JOURNAL OF MULTIVARIATE ANALYSIS **57**, 191–214 (1996)
ARTICLE NO. 0029

Estimation of the Location of the Maximum of a Regression Function Using Extreme Order Statistics*

HUNG CHEN[†]

Department of Mathematics, National Taiwan University,
Taipei, Taiwan 10764, Republic of China

MONG-NA LO HUANG AND WEN-JANG HUANG

Institute of Applied Mathematics, National Sun Yat-sen University,
Kaohsiung, Taiwan 80424, Republic of China

In this paper, we consider the problem of approximating the location, $x_0 \in C$, of a maximum of a regresion function, $\theta(x)$, under certain weak assumptions on θ. Here C is a bounded interval in R. A specific algorithm considered in this paper is as follows. Taking a random sample $X_1, ..., X_n$ from a distribution over C, we have (X_i, Y_i), where Y_i is the outcome of noisy measurement of $\theta(X_i)$. Arrange the Y_i's in nondecreasing order and take the average of the r X_i's which are associated with the r largest order statistics of Y_i. This average, $\hat{x}_0$, will then be used as an estimate of x_0. The utility of such an algorithm with fixed r is evaluated in this paper. To be specific, the convergence rates of $\hat{x}_0$ to x_0 are derived. Those rates will depend on the right tail of the noise distribution and the shape of $\theta(\cdot)$ near x_0.

1. INTRODUCTION

Let θ be a real function defined on a bounded interval $C \in R$, and suppose there is an $x_0 \in C$ with $\theta(x_0) > \theta(x)$ for any $x \neq x_0$ in C. It is further assumed that $\theta(\cdot)$ is continuous. The objective is to determine x_0 based on n samples $(X_1, Y_1), ..., (X_n, Y_n)$ with $Y_i = \theta(X_i) + \varepsilon_i$, where n is a predetermined number. Here $\{\varepsilon_i\}$ are independent and identically distributed (i.i.d.) random variables with zero expectation.

Received April 28, 1994; revised July 1995.

AMS 1991 subject classifications: primary 62G05, 62G20, 62G30; secondary 62N99.

Key words and phrases: extreme-value distribution, global function optimization from noisy samples, errors of measurement, ranking selection.

* Part of this research is supported by National Science Council of Republic of China.

[†] Hung Chen's research is also sponsored by the National Science Foundation under Grant No. DMS-9209710.

0047-259X/96 $18.00

In this paper, we study the utility of the so-called best-r-points-average method used in Changchien [5]. This method has been used to search for the optimum range of burden distribution indices of blast furnace to extract iron from large quantities of iron-bearing materials. A quick introduction on metal production using an electric furnace can be found in Lawson [12]. More explicitly, for given n samples $(X_1, Y_1), ..., (X_n, Y_n)$, where $X_1, ..., X_n$ are over C according to a certain distribution, let $X_{[i:n]}$ be the concomitants pertaining to the ordered Y-values with $Y_{1:n} \leqslant \cdots \leqslant Y_{n:n}$. Then the best-r-points-average estimator $\hat{x}_0(r)$ of x_0 is defined by $\hat{x}_0(r) = \sum_{i=1}^{r} X_{[n-i+1:n]}/r$. When $X_1, ..., X_n$ can be chosen by the experimenter, X_1 is often chosen to be uniformly distributed over C.

The best-r-points-average method considered in Changchien [5] can be thought of as a modification of the recipe in de Haan [8]. De Haan [8] uses a search recipe with uniformly distributed points over C, although it is assumed there that the regression function $\theta(\cdot)$ is observed without error (i.e., $Y_i = \theta(X_i)$). When the data has already been sampled according to some fixed designs or some distributions, as in the case we consider here, Müller [14] derives the estimate of x_0 based on a kernel estimate of $\theta(\cdot)$ over C with data-driven bandwidth.

An obvious advantage of the best-r-points-average method is its ease in implementation and the disadvantage is not using all the information contained in (X_i, Y_i). However, the convergence rate of this method is found to be faster than the one considered in Müller [14] for certain noise distributions according to the discussion at the end of Section 2.

A possible shortcoming of the non-sequential methods (such as the best-r-points-average method and the kernel method discussed in Müller [14]) is that some unproductive trials may be performed. Instead many sequential approaches have been suggested in the literature, including the Kiefer–Wolfowitz recursive stochastic approximation type procedure [10], Hotelling's stagewise approach [9], the response surface method [4], etc. Both the Kiefer–Wolfowitz procedure and the response surface methods can be viewed as extensions of the commonly used gradient methods of numerical analysis to the case where the objective function values can be observed with error. It is well known that the gradient methods may fail to approach the global maximum unless the objective function has no other stationary points (points with a vanishing derivative). Although these methods have been found to be useful in many applications, there are still many occasions that these methods are not suitable. For example, Müller [14] has considered the problem of identifying a peak in the FSH (a hypophyseal hormone) curve, where selecting the design points sequentially is not practically feasible. Also there are situations in which the data have been routinely collected as discussed in Section 4 of Banks [1]; then the method studied in this paper can be helpful.

To gain some insight on the best-r-points-average method, suppose that under certain situations the noise level is low. Intuitively, the regression function value associated with the largest order statistic of Y_i's should be large compared to $\theta(X_1)$, ..., $\theta(X_n)$. It is also clear that these central order statistics of Y_i's cannot carry asymptotic information concerning x_0 without additional global condition on $\theta(\cdot)$. On the other hand, the second largest order statistics of Y_i's should also carry information of x_0. The appropriate choice of r then becomes an interesting question.

Note that as n becomes large, so should r, or we fail to reap the benefit of an increasing sample size. Changchien [5] proposes to use r which is greater than one and demonstrates its superiority over $r=1$ via a small-scale simulation study. Again, the advantage of using $r>1$ is confirmed based on a limited simulation study presented in Section 4, not to mention that the use of $r>1$ has the potential to alleviate the problem caused by "outliers". However, as this paper is only a first attempt to gain some understanding of the merit of this algorithm, we will consider the case where r is fixed here. The case that r is a nondecreasing integer sequence tending to infinity while $\lim_n r/n=0$ is still under investigation, and the result will be reported elsewhere. According to Kiefer [11], it is expected that the analysis will be quite different from the one used in this paper since the rth largest order statistics exhibits a different behavior based on the ranges of r.

Finally, we would like to remark that the theory developed in this paper can be extended to the case where C is a bounded set in a d-dimensional Euclidean space. Also, our result might be useful in giving a partial answer to one of the questions raised in Section 4.1 of Banks [1]. The question is whether one can focus attention upon modeling the response surface to inform future efforts at process optimization using only the data that gave the 10% best quality output.

This paper is organized as follows: Section 2 describes the main results for fixed r. Section 3 collects some results of the extreme-value distributions. In Section 4 we present some heuristic arguments to illustrate why the validity of the best-r-points-average method depends on the right tail of the distribution of ε. Also, we report results from a Monte Carlo study to evaluate the finite sample property of the best-r-points-average method and to compare this estimate with the estimates in Müller [14]. The basic idea underlying all our results is presented at the beginning of Section 5. Finally, Section 5 provides a detailed proof of the main results.

2. The Main Results

The utility of the best-r-points-average method with $r=1$ depends on whether the unobservable regression function value, $\theta(\cdot)$, corresponding to the largest order statistic $Y_{n:n}$ will be close to the maximum of $\{\theta(X_1), ..., \theta(X_n)\}$. To answer this question, we reformulate the problem as follows. Suppose (Y_i, Z_i) $(i=1, 2, ..., n)$ is a random sample of n observations of a bivariate random variable (Y, Z) with $Y=Z+\varepsilon$, where Z and ε are independent variables. Note that only the Y_i's are observed and the corresponding Z_i is $\theta(X_i)$. Some notations will be introduced first. If we place the Y_i's in nondecreasing order, as $Y_{1:n} \leqslant \cdots \leqslant Y_{n:n}$, then the Z_i corresponding $Y_{i:n}$, denoted $Z_{[i:n]}$, is termed as the ith induced Z-order statistic. It is then clear that $Z_{[i:n]}$ is not the same as the ith smallest Z-value, $Z_{i:n}$. Incidentally, $Z_{[i:n]}$ is also termed as the *concomitant of the ith order statistic*. (Refer to Bhattacharya [3] for further details.) If Z has a continuous distribution function, then the $\{Z_{[i:n]}\}$ are distinct with probability 1, and the rank of $Z_{[i:n]}$ among $Z_1, ..., Z_n$ is unambiguously defined.

In the following, for a random variable T, its distribution function and density function are denoted by F_T and f_T, respectively. Also let α_T and w_T denote the left and right endpoints of F_T, respectively, which are defined as $\alpha_T = \inf\{t: F_T(t) > 0\}$ and $w_T = \sup\{t: F_T(t) \leqslant 1\}$.

