The Annals of Statistics
2000, Vol. 28, No. 4, 1083–1104

ON THE DEGREES OF FREEDOM IN SHAPE-RESTRICTED REGRESSION

BY MARY MEYER AND MICHAEL WOODROOFE[1]

University of Georgia and University of Michigan

For the problem of estimating a regression function, μ say, subject to shape constraints, like monotonicity or convexity, it is argued that the divergence of the maximum likelihood estimator provides a useful measure of the effective dimension of the model. Inequalities are derived for the expected mean squared error of the maximum likelihood estimator and the expected residual sum of squares. These generalize equalities from the case of linear regression. As an application, it is shown that the maximum likelihood estimator of the error variance σ^2 is asymptotically normal with mean σ^2 and variance $2\sigma^2/n$. For monotone regression, it is shown that the maximum likelihood estimator of μ attains the optimal rate of convergence, and a bias correction to the maximum likelihood estimator of σ^2 is derived.

1. Introduction. In shape-restricted regression problems, there are observations of the form

$$y_k = \mu(x_k) + \sigma\varepsilon_k, \qquad k = 1, \ldots, n, \tag{1}$$

where $\varepsilon_1, \ldots, \varepsilon_n$ are independent, standard normal errors, $-\infty < x_1 < \cdots < x_n < \infty$ are design points and the regression function μ is known to possess a qualitative property such as monotonicity or convexity. Let $\mathscr{F}$ denote the set of possible regression functions and suppose throughout the paper that $\mathscr{F}$ is a convex set of functions. Next, let $y = (y_1, \ldots, y_n)'$, $\theta = [\mu(x_1), \ldots, \mu(x_n)]'$, and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)'$, where $'$ denotes transpose. Then the model (1) may be written as

$$y = \theta + \sigma\varepsilon, \tag{2}$$

and the problem is to estimate θ, subject to the constraints imposed by the properties of $\mathscr{F}$. Typically, the latter constraints may be written in the form $\theta \in \Omega$, where Ω is a closed convex subset of $\mathbb{R}^n$. For example, if $\mathscr{F}$ is the class of nondecreasing functions on the interval $[x_1, x_n]$, then the constraints are $\theta_{k+1} - \theta_k \geq 0$ for $k = 1, \ldots, n-1$, and the set of θ which satisfy these constraints is a closed convex set. In fact, this Ω is a closed convex polyhedron, a set of the form $\Omega = \{\omega \in \mathbb{R}^n: \gamma_k'\omega \geq 0, k = 1, \ldots, m\}$, where $\gamma_1, \ldots, \gamma_m \in \mathbb{R}^n$. Constraints like concavity and convexity in (1) lead to convex polyhedral Ω in (2) too.

Received July 1998; revised January 2000.

[1]Supported in part by NSF Grant and U.S. Army Research Office.

AMS 1991 *subject classification*. Primary 62G08.

Key words and phrases. Asymptotic distribution, bias reduction, divergence, effective dimension, simulation, Stein's identity, variance estimation.

For any closed convex subset $\Omega \subseteq \mathbb{R}^n$, the maximum likelihood estimator (MLE), $\hat{\theta}$ say, of θ in the model (2) minimizes $\|\theta - y\|^2$ with respect to $\theta \in \Omega$. It is well known that the unique minimizing vector $\hat{\theta} = \hat{\theta}(y)$ is the projection of y onto Ω. The latter is characterized by the conditions

$$\hat{\theta} \in \Omega \quad \text{and} \quad \langle y - \hat{\theta}, \hat{\theta} - \omega \rangle \geq 0, \tag{3}$$

for all $\omega \in \Omega$. Moreover, if Ω is a convex cone, so that $c\omega \in \Omega$ for all $c \geq 0$ and all $\omega \in \Omega$, then (3) is equivalent to

$$\hat{\theta} \in \Omega, \quad \langle y - \hat{\theta}, \hat{\theta} \rangle = 0 \quad \text{and} \quad \langle y - \hat{\theta}, \omega \rangle \leq 0, \tag{4}$$

for all $\omega \in \Omega$, and the last inequality in (4) holds for all $\omega \in \mathbb{R}^n$ for which $\hat{\theta} + \varepsilon\omega \in \Omega$ for some $\varepsilon > 0$.

In this paper, we explore the role of the divergence for the MLE,

$$D = \operatorname{div}(\hat{\theta}) = \sum_{i=1}^{n} \frac{\partial}{\partial y_i} \hat{\theta}_i(y) \tag{5}$$

and argue that D provides a measure of the effective dimension of the model. To see how this conjecture generalizes simpler models, observe that if Ω is a linear space of dimension d, say, then $\hat{\theta} = Qy$, where Q is the projection matrix onto Ω, and $D(y) \equiv \operatorname{tr}(Q) = d$ for all y. Less transparently, $D(y)$ is the number of distinct values among $\hat{\theta}_1(y), \ldots, \hat{\theta}_n(y)$ for monotone regression, by an easy application of Proposition 1 below. Support for the interpretation of D takes two forms. For convex polyhedral Ω, $\hat{\theta}(y)$ is the projection of y onto a subspace of dimension $D(y)$, as described in Proposition 1. Further, there are two inequalities that generalize equalities from the linear case: $E_{\sigma,\mu}\|\hat{\theta} - \theta\|^2 \leq \sigma^2 E_{\sigma,\mu}(D)$ and $E_{\sigma,\mu}\|y - \hat{\theta}\|^2 \leq \sigma^2 E_{\sigma,\mu}(n - D)$, where $E_{\sigma,\mu}$ denotes expectation in the model (1). Both inequalities may be strict, however, even asymptotically after renormalization, so that analogies with the linear case are incomplete. Using these inequalities, it is shown that the MLE of σ^2 is asymptotically normal with mean σ^2 and variance $2\sigma^4/n$ under a mild growth condition on $E_{\mu,\sigma}(D)$. This result is obtained for a class of estimators that allows for bias reduction of the MLE: for the important special case of monotone regression, it is shown that $E_{\sigma,\mu}(D) \leq Cn^{1/3}$, where C depends only on $\Delta = [\mu(x_n) - \mu(x_1)]/\sigma$, and it follows easily from this that $\hat{\theta}$ attains the optimal rate of convergence. Under additional regularity conditions, it is also shown that there are constants $c_0 \approx 0.5$ and $c_1 \approx 1.5$ for which $E_{\sigma,\mu}\|\hat{\theta} - \theta\|^2 = c_0\sigma^2 E_{\sigma,\mu}(D) + o(n^{1/3})$ and $E_{\sigma,\mu}\|y - \hat{\theta}\|^2 = n - c_1\sigma^2 E_{\sigma,\mu}(D) + o(n^{1/3})$. In this sense, $c_0 D$ and $n - c_1 D$ are better candidates for the terms "effective dimension" and "residual degrees of freedom" than D and $n - D$. These results have implications for variance estimation. It is shown that $\tilde{\sigma}^2 = \|y - \hat{\theta}\|^2/(n - c_1 D^*)$ is asymptotically unbiased to order $n^{-2/3}$, where D^* denotes a truncated version of D. The bias corrected MLE $\tilde{\sigma}^2$ has a smaller asymptotic variance than the omnibus estimators of Gasser, Sroka and Jennen-Steinmetz (1986) and Rice (1984), though still a larger bias. The moderate sample size properties of the estimators are compared in a simulation study.

The term "effective dimension" is adapted from Hastie and Tibshirani (1990), who propose three different possible definitions for linear smoothers (of which the first is equal to the divergence of the estimator). They too find that the expected residual sum of squares may differ from σ^2 times n minus the effective dimension. There is precedent for the use of D as degrees of freedom in monotone regression. For testing H_0: $\mu(x) \equiv c$ in monotone regression with a known σ^2, the null distribution of the chi-squared statistic is a mixture of chi-squared distributions and the mixing distribution is the null distribution of D. See Theorem 2.3.1 of Robertson, Wright and Dyskstra (1988).

2. Shape restricted regression.

The divergence. For this section suppose that Ω is a closed convex set in (2). Then $\hat{\theta}$ is determined by (3). The first item of business is to show that D is well defined. Recall from Stein (1981) that a function f: $\mathbb{R}^n \to \mathbb{R}$ is said to be *almost differentiable* if there is a function g: $\mathbb{R}^n \to \mathbb{R}^n$ for which

$$f(x+y) - f(x) = \int_0^1 y' g(x+ty)\, dt \tag{6}$$

for a.e. $x \in \mathbb{R}^n$ for each $y \in \mathbb{R}^n$. Then g is essentially unique; g is called *the gradient of* f and denoted by $g = \nabla f$ or $g(y) = [\partial f(y)/\partial y_1, \ldots, \partial f(y)/\partial y_n]'$. It is not difficult to see that any Lipschitz continuous function f is almost differentiable with a bounded gradient. A simple proof is to convolve f with a normal density with mean 0 and covariance matrix $h^2 I_n$ and then show that the gradients of the convolutions are uniformly bounded in $L^\infty(\mathbb{R}^n)$ and, therefore, have a weak limit point as $h \to 0$. The details may be found in Meyer and Woodroofe (1998).

PROPOSITION 1. *The components of* $\hat{\theta}$ *are almost differentiable, and* $\nabla\hat{\theta}_i$ *is an essentially bounded function for each* $i = 1, \ldots, n$. *If* Ω *is a convex polyhedron, say*

$$\Omega = \{\omega \in \mathbb{R}^n\colon \gamma_i'\omega \geq 0, i = 1, \ldots, m\},$$

where $\gamma_1, \ldots, \gamma_m \in \mathbb{R}^n$, *then* $\hat{\theta}(y)$ *is the projection of* y *onto the linear space,*

$$K^y = \{\omega \in \mathbb{R}^n\colon \gamma_i'\omega = 0 \text{ for all } i \text{ for which } \gamma_i'\hat{\theta}(y) = 0\}$$

and $D(y) = \dim(K^y)$ *for a.e.* $y \in \mathbb{R}^n$.

