Bernoulli **23**(1), 2017, 191–218
DOI: 10.3150/15-BEJ741

Bounds for the normal approximation of the maximum likelihood estimator

ANDREAS ANASTASIOU[1,*] and GESINE REINERT[1,**]

[1]*Department of Statistics, University of Oxford, 1 South Parks Road, Oxford, OX1 3TG, UK.*
E-mail: [*]*anastasi@stats.ox.ac.uk;* [**]*reinert@stats.ox.ac.uk*

While the asymptotic normality of the maximum likelihood estimator under regularity conditions is long established, this paper derives explicit bounds for the bounded Wasserstein distance between the distribution of the maximum likelihood estimator (MLE) and the normal distribution. For this task, we employ Stein's method. We focus on independent and identically distributed random variables, covering both discrete and continuous distributions as well as exponential and non-exponential families. In particular, a closed form expression of the MLE is not required. We also use a perturbation method to treat cases where the MLE has positive probability of being on the boundary of the parameter space.

Keywords: maximum likelihood estimator; normal approximation; Stein's method

1. Introduction

This paper assesses the bounded Wasserstein distance between the distribution of the maximum likelihood estimator (MLE) and the normal distribution. We concentrate on independent and identically distributed (i.i.d.) random variables, with the case that the random variables follow an exponential family distribution as an example. We also explain how a perturbation of both the parameter and the data can be useful in specific situations. The treatment includes situations where the MLE has positive probability to be on the boundary of the parameter space. The paper also covers cases where there is not an analytic form for the MLE.

Here is the notation which is used throughout the paper. First of all, θ denotes a scalar unknown parameter found in a parametric statistical model. Let θ_0 be the true (still unknown) value of the parameter θ and let $\Theta \subset \mathbb{R}$ denote the parameter space, while $\mathbf{X} = (X_1, X_2, \dots, X_n)$ is the random sample of n i.i.d. random variables with joint density function $f(\mathbf{x}|\theta)$. For $X_i = x_i$ being some observed values, the likelihood function is $L(\theta; \mathbf{x}) = f(\mathbf{x}|\theta)$. Its natural logarithm, called the log-likelihood function is denoted by $l(\theta; \mathbf{x})$. Having a fixed set of data and a defined statistical model, a maximum likelihood estimate is a value of the parameter which maximises the likelihood function. Derivatives of the log-likelihood function, with respect to θ, are denoted by $l'(\theta; \mathbf{x}), l''(\theta; \mathbf{x}), \dots, l^{(j)}(\theta; \mathbf{x})$, for j any integer greater than 2. For many models, the MLE exists and it is also unique, in which case it is denoted by $\hat{\theta}_n(\mathbf{X})$; this is known as the "regular" case. However, uniqueness or even existence of the MLE is not always secured. Unless otherwise specified, we make the following assumptions:

(i) The log-likelihood function $l(\theta; \mathbf{x})$ is a twice continuously differentiable function with respect to θ and the parameter varies in an open interval (a, b), where $a, b \in \mathbb{R} \cup \{-\infty, \infty\}$ and $a < b$.

1350-7265 © 2017 ISI/BS

(ii) $\lim_{\theta \to a,b} l(\theta; \mathbf{x}) = -\infty$,

(iii) $l''(\theta; \mathbf{x}) < 0$ at every point $\theta \in (a, b)$ for which $l'(\theta; \mathbf{x}) = 0$.

Under the assumptions (i)–(iii) above, the MLE exists and it is unique (Makelainen *et al.* [10]). Following now Casella and Berger [2], unless otherwise stated we also make the following assumptions:

(R1) the parameter is identifiable, which means that if $\theta \neq \theta'$, then $\exists x : f(x|\theta) \not\equiv f(x|\theta')$;

(R2) the density $f(x|\theta)$ is three times differentiable with respect to θ, the third derivative is continuous in θ and $\int f(x|\theta)\,\mathrm{d}x$ can be differentiated three times under the integral sign;

(R3) for any $\theta_0 \in \Theta$ and for $\mathbb{X}$ denoting the support of $f(x|\theta)$, there exists a positive number ε and a function $M(x)$ (both of which may depend on θ_0) such that

$$\left| \frac{\mathrm{d}^3}{\mathrm{d}\theta^3} \log f(x|\theta) \right| \leq M(x) \qquad \forall x \in \mathbb{X}, \theta_0 - \varepsilon < \theta < \theta_0 + \varepsilon,$$

with $\mathrm{E}_{\theta_0}[M(X)] < \infty$;

(R4) $i(\theta_0) \neq 0$, where $i(\theta)$ is the expected Fisher Information for one random variable.

The requirement (R2) that $\int f(x|\theta)\,\mathrm{d}x$ can be differentiated three times under the integral sign is usually substituted in the literature by the assumption that integration of $f(x|\theta)$ over x and differentiation with respect to θ are three times interchangeable, so that $\int_{\mathbb{R}} \frac{\mathrm{d}^j}{\mathrm{d}\theta^j} f(x|\theta)\,\mathrm{d}x = \frac{\mathrm{d}^j}{\mathrm{d}\theta^j} \int_{\mathbb{R}} f(x|\theta)\,\mathrm{d}x = 0$, $j \in \{1, 2, 3\}$. This condition ensures that if the expressions exist, then $\mathrm{E}_\theta[l'(\theta; \mathbf{X})] = 0$ and $\mathrm{Var}_\theta[l'(\theta; \mathbf{X})] = ni(\theta)$. In addition, it is obvious from (R3) that $\{\theta : |\theta - \theta_0| < \varepsilon\} \subset \Theta$ is required. The motivation of the work presented in this paper are the results given in Theorem 1.1. The efficiency and asymptotic normality of the MLE have first been discussed in Fisher [5]. Here we present the i.i.d. case; see Hoadley [7] for the case of independent but not identically distributed random variables.

Theorem 1.1 (Casella and Berger [2], page 472). *Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables with probability density* (*or mass*) *function $f(x_i|\theta)$, where θ is the scalar parameter. Assume that the MLE exists and it is unique and* (R1)–(R4) *are satisfied. Then for $Z \sim \mathrm{N}(0, 1)$,*

$$\text{(a)} \quad \frac{1}{\sqrt{n}} l'(\theta_0; \mathbf{X}) \xrightarrow[n\to\infty]{\mathrm{d}} \sqrt{i(\theta_0)}Z, \qquad \text{(b)} \quad \sqrt{ni(\theta_0)}\big(\hat{\theta}_n(\mathbf{X}) - \theta_0\big) \xrightarrow[n\to\infty]{\mathrm{d}} Z. \qquad (1.1)$$

Theorem 1.1 gives only a qualitative result as $n \to \infty$, but in approximations the sample size, n, is always finite and it is not clear when n is "large enough" for the limiting behaviour to be a good approximation to the finite-n behaviour. The rate of convergence may also depend on the true parameter θ_0. Hence, it is of interest to obtain explicit bounds for a distributional distance related to (a) and (b) in (1.1). These bounds are given in Proposition 2.2 and Theorem 2.1, respectively. The tools we use are mainly Taylor expansions, conditional expectations, a perturbation method and a result from Stein's method as given in Lemma 1.1. Bounds are also derived in Geyer [6], using the framework of locally asymptotically mixed normal (LAMN) models, but these bounds are of asymptotic nature.

As distance, we mainly use the bounded Wasserstein distance. If F, G are two random variables with values in $\mathbb{R}$ and H is a class of separating functions, then a Zolotarev-type distance between the laws of F and G, induced by H, is given by the quantity

$$d_H(F, G) = \sup\bigl\{\bigl|\mathrm{E}\bigl[h(F)\bigr] - \mathrm{E}\bigl[h(G)\bigr]\bigr| : h \in H\bigr\}. \tag{1.2}$$

From now on, $\|\cdot\|$ denotes the supremum norm ($\|\cdot\|_\infty$) and

$$H = \bigl\{h : \mathbb{R} \to \mathbb{R} : \|h\|_{\mathrm{Lip}} + \|h\| \leq 1\bigr\}, \tag{1.3}$$

where

$$\|h\|_{\mathrm{Lip}} = \sup_{\substack{x,y\in\mathbb{R}\\ x\neq y}} \frac{|h(x) - h(y)|}{|x - y|}.$$

Using Rademacher's theorem, since $\|h\|_{\mathrm{Lip}} \leq 1$, then h is differentiable almost everywhere, with h' denoting its derivative.

Using this class of test functions, (1.2) gives the bounded Wasserstein (or Fortet–Mourier) distance between two random variables F and G, denoted from now on by

$$d_{bW}(F, G) = \sup\bigl\{\bigl|\mathrm{E}\bigl[h(F)\bigr] - \mathrm{E}\bigl[h(G)\bigr]\bigr| : h \in H\bigr\}, \tag{1.4}$$

with H as in (1.3); see, for example, Nourdin and Peccati [11]. Rachev [12] also gives a connection to the Kantorovich–Rubinstein problem. To obtain such bounds, we use the following lemma from Reinert [13] which is based on Stein's method (Stein [14]).

Lemma 1.1. *Let $Y_1, Y_2, \ldots, Y_n$ be independent random variables with* $\mathrm{E}(Y_i) = 0$, $\mathrm{Var}(Y_i) = \sigma^2 > 0$ *and* $\mathrm{E}|Y_i|^3 < \infty$. *Let* $W = \frac{1}{\sqrt{n}}\sum_{i=1}^n Y_i$ *and* $K \sim \mathrm{N}(0, \sigma^2)$. *Then for any function $h \in H$, with H given in* (1.3)

$$d_{bW}(W, K) \leq \frac{1}{\sqrt{n}}\left(2 + \frac{1}{\sigma^3}\bigl[\mathrm{E}|Y_1|^3\bigr]\right). \tag{1.5}$$

Using $Y_i = l'(\theta_0; X_i)$, we see that (1.5) is closely related to (a) in (1.1). For a bound of (b), we employ Taylor expansion.

The paper is organised as follows. Section 2 gives an upper bound on the distributional distance between the distribution of the MLE and the normal distribution in the case of i.i.d. random variables. In Section 3, the results are applied to the class of one-parameter exponential family distributions. In Section 4, we use a perturbation to treat the special case of having a random vector from a distribution where the parameter space is not an open interval and there is positive probability of the MLE to lie on the boundary of the parameter space. An example is the Poisson distribution with mean $\theta \in [0, \infty)$; the MLE could take on the value zero with positive probability, but the log-likelihood function is not differentiable at zero. In Section 5, we obtain an upper bound on the Mean Squared Error of the MLE. We use this bound in order to get an upper bound on the distributional distance to the normal distribution, even when no analytic expression of the MLE is available. We assess the quality of our results through a simulation-based study related to the Beta distribution. The R-code for the simulations and the simulation output are available at the Oxford University Research Archive (ORA). The DOI is: 10.5287/bodleian:s4655h876.

2. Bounds on the distance to normal for the MLE

In this section, we briefly relate the Kolmogorov and the bounded Wasserstein distance and we give upper bounds on the distributional distance between the distribution of the MLE and the normal distribution in terms of the bounded Wasserstein distance.

2.1. The bounded Wasserstein and the Kolmogorov distance

For $Z \sim \mathrm{N}(0,1)$, the aim is to bound

$$d_{bW}\big(\sqrt{ni(\theta_0)}\big(\hat{\theta}_n(\mathbf{X}) - \theta_0\big), Z\big), \tag{2.1}$$

with $d_{bW}(\cdot,\cdot)$ as defined in (1.4). Using $H = \{\mathbb{1}_{[\cdot \le x]}, x \in \mathbb{R}\}$ as the class of functions in (1.2), yields the Kolmogorov distance,

$$d_K\big(\sqrt{ni(\theta_0)}\big(\hat{\theta}_n(\mathbf{X}) - \theta_0\big), Z\big).$$

The next proposition links these two distances.

Proposition 2.1. *If G is any real-valued random variable and $Z \sim \mathrm{N}(0,1)$, then*

$$d_K(G,Z) \le 2\sqrt{d_{bW}(G,Z)}.$$

Proof. The proof of this proposition follows the proof of Theorem 3.3 of Chen *et al.* [3], page 48. Let $z \in \mathbb{R}$ and for $\alpha = \sqrt{d_{bW}(G,Z)}(2\pi)^{1/4}$, $z \in \mathbb{R}$, let

$$h_\alpha(w) = \begin{cases} 1, & \text{if } w \le z, \\ 1 + \dfrac{z-w}{\alpha}, & \text{if } z < w \le z+\alpha, \\ 0, & \text{if } w > z+\alpha \end{cases}$$

so that h_α is bounded Lipschitz with $\|h_\alpha\| \le 1$ and $\|h'_\alpha\| \le \frac{1}{\alpha}$. By the triangle inequality,

$$\begin{aligned} \mathbb{P}(G \le z) - \mathbb{P}(Z \le z) &\le \mathrm{E}\big[h_\alpha(G)\big] - \mathrm{E}\big[h_\alpha(Z)\big] + \mathrm{E}\big[h_\alpha(Z)\big] - \mathbb{P}(Z \le z) \\ &\le \frac{d_{bW}(G,Z)}{\alpha} + \mathbb{P}(z \le Z \le z+\alpha) \\ &\le \frac{d_{bW}(G,Z)}{\alpha} + \frac{\alpha}{\sqrt{2\pi}} \le 2\sqrt{d_{bW}(G,Z)}. \end{aligned}$$

Similarly $\mathbb{P}(G \le z) - \mathbb{P}(Z \le z) \ge -2\sqrt{d_{bW}(G,Z)}$, which completes the proof. □

The Kolmogorov distance relates directly to exact conservative confidence intervals. Our results on the bounded Wasserstein distance and Proposition 2.1 give that

$$d_K\big(\sqrt{ni(\theta_0)}\big(\hat{\theta}_n(\mathbf{X})-\theta_0\big), Z\big) \leq 2\sqrt{B_{bW}} =: B_K,$$

where B_{bW} denotes the bound for the bounded Wasserstein distance from Proposition 2.1. Therefore, for $y \in \mathbb{R}$:

$$\begin{aligned} &\big|\mathbb{P}\big(\sqrt{ni(\theta_0)}\big(\hat{\theta}_n(\mathbf{X})-\theta_0\big) \leq y\big) - \mathbb{P}(Z \leq y)\big| \leq B_K \\ &\quad\Leftrightarrow\quad -B_K \leq \mathbb{P}\big(\sqrt{ni(\theta_0)}\big(\hat{\theta}_n(\mathbf{X})-\theta_0\big) \leq y\big) - \mathbb{P}(Z \leq y) \leq B_K. \end{aligned} \tag{2.2}$$

For $\Phi^{-1}(\cdot)$ the quantile function for the standard normal distribution, applying (2.2) to $y = \Phi^{-1}(\frac{\alpha}{2} - B_K)$ and to $y = \Phi^{-1}(1 - \frac{\alpha}{2} + B_K)$ yields

$$\mathbb{P}\left(\Phi^{-1}\left(\frac{\alpha}{2} - B_K\right) \leq \sqrt{ni(\theta_0)}\big(\hat{\theta}_n(\mathbf{X})-\theta_0\big) \leq \Phi^{-1}\left(1 - \frac{\alpha}{2} + B_K\right)\right) \geq 1 - \alpha.$$

Hence, if the expected Fisher Information number for one random variable, $i(\theta_0)$, is known, then

$$\left(\hat{\theta}_n(\mathbf{X}) - \frac{\Phi^{-1}(1-\alpha/2+B_K)}{\sqrt{ni(\theta_0)}}, \hat{\theta}_n(\mathbf{X}) - \frac{\Phi^{-1}(\alpha/2-B_K)}{\sqrt{ni(\theta_0)}}\right)$$

is a conservative $100(1-\alpha)\%$ confidence interval for θ_0.