The utility of the best-r-points-average method for locating the peak will be studied by deriving the asymptotic theory for the difference between $Z_{[n-r+1:n]}$ and the right endpoint of $F_Z(z)$ under the model mentioned above; that is, $Y=Z+\varepsilon$, where Z and ε are mutually independent. Motivated by the problem of locating the maximum of a regression function, we will further assume that the support of Z $(=\theta(X))$ is an interval $I=[\alpha_Z, w_Z]$. Note that $w_Z = \theta(x_0)$.

The utility of the best-r-points-average method depends on how fast $Z_{[n-r+1:n]}$ converges to w_Z which will be shown later to be dependent on both the distribution function of Z in the neighborhood of w_Z and f_ε.

Condition R. $w_Z < \infty$ and F_Z satisfies $c_1 z^\tau \leqslant 1 - F_Z(w_Z - z) \leqslant c_2 z^\tau$ as $z \to 0^+$. Here, $0 < \tau \leqslant 1$ and c_1 and c_2 are positive constants.

We now give an example which illustrates when Condition R will hold.

Example 1. Assume $f_X(x)$ is supported on C and is bounded away from zero and infinity over C. Note that

$$\begin{aligned} P(w_Z - z \leqslant Z \leqslant w_Z) &= P(w_Z - z \leqslant \theta(X) \leqslant w_Z) \\ &= P(\theta(x_0) - z \leqslant \theta(X)) = P(\theta(x_0) - \theta(X) \leqslant z). \end{aligned}$$

Now suppose there exist some positive constants c_3, c_4, and ρ, $\rho \geqslant 1$, such that

$$c_3\,|x - x_0|^\rho \geqslant |\theta(x) - \theta(x_0)| \geqslant c_4\,|x - x_0|^\rho \qquad \text{for all} \quad x \in C. \tag{1}$$

Since

$$P\left(|X-x_0|\leqslant\left(\frac{z}{c_3}\right)^{1/\rho}\right)\leqslant P(\theta(x_0)-\theta(X)\leqslant z)$$
$$\leqslant P\left(|X-x_0|\leqslant\left(\frac{z}{c_4}\right)^{1/\rho}\right),$$

Z satisfies Condition R with $\tau=1/\rho$. As an example, if $\theta(x)$ is twice differentiable near x_0 and $\theta''(x_0)<0$, then Condition R holds with $\tau=1/2$ $(\rho=2)$.

Throughout this paper, we assume that ε satisfies either Condition E(1) or E(2) described in the following.

Condition E. (1) $w_\varepsilon<\infty$ and for some $k\geqslant 0$, f_ε and F_ε satisfy $(-1)^k f_\varepsilon^{(k)}(w_\varepsilon)>0$, $f_\varepsilon^{(j)}(w_\varepsilon)=0$ for every $0\leqslant j\leqslant k-1$, and $\lim_{t\uparrow w_\varepsilon}(w_\varepsilon-t)\, f_\varepsilon(t)/[1-F_\varepsilon(t)]=k+1$.

(2) $w_\varepsilon=\infty$ and f_ε satisfies

$$f_\varepsilon(x)\sim ABvx^{-u+v-1}\exp(-Bx^v) \qquad \text{as} \quad x\to\infty,$$

where $v>1$, $u\geqslant 0$, and A, B are positive constants.

Here "$g(x)\sim(h(x)$ as $x\to\infty$" denotes $\lim_{x\to\infty} g(x)/h(x)=1$.

THEOREM 1. *Suppose* $Y=Z+\varepsilon$, *where* Z *and* ε *are independent. Let* Z *satisfy Condition* R. *Then*

(a) $Z_{[n-l:n]}-w_Z=O_p((\log n/n)^{1/(1+(k+1)/\tau)})$ *for all* $l<r$ *under Condition* E(1);

(b) $Z_{[n-l:n]}-w_Z=O_p((\log n)^{-((v-1)/v)}(\log\log n)^\tau)$ *for all* $l<r$ *under Condition* E(2).

Since we only consider the case that r is fixed throughout this paper, for simplicity denote $\hat{x}_0(r)$ by $\hat{x}_0$. Now we describe the asymptotic behavior of $\hat{x}_0-x_0$ which follows from Theorem 1 and Example 1.

THEOREM 2. *Assume there exist some positive constants* c_3 *and* c_4 *such that* (1) *holds. Then* $\hat{x}_0-x_0=O_p((\log n/n)^{1/(1+(k+1)/\tau)})$ *or* $O_p((\log n)^{-((v-1)/v)}(\log\log n)^\tau)$ *when* ε *satisfies Condition* E(1) *or* E(2), *respectively.*

Remark 1. When ε is uniformly distributed, Condition E(1) is satisfied with $k=0$. Theorem 2 states that $\hat{x}_0-x_0=O_p((\log n)^{1/2}\,n^{-1/2})$ when $\theta(x)$

is "wedge-shaped" ($\tau=1$) around $x=x_0$. If $\theta(x)$ is twice differentiable near x_0 and $\theta''(x_0)<0$, then $\hat{x}_0-x_0=O_p((\log n)^{1/3}\,n^{-1/3})$. Müller [15] proposes an estimate of x_0, x_{M0}, by estimating $\theta(x_0)$ with the kernel smoother. If $|\theta(x)-\theta(x_0)|\geqslant c\,|x-x_0|^{\rho}$ for some $c>0$ and $\rho\geqslant 1$ in a neighborhood of x_0, then $\hat{x}_{M0}-x_0=O_p([(n\log n)^{-2/5}]^{\tau})$. Compare the estimate obtained by the best-r-points-average method with the one in Müller [15] and the so-called *passive stochastic approximation* method in Tsybakov [18], where the convergence rate is about the same as for that in Müller [15] and Tsybakov [18], it is easy to see that the estimator in this paper is *better* when ε is a uniform random variable. On the other hand, when ε is a normal random variable (i.e., $u=1$, $v=2$ in Condition E(2)), Theorem 2 states that $\hat{x}_0-x_0=O_p([(\log n)^{-1/2}]^{\tau}\,(\log\log n)^{\tau})$, which then implies that this estimator is not as good as those considered in Müller [15] and Tsybakov [18]. But the estimator based on the best-r-points-average is much simpler and easily understood so that it can be implemented in practical applications easily. Furthermore, the result derived in Müller [15] cannot be improved even when ε is known to be uniformly distributed.

Remark 2. Theorem 2 states that the best-r-points-average method for locating the peak works better under Condition E(1) than Condition E(2). By formulating the problem in the framework of the ranking selection problem as described in Section 4, the rate of $\hat{x}_0-x_0$ depends critically on whether w_ε is finite or not and the local behavior of F_ε near w_ε. In particular, the best-r-points-average method works best when $w_\varepsilon<\infty$ or ε has a short-tailed distribution. When the tail of f_ε is long as discussed in Remark 1, other estimates, such as that in Müller [15], are perhaps more suitable.

3. Extreme-Value Distributions

To facilitate the discussions in Section 4, we briefly review the aspects of the extreme-value distribution theory which can be found in Section 5.1 of Reiss [16] on the topic of the *domain of convergence*. They are summarized in two lemmas and will be used repeatedly in Sections 4 and 5.

Let $\varepsilon_1, ..., \varepsilon_n$ be independent random variables with common distribution $F_\varepsilon(\cdot)$. Let $\varepsilon_{h:n}=\max(\varepsilon_1, ..., \varepsilon_n)$. Denote the (left continuous) inverse of F_Y as $F_Y^{\leftarrow}(u)$ which is defined as $\inf\{y\colon F_Y(y)\geqslant u\}$. The distribution F_ε is said to be in the *domain of* (maximum) *attraction* of a distribution G (written $F_\varepsilon\in D(G)$) if there are $\{a_n\}$ ($a_n>0$) and $\{b_n\}$ so that

$$\lim_{n\to\infty} F_\varepsilon^n(a_n x+b_n)=G(x)$$

at every continuity point of G. It is well known (see, e.g., p. 5.3 of Reiss [16]) that G must belong to, up to location and scale, one of the three classes of extreme-value distributions described in the following lemma.

Lemma 1. *Suppose there exist $a_n > 0$, $b_n \in R$, $n \geqslant 1$, such that*

$$P[(\varepsilon_{n:n} - b_n)/a_n \leqslant x] = F_\varepsilon^n(a_n x + b_n) \to G(x),$$

weakly as $n \to \infty$, where G is assumed to be nondegenerate. Then, with suitable numbers $A > 0$ and B, $G(Ax + B)$ belongs to one of the following three classes:

$$\text{(a)} \quad \Phi_\alpha(x) = \begin{Bmatrix} \exp(-x^{-\alpha}) & \textit{if} \quad x > 0 \\ 0 & \textit{if} \quad x \leqslant 0 \end{Bmatrix} \quad \textit{for some} \quad \alpha > 0;$$

$$\text{(b)} \quad \Psi_\alpha(x) = \begin{Bmatrix} 1 & \textit{if} \quad x > 0 \\ \exp(-(-x)^\alpha) & \textit{if} \ x \leqslant 0 \end{Bmatrix} \quad \textit{for some} \quad \alpha > 0;$$

$$\text{(c)} \quad \Lambda(x) = \exp(-e^{-x}), \qquad x \in R$$

To prove Theorem 1, we need the bound for the variational distance

$$e_n = \sup_{x \in R} |F_\varepsilon^n(a_n x + b_n) - G(x)|,$$

where $F_\varepsilon \in D(G)$, between the exact and limiting distributions. The following lemma gives a prescription on the choice of normalizing constants $\{a_n\}$ and $\{b_n\}$.