PROOF. To show almost differentiability and essential boundedness, it suffices to show Lipschitz continuity; that is, if $y^i \in \mathbb{R}^n$, $i = 1, 2$, then $\|\hat{\theta}(y^2) - \hat{\theta}(y^1)\| \leq \|y^2 - y^1\|$. To see this, let $\hat{\theta}^i = \hat{\theta}(y^i)$, $i = 1, 2$. Then it follows from (3) that $\langle y^1 - \hat{\theta}^1, \hat{\theta}^2 - \hat{\theta}^1\rangle \leq 0$ and $\langle y^2 - \hat{\theta}^2, \hat{\theta}^1 - \hat{\theta}^2\rangle \leq 0$. Adding these two inequalities leads to $\langle (y^1 - y^2) - (\hat{\theta}^1 - \hat{\theta}^2), \hat{\theta}^2 - \hat{\theta}^1\rangle \leq 0$, which implies $\|\hat{\theta}^2 - \hat{\theta}^1\|^2 \leq \langle y^2 - y^1, \hat{\theta}^2 - \hat{\theta}^1\rangle \leq \|\hat{\theta}^2 - \hat{\theta}^1\| \times \|y^2 - y^1\|$ and, therefore, $\|\hat{\theta}^2 - \hat{\theta}^1\| \leq \|y^2 - y^1\|$.

For subsets $J \subseteq \{1, \ldots, m\}$, let K_J be the linear subspace $K_J = \{\omega \in \mathbb{R}^n\colon \gamma_i'\omega = 0, \text{ for all } i \in J\}$; let $\mathscr{K}$ be the collection of distinct subspaces of

this form, and let Q_K denote the projection operator onto K. Thus, each Q_K is an $n \times n$ matrix for which $Q_K y \in K$ and $\langle y - Q_K y, \xi \rangle = 0$ for all $\xi \in K$ for each $y \in \mathbb{R}^n$. Let $J_y = \{i \leq m: \gamma_i' \hat{\theta}(y) = 0\}$, and write K^y and Q^y for K_{J_y} and Q_{K^y}, as in the statement of the proposition. It is clear that $\hat{\theta} \in K^y$ and therefore suffices to show that $\langle y - \hat{\theta}, \xi \rangle = 0$ for all $\xi \in K^y$. This follows easily from (4). For if $\xi \in K^y$, then $\gamma_i'(\hat{\theta} \pm \varepsilon\xi) = 0$ for all $i \in J_y$, and $\gamma_i'(\hat{\theta} \pm \varepsilon\xi) > 0$ for all $i \notin J_y$ for all sufficiently small ε, since $\gamma_i' \hat{\theta} > 0$ for all $i \notin J_y$. That is, $\hat{\theta} \pm \varepsilon\xi \in \Omega$ for all sufficiently small $\varepsilon > 0$, so that $\langle y - \hat{\theta}, \xi \rangle = 0$, by (4).

Now let $B_K = \{y \in \mathbb{R}^n: K^y = K\}$ for $K \in \mathscr{K}$, and let $\overline{B}_K$ and B_K^o denote the closure and interior of B_K. Then it is clear that $D = \dim(K^y)$, for a.e. $y \in B_K^o$. So, it remains to show that the boundary of each B_K has measure zero, and for this it suffices to show that $\overline{B}_K \cap \overline{B}_L$ has measure zero for all $K \neq L$, since $\mathscr{K}$ is finite. This is clear, however; for if $y \in \overline{B}_K \cap \overline{B}_L$, then $Q_K y = \hat{\theta}(y) = Q_L y$ and therefore $(Q_K - Q_L)y = 0$. □

For monotone regression, the maximum likelihood estimator of θ is

$$\hat{\theta}_k = \max_{j<k} \min_{l \geq k} \bar{y}_{j,l}, \tag{7}$$

where $\bar{y}_{j,l} = (y_{j+1} + \cdots + y_l)/(l - j)$. See Robertson, Wright, and Dykstra [(1989), page 23]. For a given $y \in \mathbb{R}^n$, let $1 \leq r_1 < \cdots < r_m \leq n$ be the values of k for which $\hat{\theta}_{r_k} > \hat{\theta}_{r_k - 1}$, where $\hat{\theta}_0$ is to be interpreted as $-\infty$. Then $J_y = \{1, \ldots, r_1 - 1, r_1 + 1, \ldots, r_2 - 1, \ldots, r_m + 1, \ldots, n\}$, and $D(y) = m$, the number of distinct values of $\hat{\theta}_1, \ldots, \hat{\theta}_n$.

Risk inequalities. Recall that expectation in the model (1) is denoted by $E_{\sigma,\mu}$. Then it follows from Proposition 1 and Stein's (1981) identity that

$$E_{\sigma,\mu}[\langle y - \theta, \hat{\theta} \rangle] = \sigma^2 E_{\sigma,\mu}(D), \tag{8}$$

for any $\mu \in \mathscr{F}$ and $\sigma > 0$, where $\theta = [\mu(x_1), \ldots, \mu(x_n)]'$, as in (2). The next result provides an unbiased estimator of the risk for the case in which σ is known.

PROPOSITION 2.

$$E_{\sigma,\mu}\|\hat{\theta} - \theta\|^2 = E_{\sigma,\mu}(U), \tag{9}$$

where

$$U = \|y - \hat{\theta}\|^2 + 2\sigma^2 D - n\sigma^2.$$

Further,

$$E_{\sigma,\mu}\|\hat{\theta} - \theta\|^2 \leq \sigma^2 E_{\sigma,\mu}(D). \tag{10}$$

PROOF. Clearly,

$$\|y - \hat{\theta}\|^2 = \|y - \theta\|^2 - 2\langle y - \theta, \hat{\theta} - \theta \rangle + \|\hat{\theta} - \theta\|^2. \tag{11}$$

Using (8), the expectations of the first two terms on the right side of (11) may be computed as $E_{\sigma,\mu}\|y-\theta\|^2 = n\sigma^2$ and $E_{\sigma,\mu}[\langle y-\theta, \hat{\theta}-\theta\rangle] = \sigma^2 E_{\sigma,\mu}(D)$, and (9) then follows from substitution. For (10), observe that $0 \le \langle y-\hat{\theta}, \hat{\theta}-\theta\rangle = \langle y-\theta, \hat{\theta}-\theta\rangle - \|\hat{\theta}-\theta\|^2$, by (3). So $0 \le E_{\sigma,\mu}[\langle y-\theta, \hat{\theta}-\theta\rangle] - E_{\sigma,\mu}\|\hat{\theta}-\theta\|^2 = \sigma^2 E_{\sigma,\mu}(D) - E_{\sigma,\mu}\|\hat{\theta}-\theta\|^2$. □

If σ^2 were known, or could be estimated well, then Proposition 2 could be used to assess the quality of a fit, but this is not the primary use here. Rather, interest centers on the inequalities in (10) and Corollary 1 below.

COROLLARY 1. $n\sigma^2 - 2\sigma^2 E_{\sigma,\mu}(D) \le E_{\sigma,\mu}\|y-\hat{\theta}\|^2 \le n\sigma^2 - \sigma^2 E_{\sigma,\mu}(D)$.

PROOF. By (9), $E_{\sigma,\mu}\|y-\hat{\theta}\|^2 = n\sigma^2 - 2\sigma^2 E_{\sigma,\mu}(D) + E_{\sigma,\mu}\|\hat{\theta}-\theta\|^2$, and $0 \le E_{\sigma,\mu}\|\hat{\theta}-\theta\|^2 \le \sigma^2 E_{\sigma,\mu}(D)$, by (10). □

COROLLARY 2. *Let $\Omega_0 \subseteq \Omega_1$ be two closed convex subsets of $\mathbb{R}^n$, and let $\hat{\theta}_0$ and $\hat{\theta}_1$ be the maximum likelihood estimators for the parameter spaces Ω_0 and Ω_1. If $\theta \in \Omega_0$, then $E_{\sigma,\mu}(D_0) \le 2E_{\sigma,\mu}(D_1)$, where $D_i = \operatorname{div}(\hat{\theta}_i)$.*

PROOF. Let $r_i = E_{\sigma,\mu}\|\hat{\theta}_i-\theta\|^2$, $i = 1, 2$. Then

$$\begin{aligned} r_1 - 2\sigma^2 E_{\sigma,\mu}(D_1) &= E_{\sigma,\mu}\|y-\hat{\theta}_1\|^2 - n\sigma^2 \\ &\le E_{\sigma,\mu}\|y-\hat{\theta}_0\|^2 - n\sigma^2 = r_0 - 2\sigma^2 E_{\sigma,\mu}(D_0) \end{aligned}$$

by Proposition 1 and the assumption $\Omega_0 \subseteq \Omega_1$. So, using (10), $2E_{\sigma,\mu}(D_0) \le (r_0 - r_1)/\sigma^2 + 2E_{\sigma,\mu}(D_1) \le E_{\sigma,\mu}(D_0) + 2E_{\sigma,\mu}(D_1)$. □

Variance estimation. The MLE of σ^2 is $\hat{\sigma}^2 = \|y-\hat{\theta}\|^2/n$, and one may ask whether $\hat{\sigma}^2$ is asymptotically normal and efficient. Since $\sqrt{n}(\hat{\sigma}^2 - \sigma^2) = (\|y-\theta\|^2 - n\sigma^2)/\sqrt{n} + R/\sqrt{n}$, where $R = \|y-\hat{\theta}\|^2 - \|y-\theta\|^2 = -2\langle y-\hat{\theta}, \hat{\theta}-\theta\rangle - \|\hat{\theta}-\theta\|^2$, normality and efficiency would follow from $R = o_p(\sqrt{n})$. If Ω is a linear subspace of dimension k, then $\langle y-\hat{\theta}, \hat{\theta}-\theta\rangle = 0$ and $E_{\sigma,\mu}\|\hat{\theta}-\theta\|^2 = k$. So, one may expect that $R = o_p(\sqrt{n})$ whenever Ω can be suitably approximated by linear subspaces of dimension $k_n = o(\sqrt{n})$. This is the essence of Proposition 3 below, though the use of Stein's identity avoids any explicit approximation. It also avoids explicit smoothness assumption on μ.