2.2. Bounds in terms of the bounded Wasserstein distance

The bounded Wasserstein distance links in well with Stein's method because the Lipschitz test functions are differentiable almost everywhere. From now on, $\frac{\mathrm{d}}{\mathrm{d}\theta}\log f(X_1|\theta_0) := \frac{\mathrm{d}}{\mathrm{d}\theta}\log f(X_1|\theta)|_{\theta=\theta_0}$. The next two results provide a bound for (a) and (b) in (1.1), respectively.

Proposition 2.2. *Suppose $X_1, X_2, \dots, X_n$ are i.i.d. random variables with density or frequency function $f(x_i|\theta)$. Assume that* (R1)–(R4) *are satisfied, $Z \sim \mathrm{N}(0,1)$ and $\mathrm{E}|\frac{\mathrm{d}}{\mathrm{d}\theta}\log f(X_1|\theta_0)|^3$ exists. Then for $h: \mathbb{R} \to \mathbb{R}$, such that h is absolutely continuous and bounded*

$$\left|\mathrm{E}\left[h\left(\frac{l'(\theta_0;\mathbf{X})}{\sqrt{ni(\theta_0)}}\right)\right] - \mathrm{E}\big[h(Z)\big]\right| \leq \frac{\|h'\|}{\sqrt{n}}\left(2 + \frac{1}{[i(\theta_0)]^{3/2}}\left[\mathrm{E}\left|\frac{\mathrm{d}}{\mathrm{d}\theta}\log f(X_1|\theta_0)\right|^3\right]\right). \tag{2.3}$$

In particular,

$$d_{bW}\left(\frac{l'(\theta_0;\mathbf{X})}{\sqrt{ni(\theta_0)}}, Z\right) \leq \frac{1}{\sqrt{n}}\left(2 + \frac{1}{[i(\theta_0)]^{3/2}}\left[\mathrm{E}\left|\frac{\mathrm{d}}{\mathrm{d}\theta}\log f(X_1|\theta_0)\right|^3\right]\right). \tag{2.4}$$

Proof. Let

$$Y_i = Y_i(X_i;\theta_0) = \left(\frac{\mathrm{d}}{\mathrm{d}\theta}\log f(X_i|\theta_0)\right)\Big/\sqrt{i(\theta_0)}, \qquad i = 1, 2, \dots, n,$$

which are i.i.d. random variables as $X_1, X_2, \ldots, X_n$ are i.i.d. The regularity conditions (R1)–(R4) ensure that $\mathrm{E}_{\theta_0}[Y_i] = 0$ and $\mathrm{Var}_{\theta_0}[Y_i] = 1$. Then letting $W = W(\mathbf{X};\theta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^n Y_i = \frac{l'(\theta_0;\mathbf{X})}{\sqrt{ni(\theta_0)}}$, gives that $\mathrm{E}_{\theta_0}[W] = 0$ and $\mathrm{Var}_{\theta_0}[W] = 1$. Applying Lemma 1.1 to $K = Z \sim \mathrm{N}(0,1)$ yields the result. □

Theorem 2.1. *Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables with density or frequency function $f(x_i|\theta)$ such that the regularity conditions* (R1)–(R4) *are satisfied and that the MLE, $\hat{\theta}_n(\mathbf{X})$, exists and it is unique. Assume that $\mathrm{E}|\frac{\mathrm{d}}{\mathrm{d}\theta}\log f(X_1|\theta_0)|^3 < \infty$ and that $\mathrm{E}(\hat{\theta}_n(\mathbf{X}) - \theta_0)^4 < \infty$. Let $0 < \varepsilon = \varepsilon(\theta_0)$ be such that $(\theta_0 - \varepsilon, \theta_0 + \varepsilon) \subset \Theta$ as in* (R3) *and let $Z \sim \mathrm{N}(0,1)$. Then*

$$\begin{aligned} & d_{bW}\left(\sqrt{ni(\theta_0)}\left(\hat{\theta}_n(\mathbf{X}) - \theta_0\right), Z\right) \\ & \quad \le \frac{1}{\sqrt{n}}\left(2 + \frac{1}{[i(\theta_0)]^{3/2}}\left[\mathrm{E}\left|\frac{\mathrm{d}}{\mathrm{d}\theta}\log f(X_1|\theta_0)\right|^3\right]\right) \\ & \qquad + 2\frac{\mathrm{E}(\hat{\theta}_n(\mathbf{X}) - \theta_0)^2}{\varepsilon^2} + \frac{1}{\sqrt{ni(\theta_0)}}\Bigg\{\mathrm{E}\left(\left|R_2(\theta_0;\mathbf{X})\right|\big|\left|\hat{\theta}_n(\mathbf{X}) - \theta_0\right| \le \varepsilon\right) \\ & \qquad + \frac{1}{2}\left[\mathrm{E}\left(\left(\sup_{\theta:|\theta-\theta_0|\le\varepsilon}\left|l^{(3)}(\theta;\mathbf{X})\right|\right)^2\big|\left|\hat{\theta}_n(\mathbf{X}) - \theta_0\right| \le \varepsilon\right)\right]^{1/2}\left[\mathrm{E}\left(\hat{\theta}_n(\mathbf{X}) - \theta_0\right)^4\right]^{1/2}\Bigg\}, \end{aligned} \tag{2.5}$$

where

$$R_2(\theta_0, \mathbf{x}) = \left(\hat{\theta}_n(\mathbf{x}) - \theta_0\right)\left(l''(\theta_0;\mathbf{x}) + ni(\theta_0)\right). \tag{2.6}$$

The following lemma is useful for the conditional expectations in (2.5); the proof is in the Appendix.

Lemma 2.1. *Let $M \ge 0$ be a random variable and $\varepsilon > 0$. For every continuous function f such that $f(m)$ is increasing and $f(m) \ge 0$, for $m > 0$,*

$$\mathrm{E}\left[f(M)|M \le \varepsilon\right] \le \mathrm{E}\left[f(M)\right].$$

Proof of Theorem 2.1. For the sake of presentation, we drop the subscript θ_0 from the expectation. The regularity conditions ensure that $0 = l'(\hat{\theta}_n(\mathbf{x});\mathbf{x})$. A second order Taylor expansion of $l'(\hat{\theta}_n(\mathbf{x});\mathbf{x})$ about θ_0 gives

$$l''(\theta_0;\mathbf{x})\left(\hat{\theta}_n(\mathbf{x}) - \theta_0\right) = -l'(\theta_0;\mathbf{x}) - R_1(\theta_0;\mathbf{x}), \tag{2.7}$$

where

$$R_1(\theta_0;x) = \tfrac{1}{2}\left(\hat{\theta}_n(\mathbf{x}) - \theta_0\right)^2 l^{(3)}\left(\theta^*;\mathbf{x}\right)$$

is the remainder term with θ^* lying between $\hat{\theta}_n(\mathbf{x})$ and θ_0. The result in (2.7) gives

$$-ni(\theta_0)\left(\hat{\theta}_n(\mathbf{x}) - \theta_0\right) = -l'(\theta_0;\mathbf{x}) - R_1(\theta_0;\mathbf{x}) - \left(\hat{\theta}_n(\mathbf{x}) - \theta_0\right)\left[l''(\theta_0;\mathbf{x}) + ni(\theta_0)\right].$$

As $i(\theta_0) \neq 0$

$$\hat{\theta}_n(\mathbf{x}) - \theta_0 = \frac{l'(\theta_0;\mathbf{x}) + R_1(\theta_0;\mathbf{x}) + R_2(\theta_0,\mathbf{x})}{ni(\theta_0)},$$

with $R_2(\theta_0,\mathbf{x})$ as in (2.6). For $Z \sim \mathrm{N}(0,1)$ and $h \in H$ given in (1.3), we obtain

$$\begin{aligned}
&\left|\mathrm{E}\left[h\left(\left(\hat{\theta}_n(\mathbf{X}) - \theta_0\right)\sqrt{ni(\theta_0)}\right)\right] - \mathrm{E}\left[h(Z)\right]\right| \\
&\quad \le \left|\mathrm{E}\left[h\left(\frac{l'(\theta_0;\mathbf{X}) + R_1(\theta_0;\mathbf{X}) + R_2(\theta_0;\mathbf{X})}{\sqrt{ni(\theta_0)}}\right) - h\left(\frac{l'(\theta_0;\mathbf{X})}{\sqrt{ni(\theta_0)}}\right)\right]\right| \qquad (2.8)\\
&\qquad + \left|\mathrm{E}\left[h\left(\frac{l'(\theta_0;\mathbf{X})}{\sqrt{ni(\theta_0)}}\right)\right] - \mathrm{E}\left[h(Z)\right]\right|. \qquad (2.9)
\end{aligned}$$

The upper bound for (2.9) is given in Proposition 2.2. To bound (2.8), note that the term $R_1(\theta_0;\mathbf{X})$ is in general not uniformly bounded. For ease of presentation, let

$$C_1 = C_1(h,\theta_0;\mathbf{X}) = h\left(\frac{l'(\theta_0;\mathbf{X}) + R_1(\theta_0;\mathbf{X}) + R_2(\theta_0;\mathbf{X})}{\sqrt{ni(\theta_0)}}\right) - h\left(\frac{l'(\theta_0;\mathbf{X})}{\sqrt{ni(\theta_0)}}\right).$$

For all x the rather crude bound $|C_1| \le 2\|h\|$ is valid. If $|\hat{\theta}_n(\mathbf{X}) - \theta_0| \le \varepsilon$ then a better bound is available. Hence, we condition on whether $|\hat{\theta}_n(\mathbf{X}) - \theta_0| > \varepsilon$ or $|\hat{\theta}_n(\mathbf{X}) - \theta_0| \le \varepsilon$, with $\varepsilon > 0$ such that $(\theta_0 - \varepsilon, \theta_0 + \varepsilon) \subset \Theta$, as condition (R3) requires. Moreover, by Markov's inequality

$$\mathbb{P}_{\theta_0}\left(\left|\hat{\theta}_n(\mathbf{X}) - \theta_0\right| > \varepsilon\right) \le \frac{\mathrm{E}[\hat{\theta}_n(\mathbf{X}) - \theta_0]^2}{\varepsilon^2}. \qquad (2.10)$$

Using the law of total expectation,

$$\begin{aligned}
\left|\mathrm{E}[C_1]\right| \le{}& \mathrm{E}\left(|C_1|\,\middle|\,\left|\hat{\theta}_n(\mathbf{X}) - \theta_0\right| > \varepsilon\right)\mathbb{P}\left(\left|\hat{\theta}_n(\mathbf{X}) - \theta_0\right| > \varepsilon\right) \\
&+ \mathrm{E}\left(|C_1|\,\middle|\,\left|\hat{\theta}_n(\mathbf{X}) - \theta_0\right| \le \varepsilon\right)\mathbb{P}\left(\left|\hat{\theta}_n(\mathbf{X}) - \theta_0\right| \le \varepsilon\right).
\end{aligned}$$

Using (2.10) for the first term and a first order Taylor expansion of $h(\frac{l'(\theta_0;\mathbf{X})+R_1(\theta_0;\mathbf{X})+R_2(\theta_0;\mathbf{X})}{\sqrt{ni(\theta_0)}})$ about $\frac{l'(\theta_0)}{\sqrt{ni(\theta_0)}}$ for the second term gives

$$\begin{aligned}
\left|\mathrm{E}[C_1]\right| \le{}& 2\|h\|\frac{\mathrm{E}(\hat{\theta}_n(\mathbf{X}) - \theta_0)^2}{\varepsilon^2} \\
&+ \left|\mathrm{E}\left(\frac{R_1(\theta_0,\mathbf{X}) + R_2(\theta_0,\mathbf{X})}{\sqrt{ni(\theta_0)}}h'\left(t(\mathbf{X})\right)\middle|\left|\hat{\theta}_n(\mathbf{X}) - \theta_0\right| \le \varepsilon\right)\right| \\
\le{}& 2\|h\|\frac{\mathrm{E}(\hat{\theta}_n(\mathbf{X}) - \theta_0)^2}{\varepsilon^2} + \frac{\|h'\|}{\sqrt{ni(\theta_0)}}\mathrm{E}\left(\left|R_2(\theta_0;\mathbf{X})\right|\,\middle|\,\left|\hat{\theta}_n(\mathbf{X}) - \theta_0\right| \le \varepsilon\right) \\
&+ \frac{\|h'\|}{\sqrt{ni(\theta_0)}}\mathrm{E}\left(\frac{1}{2}\left(\hat{\theta}_n(\mathbf{X}) - \theta_0\right)^2\left|l^{(3)}(\theta^*;\mathbf{X})\right|\,\middle|\,\left|\hat{\theta}_n(\mathbf{X}) - \theta_0\right| \le \varepsilon\right),
\end{aligned}$$

where $t(\mathbf{X})$ lies between $\frac{l'(\theta_0;\mathbf{X})}{\sqrt{ni(\theta_0)}}$ and $\frac{l'(\theta_0;\mathbf{X})+R_1(\theta_0;\mathbf{x})+R_2(\theta_0;\mathbf{X})}{\sqrt{ni(\theta_0)}}$. Since for $|\hat{\theta}_n(\mathbf{X})-\theta_0|\leq\varepsilon$, $|R_1(\theta_0;\mathbf{x})|\leq\frac{1}{2}(\hat{\theta}_n(\mathbf{X})-\theta_0)^2\sup_{\theta:|\theta-\theta_0|\leq\varepsilon}|l^{(3)}(\theta;\mathbf{X})|$,

$$\begin{aligned}\bigl|\mathrm{E}[C_1]\bigr| &\leq 2\|h\|\frac{\mathrm{E}(\hat{\theta}_n(\mathbf{X})-\theta_0)^2}{\varepsilon^2}+\frac{\|h'\|}{\sqrt{ni(\theta_0)}}\mathrm{E}\bigl(\bigl|R_2(\theta_0;\mathbf{X})\bigr|\big||\hat{\theta}_n(\mathbf{X})-\theta_0\bigr|\leq\varepsilon\bigr)\\ &\quad+\frac{\|h'\|}{2\sqrt{ni(\theta_0)}}\mathrm{E}\Bigl[\sup_{\theta:|\theta-\theta_0|\leq\varepsilon}\bigl|l^{(3)}(\theta;\mathbf{X})\bigr|\bigl(\hat{\theta}_n(\mathbf{X})-\theta_0\bigr)^2\big|\bigl|\hat{\theta}_n(\mathbf{X})-\theta_0\bigr|\leq\varepsilon\Bigr].\end{aligned}$$

