*The Annals of Statistics*
1983, Vol. 11, No. 2, 498–504

# ON MODERATE DEVIATION THEORY IN ESTIMATION

BY WILBERT C. M. KALLENBERG

*Vrije Universiteit*

The performance of a sequence of estimators $\{T_n\}$ of $\theta$ can be measured by the probability concentration of the estimator in an $\varepsilon_n$-neighborhood of $\theta$. Classical choices of $\varepsilon_n$ are $\varepsilon_n = cn^{-1/2}$ (contiguous case) and $\varepsilon_n = \varepsilon$ fixed for all $n$ (non-local case). In this article all sequences $\{\varepsilon_n\}$ with $\lim_{n\to\infty} \varepsilon_n = 0$ and $\lim_{n\to\infty} \varepsilon_n n^{1/2} = \infty$ are considered. In that way the statistically important choices of small $\varepsilon$'s are investigated in a uniform sense; in that way the importance and usefulness of classical results concerning local or non-local efficiency can gather strength by extending to larger regions of neighborhoods; in that way one can investigate where optimality passes into non-optimality if for instance an estimator is locally efficient and non-locally non-efficient. The theory of moderate deviation and Cramér-type large deviation probabilities plays an important role in this context. Examples of the performance of particularly maximum likelihood estimators are presented in $k$-parameter exponential families, a curved exponential family and the double-exponential family.

**1. Introduction.** The performance of an estimator $T$ of an unknown real valued parameter $\theta$ can be measured by

$$a(\varepsilon, \theta, T) = P_\theta\{|T - \theta| > \varepsilon\},$$

the probability dispersion of the estimator outside an $\varepsilon$-neighborhood of $\theta$. Unfortunately, in most practical cases $a(\varepsilon, \theta, T)$ can not be handled exactly and therefore one or another asymptotic approach is exploited.

Two frequently applied approaches are as follows. Let $\{T_n^{(1)}\}$ and $\{T_n^{(2)}\}$ denote two competing sequences of estimators where $T_n^{(i)}$ is based on $n$ observations, $i = 1, 2$. If $T_n^{(i)}$ is asymptotically normal $N(\theta, n^{-1}\sigma_i^2(\theta))$ as $n \to \infty$, then for all $c > 0$

$$\lim_{n\to\infty} a(n^{-1/2}c, \theta, T_n^{(i)}) = 2\Phi(-c/\sigma_i(\theta)), \qquad i = 1, 2, \tag{1.1}$$

where $\Phi$ denotes the standard normal distribution function. In that case the asymptotic relative efficiency of $\{T_n^{(1)}\}$ w.r.t. $\{T_n^{(2)}\}$ equals $\sigma_2^2(\theta)/\sigma_1^2(\theta)$. If $\sigma_2^2(\theta)/\sigma_1^2(\theta) > 1$ the probability dispersion of $T_n^{(1)}$ is asymptotically smaller than the probability dispersion of $T_n^{(2)}$, when sequences of $\varepsilon_n$-neighborhoods are considered with $\varepsilon_n = cn^{-1/2}$.

For fixed $\varepsilon > 0$ however, the probability dispersion $a(\varepsilon, \theta, T_n)$ tends to zero as $n \to \infty$ if $\{T_n\}$ is consistent. In typical cases the convergence is exponentially fast. Therefore the second approach is to measure the performance of $\{T_n\}$ by the inaccuracy rate

$$e(\varepsilon) = \lim_{n\to\infty} -n^{-1}\log a(\varepsilon, \theta, T_n). \tag{1.2}$$

The two approaches mentioned above can also be discussed in terms of sample sizes needed for the two sequences of estimators to "perform equivalently"; cf. Serfling (1980), Section 1.15.4. From a practical and philosophical point of view it seems rather peculiar to consider in such approaches either sequences $\{\varepsilon_n\}$ with $\varepsilon_n$ of order $n^{-1/2}$ or sequences $\{\varepsilon_n\}$ with $\varepsilon_n = \varepsilon$ is fixed. It is interesting to investigate $a(\varepsilon_n, \theta, T_n)$ for sequences $\{\varepsilon_n\}$ with

$$\lim_{n\to\infty} \varepsilon_n = 0 \quad \text{and} \quad \lim_{n\to\infty} \varepsilon_n n^{1/2} = \infty, \tag{1.3}$$

Received November 1981; revised October 1982.

