The Annals of Statistics
1994, Vol. 22, No. 1, 271–289

ON MINIMAX ESTIMATION OF A SPARSE NORMAL MEAN VECTOR[1]

BY IAIN M. JOHNSTONE

Stanford University

Mallows has conjectured that among distributions which are Gaussian but for occasional contamination by additive noise, the one having least Fisher information has (two-sided) geometric contamination. A very similar problem arises in estimation of a nonnegative vector parameter in Gaussian white noise when it is known also that most [i.e., $(1-\varepsilon)$] components are zero.

We provide a partial asymptotic expansion of the minimax risk as $\varepsilon \to 0$. While the conjecture seems unlikely to be exactly true for finite ε, we verify it asymptotically up to the accuracy of the expansion. Numerical work suggests the expansion is accurate for ε as large as 0.05. The best l_1-estimation rule is first- but not second-order minimax. The results bear on an earlier study of maximum entropy estimation and various questions in robustness and function estimation using wavelet bases.

1. Introduction. In many estimation settings, there is definite prior information concerning the values of a parameter vector θ. One may have bounds on the individual components θ_i ("all θ_i lie between 0 and 1"), or on particular functionals of the whole vector ("the squared length of θ is at most c" or "most of the θ_i are zero"). Many estimation methods have been developed to capitalize on such prior information, either explicitly in the form of constraints on an optimization procedure (e.g., positivity-constrained least squares) or implicitly in the sense that the estimator performs well "if and only if" θ satisfies the prior constraints. An example of the latter are maximum entropy regularization estimates in the case of "nearly black" θ [e.g., Frieden (1972) and Gull and Daniel (1978)].

How does one compare the performance of various possible estimators when such prior information is present? One common, admittedly conservative, approach is the worst-case analysis: given some error measure, compute the maximum expected error over the restricted parameter space, and then seek the estimator that minimizes this maximum risk. The resulting best or minimax risk provides (i) a benchmark against which to measure other estimators and (ii) a measure of the value of the prior information (by comparison with the minimax risk computed ignoring the prior information).

Received July 1991; revised May 1993.

[1]This work, begun in 1986, has been supported in part by NSF Grants DMS-84-51750 and DMS-92-09130, by NIH PHS GM 21215-12 and by the Sloan Foundation.

AMS 1991 *subject classifications.* Primary 62C20; secondary 62C10, 62G05.

Key words and phrases. Fisher information, minimax decision theory, least favorable prior, nearly black object, robustness, white noise model.

Institute of Mathematical Statistics is collaborating with JSTOR to digitize, preserve, and extend access to *The Annals of Statistics.* www.jstor.org

This paper presents an asymptotic evaluation of this minimax risk in a simple, but hopefully representative, context: estimation of nonnegative signals that are mostly zero, such as spectra, star maps and the like. Our model is idealized in certain significant ways: A signal-plus-noise model is adopted, and the noise is assumed to be i.i.d. Gaussian. These assumptions permit a more detailed analysis, but at the price of a degree of nonrobustness to departures from the model.

To state the problem precisely, let $X \sim N(\theta, 1)$. Denote the mean squared error, or risk, of an estimator $d(x)$ by $R(\theta, d) = E_\theta(d(X) - \theta)^2$. If $G(d\theta)$ is a prior probability distribution, denote the integrated risk of d by

$$r(G, d) = \int R(\theta, d) G(d\theta).$$

Let δ_a denote point mass at a, and let $\mathcal{P}([0, \infty))$ denote the class of probability measures supported on $[0, \infty)$. Consider a class of priors on $[0, \infty)$ with an atom at 0:

$$\mathcal{G}_\varepsilon = \{G = (1 - \varepsilon)\delta_0 + \varepsilon H: H \in \mathcal{P}([0, \infty))\}.$$

This paper is concerned with asymptotic evaluation of the (restricted Bayes) minimax risk (see Section A1)

$$m(\varepsilon) = \sup_{\mathcal{G}_\varepsilon} \inf_d r(G, d) = \inf_d \sup_{\mathcal{G}_\varepsilon} r(G, d), \tag{1}$$

as $\varepsilon \to 0$. This problem or a close relative arises in a variety of contexts, some of which we now review.