Lemma 2. *Assume that f_ε is positive on (t, w_ε), where $t < w_\varepsilon$.*

(a) *If $w_\varepsilon = \infty$, and*

$$\lim_{t \to \infty} \frac{t f_\varepsilon(t)}{1 - F_\varepsilon(t)} = \alpha, \tag{2}$$

with some $\alpha > 0$, then there are constants $a_n > 0$ and b_n such that the distribution of $(\varepsilon_{n:n} - b_n)/a_n$ converges to Φ_α. Moreover, the constants can be chosen as $a_n = F_\varepsilon^{\leftarrow}(1 - 1/n)$ and $b_n = 0$.

(b) *If $w_\varepsilon < \infty$ and $\lim_{t \uparrow w_\varepsilon}(w_\varepsilon - t) f_\varepsilon(t)/[1 - F_\varepsilon(t)] = \alpha$, then there are constants $a_n > 0$ and b_n such that the distribution of $(\varepsilon_{n:n} - b_n)/a_n$ converges to Ψ_α. Moreover, the constants can be chosen as $a_n = w_\varepsilon - F_\varepsilon^{\leftarrow}(1 - 1/n)$ and $b_n = w_\varepsilon$.*

(c) *If $\int_{-\infty}^{w_\varepsilon} (1 - F_\varepsilon(t))\, dt < \infty$ and*

$$\lim_{t \uparrow w_\varepsilon} \frac{f_\varepsilon(t)}{[1 - F_\varepsilon(t)]^2} \int_t^{w_\varepsilon} [1 - F_\varepsilon(u)]\, du = 1,$$

then there are constants $a_n > 0$ *and* b_n *such that the distribution of* $(\varepsilon_{n:n} - b_n)/a_n$ *converges to* Λ. *Moreover, the constants can be chosen as* $a_n = [nf_\varepsilon(b_n)]^{-1}$ *and* $b_n = F_\varepsilon^{\leftarrow}(1 - 1/n)$.

(d) $e_n = o(1)$ *if one of* (a), (b), *and* (c) *applies.*

If, to $F_\varepsilon(t)$, *none of* (a), (b), *and* (c) *applies, then there are no constants* $a_n > 0$ *and* b_n *such that the distribution of* $(\varepsilon_{n:n} - b_n)/a_n$ *would converge.*

4. Discussion and Monte Carlo Study

In this section, we give a heuristic argument to illuminate when and why the best-r-points-average method for locating the maximum works. In order to get some ideas on the finite sample property of the best-r-points-average method, we also run a Monte Carlo study as in Müller [15] with $r = 1, 5$. The results are summarized in Tables II and III which indicate the advantage of using $r > 1$. We also compare our simulation results with the one in Müller [15]. Theoretical development in this paper states that the best-r-points-average method can be useful in locating global maximum of a regression function with local maximum. A Monte Carlo experiment is conducted to confirm it.

We now give a heuristic argument to explain why the convergence rates of the proposed estimate should depend on the behavior of the tail of $1 - F_\varepsilon(x)$ as x increases. This argument is essentially used in Section 5 to prove Theorem 1. Although Z is assumed to be a continuous random variable in Section 2, we here consider the case where Z takes values on discrete levels $0 \leqslant \theta_{n1} < \cdots < \theta_{nK} \leqslant 1$, for some integer $K \geqslant 1$. Let $n = KN$, where N is an integer. For each θ_{nj}, we further assume that there are N samples from $Y = \theta_{nj} + \varepsilon$. In other words, we have i.i.d. random variables $Y_{j1}, ..., Y_{jN}$, from the jth population with distribution $F_\varepsilon(\cdot - \theta_{nj})$ for $1 \leqslant j \leqslant K$. It is then clear that the utility of the best-r-points-average method with $r = 1$ depends on whether the location parameter, associated with the population yielding $Y_{n:n}$, is close to θ_{nK}. This problem can then be viewed as to use the largest order statistics from each population to discriminate among location parameter families $F_\varepsilon(\cdot - \theta)$ for $\theta \in \{\theta_{n1}, ..., \theta_{nK}\}$. Obviously, this problem is related to the ranking selection problem as introduced by Bechhofer [2].

When $Y_{j1}, ..., Y_{jN}$ follow the distribution $F_\varepsilon(\cdot - \theta_{nj})$, its sample mean is the complete sufficient statistics of θ_{nj} when ε is normally distributed. In this case, it seems reasonable to use the sample means from those K populations to discriminate the location parameter families $F_\varepsilon(\cdot - \theta)$. At the above setting, Müller's curve fitting approach [15] reduces to discriminate the location parameter families $\{F_\varepsilon(\cdot - \theta); \theta = \theta_{n1}, ..., \theta_{nK}\}$ with sample means. When ε is uniformly distributed, the largest order statistics from those K populations can be used to discriminate the location parameter families $F_\varepsilon(\cdot - \theta)$ effectively.

It is then expected that the estimate derived by Müller's curve fitting approach will converge to x_0 with a faster rate than the estimate derived from the best-r-points-average method with $r=1$ for normal error. Also, Müller's estimate should have a slower convergence rate than the estimate obtained by the best-r-points-average method with $r=1$ for uniform error. This conjecture is confirmed by Theorem 2 and the discussions at the end of Section 2.

Next, we will demonstrate that the best-r-points-average method fails when the right tail of the error distribution is heavy. Let now $\varepsilon_1, \varepsilon_2, ..., \varepsilon_N$ be a random sample of size N from a unit double exponential distribution. Then $\varepsilon_{N:N} \in D(\Lambda)$ with $a_N = 1$ and $b_N = \log N$ by Lemmas 1 and 2. Therefore, the largest order statistic from the Kth population (with location parameter θ_{nK}) is not necessarily greater than the largest order statistic from the first population (with location parameter θ_{n1}) with probability 1, even when $\theta_{n1} = 0$ and $\theta_{nK} = 1$. According to the above discussion, it is expected that the best-r-points-average method will fail to give a consistent estimate of x_0 when the limit of a_n is nonzero.

By Lemma 2(a), we have $\lim_{N \to \infty} a_N = \infty$ for $F \in D(\Phi_\alpha)$. Hence, we exclude those $F \in D(\Phi_\alpha)$ from our study on the utility of the best-r-points-average method. Also by Lemma 2(b), b_N is finite and $\lim_{N \to \infty} a_N = 0$ for $F \in D(\Psi_\alpha)$. Therefore, we consider a class of distributions in $D(\Psi_\alpha)$ with $a_N = O(N^{-1/(k+1)})$ as described in Theorem 1(a). When $F \in D(\Lambda)$, $\lim_{N \to \infty} a_N$ may take any nonnegative value, as are the cases for double exponential distribution with $a_N = 1$ and normal distribution with $a_N = (2 \log N)^{-1/2}$. Hence, we consider a class of distributions in $D(\Lambda)$, as described in Theorem 1(b), whose tail is "lighter" than the double exponential distribution (with $a_N = (B^{-1} \log N)^{(1-v)/v}$ for $v > 1$).

Motivated by the problem of estimating distance to a stellar system from measurements on the apparent magnitude of a few of the brightest objects in the system, Rohatgi [17] considers the problem to determine which of the objects should be observed. She finds that the extreme order statistics are asymptotically sufficient for estimating distance when the distribution of the apparent magnitude of star in that galaxy is known up to a location parameter. Her conclusion is close to the above discussion in spirit.

In order to have some ideas on how well our asymptotic results of the best-r-points-average method predicted what would transpire for finite samples, we consider a Monte Carlo study as in Müller [15] with $r = 1, 5$ and sample size $n = 50$. In this study, $\theta(x) = 1 + 3\exp(-(x-0.5)^2/0.01)$ (symmetric peak at $(0.5, 4.0) = (x_0, \theta(x_0))$) with 50 points X_i's from the uniform distribution over $[0, 1]$. Since the performance of the proposed

TABLE I

Müller's Adaptive Procedure for Peak Estimation

	$\sigma^2 = 0.25$	$\sigma^2 = 0.5$	$\sigma^2 = 1.0$
Average	0.5005	0.5012	0.5012
ASE	1.247^{-4}	1.893^{-4}	2.860^{-4}

estimate $\hat{x}_0(r)$ depends on the error distribution, we consider three error distributions, which are uniform, normal, and double exponential. Note that Müller [15] only reports results for normal error distribution. For ease of comparing with Müller's study, we also consider three noise levels with variances 0.25, 0.5, and 1.0. The number of Monte Carlo runs is 200.