The next step is based on the Cauchy–Schwarz inequality and the fact that

$$\mathrm{E}\bigl[\bigl(\hat{\theta}_n(\mathbf{X})-\theta_0\bigr)^4\big|\bigl|\hat{\theta}_n(\mathbf{X})-\theta_0\bigr|\leq\varepsilon\bigr]\leq\mathrm{E}\bigl[\bigl(\hat{\theta}_n(\mathbf{X})-\theta_0\bigr)^4\bigr], \tag{2.11}$$

due to Lemma 2.1, giving

$$\begin{aligned}\bigl|\mathrm{E}[C_1]\bigr| &\leq 2\|h\|\frac{\mathrm{E}(\hat{\theta}_n(\mathbf{X})-\theta_0)^2}{\varepsilon^2}+\frac{\|h'\|}{\sqrt{ni(\theta_0)}}\Bigl\{\mathrm{E}\bigl(\bigl|R_2(\theta_0;\mathbf{X})\bigr|\big|\bigl|\hat{\theta}_n(\mathbf{X})-\theta_0\bigr|\leq\varepsilon\bigr)\\ &\quad+\frac{1}{2}\Bigl[\mathrm{E}\Bigl(\Bigl(\sup_{\theta:|\theta-\theta_0|\leq\varepsilon}\bigl|l^{(3)}(\theta;\mathbf{X})\bigr|\Bigr)^2\big|\bigl|\hat{\theta}_n(\mathbf{X})-\theta_0\bigr|\leq\varepsilon\Bigr)\Bigr]^{1/2}\\ &\quad\times\bigl[\mathrm{E}\bigl(\hat{\theta}_n(\mathbf{X})-\theta_0\bigr)^4\bigr]^{1/2}\Bigr\}.\end{aligned} \tag{2.12}$$

The result of the theorem is obtained using (2.4) and (2.12) and the fact that $\|h\|\leq 1$ and $\|h'\|\leq 1$. □

Remark 2.1. (1) If $l''(\theta_0;\mathbf{x})\equiv -ni(\theta_0)$ then in (2.6), $R_2(\theta_0;\mathbf{x})\equiv 0$ and the bound given in Theorem 2.1 simplifies.

(2) The rate of convergence of the Mean Squared Error, $\mathrm{E}(\hat{\theta}_n(\mathbf{X})-\theta_0)^2$, is $\mathcal{O}(\frac{1}{n})$. This result is obtained using that

$$\mathrm{E}\bigl(\hat{\theta}_n(\mathbf{X})-\theta_0\bigr)^2=\mathrm{Var}\bigl[\hat{\theta}_n(\mathbf{X})\bigr]+\mathrm{bias}^2\bigl[\hat{\theta}_n(\mathbf{X})\bigr]. \tag{2.13}$$

Under the standard asymptotics (from the regularity conditions (R1)–(R4)) the MLE is asymptotically efficient,

$$n\,\mathrm{Var}\bigl[\hat{\theta}_n(\mathbf{X})\bigr]\xrightarrow[n\to\infty]{}\bigl[i(\theta_0)\bigr]^{-1},$$

and hence the variance of the MLE is of order $\frac{1}{n}$. In addition, from Theorem 1.1 the bias of the MLE is of order $\frac{1}{\sqrt{n}}$; see also Cox and Snell [4], where no explicit conditions are given. Combining these two results and using (2.13) shows that the Mean Squared Error of the MLE is of order $\frac{1}{n}$. In the examples that follow, the remaining terms in the bound are of order at most $\frac{1}{\sqrt{n}}$.

(3) When the calculation of $\mathrm{E}(|\frac{\mathrm{d}}{\mathrm{d}\theta}\log f(X_1|\theta_0)|^3)$ is awkward, Hölder's inequality can be used, giving $\mathrm{E}(|\frac{\mathrm{d}}{\mathrm{d}\theta}\log f(X_1|\theta_0)|^3) \leq [\mathrm{E}(\frac{\mathrm{d}}{\mathrm{d}\theta}\log f(X_1|\theta_0))^4]^{3/4}$.

3. One-parameter exponential families

This section specifies Theorem 2.1 for the distribution of the MLE for one-parameter exponential family distributions. Many popular distributions which have the same underlying structure based on relatively simple properties are exponential families, such as the normal, Gamma and Laplace distributions. The case of the Poisson distribution with $\theta \in [0, \infty)$ is treated in Section 4.2. Generalisations of exponential families can be found in Lauritzen [9] and Berk [1]. The density or frequency function is of the form

$$f(x|\theta) = \exp\bigl\{k(\theta)T(x) - A(\theta) + S(x)\bigr\}\mathbb{1}_{\{x\in B\}},$$

where the set $B = \{x : f(x|\theta) > 0\}$ is the support of X and does not depend on θ; $k(\theta)$ and $A(\theta)$ are functions of the parameter; $T(x)$ and $S(x)$ are functions only of the data. The choice of the functions $k(\theta)$ and $T(X)$ is not unique. The case $k(\theta) = \theta$ is the so-called *canonical case*. In this case, θ and $T(X)$ are called the *natural parameter* and *natural observation* (Casella and Berger [2]). We make the following assumptions, where (Ass.Ex.1)–(Ass.Ex.3) are necessary for the existence and uniqueness of the MLE and (A1)–(A4) follow from the regularity conditions in Section 1.

(Ass.Ex.1) $\Theta \subset \mathbb{R}$ is open and connected;
(Ass.Ex.2) $\lim_{\theta\to\partial\Theta} k(\theta)\sum_{i=1}^n T(x_i) - nA(\theta) + \sum_{i=1}^n S(x_i) = -\infty$;
(Ass.Ex.3) We have $k''(\theta)\sum_{i=1}^n T(x_i) - nA''(\theta) < 0$ at every point $\theta \in \Theta$ for which it holds that $k'(\theta)\sum_{i=1}^n T(x_i) - nA'(\theta) = 0$;
(A1) $k'(\theta) \neq 0, \forall\theta \in \Theta$ and $D(\theta) = \frac{A'(\theta)}{k'(\theta)}$ is invertible;
(A2) $l(\theta; x)$ is thrice continuously differentiable with respect to θ, meaning that both $k^{(3)}(\theta)$ and $A^{(3)}(\theta)$ exist and they are continuous. In addition, integration of the density function over x and differentiation with respect to θ are three times interchangeable;
(A3) for any $\theta_0 \in \Theta$, there exists a positive number ε and a function $M(x)$ (both of which may depend on θ_0) such that

$$\bigl|k^{(3)}(\theta)T(x) - A^{(3)}(\theta)\bigr| \leq M(x) \qquad \forall x \in B, \theta_0 - \varepsilon < \theta < \theta_0 + \varepsilon,$$

with $\mathrm{E}[M(X)] < \infty$;
(A4) $\mathrm{Var}[T(X)] > 0$;
(A5) $\mathrm{E}|T(X) - D(\theta_0)|^3$ exists. This assumption is required for meaningful bounds.

Corollary 3.1. *Let $X_1, X_2, \dots, X_n$ be i.i.d. random variables with the density or frequency function of a single-parameter exponential family. Assume that* (A1)–(A5) *are satisfied and that* (Ass.Ex.1)–(Ass.Ex.3) *also hold. With $Z \sim \mathrm{N}(0, 1)$, $h \in H$, $R_2(\theta_0; \mathbf{X})$ as in* (2.6) *and also*

$0 < \varepsilon = \varepsilon(\theta_0)$ *such that* $(\theta_0 - \varepsilon, \theta_0 + \varepsilon) \subset \Theta$ *as in* (A3), *it holds that*

$$\begin{aligned} &d_{bW}\big(\sqrt{ni(\theta_0)}\big(\hat{\theta}_n(\mathbf{X}) - \theta_0\big), Z\big) \\ &\quad \leq \frac{1}{\sqrt{n}}\left(2 + \frac{\mathrm{E}|T(X_1) - D(\theta_0)|^3}{[\mathrm{Var}[T(X_1)]]^{3/2}}\right) \\ &\qquad + 2\frac{\mathrm{E}(\hat{\theta}_n(\mathbf{X}) - \theta_0)^2}{\varepsilon^2} + \frac{1}{\sqrt{ni(\theta_0)}}\Big\{\mathrm{E}\big(\big|R_2(\theta_0;\mathbf{X})\big|\big|\big|\hat{\theta}_n(\mathbf{X}) - \theta_0\big| \leq \varepsilon\big) \\ &\qquad + \frac{1}{2}\Big[\mathrm{E}\Big(\Big(\sup_{\theta:|\theta-\theta_0|\leq\varepsilon}\big|l^{(3)}(\theta;\mathbf{X})\big|\Big)^2\big|\big|\hat{\theta}_n(\mathbf{X}) - \theta_0\big| \leq \varepsilon\Big)\Big]^{1/2}\big[\mathrm{E}\big(\hat{\theta}_n(\mathbf{X}) - \theta_0\big)^4\big]^{1/2}\Big\}. \end{aligned} \tag{3.1}$$

Proof. For the first term of the bound, let

$$Y_i = Y_i(X_i;\theta_0) = \left(\frac{\mathrm{d}}{\mathrm{d}\theta}\log f(X_i|\theta_0)\right)\Big/\sqrt{i(\theta_0)}, \qquad i = 1, 2, \dots, n.$$

Using Proposition 2.2, we calculate $\mathrm{E}|Y_1|^3$. Now

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\log f(X_i|\theta)\Big|_{\theta=\theta_0} = k'(\theta_0)T(X_i) - A'(\theta_0)$$

yields

$$\begin{aligned} \mathrm{E}\left|\frac{\mathrm{d}}{\mathrm{d}\theta}\log f(X_i|\theta_0)\right|^3 &= \mathrm{E}\big|k'(\theta_0)T(X_i) - A'(\theta_0)\big|^3 = \mathrm{E}\big|k'(\theta_0)\big(T(X_i) - D(\theta_0)\big)\big|^3 \\ &= \big|k'(\theta_0)\big|^3\mathrm{E}\big|T(X_i) - D(\theta_0)\big|^3 \qquad \forall i \in \{1, 2, \dots, n\}. \end{aligned}$$

In addition, $i(\theta_0) = \mathrm{Var}[\frac{\mathrm{d}}{\mathrm{d}\theta}\log f(X_i|\theta_0)] = [k'(\theta_0)]^2\,\mathrm{Var}[T(X_i)] > 0$ from (A1) and (A4). These quantities can now be applied to get the first term of the bound in (3.1) while the rest of the terms are as in Theorem 2.1. □

Remark 3.1. In the canonical case, $ni(\theta_0) \equiv nA''(\theta_0) \equiv -l''(\theta_0;\mathbf{x})$. So $R_2(\theta_0;\mathbf{x}) \equiv 0$.

3.1. Example: The exponentially distributed random variable

In this section, we consider two examples using the exponential distribution, first, its canonical form, and then under a change of parameterisation.

3.1.1. *The canonical case*

In the case of $X_1, X_2, \dots, X_n$ exponentially distributed, $\mathrm{Exp}(\theta)$, i.i.d. random variables where $\theta > 0$ the probability density function is

$$f(x|\theta) = \theta\exp\{-\theta x\} = \exp\{\log\theta - \theta x\} = \exp\big\{k(\theta)T(x) - A(\theta) + S(x)\big\}\mathbb{1}_{\{x\in B\}},$$

where $B=(0,\infty)$, $\theta \in \Theta=(0,\infty)$, $T(x)=-x$, $k(\theta)=\theta$, $A(\theta)=-\log\theta$ and $S(x)=0$. Hence, $\mathrm{Exp}(\theta)$ is a single-parameter canonical exponential family. Moreover,

$$l'(\theta;\mathbf{x})=\frac{n}{\theta}-\sum_{i=1}^{n}x_i, \qquad l''(\theta;\mathbf{x})=-\frac{n}{\theta^2}.$$

Thus, it is easy to see that the MLE exists, it is unique, equal to $\hat{\theta}_n(\mathbf{X})=\frac{1}{\bar{X}}$ and (A1)–(A5) are satisfied. Corollary 3.1 gives

$$\begin{aligned} d_{bW}\bigl(\sqrt{ni(\theta_0)}\bigl(\hat{\theta}_n(\mathbf{X})-\theta_0\bigr),Z\bigr) &\le \frac{4.41456}{\sqrt{n}}+\frac{8(n+2)}{(n-1)(n-2)} \\ &\quad+\frac{8\sqrt{n}(n+2)}{(n-1)(n-2)}. \end{aligned} \tag{3.2}$$

For $\varepsilon>0$, since $\Theta=(0,\infty)$ simple calculations yield that $0<\varepsilon<\theta_0$ to apply (A3) and moreover $\sup_{\theta:|\theta-\theta_0|\le\varepsilon}|l^{(3)}(\theta;\mathbf{x})|=\frac{2n}{(\theta_0-\varepsilon)^3}$. Choosing $\varepsilon=\frac{\theta_0}{2}$, gives that $\sup_{\theta:|\theta-\theta_0|\le\varepsilon}|l^{(3)}(\theta;\mathbf{x})|=\frac{16n}{\theta_0^3}$. In addition, since $X_i\sim\mathrm{Exp}(\theta)$, $\forall i\in\{1,2,\dots,n\}$ then $\bar{X}\sim\mathrm{G}(n,n\theta)$, with $G(\alpha,\beta)$ being the Gamma distribution with shape parameter α and rate parameter β. Basic calculations of integrals show that $\mathrm{E}|T(X)-D(\theta_0)|^3=\mathrm{E}|\frac{1}{\theta_0}-X|^3\le\frac{2.41456}{\theta_0^3}$ and

$$\mathrm{E}\bigl[\bigl(\hat{\theta}_n(\mathbf{X})-\theta_0\bigr)^2\bigr]=\frac{(n\theta_0)^2}{(n-1)(n-2)}-\frac{2n\theta_0^2}{n-1}+\theta_0^2=\frac{(n+2)\theta_0^2}{(n-1)(n-2)}.$$

Since $\sup_{\theta:|\theta-\theta_0|\le\varepsilon}|l^{(3)}(\theta)|$ does not depend on the sample, it is not necessary to use (2.11). Thus, $\varepsilon=\frac{\theta_0}{2}$ yields the result in (3.2).

