions will be generally satisfied provided that $\ell_x(\theta)$ is smooth enough in θ.

For instance, let us briefly consider the case when the family of densities (p_θ) is a location family, so that $\ell_x(\theta) = \lambda(x - \theta)$ for all $(x, \theta) \in \mathcal{X} \times \Theta = \mathbb{R}^2$, where λ is a smooth enough function. If the densities p_θ have power-like tails, then for some positive real constants c_+ and c_- one has $\lambda(x) \sim -c_\pm \ln|x|$ as $x \to \pm\infty$, in which case typically $|\lambda^{(k)}(x)| \sim -c_\pm k!|x|^{-k} \ln|x|$ for $k = 0, 1, \dots$ as $x \to \pm\infty$. So, conditions (III) and (IV) will hold, since $|\ell_x^{(k)}(\theta)| = |\lambda^{(k)}(x-\theta)|$. If the tails of the densities p_θ are lighter than power-like tails, so that (say) $\lambda(x) \sim -c_\pm|x|^\alpha$ for some real $\alpha > 0$ as $x \to \pm\infty$, then typically $|\lambda^{(k)}(x)| \sim -c_\pm k!|x|^{\alpha-k}$ for $k = 0, 1, \dots$ as $x \to \pm\infty$, so that conditions (III) and (IV) will again hold.

The case of a scale family is quite similar to that of a location family. Alternatively, the "scale" case can be reduced to the "location" one by logarithmic rescaling in both x and θ.

At this point, consider also the case when the family of densities (p_θ) is an exponential family, so that $\ell_x(\theta) = w(\theta)T(x) + d(\theta)$ for some functions w, T, and d and for all $(x, \theta) \in \mathcal{X} \times \Theta = \mathbb{R}^2$, where the functions w and d are smooth enough, with $w'(\theta_0) \neq 0$. Then $\ell_x^{(k)}(\theta) = w^{(k)}(\theta)T(x) + d^{(k)}(\theta)$. So, conditions (III) and (IV) will hold in this case as well, since $\mathsf{E}\,|T(X)|^\alpha = \int_{\mathcal{X}} |T(x)|^\alpha \exp\{w(\theta_0)T(x) + d(\theta_0)\}\mu(\mathrm{d}x)$ for $\alpha > 0$, $|T(x)|^\alpha = O(e^{hT(x)} + e^{-hT(x)})$ for any given real $\alpha > 0$ and any given nonzero real h, and the conditions $\theta_0 \in \Theta^\circ$ and $w'(\theta_0) \neq 0$ imply that $\int_{\mathcal{X}} \exp\{[w(\theta_0) + h]T(x) + d(\theta_0)\}\mu(\mathrm{d}x) < \infty$ for all real h close enough to 0. □

Let

$$\ell_{\mathbf{X}}(\theta) := \sum_{i=1}^n \ell_{X_i}(\theta) \tag{2.2}$$

for $\theta \in \Theta$, the log-likelihood of the sample $\mathbf{X} = (X_1, \dots, X_n)$.

3. Tight bracketing of the MLE between two functions of the sum of independent random vectors

Without loss of generality (w.l.o.g.), $\mathcal{X}_{>0} = \mathcal{X}$. Then on the event

$$G := \{\hat\theta \in [\theta_0 - \delta, \theta_0 + \delta]\} \tag{3.1}$$

(G for "good event") one must have

$$\begin{aligned} 0 = \ell'_{\mathbf{X}}(\hat\theta) =& \ell'_{\mathbf{X}}(\theta_0) + (\hat\theta - \theta_0)\, \ell''_{\mathbf{X}}(\theta_0) + \frac{(\hat\theta-\theta_0)^2}{2}\, \ell'''_{\mathbf{X}}(\theta_0 + \xi(\hat\theta - \theta_0)) && (3.2)\\ =& n\Big(\overline{Z} - (\hat\theta-\theta_0)\,\overline{U} + \frac{(\hat\theta-\theta_0)^2}{2}\,\overline{R}\Big) && (3.3) \end{aligned}$$

for some $\xi \in (0,1)$, depending on the values of the X_i's, where $\overline{Z} := \frac{1}{n}\sum_{i=1}^n Z_i$, $\overline{U} := \frac{1}{n}\sum_{i=1}^n U_i$, $\overline{R} := \frac{1}{n}\sum_{i=1}^n R_i$, $\overline{R^*} := \frac{1}{n}\sum_{i=1}^n R_i^*$,

$$\begin{aligned} Z_i &:= \ell'_{X_i}(\theta_0), \quad U_i := -\ell''_{X_i}(\theta_0), \\ R_i &:= \ell'''_{X_i}(\theta_0 + \xi(\hat\theta-\theta_0)) \in [-R_i^*, R_i^*], \quad R_i^* := \sup_{\theta\in[\theta_0-\delta,\theta_0+\delta]} |\ell'''_{X_i}(\theta)|. \end{aligned} \tag{3.4}$$

Note that the Z_i's are i.i.d. r.v.'s, and so are the U_i's and the R_i^*'s (but not necessarily the R_i's).

Equalities (3.2) and (3.3) provide a quadratic equation for $\hat\theta$. So, on the event G one has

$$\begin{aligned} \hat\theta - \theta_0 &= \frac{\overline{Z}}{\overline{U}} && \text{if } \overline{R} = 0 \;\&\; \overline{U} \neq 0, \\ \hat\theta - \theta_0 &\in \{d_+, d_-\} && \text{if } \overline{R} \neq 0, \end{aligned} \tag{3.5}$$

where

$$d_\pm := \frac{\overline{U} \pm \sqrt{\overline{U}^2 - 2\overline{Z}\,\overline{R}}}{\overline{R}}.$$

Letting

$$\begin{aligned} &B := B_1 \cup B_2, \quad \text{where} \\ &B_1 := \{\overline{R} \neq 0,\ \hat\theta - \theta_0 = d_+\} \cup \{\overline{U} \leqslant 0\} \quad \text{and} \quad B_2 := \{\overline{U}^2 \leqslant 2|\overline{Z}|\,\overline{R^*}\} \end{aligned} \tag{3.6}$$

(B for "bad event"), on the event $B_1 \cap \{\overline{U} > 0\}$ one has $|\hat\theta - \theta_0| = |d_+| \geqslant \overline{U}/|\overline{R}| \geqslant \overline{U}/\overline{R^*}$, whence, by (3.1),

$$\mathsf{P}(G \cap B_1) \leqslant \mathsf{P}\Big(\overline{U} \leqslant 0 \text{ or } \frac{\overline{U}}{\overline{R^*}} \leqslant \delta\Big) = \mathsf{P}\Big(\frac{\overline{U}}{\overline{R^*}} \leqslant \delta\Big) = \mathsf{P}\Big(\sum_{i=1}^n (U_i - \delta R_i^*) \leqslant 0\Big). \tag{3.7}$$