This experiment is repeated for 100 times. Tables II–IV show a typical result from one of these one hundred experiments. The notations used in the tables are defined as follows. Let σ^2, *Average*, *ASE*, and *Range* denote the variance of noise variable, *average estimated location*, *average squared error for location*, and *range of estimated location*, respectively. Also, 5.272^{-4} should be read as 5.272×10^{-4}. We first give Table I which is taken from Table 1 of Müller [15] which reflects the performance of his proposed procedure for adaptive peak estimation in that paper.

Tables I, II,.and III indicate that:

- The best-r-points-average method performs better when the error is uniform from the fact that it has smaller ASE and tighter range than that when the error is normal. This is consistent with Theorem 2(a) and (b) qualitatively, since according to Theorem 2, $x_0 - x_0 = O_p((\log n)^{1/3} n^{-1/3})$ and $O_p((\log n)^{-2} (\log \log n)^2)$ when the error distribution is uniform and normal, respectively. Tables II and III also illustrate the advantage of using $r > 1$.

- The best-r-points-average method with $r = 1, 5$ is not as good as the adaptive procedure in Müller [15] from an ASE standpoint in this

TABLE II

Uniform Error, $U[-\sigma\sqrt{3}, \sigma\sqrt{3}]$

	$\sigma^2 = 0.25$		$\sigma^2 = 0.5$		$\sigma^2 = 1.0$	
	$r = 1$	$r = 5$	$r = 1$	$r = 5$	$r = 1$	$r = 5$
Average	0.4998	0.4982	0.5006	0.4958	0.4988	0.4965
ASE	6.627^{-4}	3.211^{-4}	9.290^{-4}	6.157^{-4}	1.283^{-3}	1.189^{-3}
Range	(0.441, 0.586)	(0.452, 0.547)	(0.428, 0.586)	(0.373, 0.556)	(0.386, 0.586)	(0.373, 0.651)

TABLE III

Normal Error, $N(0, \sigma^2)$

	$\sigma^2 = 0.25$		$\sigma^2 = 0.5$		$\sigma^2 = 1.0$	
	$r = 1$	$r = 5$	$r = 1$	$r = 5$	$r = 1$	$r = 5$
Average	0.5012	0.5014	0.5017	0.4995	0.5008	0.4966
ASE	7.521^{-4}	5.310^{-4}	1.106^{-3}	1.311^{-3}	2.480^{-3}	2.470^{-3}
Range	(0.437, 0.565)	(0.414, 0.601)	(0.387, 0.575)	(0.360, 0.668)	(0.270, 0.884)	(0.306, 0.669)

particular setting. Also, for $r = 1$, the faster rate of the best-1-average method as claimed in Section 2 is not realized, based on the comparison of Tables I and II, perhaps due to the fact that the sample $n = 50$ is not large enough for reflecting the asymptotic results.

As a consequence, it is recommended to use $r > 1$ and derive better asymptotic results such as the asymptotic distribution of $\hat{x}_0 - x_0$. Research on the asymptotic distribution of the estimate based on the best-r-points-average method the practical choice of r is underway and the result will be reported elsewhere.

As a remark, Chen [6] shows that a modified Kiefer–Wolfowitz procedure in Fabian [7] achieves the optimal rates of convergence. However, it is known that the finite-sample performance of the Kiefer–Wolfowitzs procedure depends crucially on the choice of a starting point. The best-r-points-average method has the potential to be used in determining a "good" starting point for the Kiefer–Wolfowitz procedure.

According to the discussion at the beginning of this section, a problem with the best-r-points-average method is that it is not consistent when the tail of the error distribution is *heavy*. Table IV summarizes the Monte Carlo results at the setting when the error distribution is double exponential.

Table IV supports the discussion on the failure of the best-1-points-average method as the range of the estimator is much wider than the other

TABLE IV

Double Exponential Error, $(\sigma/\sqrt{2})\ DE(1)$, with Peak (0.5, 4.0)

	$\sigma^2 = 0.25$		$\sigma^2 = 0.5$		$\sigma^2 = 1.0$	
	$r = 1$	$r = 5$	$r = 1$	$r = 5$	$r = 1$	$r = 5$
Average	0.4968	0.4980	0.4909	0.5008	0.5076	0.5040
ASE	1.202^{-3}	6.616^{-4}	5.261^{-3}	1.522^{-3}	1.088^{-2}	4.124^{-3}
Range	(0.399, 0.728)	(0.381, 0.590)	(0.012, 0.803)	(0.384, 0.637)	(0.171, 0.947)	(0.333, 0.675)

TABLE V

Double Exponential Error with Peak (0.7, 4.0)

	$\sigma^2 = 0.25$		$\sigma^2 = 0.5$		$\sigma^2 = 1.0$	
	$r = 1$	$r = 5$	$r = 1$	$r = 5$	$r = 1$	$r = 5$
Average	0.6968	0.6969	0.6926	0.6886	0.6708	0.6665
ASE	2.680^{-3}	6.525^{-4}	4.118^{-3}	1.828^{-3}	1.689^{-2}	4.958^{-3}
Range	(0.097, 0.793)	(0.565, 0.760)	(0.097, 0.799)	(0.508, 0.788)	(0.024, 0.976)	(0.333, 0.770)

cases and either the left or the right end point of the interval for the range is close to one of the boundary points of the design interval from where the sample is taken.

But, based on a similar discussion above about the range of the estimator, it indicates that the best-5-points-average method might work. It is actually an artifact due to the facts that the peak is at 0.5 and the design points are uniformly distributed over [0, 1]. This explanation is supported by the following simulation study. In this study, we consider the case that $\theta(x) = 1 + 3\exp(-(x - 0.7)^2/0.01)$ with the peak at (0.7, 4.0) and the rest of settings remain the same. The results are summarized in Table V.

Table V supports the preceding explanation for results summarized as in Table IV. As the average of the estimators is shifting away from 0.7, which is the value of the maximizer of this example, when the variance is getting larger. Now, Tables IV and V clearly indicate that the best-r-points-average method is not consistent for the double exponential error. It supports the discussion on the failure of the best-r-points-average method when the tail of the error distribution is *heavy*.

As discussed in Section 1, some commonly used sequential approaches for estimating x_0 may fail to approach the global maximum if the regression function has multiple stationary points. We now assess the performance of the best-r-points-average method when the regression function has two well-separated stationary points. Here we consider $\theta(x) = 3.2\exp(-(x - 0.4)^2/0.01) + 4\exp(-(x - 0.6)^2/0.01)$ with the peak (0.5769, 3.9365) and a local maximum (0.4053, 3.3811). The error distribution is uniform with variance 0.25. Based on 200 Monte Carlo runs with $n = 50$ and $r = 1$, there are 9 runs falling in (0.3900, 0.4380) and the rest are between 0.5036 and 0.6453. When $n = 100$, the number reduces to only 3 out of 200 runs are within 0.02 distance of the local maximizer 0.4053. This indicates that the best-r-points-average method can pick up the global maximum when the signal to noise ratio is large enough.

5. Proof of Theorem 1

Let $\{K_n\}$ denote a sequence of positive integers such that $K_n \to \infty$ and $K_n/n \to \infty$ and $K_n/n \to 0$. For brevity we omit the subscript of K_n later on. Given $K+1$ knots $\alpha_Z = t_0 < t_1 < \cdots < t_{K-1} < t_K = w_Z$, let $I = [\alpha_Z, w_Z]$ be partitioned into subintervals

$$I_{Kj} = [t_{k-1}, t_k) \qquad \text{for} \quad 1 \leqslant j < K, \qquad I_{KK} = [t_{K-1}, t_K].$$

Set $\mathscr{I}_{Kj} = \{i\colon 1 \leqslant i \leqslant n \text{ and } Z_i \in I_{Kj}\}$ and denote the cardinality of $\mathscr{I}_{Kj}$ as $N_j(K)$. Assume that n/K ($\equiv N$) is an integer. Consider a particular choice of knots $\{t_{n1}, ..., t_{n,K-1}\}$ such that $N_j(K) \equiv N$ for $1 \leqslant j \leqslant K$. For $i \in \mathscr{I}_{Kj}$, denote those Z_i's by $Z_{k1}, ..., Z_{kN}$. We also denote those associated Y_i's and ε_i's by $Y_{j1}, ..., Y_{jN}$ and $\varepsilon_{j1}, ..., \varepsilon_{jN}$, respectively. Arrange the Y_{jl} (ε_{jl}, respectively) in nondecreasing order as the order statistics $Y_{l:N,j}$ ($\varepsilon_{l:N,j}$, respectively) for $1 \leqslant l \leqslant N$.

Suppose that the following statement holds for $r < J$.

$$\lim_n P(\inf_{K-r<l\leqslant K} Y_{N:N,l} \geqslant \sup_{1\leqslant j\leqslant K-J} Y_{N:N,j}) = 1. \tag{3}$$

By (3) and the definition of $Y_{N:N,j}$, we have $w_Z - Z_{[n-r+1:n]} \leqslant (J-r)/K_n$. In other words, $Z_{[n-r+1:n]} - w_Z = O_p((J-r)\,K_N^{-1})$.