To allow for bias reduction, consider estimators of the form

$$\tilde{\sigma}^2 = \frac{\|y-\hat{\theta}\|^2}{n - CD}, \tag{12}$$

where $0 \le C = C(y) < n/D$ is a measurable function. The choice of C is discussed in the next section. The following simple result is valid for any choice of C. Write $\hat{\theta}^n$, C_n, D_n and $\tilde{\sigma}_n^2$ for $\hat{\theta}$, C, D and $\tilde{\sigma}^2$ to emphasize the dependence on n and suppose that

$$E_{\sigma,\mu}(D_n) = o(n^\alpha) \tag{13}$$

as $n \to \infty$ for appropriate α. It is shown below that (13) holds for any $\alpha > 1/3$ for monotone regression, and it then follows from Corollary 2 that (13) holds whenever $\mathscr{F}$ consists entirely of monotone functions.

PROPOSITION 3. *Suppose that $C_n \geq 0$, that $C_n D_n < n$ w.p.1 and that C_n are stochastically bounded. If* (13) *holds with $\alpha = 1$, then $\tilde{\sigma}_n^2 \to \sigma^2$ in probability, and if* (13) *holds with $\alpha = 1/2$, then $\sqrt{n}(\tilde{\sigma}_n^2 - \sigma^2)$ is asymptotically normal with mean 0 and variance $2\sigma^4$.*

PROOF. First consider $\hat{\sigma}_n^2$ and write $\hat{\sigma}_n^2 - \sigma^2 = (\|y^n - \theta^n\|^2)/n + R_n/n$, where $R_n = \|y^n - \hat{\theta}^n\|^2 - \|y^n - \theta^n\|^2$, as above. Then, from (10) and the definition of $\hat{\theta}^n$,

$$\begin{aligned} |R_n| &= \|y^n - \theta^n\|^2 - \|y^n - \hat{\theta}^n\|^2 \\ &= 2\langle y^n - \theta^n, \hat{\theta}^n - \theta^n\rangle - \|\hat{\theta}^n - \theta^n\|^2 \leq 2\langle y^n - \theta^n, \hat{\theta}^n - \theta^n\rangle \end{aligned}$$

and

$$E_{\sigma,\mu}[\langle y^n - \theta^n, \hat{\theta}^n - \theta^n\rangle] = \sigma^2 E_{\sigma,\mu}(D_n) = o(n^\alpha),$$

by (13). Thus, $|R_n| = o_p(n^\alpha)$ as $n \to \infty$. The proposition follows for $\hat{\sigma}_n^2$, since $(\|y^n - \theta^n\|^2 - n\sigma^2)/\sqrt{n}$ is asymptotically normal with mean 0 and variance $2\sigma^4$. For $\tilde{\sigma}_n^2$,

$$\tilde{\sigma}_n^2 - \hat{\sigma}_n^2 = \frac{\|y^n - \hat{\theta}^n\|^2}{n(n - C_n D_n)} C_n D_n,$$

which is $o_p(n^{\alpha-1})$, under the conditions of the proposition. □

Rice (1984) suggested two simple and general ways to estimate the residual variance following a nonparametric regression, one based on differences of successive points and one based on the residuals from straightline fits to successive triples, and Rice's suggestions were studied further by Gasser, Sroka, and Jennen-Steinmetz (1986). These estimators are

$$\frac{1}{2(n-1)} \sum_{i=2}^{n} (y_i - y_{i-1})^2 \tag{14}$$

and

$$\frac{1}{2(n-2)} \sum_{i=2}^{n-1} \frac{(a_i y_{i+1} + b_i y_{i-1} - y_i)^2}{a_i^2 + b_i^2 + 1}, \tag{15}$$

where $a_i = (x_i - x_{i-1})/(x_{i+1} - x_{i-1})$ and $b_i = (x_{i+1} - x_i)/(x_{i+1} - x_{i-1})$ for $i = 2, \ldots, n-1$. Compared to these estimations, the estimator $\tilde{\sigma}^2$ of Proposition 3 has a smaller asymptotic variance, though a larger bias. For example, the asymptotic variance of (14) is $3\sigma^3/n$ and, for equally spaced $x_1, \ldots, x_n$, the asymptotic variance of (15) is $(35/9n)\sigma^4$. The bias of $\tilde{\sigma}^2$ is considered in more detail in the next section, and a bias-corrected MLE is compared to (14), (15) and the MLE in Section 4. Preliminary versions of the Propositions 2 and 3 appear in Meyer (1996).

3. Monotone regression. In this section suppose that $\mathscr{F}$ is the class of nondecreasing functions on the interval [0, 1] and that $0 \le x_1 < \cdots < x_n \le 1$.

Approximations and bounds for D_n. Recall the expression (7) for the MLE in the case of monotone regression and that D is the number of distinct value of $\hat{\theta}_1, \ldots, \hat{\theta}_n$ in this case. The following result is proved in Section 5, using (7) and some properties of random walks.

THEOREM 1. *There are absolute constants κ_0, κ_1 and κ_2 for which*

$$E_{\sigma,\mu}(D_n) \le \kappa_0[\Delta + \log(n)] + \kappa_1 \Delta^{2/3} n^{1/3} \tag{16}$$

and

$$E_{\sigma,\mu}(D_n^2) \le \kappa_2^2[\Delta^2 + (1+\Delta)^{4/3} n] \tag{17}$$

for all $n \ge 3$, where

$$\Delta = \frac{\mu(1) - \mu(0)}{\sigma}.$$

The following corollary shows that the maximum likelihood estimator attains the optimal rate of convergence, as described by Donoho and Johnstone (1995) and Efromovich (1997). The result itself is known; it appears in the unpublished manuscript of Donoho (1990) and a closely related result appears in Van der Geer (1990). The proof given here is quite different from earlier ones, however. Consider a sequence of regression problems, as in (13). Then (10) and Theorem 1 combine as follows.

COROLLARY 3. *For any $K > 0$, $E_{\sigma,\mu}\|\hat{\theta}^n - \theta^n\|^2 = O(n^{1/3})$ as $n \to \infty$, uniformly with respect to μ for which $\mu(1) - \mu(0) \le K$.*

The last corollary does not require any smoothness of μ. More detailed conclusions are possible when μ is smooth. Let F_n denote the design distribution function

$$F_n(s) = \frac{\#\{k \le n\colon x_{nk} \le s\}}{n}, \qquad 0 \le s \le 1.$$

Further, let $\mathbb{B}$ denote a standard two-sided Brownian motion; let

$$W(t) = \mathbb{B}(t) + \tfrac{1}{2}t^2, \qquad -\infty < t < \infty; \tag{18}$$

let $\widetilde{W}$ denote the greatest convex minorant of W and let $a = -E[\widetilde{W}(0)]$ and $b = E[\widetilde{W}'(0)^2]$. Then $0 < a, b < \infty$. See Groeneboom (1985, 1989). Let

$$\Delta_0 = \sigma^{4/3} \int_0^1 \nu'(s)^{2/3}\, ds.$$

THEOREM 2. *Suppose also that there is a continuous strictly increasing distribution function F on $[0, 1]$ for which*

$$\Gamma_n := \sup_{0 \le s \le 1} |F_n(s) - F(s)| = o(n^{-1/3}) \tag{19}$$

as $n \to \infty$. Let $\nu(s) = \mu \circ F^{-1}(s)$, $0 \le s \le 1$. Suppose that ν has a positive continuous derivative on $[0, 1]$ and let $\Delta_0 = \sigma^{4/3} \int_0^1 \nu'(s)^{2/3}\, ds$. Then

$$E_{\sigma,\mu}[\langle \theta^n, y^n - \hat{\theta}^n \rangle] \sim -a\Delta_0 n^{1/3} \tag{20}$$

and

$$E_{\sigma,\mu}\|\hat{\theta}^n - \theta^n\|^2 \sim b\Delta_0 n^{1/3} \tag{21}$$

as $n \to \infty$, where $\sim$ means that the ratio of the two sides approaches one as $n \to \infty$.

The theorem is proved in Section 6 by finding the asymptotic distributions of $\hat{\theta}^n$ and $y - \hat{\theta}^n$, suitably normalized, and establishing uniform integrability. Here are some consequences.