By definitions (3.4) and conditions (II), (III), and (IV),

$$\mathsf{E}\, U_1 > 0, \quad \mathsf{E}\,|Z_1|^3 < \infty, \quad \mathsf{E}\,|U_1|^3 < \infty, \quad \mathsf{E}(R_1^*)^3 < \infty, \tag{3.8}$$

and hence $\mathsf{E}\,R_1^* < \infty$. So, w.l.o.g. one may choose $\delta > 0$ to be small enough so that

$$\delta_1 := \mathsf{E}(U_i - \delta R_i^*) > 0.$$

Then, letting $Y_i := (U_i - \delta R_i^*) - \mathsf{E}(U_i - \delta R_i^*)$ and using (3.7), Markov's inequality, and a Rosenthal-type inequality (see e.g. [18, Theorem 1.5]), we have

$$\begin{aligned} \mathsf{P}(G \cap B_1) \leqslant \mathsf{P}\Big(\sum_{i=1}^n Y_i \leqslant -n\delta_1\Big) &\leqslant \frac{1}{(n\delta_1)^3}\,\mathsf{E}\Big|\sum_{i=1}^n Y_i\Big|^3 \\ &\leqslant \frac{n\,\mathsf{E}\,|Y_1|^3 + \sqrt{8/\pi}\,(n\,\mathsf{E}\,Y_1^2)^{3/2}}{(n\delta_1)^3} \leqslant \frac{\mathfrak{C}}{n^{3/2}}, \end{aligned} \tag{3.9}$$

where $\mathfrak{C} := \big(\mathsf{E}\,|Y_1|^3 + \sqrt{8/\pi}\,(\mathsf{E}\,Y_1^2)^{3/2}\big)/\delta_1^3$, which depends on $\delta_1 > 0$, $\mathsf{E}\,Y_1^2 < \infty$, and $\mathsf{E}\,|Y_1|^3 < \infty$ – but not on n.

Next, the occurrence of B_2 implies the occurrence of at least one of the following events: $B_{21} := \{\overline{U} \leqslant \frac12\,\mathsf{E}\,U_1\}$, $B_{22} := \{\overline{R^*} \geqslant 1 + \mathsf{E}\,R_1^*\}$, or $B_{23} := \{|\overline{Z}| \geqslant \frac18\,(\mathsf{E}\,U_1)^2/(1 + \mathsf{E}\,R_1^*)\}$. So,

$$\mathsf{P}(B_2) \leqslant \mathsf{P}(B_{21}) + \mathsf{P}(B_{22}) + \mathsf{P}(B_{23}). \tag{3.10}$$

In view of (3.8), the bounding of each of the probabilities $\mathsf{P}(B_{21})$, $\mathsf{P}(B_{22})$, $\mathsf{P}(B_{23})$ is quite similar to the bounding of $\mathsf{P}(G \cap B_1)$ in (3.9) – because $\mathsf{P}(B_{21}) = \mathsf{P}(\sum_{i=1}^n Y_{i,21} \leqslant -n\delta_{21})$, $\mathsf{P}(B_{22}) = \mathsf{P}(\sum_{i=1}^n Y_{i,22} \geqslant n\delta_{22})$, and $\mathsf{P}(B_{23}) = \mathsf{P}(\sum_{i=1}^n |Y_{i,23}| \geqslant n\delta_{23})$, where $Y_{i,21} := U_i - \mathsf{E}\,U_1$, $\delta_{21} := \frac12\,\mathsf{E}\,U_1 > 0$, $Y_{i,22} := R_i^* - \mathsf{E}\,R_1^*$, $\delta_{22} := 1 > 0$, $Y_{i,23} := Z_i - \mathsf{E}\,Z_1 = Z_i$, $\delta_{23} := \frac18\,(\mathsf{E}\,U_1)^2/(1 + \mathsf{E}\,R_1^*) > 0$.

Thus, by (3.6), (3.9), and (3.10),

$$\mathsf{P}(G \cap B) \leqslant \mathsf{P}(G \cap B_1) + \mathsf{P}(B_2) \leqslant \frac{\mathfrak{C}}{n^{3/2}}, \tag{3.11}$$

where $\mathfrak{C}$ depends on the likelihood function, the measure μ, and the choice of θ_0 – but not on n.

On the other hand, if $\overline{R} \neq 0$ and $\overline{U} > 0$, then $d_- = \frac{2\overline{Z}}{\overline{U} + \sqrt{\overline{U}^2 - 2\overline{Z}\,\overline{R}}}$; here, the condition $\overline{U} > 0$ was used only to ensure that the denominator of the latter ratio is nonzero. Hence, on the event $G \setminus B$ one has

$$\overline{U} > 0 \quad \text{and} \quad \hat\theta - \theta_0 = \frac{2\overline{Z}}{\overline{U} + \sqrt{\overline{U}^2 - 2\overline{Z}\,\overline{R}}} \in [T_-, T_+], \tag{3.12}$$

where

$$T_\pm := \frac{2\overline{Z}}{\overline{U} + \sqrt{\overline{U}^2 \mp 2|\overline{Z}|\,\overline{R^*}}}; \tag{3.13}$$

note that, when $\overline{R} = 0$ and $\overline{U} > 0$, the expression of $\hat\theta - \theta_0$ in (3.12) is in agreement with the corresponding expression in (3.5).

Now that the desired bracketing of $\hat\theta - \theta_0$ between T_- and T_+ is obtained in (3.12), we are ready to apply some of the mentioned general results of [20], presented in the next section.

4. General uniform and nonuniform bounds from [20] on the rate of convergence to normality for smooth nonlinear functions of sums of independent random vectors

The standard normal distribution function (d.f.) will be denoted by Φ. For any $\mathbb{R}^d$-valued random vector ζ, we use the norm notation

$$\|\zeta\|_p := \big(\mathsf{E}\,\|\zeta\|^p\big)^{1/p} \text{ for any real } p \geqslant 1,$$

where $\|\cdot\|$ denotes the Euclidean norm on $\mathbb{R}^d$.

Take any Borel-measurable functional $f\colon \mathbb{R}^d \to \mathbb{R}$ satisfying the following smoothness condition: there exist $\epsilon \in (0,\infty)$, $M_\epsilon \in (0,\infty)$, and a linear functional $L\colon \mathbb{R}^d \to \mathbb{R}$ such that

$$|f(\mathbf{x}) - L(\mathbf{x})| \leqslant \frac{M_\epsilon}{2}\,\|\mathbf{x}\|^2 \text{ for all } \mathbf{x} \in \mathbb{R}^d \text{ with } \|\mathbf{x}\| \leqslant \epsilon. \tag{4.1}$$

Thus, $f(\mathbf{0}) = 0$ and L necessarily coincides with the first Fréchet derivative, $f'(\mathbf{0})$, of the function f at $\mathbf{0}$. Moreover, for the smoothness condition (4.1) to hold, it is enough that

$$M_\epsilon \geqslant M_\epsilon^* := \sup\left\{\frac{1}{\|\mathbf{x}\|^2}\left|\frac{\mathrm{d}^2}{\mathrm{d}t^2}\, f(\mathbf{x}+t\mathbf{x})\Big|_{t=0}\right| : \mathbf{x} \in \mathbb{R}^d,\ 0 < \|\mathbf{x}\| \leqslant \epsilon\right\}; \tag{4.2}$$

it is not necessary that f be twice differentiable at $\mathbf{0}$. E.g., if $d = 1$ and $f(x) = \frac{x}{1+|x|}$ for $x \in \mathbb{R}$, then $f(0) = 0$, $f'(0) = 1$, and $f''(x) = -\frac{2\,\operatorname{sign} x}{(1+|x|)^3}$ for real $x \neq 0$; so, (4.1) holds for any real $\epsilon > 0$ with $L(x) \equiv x$ and $M_\epsilon = 2$, whereas $f''(0)$ does not exist.