Recall that $Y = Z + \varepsilon$. It follows easily that for $1 \leqslant j \leqslant K$,

$$\varepsilon_{N:N,j} + t_{nj} > Y_{N:N,j} \geqslant \varepsilon_{N:N,j} + t_{n,j-1}. \tag{4}$$

By (4) and the Bonferroni inequality,

$$\begin{aligned} &P(\inf_{K-r<l\leqslant K} Y_{N:N,l} \geqslant \sup_{1\leqslant j\leqslant K-J} Y_{N:N,j}) \\ &\qquad \geqslant 1 - \sum_{l=K-r+1}^{K} \sum_{j=1}^{K-J} P(\varepsilon_{N:N,l} - \varepsilon_{N:N,j} \leqslant t_{nj} - t_{n,l-1}). \end{aligned} \tag{5}$$

Note that the Z_i's are independent of the ε_i's. This implies that $\{\varepsilon_{jl}\}_{1\leqslant l\leqslant N;\, 1\leqslant j\leqslant K}$ are independent since the new label jl attached to the ε's are determined by Z_i. Hence, the sample maxima $\varepsilon_{N:N,j}$ for $1 \leqslant j \leqslant K$ are i.i.d. random variables. Denote by $d_{nlj} = t_{n,l-1} - t_{nj} > 0$. Write

$$P(\varepsilon_{N:N,l} - d\varepsilon_{N:N,j} \leqslant -d_{nlj}) = \int_{\alpha_\varepsilon}^{w_\varepsilon} [F_\varepsilon(t - d_{nlj})]^N \, N[F_\varepsilon(t)]^{N-1} f_\varepsilon(t)\, dt, \tag{6}$$

where $N[F_\varepsilon(t)]^{N-1} f_\varepsilon(t)$ is the density function of $\varepsilon_{N:N,j}$. Unless the distribution of $\varepsilon_{N:N,j}$ (suitably normalized) can be approximated by a nondegenerate distribution, the derivation of (6) would be quite difficult. From now on, we assume that F_ε belongs to the domain of attraction of an extreme-value distribution. Then the limiting distribution of extreme order statistics must be one of the three forms of limiting extreme-value distribution. Refer to Lemmas 1 and 2 for the details.

Assume that $F_\varepsilon^n(a_n t + b_n) \to G(t)$, where $G(t)$ is one of the extreme-value distribution functions described in Lemma 1. The right-hand side of (6) will be evaluated via the following:

$$\begin{aligned}
&\int_{a_0}^{b_0} [F_\varepsilon(t-d_{nlj})]^N N[F_\varepsilon(t)]^{N-1} f_\varepsilon(t)\, dt \\
&\quad = \int_{(a_0-b_N-d_{nlj})/a_N}^{(b_0-b_N-d_{nlj})/a_N} [F_\varepsilon(a_N u + b_N)]^N \\
&\qquad \times N[F_\varepsilon(a_N u + b_N + d_{nlj})]^{N-1} f_\varepsilon(a_N u + b_N + d_{nlj})\, a_N\, du \\
&\quad \leqslant e_N + N a_N \int_{(a_0-b_N-d_{nlj})/a_N}^{(b_0-b_N-d_{nlj})/a_N} G(u)[F_\varepsilon(a_N u + b_N + d_{nlj})]^{N-1} \\
&\qquad \times f_\varepsilon(a_N u + b_N + d_{nlj})\, du \\
&\quad = e_N + N a_{N-1} \int_{(a_0-b_{N-1})/a_{N-1}}^{(b_0-b_{N-1})/a_{N-1}} G\left(\frac{a_{N-1}t + b_{N-1} - b_N - d_{nlj}}{a_N}\right) \\
&\qquad \times [F_\varepsilon(a_{N-1}t + b_{N-1})]^{N-1} f_\varepsilon(a_{N-1}t + b_{N-1})\, dt \\
&\quad \leqslant e_N + N a_{N-1} \int_{(a_0-b_{N-1})/a_{N-1}}^{(b_0-b_{N-1})/a_{N-1}} G\left(\frac{a_{N-1}t + b_{N-1} - b_N - d_{nlj}}{a_N}\right) \\
&\qquad \times G(t)\, f_\varepsilon(a_{N-1}t + b_{N-1})\, dt \\
&\qquad + e_{N-1} N a_{N-1} \int_{(a_0-b_{N-1})/a_{N-1}}^{(b_0-b_{N-1})/a_{N-1}} G\left(\frac{a_{N-1}t + b_{N-1} - b_N - d_{nlj}}{a_N}\right) \\
&\qquad \times f_\varepsilon(a_{N-1}t + b_{N-1})\, dt \\
&\quad = (\mathrm{I}) + (\mathrm{II}) + (\mathrm{III}).
\end{aligned} \tag{7}$$

Recall that t_{nj} is the jK^{-1} quantile of $F_Z(\cdot)$. A bound of $t_{nj} - F_Z^{\leftarrow}(jK_n^{-1})$ can be obtained from the following lemma on sample quantiles, which can be found as Proposition 2 in Lo [13].

LEMMA 3. *Let $F(z)$ be a continuous distribution on the real line and let $\{p_n\}$ be a positive monotone increasing sequence between* 0 *and* 1, *let ξ_{P_n} denote the p_nth quantile of the distribution F, and let $\hat{\xi}_{p_n} = F_n^{\leftarrow}(p_n)$ be the* *sample quantile. Here F_n is the usual empirical distribution function of F. Then, for* $(\log n)(n(1-p_n))^{-1}=O(1)$,

$$P(|\hat{\xi}_{p_n}-\xi_{p_n}|>3\sqrt{2}\,(\log n)^{1/2}\,(1-p_n)^{1/2}\,n^{-1/2})$$
$$\leqslant 4\exp(-2\log n)=O(n^{-2}).$$

5.1. *A Family of Distributions in the Domain of Attraction of* $\Psi_\alpha(x)$

In this section we consider the case that ε satisfies Condition E(1), which would imply that ε is a random variable with $w_\varepsilon<\infty$ and the density function, f_ε, in a neighborhood of w_ε behaves like $c(w_\varepsilon-x)^k$ for some constant c and nonnegative integer k. This particular setting is used to illustrate how the behavior of f_ε near w_ε affects the behavior of $Z_{[n-l+1:n]}-w_\varepsilon$. Before we prove Theorem 1(a), we need a preliminary result on a_n and b_n.

Lemma 4. $F_\varepsilon\in\Psi_{k+1}(x)$ *with* $a_n=cn^{-1/(k+1)}$ *and* $b_n=w_\varepsilon$, *where* $c=[(-1)^k\,(k+1)\,!/f_\varepsilon^{(k)}(w_\varepsilon)]^{1/(k+1)}$.

Proof. According to Condition E(1) and Lemma 2(b), $F_\varepsilon\in\Psi_{k+1}(x)$, $a_n=w_\varepsilon-F_\varepsilon^{\leftarrow}(1-1/n)$, and $b_n=w_\varepsilon$. Note that $F_\varepsilon(w_\varepsilon)-F_\varepsilon(F_\varepsilon^{\leftarrow}(1-1/n))=1/n$. Set $c_n=F_\varepsilon^{\leftarrow}(1-1/n)$. Hence,

$$n^{-1}=\int_{c_n}^{w_\varepsilon}\left[f_\varepsilon(t)-\frac{f_\varepsilon^{(k)}(w_\varepsilon)}{k!}\,(t-w_\varepsilon)^k\right]dt+\frac{f_\varepsilon^{(k)}(w_\varepsilon)}{k!}\int_{c_n}^{w_\varepsilon}(t-w_\varepsilon)^k\,dt,$$

$$n^{-1}=O\left(\frac{f_\varepsilon^{(k)}(w_\varepsilon)}{(k+1)!}\,(c_n-w_\varepsilon)^{k+1}\right).$$