COROLLARY 4. *Under the conditions of Theorem 2,*

$$\begin{aligned} \sigma^2 E_{\sigma,\mu}(D_n) &\sim (a+b)\Delta_0, \\ E_{\sigma,\mu}\|\hat{\theta}^n - \theta\|^2 &= c_0\sigma^2 E_{\sigma,\mu}(D_n) + o(n^{1/3}), \\ E_{\sigma,\mu}\|y^n - \hat{\theta}^n\|^2 &= n\sigma^2 - c_1\sigma^2 E_{\sigma,\mu}(D_n) + o(n^{1/3}), \end{aligned}$$

where

$$c_0 = \frac{b}{a+b} \quad \text{and} \quad c_1 = \frac{2a+b}{a+b}. \tag{22}$$

PROOF. Since $\|y^n - \theta^n\|^2 = \|y^n - \hat{\theta}^n\|^2 - 2\langle y^n - \hat{\theta}^n, \theta\rangle + \|\hat{\theta}^n - \theta^n\|^2$, it follows directly from the theorem that

$$E_{\sigma,\mu}\|y^n - \hat{\theta}^n\|^2 = n\sigma^2 - (2a+b)\Delta_0 n^{1/3} + o(n^{1/3})$$

as $n \to \infty$, and since $\|\hat{\theta}^n - \theta^n\|^2 + \langle y^n - \hat{\theta}^n, \hat{\theta}^n - \theta^n\rangle = \langle y - \theta^n, \hat{\theta}^n - \theta^n\rangle$,

$$\sigma^2 E_{\sigma,\mu}(D_n) = E_{\sigma,\mu}[\langle y - \theta^n, \hat{\theta}^n - \theta^n\rangle] = (a+b)\Delta_0 n^{1/3} + o(n^{1/3})$$

as $n \to \infty$. The corollary follows directly from these two observations, (20) and (21). □

Bias reduction in variance estimation. Let

$$D_n^* = \min\left[D_n, \frac{n}{2c_1}\right]$$

and

$$\tilde{\sigma}_n^2 = \frac{\|y^n - \hat{\theta}^n\|^2}{n - c_1 D_n^*}.$$

Then $\tilde{\sigma}_n^2$ is of the form (12) with $C_n = c_1 D_n^*/D_n$.

PROPOSITION 4. *Under the conditions of Theorem* 2, $E_{\sigma,\mu}(D_n - D_n^*) = O(1)$, *and*

$$E_{\sigma,\mu}(\tilde{\sigma}_n^2 - \sigma^2) = o(n^{-2/3}). \tag{23}$$

PROOF. For the first assertion,

$$\begin{aligned} 0 \le E_{\sigma,\mu}(D_n - D_n^*) &\le n P_{\sigma,\mu}\left\{D_n > \frac{n}{2c_1}\right\} \\ &\le n\left(\frac{2c_1}{n}\right)^2 E_{\sigma,\mu}(D_n^2) = O(1), \end{aligned}$$

as $n \to \infty$, by (17). For (23),

$$\tilde{\sigma}_n^2 - \sigma^2 = \frac{\|y^n - \hat{\theta}^n\|^2 - (n - c_1 D_n^*)\sigma^2}{(n - c_1 D_n^*)} = I_n + II_n,$$

where

$$I_n = \frac{\|y^n - \hat{\theta}^n\|^2 - (n - c_1 D_n^*)\sigma^2}{n}$$

and

$$II_n = \frac{\|y^n - \hat{\theta}^n\|^2 - n\sigma^2 + c_1 D_n^* \sigma^2}{n(n - c_1 D_n^*)} c_1 D_n^*.$$

Clearly, $E_{\sigma,\mu}(I_n) = o(n^{-2/3})$, by Corollary 4. Since $n - c_1 D_n^* \ge n/2$, Proposition 4 then follows from

$$E|II_n| \le \frac{2c_1}{n^2}\sqrt{E[(\|y^n - \hat{\theta}^n\|^2 - n\sigma^2)^2]}\sqrt{E(D_n^2)} + \frac{2c_1^2\sigma^2}{n^2}E(D_n^2),$$

(20) and

$$\begin{aligned} E_{\sigma,\mu}[(\|y^n - \hat{\theta}^n\|^2 - n\sigma^2)^2] &\le E_{\sigma,\mu}[\|y^n - \theta^n\|^4] - 2n\sigma^2 E_{\sigma,\mu}[\|y^n - \hat{\theta}^n\|^2] \\ &\quad + n^2\sigma^4 \\ &= O(n^{4/3}), \end{aligned}$$

where Corollary 1 and the relation $E_{\sigma,\mu}\|y^n - \theta^n\|^4 = n(n+2)\sigma^4$ were used to obtain the final equality. □

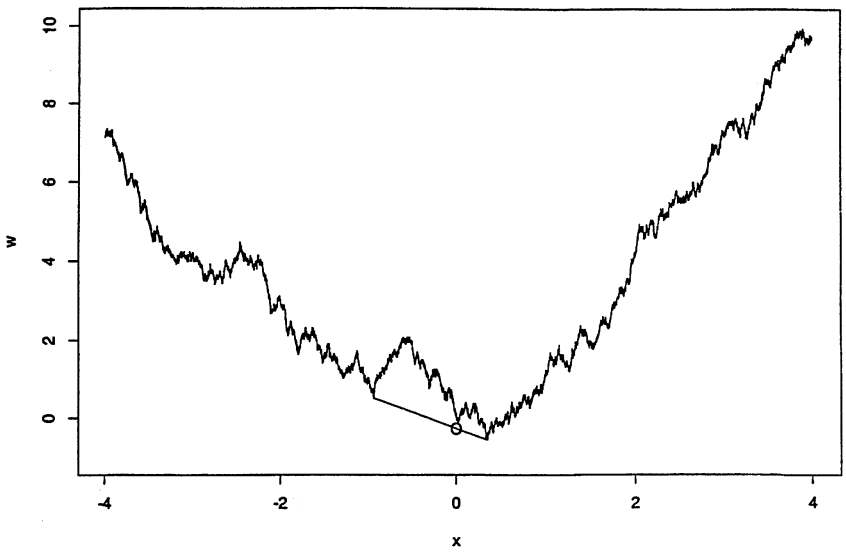

FIG. 1. *An example of the simulated Brownian motion with parabolic drift.*

4. Simulations. A simulation study was conducted to determine the constants c_0 and c_1 described in Section 3, assess the quality of normal approximation to $\tilde{\sigma}^2$, and compare several estimators of σ^2, all in the context of monotone regression. The results are reported in this section.

To use $\tilde{\sigma}^2$, the value of c_1 must be computed, and this was done by simulation. A simulated Brownian motion with parabolic drift is shown in Figure 1 with the greatest convex minorant marked. The step size in this simulation was 0.0001. From this simulation, it is possible to determine the values of $\widetilde{W}(0)$ and $\widetilde{W}'(0)$. Repeating the simulation 100,000 times then leads to confidence intervals for $a = -E[\widetilde{W}(0)]$ and $b = E[\widetilde{W}'(0)^2]$. These were $(0.637, 0.642)$ for a and $(0.644, 0.655)$ for b; a and b were taken to be the midpoints of these intervals, and $c_0 = 0.504$ and $c_1 = 1.496$ were computed from (22).

The bias corrected MLE (BCMLE) was compared to the MLE and the estimators (14) and (15) for three possible regression functions,

$$\mu_1(x) = x,$$
$$\mu_2(x) = e^{16x-8}/(1 + e^{16x-8}),$$
$$\mu_3(x) = (2x - 1)^3 + 1,$$

and the values $\sigma = 0.25, 0.5, 1.0, 2.0$ and $n = 25, 50, 100, 250, 500$. Selected results are reported in Tables 1–3. Reported in these tables are the relative biases, standard deviations and mean squared errors, $E_{\sigma,\mu}(\breve{\sigma}^2)/\sigma^2 - 1$,

$\sqrt{\mathrm{Var}_{\sigma,\mu}(\check{\sigma}^2)}/\sigma^2$, and $E_{\sigma,\mu}[(\check{\sigma}^2-\sigma^2)^2]/\sigma^4$ for each of the four estimators $\check{\sigma}^2 = \tilde{\sigma}^2, \hat{\sigma}^2, (14)$ and (15), along with the risk and Kolmogorov–Smirov distance described below. Each entry was approximated by 2500 Monte Carlo runs.

There were several consistent patterns in the tables. The bias corrections to the MLE overcorrected in the cases considered, leaving $\tilde{\sigma}^2$ with a positive bias that was smaller than the absolute bias of the MLE, but larger than the biases of (14) and (15). Unsurprisingly, the BCMLE had a larger variance than the MLE and a smaller variance than (14) and (15) for larger n($n \geq 100$ in all cases considered and $n \geq 50$ for $\sigma \geq 0.50$). The total mean squared errors of the BCMLE and MLE were quite similar, though the BCMLE appears to do better for larger n. Mean square error, however, is suspect for variance estimation, because it penalizes overestimation much more severely than underestimation. For that reason, another possible loss function was included in the study, $K(\check{\sigma}^2/\sigma^2)$, where

$$K(x) = x - 1 - \log(x).$$

The row labelled "risk" contains $E_{\sigma,\mu}[K(\check{\sigma}^2/\sigma^2)]$. In terms of risk, the BCMLE outperforms the MLE for the cases considered. There are two other consistent relations in the cases considered: the estimator (14) consistently outperforms (15) in terms of variance, mean squared error, and risk, and BCMLE was consistently better (respectively, worse) than (14) for large (respectively, small) n. For $n \geq 100$, BCMLE had smaller mean square error and risk than (14) in all but one of the cases considered; for $n = 25$ (14) had smaller mean square error and risk than BCMLE in all but one of the cases considered. For $n = 50$, there was not much difference between the two. These consistent relations seem intuitive. Both (14) and (15) have small bias, and (14) has smaller asymptotic variance. Further, (14) is simpler and not dependent on asymptotic approximations, while BCMLE is asymptotically efficient, but highly dependent on asymptotic approximations.

The speed of convergence to normality was also considered in the simulations. To reduce the effect of skewness, we considered the distributions of

$$\sqrt{C_n}\,[\log(\check{\sigma}^2) - \log(\sigma^2)], \tag{24}$$

where $C_n = 0.5(n - c_1 D^*)$ for $\tilde{\sigma}^2$, $C_n = 0.5n$ for $\tilde{\sigma}^2$, $C_n = n/3$ for (14) and $C_n = 9n/35$ for (15). The distribution of (24) appears to converge much faster for the BCMLE than for the MLE, but even for the BCMLE the convergence is slow in absolute terms. The convergence is faster for (14) and (15) though still not terribly fast. With such a large Monte Carlo sample size, all values of the Kolmogorov–Smirov statistic are significant at the 5% level.