Let $V, V_1, \dots, V_n$ be i.i.d. random vectors in $\mathbb{R}^d$, with $\mathsf{E}\,V = \mathbf{0}$ and

$$\overline{V} := \frac{1}{n}\sum_{i=1}^n V_i.$$

Further let

$$\tilde{\sigma} := \|L(V)\|_2, \quad v_3 := \|V\|_3, \quad \text{and} \quad \varsigma_3 := \frac{\|L(V)\|_3}{\tilde{\sigma}}. \tag{4.3}$$

Theorem 4.1. [20] *Suppose that (4.1) holds, and that $\tilde{\sigma} > 0$ and $v_3 < \infty$. Then for all $z \in \mathbb{R}$*

$$\left|\mathsf{P}\left(\frac{f(\overline{V})}{\tilde{\sigma}/\sqrt{n}} \leqslant z\right) - \Phi(z)\right| \leqslant \frac{\mathfrak{C}}{\sqrt{n}}, \tag{4.4}$$

where $\mathfrak{C}$ is a finite positive expression that depends only on the function f (through (4.1)) and the moments $\tilde{\sigma}$, ς_3, and v_3. Moreover, for any $\omega \in (0,\infty)$ and for all

$$z \in \big(0, \omega\sqrt{n}\,\big] \tag{4.5}$$

one has

$$\left|\mathsf{P}\left(\frac{f(\overline{V})}{\tilde{\sigma}/\sqrt{n}} \leqslant z\right) - \Phi(z)\right| \leqslant \frac{\mathfrak{C}_\omega}{z^3\sqrt{n}}, \tag{4.6}$$

where $\mathfrak{C}_\omega$ is a finite positive expression that depends only on the function f (through (4.1)), the moments $\tilde{\sigma}$, ς_3, and v_3, and also on ω.

The restriction (4.5) cannot be relaxed in general; see [20].

To simplify the presentation, in what follows let $\mathfrak{C}$ stand for various finite positive expressions whose values do not depend on n or z; that is, $\mathfrak{C}$ will denote various positive real constants – with respect to n and z. However, $\mathfrak{C}$ may depend on other attributes of the setting, including the model $(\mathsf{P}_\theta)_{\theta\in\Theta}$ under consideration, the P_{θ_0}-distribution of X_1, the measure μ, and the values of parameters freely chosen in a given range (such as ω in (4.5) and ε in (4.1)).

5. Making the bracketing work: Applying the general bounds of [20]

Now let $d = 3$ and then let

$$\mathcal{D} := \{\mathbf{x} = (x_1, x_2, x_3) \in \mathbb{R}^d = \mathbb{R}^3 \colon x_2 + \mathsf{E}\,U_1 > 0,\ (x_2 + \mathsf{E}\,U_1)^2 > 2|x_1|\,|x_3 + \mathsf{E}\,R_1^*|\}.$$

By (3.4) and conditions (II) and (IV), $\mathsf{E}\,U_1 = I(\theta_0) \in (0,\infty)$ and $\mathsf{E}\,R_1^* \in [0,\infty)$. So, for some real $\epsilon > 0$, the set $\mathcal{D}$ contains the ϵ-neighborhood of the origin $\mathbf{0}$ of $\mathbb{R}^3$.

Define functions $f_\pm \colon \mathbb{R}^3 \to \mathbb{R}$ by the formula

$$f_\pm(\mathbf{x}) = f_\pm(x_1, x_2, x_3) = \frac{2x_1}{x_2 + \mathsf{E}\,U_1 + \sqrt{(x_2 + \mathsf{E}\,U_1)^2 \mp 2|x_1|\,|x_3 + \mathsf{E}\,R_1^*|}} \tag{5.1}$$

for $\mathbf{x} = (x_1, x_2, x_3) \in \mathcal{D}$, and let $f(\mathbf{x}) := 0$ if $\mathbf{x} \in \mathbb{R}^3 \setminus \mathcal{D}$. Clearly, $f_\pm(\mathbf{0}) = 0$,

$$L_\pm(\mathbf{x}) := f'_\pm(\mathbf{0})(\mathbf{x}) = \frac{x_1}{\mathsf{E}\,U_1} = \frac{x_1}{I(\theta_0)} \tag{5.2}$$

for $\mathbf{x} = (x_1, x_2, x_3) \in \mathbb{R}^3$, and, in accordance with (4.2), the smoothness condition (4.1) holds for some ϵ and M_ϵ in $(0,\infty)$ – because, as was noted above, $\mathsf{E}\,U_1 = I(\theta_0) \in (0,\infty)$ and $\mathsf{E}\,R_1^* \in [0,\infty)$, and hence the denominator of the ratio in (5.1) is bounded away from 0 for $\mathbf{x} = (x_1, x_2, x_3)$ in a neighborhood of $\mathbf{0}$.

Next, let

$$V_i := (Z_i, U_i - \mathsf{E}\,U_i, R_i^* - \mathsf{E}\,R_i^*) \tag{5.3}$$

for $i = 1, \dots, n$, with Z_i, U_i, R_i^* as defined in (3.4). Then, by (4.3), (5.2), and condition (II), for $f = f_\pm$,

$$\tilde{\sigma} = \sqrt{\frac{\mathsf{E}\,Z_1^2}{I(\theta_0)^2}} = \frac{1}{\sqrt{I(\theta_0)}} > 0 \tag{5.4}$$

and $v_3^3 = \mathsf{E}\,\|V\|^3 < \infty$ by conditions (III) and (IV). So, all the conditions of Theorem 4.1 are satisfied for $f = f_\pm$.

Moreover, by (3.13), (5.1), and (5.3),

$$T_\pm = f_\pm(\overline{V})$$

on the event $G \setminus B$. So, by the inclusion relation in (3.12) (which holds on the event $G \setminus B = (G^{\mathsf{c}} \cup B)^{\mathsf{c}}$, where ${}^{\mathsf{c}}$ denotes the complement) and (5.4), inequality (4.4) in Theorem 4.1 implies

$$\begin{aligned}\mathsf{P}\left(\sqrt{nI(\theta_0)}\,(\hat\theta - \theta_0) \leqslant z\right) &\leqslant \mathsf{P}\left(\sqrt{nI(\theta_0)}\,f_-(\overline{V}) \leqslant z\right) + \mathsf{P}(G^{\mathsf{c}} \cup B)\\ &\leqslant \Phi(z) + \frac{\mathfrak{C}}{\sqrt{n}} + \mathsf{P}(G^{\mathsf{c}} \cup B)\end{aligned}$$