Nearly black objects. As part of a study of maximum entropy estimation, Donoho, Johnstone, Hoch and Stern (1992) consider the problem of estimating a *nonnegative* vector $\boldsymbol{\theta} = (\theta_i)_{i=1}^n$ from noisy data,

$$X_i = \theta_i + Z_i, \qquad i = 1, \dots, n, \tag{2}$$

where the noise terms Z_i are i.i.d. $N(0, \sigma^2)$. They show that a maximum entropy rule, defined as a minimizer $\widehat{\boldsymbol{\theta}}_{\mathrm{ME}, \lambda}(x)$ of

$$\sum_i (\theta_i - x_i)^2 + 2\lambda \sum_i \theta_i \log \theta_i, \tag{3}$$

achieves significant savings over linear rules in mean squared error when and only when most components θ_i of the unknown object are nearly zero. It is then natural to define the class of "ε-black images" $\Theta_n(\varepsilon)$ as the set of nonnegative sequences (θ_i) of length n satisfying $\#\{i: \theta_i > 0\} \le n\varepsilon$. The worst-case error of an estimator $\widehat{\theta}_n(x)$ is

$$m_n(\widehat{\theta}_n, \varepsilon) = \sup_{\Theta_N(\varepsilon)} E_\theta n^{-1} \sum_1^n \left(\widehat{\theta}_{n,i}(X) - \theta_i\right)^2.$$

A benchmark against which particular estimators may be measured is the minimax risk

$$m_n(\varepsilon) = \inf_{\widehat{\theta}_n} m_n(\widehat{\theta}_n, \varepsilon). \tag{4}$$

As n becomes large, $m_n(\varepsilon)$ approaches a limit $m(\varepsilon)$, and the evaluation of $m(\varepsilon)$ is exactly equivalent to the problem (1) [Donoho, Johnstone, Hoch and Stern (1992), Theorem 1].

Robust estimation. Denote the Fisher information of an absolutely continuous distribution F with density f by $I(F) = \int f'^2/f$. In Huber's (1964, 1981) asymptotic minimax approach to robust estimation, there arises the problem of minimizing Fisher information over neighborhoods of the standard Gaussian distribution. One possibility is to consider Gaussian variables that are occasionally contaminated by additive noise, that is, distributions $F = \Phi * G$, where G belongs to $\mathcal{G}'_\varepsilon = \{G = (1-\varepsilon)\delta_{\{0\}} + \varepsilon H, H \in \mathcal{P}((-\infty,\infty))\}$. Here Φ is the standard Gaussian distribution and $*$ denotes convolution. Mallows (1978) noted that the distribution G_0 minimizing $I(\Phi * G)$ would be symmetric and discrete: $G_0 = \Sigma p_j \delta_{g_j}$, with $G_0(\{0\}) = 1-\varepsilon$. He further stated that "a plausible guess is that $p_j = cp^j$, $g_j = jg$ for $j > 0$," that is, that G_0 is a two-sided geometric distribution. This problem is connected with (1) through Brown's (1971) identity

$$r(G) \stackrel{\triangle}{=} \inf_d r(G,d) = 1 - I(\Phi * G).$$

Thus Mallows' problem is identical to (1) except that $\mathcal{G}'_\varepsilon$ allows two-sided contamination distributions.

Parametric robustness. Bickel (1983, 1984) studied the question of optimal (minimax) estimation of θ subject to good risk properties at $\theta = 0$, namely, to calculate

$$\inf_d \left\{ \sup_\theta R(\theta, d) \colon R(0,d) \le s \right\}. \tag{5}$$

Bayes-minimax compromises (of which this is an important special case) were earlier studied by Hodges and Lehmann (1952) and Efron and Morris (1971). For related work, see also Berger (1982) and Marazzi (1980). As Bickel notes, this problem is central to estimation of a parameter η in the presence of a nuisance parameter θ, when one believes that $\theta = 0$ but desires robustness against the possibility of error.

Introducing a Lagrange multiplier ε shows an equivalent form of (5) to be

(1), using $\mathcal{G}'_\varepsilon$ in place of $\mathcal{G}_\varepsilon$:

$$\begin{aligned}
&\inf_d \left\{ (1-\varepsilon)R(0,d) + \varepsilon \sup_\theta R(\theta,d) \right\} \\
&= \inf_d \left\{ (1-\varepsilon)R(0,d) + \varepsilon \sup_{H\in\mathcal{P}(-\infty,\infty)} R(H,d) \right\} \\
&= \inf_d \sup_{\mathcal{G}'_\varepsilon} R(G,d).
\end{aligned} \tag{6}$$

Equivalence of the two problems means that, for given s and optimal rule d_s^* in (5), there exists $\varepsilon = \varepsilon(s)$ such that the optimal rule d_ε^* in (6) equals d_s^*. Bickel also describes in greater detail the connections between (6), the identities of Brown and Stein and Mallows' problem.