5. Proof of Theorem 1. It suffices to prove Theorem 1 in the case that $\sigma = 1$ and $\Delta = \mu(1) - \mu(0) > 0$, since σ may be absorbed into μ and the result is known for $\Delta = 0$. See Robertson, Wright and Dykstra [(1988), pages 81, 82]. Since n is fixed throughout this section, it is omitted from the notation. Recall the expression (7) for the maximum likelihood estimator of θ and let A_k^o be

TABLE 1
Comparison of estimators for $\mu_1(x) = x$

σ	n	Measure	BCMLE	MLE	(14)	(15)
0.25	50	bias	0.0779	−0.3004	0.0037	−0.0008
		sd	0.2588	0.1663	0.2443	0.2784
		mse	0.0731	0.1179	*0.0597*	0.0775
		risk	0.0317	0.0854	*0.0296*	0.0385
		Ks	0.0976	0.6167	0.0407	0.0575
	100	bias	0.0368	−0.2004	0.0031	0.0017
		sd	0.1632	0.1263	0.1726	0.1965
		mse	*0.0280*	0.0561	0.0298	0.0386
		risk	*0.0131*	0.0358	0.0149	0.0193
		Ks	0.0782	0.5678	0.0292	0.0395
	250	bias	0.0153	−0.1146	0.0019	0.0011
		sd	0.0970	0.0850	0.1110	0.1259
		mse	*0.0097*	0.0204	0.0123	0.0159
		risk	*0.0047*	0.0117	0.0061	0.0079
		Ks	0.0505	0.5042	0.0275	0.0306
0.50	50	bias	0.0597	−0.2179	0.0013	−0.0008
		sd	0.2386	0.1767	0.2443	0.2784
		mse	0.0605	0.0787	*0.0597*	0.0775
		risk	*0.0272*	0.0536	0.0297	0.0385
		Ks	0.0771	0.4644	0.0461	0.0575
	100	bias	0.0292	−0.1417	0.0025	0.0017
		sd	0.1568	0.1312	0.1726	0.1965
		mse	*0.0254*	0.0373	0.0298	0.0386
		risk	*0.0121*	0.0229	0.0149	0.0193
		Ks	0.0582	0.4166	0.0304	0.0395
	250	bias	0.0130	−0.0787	0.0018	0.0011
		sd	0.0950	0.0869	0.1110	0.1259
		mse	*0.0092*	0.0137	0.0123	0.0159
		risk	*0.0045*	0.0077	0.0061	0.0079
		Ks	0.0491	0.3632	0.0380	0.0306
1.00	50	bias	0.0529	−0.1634	0.0007	−0.0008
		sd	0.2304	0.1836	0.2443	0.2784
		mse	*0.0559*	0.0604	0.0597	0.0775
		risk	*0.0254*	0.0393	0.0298	0.0385
		Ks	0.0694	0.3596	0.0476	0.0575
	100	bias	0.0266	−0.1033	0.0023	0.0017
		sd	0.1535	0.1345	0.1726	0.1965
		mse	*0.0243*	0.0288	0.0298	0.0386
		risk	*0.0116*	0.0171	0.0149	0.0193
		Ks	0.0650	0.3057	0.0305	0.0395
	250	bias	0.0129	−0.0556	0.0017	0.0011
		sd	0.0941	0.0881	0.1110	0.1259
		mse	*0.0090*	0.0109	0.0123	0.0159
		risk	*0.0044*	0.0060	0.0061	0.0079
		Ks	0.0502	0.2651	0.0281	0.0306

Relative bias, standard deviation, mean squared error and risk for several estimators. Lowest values of mean squared error and risk are italicized.

TABLE 2
Comparison of estimators for $\mu_2(x) = e^{16x-8}/(1+e^{16x-8})$

σ	n	Measure	BCMLE	MLE	(14)	(15)
0.25	50	bias	0.0987	−0.2703	0.0092	−0.0008
		sd	0.2583	0.1713	0.2443	0.2784
		mse	0.0765	0.1024	*0.0598*	0.0775
		risk	0.0323	0.0727	*0.0293*	0.0385
		Ks	0.1253	0.5651	0.0319	0.0573
	100	bias	0.0481	−0.1772	0.0044	0.0017
		sd	0.1628	0.1287	0.1726	0.1965
		mse	*0.0288*	0.0480	0.0298	0.0386
		risk	*0.0133*	0.0302	0.0149	0.0193
		Ks	0.1064	0.5145	0.0274	0.0395
	250	bias	0.0198	−0.0998	0.0021	0.0011
		sd	0.0969	0.0861	0.1110	0.1259
		mse	*0.0098*	0.0174	0.0123	0.0159
		risk	*0.0047*	0.0099	0.0061	0.0079
		Ks	0.0744	0.4463	0.0270	0.0306
0.50	50	bias	0.0763	−0.2090	0.0026	−0.0008
		sd	0.2418	0.1788	0.2443	0.2784
		mse	0.0643	0.0756	*0.0597*	0.0775
		risk	*0.0281*	0.0512	0.0296	0.0385
		Ks	0.1039	0.4450	0.0430	0.0574
	100	bias	0.0387	−0.1329	0.0028	0.0017
		sd	0.1571	0.1322	0.1726	0.1965
		mse	*0.0262*	0.0351	0.0298	0.0386
		risk	*0.0123*	0.0214	0.0149	0.0193
		Ks	0.0913	0.3924	0.0298	0.0395
	250	bias	0.0171	−0.0721	0.0018	0.0011
		sd	0.0955	0.0875	0.1110	0.1259
		mse	*0.0094*	0.0128	0.0123	0.0159
		risk	*0.0045*	0.0072	0.0061	0.0079
		Ks	0.0652	0.3346	0.0279	0.0306
1.00	50	bias	0.0631	−0.1649	0.0010	−0.0008
		sd	0.2324	0.1840	0.2443	0.2784
		mse	*0.0580*	0.0611	0.0597	0.0775
		risk	*0.0259*	0.0398	0.0297	0.0385
		Ks	0.0828	0.3602	0.0467	0.0574
	100	bias	0.0331	−0.1022	0.0024	0.0017
		sd	0.1546	0.1347	0.1726	0.1965
		mse	*0.0250*	0.0286	0.0298	0.0386
		risk	*0.0118*	0.0170	0.0149	0.0193
		Ks	0.0746	0.3039	0.0305	0.0395
	250	bias	0.0159	−0.0535	0.0018	0.0011
		sd	0.0944	0.0884	0.1110	0.1259
		mse	*0.0092*	0.0107	0.0123	0.0159
		risk	*0.0044*	0.0058	0.0061	0.0079
		Ks	0.0641	0.2566	0.0282	0.0306

Relative bias, standard deviation, mean squared error and risk for several estimators. Lowest values of mean squared error and risk are italicized.

TABLE 3
Comparison of estimators for $\mu_3(x) = (2x-1)^3 + 1$

σ	n	Measure	BCMLE	MLE	(14)	(15)
0.25	50	bias	0.1134	−0.3529	0.0222	−0.0007
		sd	0.2801	0.1602	0.2446	0.2783
		mse	0.0913	0.1502	*0.0603*	0.0775
		risk	0.0377	0.1132	*0.0288*	0.0385
		Ks	0.13399	0.7023	0.0202	0.0575
	100	bias	0.0483	−0.2464	0.0078	0.0017
		sd	0.1689	0.1226	0.1726	0.1965
		mse	0.0309	0.0758	*0.0298*	0.0386
		risk	*0.0141*	0.0498	0.0148	0.0193
		Ks	0.0955	0.6718	0.0231	0.0395
	250	bias	0.0202	−0.1460	0.0026	0.0011
		sd	0.1002	0.0835	0.1110	0.1259
		mse	*0.0105*	0.0283	0.0123	0.0159
		risk	*0.0050*	0.0166	0.0061	0.0079
		Ks	0.0689	0.6256	0.0244	0.0306
0.50	50	bias	0.0688	−0.2580	0.0060	−0.0008
		sd	0.2477	0.1720	0.2444	0.2783
		mse	0.0661	0.0961	*0.0598*	0.0775
		risk	*0.0292*	0.0674	0.0295	0.0385
		Ks	0.0896	0.5376	0.0372	0.0573
	100	bias	0.0311	−0.1742	0.0037	0.0017
		sd	0.1603	0.1288	0.1726	0.1965
		mse	*0.0267*	0.0470	0.0298	0.0386
		risk	*0.0126*	0.0295	0.0149	0.0193
		Ks	0.0624	0.4983	0.0284	0.0395
	250	bias	0.0146	−0.1000	0.0020	0.0011
		sd	0.0965	0.0857	0.1110	0.1259
		mse	*0.0095*	0.0174	0.0123	0.0159
		risk	*0.0046*	0.0099	0.0061	0.0079
		Ks	0.0520	0.4468	0.0268	0.0306
1.00	50	bias	0.0532	−0.1900	0.0019	−0.0008
		sd	0.2335	0.1804	0.2443	0.2784
		mse	0.0574	0.0686	0.0597	0.0775
		risk	0.0261	0.0456	0.0297	0.0385
		Ks	0.0693	0.4597	0.0439	0.0574
	100	bias	0.0274	−0.1243	0.0026	0.0017
		sd	0.1550	0.1329	0.1726	0.1965
		mse	*0.0248*	0.0331	0.0298	0.0386
		risk	*0.0118*	0.0200	0.0149	0.0193
		Ks	0.0550	0.3672	0.0300	0.0395
	250	bias	0.0122	−0.0693	0.0018	0.0011
		sd	0.0944	0.0873	0.1110	0.1259
		mse	*0.0091*	0.0124	0.0123	0.0159
		risk	*0.0044*	0.0069	0.0061	0.0079
		Ks	0.0433	0.3246	0.0276	0.0306