and, quite similarly,

$$\begin{aligned}\mathsf{P}\left(\sqrt{nI(\theta_0)}\,(\hat\theta - \theta_0) \leqslant z\right) &\geqslant \mathsf{P}\left(\sqrt{nI(\theta_0)}\,f_+(\overline{V}) \leqslant z\right) - \mathsf{P}(G^{\mathsf{c}} \cup B)\\ &\geqslant \Phi(z) - \frac{\mathfrak{C}}{\sqrt{n}} - \mathsf{P}(G^{\mathsf{c}} \cup B),\end{aligned}$$

for all real z. Note that $\mathsf{P}(G^{\mathsf{c}} \cup B) = \mathsf{P}(G^{\mathsf{c}}) + \mathsf{P}(G \cap B)$. It follows now by (3.1) and (3.11) that

$$\left|\,\mathsf{P}\left(\sqrt{nI(\theta_0)}\,(\hat\theta - \theta_0) \leqslant z\right) - \Phi(z)\right| \leqslant \frac{\mathfrak{C}}{\sqrt{n}} + \mathsf{P}(|\hat\theta - \theta_0| > \delta) \tag{5.5}$$

for all real z. Quite similarly, but using (4.6) instead of (4.4), one has

$$\left|\,\mathsf{P}\left(\sqrt{nI(\theta_0)}\,(\hat\theta - \theta_0) \leqslant z\right) - \Phi(z)\right| \leqslant \frac{\mathfrak{C}}{z^3\sqrt{n}} + \mathsf{P}(|\hat\theta - \theta_0| > \delta) \tag{5.6}$$

for z as in (4.5).

Typically, given rather standard regularity conditions, the remainder term $\mathsf{P}(|\hat\theta - \theta_0| > \delta)$ decreases exponentially fast in n and thus is negligible as compared with the “error” term $\frac{\mathfrak{C}}{\sqrt{n}}$, and even with the “error” term $\frac{\mathfrak{C}}{z^3\sqrt{n}}$ – under condition (4.5). Some details on this can be found in the following section.

6. Exponentially small bounds on the remainder term $\mathsf{P}(|\hat\theta - \theta_0| > \delta)$

6.1. Bounding the remainder: Log-concave case

In this subsection, suppose that the log-likelihood $\ell_x(\theta)$ is concave in $\theta \in \Theta$, for each $x \in \mathcal{X}$. By condition (II), $\mathsf{E}\,\ell''_X(\theta_0) \neq 0$. Hence, $\mathsf{P}\left(p_{\theta_0+h}(X) \neq p_{\theta_0}(X)\right) =$

$\mathsf{P}\left(\ell_X(\theta_0+h)\neq\ell_X(\theta_0)\right)>0$ for some $h\in(0,\delta)$. The concavity of $\ell_x(\theta)$ in θ implies that of $\ell_{\mathbf{X}}(\theta)$. So, if $\hat\theta>\theta_0+\delta$, then $\ell_{\mathbf{X}}(\theta_0+h)\geqslant\ell_{\mathbf{X}}(\theta_0)$. Therefore,

$$\mathsf{P}(\hat\theta>\theta_0+\delta)\leqslant\mathsf{P}\left(\ell_{\mathbf{X}}(\theta_0+h)\geqslant\ell_{\mathbf{X}}(\theta_0)\right)=\mathsf{P}\Big(\prod_{i=1}^n\sqrt{\frac{p_{\theta_0+h}(X_i)}{p_{\theta_0}(X_i)}}\geqslant1\Big)$$
$$\leqslant\mathsf{E}\prod_{i=1}^n\sqrt{\frac{p_{\theta_0+h}(X_i)}{p_{\theta_0}(X_i)}}=\lambda_+^n,$$

where

$$\lambda_+:=\mathsf{E}\sqrt{\frac{p_{\theta_0+h}(X)}{p_{\theta_0}(X)}}<\sqrt{\mathsf{E}\,\frac{p_{\theta_0+h}(X)}{p_{\theta_0}(X)}}=\sqrt{\mathsf{E}_{\theta_0}\,\frac{p_{\theta_0+h}(X)}{p_{\theta_0}(X)}}=1;$$

the inequality here is an instance of a strict version of the Cauchy–Schwarz inequality, which holds because, as was noted, $\mathsf{P}\left(p_{\theta_0+h}(X)\neq p_{\theta_0}(X)\right)>0$. Quite similarly, $\mathsf{P}(\hat\theta<\theta_0-\delta)\leqslant\lambda_-^n$ for some $\lambda_-\in[0,1)$, and so,

$$\mathsf{P}(|\hat\theta-\theta_0|>\delta)\leqslant2\lambda^n \tag{6.1}$$

for $\lambda:=\max(\lambda_+,\lambda_-)\in[0,1)$.

In particular, the condition of the concavity of the log-likelihood $\ell_x(\theta)=\ln p_\theta(x)$ in θ is fulfilled in the important case when the densities p_θ form an exponential family with θ as the natural parameter, so that

$$p_\theta(x)=e^{\theta g(x)-\psi(\theta)}$$

for some function $\psi\colon\Theta\to\mathbb{R}$ and all $\theta\in\Theta$ and $x\in\mathcal{X}$. Here, $g\colon\mathcal{X}\to\mathbb{R}$ is a measurable function. Then necessarily $\psi(\theta)=\ln\int_{\mathcal{X}}e^{\theta g(x)}\mu(dx)$, which is convex in θ – because any mixture of log-convex functions is log-convex, as is well known – see e.g. [8, page 66, Theorem 5.4C]. So, $\ell_x(\theta)=\ln p_\theta(x)=\theta g(x)-\psi(\theta)$ is indeed concave in θ. In the case of multivariate exponential families, an exponentially decreasing bound of a form more complicated than that of the bound in (6.1) was given in [11].

6.2. Bounding the remainder: General case

Upper bounds on the large-deviation probability $\mathsf{P}(|\hat\theta-\theta_0|>\delta)$ that are exponentially decreasing in n without the assumption of the concavity of the log-likelihood function were presented e.g. in [22, 21, 15, 5, 13]. However, the parameter space Θ was assumed in [22, 21, 5] to be bounded, whereas in [13] the distributions P_θ were assumed to be subgaussian (cf. Theorems 2.1, 2.2, and 3.3 in [13]). Conditions in [15] appear to be difficult to verify, including the strict positivity of the infimum of the rate function, needed for an actual exponential decrease.

Related is the work [7], containing a result on so-called moderate deviation probabilities for MLEs, which decrease slower than exponentially but still faster than any powers. So, such a result would be enough for our conclusions in Theorem 7.1 in the next section (cf. Remark 7.2 there), if it were not assumed in [7] (as in [22, 21, 5]) that Θ is bounded.

Here we modify the method of [5] to get rid of the condition that Θ is bounded. Consider the (squared) Hellinger distance

$$H(\theta,\theta_0) := \int_{\mathcal{X}} \left(\sqrt{p_\theta} - \sqrt{p_{\theta_0}}\right)^2 \mathrm{d}\mu \tag{6.2}$$

between the probability measures P_θ and P_{θ_0}.

Assume now the following conditions:

(B) The set Θ is a (possibly infinite) interval, and the Fisher information $I(\theta)$ is well defined and satisfies the boundedness condition

$$I(\theta) \leqslant c_1 + c_2|\theta - \theta_0|^\alpha \tag{6.3}$$

for some positive real constants c_1, c_2, α and all $\theta \in \Theta$. (If a point θ in Θ is an endpoint of the interval Θ, then $I(\theta)$ is naturally understood in terms of the corresponding one-sided derivative of $p_\theta(x)$ in θ.)