This proves the result for a_n. ▮

Proof of Theorem 1(a). According to the discussion at the beginning of Section 5, it remains to study $P(\varepsilon_{N:N,l}\leqslant -d_{nlj})$. It will be evaluated by dividing the interval of integration $(\alpha_\varepsilon, w_\varepsilon)$ in (6), into two intervals $(\alpha_\varepsilon, a_0)$ and $[a_0, w_\varepsilon)$ with $a_0=F_\varepsilon^{\leftarrow}(\frac{1}{2})$. Using (7) with $b_0=w_\varepsilon$, $a_N=cN^{-1/(k+1)}$, $b_N=w_\varepsilon$, and $G(\cdot)=\Psi_{k+1}(\cdot)$, we have

$$\int_{a_0}^{w_\varepsilon}[F_\varepsilon(t-d_{nlj})]^N\,N[F_\varepsilon(t)]^{N-1}\,f_\varepsilon(t)\,dt$$
$$\leqslant e_N+\frac{ce_{N-1}N^{k/(k+1)}}{2}\int_{((a_0-w_\varepsilon)/c)\,N^{1/(k+1)}}^{0}\Psi_{k+1}$$
$$\left(\left(\frac{N}{N-1}\right)^{1/(k+1)u}u-\frac{N^{1/(k+1)}d_{nlj}}{c}\right)$$
$$\times f_\varepsilon(c(N-1)^{-1/(k+1)}\,u+w_\varepsilon)\,du$$

$$+\frac{cN^{k/(k+1)}}{2}\int_{((a_0-w_\varepsilon)/c)\,N^{1/(k+1)}}^{0}\Psi_{k+1}\left(\left(\frac{N}{N-1}\right)^{1/(k+1)}u-\frac{N^{1/(k+1)}d_{nlj}}{c}\right)\Psi_{k+1}(u)$$
$$\times f_\varepsilon(c(N-1)^{-1/(k+1)}u+w_\varepsilon)\,du. \tag{8}$$

We will study the third term and the second term at the right-hand side of (8), respectively. Note that $\Psi_{k+1}(\cdot)$ is a nondecreasing function. We then have

$$\int_{((a_0-w_\varepsilon)/c)\,N^{1/(k+1)}}^{0}\Psi_{k+1}\left(\left(\frac{N}{N-1}\right)^{1/(k+1)}u-\frac{N^{1/(k+1)}d_{nlj}}{c}\right)$$
$$\times\Psi_{k+1}(u)\,f_\varepsilon(c(N-1)^{-1/(k+1)}u+w_\varepsilon)\,du$$
$$=\int_{((a_0-w_\varepsilon)/c)\,N^{1/(k+1)}}^{0}\exp\left((-1)^k\left\{\left[\left(\frac{N}{N-1}\right)^{1/(k+1)}u-\frac{N^{1/(k+1)}d_{nlj}}{c}\right]^{k+1}+u^{k+1}\right\}\right)$$
$$\times f_\varepsilon(c(N-1)^{-1/(k+1)}u+w_\varepsilon)\,du$$
$$\leqslant\exp\left(-\frac{(-1)^k f_\varepsilon^{(k)}(w_\varepsilon)}{(k+1)!}Nd_{nlj}^{k+1}\right) \tag{9}$$

and

$$e_{N-1}\int_{((a_0-w_\varepsilon)/c)\,N^{1/(k+1)}}^{0}\Psi_{k+1}\left(\left(\frac{N}{N-1}\right)^{1/(k+1)}u-\frac{N^{1/(k+1)}d_{nlj}}{c}\right)$$
$$\times f_\varepsilon\left(c(N-1)^{-1/(k+1)}u+\frac{1}{2}\right)du$$
$$\leqslant e_{N-1}\exp\left(-\frac{(-1)^k f_\varepsilon^{(k)}(w_\varepsilon)}{(k+1)!}Nd_{nlj}^{k+1}\right). \tag{10}$$

Since $e_{N-1}=o(1)$ by Lemma 2(d), the right-hand side of (10) is of smaller order than the right-hand side of (9).

By (8), (9), and (10), it is clear that $P(\varepsilon_{N:N,l}-\varepsilon_{N:N,j}\leqslant -d_{nlj})$ tends to zero if Nd_{nlj}^{k+1} tends to infinity. Recall that $(-1)^k f_\varepsilon^{(k)}(w_\varepsilon)>0$. Observe that

$$\int_{\alpha_\varepsilon}^{a_0}[F_\varepsilon(t-d_{nlj})]^N N[F_\varepsilon(t)]^{N-1}f_\varepsilon(t)\,dt\leqslant N[F_\varepsilon(a_0)]^{2N}.$$

Hence, for $r < J$, we have

$$\begin{aligned} P(\inf_{K-r<l\leqslant K} Y_{N:N,K} \geqslant \sup_{1\leqslant j\leqslant K-J} Y_{N:N,j}) \\ \geqslant 1 + e_N - \sum_{l=K-r+1}^{K} \sum_{j=1}^{K-J} N[F_\varepsilon(w_\varepsilon)]^{2N} \\ - \frac{c}{2} \sum_{l=K-r+1}^{K} \sum_{j=1}^{K-J} \exp\left(-\frac{(-1)^k f_\varepsilon^{(k)}(w_\varepsilon)}{(k+1)!} Nd_{nlj}^{k+1}\right) \\ + r(K-J)\, O(N^{-1}). \end{aligned}$$

Finally, we evaluate the magnitude of d_{nlj}. Recall that $d_{nlj} = t_{n,l-1} - t_{nj}$ and that t_{nj} is the jK^{-1} quantile of $F_Z(\cdot)$. Note that $(\log n)(n(1 - jK_n^{-1}))^{-1} = O(1)$ for $1 \leqslant j \leqslant K$ when $\log n/N \to 0$. It follows from Lemma 3 that

$$\begin{aligned} P(\sup_{1\leqslant j\leqslant K} |t_{nj} - F_Z^{\leftarrow}(jK^{-1})| > 3\sqrt{2}(\log n)^{1/2} n^{-1/2}) \\ \leqslant 4K \exp(-2 \log n) = O(Kn^{-2}) = O(n^{-\gamma}) \end{aligned}$$

for some $\gamma > 1$. By the Borel–Cantelli lemma, we have

$$t_{nj} - F_Z^{\leftarrow}(jK_n^{-1}) = O((\log n)^{1/2} n^{-1/2}) \qquad \text{a.s.} \tag{11}$$

For the ease of presentation, we first consider the case $\tau = 1$. It follows from (11) and Condition R that $F_Z^{\leftarrow}(jK^{-1})/[jK^{-1}]$ is bounded away from zero and infinity when j is large. Hence, we have

$$\begin{aligned} \sum_{l=K-r+1}^{K} \sum_{j=1}^{K-J} \exp\left(-\frac{(-1)^k f_\varepsilon^{(k)}(w_\varepsilon)}{(k+1)!} Nd_{nlj}^{k+1}\right) \\ \leqslant rK \exp\left(-M_1 \frac{(-1)^k f_\varepsilon^{(k)}(w_\varepsilon)}{(k+1)!} N[(J-r)\, K^{-1}]^{k+1}\right) \end{aligned}$$

for some positive constant M_1. Set $K = O((n/\log n)^{1/(k+2)})$. Note that

$$\sum_{l=K-r+1}^{K} \sum_{j=1}^{K-J} N[F_\varepsilon(a_0)]^{2N} \leqslant rn[F_\varepsilon(a_0)]^{2N} \to 0.$$

It follows easily that $\lim_n rN^{k/(k+1)}K \exp(-M((-1)^k f_\varepsilon^{(k)}(w_\varepsilon)/(k+1!)\, N[(J-r)\, K^{-1}]^{k+1}) = 0$ when $J \to \infty$. Since $r(K-J)\, O(N^{-1}) = O(n^{-k/(k+2)}(\log n)^{-1/2(k+2)}) = o(1)$ and $e_N = o(1)$ by Lemma 2(d), the above discussions conclude that $Z_{[n:n]} - w_Z = O_p((\log n/n)^{1/(k+2)})$ for all $l < r$ under Condition R with $\tau = 1$.

For general τ, $F_Z^{\leftarrow}(1-jK^{-1})=w_Z-O((jK^{-1})^{1/\tau})$ when j is small and Condition R holds, which implies $d_{n,K-r+1,K-J}=O([J^{1/\tau}-(r-1)^{1/\tau}]\,K^{-1/\tau})$. It follows that

$$N^{k/(k+1)}\sum_{l=K-r+1}^{K}\sum_{j=1}^{K-J}\exp\left(-\frac{(-1)^k f_\varepsilon^{(k)}(w_\varepsilon)}{(k+1)!}Nd_{nlj}^{k+1}\right)$$
$$\leqslant rN^{k/(k+1)}K\exp\left(-M_2\frac{(-1)^k f_\varepsilon^{(k)}(w_\varepsilon)}{(k+1)!}N[(J^{1/\tau}-r^{1/\tau})\,K^{-1/\tau}]^{k+1}\right)$$

for some positive constant M_2. Set $K_n=O((n/\log n)^{1/[1+(k+1)/\tau]})$. It follows easily that $\lim_n rN^{k/(k+1)}K\exp(-M_2(-1)^k f_\varepsilon^{(k)}(w_\varepsilon)/(k+1)!)$ $N[(J^{1/\tau}-r^{1/\tau})\,K^{-1/\tau}]^{k+1})=0$ when $J\to\infty$. Since $r(K-J)\,O(N^{-1})=O((\log n)^{-2\tau/(k+1+\tau)}\,n^{-(k+1-\tau)/(k+1+\tau)})=o(1)$, the above discussions conclude that $Z_{[n-l+1:n]}-w_Z=O_p((\log n/n)^{1/[1+(k+1)/\tau]})$ for all $l\leqslant r$ under Condition R. ▌

5.2. *A Family of Distributions in the Domain of Attraction of $\Lambda(x)$*

In this section, we consider the case that ε satisfies Condition E(2). Before we prove Theorem 1(b), we need the following lemma.