The remainder of this note is organized as follows. Section 2 presents a three-term asymptotic expansion for the minimax risk (1) and compares it with numerical approximations. Higher-order properties of the simple l_1-rules studied in Donoho, Johnstone, Hoch and Stern (1992) are briefly described. Section 3 outlines the proof of the main result, and Section 4 briefly describes connections with other constraint sets (l_p-balls) and with wavelet transforms.

2. Main results. Introduce, following Mallows, the geometric prior

$$G_\varepsilon = (1-\varepsilon)\sum_{k=0}^{\infty} \varepsilon^k \delta_{kn}.$$

The lattice spacing $n = n(\varepsilon)$ is defined implicitly along with an additional parameter $a = a(n)$ by the two equations

$$\phi(n+a) = \varepsilon\phi(a), \tag{7}$$

$$(n+a)\phi(a) = 2\Phi(a). \tag{8}$$

Here Φ and ϕ are the standard Gaussian distribution and density functions, respectively. The origin of these equations is discussed later, but for now we note that the orders of magnitude of $n(\varepsilon)$ and $a_n = a(n)$ are given by

$$n \sim \sqrt{2\log\varepsilon^{-1}}, \qquad a_n \sim \sqrt{2\log c_0 n}, \qquad c_0 = (2\sqrt{2\pi})^{-1}. \tag{9}$$

Define the Bayes risk of a prior G and the maximum risk (relative to $\mathcal{G}_\varepsilon$) of an estimator d, respectively, by

$$r(G) = \inf_d r(G,d), \qquad m(d,\varepsilon) = \sup_{\mathcal{G}_\varepsilon} r(G,d).$$

Expression (1) for minimax risk now takes the simpler form

$$\sup_{\mathcal{G}_\varepsilon} r(G) = m(\varepsilon) = \inf_d m(d,\varepsilon).$$

THEOREM 1. *The minimax risk*

$$m(\varepsilon) = \varepsilon n^2 \Phi(a_n) - \frac{\pi^2}{6}\varepsilon a_n \phi(a_n) + O\Big(\frac{\varepsilon a_n^2}{n^2}\Big). \tag{10}$$

The geometric prior G_ε is asymptotically least favorable and the Bayes estimator d_{G_ε} is asymptotically minimax to this order:

$$m(d_{G_\varepsilon}, \varepsilon) - r(G_\varepsilon) = O\Big(\frac{\varepsilon a_n^2}{n^2}\Big). \tag{11}$$

Formula (10) will be checked numerically below. For theoretical interpretation, some remarks are in order. Definition (8) shows that $n\phi(a_n) = 2 + O(a_n/n)$, which converts the second term to $(\pi^2/3)(\varepsilon a_n/n)$ without increasing the order of error. Setting $\widetilde{\Phi} = 1 - \Phi$, we obtain

$$m(\varepsilon) = \varepsilon n^2 - \varepsilon n^2 \widetilde{\Phi}(a_n) - \Big(\frac{\pi^2}{3}\Big)\frac{\varepsilon a_n}{n} + O\Big(\frac{\varepsilon a_n^2}{n^2}\Big). \tag{12}$$

Since $n^2\widetilde{\Phi}(a_n) \sim 2n/a_n$, this shows that (10) is actually a *third*-order expansion of $m(\varepsilon)$ and that the geometric prior is *third*-order minimax. We are unaware of any previous settings in which third-order minimaxity in a parameter other than sample size has been established, although Levit (1987) gives bounds on the third term of the minimax risk for estimating a Gaussian mean restricted to an interval as the noise level decreases.

Finally, we note that, expressed in terms of ε, the order of error is

$$\varepsilon a_n^2 n^{-2} \sim \varepsilon\Big(\frac{\log\log\varepsilon^{-1}}{\log\varepsilon^{-1}}\Big).$$