(D$_0$) For each bounded neighborhood U of θ_0,

$$H(\theta,\theta_0) \gtrsim (\theta - \theta_0)^2 \tag{6.4}$$

over all $\theta \in U$.

(D$_1$) For some real constant $\gamma > 0$ and some bounded neighborhood V of θ_0,

$$J(\theta,\theta_0) := 1 - \tfrac{1}{2}\,H(\theta,\theta_0) = \int_{\mathcal{X}} \sqrt{p_\theta}\sqrt{p_{\theta_0}}\,\mathrm{d}\mu \lesssim |\theta - \theta_0|^{-\gamma} \tag{6.5}$$

over all $\theta \in \Theta \setminus V$.

Here and in the sequel, for any two expressions $E_1 > 0$ and $E_2 \geqslant 0$ whose values depend on some variables, the relation $E_1 \gtrsim E_2$ and its equivalent $E_2 \lesssim E_1$ mean that $\sup(E_2/E_1) < \infty$, where the supremum is taken over the corresponding specified range of values of the variables.

Conditions (D$_0$) and (D$_1$) may be referred to as distinguishability conditions: (D$_0$) means that the probability measures P_θ and P_{θ_0} are not too close to each other for θ in a punctured neighborhood of θ_0 – whereas (D$_1$) implies that for θ far away from θ_0, the probability measures P_θ and P_{θ_0} are almost mutually singular, and thus, easily distinguishable, at least in principle.

Remark 6.1. In the particular case when the parameter space Θ is compact (or just bounded), condition (D$_1$) trivially holds. Moreover, as shown in [5, Section 31], if Θ is compact and the Fisher information $I(\theta)$ is continuous in $\theta \in \Theta$ and strictly positive for $\theta \in \Theta$, then (6.4) holds over all $\theta \in \Theta$. So, condition (D$_0$) holds (whether the set Θ is bounded or not) whenever the Fisher information $I(\cdot)$ is continuous and strictly positive on Θ.

However, since $H(\theta,\theta_0)$ is always bounded from above by 2, it is clear that condition (6.4) cannot possibly hold over all $\theta\in\Theta$ if the parameter space Θ is unbounded. In such a case, we need to complement condition (D_0) by condition (D_1), which latter appears to be natural, and it is indeed commonly satisfied. In particular, conditions (D_0) and (D_1) (as well as regularity conditions (I)–(IV)) hold if p_θ is the density belonging to any one of the following families of probability distributions:

(a) $\mathrm{N}(\theta,\sigma^2)$ – with $\sigma>0$ known, $\Theta=\mathbb{R}$, $H(\theta,\theta_0)=2-2\exp\big\{-\frac{(\theta-\theta_0)^2}{8\sigma^2}\big\}$;
(b) $\mathrm{N}(\mu,\theta^2)$ – with $\mu>0$ known, $\Theta=(0,\infty)$, $H(\theta,\theta_0)=2-2\sqrt{\frac{2\theta\theta_0}{\theta^2+\theta_0^2}}$;
(c) $\mathrm{Exp}(\theta)$ – with $\Theta=(0,\infty)$, $H(\theta,\theta_0)=2-\frac{4\sqrt{\theta\theta_0}}{\theta+\theta_0}$;
(d) more generally, Weibull distributions $\mathrm{W}(k,\theta)$ – with $p_\theta(x)\equiv\frac{k}{\theta}(\frac{x}{\theta})^{k-1}e^{-(x/\theta)^k}I\{x>0\}$, $k>0$ known, $\Theta=(0,\infty)$, $H(\theta,\theta_0)=2-\frac{4(\theta\theta_0)^{k/2}}{\theta^k+\theta_0^k}$;
(e) $\mathrm{Gamma}(\theta,\beta)$ – with scale parameter $\beta>0$ known, $\Theta=(0,\infty)$, $H(\theta,\theta_0)=2-\frac{2\Gamma((\theta+\theta_0)/2)}{\sqrt{\Gamma(\theta)\Gamma(\theta_0)}}$;
(f) $\mathrm{Gamma}(\alpha,\theta)$ – with shape parameter $\alpha>0$ known, $\Theta=(0,\infty)$, $H(\theta,\theta_0)=2-\frac{2^{1+\alpha}(\theta\theta_0)^{\alpha/2}}{(\theta+\theta_0)^\alpha}$;
(g) $\mathrm{Poisson}(\theta)$ – with $\Theta=(0,\infty)$, $H(\theta,\theta_0)=2-2e^{-(\sqrt{\theta}-\sqrt{\theta_0})^2/2}$;
(h) $\mathrm{Beta}(s\theta,s(1-\theta))$ – with $s>0$ known, $\Theta=(0,1)$, $H(\theta,\theta_0)=2-\frac{2\mathrm{B}\big(\frac12 s(\theta+\theta_0),\frac12 s(2-\theta-\theta_0)\big)}{\sqrt{\mathrm{B}(s\theta,s-s\theta)\mathrm{B}(s\theta_0,s-s\theta_0)}}$, where $\mathrm{B}(\cdot,\cdot)$ is the Beta function;
(i) $\mathrm{Beta}(\alpha\theta,\beta\theta)$ – with $\alpha,\beta>0$ known, $\Theta=(0,\infty)$, $H(\theta,\theta_0)=2-\frac{2\mathrm{B}\big(\frac12\alpha(\theta+\theta_0),\frac12\beta(\theta+\theta_0)\big)}{\sqrt{\mathrm{B}(\alpha\theta,\beta\theta)\mathrm{B}(\alpha\theta_0,\beta\theta_0)}}$ (in this case, by Stirling's formula, condition (D_1) holds with $\gamma=1/4$).

Item (a) above, concerning the normal location family, can be quite broadly generalized:

Proposition 6.2. *Suppose that $(p_\theta)_{\theta\in\Theta}$ is a location family over $\mathbb{R}$, so that $p_\theta(x)=p(x-\theta)$ for all $x\in\mathbb{R}$ and $\theta\in\Theta$, where p is a pdf (with respect to the Lebesgue measure over $\mathbb{R}$). Suppose also that*

$$p(u)\lesssim(1+|u|)^{-\alpha}\tag{6.6}$$

for some real $\alpha>1$ and all real u. Then condition (D_1) holds.

Note that the restriction $\alpha>1$, together with (6.6), implies the integrability of the nonnegative function p.