*AMS* 1970 *subject classification.* Primary 62F20; secondary 62F10, 60F10.

*Key words and phrases.* First and second order efficiency; moderate and Cramér-type large deviations; probability concentration; maximum likelihood estimator.

in that way filling the gap between the two well-known extremes. By doing so a variety of statistically important options of $\varepsilon$'s are supplied, which enables the statistician to see what happens if the comparison is based on $\varepsilon$-neighborhoods intermediate between relatively small ($\varepsilon = cn^{-1/2}$) or relatively large ($\varepsilon$ fixed for all $n$) neighborhoods. Moreover, there are examples of estimators which are optimal w.r.t. the local criterion of asymptotic variances and non-optimal in the non-local case of fixed $\varepsilon$. One example is the sample median in estimating the location parameter of a double exponential distribution, cf. Example 2.2 (continued). It is interesting to investigate where the optimality passes into non-optimality when one runs from the contiguous case to the non-local case.

It is shown by Bahadur and Fu (1975) that, whatever the consistent sequence $\{T_n\}$, $a(\varepsilon, \theta, T_n)$ cannot tend to zero at a rate faster than a certain exponential rate; that is, there exists an upper bound for $e(\varepsilon)$, cf. (1.2). Partly for technical reasons (the large deviation probabilities involved are rather hard to handle, except for maximum likelihood estimators in exponential families, cf. Kester (1981)), and partly because of practical considerations (statistically $\varepsilon$ must be small) comparison with an expansion of the upper bound is studied as $\varepsilon \to 0$.

The higher order contact the exponential rate of an estimator has with the expansion of the upper bound as $\varepsilon \to 0$, the better the estimator. This concept suggests to be an optimality criterion for the case of small $\varepsilon$ and large $n$. However here is a uniformity problem. It is not sure that taking first the limit w.r.t. $n$ (to $\infty$) and afterwards the limit w.r.t. $\varepsilon$ (to 0) yields the same as sending $\varepsilon$ and $n$ *simultaneously* to 0 and $\infty$. This is only valid if the remainder terms, which disappear by taking limits, are small *uniformly* in $\varepsilon$ and $n$. Therefore if one wants to investigate the case of small $\varepsilon$ and large $n$ rigorously, it is more natural to derive *directly* an upper bound for $-n^{-1}\log a(\varepsilon_n, \theta, T_n)$ with $\varepsilon_n$ in the range of intermediate $\varepsilon$'s and to make comparison with *this* upper bound, which is adjusted to the range of $\varepsilon$'s under consideration. Accomplishing this for all sequences $\{\varepsilon_n\}$ satisfying (1.3), the region of small but not contiguous $\varepsilon$-neighborhoods is covered in a uniform sense. Especially in studying second order Bahadur efficiency (Fu, 1982) the classical indirect approach and the here proposed direct approach are different, see also Remark 2.1.

In addition another phenomenon has to be mentioned. Suppose that an estimator is efficient in the sense of asymptotic variances, and also efficient in the sense of achieving the upper bound for all sequences $\{\varepsilon_n\}$ satisfying (1.3). Then the local optimality is extended to a much larger class of $\varepsilon$-neighborhoods. The importance and usefulness of local optimality gathers strength by such a result. A similar statement with respect to non-local optimality can be made.

Apart from studying optimality one may also wish to compare two estimators: in Section 2 a method of comparison is presented. The performance of particularly maximum likelihood (m.l.) estimators is investigated in the examples in Section 2. Three examples are presented: (1) It is shown that the m.l. estimator is (second order) optimal in full exponential families (Example 2.1 (continued)); (2) In curved exponential families the m.l. estimator is typically not second order optimal (due to non-convexity); this is exemplified in Example 2.2 (continued); (3) Example 2.3 (continued) concerns the double-exponential family as a famous example of local optimality and non-local non-optimality of the m.l. estimator.

The role of central limit theorems in the contiguous case (cf. (1.1)) and Chernoff-type large deviation theorems in the non-local case (cf. (1.2)) is here taken over by moderate and Cramér-type large deviation theorems. The moderate deviation theory approach in statistics is not new; see for instance Rubin and Sethuraman (1965), Johnson and Truax (1974, 1978), Groeneboom (1980, Section 3.4), Kallenberg (1983). Although a lot of moderate and Cramér-type large deviation theorems are available (for a recent review see Vandemaele, 1981), the domain $cn^{1/6} < n^{1/2}\varepsilon_n = o(n^{1/2})$ is as yet rather unexploited.