Simpler approximations. The dependence of $m(\varepsilon)$ on ε in (10) is implicit, since it involves the solutions of equations (7) and (8). Two cruder approximations may be derived (see Section A2), the second involving only elementary functions. To this end, let $n_0^2 = 2\log\varepsilon^{-1}$ and recall that $c_0 = (2\sqrt{2\pi})^{-1}$;

$$m(\varepsilon) = \varepsilon n_0^2 \Phi\big((2\log c_0 n_0)^{1/2}\big) - 2\varepsilon n_0 (2\log c_0 n_0)^{1/2} + O(\varepsilon \log n_0) \tag{13}$$

$$= \varepsilon n_0^2 - 2\varepsilon n_0(2\log c_0 n_0)^{1/2} - \varepsilon n_0 (2\log c_0 n_0)^{-1/2} + O\big(\varepsilon n_0 (\log n_0)^{-3/2}\big). \tag{14}$$

First-order minimax rules. In the two-sided version of (1) that uses $\mathcal{G}'_\varepsilon$ in place of $\mathcal{G}_\varepsilon$, the first-order asymptotic minimax behavior of $m(\varepsilon)$ was described by Bickel (1983). The corresponding result for (1) is proved in Donoho, Johnstone, Hoch and Stern (1992) and says simply that

$$m(\varepsilon) = \varepsilon n_0^2 (1 + o(1)). \tag{15}$$

Denote by l_1-rule an estimator of the form $d_\lambda(x) = \max(x - \lambda, 0)$, for $\lambda > 0$. The name arises because $\widehat{\theta}_\lambda = (d_\lambda(x_i))$ is the coordinatewise solution to the problem

$$\min_{\theta \geq 0} \sum_i (\theta_i - x_i)^2 + 2\lambda \sum_i \theta_i.$$

Both Bickel (1983) and Donoho, Johnstone, Hoch and Stern (1992) show for their respective settings that l_1-rules with $\lambda(\varepsilon) \sim n_0(\varepsilon) = \left(2\log\varepsilon^{-1}\right)^{1/2}$ are first-order asymptotically minimax:

$$m(d_{\lambda_\varepsilon}, \varepsilon) \sim m(\varepsilon).$$

However, as Bickel notes, the first-order approximation is rather crude and not practically useful. The next result gives higher-order behavior of the best l_1-rule.

THEOREM 2. *Let λ_ε minimize $m(d_\lambda, \varepsilon)$. Then*

$$\lambda_\varepsilon = \left(n_0^2 - 6\log n_0 - \log 2\pi\right)^{1/2} + O\left(n_0^{-3}\log n_0\right) \tag{16}$$

and

$$m(d_{\lambda_\varepsilon}, \varepsilon) = \varepsilon\left[n_0^2 - 6\log n_0 - \log 2\pi + 3 + 18n_0^{-2}\log n_0 + O\left(n_0^{-2}\right)\right]. \tag{17}$$

Discussion of (17). A simple relation between n_0 and n results from putting the definition $\varepsilon = \exp(-n_0^2/2)$ into (7):

$$n_0^2 = n^2 + 2na_n. \tag{18}$$

Substituting (18) into (17) and comparing with (10) shows immediately that the best l_1-rule is not even second-order minimax. Although hardly needed here for theoretical purposes, the extra terms in expansion (17) occur naturally in the proof and are retained to provide possibly greater numerical accuracy.

Numerical evaluation. Table 1 compares the various approximations for a range of values of ε. Note, for example, that $\varepsilon = 10^{-6}$ corresponds to a single nonzero pixel in a 1000×1000 screen image. To calculate approximation (10), values of n and a_n were obtained by solving equations (7) and (8) numerically.

As a check on these values, one can take geometric priors

$$G_{\nu,\varepsilon} = (1 - \varepsilon)\sum_0^\infty \varepsilon^k \delta_{k\nu},$$

and compute the Bayes risk $r(G_{\nu,\varepsilon})$ by numerical integration and summation. By numerical minimization over ν, one obtains an optimal spacing $\nu = n_{\text{lo}}(\varepsilon)$ and thus a lower bound $m_{\text{lo}}(\varepsilon)$ to $m(\varepsilon)$ that Mallow's conjecture suggests ought to be quite sharp [and would in fact equal $m(\varepsilon)$ were the conjecture to be exactly true]. The results for selected values of ε are displayed in Table 2.