Lemma 5. (a) $1-F_\varepsilon(x)\sim Ax^{-u}\exp(-Bx^v)$ *as* $x\to\infty$.

(b) $F_\varepsilon\in D(\Lambda)$ *with* $a_n=(nf(b_n))^{-1}\sim(Bv)^{-1}\,b_n^{1-v}$ *as* $n\to\infty$ *and* $b_n=(B^{-1}\log n)^{1/v}-u\log(B^{-1}\log n)/v^2B^{1/v}(\log n)^{(v-1)/v}$.

Proof. When $u=0$, (a) follows easily. When $u>0$ for $t>0$,

$$\frac{1}{t^v}\int_t^\infty x^{-u+v-1}\exp(-Bx^v)\,dx$$
$$>\frac{1}{u}t^{-u}\exp(-Bt^v)-\frac{Bv}{u}\int_t^\infty x^{-u+v-1}\exp(-Bx^v)\,dx,$$

whence

$$At^{-u}\exp(-Bt^v)=\int_t^\infty\left(1+\frac{u}{Bv}x^{-(v-1)}\right)\frac{A}{Bv}x^{-u+v-1}\exp(-Bx^v)\,dx$$
$$>1-F_\varepsilon(t)>ABv\left(\frac{Bv}{u}+\frac{1}{t^v}\right)^{-1}\frac{1}{u}t^{-u}\exp(-Bt^v).$$

The conclusion of (a) follows again for $u>0$.

By Lemma 2(c), Condition E(2), and (a), $F_\varepsilon\in D(\Lambda)$ by simple algebra. We now find the acceptable choices of norming constants. Since $1-F_\varepsilon(x)\sim Ax^{-u}$

$\exp(-Bx^v)$, taking the logarithm of both sides of $Ab_n^{-u}\exp(-Bb_n^v)=n^{-1}$ gives

$$-\log A+u\log b_n+Bb_n^v=\log n. \tag{12}$$

Hence $b_n\to\infty$ and $b_n\sim(B^{-1}\log n)^{1/v}$ by dividing both sides of (12) by b_n^v. Since $a_n=(nf(b_n))^{-1}$ we see that an acceptable choice for a_n is $(Bv)^{-1}(B^{-1}\log n)^{(1-v)/v}$.

Next, try an expansion of b_n by writing $b_n=(B^{-1}\log n)^{1/v}+r_n$, where r_n is a remainder which is $o((\log n)^{1/v})$. Substitute this b_n into (12) and we find

$$o(1)-\frac{u}{v}\log(B^{-1}\log n)+(\log n)\left\{\left[1+\frac{r_n}{(B^{-1}\log n)^{1/v}}\right]^v-1\right\}=0.$$

Hence we conclude that

$$b_n=(B^{-1}\log n)^{1/v}-\frac{u\log(B^{-1}\log n)}{v^2B^{1/v}(\log n)^{(v-1)/v}}. \quad \blacksquare$$

Proof of Theorem 1(b). Recall that

$$P(\varepsilon_{N:N,l}-\varepsilon_{N:N,j}\leqslant -d_{nj})=\int_{-\infty}^{\infty}[F_\varepsilon(t-d_{nlj})]^N N[F_\varepsilon(t)]^{N-1}f_\varepsilon(t)\,dt,$$

where $N[F_\varepsilon(t)]^{N-1}f_\varepsilon(t)$ is the density function of $\varepsilon_{N:N,j}$. The proof argument is motivated by the following heuristic. Since $a_N^{-1}(\varepsilon_{N:N,j}-b_N)\to\Lambda$ in distribution by Lemma 2, it is then expected that $P(\varepsilon_{N:,l}-\varepsilon_{N:N,j}\leqslant -d_{nlj})\to 0$ when $d_{nlj}a_N^{-1}\to\infty$. To avoid notational complexity, we only consider $r=1$. For fixed r, the result can be derived accordingly.

The above-mentioned probability will be evaluated by dividing $(-\infty,\infty)$ into three intervals $(-\infty,a_0)$, $[a_0,b_0)$, and $[b_0,\infty)$ with $a_0=0$ and $b_0=b_N+c_Na_N$, where $c_N=[4(v-1)/v]\log\log N$. Observe that

$$\begin{aligned}&\int_{b_0}^{\infty}[F_\varepsilon(t-d_{nKj})]^N N[F_\varepsilon(t)]^{N-1}f_\varepsilon(t)\,dt\\ &\qquad\leqslant N\int_{b_0}^{\infty}[F_\varepsilon(t)]^{2N-1}f_\varepsilon(t)\,dt\leqslant\frac{1}{2}[1-F_\varepsilon^{2N}(b_0)]\end{aligned}$$

and

$$\begin{aligned}&\int_{-\infty}^{a_0}[F_\varepsilon(t-d_{nKj})]^N N[F_\varepsilon(t)]^{N-1}f_\varepsilon(t)\,dt\\ &\qquad\leqslant[F_\varepsilon^{2N-1}(0)]\int_{-\infty}^{a_0}f_\varepsilon(t)\,dt\leqslant N[F_\varepsilon(0)]^{-2N}.\end{aligned} \tag{13}$$

Since $1-F_\varepsilon(t) \sim At^{-u}\exp(-Bt^v)$ and $b_0 \to \infty$, we have

$$\begin{aligned} F_\varepsilon(b_0) &\geqslant 1-2Ab_N^{-u}e^{-Bb_0^v} \\ &\geqslant 1-\frac{2A}{(B^{-1}\log N)^{u/v}}\exp(-Bb_N^v)\exp(-Bvc_Na_Nb_N^{v-1}/2) \\ &\geqslant 1-\frac{1}{N}\frac{A}{(\log N)^{(u+2v-2)/v}} \end{aligned}$$

and

$$\begin{aligned} 1-F_\varepsilon^{2N}(b_0) &\leqslant \left(1-\frac{A}{N(\log N)^{(u+2v-2)/v}}\right)^{2N} \\ &\leqslant 2A(\log N)^{-(u+2v-2)/v}. \end{aligned}$$

Hence

$$\int_{b_0}^{\infty}[F_\varepsilon(t-d_{nKj})]^N N[F_\varepsilon(t)]^{N-1}f_\varepsilon(t)\,dt \leqslant 2A(\log N)^{-(u+2v-2)/v}. \qquad (14)$$

Note that (II) in (7) can be written as

$$\begin{aligned} &Na_{N-1}\int_{(a_0-b_{N-1})/a_{N-1}}^{(b_0-b_{N-1})/a_{N-1}}\Lambda\left(\frac{a_{N-1}t+b_{N-1}-b_N-d_{nKj}}{a_N}\right) \\ &\qquad \times\Lambda(t)\,f_\varepsilon(a_{N-1}t+b_{N-1})\,dt \\ &\quad N\int_0^{b_0}\Lambda\left(\frac{t-b_N-d_{nKj}}{a_N}\right)\Lambda\left(\frac{t-b_{N-1}}{a_{N-1}}\right)f_\varepsilon(t)\,dt \\ &\quad = N\int_0^{b_0}\left[\Lambda\left(\frac{t-b_N}{a_N}\right)\right]^{\exp(a_N^{-1}d_{nKj})}\Lambda\left(\frac{t-b_{N-1}}{a_{N-1}}\right)f_\varepsilon(t)\,dt. \qquad (15) \end{aligned}$$

Observe that $\Lambda(0)=\exp(-1)$,

$$\Lambda\left(\frac{b_0-b_N}{a_N}\right)=\Lambda(c_N)=\exp(-e^{-c_N}) \qquad (16)$$

$$\begin{aligned} \Lambda\left(\frac{cb_N-b_N}{a_N}\right) &\leqslant 2\Lambda(v(c-1)\log N) \\ &= 2\exp(-N^{v(1-c)}) \qquad \text{for} \quad 0<c<1, \qquad (17) \end{aligned}$$

$$\Lambda\left(\frac{b_0^*-b_N}{a_N}\right)=\Lambda(\log 4)=\exp(-0.25),$$

where $b_0^* = b_N + a_N \log 4$; (17) is derived by using $a_N^{-1} \sim Bvb_N^{v-1}$ and $Bb_N^v = \log N$.