**2. Results and examples.** In Bahadur, Gupta and Zabell (1980) a lower bound for large deviation probabilities is presented. Here we give a slightly different version of their theorem adjusted to our situation.

Let $S$ be a space of points $s$, $\mathscr{A}$ a $\sigma$-field of subsets of $S$. For each $n = 1, 2, \cdots$ let $\mathscr{B}_n$ be a $\sigma$-field such that $\mathscr{B}_n \subset \mathscr{A}$, $n = 1, 2, \cdots$. For each probability measure $P$ we denote by $P_n$ the inner measure on $S$ determined by the restriction of $P$ to $\mathscr{B}_n$, i.e. $P_n(B) = \sup\{P(C): C \in \mathscr{B}_n, C \subset B\}$ for all $B \subset S$. Note that if $B \in \mathscr{B}_n$ then $P_n(B) = P(B)$. For a more extensive description of this framework cf. Section 2 of Bahadur, Gupta and Zabell (1980).

PROPOSITION 2.1. *Let $\{A_n\}$ be a sequence of subsets of $S$ and let $P$, $Q^{(n)}$, $n = 1, 2, \cdots$, be probability measures on $\mathscr{A}$ such that $Q^{(n)} \ll P$ on $\mathscr{B}_n$. Denote by $r_n(s)$ a $\mathscr{B}_n$-measurable function such that $0 \le r_n(s) < \infty$ and such that $dQ^{(n)}(s) = r_n(s)\, dP(s)$ on $\mathscr{B}_n$. Define $B_n(t) = \{s: r_n(s) > e^{nt}\}$, then*

$$\liminf_{n\to\infty} P_n(A_n)e^{nt_n} \ge \liminf_{n\to\infty}\{Q^{(n)}(A_n) - Q^{(n)}(B_n(t_n))\} \tag{2.1}$$

*for all sequences $\{t_n\}$. In particular if $\liminf_{n\to\infty} Q^{(n)}(A_n) > \limsup_{n\to\infty} Q^{(n)}(B_n(t_n))$, then*

$$\liminf_{n\to\infty} P_n(A_n)e^{nt_n} > 0. \tag{2.2}$$

Because of the above mentioned correspondence the proof of Proposition 2.1 is omitted. Proposition 2.1 will be applied to obtain a suitable upper bound for the probability dispersion of estimators. Next the considerations of Section 1 are made more precise by studying the following estimation problem.

Let the parameter space $\Theta$ be a subset of $\mathbb{R}^k$ (unlike the situation in Section 1, $k$ may be greater than 1). Let $S = (X_1, X_2, \cdots)$ be a sequence of i.i.d. random elements in the probability space $(\mathscr{X}, \mathscr{A}, P_\theta)$, $\theta \in \Theta$. It is assumed that $P_\theta \ne P_{\theta'}$ if $\theta \ne \theta'$. The distribution of $S$ is denoted by $\mathbb{P}_\theta$, when $\theta \in \Theta$ obtains. For each $n = 1, 2, \cdots$ let $T_n$ be a measurable function of $S$ into $\mathbb{R}^k$ depending on $X_1, \cdots, X_n$, to be thought of as an estimator of $\theta$. An important role is played by the Kullback Leibler information number defined by

$$K(\theta, \theta_0) = \begin{cases} E_\theta \log\,(dP_\theta/dP_{\theta_0}) & \text{if } P_\theta \ll P_{\theta_0}, \\ \infty & \text{otherwise.} \end{cases} \tag{2.3}$$

Let $\theta_0 \in \operatorname{int}\Theta$ be fixed. The behavior of $T_n$ in a neighborhood of $\theta_0$ will be discussed.

Suppose that the following regularity assumptions are fulfilled.