Remark 3.2. (1) The rate of convergence of the bound is $\mathcal{O}(\frac{1}{\sqrt{n}})$. Note also that the bound does not depend on the value of θ_0.

(2) Note that the calculation of $\mathrm{E}|\frac{1}{\theta_0}-X|^3$ requires a significant amount of steps. Therefore, one could use Hölder's inequality with $\mathrm{E}|\frac{1}{\theta_0}-X|^3\le[\mathrm{E}(\frac{1}{\theta_0}-X)^4]^{3/4}=\frac{9^{3/4}}{\theta_0^3}$ using the results in pages 70–73 of Kendall and Stuart [8].

3.1.2. *The non-canonical case*

Let $X_1,X_2,\dots,X_n$ be i.i.d. random variables from $\mathrm{Exp}(\frac{1}{\theta})$, with p.d.f.

$$\begin{aligned} f(x|\theta)&=\frac{1}{\theta}\exp\biggl\{-\frac{1}{\theta}x\biggr\}=\exp\biggl\{-\log\theta-\frac{1}{\theta}x\biggr\} \\ &=\exp\bigl\{k(\theta)T(x)-A(\theta)+S(x)\bigr\}\mathbb{1}_{\{x\in B\}}, \end{aligned} \tag{3.3}$$

where $B=(0,\infty)$, $\theta\in\Theta=(0,\infty)$, $T(x)=-x$, $k(\theta)=\frac{1}{\theta}$, $A(\theta)=\log\theta$ and $S(x)=0$. Again, it is easy to show that the MLE exists, it is unique, equal to $\hat{\theta}_n(\mathbf{X})=\bar{X}$ and (A1)–(A5) are satisfied.

For ε as before and $h \in H$, Corollary 3.1 gives

$$d_{bW}\big(\sqrt{ni(\theta_0)}\big(\hat{\theta}_n(\mathbf{X}) - \theta_0\big), Z\big) \leq \frac{4.41456}{\sqrt{n}} + \frac{8}{n} + \frac{2}{\sqrt{n}} + \frac{1}{\sqrt{n}}\left(80\left[3\left(\frac{2}{n} + 1\right)\right]^{1/2}\right). \tag{3.4}$$

The Mean Squared Error is found to be $\mathrm{E}(\hat{\theta}_n(\mathbf{X}) - \theta_0)^2 = \mathrm{E}(\bar{X} - \theta_0)^2 = \frac{\theta_0^2}{n}$. Also (3.3) gives that $l^{(3)}(\theta; \mathbf{X}) = -\frac{2n}{\theta^3} + \frac{6}{\theta^4}\sum_{i=1}^n X_i = \frac{2n}{\theta^4}(3\hat{\theta}_n(\mathbf{X}) - \theta)$ and the triangle inequality yields

$$\sup_{\theta:|\theta-\theta_0|\leq\varepsilon} \left|l^{(3)}(\theta; \mathbf{X})\right| \leq \sup_{\theta:|\theta-\theta_0|\leq\varepsilon}\left[\left|\frac{6n\hat{\theta}_n(\mathbf{X})}{\theta^4}\right| + \left|\frac{2n}{\theta^3}\right|\right] = \frac{2n}{(\theta_0 - \varepsilon)^4}\big(3\hat{\theta}_n(\mathbf{X}) + \theta_0 - \varepsilon\big).$$

Therefore,

$$\begin{aligned}
&\left[\mathrm{E}\left(\Big(\sup_{\theta:|\theta-\theta_0|\leq\varepsilon}\left|l^{(3)}(\theta; \mathbf{X})\right|\Big)^2 \big|\big|\hat{\theta}_n(\mathbf{X}) - \theta_0\big| \leq \varepsilon\right)\right]^{1/2} \\
&\leq \left[\mathrm{E}\left(\left(\frac{2n}{(\theta_0 - \varepsilon)^4}\big(3\hat{\theta}_n(\mathbf{X}) + \theta_0 - \varepsilon\big)\right)^2 \Big|\big|\hat{\theta}_n(\mathbf{X}) - \theta_0\big| \leq \varepsilon\right)\right]^{1/2} \\
&\leq \frac{2n}{(\theta_0 - \varepsilon)^4}\left[\mathrm{E}\big(\big(3\big|\hat{\theta}_n(\mathbf{X}) - \theta_0\big| + 4\theta_0 - \varepsilon\big)^2 \big|\big|\hat{\theta}_n(\mathbf{X}) - \theta_0\big| \leq \varepsilon\big)\right]^{1/2} \\
&\leq \frac{2n}{(\theta_0 - \varepsilon)^4}\left[(2\varepsilon + 4\theta_0)^2\right]^{1/2} = \frac{4n(2\theta_0 + \varepsilon)}{(\theta_0 - \varepsilon)^4}.
\end{aligned}$$

The quantity $[\mathrm{E}(\hat{\theta}_n(\mathbf{X}) - \theta_0)^4]^{1/2}$ is calculated using the results in page 73 and the equations (3.38), page 70 of Kendall and Stuart [8] along with the fact that $\hat{\theta}_n(\mathbf{X}) = \bar{X} \sim G(n, \frac{n}{\theta_0})$, yielding that $\mathrm{E}(\hat{\theta}_n(\mathbf{X}) - \theta_0)^4 = \frac{3\theta_0^4}{n^2}(\frac{2}{n} + 1)$. Therefore,

$$\begin{aligned}
&\left[\mathrm{E}\left(\Big(\sup_{\theta:|\theta-\theta_0|\leq\varepsilon}\left|l^{(3)}(\theta; \mathbf{X})\right|\Big)^2 \big|\big|\hat{\theta}_n(\mathbf{X}) - \theta_0\big| \leq \varepsilon\right)\right]^{1/2}\left[\mathrm{E}\big(\hat{\theta}_n(\mathbf{X}) - \theta_0\big)^4\right]^{1/2} \\
&\leq \frac{4n(2\theta_0 + \varepsilon)}{(\theta_0 - \varepsilon)^4}\left[\frac{3\theta_0^4}{n^2}\left(\frac{2}{n} + 1\right)\right]^{1/2} = \frac{4(2\theta_0 + \varepsilon)}{(\theta_0 - \varepsilon)^4}\left[3\theta_0^4\left(\frac{2}{n} + 1\right)\right]^{1/2}.
\end{aligned}$$

To find an upper bound for $\mathrm{E}(|R_2(\theta_0; \mathbf{X})|||\hat{\theta}_n(\mathbf{X}) - \theta_0| \leq \varepsilon)$,

$$\begin{aligned}
R_2(\theta_0; \mathbf{X}) &= \big(\hat{\theta}_n(\mathbf{X}) - \theta_0\big)\left(\frac{n}{\theta_0^2} - \frac{2n\bar{X}}{\theta_0^3} + \frac{n}{\theta_0^2}\right) = \big(\hat{\theta}_n(\mathbf{X}) - \theta_0\big)\left(\frac{2n}{\theta_0^2} - \frac{2n\bar{X}}{\theta_0^3}\right) \\
&= -\frac{2n(\hat{\theta}_n(\mathbf{X}) - \theta_0)^2}{\theta_0^3}.
\end{aligned}$$

Using Lemma 2.1 for $f(x) = x^2$ gives

$$\mathrm{E}\big[\big(\hat{\theta}_n(\mathbf{X}) - \theta_0\big)^2 \big| \big|\hat{\theta}_n(\mathbf{X}) - \theta_0\big| \le \varepsilon\big] \le \mathrm{E}\big[\big(\hat{\theta}_n(\mathbf{X}) - \theta_0\big)^2\big].$$

Finally,

$$\begin{aligned} \mathrm{E}\big(\big|R_2(\theta_0; \mathbf{x})\big| \big| \big|\hat{\theta}_n(\mathbf{X}) - \theta_0\big| \le \varepsilon\big) &= \mathrm{E}\bigg(\frac{2n}{\theta_0^3}\big(\hat{\theta}_n(\mathbf{X}) - \theta_0\big)^2 \Big| \big|\hat{\theta}_n(\mathbf{X}) - \theta_0\big| \le \varepsilon\bigg) \\ &\le \frac{2n}{\theta_0^3}\mathrm{E}\big[\big(\hat{\theta}_n(\mathbf{X}) - \theta_0\big)^2\big] = \frac{2}{\theta_0}. \end{aligned}$$

Applying now the general result of Corollary 3.1 for $\varepsilon = \frac{\theta_0}{2}$ yields the result in (3.4).

Remark 3.3. (1) In this case, the speed of convergence related to the sample size of the above upper bound is $\mathcal{O}(\frac{1}{\sqrt{n}})$ and the bound does not depend on θ_0.

(2) Comparing the upper bound in (3.4) with that in (3.2) for the canonical case we see that the first term is the same. However, the rest of the bound is larger in (3.4) than in (3.2) $\forall n \in \mathbb{N}$.

(3) In the specific occasion of independent, exponentially distributed random variables with rate parameter $\frac{1}{\theta_0}$, the MLE exists, it is unique and equal to $\bar{X}$. Define $W = \frac{\sqrt{n}(\bar{X} - \theta_0)}{\theta_0} = \frac{1}{\sqrt{n}}\sum_{i=1}^n Y_i$, where $Y_i = \frac{X_i - \theta_0}{\theta_0}$ are independent, zero mean and unit variance random variables. Also, $\mathrm{E}(W) = 0$ and $\mathrm{Var}(W) = \frac{1}{n\theta_0^2}\sum_{i=1}^n \mathrm{Var}(X_i) = 1$. Therefore, (1.5) can be used to show

$$d_{bW}\big(\sqrt{ni(\theta_0)}\big(\hat{\theta}_n(\mathbf{X}) - \theta_0\big), Z\big) \le \frac{1}{\sqrt{n}}\bigg(2 + \frac{1}{\theta_0^3}\mathrm{E}|X_1 - \theta_0|^3\bigg) \le \frac{4.41456}{\sqrt{n}}. \tag{3.5}$$

The upper bound given in (3.5) as a result of the direct use of Stein's method is smaller than the upper bound given in (3.4) using the general method explained in Section 2. However, in order to apply Stein's method directly, the quantity $(\hat{\theta}_n(\mathbf{x}) - \theta_0)\sqrt{ni(\theta_0)}$ is assumed to be a sum of independent random variables. The general method, on the other hand, gives an upper bound for (2.1), whatever the MLE is, as long as the assumptions expressed in the beginning of the section hold.

3.1.3. *Empirical results*

In this subsection, we study the accuracy of our bounds by simulations. We start by generating 10 000 trials of n random independent observations, x, from the exponential distribution. The means for the canonical and the non-canonical case are equal to 1 and 2, respectively. We evaluate the MLE, $\hat{\theta}_n(\mathbf{X})$, of the parameter in each trial, which in turn gives a vector of 10 000 values. We standardise these values and we apply to them the function $h(x) = \frac{1}{x^2+2}$ with $h \in H$ and $\|h\| = 0.5$, $\|h'\| = \frac{3\sqrt{1.5}}{16}$ to calculate the expressions in (2.3) and (2.12). Finally, we compare $|\mathrm{E}[h(\sqrt{ni(\theta_0)}(\hat{\theta}_n(\mathbf{X}) - \theta_0))] - \mathrm{E}[h(Z)]|$ with the sum of the right-hand sides of (2.3) and (2.12), using the difference between their values as a measure of the error. The results presented in the following tables are based on this particular function h while the bounded

Table 1. Results taken by simulations from the Exp(1) distribution

n	$\vert\hat{\mathrm{E}}[h((\hat{\theta}_n(\mathbf{X})-\theta_0)\sqrt{ni(\theta_0)})]-\mathrm{E}[h(Z)]\vert$	Upper bound	Error
10	0.007	1.955	1.948
100	0.002	0.336	0.334
1000	0.001	0.094	0.093
10 000	0.0002	0.029	0.0288
100 000	0.0001	0.009	0.0089

Wasserstein metric is a supremum over a broader class of test functions, given in (1.3). Here, $\mathrm{E}[h(Z)] = 0.379$ and the results from the simulations are shown in Tables 1 and 2. The tables indicate that $|\hat{\mathrm{E}}[h((\hat{\theta}_n(\mathbf{X})-\theta_0)\sqrt{ni(\theta_0)})]-\mathrm{E}[h(Z)]|$, the bound and the error, decrease as the sample size gets larger. All the values in Table 1 are smaller than the respective ones in Table 2, as expected from Remark 3.3. The bounds are not very good for $n = 100$. The reason might be due to the crude upper bound related to the second term of the bound in (3.1). However, when $n \geq 1000$ the bounds are informative. For the non-canonical case the bounds using directly Lemma 1.1 are, as expected, much better than those from the general approach. The bounds are conceptual and better constraints may be possible.

4. Discrete distributions: The boundary issue

In this section, we use a perturbation method for any discrete distribution that faces the problem of the MLE having positive probability of being on the boundary of the parameter space. We also illustrate the perturbation for the specific example of the Poisson distribution.

4.1. The perturbation approach

A perturbation method based on a perturbation function, should be such that first of all, the function should perturb the quantity of interest in a way that ensures it will be interior to its domain. The second requirement is that the perturbed quantity should be as close as possible

Table 2. Results taken by simulations from the Exp(0.5) distribution treated as a non-canonical exponential family

n	$\vert\hat{\mathrm{E}}[h((\hat{\theta}_n(\mathbf{X})-\theta_0)\sqrt{ni(\theta_0)})]-\mathrm{E}[h(Z)]\vert$	Bound	Error	Bound using Lemma 1.1
10	0.004	11.888	11.884	0.321
100	0.003	3.401	3.398	0.101
1000	0.002	1.058	1.056	0.032
10 000	0.001	0.333	0.332	0.010
100 000	0.0005	0.105	0.1045	0.003

to the initial quantity. Let X be a random variable with support B, the connected closed (semi-closed) interval $[a,b]$ ($(a,b]$ or $[a,b)$), where $-\infty < a < b < \infty$. For $0 < \varepsilon < \frac{b-a}{2}$, we are looking for a perturbation function, $q : B \to \overset{\circ}{B}$ (where in this case, $\overset{\circ}{B}$ denotes the interior of the set B) with $q(x) = kx + d$, such that:

(1) $q(a) = a + \varepsilon$ and $q(b) = b - \varepsilon$.
(2) $\sup_x |q(x) - x|$ is minimum, $x \in B$.