Now, we find an upper bound for the right-hand side of (15) by dividing the interval of integration $[0, b_0]$ into $[0, cb_N)$, and $[b_N, b_0]$. By (16), (17), $N[1 - F_\varepsilon(b_N)] = 1$, and $\Lambda(\cdot) \leqslant 1$, we have

$$N \int_{b_N}^{b_0} \left[\Lambda\left(\frac{t - b_N}{a_N}\right)\right]^{\exp(a_N^{-1} d_{nKj})} \Lambda\left(\frac{t - b_{N-1}}{a_{N-1}}\right) f_\varepsilon(t)\, dt$$

$$\leqslant [\Lambda(c_N)]^{\exp(a_N^{-1} d_{nKj})} N[F_\varepsilon(b_0) - F_\varepsilon(b_N)] \leqslant \exp(-e^{a_N^{-1} d_{nKj} - c_N}), \qquad (18)$$

$$N \int_{cb_N}^{b_N} \left[\Lambda\left(\frac{t - b_N}{a_N}\right)\right]^{\exp(a_N^{-1} d_{nKj})} \Lambda\left(\frac{t - b_{N-1}}{a_{N-1}}\right) f_\varepsilon(t)\, dt$$

$$\leqslant [\exp(-\exp(a_N^{-1} d_{nKj}))][Nf_\varepsilon(b_N)]$$

$$\times \frac{[F_\varepsilon(b_N) - F_\varepsilon(cb_N)]/(b_N - cb_N)}{f_\varepsilon(b_N)} (1-c)\, b_N$$

$$\leqslant 2(1-c) \frac{b_N}{a_N} \exp(-\exp(a_N^{-1} d_{nKj}))$$

$$\leqslant 2v(1-c)(\log N) \exp(-\exp(a_N^{-1} d_{nKj})), \qquad (19)$$

$$N \int_0^{cb_N} \left[\Lambda\left(\frac{t - b_N}{a_N}\right)\right]^{\exp(a_N^{-1} d_{nKj})} \Lambda\left(\frac{t - b_{N-1}}{a_{N-1}}\right) f_\varepsilon(t)\, dt$$

$$< N[\exp(-N^{v(1-c)})]^{\exp(a_N^{-1} d_{nKj})}. \qquad (20)$$

Note that (III) in (7) can be evaluated similarly. We have

$$e_{N-1} N a_{N-1} \int_{(a_0 - b_{N-1})/a_{N-1}}^{(b_0 - b_{N-1})/a_{N-1}} \Lambda\left(\frac{a_{N-1} t + b_{N-1} - b_N - d_{nKj}}{a_N}\right)$$

$$\times f_\varepsilon(a_{N-1} t + b_{N-1})\, dt$$

$$\leqslant e_{N-1}\{\exp(-\exp(a_N^{-1} d_{nKj} - c_N))$$

$$+ 2v(1-c)(\log N) \exp(-\exp(a_N^{-1} d_{nKj}))$$

$$+ N[\exp(-N^{v(1-c)})]^{\exp(a_N^{-1} d_{nKj})}\}. \qquad (21)$$

Note that $e_{N-1} = o(1)$ by Lemma 2(d). The magnitude of (III) in (7) is of small order (II) in (7).

It follows from (5), (6), (13)–(15), and (18)–(21) that

$$\begin{aligned} P\Big(Y_{N:N,K} \geqslant \sup_{1\leqslant j\leqslant K-J} Y_{N:N,j}\Big) &\geqslant 1-\left[\frac{K}{(\log N)^{(u+2v-2)/v}}+N[F_\varepsilon(0)]^{2N}+O(e_N)\right] \\ &\quad -\sum_{j=1}^{K-J}\Bigg[\exp(-\exp(a_N^{-1}d_{nKj}-c_N)) \\ &\quad +2v(1-c)(\log N)\left(\frac{1}{e}\right)^{\exp(a_N^{-1}d_{nKj})} \\ &\quad +N[\exp(-N^{v(1-c)})]^{\exp(a_N^{-1}d_{nKj})}\Bigg]. \end{aligned}$$

Again, when Condition R holds with $\tau=1$, we have

$$\begin{aligned} &(\log N)\sum_{j=1}^{K-J} e^{-\exp(a_N^{-1}d_{nKj})} \\ &\qquad \leqslant K(\log N)\exp\left[-\exp\left(Bv\left(\frac{\log N}{B}\right)^{(v-1)/v} JK^{-1}\right)\right], \end{aligned}$$

by the same argument used in Section 5.1. Set $K=Bv(B^{-1}\log n)^{(v-1)/v}(\log\log n)^{-1}$. Hence, $e_N=o(1)$ by Lemma 2(d). It follows easily that $K(\log N)^{-(u+2v-2)/v}=o(1)$, $K\log N\sum_{j=1}^{K-J}\exp(-\exp(a_N^{-1}d_{nKj}))=o(1)$, $N[F_\varepsilon(0)]^{2N}=o(1)$,

$$\begin{aligned} &\sum_{j=1}^{K-J} e^{-\exp(a_N^{-1}d_{nKj}-c_N)} \\ &\qquad \leqslant K\exp\left(-\exp\left(\left(\frac{\log N}{B}\right)^{(v-1)/v}\frac{J\log\log n}{(B^{-1}\log n)^{(v-1)/v}}-\log\log n\right)\right) \\ &\qquad \leqslant K\exp(-(\log n)^{J/2-1})=o(1) \end{aligned}$$

and

$$\begin{aligned} &N\sum_{j=1}^{K-J}[\exp(-N^{v(1-c)})]^{\exp(a_N^{-1}d_{nKj})} \\ &\qquad \leqslant n[\exp(-N^{v(1-c)})]^{\exp(a_N^{-1}d_{nKj})}=o(1). \end{aligned}$$

The above discussions conclude that $Z_{[n:n]}-w_Z=O_p((\log n)^{-(v-1)/v}\log\log n)$ under Condition R with $\tau=1$.

For general τ, we have

$$(\log N)\sum_{j=1}^{K-J} e^{-\exp(a_N^{-1}d_{nKj})} \leqslant K(\log N)\exp[-\exp((2\log N)^{(v-1)/v}\, JK^{-1/\tau})].$$

Set $K=(\log n)^{((v-1)/v)\tau}(\log\log n)^{-\tau}$. Applying the same argument used in deriving the result with $\tau=1$, we have $Z_{[n:n]}-w_Z=O_p((\log n)^{-((v-1)/v)\tau}(\log\log n)^{\tau})$ under Condition R. ▮

Acknowledgments

The authors thank the referees for helpful comments which led to significant improvements of this paper.

References

[1] Banks, D. (1993). Is industrial statistics out of control? (with discussions). *Statist. Sci.* **8** 356–409.

[2] Bechhofer, R. E. (1954). A single sample multiple procedure for ranking means of normal populations with known variances. *Ann. Math. Statist.* **25** 16–39.

[3] Bhattacharya, P. K. (1984). Induced order statistics: Theory and applications. In *Handbook of Statistics* (P. R. Krishnaiah and P. K. Sen, Eds.), Vol. 4, pp. 383–403, Elsevier, Amsterdam.

[4] Box, G. E. P., and Wilson, K. B. (1951). On the experimental attainment of optimum conditions. *J. Roy. Statist. Soc. Ser. B* **13** 1–45.

[5] Changchien, G. M. (1990). Optimization of blast furnace burden distribution. In *Proceedings of the 1990 Taipei Symposium in Statistics, June 28-30, 1990* (M. T. Chao and P. E. Cheng, Eds.), pp. 63–78.

[6] Chen, H. (1988). Lower rate of convergence for locating a maximum of a function. *Ann. Statist.* **16** 1330–1334.

[7] Fabian, V. (1967). Stochastic approximation of minima with improved asymptotic speed. *Ann. Math. Statist.* **38** 91–200.

[8] Haan, L. de (1981). Estimation of the minimum of a function using order statistics. *J. Amer. Statist. Assoc.* **76** 467–469.

[9] Hotelling, H. (1941). Experimental determination of the maximum of a function. *Ann. Math. Statist.* **12** 20–45.

[10] Kiefer, J., and Wolfowitz, J. (1952). Stochastic estimation of the maximum of a regression function. *Ann. Math. Statist.* **23** 462–466.

[11] Kiefer, J. (1972). Iterated logarithm analogues of sample quantiles when $p_n \downarrow 0$. In *Proc. Sixth Berkeley Symposium on Math. Statist. and Probab.*, Vol. 1, pp. 227–244. University of California Pres, Berkeley.

[12] Lawson, JU. S. (1988). A case study of effective use of statistical experimental design in a smoke stack industry. *J. Quality Technol.* **20** 51–62.

[13] Lo, S. H. (1989). On some representations of the bootstrap. *Probab. Theory Related Fields* **83** 411–418.

[14] Müller, H. G. (1985). Kernel estimators of zeros and of location and size of extrema of regression functions. *Scand. J. Statist.* **12** 221–232.

[15] Müller, H. G. (1989). Adaptive nonparametric peak estimation. *Ann. Statist.* **17** 1053–1069.

[16] Reiss, R. D. (1989). *Approximate Distributions of Order Statistics: With Applications to Nonparamatric Statistics.* Springer-Verag, New York.

[17] Rohatgi, M. S. (1962). On the asymptotic sufficiency of certain order statistics. *J. Roy. Statist. Soc. Ser. B* **24** 167–176.

[18] Tsybakov, A. B. (1988). *Passive Stochastic Approximation.* University of Bonn, SFB 303 Discussion Paper.