CONDITION 1.

$$K(\theta, \theta_0) \to 0 \Leftrightarrow \theta \to \theta_0. \tag{2.4}$$

CONDITION 2.

$$0 < A_1(\theta_0) = \liminf_{\theta\to\theta_0} \frac{K(\theta, \theta_0)}{\|\theta - \theta_0\|^2} \le \limsup_{\theta\to\theta_0} \frac{K(\theta, \theta_0)}{\|\theta - \theta_0\|^2} = A_2(\theta_0) < \infty. \tag{2.5}$$

CONDITION 3. For each sequence $\{\theta_n\}$ such that $K(\theta_n, \theta_0) \to 0$ and $nK(\theta_n, \theta_0) \to \infty$ (i.e. $\theta_n \to \theta_0$ and $n\|\theta_n - \theta_0\|^2 \to \infty$)

$$\frac{\sum_{t=1}^n \{\log(dP_{\theta_n}/dP_{\theta_0})(X_t)\} - nK(\theta_n, \theta_0)}{\{2nK(\theta_n, \theta_0)\}^{1/2}} \xrightarrow{D_{\theta_n}} U, \tag{2.6}$$

where $\xrightarrow{D_{\theta_n}}$ denotes convergence in distribution under $\mathbb{P}_{\theta_n}$, and $U$ has a standard normal distribution.

Note that $E_{\theta_n}\log(dP_{\theta_n}/dP_{\theta_0})(X_1) = K(\theta_n, \theta_0)$ and that in regular cases $\operatorname{Var}_{\theta_n}\log(dP_{\theta_n}/dP_{\theta_0})(X_1) \approx (\theta_n - \theta_0)'J_0(\theta_n - \theta_0) \approx 2K(\theta_n, \theta_0)$. Here $J_0$ is the Fisher information matrix at $\theta_0$. Further it has to be noted that Conditions 1, 2 and 3 are in terms of the family of distributions and not in terms of the estimators.

EXAMPLE 2.1. Let $\{P_\theta: \theta \in \Theta\}$ be a full $k$-parameter exponential family with canonical parametrization. If $\theta_0$ is an interior point of the natural parameter space then it is easily seen that Conditions 1, 2 and 3 are fulfilled.

EXAMPLE 2.2. Consider the bivariate normal distribution with covariance matrix $I$, the identity, and mean vector $(\theta, \frac{1}{2}\gamma_0\theta^2)$, $\theta \in \Theta = \mathbb{R}$, $\gamma_0$ a constant (cf. Efron, 1975). Conditions 1, 2 and 3 are fulfilled (the distribution of the left-hand side of (2.6) is exactly standard normal).

EXAMPLE 2.3. Consider the double exponential distribution

$$dP_\theta(x) = \tfrac{1}{2}\exp(-|x - \theta|),\ x, \theta \in \mathbb{R}.$$

Then

$$K(\theta, \theta_0) = |\theta - \theta_0| - 1 + e^{-|\theta - \theta_0|},\quad \theta, \theta_0 \in \mathbb{R}. \tag{2.7}$$

Again it is easily seen that Conditions 1, 2 and 3 are fulfilled.

EXAMPLE 2.4. Let $P_\theta$ denote the uniform $(0, \theta)$ distribution, $\theta > 0$. Then Conditions 1, 2 and 3 fail.

Define

$$\Delta(\varepsilon) = \{\theta \in \Theta : \|\theta - \theta_0\| > \varepsilon\} \tag{2.8}$$

and

$$K(\varepsilon) = \inf\{K(\theta, \theta_0) : \theta \in \Delta(e)\}. \tag{2.9}$$

THEOREM 2.2. *Let* $\{\varepsilon_n\}$ *be a sequence of real numbers with* $\lim_{n\to\infty}\varepsilon_n = 0$, $\lim_{n\to\infty} \varepsilon_n n^{1/2} = \infty$. *If Conditions* 1, 2 *and* 3 *are fulfilled, then we have for all* $u \in \mathbb{R}$

$$\begin{aligned}\liminf_{n\to\infty}&\mathbb{P}_{\theta_0}(\|T_n - \theta_0\| > \varepsilon_n)\exp\{nK(\varepsilon_n) + u\sqrt{2nK(\varepsilon_n)}\} \\ &\geq \liminf_{n\to\infty}\inf_{\theta\in\Delta(\varepsilon_n)}\mathbb{P}_\theta(\|T_n - \theta_0\| > \varepsilon_n) - \Phi(-u).\end{aligned} \tag{2.10}$$