Solving this problem for k and d, gives $k = 1 - \frac{2\varepsilon}{b-a}$ and $d = \varepsilon + \frac{2a}{b-a}\varepsilon$. There is only one solution, which is minimal. Thus, the second requirement is also satisfied. Choose $\varepsilon = \varepsilon(n) = \frac{c}{n}$ and $0 < c < \frac{n(b-a)}{2}$. Finally, the perturbation function is

$$q(x) = x + \frac{c}{n} - \frac{2c}{n}\left(\frac{x-a}{b-a}\right), \qquad x \in B, 0 < c < \frac{n(b-a)}{2}. \tag{4.1}$$

In the case where $B = (-\infty, b]$ or $B = [a, \infty)$, then $q(x) = x - \frac{c}{n}$ or $q(x) = x + \frac{c}{n}$, respectively.

Assuming existence and uniqueness of the MLE, $\hat{\theta}_n(\mathbf{X})$, for the parameter θ_0, of a discrete distribution with parameter space as in the previous paragraph, the aim is to find an upper bound on

$$d_{bW}\big(\sqrt{n}\big(\hat{\theta}_n(\mathbf{X}) - \theta_0\big), K\big),$$

where $K \sim \mathrm{N}(0, \frac{1}{i(\theta_0)})$. Note that $\mathrm{N}(0,0)$ is point mass at 0. The quantity we will bound is not exactly the one shown in (2.1) because the Expected Fisher Information number might not exist or not be finite when θ_0 lies on the boundary of the parameter space. For this purpose, we will use the perturbation function in (4.1) for both the parameter and the data.

First, we introduce some notations. For S being the discrete sample space, let $a := \inf\Theta$, $b := \sup\Theta$, $S_1 := \inf S$, $S_p := \sup S$ and $0 < c_1 < \frac{n(b-a)}{2}$, $0 < c_2 < \frac{n(S_p - S_1)}{2}$. In addition, $\theta_0^* = \theta_0 + \frac{c_1}{n} - \frac{2c_1}{n}(\frac{\theta_0 - a}{b-a})$ is the perturbed parameter and

$$q(x_i) = x_i + \frac{c_2}{n} - \frac{2c_2}{n}\left(\frac{x_i - S_1}{S_p - S_1}\right) \tag{4.2}$$

is the perturbed data. The perturbed MLE is denoted by $\hat{\theta}_n^*(\mathbf{x}) := \hat{\theta}_n(\mathbf{x})|_{\mathbf{x}=q(\mathbf{x})}$. Also,

$$l'\big(\theta_0^*; q(\mathbf{x})\big) := l'(\theta; \mathbf{x})\Big|_{\substack{\theta=\theta_0^* \\ \mathbf{x}=q(\mathbf{x})}}, \qquad l''\big(\theta_0^*; q(\mathbf{x})\big) := l''(\theta; \mathbf{x})\Big|_{\substack{\theta=\theta_0^* \\ \mathbf{x}=q(\mathbf{x})}},$$
$$l^{(3)}\big(\theta; q(\mathbf{x})\big) = l^{(3)}(\theta; \mathbf{x})\big|_{\mathbf{x}=q(\mathbf{x})}.$$

For ease of presentation, abbreviate $Y_i = \frac{l'(\theta_0^*; q(X_i))}{\sqrt{n}\, i(\theta_0^*)}$, $i \in \{1, \ldots, n\}$ while $w_1 := w_1(n, \theta_0^*)$ and $w_2 := w_2(n, \theta_0^*)$ are its expectation and variance, respectively.

Theorem 4.1. *Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables from a single-parameter discrete distribution with parameter space the connected, closed or semi-closed interval $\Theta \subset \mathbb{R}$ and discrete sample space S. Assume that $i(\theta_0) > 0$ and let $\frac{1}{i(\theta_0)} = 0$ to be the continuous extension of*

$\frac{1}{i(\theta)}$ to $\theta \to \theta_0$ when θ_0 is such that $i(\theta_0)$ does not exist or it is equal to infinity. Let $h \in H$ and $0 < \varepsilon = \varepsilon(\theta_0^*)$ such that $(\theta_0^* - \varepsilon, \theta_0^* + \varepsilon) \subset \overset{\circ}{\Theta}$. Then

$$
\begin{aligned}
&d_{bW}\big(\sqrt{n}\big(\hat{\theta}_n(\mathbf{X}) - \theta_0\big), K\big) \\
&\quad \leq \frac{c_1}{\sqrt{n}}\left|1 - 2\left(\frac{\theta_0 - a}{b - a}\right)\right| + \sqrt{n}\mathrm{E}\big|\hat{\theta}_n(\mathbf{X}) - \hat{\theta}_n^*(\mathbf{X})\big| \\
&\qquad + \left[\left|1 - \frac{1}{\sqrt{w_2 n i(\theta_0)}}\right|\sqrt{n w_2 + (n w_1)^2} + \frac{\sqrt{n}|w_1|}{\sqrt{w_2 i(\theta_0)}}\right]\mathbb{1}\left\{\frac{1}{i(\theta_0)} > 0\right\} \\
&\qquad + \frac{1}{\sqrt{n}}\left(2 + \frac{1}{(w_2)^{3/2}}\mathrm{E}|Y_1 - w_1|^3\right)\mathbb{1}\left\{\frac{1}{i(\theta_0)} > 0\right\} + 2\frac{\mathrm{E}(\hat{\theta}_n^*(\mathbf{X}) - \theta_0^*)^2}{\varepsilon^2} \\
&\qquad + \frac{1}{\sqrt{n i(\theta_0^*)}}\bigg\{\mathrm{E}\big(\big|\big(\hat{\theta}_n^*(\mathbf{x}) - \theta_0^*\big)\big[l''\big(\theta_0^*; q(\mathbf{x})\big) + n i\big(\theta_0^*\big)\big]\big|\,\big|\,\big|\hat{\theta}_n^*(\mathbf{X}) - \theta_0^*\big| \leq \varepsilon\big) \\
&\qquad + \frac{1}{2}\bigg[\mathrm{E}\bigg(\Big(\sup_{\theta:|\theta - \theta_0^*| \leq \varepsilon}\big|l^{(3)}\big(\theta; q(\mathbf{X})\big)\big|\Big)^2\,\Big|\,\big|\hat{\theta}_n^*(\mathbf{X}) - \theta_0^*\big| \leq \varepsilon\bigg)\bigg]^{1/2}\big[\mathrm{E}\big(\hat{\theta}_n^*(\mathbf{X}) - \theta_0^*\big)^4\big]^{1/2}\bigg\}.
\end{aligned}
\tag{4.3}
$$

Proof. *Step* 1: *Perturbation of* θ_0. Using the triangle inequality and then a first order Taylor expansion of $h(\sqrt{n}(\hat{\theta}_n(\mathbf{X}) - \theta_0))$ about $\sqrt{n}(\hat{\theta}_n(\mathbf{X}) - \theta_0^*)$ gives

$$
\begin{aligned}
&\big|\mathrm{E}\big[h\big(\sqrt{n}\big(\hat{\theta}_n(\mathbf{X}) - \theta_0\big)\big)\big] - \mathrm{E}\big[h(K)\big]\big| \\
&\quad \leq \big|\mathrm{E}\big[h\big(\sqrt{n}\big(\hat{\theta}_n(\mathbf{X}) - \theta_0^*\big)\big)\big] - \mathrm{E}\big[h(K)\big]\big| \\
&\qquad + \big|\mathrm{E}\big[h\big(\sqrt{n}\big(\hat{\theta}_n(\mathbf{X}) - \theta_0\big)\big) - h\big(\sqrt{n}\big(\hat{\theta}_n(\mathbf{X}) - \theta_0^*\big)\big)\big]\big| \\
&\quad \leq \big|\mathrm{E}\big[h\big(\sqrt{n}\big(\hat{\theta}_n(\mathbf{X}) - \theta_0^*\big)\big)\big] - \mathrm{E}\big[h(K)\big]\big| + \sqrt{n}\|h'\|\mathrm{E}\big|\theta_0^* - \theta_0\big| \\
&\quad = \big|\mathrm{E}\big[h\big(\sqrt{n}\big(\hat{\theta}_n(\mathbf{X}) - \theta_0^*\big)\big)\big] - \mathrm{E}\big[h(K)\big]\big| + \frac{\|h'\|c_1}{\sqrt{n}}\left|1 - 2\left(\frac{\theta_0 - a}{b - a}\right)\right|.
\end{aligned}
\tag{4.4}
$$

Step 2: *Perturbation of the MLE*. To perturb the MLE, we perturb the data. The perturbed data is denoted by $q(\mathbf{x}) = (q(x_1), q(x_2), \ldots, q(x_n))$, with $q(x_i)$ given in (4.2). This construction ensures that the MLE evaluated at $q(\mathbf{x})$ is not on the boundary of the parameter space. Following the same process as in (4.4), using the triangle inequality and a first order Taylor expansion of $h(\sqrt{n}(\hat{\theta}_n(\mathbf{X}) - \theta_0^*))$ about $\sqrt{n}(\hat{\theta}_n^*(\mathbf{X}) - \theta_0^*)$ gives

$$
\begin{aligned}
&\big|\mathrm{E}\big[h\big(\sqrt{n}\big(\hat{\theta}_n(\mathbf{X}) - \theta_0^*\big)\big)\big] - \mathrm{E}\big[h(K)\big]\big| \\
&\quad \leq \big|\mathrm{E}\big[h\big(\sqrt{n}\big(\hat{\theta}_n^*(\mathbf{X}) - \theta_0^*\big)\big)\big] - \mathrm{E}\big[h(K)\big]\big| \\
&\qquad + \big|\mathrm{E}\big[h\big(\sqrt{n}\big(\hat{\theta}_n(\mathbf{X}) - \theta_0^*\big)\big) - h\big(\sqrt{n}\big(\hat{\theta}_n^*(\mathbf{X}) - \theta_0^*\big)\big)\big]\big| \\
&\quad \leq \big|\mathrm{E}\big[h\big(\sqrt{n}\big(\hat{\theta}_n^*(\mathbf{X}) - \theta_0^*\big)\big)\big] - \mathrm{E}\big[h(K)\big]\big| + \sqrt{n}\|h'\|\mathrm{E}\big|\hat{\theta}_n(\mathbf{X}) - \hat{\theta}_n^*(\mathbf{X})\big|.
\end{aligned}
\tag{4.5}
$$

Step 3: *The final bound.* It remains to bound

$$\big|\mathrm{E}\big[h\big(\sqrt{n}\big(\hat{\theta}_n^*(\mathbf{X}) - \theta_0^*\big)\big)\big] - \mathrm{E}\big[h(K)\big]\big|.$$

Since both θ_0^* and $\hat{\theta}_n^*(\mathbf{x})$ are interior to Θ, a second-order Taylor expansion of $l'(\hat{\theta}_n^*(\mathbf{x}); q(\mathbf{x}))$ about θ_0^* yields

$$0 = l'\big(\theta_0^*; q(\mathbf{x})\big) + \big(\hat{\theta}_n^*(\mathbf{x}) - \theta_0^*\big) l''\big(\theta_0^*; q(\mathbf{x})\big) + R_1\big(\theta_0^*; q(\mathbf{x})\big), \tag{4.6}$$

where, similarly as in Section 2,

$$R_1\big(\theta_0^*; q(\mathbf{x})\big) = \tfrac{1}{2}\big(\hat{\theta}_n^*(\mathbf{x}) - \theta_0^*\big)^2 l^{(3)}\big(\tilde{\theta}; q(\mathbf{x})\big)$$

with

$$l^{(3)}\big(\tilde{\theta}; q(\mathbf{x})\big) = l^{(3)}(\theta; \mathbf{x})\Big|_{\substack{\theta=\tilde{\theta} \\ \mathbf{x}=q(\mathbf{x})}}$$

for $\tilde{\theta}$ between $\hat{\theta}_n^*(\mathbf{x})$ and θ_0^*. A simple rearrangement of the terms in (4.6), leads to $\hat{\theta}_n^*(\mathbf{x}) - \theta_0^* = \frac{-l'(\theta_0^*; g(\mathbf{x})) - R_1(\theta_0^*; g(\mathbf{x}))}{l''(\theta_0^*; g(\mathbf{x}))}$. Since, in general $l''(\theta_0^*; q(\mathbf{x})) \neq -ni(\theta_0^*)$, using the results in the proof of Theorem 2.1 gives

$$\hat{\theta}_n^*(\mathbf{x}) - \theta_0^* = \frac{l'(\theta_0^*; q(\mathbf{x})) + R_1(\theta_0^*; q(\mathbf{x})) + R_2(\theta_0^*; q(\mathbf{x}))}{ni(\theta_0^*)},$$

where

$$R_2\big(\theta_0^*; q(\mathbf{x})\big) = \big(\hat{\theta}_n^*(\mathbf{x}) - \theta_0^*\big)\big[l''\big(\theta_0^*; q(\mathbf{x})\big) + ni\big(\theta_0^*\big)\big].$$

Using that $q(\mathbf{X}) = (q(X_1), q(X_2), \dots, q(X_n))$, the triangle inequality gives

$$\begin{aligned}
&\big|\mathrm{E}\big[h\big(\sqrt{n}\big(\hat{\theta}_n^*(\mathbf{X}) - \theta_0^*\big)\big)\big] - \mathrm{E}\big[h(K)\big]\big| \\
&\quad \leq \left|\mathrm{E}\left[h\left(\frac{l'(\theta_0^*; q(\mathbf{X}))}{\sqrt{n}i(\theta_0^*)}\right)\right] - \mathrm{E}\big[h(K)\big]\right| \\
&\qquad + \left|\mathrm{E}\left[h\left(\frac{l'(\theta_0^*; q(\mathbf{X})) + R_1(\theta_0^*; q(\mathbf{X})) + R_2(\theta_0^*; q(\mathbf{X}))}{\sqrt{n}i(\theta_0^*)}\right) - h\left(\frac{l'(\theta_0^*; q(\mathbf{X}))}{\sqrt{n}i(\theta_0^*)}\right)\right]\right|.
\end{aligned} \tag{4.7}$$

(A) To find an upper bound on the first quantity on the right-hand side of (4.7) using Lemma 1.1, note that

$$\frac{l'(\theta_0^*; q(\mathbf{X}))}{\sqrt{n}i(\theta_0^*)} = \sum_{i=1}^{n} Y_i, \qquad \text{where } Y_i = \frac{l'(\theta_0^*; q(X_i))}{\sqrt{n}i(\theta_0^*)}.$$

Denote by $w_1 := w_1(n)$ and $w_2 := w_2(n)$ the expectation and the variance of $Y_i, i = 1, 2, \dots, n$, respectively. These quantities depend on the sample size and on the perturbed values (θ_0^* and