Proof of Proposition 6.2. Without loss of generality, $\theta_0=-\theta$ and $\theta>0$, so that $\theta-\theta_0=2\theta>0$ and

$$J(\theta,\theta_0)=\int_{\mathbb{R}}\sqrt{p(x+\theta)}\sqrt{p(x-\theta)}\,\mathrm{d}x=\int_{|x|\geqslant2\theta}\cdots+\int_{|x|<2\theta}\cdots.\tag{6.7}$$

Since $|x| \geqslant 2\theta$ implies $|x \pm \theta| \gtrsim |x|$, condition (6.6) yields

$$\int_{|x|\geqslant 2\theta} \cdots \lesssim \int_{|x|\geqslant 2\theta} |x|^{-\alpha}\, \mathrm{d}x \lesssim \theta^{1-\alpha}. \tag{6.8}$$

Since $0 \leqslant x < 2\theta$ implies $x + \theta \geqslant \theta$ and $-\theta \leqslant x - \theta < \theta$, condition (6.6) yields

$$\int_0^{2\theta} \ldots \lesssim \theta^{-\alpha/2} \int_{-\theta}^{\theta} (1 + |u|)^{-\alpha/2}\, \mathrm{d}u \lesssim \theta^{-\alpha/2}\, \theta^{0\vee(1-\alpha/2)} \ln\theta \tag{6.9}$$

for (say) $\theta \geqslant 2$; the factor $\ln\theta$ is actually needed here only in the case when $\alpha = 2$. The integral $\int_{-2\theta}^{0} \cdots$ can be bounded quite similarly. So,

$$\int_{|x|<2\theta} \ldots \lesssim \theta^{-\alpha/2}\, \theta^{0\vee(1-\alpha/2)} \ln\theta \tag{6.10}$$

for $\theta \geqslant 2$. Thus, (D_1) holds for any $\gamma \in (0, \gamma_\alpha)$, where $\gamma_\alpha := \frac{\alpha}{2} - \big(0 \vee (1 - \frac{\alpha}{2})\big) = \frac{\alpha}{2} \wedge (\alpha - 1) > 0$. □

The problem concerning the possibility of a non-compact parameter space Θ may be illustrated by the following simple example:

Example 6.3. For $\theta \in \Theta = (-1, \infty)$, let p_θ be the density (with respect to the Lebesgue measure on $\mathbb{R}$) of the normal distribution with mean $\mu(\theta) := \frac{\theta}{1+\theta^2}$ and variance $\sigma^2(\theta) := \frac{(1+\theta)^3-\theta}{1+\theta^3}$, and let $\theta_0 = 0$, so that $\theta_0 \in \Theta^\circ = \Theta$. Then for any two distinct θ and τ in Θ the equality $\mu(\tau) = \mu(\theta)$ implies $\theta \notin \{0, 1\}$ and $\tau = 1/\theta > 0$, whence $\sigma^2(\tau) \neq \sigma^2(\theta)$. So, $p_\tau \neq p_\theta$ for any two distinct θ and τ in Θ. However, $\mu(\theta) \underset{\theta\to\infty}{\longrightarrow} 0 = \mu(0)$ and $\sigma^2(\theta) \underset{\theta\to\infty}{\longrightarrow} 1 = \sigma^2(0)$, so that p_0 is almost indistinguishable from p_θ for large θ. More specifically, it is not hard to check that here

$$J(\theta, \theta_0) = \int_{\mathbb{R}} \sqrt{p_\theta(x)}\sqrt{p_0(x)}\, \mathrm{d}x = \sqrt{\frac{2\sigma(\theta)}{\sigma^2(\theta) + 1}} \exp\Big(-\frac{\mu(\theta)^2}{4\sigma^2(\theta) + 4}\Big) \underset{\theta\to\infty}{\longrightarrow} 1,$$

so that this situation is excluded by condition (6.5).

Now we are well prepared to state the main result of this subsection:

Proposition 6.4. *Under conditions* (B), (D_0), *and* (D_1),

$$\mathsf{P}(|\hat\theta - \theta_0| > \delta) \leqslant c\, \lambda^n \tag{6.11}$$

for some real constants $c > 0$ and $\lambda \in [0, 1)$ (depending on $\gamma, c_0, \alpha, c_1, c_2$) and all natural n; cf. (6.1).

Inequality (6.11) is similar to inequality (6) in [5, Section 33.2, Theorem 3], with the following main differences.

(i) It is assumed in [5] that Θ is compact, in addition to the assumption that $I(\theta)$ is continuous in $\theta\in\Theta$ and strictly positive for $\theta\in\Theta$. Under these assumptions, condition (D$_0$) is, not assumed, but derived in [5]. As noted above, if the parameter space Θ is compact, then condition (D$_1$) is trivial.

(ii) As we do not assume that Θ is compact (or even bounded), we need to control the behavior of log-likelihood $\ell_{\mathbf{X}}(\theta)$ for θ far from θ_0. This is done using condition (D$_1$).

(iii) In [5], instead of condition (B) above, it is assumed that the Fisher information $I(\theta)$ is just bounded over all $\theta\in\Theta$. However, mainly following the lines of proof in [5], one can see that the more general condition (B) suffices, given conditions (D$_0$) and (D$_1$).

For the readers' convenience here is

Proof of Proposition 6.4. Let

$$Z(u):=\frac{p_{\theta_0+u}(\mathbf{X})}{p_{\theta_0}(\mathbf{X})}=\prod_{i=1}^n\frac{p_{\theta_0+u}(X_i)}{p_{\theta_0}(X_i)}=\exp\{\ell_{\mathbf{X}}(\theta_0+u)-\ell_{\mathbf{X}}(\theta_0)\},\tag{6.12}$$

where $p_\theta(\mathbf{X}):=\prod_{i=1}^n p_\theta(X_i)=\exp\ell_{\mathbf{X}}(\theta)$ and $\ell_{\mathbf{X}}$ is the log-likelihood function, as defined in (2.2); here and subsequently in this proof, u is a real number such that $\theta_0+u\in\Theta$.

By conditions (D$_1$) and (D$_0$), there exist real $C_1>0$,

$$u_*>C_1^{1/\gamma}\vee\delta,\tag{6.13}$$

and $C_0>0$ such that

$$\mathsf{E}\,Z(u)^{1/2}=\mathsf{E}_{\theta_0}\,Z(u)^{1/2}=J(\theta_0,\theta_0+u)^n\leqslant C_1^n u^{-n\gamma}\quad\text{if }|u|>u_*\tag{6.14}$$

and

$$\mathsf{E}\,Z(u)^{1/2}=\big(1-\tfrac12\,H(\theta_0,\theta_0+u)\big)^n\leqslant(1-u^2/C_0)^n\leqslant e^{-nu^2/C_0}\quad\text{if }|u|\leqslant u_*.\tag{6.15}$$

Note also that $\mathsf{E}\,Z(u)=1$. So, introducing

$$P(u):=Z(u)^{3/4},\tag{6.16}$$

by the Cauchy–Schwarz inequality one has

$$\mathsf{E}\,P(u)\leqslant\sqrt{\mathsf{E}\,Z(u)\;\mathsf{E}\,Z(u)^{1/2}}=\sqrt{\mathsf{E}\,Z(u)^{1/2}}.\tag{6.17}$$

Further, $P'(u)=\frac34\,\ell'_{\mathbf{X}}(\theta_0+u)Z(u)^{3/4}$, whence, again by the Cauchy–Schwarz inequality,

$$\begin{aligned}\mathsf{E}\,|P'(u)|&\leqslant\tfrac34\sqrt{\mathsf{E}\,\ell'_{\mathbf{X}}(\theta_0+u)^2Z(u)\;\mathsf{E}\,Z(u)^{1/2}}\\&=\tfrac34\sqrt{\mathsf{E}_{\theta_0+u}\,\ell'_{\mathbf{X}}(\theta_0+u)^2\;\mathsf{E}\,Z(u)^{1/2}}\\&=\tfrac34\sqrt{nI(\theta_0+u)\;\mathsf{E}\,Z(u)^{1/2}}.\end{aligned}\tag{6.18}$$