PROOF. Since $\theta_0 \in \operatorname{int}\Theta$ and $\varepsilon_n \to 0$ it follows by (2.4), (2.8) and (2.9) that $K(\varepsilon_n) \to 0$. Let $\theta_n \in \Delta(\varepsilon_n)$ satisfy $K(\theta_n, \theta_0) < K(\varepsilon_n) + n^{-2}$. Then $K(\theta_n, \theta_0) \to 0$ and by (2.4) $\theta_n \to \theta_0$; hence by (2.5) $nK(\theta_n, \theta_0) \to \infty$, because $n\varepsilon_n^2 \to \infty$. In view of Condition 3, application of Proposition 2.1 with $P = \mathbb{P}_{\theta_0}$, $A_n = \{s : \|T_n(s) - \theta_0\| > \varepsilon_n\}$, $Q^{(n)} = \mathbb{P}_{\theta_n}$ and $t_n = K(\theta_n, \theta_0)[1 + u/\{\frac{1}{2}nK(\theta_n, \theta_0)\}^{1/2}]$ yields

$$\begin{aligned}\liminf_{n\to\infty}&\mathbb{P}_{\theta_0}(\|T_n - \theta_0\| > \varepsilon_n)e^{nK(\theta_n,\theta_0)+u\sqrt{2nK(\theta_n,\theta_0)}} \\ &\geq \liminf_{n\to\infty}\mathbb{P}_{\theta_n}(\|T_n - \theta_0\| > \varepsilon_n) - \Phi(-u).\end{aligned} \tag{2.11}$$

Since $K(\theta_n, \theta_0) < K(\varepsilon_n) + n^{-2}$ the desired result is obtained. □

The behavior of the inaccuracy function is studied for local but non-contiguous $\varepsilon$. Therefore we consider not the class of asymptotically normal estimators nor the class of consistent estimators, but something in between. Bahadur, Gupta and Zabell (1980) use the consistency to obtain that $\mathbb{P}_\theta(\|T_n - \theta\| > \varepsilon) \to 1$ if $\|\theta - \theta_0\| > \varepsilon$. So the class of consistent estimators is replaced by the class of estimators satisfying

$$\liminf_{n\to\infty}\inf_{\theta\in\Delta(\varepsilon_n)}\mathbb{P}_\theta(\|T_n - \theta_0\| > \varepsilon_n) > 0, \tag{2.12}$$

where $\varepsilon_n \to 0$ in such a way that $n\varepsilon_n^2 \to \infty$ as $n \to \infty$. For this class of estimators a lower bound for the probability dispersion is provided by the following corollary.

COROLLARY 2.3. *Let* $\{\varepsilon_n\}$ *be a sequence of real numbers with* $\lim_{n\to\infty}\varepsilon_n = 0$, $\lim_{n\to\infty} \varepsilon_n n^{1/2} = \infty$. *If Conditions* 1, 2 *and* 3 *are fulfilled, then for all sequences of estimators* $\{T_n\}$ *satisfying* (2.12) *we have*

$$\mathbb{P}_{\theta_0}(\|T_n - \theta_0\| > \varepsilon_n) > e^{-nK(\varepsilon_n) - O(1)\sqrt{nK(\varepsilon_n)}} \quad \text{as } n \to \infty. \tag{2.13}$$

Because the corollary immediately follows from Theorem 2.2 the proof is omitted.

Next we discuss how to compare sequences of estimators and how to define optimality in this context. Let $\{T_n^{(1)}\}$ and $\{T_n^{(2)}\}$ denote two competing sequences of estimators, where $T_n^{(i)}$ is based on $n$ observations, $i = 1, 2$. In typical cases we have for all sequences $\{\varepsilon_n\}$ satisfying (1.3)

$$\frac{-\log \mathbb{P}_{\theta_0}(\| T_n^{(i)} - \theta_0 \| > \varepsilon_n)}{nK(\varepsilon_n)} \to c_i(\theta_0) \quad \text{as } n \to \infty,\ i = 1, 2. \tag{2.14}$$

In such a case the two estimators may be said to perform "first order equivalently" at respective sample sizes $n_1$ and $n_2$ satisfying

$$\frac{K(\varepsilon_{n_1})^{-1}\log \mathbb{P}_{\theta_0}(\| T_{n_1}^{(1)} - \theta_0 \| > \varepsilon_{n_1})}{K(\varepsilon_{n_2})^{-1}\log \mathbb{P}_{\theta_0}(\| T_{n_2}^{(2)} - \theta_0 \| > \varepsilon_{n_2})} \to 1.$$