$q(x_i)$). Define $\tilde{Y}_i = \frac{Y_i - w_1}{\sqrt{w_2 i(\theta_0)}}$, $\forall i \in \{1, 2, \dots, n\}$ with $\mathrm{E}(\tilde{Y}_i) = 0$ and $\mathrm{Var}(\tilde{Y}_i) = \frac{1}{i(\theta_0)}$. As a consequence of $X_1, X_2, \dots, X_n$ being i.i.d. random variables, $\tilde{Y}_1, \tilde{Y}_2, \dots, \tilde{Y}_n$ are i.i.d. random variables too. Using the triangle inequality and that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \tilde{Y}_i = \frac{1}{\sqrt{w_2 n i(\theta_0)}} \left(\frac{l'(\theta_0^*; q(\mathbf{X}))}{\sqrt{n i(\theta_0^*)}} - n w_1 \right)$$

gives

$$\begin{aligned} &\left| \mathrm{E}\left[h\left(\frac{l'(\theta_0^*; q(\mathbf{X}))}{\sqrt{n} i(\theta_0^*)} \right) \right] - \mathrm{E}\left[h(K)\right] \right| \\ &\quad \leq \left| \mathrm{E}\left[h\left(\frac{1}{\sqrt{w_2 n i(\theta_0)}} \left(\frac{l'(\theta_0^*; q(\mathbf{X}))}{\sqrt{n} i(\theta_0^*)} - n w_1 \right) \right) \right] - \mathrm{E}\left[h(K)\right] \right| \\ &\qquad + \left| \mathrm{E}\left[h\left(\frac{l'(\theta_0^*; q(\mathbf{X}))}{\sqrt{n} i(\theta_0^*)} \right) - h\left(\frac{1}{\sqrt{w_2 n i(\theta_0)}} \left(\frac{l'(\theta_0^*; q(\mathbf{X}))}{\sqrt{n} i(\theta_0^*)} - n w_1 \right) \right) \right] \right|. \end{aligned} \tag{4.8}$$

The first term of the bound in (4.8) will be bounded using Lemma 1.1 with $W = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \tilde{Y}_i$. Thus,

$$\begin{aligned} &\left| \mathrm{E}\left[h\left(\frac{1}{\sqrt{w_2 n i(\theta_0)}} \left(\frac{l'(\theta_0^*; q(\mathbf{X}))}{\sqrt{n} i(\theta_0^*)} - n w_1 \right) \right) \right] - \mathrm{E}\left[h(K)\right] \right| \\ &\quad \leq \frac{\|h'\|}{\sqrt{n}} \left(2 + \left[i(\theta_0)\right]^{3/2} \mathrm{E}|\tilde{Y}_1|^3\right) = \frac{\|h'\|}{\sqrt{n}} \left(2 + \frac{1}{(w_2)^{3/2}} \mathrm{E}|Y_1 - w_1|^3 \right). \end{aligned} \tag{4.9}$$

For the second term of the upper bound in (4.8) a first-order Taylor expansion and the Cauchy–Schwarz inequality yield

$$\begin{aligned} &\left| \mathrm{E}\left[h\left(\frac{l'(\theta_0^*; q(\mathbf{X}))}{\sqrt{n} i(\theta_0^*)} \right) - h\left(\frac{1}{\sqrt{w_2 n i(\theta_0)}} \left(\frac{l'(\theta_0^*; q(\mathbf{X}))}{\sqrt{n} i(\theta_0^*)} - n w_1 \right) \right) \right] \right| \\ &\quad \leq \|h'\| \left| 1 - \frac{1}{\sqrt{w_2 n i(\theta_0)}} \right| \mathrm{E} \left| \frac{l'(\theta_0^*; q(\mathbf{X}))}{\sqrt{n} i(\theta_0^*)} \right| + \frac{\|h'\| \sqrt{n} |w_1|}{\sqrt{w_2 i(\theta_0)}} \\ &\quad \leq \|h'\| \left| 1 - \frac{1}{\sqrt{w_2 n i(\theta_0)}} \right| \left(\mathrm{Var}\left(\frac{l'(\theta_0^*; q(\mathbf{X}))}{\sqrt{n} i(\theta_0^*)} \right) + \frac{[\mathrm{E}(l'(\theta_0^*; q(\mathbf{X})))]^2}{n [i(\theta_0^*)]^2} \right)^{1/2} \\ &\qquad + \frac{\|h'\| \sqrt{n} |w_1|}{\sqrt{w_2 i(\theta_0)}} \\ &\quad = \|h'\| \left[\left| 1 - \frac{1}{\sqrt{w_2 n i(\theta_0)}} \right| \sqrt{n w_2 + (n w_1)^2} + \frac{\sqrt{n} |w_1|}{\sqrt{w_2 i(\theta_0)}} \right]. \end{aligned} \tag{4.10}$$

When $\frac{1}{i(\theta_0)} = 0$ then $\tilde{Y}_i = 0, \forall i \in \{1, 2, \dots, n\}$ and by following the above process, the first term on the right-hand side of (4.7) is equal to zero.

(B) To complete the proof, it remains to find an upper bound for the second term on the right-hand side of (4.7). The idea is the same as the one used for (2.12). We condition on whether $|\hat{\theta}_n^*(\mathbf{X}) - \theta_0^*| > \varepsilon$ or $|\hat{\theta}_n^*(\mathbf{X}) - \theta_0^*| \leq \varepsilon$, where now $\varepsilon = \varepsilon(\theta_0^*)$ and $0 < \varepsilon$ $(\theta_0^* - \varepsilon, \theta_0^* + \varepsilon) \subset \overset{\circ}{\Theta}$. Following the same process as in Section 2 yields

$$\begin{aligned}
&\left|\mathrm{E}\left[h\left(\frac{l'(\theta_0^*; g(\mathbf{X})) + R_1(\theta_0^*; g(\mathbf{X})) + R_2(\theta_0^*; g(\mathbf{X}))}{\sqrt{n}i(\theta_0^*)}\right) - h\left(\frac{l'(\theta_0^*; g(\mathbf{X}))}{\sqrt{n}i(\theta_0^*)}\right)\right]\right| \\
&\quad \leq 2\|h\|\frac{\mathrm{E}(\hat{\theta}_n^*(\mathbf{X}) - \theta_0^*)^2}{\varepsilon^2} + \frac{\|h'\|}{\sqrt{n}i(\theta_0^*)}\Big\{\mathrm{E}\big(\big|R_2\big(\theta_0^*; g(\mathbf{X})\big)\big|\big||\hat{\theta}_n^*(\mathbf{X}) - \theta_0^*| \leq \varepsilon\big) \qquad (4.11) \\
&\qquad + \frac{\|h'\|}{2}\Big[\mathrm{E}\Big(\Big(\sup_{\theta:|\theta-\theta_0^*|\leq\varepsilon}\big|l^{(3)}\big(\theta; g(\mathbf{X})\big)\big|\Big)^2\big||\hat{\theta}_n^*(\mathbf{X}) - \theta_0^*| \leq \varepsilon\Big)\Big]^{1/2}\big[\mathrm{E}\big(\hat{\theta}_n^*(\mathbf{X}) - \theta_0^*\big)^4\big]^{1/2}\Big\}.
\end{aligned}$$

Combining (4.4), (4.5), (4.9), (4.10) and (4.11) and the fact that $\|h\| \leq 1$, $\|h'\| \leq 1$ gives the result in (4.3). □

Remark 4.1. (1) In order for the above bound to approach zero as the sample size, n, increases we require that $\mathrm{E}|\hat{\theta}_n(\mathbf{X}) - \hat{\theta}_n^*(\mathbf{X})| = o(\frac{1}{\sqrt{n}})$.

(2) When both endpoints of the parameter space are not finite, then parameter perturbation is not necessary. In the case where one of the two endpoints of the now semi-closed parameter space is infinite, then it suffices to change the form of the perturbed parameter, which now becomes

$$\theta_0^* = \theta_0 - \frac{c_1}{n} \qquad \text{if the left endpoint is equal to } -\infty,$$

$$\theta_0^* = \theta_0 + \frac{c_1}{n} \qquad \text{if the right endpoint is equal to } \infty.$$

The same holds regarding the sample space and the relevant perturbation of the data.

4.2. Example: The Poisson distribution

In this subsection, we consider the Poisson distribution with parameter $\theta \in \Theta = [0, \infty)$. The value $\theta = 0$ must be in the parameter space in order for the MLE, $\hat{\theta}_n(\mathbf{X}) = \bar{X}$, to exist and to be unique. The Poisson(θ) distribution with the aforementioned parameter space is not a single-parameter exponential family. When $\theta = 0$ is included in the parameter space the requirements of an exponential family are not satisfied as the set of values x for which the relevant probability mass function

$$f(x|\theta) = \frac{\mathrm{e}^{-\theta}\theta^x}{x!}, \qquad \theta \in [0, \infty), x \in \mathbb{Z}_0^+$$

is positive, is different for $\theta = 0$ than for any other value of the parameter θ; the support of the distribution depends on the parameter. Following the steps of the proof of Theorem 4.1, using also Hölder's inequality for the third absolute moment in the third term of the bound in (4.3) and taking $0 < c = c_1 = c_2$, which minimizes the bound, gives the next result.

Corollary 4.1. *Let $X_1, X_2, \dots, X_n$ be i.i.d. random variables which follow the* Poisson(θ_0) *distribution, with $\theta_0 \in [0, \infty)$. For $K \sim \mathrm{N}(0, \theta_0)$, $h \in H$ and $c > 0$ a positive constant,*

(1) *if $\theta_0 > 0$ then*

$$d_{bW}\big(\sqrt{n}\big(\hat{\theta}_n(\mathbf{X}) - \theta_0\big), K\big) \\ \leq \frac{2c}{\sqrt{n}} + \frac{1}{\sqrt{n}}\left[2 + \frac{(3\theta_0 + 1)^{3/4}}{\theta_0^{3/4}}\right] \\ + \frac{8\theta_0}{n(\theta_0 + c/n)^2} + \frac{\theta_0}{\sqrt{n}(\theta_0 + c/n)} + \frac{12}{\sqrt{n}(\theta_0 + c/n)}\left[\frac{\theta_0}{n} + 3\theta_0^2\right]^{1/2}; \tag{4.12}$$

(2) *if $\theta_0 = 0$ then*

$$d_{bW}\big(\sqrt{n}\hat{\theta}_n(\mathbf{X}), K\big) = 0.$$

Remark 4.2. (1) The upper bound expressed in (4.12) for the distributional distance between the actual distribution of the MLE and the normal distribution in the case of i.i.d. random variables following the Poisson(θ) distribution, with $\theta \in [0, \infty)$ is of order at most $\frac{1}{\sqrt{n}}$.

(2) Since the MLE is unique and equal to $\hat{\theta}_n(\mathbf{X}) = \bar{X}$, Lemma 1.1 could be used directly for $\bar{X}$. Define $W = \sqrt{n}(\bar{X} - \theta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^n Y_i$, where $Y_i = X_i - \theta_0$ are independent, zero mean random variables. Also, $\mathrm{E}(W) = 0$ and $\mathrm{Var}(W) = n\,\mathrm{Var}(\bar{X}) = \frac{1}{n}\sum_{i=1}^n \mathrm{Var}(X_i) = \theta_0$. Therefore, (1.5) for $K \sim \mathrm{N}(0, \theta_0)$ and Hölder's inequality give for $\theta_0 > 0$

$$d_{bW}\big(\sqrt{n}\big(\hat{\theta}_n(\mathbf{X}) - \theta_0\big), K\big) \leq \frac{1}{\sqrt{n}}\left(2 + \frac{1}{\theta_0^{3/2}}\big[\mathrm{E}(Y_1)^4\big]^{3/4}\right) = \frac{1}{\sqrt{n}}\left(2 + \frac{(3\theta_0 + 1)^{3/4}}{\theta_0^{3/4}}\right).$$

This bound, obtained by the direct application of Stein's method, is smaller than the bound given in Corollary 4.1. However, the interest in the example treated in this section, where $\Theta = [0, \infty)$, is in adapting the approach to such cases where the MLE could be on the boundary of the parameter space with positive probability when it is not assumed that the MLE is a sum of random variables.

5. Bounds on the Mean Squared Error of the MLE

This section focuses on the situation when an analytic form for the MLE is not available. In the proof for the final upper bound in Theorem 2.1, an explicit form of the MLE was not used. However, if the MLE is not known, then the MSE, $\mathrm{E}(\hat{\theta}_n(\mathbf{X}) - \theta_0)^2$, appearing in the bound for (2.1) should be bounded by a quantity which is independent of $\hat{\theta}_n(\mathbf{X})$.

Let $X_1, X_2, \dots, X_n$ be i.i.d. random variables. Apart from the regularity conditions, first defined in Section 1, we make the following further assumptions that make the steps and the calculations easier and ensure a meaningful upper bound:

(Fur.1) The support, S, is bounded;

(Fur.2) For $\varepsilon = \varepsilon(\theta_0) > 0$ such that $(\theta_0 - \varepsilon, \theta_0 + \varepsilon) \subset \Theta$, we require that there is a constant $C_1 = C_1(\theta_0)$ which depends on the unknown parameter θ_0 such that $\sup_{\theta:|\theta-\theta_0|\leq\varepsilon} |l^{(3)}(\theta; x_1)| \leq C_1$, where $C_1 = C_1(\theta_0)$ is a constant that depends on the unknown parameter θ_0;

(Fur.3) $\exists N \in \mathbb{N}$ such that $\forall n \geq N$ we have $1 - 2\frac{\|x^2\|}{ni(\theta_0)\varepsilon^2} - \frac{\|x\|C_1}{\sqrt{n}[i(\theta_0)]^{3/2}} > 0$ for ε as in (Fur.2). Solving the quadratic inequality, with unknown the $\sqrt{n}$ yields that n, the sample size, should satisfy

$$n \geq \frac{\|x\|^2[C_1\varepsilon + \sqrt{(C_1\varepsilon)^2 + 8[i(\theta_0)]^2}]^2}{4[i(\theta_0)]^3\varepsilon^2}.$$

For ease of presentation, let $D_1 = D_1(\theta_0, x, n) = 1 - 2\frac{\|x^2\|}{ni(\theta_0)\varepsilon^2} - \frac{\|x\|C_1}{\sqrt{n}[i(\theta_0)]^{3/2}}$.