For $u>\delta$, one has $P(u)\leqslant P(\delta)+\int_{\Theta\cap(\delta,\infty)}|P'(t)|\,dt$. So, by (6.17), (6.15), (6.18), (6.3), (6.13), and (6.14),

$$\mathsf{E}\sup_{u>\delta}P(u)\leqslant e^{-n\delta^2/(2C_0)}+I_0+I_1=\lambda_*^n+I_0+I_1,$$

where $\lambda_*:=e^{-\delta^2/(2C_0)}\in(0,1)$,

$$I_0:=\int_\delta^{u_*}\sqrt{n(c_1+c_2u^\alpha)}\,e^{-nu^2/(2C_0)}\,du\lesssim\int_\delta^\infty\sqrt{nu^\alpha}\,e^{-nu^2/(2C_0)}\,du\lesssim\lambda_0^n$$

for any fixed $\lambda_0\in(\lambda_*,1)$, and

$$I_1:=\int_{u_*}^\infty\sqrt{n(c_1+c_2u^\alpha)C_1^nu^{-n\gamma}}\,du\lesssim\lambda_1^{n/2}$$

for any fixed $\lambda_1\in(C_1/u_*^\gamma,1)$ – note that the latter interval is nonempty, in view of (6.13). Thus, $\mathsf{E}\sup_{u>\delta}P(u)\lesssim\lambda^n$ for $\lambda:=\lambda_0\vee\sqrt{\lambda_1}\in(0,1)$. Quite similarly, $\mathsf{E}\sup_{u<-\delta}P(u)\lesssim\lambda^n$ and hence $\mathsf{E}\sup_{|u|>\delta}P(u)\lesssim\lambda^n$. So,

$$\mathsf{P}(|\hat\theta-\theta_0|>\delta)\leqslant\mathsf{P}(\sup_{|u|>\delta}Z(u)\geqslant Z(0))=\mathsf{P}(\sup_{|u|>\delta}P(u)\geqslant1)\leqslant\mathsf{E}\sup_{|u|>\delta}P(u)\lesssim\lambda^n,$$

which completes the proof of Proposition 6.4. □

7. Conclusion

Inequalities (5.5) and (5.6) together with (6.1) and Proposition 6.4 yield

Theorem 7.1. *Suppose that conditions* (I)*–*(IV) *hold. Suppose also that either (i) the log-likelihood $\ell_x(\theta)$ is concave in $\theta\in\Theta$, for each $x\in\mathcal{X}$, or (ii) conditions* (B), (D$_0$)*, and* (D$_1$) *hold. Then*

$$\left|\mathsf{P}\left(\sqrt{nI(\theta_0)}\,(\hat\theta-\theta_0)\leqslant z\right)-\Phi(z)\right|\leqslant\frac{\mathfrak{C}}{\sqrt{n}}\tag{7.1}$$

for all real z, and

$$\left|\mathsf{P}\left(\sqrt{nI(\theta_0)}\,(\hat\theta-\theta_0)\leqslant z\right)-\Phi(z)\right|\leqslant\frac{\mathfrak{C}}{z^3\sqrt{n}}\tag{7.2}$$

for z as in (4.5)*. Here, as before, each of the two instances of the symbol $\mathfrak{C}$ stands for a finite positive expression whose values do not depend on n or z, in accordance with the last paragraph of Section 4.*

Remark 7.2. It should be clear that the conditions assumed in the second sentence of Theorem 7.1 can be replaced by any other conditions that imply (6.11) for some real constants $c>0$ and $\lambda\in[0,1)$ not depending on n. Actually, a much weaker bound, of the form c/n^2, instead of the exponentially fast decreasing upper bound $c\lambda^n$ in (6.11), will already suffice.

Remark 7.3. It is shown in Proposition A.1 of the arXiv version [19] of this paper that, under general regularity conditions, (1.1) (or even a relaxed version of it) implies that the family of densities (p_θ) is a one-parameter exponential one; thus, condition (1.1) is quite restrictive. This allows one to give any number of examples where Theorem 7.1 of the present paper is applicable, whereas [20, Theorem 3.16] is not. Indeed, taking almost any smooth enough location family (cf. Remarks 2.1 and 6.1 and Proposition 6.2), one has an example where Theorem 7.1 of the present paper is applicable, whereas [20, Theorem 3.16] is not. For instance, one may take the Cauchy location family, with $p_\theta(x) = \frac{1}{\pi}\frac{1}{1+(x-\theta)^2}$, or the location family defined by the formula $p_\theta(x) = \frac{1}{2\Gamma(5/4)}\exp\{-(x-\theta)^4\}$, for all real x and θ. (One may note that, whereas the tails in the first example here, of the Cauchy family, are very heavy, they are very light in the second example.) An additional advantage of Theorem 7.1 of the present paper over [20, Theorem 3.16] is that now one does not have to check a special, restrictive condition of the form (1.1) even when it holds. □

Theorem 7.1 can be extended to the more general case of M-estimators. Indeed, the condition that p_θ is a pdf for $\theta \neq \theta_0$ is used in our proofs only in order to state that $\mathsf{E}_\theta\, \ell'_X(\theta) = 0$ and $\mathsf{E}_\theta\, \ell'_X(\theta)^2 = -\,\mathsf{E}_\theta\, \ell''_X(\theta) = I(\theta) \in (0,\infty)$. In the case of M-estimators, the corresponding conditions will have to be just assumed, with some other expressions in place of the Fisher information $I(\theta)$, as it is done e.g. in [16, 17], where uniform bounds of optimal order $O(1/\sqrt{n})$ for M-estimators were obtained; M-estimators were referred to as minimum contrast estimates in [14, 16, 17]. We have chosen to restrict the consideration here to MLEs in order not to obscure the novelty elements in our result.

The most significant novelty in our Theorem 7.1, as compared with the results of [14, 16, 17], is that, in addition to the uniform bound in (7.1), inequality (7.2) in Theorem 7.1 also provides a nonuniform Berry–Esseen-type bound for MLEs in general, which latter appears to be the first such result in the literature – except for the already mentioned special case considered recently in [20]. On the other hand, paper [17] treats the case of a multidimensional parameter θ. The uniform bound in [14] was of the form $O(\sqrt{\ln n}/\sqrt{n})$, rather than of the optimal order $O(1/\sqrt{n})$.

Another notable distinction is that condition [14, (1)] (the same as the corresponding conditions on page 73 in [16] and on page 173 in [17]) effectively reduces the consideration to the case when the parameter space Θ is compact in $[-\infty,\infty]$. This obviates the need in a condition such as (D_1), which is there to control the behavior of the likelihood $\ell_{\mathbf{X}}(\theta)$ for large $|\theta|$. However, as pointed out in [14, page 75] concerning the main result there, the nonconstructive compactification condition used in [14, 16, 17] "gives no method for determining [the] value [of the constant in the Berry–Esseen-type bound] for a given family of probability measures."