TABLE 1
Various approximations to the minimax risk $m(\varepsilon)$: n and a_n are the solutions to equations (7) *and* (8); $n_0 = (2\log\varepsilon^{-1})^{1/2}$; *equation* (10) *is the third-order approximation provided by Theorem* 1; (15) *is the first-order approximation*; (13) *and* (14) *are simpler, but less accurate, versions of* (10)

ε	n	n_0	a_n	$m(\varepsilon)$(10)	$m(\varepsilon)$(15)	$m(\varepsilon)$(13)	$m(\varepsilon)$(14)
7.5e−2	2.40	2.27	−0.12	2.00e−1	3.88e−1	NA	NA
5.0e−2	2.47	2.44	−0.02	1.50e−1	2.99e−1	NA	NA
4.0e−2	2.52	2.53	0.01	1.28e−1	2.57e−1	NA	NA
3.0e−2	2.58	2.64	0.06	1.04e−1	2.10e−1	NA	NA
2.0e−2	2.66	2.79	0.13	7.69e−2	1.56e−1	NA	NA
1.0e−2	2.81	3.03	0.23	4.52e−2	9.21e−2	NA	NA
5.0e−3	2.96	3.25	0.30	2.62e−2	5.29e−2	NA	NA
2.0e−3	3.15	3.52	0.39	1.25e−2	2.48e−2	NA	NA
1.0e−3	3.29	3.71	0.45	7.04e−3	1.38e−2	NA	NA
1.0e−4	3.73	4.29	0.59	9.78e−4	1.84e−3	NA	NA
1.0e−5	4.15	4.79	0.69	1.26e−4	2.30e−4	NA	NA
1.0e−6	4.53	5.25	0.77	1.57e−5	2.76e−5	1.39e−5	7.31e−6
1.0e−7	4.89	5.67	0.84	1.88e−6	3.22e−6	1.66e−6	1.51e−6
1.0e−8	5.23	6.06	0.90	2.19e−7	3.68e−7	1.94e−7	1.95e−7
1.0e−9	5.55	6.43	0.94	2.52e−8	4.14e−8	2.24e−8	2.32e−8
1.0e−10	5.86	6.78	0.98	2.84e−9	4.60e−9	2.54e−9	2.67e−9
1.0e−11	6.16	7.11	1.02	3.18e−10	5.06e−10	2.85e−10	3.02e−10
1.0e−12	6.45	7.43	1.05	3.51e−11	5.52e−11	3.17e−11	3.36e−11
1.0e−15	7.24	8.31	1.14	4.54e−14	6.90e−14	4.14e−14	4.40e−14
1.0e−18	7.98	9.10	1.20	5.60e−17	8.28e−17	5.16e−17	5.46e−17

A corresponding numerical upper bound for $m(\varepsilon)$ was obtained by locating the maximum $\theta_{\max}$ of the risk function of the Bayes rule corresponding to $G_{n_{lo}(\varepsilon),\,\varepsilon}$ and evaluating the left-hand side of (6) using this Bayes rule and $\theta_{\max}$ to obtain the upper bound $m_{up}(\varepsilon)$. Table 2 shows that even at $\varepsilon = 0.2$ the upper and lower bounds differ by only 2.5%, and the bounds become tighter as ε decreases.

The agreement between Table 2 and (10) is remarkable at $\varepsilon = 0.02$ and

TABLE 2
Numerically computed approximations to the minimax risk $m(\varepsilon)$ and risk of the best l_1-rule. Compare with asymptotic approximations in Table 1 [*column* (10)] *and Table* 3 [*column* (17)], *respectively*; $n_{lo}(\varepsilon)$ *is the numerical approximation to $n(\varepsilon)$ corresponding to $m_{lo}(\varepsilon)$*

ε	$m_{lo}(\varepsilon)$	$m_{up}(\varepsilon)$	$m(d_{\lambda_\varepsilon}, \varepsilon)$	$n_{lo}(\varepsilon)$	$\theta_{\max}$
0.200	0.3907	0.4100	0.4100	2.4781	4.2655
0.100	0.2481	0.2604	0.2600	2.5890	4.6248
0.050	0.1534	0.1602	0.1670	2.7018	4.9684
0.020	0.0784	0.0811	0.0873	2.8564	5.4028
0.010	0.0461	0.0474	0.0522	2.9776	5.7182
0.005	0.0267	0.0272	0.0306	3.1012	6.0227
0.002	0.0127	0.0128	0.0148	3.2662	6.4096
0.001	0.0071	0.0072	0.0084	3.3911	6.6913

TABLE 3

Maximum risk over ε-black objects for the best l_1-rule; λ_ε and $m(d_{\lambda_\varepsilon}, \varepsilon)$ are the approximations given in Theorem 2 by (16) and (17), respectively; λ_t and $m(d_{\lambda_t}, \varepsilon)$ denote the numerically determined optimal values [λ_t is the root of (44) and $m(d_{\lambda_t}, \varepsilon) = H(\lambda_t)$]. For comparison, $m(\varepsilon)$ given by (10) is the third-order approximation to minimax risk