Theorem 5.1. *Let $X_1, X_2, \dots, X_n$ be i.i.d. random variables with density or frequency function $f(x_i|\theta)$. Assume that the regularity conditions* (R1)–(R4), *as well as the assumptions* (Fur.1)–(Fur.3) *are satisfied. Also assume that the MLE exists and that it is unique. Then $A_1 = A_1(\theta_0, n)$ is an upper bound for $\sqrt{\mathrm{E}(\hat{\theta}_n(\mathbf{X}) - \theta_0)^2}$, where for ε as in* (Fur.2),

$$\begin{aligned} A_1 = [2D_1]^{-1}\Bigg\{ & \frac{2\|x\|\sqrt{\mathrm{Var}[l''(\theta_0; X_1)]}}{n[i(\theta_0)]^{3/2}} \\ & + \Bigg[4\frac{\|x\|^2\,\mathrm{Var[1''(\theta_0; X_1)]}}{n^2[i(\theta_0)]^3} \\ & + \frac{4D_1}{ni(\theta_0)}\Bigg[1 + 2\frac{\|x\|}{\sqrt{n}}\bigg(2 + \frac{\mathrm{E}|l'(\theta_0; X_1)|^3}{[i(\theta_0)]^{3/2}}\bigg)\Bigg]\Bigg]^{1/2}\Bigg\}. \end{aligned} \tag{5.1}$$

Proof. Using the notations for the remainder terms, the triangle inequality, conditional expectations, Markov's inequality and Stein's method, the same way as in Section 2, gives

$$\begin{aligned} & \big|\mathrm{E}\big[h\big(\big(\hat{\theta}_n(\mathbf{X}) - \theta_0\big)\sqrt{ni(\theta_0)}\big)\big] - \mathrm{E}\big[h(Z)\big]\big| \\ & \quad \leq \frac{\|h'\|}{\sqrt{n}}\bigg(2 + \frac{\mathrm{E}|l'(\theta_0; X_1)|^3}{[i(\theta_0)]^{3/2}}\bigg) \\ & \qquad + \frac{2\|h\|\mathrm{E}(\hat{\theta}_n(\mathbf{X}) - \theta_0)^2}{\varepsilon^2} \\ & \qquad + \frac{\|h'\|}{\sqrt{ni(\theta_0)}}\big|\mathrm{E}\big[R_2(\theta_0; \mathbf{X})\big|\big|\hat{\theta}_n(\mathbf{X}) - \theta_0\big| \leq \varepsilon\big]\big|\mathbb{P}\big(\big|\hat{\theta}_n(\mathbf{X}) - \theta_0\big| \leq \varepsilon\big) \\ & \qquad + \frac{\|h'\|}{2\sqrt{ni(\theta_0)}}\mathrm{E}\Big(\big(\hat{\theta}_n(\mathbf{X}) - \theta_0\big)^2 \sup_{\theta:|\theta-\theta_0|\leq\varepsilon}\big|l^{(3)}(\theta; \mathbf{X})\big|\big|\big|\hat{\theta}_n(\mathbf{X}) - \theta_0\big| \leq \varepsilon\Big). \end{aligned}$$

Using the definition of $R_2(\theta_0; \mathbf{x})$ and the Cauchy–Schwarz inequality yields

$$\begin{aligned}
&\big|\mathrm{E}\big[R_2(\theta_0; \mathbf{X})\big|\big|\hat{\theta}_n(\mathbf{X}) - \theta_0\big| \le \varepsilon\big]\big|\mathbb{P}\big(\big|\hat{\theta}_n(\mathbf{X}) - \theta_0\big| \le \varepsilon\big)\\
&\quad\le \mathrm{E}\big|\big(ni(\theta_0) + l''(\theta_0; \mathbf{X})\big)\big(\hat{\theta}_n(\mathbf{X}) - \theta_0\big)\big|\\
&\quad\le \sqrt{\mathrm{E}\big[ni(\theta_0) + l''(\theta_0; \mathbf{X})\big]^2 \mathrm{E}\big[\hat{\theta}_n(\mathbf{X}) - \theta_0\big]^2}\\
&\quad= \sqrt{n \operatorname{Var}\big(l''(\theta_0; X_1)\big)}\sqrt{\mathrm{E}\big[\hat{\theta}_n(\mathbf{X}) - \theta_0\big]^2},
\end{aligned}$$

which leads to

$$\begin{aligned}
&\big|\mathrm{E}\big[h\big(\big(\hat{\theta}_n(\mathbf{X}) - \theta_0\big)\sqrt{ni(\theta_0)}\big)\big] - \mathrm{E}\big[h(Z)\big]\big|\\
&\quad\le \frac{\|h'\|}{\sqrt{n}}\left(2 + \frac{\mathrm{E}|l'(\theta_0; X_1)|^3}{[i(\theta_0)]^{3/2}}\right)\\
&\qquad+ \frac{2\|h\|\mathrm{E}(\hat{\theta}_n(\mathbf{X}) - \theta_0)^2}{\varepsilon^2} + \frac{\|h'\|nC_1}{2\sqrt{ni(\theta_0)}}\mathrm{E}\big(\hat{\theta}_n(\mathbf{X}) - \theta_0\big)^2\\
&\qquad+ \frac{\|h'\|\sqrt{\operatorname{Var}(l''(\theta_0; X_1))}\sqrt{\mathrm{E}(\hat{\theta}_n(\mathbf{X}) - \theta_0)^2}}{\sqrt{i(\theta_0)}}.
\end{aligned} \tag{5.2}$$

Straightforward calculations and denoting with B_{x^2} the upper bound for (2.1) when $h(x) = x^2$, lead to

$$\begin{aligned}
\mathrm{E}\big(\hat{\theta}_n(\mathbf{X}) - \theta_0\big)^2 &= \frac{1}{ni(\theta_0)}\big|\mathrm{E}\big[\sqrt{ni(\theta_0)}\big(\hat{\theta}_n(\mathbf{X}) - \theta_0\big)\big]^2 - \mathrm{E}\big(Z^2\big) + \mathrm{E}\big(Z^2\big)\big|\\
&\le \frac{1}{ni(\theta_0)}(B_{x^2} + 1),
\end{aligned} \tag{5.3}$$

where

$$\begin{aligned}
B_{x^2} &\le 2\frac{\|x\|}{\sqrt{n}}\left(2 + \frac{\mathrm{E}|l'(\theta_0; X_1)|^3}{[i(\theta_0)]^{3/2}}\right) + \frac{2\|x^2\|\mathrm{E}(\hat{\theta}_n(\mathbf{X}) - \theta_0)^2}{\varepsilon^2} + \frac{\|x\|\sqrt{n}C_1}{\sqrt{i(\theta_0)}}\mathrm{E}\big(\hat{\theta}_n(\mathbf{X}) - \theta_0\big)^2\\
&\quad+ 2\frac{\|x\|\sqrt{\operatorname{Var}(l''(\theta_0; X_1))}\sqrt{\mathrm{E}(\hat{\theta}_n(\mathbf{X}) - \theta_0)^2}}{\sqrt{i(\theta_0)}}.
\end{aligned}$$

Now B_{x^2} also includes $\mathrm{E}(\hat{\theta}_n(\mathbf{X}) - \theta_0)^2$ and its positive root. Therefore, the next step is to solve the simple quadratic inequality (5.3), with unknown $\sqrt{\mathrm{E}(\hat{\theta}_n(\mathbf{X}) - \theta_0)^2}$. Using (Fur.3), after basic calculations we obtain that $0 < \sqrt{\mathrm{E}(\hat{\theta}_n(\mathbf{X}) - \theta_0)^2} \le A_1$. □

Remark 5.1. (1) Using this result, the final upper bound for (2.1) which is useful when no analytic expression of the MLE is available, becomes

$$d_{bW}\left(\sqrt{ni(\theta_0)}\left(\hat{\theta}_n(\mathbf{X})-\theta_0\right),Z\right)\leq\frac{1}{\sqrt{n}}\left(2+\frac{\mathrm{E}|l'(\theta_0;X_1)|^3}{[i(\theta_0)]^{3/2}}\right)+\frac{2(A_1)^2}{\varepsilon^2}$$
$$+\frac{\sqrt{n}C_1(A_1)^2}{2\sqrt{i(\theta_0)}}+\frac{\sqrt{\mathrm{Var}[l''(\theta_0;X_1)]}A_1}{\sqrt{i(\theta_0)}}. \tag{5.4}$$

(2) The order of A_1 in terms of the sample size is $\frac{1}{\sqrt{n}}$ and hence the order of the final upper bound in (5.4) is also $\frac{1}{\sqrt{n}}$.

Example (The Beta distribution). Consider the example of i.i.d random variables from the Beta distribution with one of the two shape parameters being unknown. In this case, the MLE can only be expressed in terms of the inverse of the digamma function, $\Psi(\theta)=\frac{\mathrm{d}}{\mathrm{d}\theta}\log\Gamma(\theta)$. We use the general result in Theorem 5.1, in order to obtain an upper bound for the MSE and use it to get an upper bound for (2.1). The following corollary gives the result.

Corollary 5.1. *Let* $X_1,X_2,\dots,X_n$ *be i.i.d. random variables from the* $\mathrm{Beta}(\theta_0,\beta)$ *distribution, where* β *is known and* θ_0 *is unknown. Let* $B_1=B_1(\theta_0)=8(\Psi_3(\theta_0)+\Psi_3(\theta_0+\beta)+3[\Psi_1(\theta_0)]^2+3[\Psi_1(\theta_0+\beta)]^2)$, *where* $\Psi_j(\theta)$, $j\in\mathbb{N}$ *is the* j*th derivative of the digamma function*, $\Psi(\theta)$. *Also, let* $B_2=B_2(\theta_0)=\frac{96\beta+6.6\beta\theta_0^4}{\theta_0^4}$, $D_{\Psi 1}=D_{\Psi 1}(\theta_0,\beta)=\Psi_1(\theta_0)-\Psi_1(\theta_0+\beta)$ *and*

$$B_3=B_3(\theta_0,n)=\left[\left(4+\frac{8}{\sqrt{n}}\left(2+\frac{(B_1)^{3/4}}{D_{\Psi 1}^{3/2}}\right)\right)\left(1-\frac{8}{n\theta_0^2D_{\Psi 1}}-\frac{B_2}{\sqrt{n}D_{\Psi 1}^{3/2}}\right)\right]^{1/2}$$
$$\times\left(2\left(\sqrt{D_{\Psi 1}}-\frac{8}{n\theta_0^2\sqrt{D_{\Psi 1}}}-\frac{B_2}{\sqrt{n}D_{\Psi 1}}\right)\right)^{-1}. \tag{5.5}$$

Let

$$n\geq\left[B_2\frac{\theta_0}{2}+\sqrt{\frac{(B_2\theta_0)^2}{4}+8\left[\Psi_1(\theta_0)-\Psi_1(\theta_0+\beta)\right]^2}\right]^2\left(\left[\Psi_1(\theta_0)-\Psi_1(\theta_0+\beta)\right]^3\theta_0^2\right)^{-1}.$$

Then for $Z\sim\mathrm{N}(0,1)$

$$d_{bW}\left(\sqrt{ni(\theta_0)}\left(\hat{\theta}_n(\mathbf{X})-\theta_0\right),Z\right)\leq\frac{1}{\sqrt{n}}\left(2+\frac{(B_1)^{3/4}}{[\Psi_1(\theta_0)-\Psi_1(\theta_0+\beta)]^{3/2}}\right)$$
$$+\frac{8}{n\theta_0^2}(B_3)^2+\frac{B_2(B_3)^2}{2\sqrt{n}[\Psi_1(\theta_0)-\Psi_1(\theta_0+\beta)]^{1/2}}. \tag{5.6}$$

Proof. See the Appendix. □

Now, we study the accuracy of our bound for the MSE of the MLE by simulations. For the simulations, $\theta_0=1.5$, $\beta=1$ and in this case of β being equal to 1, the MLE is $\hat{\theta}_n(\mathbf{X})=-\frac{n}{\sum_{i=1}^n\log X_i}$.

Table 3. Part of the results taken by simulations from the Beta(1.5, 1) distribution

n	$\hat{\mathrm{E}}(\hat{\theta}_n(\mathbf{X}) - \theta_0)^2$	Upper bound	Error
7500	0.0002	0.2517	0.2515
7700	0.0002	0.0416	0.0414
7900	0.0002	0.0223	0.0221
8100	0.0002	0.0151	0.0149
8300	0.0002	0.0112	0.00110

We find that $n \geq 7460$, in order for (Fur.3) to be satisfied. The process to simulate is quite simple. Let $n \in \{7460, 7461, \dots, 8459\}$ and for each n, start by generating 10 000 trials of n random independent observations, x, from the Beta distribution with parameter values as above. We evaluate the MLE, $\hat{\theta}_n(\mathbf{X})$, of the parameter in each trial, which in turn gives a vector of 10 000 values. Thus, for each n from 7460 to 8459, we evaluate the sample MSE, $\hat{\mathrm{E}}(\hat{\theta}_n(\mathbf{X}) - \theta_0)^2 = \frac{1}{10\,000}\sum_{i=1}^{10\,000}[\hat{\theta}_n(\mathbf{x})[i] - \theta_0]^2$ and compare it with its upper bound, $(\frac{B_3}{\sqrt{n}})^2$, where B_3 is given in (5.5). The difference between their values measures the error of our bound on the MSE. Part of the results from the simulations is shown in Table 3. The table indicates that the bound and the error decrease as the sample size increases, as expected, since the order of the upper bound for the MSE is $\frac{1}{n}$. In addition, it is reasonable that the smaller the sample size is, the larger the bound is. The bounds are considerably larger than the estimated MSE and they are not numerically sharp. In addition, because of the relatively strong requirement that $n \geq 7460$, these bounds on the MSE are more of theoretical interest.

Remarks. Several interesting paths lead from the work explained in this paper. When the dimension of the parameter is $d > 1$, Stein bounds are available in Chen *et al.* [3], which can be employed to get upper bounds related to the distribution of the MLE in a multi-parameter setting (work in progress). In addition, one of the main advantages of Stein's method is that it can be used in situations where dependence comes into play. Upper bounds on the distributional distance between the distribution of the MLE and the normal distribution in the case of dependent random variables are also work in progress.