The problem of controlling the likelihood over far-away zones of a noncompact parameter space Θ was illustrated in Example 6.3, where the "bad" situation was excluded by condition (6.5). That same situation – with $f_\theta = -\ln p_\theta$ for $\theta \in \overline{\Theta} = [-1, \infty]$ and $\mu(\infty) := \lim_{\theta\to\infty} \mu(\theta) = 0 = \mu(0)$ and variance $\sigma^2(\infty) := \lim_{\theta\to\infty} \sigma^2(\theta) = 1 = \sigma^2(0)$ – was also excluded by the mentioned compactification condition in [14, 16, 17].

As was pointed out, the method of the present paper is based on the general Berry–Esseen bounds for the multivariate delta method obtained in [20], which were apllied here via the bracketing argument delineated in Section 3. As such, this method is quite different from the methods in [14, 16, 17], specialized to deal with MLEs. Partly because of this difference in the methods, there are many differences between the conditions in [14, 16, 17] and those in the present paper. Most of these differences – apart from the ones discussed above – are rather minor. Since the result of [16] is apparently the closest to ours in the literature, let us further discuss the regularity conditions in [16], in comparison with ours, in some detail:

Remark 7.4. Condition (I) in the present paper can be replaced by the condition that $p_\theta > 0$ everywhere on $\mathcal{X}$. The latter condition is necessary in order for $\ell_x(\theta) = \ln p_\theta(x)$ to be defined for all $x \in \mathcal{X}$; cf. the first paragraph on page 83 in [16].

Our condition (II) follows, by Remark 2.1, from regularity conditions (iv), (v)(a), (vi) on pages 83–84 in [16] – for $f_\theta := -\ell_\theta$.

Next, condition (III) follows from [16, (vi)]. Here and in the rest of this remark, the lower-case Roman numerals and letters in parentheses refer to the regularity conditions on pages 83–84 in [16] – again for $f_\theta := -\ell_\theta$.

Condition (IV) is, in main, a bit stronger than [16, (viii)]. Of course, condition (IV) can be relaxed, for the price of making it more complicated.

By Remark 6.1, our condition (D$_0$) will hold if the Fisher information $I(\cdot)$ is continuous and strictly positive on Θ, for which conditions (ix) and (v)(a), respectively, in [16] will be more than enough.

Next, our condition (D$_1$), to control the behavior of the likelihood $\ell_{\mathbf{X}}(\theta)$ for large $|\theta|$, was already discussed at length, versus the compactification condition used in [14, 16, 17].

In the case when Θ is compact, for our condition (B) to hold, either one of regularity conditions (vi)(a) or (vi)(b) in [16] will be more than enough. More generally, condition (B) together with condition (D$_1$) replace the just mentioned compactification condition in [14, 16, 17].

In the present paper, no explicit analogues of regularity conditions (i), (ii), (iii), (vii) of [16] are imposed.

So, quite predictably, neither our conditions imply those in [14, 16, 17], nor vice versa. However, our conditions appear to be a bit simpler and more explicit overall than those in [14, 16, 17]. It should also be mentioned that in [14, 16] both the relevant conditions and the corresponding results are stated uniformly over compact subsets of Θ. Of course, a similar modification of our conditions and results can be done.

References

[1] Anastasiou, A. and Ley, C. (2015). New simpler bounds to assess the asymptotic normality of the maximum likelihood estimator. http://arxiv.org/abs/1508.04948.

[2] Anastasiou, A. and Reinert, G. (2017). Bounds for the normal approximation of the maximum likelihood estimator. *Bernoulli* **23** 191–218. MR3556771

[3] Bentkus, V., Bloznelis, M. and Götze, F. (1996). A Berry-Esséen bound for Student's statistic in the non-i.i.d. case. *J. Theoret. Probab.* **9** 765–796. MR1400598 (97e:60036)

[4] Bentkus, V. and Götze, F. (1996). The Berry-Esseen bound for Student's statistic. *Ann. Probab.* **24** 491–503. MR1387647 (97f:62021)

[5] Borovkov, A. A. (1998). *Mathematical statistics*. Gordon and Breach Science Publishers, Amsterdam Translated from the Russian by A. Moullagaliev and revised by the author. MR1712750 (2000f:62003)

[6] Chen, L. H. Y. and Shao, Q.-M. (2007). Normal approximation for nonlinear statistics using a concentration inequality approach. *Bernoulli* **13** 581–599. MR2331265

[7] Ibragimov, I. A. and Radavichyus, M. È. (1981). On large deviation probabilities for maximum likelihood estimators. *Dokl. Akad. Nauk SSSR* **257** 1048–1052. MR614036

[8] Keilson, J. (1979). *Markov chain models—rarity and exponentiality*. *Applied Mathematical Sciences* **28**. Springer-Verlag, New York-Berlin. MR528293 (80f:60061)

[9] 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, 139–142. Academic Press, New York-London. MR0234539

[10] Koroljuk, V. S. and 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 (98e:60033)

[11] Kourouklis, S. (1984). A large deviation result for the likelihood ratio statistic in exponential families. *Ann. Statist.* **12** 1510–1521. MR760703

[12] Lehmann, E. L. and Casella, G. (1998). *Theory of point estimation*, second ed. *Springer Texts in Statistics*. Springer-Verlag, New York. MR1639875

[13] Miao, Y. (2010). Concentration inequality of maximum likelihood estimator. *Appl. Math. Lett.* **23** 1305–1309. MR2665616

[14] Michel, R. and Pfanzagl, J. (1971). The accuracy of the normal approximation for minimum contrast estimates. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete* **18** 73–84. MR0288897

[15] Mogul′skiĭ, A. A. (1988). Large deviations for the maximum likelihood estimators. In *Probability theory and mathematical statistics (Kyoto, 1986)*. *Lecture Notes in Math.* **1299** 326–331. Springer, Berlin. MR936005

[16] Pfanzagl, J. (1971). The Berry-Esseen bound for minimum contrast estimates. *Metrika* **17** 82–91. MR0295467 (45 ##4533)

[17] Pfanzagl, J. (1972/73). The accuracy of the normal approximation for estimates of vector parameters. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete* **25** 171–198. MR0329093 (48 ##7435)

[18] Pinelis, I. (2015). Exact Rosenthal-type bounds. *Ann. Probab.* **43** 2511–2544. MR3395468

[19] Pinelis, I. (2016). Optimal-order bounds on the rate of convergence to normality for maximum likelihood estimators. `http://arxiv.org/abs/1601.02177`.

[20] Pinelis, I. and Molzon, R. (2016). Optimal-order bounds on the rate of convergence to normality in the multivariate delta method. *Electron. J. Stat.* **10** 1001–1063. MR3486424

[21] Radavichyus, M. È. (1983). Probabilities of large deviations for maximum likelihood estimators. *Dokl. Akad. Nauk SSSR* **268** 551–556. MR691093

[22] Radavičjus, M. È. (1981). Probabilities of large and moderate deviations for maximum likelihood estimates. *Zap. Nauchn. Sem. Leningrad. Otdel. Mat. Inst. Steklov. (LOMI)* **108** 154–169, 196, 199. Studies in mathematical statistics, V. MR629406