Relative bias, standard deviation, mean squared error and risk for several estimators. Lowest values of mean squared error and risk are italicized.

the event that $\hat{\theta}_{k-1} < \hat{\theta}_k$, $k = 2, \ldots, n$. Then D is one plus the sum of the indicators of A_k^o, $k = 2, \ldots, n$. An upper bound is sought for $E_\mu(D) = 1 + P_\mu(A_2^o) + \cdots + P_\mu(A_n^o)$. The bound is obtained by bounding $P_\mu(A_k^o)$ for each k. If $0 \le j_k < k$ and $k < l_k \le n$ are integers, to be specified later, then

$$A_k^o = \left\{ \max_{j<k} \bar{y}_{j,k} < \min_{l>k} \bar{y}_{k,l} \right\} \subseteq \left\{ \max_{j_k \le j<k} \bar{y}_{j,k} < \min_{k<l\le l_k} \bar{y}_{k,l} \right\} = A_k,$$

say, where $\bar{y}_{j,l} = (y_{j+1} + \cdots + y_l)/(l-j)$. Observe that $\bar{\theta}_{j,l} = [\theta_{j+1} + \cdots + \theta_l]/(l-j)$ is nondecreasing in $l = k+1, \ldots, n$ and nonincreasing in $j = 1, \ldots, k$. So $\bar{\theta}_{k,l} \le \bar{\theta}_{k,l_k}$ for all $l = k+1, \ldots, l_k$ and $\bar{\theta}_{j,k} \ge \bar{\theta}_{j_k,k}$ for $j = j_k, \ldots, k-1$ and $k = 2, \ldots, n$. So, since $\bar{y}_{j,l} = \bar{\varepsilon}_{j,l} + \bar{\theta}_{j,l}$ for $0 \le j \le k \le l \le n$,

$$A_k \subseteq \left\{ \max_{j_k \le j<k} \bar{\varepsilon}_{j,k} - \min_{k<l\le l_k} \bar{\varepsilon}_{k,l} < \bar{\theta}_{k,l_k} - \bar{\theta}_{j_k,k} \right\}. \tag{25}$$

This leads to the following side problem: given i.i.d standard normal random variables $X_1, X_2, \ldots, Y_1, Y_2, \ldots$ and positive integers m and n, find bounds for

$$H_{m,n}(z) := P\left\{ \max_{j\le m} \overline{X}_j + \max_{k\le n} \overline{Y}_k \le z \right\} \tag{26}$$

for real z. Let P_z denote a probability distribution under which $X_1, X_2, \ldots, Y_1, Y_2 \ldots$ are i.i.d. normally distributed random variables with common mean z and unit variance, so that $P = P_0$. Further, let $S_k = X_1 + \cdots + X_k$, $k = 1, 2, \ldots$, and let τ denote the first ladder epoch $\tau = \inf\{k \ge 1\colon S_k > 0\}$. Some properties of τ and S_τ are needed. From Feller [(1971), Chapter 12], it is known that

$$P_0\{\tau > n\} = \binom{2n}{n} 4^{-n} \le \frac{1}{\sqrt{n}}$$

for all n, and $E_0(S_\tau) = 1/\sqrt{2}$ and $E_z(S_\tau) = z \exp[\sum_{k=1}^\infty k^{-1} \Phi(-z\sqrt{k})]$ from Feller [(1971), Chapter 18]. Moreover, from Klass (1983) or direct analysis, $\lim_{z\to 0} E_z(S_\tau) = E_0(S_\tau)$. So there is a $c > 0$ for which $E_z(S_\tau) \le c$ for $0 \le z \le 1$. From Woodroofe [(1982), page 33], this holds with $c = 1.5$.

PROPOSITION 5. *For all* $z > 0$,

$$H_{m,n}(z) \le 18z^2 + \frac{9}{m} + \frac{9}{n}.$$

PROOF. It suffices to prove the proposition for $0 < z \le 1$. First, observe that $P_0\{\max_{j\le m} \overline{X}_j \le -z\} = P_z\{\max_{j\le m} \overline{X}_j \le 0\} = P_z\{\tau > m\}$. So, since $\tau > m$ implies $S_m \le 0$,

$$\begin{aligned} P_0\left\{ \max_{j\le m} \overline{X}_j \le -z \right\} &= \int_{\{\tau>m\}} e^{zS_m - (1/2)mz^2} dP_0 \\ &\le e^{-(1/2)mz^2} P_0\{\tau > m\} \le \frac{1}{\sqrt{m}} e^{-(1/2)mz^2}. \end{aligned} \tag{27}$$

Continuing, $P_0\{0 < \max_{j\le m} \overline{X}_j \le z\} = P_{-z}\{\max_{j\le m} \overline{X}_j \le 0\} - P_0\{\max_{j\le m} \overline{X}_j \le 0\} = P_{-z}\{\tau > m\} - P_0\{\tau > m\} = P_0\{\tau \le m\} - P_{-z}\{\tau \le m\}$. So since $P_0\{\tau \le m\} \le P_z\{\tau \le m\}$,

$$P_0\left\{0 < \max_{j\le m} \overline{X}_j \le z\right\} \le P_z\{\tau \le m\} - P_{-z}\{\tau \le m\} \le \int_{\{\tau\le m\}} [1 - e^{-2zS_\tau}]dP_z \le 2zE_z(S_\tau) \le 3z. \tag{28}$$

In the remainder of the proof, write $X = \max_{j\le m} \overline{X}_j$ and $Y = \max_{k\le n} \overline{Y}_k$, and let F and G denote the distribution functions of X and Y, respectively. Then, since $X + Y \le z$ implies $\min(X, Y) \le z$,

$$P\{X + Y \le z\} \le P\{X \le z, Y \le z - X\} + P\{Y \le z, X \le z - Y\}.$$

Here

$$\begin{aligned} P\{X \le z, Y \le z - X\} &\le \frac{1}{\sqrt{m}} e^{-(1/2)m} + P\{-1 < X \le z, Y \le z - X\} \\ &= \frac{1}{\sqrt{m}} e^{-(1/2)m} + \int_{-1}^{z} G(z - x)\, dF(x) \\ &\le \frac{1}{\sqrt{m}} e^{-(1/2)m} + \int_{-1}^{z} \left[\frac{1}{\sqrt{n}} + 3(z - x)\right] dF(x) \\ &\le \frac{1}{\sqrt{m}} e^{-(1/2)m} + \frac{1}{\sqrt{n}} F(z) + 3\int_{-1}^{z} F(y)dy, \end{aligned}$$

by (27) and (28) for $0 \le z \le 1$ and, since $e^{-(1/2)m} \le 1/m$, the last line is at most

$$\frac{1}{\sqrt{m}} e^{-(1/2)m} + \frac{1}{\sqrt{n}}\left[\frac{1}{\sqrt{m}} + 3z\right] + 3\int_{-1}^{0} \frac{1}{\sqrt{m}} e^{-(1/2)my^2}\, \mathrm{dy} + 3\int_{0}^{z} \left[\frac{1}{\sqrt{m}} + 3y\right] dy$$

$$\le \frac{1}{\sqrt{mn}} + \frac{3z}{\sqrt{n}} + \frac{5}{m} + \frac{3z}{\sqrt{m}} + \frac{9}{2} z^2.$$

A similar bound may be obtained for $P\{Y \le z, X \le z - Y\}$, and the proposition then follows by collecting terms and using the inequality $2ab \le a^2 + b^2$. □

PROOF OF (16). Recall that $\theta = [\mu(x_1), \ldots, \mu(x_n)]'$ and let $K = \{k\colon 2 \le k \le n - 1\colon \theta_{k+1} - \theta_{k-1} \le 1\}$. Then, clearly, $\#K^c \le 2\Delta$, where $\#K^c$ denotes the number of elements that are not in K. So,

$$E_\mu(D) \le 2 + 2\Delta + \sum_{k\in K} P(A_k).$$

Let R be the least integer that exceeds $\Delta^{-2/3} n^{2/3}$. For $k \in K$, let l_k be the largest l for which $k < l \le n, l - k \le R$ and $\bar{\theta}_{k,l} - \theta_k \le 1/\sqrt{l - k}$, and let j_k be the smallest j for which $0 \le j < k, k - j \le R$ and $\theta_k - \bar{\theta}_{j,k} \le 1/\sqrt{k - j}$.

Further, let $m_k = k - j_k$ and $n_k = l_k - k$. Then j_k and l_k are well defined by definition of K, and $\bar{\theta}_{k,l_k} - \bar{\theta}_{j_k,k} \le 1/\sqrt{m_k} + 1/\sqrt{n_k}$ for all $k = 2, \ldots, n-1$. By (25) and Proposition 5,

$$\sum_{k\in K} P_\mu(A_k) \le \sum_{k\in K}\left[18(\bar{\theta}_{k,l_k} - \bar{\theta}_{j_k,k})^2 + 9\left(\frac{1}{m_k} + \frac{1}{n_k}\right)\right]$$
$$\le 45\sum_{k\in K}\left(\frac{1}{m_k} + \frac{1}{n_k}\right).$$

Let $J_r = \{k \in K\colon n_k < r\}$ for $r \ge 1$. Then

$$\sum_{k\in K}\frac{1}{n_k} = \sum_{r=1}^{n-1}\frac{1}{r}\bigl[\#J_{r+1} - \#J_r\bigr] \le 1 + \sum_{r=2}^{n}\frac{1}{r(r-1)}\#J_r,$$

If $k \in J_r$, where $r \le R$, then either $k > n-r$ or $k \le n-r$ and $\theta_{k+r} - \theta_k > 1/\sqrt{r}$. So

$$\#J_r \le r + \sqrt{r}\sum_{k\in J_r,\, k\le n-r}[\theta_{k+r} - \theta_k]$$
$$= r + \sqrt{r}\sum_{k\in J_r,\, k\le n-r}\sum_{j=1}^{r}(\theta_{k+j} - \theta_{k+j-1}) \le r + \Delta r^{3/2},$$

for $r \le R$, where the last inequality follows by reversing the order of summation. Since $\#J_r \le n$ for $r > R$,

$$\sum_{r=2}^{n}\frac{1}{r(r-1)}\#J_r \le \sum_{r=2}^{n}\frac{1}{r-1} + \sum_{r=2}^{R}\frac{1}{r(r-1)}\Delta r^{3/2} + \sum_{r=R+1}^{n}\frac{1}{r(r-1)}n$$
$$\le 1 + \log(n) + 2\Delta\sum_{r=1}^{R-1}\frac{1}{\sqrt{r}} + \frac{n}{R}$$
$$\le 1 + \log(n) + 5\Delta^{2/3}n^{1/3}.$$

Of course, a similar bound may be obtained for $\sum_{k=2}^{n-1} 1/m_k$, and (16) then follows by collecting terms.