Appendix: Some proofs

Proof of Lemma 2.1. Let $\varepsilon > 0$ and f a continuous increasing function with $f(m) \geq 0$ for $m > 0$. Then,

$$\begin{aligned}
\mathrm{E}\big[f(M)\big] &= \mathrm{E}\big[f(M)|M \leq \varepsilon\big]\mathbb{P}(M \leq \varepsilon) + \mathrm{E}\big[f(M)|M > \varepsilon\big]\mathbb{P}(M > \varepsilon) \\
&= \mathrm{E}\big[f(M)|M \leq \varepsilon\big]\big(1 - \mathbb{P}(M > \varepsilon)\big) + \mathrm{E}\big[f(M)|M > \varepsilon\big]\mathbb{P}(M > \varepsilon) \\
&= \mathrm{E}\big[f(M)|M \leq \varepsilon\big] + \mathbb{P}(M > \varepsilon)\big(\mathrm{E}\big[f(M)|M > \varepsilon\big] - \mathrm{E}\big[f(M)|M \leq \varepsilon\big]\big) \\
&\geq \mathrm{E}\big[f(M)|M \leq \varepsilon\big] \qquad \text{as } f(m) \text{ is increasing.}
\end{aligned}$$

□

Proof of Corollary 5.1. The probability density function is

$$f(x|\theta) = \frac{\Gamma(\theta+\beta)}{\Gamma(\theta)\Gamma(\beta)} x^{\theta-1}(1-x)^{\beta-1}, \tag{A.1}$$

with $\theta > 0$ and $x \in [0, 1]$. Hence

$$\begin{aligned} l(\theta; \mathbf{x}) &= n\big[\log\big(\Gamma(\theta+\beta)\big) - \log\big(\Gamma(\theta)\big) - \log\big(\Gamma(\beta)\big)\big] \\ &\quad + (\theta-1)\sum_{i=1}^{n} \log x_i + (\beta-1)\sum_{i=1}^{n} \log(1-x_i) \end{aligned} \tag{A.2}$$

and

$$\begin{aligned} l'(\theta; \mathbf{x}) &= n\big[\Psi(\theta+\beta) - \Psi(\theta)\big] + \sum_{i=1}^{n} \log x_i \\ l^{(j)}(\theta; \mathbf{x}) &= n\big(\Psi_{j-1}(\theta+\beta) - \Psi_{j-1}(\theta)\big), \qquad j \in \mathbb{N} \setminus \{1\}. \end{aligned}$$

Now we show that the conditions (R1)–(R4) and the assumptions (Fur.1)–(Fur.3) are satisfied. For (R1) it is obvious. As for (R2), the three times differentiability of the density function can be verified from (A.2). In addition, using (A.1) and the expressions for the logarithmic expectations of a Beta distributed random variable, it is straightforward to verify $\int_0^1 \frac{\mathrm{d}^j}{\mathrm{d}\theta^j} f(x|\theta)\,\mathrm{d}x = \frac{\mathrm{d}^j}{\mathrm{d}\theta^j}\int_0^1 f(x|\theta)\,\mathrm{d}x = 0$, $j \in \{1, 2, 3\}$ for (R2). Let $\varepsilon = \varepsilon(\theta_0) > 0$ such that $\theta \in (\theta_0-\varepsilon, \theta_0+\varepsilon) \subset \Theta$. Since in this case $\Theta = (0, \infty)$, indeed $0 < \varepsilon < \theta_0$. Using a first order Taylor expansion and the fact that

$$\Psi_m(z) = (-1)^{m+1} m! \sum_{k=0}^{\infty} \frac{1}{(z+k)^{m+1}} \tag{A.3}$$

gives

$$\Psi_3(z) = 6\sum_{k=0}^{\infty} \frac{1}{(z+k)^4} \qquad \text{for } z \in \mathbb{C} \setminus \big\{\mathbb{Z}^-\big\} \text{ and } m > 0,$$

with $\Psi_3(z)$ being a decreasing function of z. For $\theta \in (\theta_0 - \varepsilon, \theta_0 + \varepsilon)$,

$$\begin{aligned} \left|\frac{\mathrm{d}^3}{\mathrm{d}\theta^3} \log f(x|\theta)\right| &= \big|\Psi_2(\theta+\beta) - \Psi_2(\theta)\big| \\ &\leq \beta\big|\Psi_3\big(\theta^*\big)\big| \leq \beta\big|\Psi_3(\theta_0-\varepsilon)\big| = M(x), \end{aligned} \tag{A.4}$$

with $\mathrm{E}[M(X)] < \infty$. Hence, (R3) holds as well. Also, $i(\theta_0) = \Psi_1(\theta_0) - \Psi_1(\theta_0+\beta)$ which is positive since it is obvious from (A.3) that $\Psi_1(z)$ is a decreasing function. The assumption (Fur.1) obviously holds with $\|x\| \leq 1$. Using (A.4) and the fact that $\sum_{i=1}^{\infty} \frac{1}{i^4} = \frac{\pi^4}{90} < 1.1$ gives

$$\begin{aligned} \sup_{\theta:|\theta-\theta_0|\leq\varepsilon} \left|l^{(3)}(\theta; X_1)\right| &\leq \beta\left|\Psi_3(\theta_0 - \varepsilon)\right| = 6\beta \sum_{k=0}^{\infty} \frac{1}{(\theta_0 - \varepsilon + k)^4} \\ &\leq 6\beta\left[\frac{1}{(\theta_0 - \varepsilon)^4} + \sum_{k=1}^{\infty} \frac{1}{k^4}\right] < \frac{6\beta}{(\theta_0 - \varepsilon)^4} + 6.6\beta = C_1. \end{aligned} \tag{A.5}$$

Thus, (Fur.2) is also satisfied. Now, since $i(\theta_0) = \Psi_1(\theta_0) - \Psi_1(\theta_0 + \beta)$ take

$$n \geq \frac{[C_1\varepsilon + \sqrt{(C_1\varepsilon)^2 + 8[\Psi_1(\theta_0) - \Psi_1(\theta_0 + \beta)]^2}]^2}{4\varepsilon^2[\Psi_1(\theta_0) - \Psi_1(\theta_0 + \beta)]^3}$$

in order for (Fur.3) to be satisfied. To find B_3, firstly, as $\mathrm{E}|l'(\theta_0; X_1)|^3$ is not straightforward to evaluate due to the absolute value in the expectation, it is easily seen that using Hölder's inequality $\mathrm{E}|l'(\theta_0; X_1)|^3 \leq [\mathrm{E}(l'(\theta_0; X_1))^4]^{3/4}$ we find an upper bound for

$$\begin{aligned} \mathrm{E}\big[l'(\theta_0; X_1)\big]^4 &= \mathrm{E}\big[\log X_1 + \Psi(\theta_0 + \beta) - \Psi(\theta_0)\big]^4 \\ &= \mathrm{E}\big[\log X_1 - \mathrm{E}(\log X_1)\big]^4. \end{aligned}$$

If $G_1 \sim \Gamma(\theta_0, \lambda)$ and $G_2 \sim \Gamma(\beta, \lambda)$ independent, then $\frac{G_1}{G_1+G_2} \sim \mathrm{Beta}(\theta_0, \beta)$. Thus, with $X_1 = \frac{G_1}{G_1+G_2}$

$$\begin{aligned} \mathrm{E}\big[l'(\theta_0; X_1)\big]^4 &= \mathrm{E}\big[\big(\log G_1 - \mathrm{E}[\log G_1]\big) + \big(\mathrm{E}\big[\log(G_1 + G_2)\big] - \log(G_1 + G_2)\big)\big]^4 \\ &\leq 8\big[\mathrm{E}\big(\log G_1 - \mathrm{E}(\log G_1)\big)^4 + \mathrm{E}\big(\log(G_1 + G_2) - \mathrm{E}\big(\log(G_1 + G_2)\big)\big)^4\big]. \end{aligned} \tag{A.6}$$

Now we calculate the fourth central moment of the logarithm of a Gamma distributed random variable. Using that $\int_0^{\infty} \frac{z^{\alpha-1}\mathrm{e}^{-z}(\log z)^k}{\Gamma(\alpha)}\,\mathrm{d}z = \frac{\Gamma^{(k)}(\alpha)}{\Gamma(\alpha)}$, for any $\alpha > 0$ and $k \in \mathbb{N}$ gives that for $Y \sim \Gamma(\alpha, \lambda)$

$$\mathrm{E}(\log Y) = \Psi(\alpha) - \log\lambda.$$

Using again $z = \lambda y$,

$$\begin{aligned} \mathrm{E}\big[\log Y - \mathrm{E}(\log Y)\big]^4 &= \int_0^{\infty} \frac{z^{\alpha-1}\mathrm{e}^{-z}}{\Gamma(\alpha)}\left(\log\left(\frac{z}{\lambda}\right) - \mathrm{E}\left(\log\left(\frac{Z}{\lambda}\right)\right)\right)^4 \mathrm{d}z \\ &= \int_0^{\infty} \frac{z^{\alpha-1}\mathrm{e}^{-z}}{\Gamma(\alpha)}\big(\log z - \mathrm{E}(\log Z)\big)^4 \mathrm{d}z \\ &= \frac{1}{\Gamma(\alpha)} \sum_{k=0}^{4} \binom{4}{k} (-1)^k\big[\Psi(\alpha)\big]^{4-k} \int_0^{\infty} z^{\alpha-1}\mathrm{e}^{-z}(\log z)^k\,\mathrm{d}z \\ &= -3\big[\Psi(\alpha)\big]^4 + 6\big[\Psi(\alpha)\big]^2\frac{\Gamma''(\alpha)}{\Gamma(\alpha)} - 4\Psi(\alpha)\frac{\Gamma^{(3)}(\alpha)}{\Gamma(\alpha)} + \frac{\Gamma^{(4)}(\alpha)}{\Gamma(\alpha)}. \end{aligned}$$

At this point, the digamma function can be used in order to simplify the expression above. Following simple steps it can be easily verified that

$$\frac{\Gamma''(\alpha)}{\Gamma(\alpha)} = \Psi_1(\alpha) + \big[\Psi(\alpha)\big]^2, \qquad \frac{\Gamma^{(3)}(\alpha)}{\Gamma(\alpha)} = \Psi_2(\alpha) + 3\Psi(\alpha)\Psi_1(\alpha) + \big[\Psi(\alpha)\big]^3,$$

$$\frac{\Gamma^{(4)}(\alpha)}{\Gamma(\alpha)} = \Psi_3(\alpha) + 4\Psi_2(\alpha)\Psi(\alpha) + 6\Psi_1(\alpha)\big[\Psi(\alpha)\big]^2 + 3\big[\Psi_1(\alpha)\big]^2 + \big[\Psi(\alpha)\big]^4.$$

Hence for $Y \sim \Gamma(\alpha, \lambda)$

$$\mathrm{E}\big[\log Y - \mathrm{E}(\log Y)\big]^4 = \Psi_3(\alpha) + 3\big[\Psi_1(\alpha)\big]^2$$

and therefore, from (A.6),

$$\mathrm{E}\big[l'(\theta_0; X_1)\big]^4 \le 8\big(\Psi_3(\theta_0) + \Psi_3(\theta_0+\beta) + 3\big[\Psi_1(\theta_0)\big]^2 + 3\big[\Psi_1(\theta_0+\beta)\big]^2\big) = B_1.$$

With C_1 as in (A.5), taking $\varepsilon = \frac{\theta_0}{2}$, we conclude that

$$\sup_{\theta:|\theta-\theta_0|\le\varepsilon} \big|l^{(3)}(\theta; X_1)\big| \le \frac{96\beta}{\theta_0^4} + 6.6\beta = B_2.$$

Using (A.2), gives

$$\mathrm{Var}\big(l''(\theta_0; X_1)\big) = \mathrm{Var}\big(\Psi_1(\theta_0+\beta) - \Psi_1(\theta_0)\big) = 0.$$

Having found all the necessary quantities, we calculate the upper bound in (5.1) and multiply it by $\sqrt{n}$. This is equal to B_3 shown in (5.5), which is an upper bound for $\sqrt{n\mathrm{E}(\hat{\theta}_n(\mathbf{X}) - \theta_0)^2}$ in the specific case of i.i.d. random variables from the Beta distribution. Using this bound in (5.2) gives the result in (5.6). □

Acknowledgements

The authors would like to thank Robert E. Gaunt for various insightful comments and for his idea regarding a perturbation to treat the Poisson example. This was the motivation behind the necessity to generalize the perturbation idea in Section 4 for any discrete distribution. We would also like to thank Larry Goldstein for his comments and Adrian Röllin for suggesting the Poisson example. In addition, the authors sincerely thank two anonymous reviewers for suggestions that lead to an improvement of the paper. Andreas Anastasiou was supported by a Teaching Assistantship Bursary from the Department of Statistics, University of Oxford, and an Engineering and Physical Sciences Research Council (EPSRC) Scholarship. Gesine Reinert was supported in part by EPSRC grant EP/K032402/1.

References

[1] Berk, R.H. (1972). Consistency and asymptotic normality of MLE's for exponential models. *Ann. Mat. Statist.* **43** 193–204. MR0298810

[2] Casella, G. and Berger, R.L. (2002). *Statistical Inference*, 2nd ed. Pacific Grove, CA: Duxbury Press.

[3] Chen, L.H.Y., Goldstein, L. and Shao, Q.-M. (2011). *Normal Approximation by Stein's Method*. *Probability and Its Applications* (*New York*). Heidelberg: Springer. MR2732624

[4] Cox, D.R. and Snell, E.J. (1968). A general definition of residuals. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **30** 248–275. MR0237052

[5] Fisher, R.A. (1925). In theory of statistical estimation. *Mathematical Proceedings of the. Cambridge Philosophical Society* **22** 700–725.

[6] Geyer, C.J. (2013). Asymptotics of maximum likelihood without the LLN or CLT or sample size going to infinity. *Institute of Mathematical Statistics* **10** 1–24.

[7] Hoadley, B. (1971). Asymptotic properties of maximum likelihood estimators for the independent not identically distributed case. *Ann. Math. Statist.* **42** 1977–1991. MR0297051

[8] Kendall, M.G. and Stuart, A. (1969). *The Advanced Theory of Statistics*, *Volume* 1 *Distribution Theory* 3rd ed. London: Charles Griffin and Company Limited.

[9] Lauritzen, S.L. (1988). *Extremal Families and Systems of Sufficient Statistics*. *Lecture Notes in Statistics* **49**. New York: Springer. MR0971253

[10] Mäkeläinen, T., Schmidt, K. and Styan, G.P.H. (1981). On the existence and uniqueness of the maximum likelihood estimate of a vector-valued parameter in fixed-size samples. *Ann. Statist.* **9** 758–767. MR0619279

[11] Nourdin, I. and Peccati, G. (2012). *Normal Approximations with Malliavin Calculus*. *Cambridge Tracts in Mathematics* **192**. Cambridge: Cambridge Univ. Press. MR2962301

[12] Rachev, S.T. (1991). *Probability Metrics and the Stability of Stochastic Models*. *Wiley Series in Probability and Mathematical Statistics*: *Applied Probability and Statistics*. Chichester: Wiley. MR1105086

[13] Reinert, G. (1998). Couplings for normal approximations with Stein's method. In *Microsurveys in Discrete Probability* (*Princeton*, *NJ*, 1997) (D. Aldous and J. Propp, eds.). *DIMACS Ser. Discrete Math. Theoret. Comput. Sci.* **41** 193–207. Providence, RI: Amer. Math. Soc. MR1630415

[14] 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* (*Univ. California*, *Berkeley*, *Calif.*, 1970/1971), *Vol. II*: *Probability Theory* 583–602. Berkeley, CA: Univ. California Press. MR0402873

Received December 2014 and revised June 2015