In this case

$$\frac{n_2}{n_1} \sim \frac{-(n_1K(\varepsilon_{n_1}))^{-1}\log \mathbb{P}_{\theta_0}(\| T_{n_1}^{(1)} - \theta_0 \| > \varepsilon_{n_1})}{-(n_2K(\varepsilon_{n_2}))^{-1}\log \mathbb{P}_{\theta_0}(\| T_{n_2}^{(2)} - \theta_0 \| > \varepsilon_{n_2})} \to \frac{c_1(\theta_0)}{c_2(\theta_0)}, \tag{2.15}$$

yielding $c_1(\theta_0)/c_2(\theta_0)$ as a measure of first order asymptotic relative efficiency of $\{T_n^{(1)}\}$ relative to $\{T_n^{(2)}\}$. If $c_1(\theta_0) > c_2(\theta_0)$, then $\{T_n^{(1)}\}$ is better than $\{T_n^{(2)}\}$ because asymptotically the same result is obtained with less observations for $\{T_n^{(1)}\}$ than for $\{T_n^{(2)}\}$. In view of (2.13) an estimator $\{T_n\}$ is called *first order asymptotically optimal* if

$$\frac{-\log \mathbb{P}_{\theta_0}(\| T_n - \theta_0 \| > \varepsilon_n)}{nK(\varepsilon_n)} \to 1 \quad \text{as } n \to \infty. \tag{2.16}$$

If the estimators $\{T_n^{(1)}\}$ and $\{T_n^{(2)}\}$ are first order equivalent, i.e. $c_1(\theta_0) = c_2(\theta_0)$, then $\{T_n^{(1)}\}$ is at second order asymptotically better than $\{T_n^{(2)}\}$ if

$$\frac{-\log \mathbb{P}_{\theta_0}(\| T_n^{(1)} - \theta_0 \| > \varepsilon_n) + \log \mathbb{P}_{\theta_0}(\| T_n^{(2)} - \theta_0 \| > \varepsilon_n)}{\{nK(\varepsilon_n)\}^{1/2}} \to \infty. \tag{2.17}$$

In view of Corollary 2.3 a sequence $\{T_n\}$ is called *second order asymptotically optimal* if

$$\limsup_{n\to\infty} \frac{nK(\varepsilon_n) + \log \mathbb{P}_{\theta_0}(\| T_n - \theta_0 \| > \varepsilon_n)}{\{nK(\varepsilon_n)\}^{1/2}} < \infty. \tag{2.18}$$

EXAMPLE 2.1 (continued). For the maximum likelihood (m.l.) estimator $\hat{\theta}_n$ we have for some constant $c > 0$

$$\mathbb{P}_{\theta_0}(\| \hat{\theta}_n - \theta_0 \| > \varepsilon_n) \le \mathbb{P}_{\theta_0}(K(\hat{\theta}_n, \theta_0) \ge K(\varepsilon_n)) \le c\{nK(\varepsilon_n)\}^{(1/2)(k-2)}\exp\{-nK(\varepsilon_n)\};$$

c.f. Lemma 3.2 in Kallenberg (1981). Hence the m.l. estimator is (first and) second order optimal in $k$-parameter exponential families,

EXAMPLE 2.2 (continued). Denote the sample by $(X_1, Y_1), \cdots, (X_n, Y_n)$, where $X_i$ and $Y_i$ are independent and where their distributions are normal $N(\theta, 1)$ and $N(\frac{1}{2}\gamma_0\theta^2, 1)$, respectively. Consider the m.l. estimator $\hat{\theta}_n$ and the estimator $\bar{X}_n = n^{-1}\sum_{i=1}^n X_i$. Note that

$$|\hat{\theta}_n - \theta_0| \le \varepsilon \Leftrightarrow \begin{cases} \bar{Y}_n \le -[\gamma_0(\theta_0 + \varepsilon)]^{-1}[\bar{X}_n - (\theta_0 + \varepsilon)] + \frac{1}{2}\gamma_0(\theta_0 + \varepsilon)^2 \\ \text{and} \\ \bar{Y}_n \ge -[\gamma_0(\theta_0 - \varepsilon)]^{-1}[\bar{X}_n - (\theta_0 - \varepsilon)] + \frac{1}{2}\gamma_0(\theta_0 - \varepsilon)^2 \end{cases}$$

and

$$K(\varepsilon_n) = \tfrac{1}{2}\varepsilon_n^2[1 + \theta_0^2\gamma_0^2 - \gamma_0^2\theta_0\varepsilon_n + \tfrac{1}{4}\gamma_0^2\varepsilon_n^2].$$