PROOF OF (17). For the proof of (17), let R be the least integer that exceeds $(1+\Delta)^{-2/3}n^{2/3}$. Then $D \le 1 + 1_{A_1} + \cdots + 1_{A_n}$, and since A_j and A_k are independent when $|j-k| > 2R$,

$$E_\mu(D^2) \le 2 + 2\left[\sum_{|j-k|\le 2R} + \sum_{|j-k|>2R}\right]P_\mu(A_j \cap A_k)$$
$$\le 2R\sum_{j=2}^{n}P_\mu(A_j) + 2\left[\sum_{j=2}^{n}P_\mu(A_j)\right]^2.$$

The summation $\sum_{j=2}^{n} P_\mu(A_j)$ may be bounded as in the proof of (16), and (17) results. □

6. Proof of Theorem 2. As in the last section, it suffices to prove Theorem 2 when $\sigma = 1$. Recall that $\Gamma_n = \sup_{0\le x\le 1}|F_n(x) - F(x)|$ and suppose that $\Gamma_n = o(n^{-1/3})$, as in (19).

LEMMA 1. *If* $0 \le a < b \le 1$, *then*

$$\left|\int_a^b [\mu(x) - \mu(a)]\,dF_n(x) - \int_a^b [\mu(x) - \mu(a)]\,dF(x)\right| \le 2\Gamma_n[\mu(b) - \mu(a)].$$

Further, if $\nu'(x) \le C$ *for* $0 \le x \le 1$, *then*

$$\int_{x_m}^{x_{m+k}} [\mu(x) - \mu(x_m)]\,dF_n(x) \le 2C\frac{k^2}{n^2} + 9C\Gamma_n^2 \tag{29}$$

for all m *and* k, *and if* $\nu'(x) \ge 6\delta > 0$, *then*

$$\int_{x_m}^{x_{m+k}} [\mu(x) - \mu(x_m)]\,dF_n(x) \ge \delta\frac{k^2}{n^2} \tag{30}$$

for $k \ge n^{2/3}$ *and sufficiently large* n.

PROOF. The first assertion follows from a transparent integration by parts. For the second, let $I = \int_{x_m}^{x_{m+k}} [\mu(x) - \mu(x_m)]\,dF_n(x)$. Then

$$\begin{aligned} I &\le \int_{F(x_m)}^{F(x_{m+k})} \{\nu(x) - \nu[F(x_m)]\}\,dx + 2\Gamma_n\{\nu[F(x_{m+k})] - \nu[F(x_m)]\} \\ &\le \tfrac{1}{2}C[F(x_{m+k}) - F(x_m)]^2 + 2C\Gamma_n[F(x_{m+k}) - F(x_m)]. \end{aligned}$$

Now, $|F(x_k) - k/n| = |F(x_k) - F_n(x_k)| \le \Gamma_n$ for all k. So,

$$\begin{aligned} I &\le \frac{1}{2}C\left[\frac{k}{n} + 2\Gamma_n\right]^2 + 2C\Gamma_n\left[\frac{k}{n} + 2\Gamma_n\right] \\ &\le C\left\{\left[\left(\frac{k}{n}\right)^2 + 4\Gamma_n^2\right] + 2\Gamma_n\frac{k}{n} + 4\Gamma_n^2\right\} \\ &\le 2C\frac{k^2}{n^2} + 9C\Gamma_n^2, \end{aligned}$$

establishing (29). Relation (30) may be established similarly. □

In the proof of (21) use is made of the following inequalities: if $Z_1, Z_2, \ldots,$ are i.i.d. standard normal random variables, then

$$P\left(\max_{k\le n} Z_1 + \cdots + Z_k > z\right) \le 2\left[1 - \Phi\left(\frac{z}{\sqrt{n}}\right)\right],$$

$$P\{Z_1 + \cdots + Z_k \ge a + bk, \text{ for some } k \ge 1\} \le e^{-2ab},$$

which may be proved by comparison with Brownian motion. See, for example, Breiman [(1968), pages 258, 289].

PROOF OF (21). It follows from Wright (1981) that if $m = m_n \to \infty$ and $m/n \to x_0 \in (0,1)$, then

$$n^{1/3}\big(\hat{\theta}_m^n - \theta_m^n\big) \Rightarrow \nu'(x_0)^{1/3}\widetilde{W}'(0),$$

where $\Rightarrow$ denotes convergence in distribution. Uniform integrability is established below, along with bounds for small and large m. It follows that $E_{\sigma,\mu}[(\hat{\theta}_m^n - \theta_m^n)^2] \sim b\nu'(x_0)^{2/3}n^{-2/3}$ as $n \to \infty$, if $m/n \to x_0 \in (0,1)$. Relation (21) then follows from $E_{\sigma,\mu}\|\hat{\theta}^n - \theta^n\|^2 = \sum_{m=1}^n E_{\sigma,\mu}[(\hat{\theta}_m^n - \theta_m^n)^2]$. Let $l_m^n = \min(m, n-m, \lfloor n^{2/3}\rfloor)$, where $\lfloor x\rfloor$ denotes the greatest integer that is less than or equal to x. It is first shown that there is a sequence $\gamma_n, n \geq 1$, for which $\gamma_n = o(n^{1/3})$ and

$$\left[\sqrt{l_m^n}\,|\hat{\theta}_m^n - \theta_m^n| - \frac{\gamma_n}{\sqrt{l_m^n}}\right]_+, \tag{31}$$

$$1 \leq m \leq n-1, \text{ are uniformly square integrable,}$$

where $a_+ = \max(0,a)$. Relation (21) is a consequence of (31). For if $\varepsilon > 0$, then (31) implies that $E_\mu\big[\big(\hat{\theta}_m^n - \theta_m^n\big)^2\big] \sim b\nu'(m/n)^{2/3}n^{-2/3}$ uniformly in $\varepsilon n \leq m \leq (1-\varepsilon)n$ as $n \to \infty$ and therefore that

$$\sum_{\varepsilon n \leq m \leq (1-\varepsilon)n} E_\mu\Big[\Big(\hat{\theta}_m^n - \theta_m^n\Big)^2\Big] \sim bn^{1/3}\int_\varepsilon^{1-\varepsilon} \nu'(x)^{2/3}dx$$

as $n \to \infty$. Moreover, (31) implies that $l_m^n E\mu\big[\big(\hat{\theta}_m^n - \theta_m^n\big)^2\big] - \gamma_n^2/l_m^n$ is bounded in $m = 1, \ldots, n-1$ and $n \geq 2$. Since $E_\mu\big[\big(\hat{\theta}_m^n - \theta_m^n\big)^2\big]$ is also (easily seen to be) bounded, there is a $0 < K < \infty$ for which $E_\mu\big[\big(\hat{\theta}_m^n - \theta_m^n\big)^2\big] \leq K\min\big[1, 1/l_m^n + \big(\gamma_n/l_m^n\big)^2\big]$. So, for any $0 < \delta < \varepsilon$ and sufficiently large n,

$$\begin{aligned}
&n^{-\frac{1}{3}}\sum_{m \leq \varepsilon n} E_\mu\Big[\Big(\hat{\theta}_m^n - \theta_m^n\Big)^2\Big] \\
&\quad = n^{-\frac{1}{3}}\Bigg(\sum_{m \leq \delta n^{\frac{1}{3}}} + \sum_{\delta n^{\frac{1}{3}} < m \leq n^{\frac{2}{3}}} + \sum_{n^{\frac{2}{3}} < m \leq \varepsilon n}\Bigg)E_\mu\Big[\Big(\hat{\theta}_m^n - \theta_m^n\Big)^2\Big] \\
&\quad \leq K\delta + Kn^{-1/3}\sum_{\delta n^{1/3} < m \leq n^{2/3}}\left[\frac{1}{m} + \frac{\gamma_n^2}{m^2}\right] + 2K\varepsilon \\
&\quad \leq K\delta + 2Kn^{-1/3}\log(n) + \frac{2K}{\delta}n^{-2/3}\gamma_n^2 + 2K\varepsilon,
\end{aligned}$$

which approaches zero as $n \to \infty$, $\delta \to 0$ and $\varepsilon \to 0$ in that order. Of course, the right endpoint may be handled similarly, and relation (21) follows from (31).

To prove (31) it will be shown that

$$\left[\sqrt{l_m^n}\left[\hat{\theta}_m^n - \theta_m^n\right]_+ - \frac{\gamma_n}{\sqrt{l_m^n}}\right]_+, \tag{32}$$

$1 \le m \le n-1$ are uniformly square integrable.