The following results hold: $\hat{\theta}_n$ is first order optimal for all $\theta_0$ (cf. (2.16)); $\bar{X}_n$ is first order optimal iff $\theta_0 = 0$; $\hat{\theta}_n$ is second order optimal for all $\theta_0$ if $\varepsilon_n n^{1/2} = O(n^{1/3})$, and $\bar{X}_n$ is second order optimal at $\theta_0 = 0$ if $\varepsilon_n n^{1/2} = O(n^{1/3})$, because the left-hand side of (2.18) $\sim (\sqrt{2}/8)\ \gamma_0^2 n^{1/2}\varepsilon_n^3$. So $\hat{\theta}_n$ is better than $\bar{X}_n$ except in the point $\theta_0 = 0$, where they are equivalent.

In contrast with the situation of a full exponential family, here the m.l. estimator is not second order optimal; it is due to the non-convexity of $\{(\theta, \tfrac{1}{2}\gamma_0\theta^2) : \theta \in \mathbb{R}\}$ as subset of the full parameter space $\mathbb{R}^2$ of the bivariate normal distribution. This is typical for curved exponential families.

EXAMPLE 2.3 (continued). For the m.l. estimator $M_n$ = median $(X_1, \cdots, X_n)$ it holds that

$$\log \mathbb{P}_{\theta_0}(|M_n - \theta_0| > \varepsilon_n) + nK(\varepsilon_n) = \tfrac{1}{3}\, n\varepsilon_n^3\{1 + O(\varepsilon_n)\},$$

while $K(\varepsilon_n) \sim \tfrac{1}{2}n\varepsilon_n^2$, cf. (2.7). So $M_n$ is first order optimal; it is second order optimal if $\varepsilon_n n^{1/2} = O(n^{1/4})$, but not second order optimal if $\varepsilon_n n^{1/4} \to \infty$.

It is well-known that the sample median is locally efficient and non-locally non-efficient (cf. Sievers, 1978). The result here shows that at first order the optimality remains true and that even second order optimality holds if $\varepsilon_n$ is not too large.

We conclude with some remarks.

REMARK 2.1. It is seen from Corollary 2.3 that $K(\varepsilon_n) + O(\sqrt{K(\varepsilon_n)/n})$ is an upper bound of $-n^{-1}\log \mathbb{P}_\theta(\| T_n - \theta\| > \varepsilon_n)$. If $\varepsilon_n = \varepsilon$ is kept fixed and the limit w.r.t. $n$ is taken, the $O$-term of course disappears. In typical cases $K(\varepsilon) = c_1\varepsilon^2 + c_2\varepsilon^3 + c_3\varepsilon^4 + \cdots$ as $\varepsilon \to 0$. Considering such an expansion up to and including $c_3\varepsilon^4$, say, as an approximation of the upper bound for small $\varepsilon$ and large $n$ without referring to the $O$-term seems not reasonable if $\varepsilon_n^3\sqrt{n} = O(1)$. Because in that case the $O$-term is of the same order as the term $c_3\varepsilon_n^4$ and can therefore not be omitted. It is seen that w.r.t. first order approximation problems such as these do not arise. However, in studying second order Bahadur efficiency (cf. Fu, 1982) the classical indirect approach and the here proposed direct approach are different.

REMARK 2.2. The sharpness of (2.13) is seen by the following example. Let $X_1, X_2, \cdots$ be i.i.d. normal $N(\theta, 1)$ random variables, where $\theta \in \mathbb{R}$ is unknown. Define $T_n = \{1 - A(\varepsilon_n n^{1/2})^{-1}\}\bar{X}_n$, where $A$ is a positive constant. Choosing $\theta_0 = 0$ we have $\mathbb{P}_0(|T_n| > \varepsilon_n) = 2\Phi(-\varepsilon_n\sqrt{n}/\{1 - A(\varepsilon_n n^{1/2})^{-1}\}) = \exp[-nK(\varepsilon_n) - \sqrt{2A\sqrt{nK(\varepsilon_n)}}\ \{1 + o(1)\}]$.