The assertion (31) follows from this and a dual argument for the negative parts. To begin, fix a $1 \le m \le n-1$ and recall that $\hat{\theta}_m^n = \max_{j<m} \min_{k\ge m} \bar{y}_{j,k}^n$, where $\bar{y}_{j,k}^n = \left(y_{j+1}^n + \cdots + y_k^n\right)/(k-j)$ for $0 \le j < k \le n$. Thus,

$$\left\{\hat{\theta}_m^n - \theta_m^n > z\right\} = \bigcup_{j=0}^{m-1} \bigcap_{k=m}^{n} \left\{\bar{y}_{j,k}^n - \theta_m^n > z\right\}$$

for $0 < z < \infty$. Next, fix n for the moment and let $S_k = y_1^n + \cdots + y_k^n$, $1 \le k \le n$. $S'_j = S_m - S_{m-j} - \theta_m^n j$ for $j \le m$ and $S''_k = S_{m+k} - S_m - \theta_m^n k$ for $k \le n-m$. Then $\left\{\hat{\theta}_m^n - \theta_m^n > z\right\} = \bigcup_{j=1}^{m} \bigcap_{k=0}^{n-m} \left\{S'_j + S''_k > z(j+k)\right\}$ and, therefore,

$$\begin{aligned} P_\mu\left\{\hat{\theta}_m^n - \theta_m^n > z\right\} &\le P_\mu\left\{S'_j + S''_l > (j+l)z, \text{ for some } j \le m\right\} \\ &\le P_\mu\left\{S''_l > \tfrac{1}{2}zl\right\} + P_\mu\left\{S'_j > \tfrac{1}{2}z(j+l), \text{ for some } j \le m\right\} \end{aligned}$$

for every $l = 0, \ldots, n-m$. Let $M'_j = E_\mu(S'_j)$ and $M''_k = E_\mu(S''_k)$ denote the means and let C denote an upper bound for ν'. Then M'_j are nonpositive so that

$$\begin{aligned} P_\mu\left\{S'_j > \tfrac{1}{2}z(j+l), \text{ for some } j \le m\right\} &\le P_0\left\{S_j > \tfrac{1}{2}z(j+l), \text{ for some } j \le m\right\} \\ &\le \exp\left[-\tfrac{1}{2}lz^2\right], \end{aligned}$$

by the second boundary crossing probability. Let C denote an upper bound for ν'. Then $M''_k \le 2Ck^2/n + 9Cn\Gamma_n^2$ for all $k = 0, \ldots, n-m$, by Lemma 3. Let $\gamma_n = 9C\Gamma_n^2$, so that $\gamma_n = o(n^{1/3})$ as $n \to \infty$, by (19). Then

$$\begin{aligned} P\left(S''_l > \frac{1}{2}zl\right) &= (1-\Phi)\left(\frac{\frac{1}{2}zl - M_l}{\sqrt{l}}\right) \\ &\le (1-\Phi)\left(\frac{1}{2}z\sqrt{l} - 2C\frac{l^{3/2}}{n} - 9Cn\frac{\Gamma_n^2}{\sqrt{l}}\right). \end{aligned}$$

Now let $l = \min(\lfloor n^{2/3} \rfloor, n-m] \ge l_m^n$. Then

$$P_\mu\left(\sqrt{l}\left(\hat{\theta}_m^n - \theta_m^n\right) > z + \frac{\gamma_n}{\sqrt{l}}\right) \le (1-\Phi)\left(\tfrac{1}{2}z - 2C\right) + e^{-(1/2)z^2}$$

for all $z \ge 0$. Relation (32) follows, completing the proof of (21). □

PROOF OF (20). For (20), let $S_k = y_1 + \cdots + y_k$, $k = 1, \ldots, n$, and let $\widetilde{S}_k$ denote the greatest convex minorant of S_k. That is, $\widetilde{S}_k = \widetilde{S}(k)$, where $\widetilde{S}$ is the

greatest convex minorant of the continuous piecewise linear function S with knots at $0, \ldots, n$ and values $S(k) = S_k$. Then

$$\langle y^n - \hat{\theta}^n, \theta^n \rangle = \sum_{m=1}^{n-1} \big(\widetilde{S}_m - S_m\big)\big[\mu(x_{n,m+1}) - \mu(x_{n,m})\big].$$

It is implicit in the proof of (21) that $n^{1/3}\big(\widetilde{S}_m - S_m\big) \Rightarrow \nu'(x_0)^{-1/3}\widetilde{W}(0)$, if $m/n \to x_0 \in (0,1)$. Relation (20) then follows by showing that

$$n^{-1/3}\big(\widetilde{S}_m - S_m\big), 1 \le m \le n, \text{ are uniformly integrable.} \tag{33}$$

and summing over m, as above. For any fixed m, $\widetilde{S}_k \ge (k-m)\mu(x_m) + \min_{0\le j\le n}[S_j - (j-m)\mu(x_m)]$, so that $0 \ge \big(\widetilde{S}_m - S_m\big) \ge \inf_{-m\le k\le n-m} S_{m+k} - S_m - k\mu(x_m)$. Let $S'_k = S_{m+k} - S_m - k\mu(x_m)$, and $M'_k = E\mu(S'_k)$, and $l = \min(n-m, \lfloor n^{2/3} \rfloor)$. Then $M'_k \ge 0$ for all $k = 1, \ldots, n$, so that

$$P_\mu\left\{n^{-1/3}\min_{0\le k\le l} S'_k \le -z\right\} \le P_0\left\{\min_{0\le k\le l} S_k \le -zn^{1/3}\right\} \le 2\Phi(-z). \tag{34}$$

Let $6\delta = \inf_{0\le x\le 1}\nu'(x) > 0$. If $l < n-m$, then $M'_l \ge \delta l^2/n^2$ for sufficiently large n and therefore,

$$\begin{aligned} &P_\mu\Big(n^{-1/3}\min_{k>l} S'_k \le -z\Big) \\ &\quad\le P_\mu\Big\{S'_k - M'_k \le -\Big[n^{1/3}z + \frac{M'_l}{l}k\Big], \text{for some } k \ge 1\Big\} \\ &\quad\le \exp\Big[-2\frac{M'_l}{l}n^{1/3}z\Big] \end{aligned}$$

for all $z \ge 1$. Combining (34) and (35) with dual arguments for $k \le 0$, it follows that

$$\max_{1\le m\le n} P_\mu\Big(\min_{-m\le k\le n-m} n^{-1/2}S_{m+k} - S_m - \mu(x_m)k \le -z\Big) \le 4\Phi(-z) + 2e^{-z}$$

for all $z \ge 0$ for all sufficiently large n, establishing (33). It follows easily that

$$\begin{aligned} &E_\mu\Big\{\sum_{k=1}^{n-1}\big(\widetilde{S}_k - S_k\big)\Big\}\big(\theta^n_k - \theta^n_{k-1}\big) \\ &\quad= -an^{1/3}\sum_{k=1}^{n-1}\Big[\nu'\Big(\frac{k}{n}\Big)\Big]^{-1/3}\big[\mu(x_{n,k-1}) - \mu\big(x_{n,k}\big)\big] + o(n^{1/3}), \end{aligned}$$

and the sum on the right is $\int_0^1 \nu'(x)^{2/3}dx + o(1)$, by (19). □

Acknowledgments. Thanks to two referees, an Associate Editor and an Editor for their comments.

REFERENCES

BREIMAN, L. (1968). *Probability*. Addison-Weseley, Reading, MA.

DONOHO, D. (1990). Gelfand n-widths and the method of least squares. Unpublished manuscript.

DONOHO, D. and JOHNSTONE, I. (1995). Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.* **90** 1200–1225.

EFROMOVICH, S. (1997). On quasi-linear wavelet estimation. *J. Amer. Statist. Assoc.* **94** 189–204.

FELLER, W. (1971). *An Introduction to Probability Theory and Its Applications* **2**. Wiley, New York.

GASSER, T. SROKA, L. and JENNEN-STEINMETZ, C. (1986). Residual variance and residual pattern in non-linear regression. *Biometrika* **73** 625–633.

GROENEBOOM, P. (1985). Estimating a monotone density. In *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer* (L. Le Cam and R. Olshen, eds.) **2** 539–558. Wadsworth and Brooks/Cole, Belmont, CA.

GROENEBOOM, P. (1989). Brownian motion with a parabolic drift and airy functions. *Probab. Theory Related Fields* **81** 79–109.

HASTIE, T. and TIBSHIRANI, R. (1990). *Generalized Additive Models*. Chapman and Hall, London.

KLASS, M. (1983). On the maximum of a random walk with a small negative drift. *Ann. Probab.* **11** 491–505.

MEYER, M. (1996). Shape restricted inference with applications to nonparametric regression, smooth nonparametric regression, and density estimation. Ph.D. thesis, Dept. statistics, Univ. Michigan.

MEYER, M. and WOODROOFE, M. (1998). Variance estimation in monotone regression. Technical Report 322, Dept. Statistics, Univ. Michigan.

RICE, J. (1984). Bandwidth choice for nonparametric regression. *Ann. Statist*. **12** 1215–1230.

ROBERTSON, T., WRIGHT, F. and DYKSTRA, R. (1988). *Order Restricted Inference*, Wiley, New York.

STEIN, C. (1981). Estimation of the mean of a multivariate normal distribution. *Ann. Statist.* **9** 1135–1151.

VAN DER GEER, S. (1990). Estimating a regression function. *Ann. Statist.* **18** 907–924.

WOODROOFE, M. (1982). *Non-linear Renewal Theory in Sequential Analysis*. SIAM, Philadelphia.

WRIGHT, F. T. (1981). The asymptotic behavior of monotone regression estimates. *Ann. Statist*. **9** 449–453.

STATISTICS DEPARTMENT
UNIVERSITY OF GEORGIA
ATHENS, GEORGIA
E-MAIL: mmeyer@stat.uga.edu

STATISTICS DEPARTMENT
UNIVERSITY OF MICHIGAN
ANN ARBOR, MICHIGAN 48109
E-MAIL: michaelw@umich.edu