REMARK 2.3. Some obvious generalizations (with adjusted conditions) are possible. For instance, the parameter space $\Theta$ can be taken more general; estimating $\theta$ can be replaced by estimating $g(\theta)$, where $g$ is some (not necessarily $1 - 1$) function; the random variable $U$ in (2.6) may be any random variable with $\Pr(U < \infty) = 1$ without affecting Corollary 2.3; Conditions 1 and 2 can be left out if (2.6) holds for a sequence $\{\theta_n\}$ with $\theta_n \in \Delta(\varepsilon_n)$ and $K(\theta_n, \theta_0) - K(\varepsilon_n)$ sufficiently small.

REMARK 2.4. In effect Proposition 2.1 and therefore also Theorem 2.2, is an application of the Neyman-Pearson lemma, so it is not surprising and very natural that in Condition 3 the likelihood ratio $dP_{\theta_n}/dP_{\theta_0}$ appears. Verification of (2.6) can be done by using central limit theorems for a triangular array, cf. Serfling (1980, Section 1.9.3).

REMARK 2.5. If the class of estimators satisfying (2.12) is restricted to the class of uniformly asymptotically normally distributed estimators, that is, if

$$\{T_n - \theta_n\}\sqrt{n} \xrightarrow{D_{\theta_n}} U$$

where $U$ has a $k$-dimensional normal $N(0; \Sigma)$ distribution with some positive definite covariance matrix $\Sigma$, then (2.13) can be sharpened: $O(1)$ can be replaced by $o(1)$. For a proof cf. (2.11).

**Acknowledgment.** I am indebted to a referee for comments which led to a considerable improvement in presentation.

## REFERENCES

BAHADUR, R. R. (1967). Rates of convergence of estimates and test statistics. *Ann. Math. Statist.* **38** 303–324.

BAHADUR, R. R. (1971). *Some Limit Theorems in Statistics*. SIAM, Philadelphia.

BAHADUR, R. R., GUPTA, J. C. and ZABELL, S. L. (1980). Large deviations, tests and estimates. In *Asymptotic Theory of Statistical Tests and Estimation* (I. M. Chakravarti, ed.). Academic, New York, 33–64.

EFRON, B. (1975). Defining the curvature of a statistical problem. *Ann. Statist.* **3** 1189–1242.

FU, J. C. (1975). The rate of convergence of consistent point estimators. *Ann. Statist.* **3** 234–240.

FU, J. C. (1982). Large sample point estimation: a large deviation approach. *Ann. Statist.* **10** 762–771.

GROENEBOOM, P. (1980). *Large Deviations and Asymptotic Efficiencies*. Mathematical Centre Tracts 118, Amsterdam.

JOHNSON, B. R. and TRUAX, D. R. (1974). Asymptotic behavior of Bayes tests and Bayes Risk. *Ann. Statist.* **2** 278–294.

JOHNSON, B. R. and TRUAX, D. R. (1978). Asymptotic behavior of Bayes procedures for testing simple hypotheses in multiparameter exponential families. *Ann. Statist.* **6** 346–351.

KALLENBERG, W. C. M. (1981). Bahadur deficiency of likelihood ratio tests in exponential families. *J. Multivariate Anal.* **11** 506–531.

KALLENBERG, W. C. M. (1983). Intermediate efficiency, theory and examples. *Ann. Statist.*, to appear.

KESTER, A. D. M. (1981). Large deviation optimality of MLE's in exponential families. Report No. 160, Wiskundig Seminarium, Vrije Universiteit, Amsterdam.

RUBIN, H. and SETHURAMAN, J. (1965). Bayes risk efficiency. *Sankhyā Ser. A* **27** 347–356.

SERFLING, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley, New York.

SIEVERS, G. L. (1978). Estimates of Location: a large deviation comparison. *Ann. Statist.* **6** 610–618.

VANDEMAELE, M. (1981). Kansen op grote en matige afwijkingen voor $U$- en $L$-statistieken. Ph.D. thesis, Leuven (in Dutch).

MATHEMATICS DEPARTMENT
FREE UNIVERSITY
DE BOELELAAN 1081
1081 HV AMSTERDAM
THE NETHERLANDS