ε	λ_ε	λ_t	$m(d_{\lambda_\varepsilon}, \varepsilon)$	$m(d_{\lambda_t}, \varepsilon)$	$m(\varepsilon)$
7.5e−2	NA	1.00	NA	0.219	0.195
5.0e−2	NA	1.15	NA	0.167	0.148
4.0e−2	NA	1.23	NA	0.143	0.126
3.0e−2	NA	1.34	NA	0.116	0.103
2.0e−2	NA	1.48	NA	8.73e−02	7.65e−2
1.0e−2	0.84	1.72	5.87e−2	5.22e−02	4.51e−2
5.0e−3	1.29	1.94	3.33e−2	3.06e−02	2.62e−2
2.0e−3	1.74	2.23	1.56e−2	1.48e−02	1.25e−2
1.0e−3	2.02	2.43	8.81e−3	8.41e−03	7.04e−3
5.0e−4	2.28	2.63	4.89e−3	4.72e−03	3.92e−3
2.0e−4	2.58	2.88	2.22e−3	2.17e−03	1.79e−3
1.0e−4	2.80	3.06	1.22e−3	1.19e−03	9.78e−4
5.0e−5	3.00	3.23	6.67e−4	6.53e−04	5.32e−4

$\varepsilon = 0.01$, suggesting that (10) is likely to be quite accurate for smaller ε. Over the range of numerical calculations, approximation (10) is typically smaller than $m_{\text{lo}}(\varepsilon)$, but by an amount less than the difference between the upper and lower bounds $m_{\text{up}} - m_{\text{lo}}$. This implies a relative error in $m(\varepsilon)$ of about 6% at $\varepsilon = 0.05$, dropping to about 2.5% at $\varepsilon = 0.001$. By contrast, the first-order expression (15) is too large by a factor of 2 for plausible values of ε, and the approximations (13) and (14), while considerably better for small ε, are useless above $\varepsilon \sim 10^{-6}$, due to the logarithms being undefined.

Table 3 shows the approximations to λ_ε and $m(d_{\lambda_\varepsilon}, \varepsilon)$ as given by Theorem 2. Formulas (34) and (44) in the proof of Theorem 2 show that it is in principle not difficult to calculate $m(d_{\lambda_\varepsilon}, \varepsilon)$ directly by evaluating the critical λ_ε numerically and then substituting into (34). This was done to provide a check on (17). For $\varepsilon \geq 0.02$, the approximation is undefined and, indeed, the asymptotic formula is barely satisfactory over the range shown. It would presumably improve for $\varepsilon \leq 10^{-6}$. In any case, the table shows clearly the higher-order suboptimality of the l_1-rule over a wide range of ε.

3. Approximating minimax rules. This section outlines the proofs of Theorems 1 and 2, with some details deferred to the Appendix. The usual device for proving that the geometric prior G_ε is asymptotically minimax is to show that its Bayes risk is close to the maximum risk $m(d_{G_\varepsilon}, \varepsilon)$ of the Bayes rule corresponding to G_ε. For G_ε, the Bayes risk

$$r(G_\varepsilon) = (1 - \varepsilon)R(0, d_{G_\varepsilon}) + \varepsilon(1 - \varepsilon)\sum_{k=1}^{\infty} \varepsilon^{k-1}R(kn, d_{G_\varepsilon}). \tag{19}$$

Since the "maximum risk" is given by

$$m(d,\varepsilon) = \sup\left\{(1-\varepsilon)R(0,d) + \varepsilon \int R(\theta,d)H(d\theta) \colon H \in \mathcal{P}([0,\infty))\right\},$$

it may be rewritten for the Bayes estimator as

$$m(d_{G_\varepsilon},\varepsilon) = (1-\varepsilon)R(0,d_{G_\varepsilon}) + \varepsilon \sup_{\theta\geq 0} R(\theta,d_{G_\varepsilon}). \tag{20}$$

To establish (11), the method is to construct a lattice spacing n so that up to terms of order $O(a_n^2/n^2)$, the maximum of $R(\theta,d_{G_\varepsilon})$ is attained at each of the support points $\{kn;\, k=1,2,\ldots\}$ of the positive component of G_ε.

To this end, approximations to the risk of d_{G_ε} are useful. The posterior distribution $G_\varepsilon(\{kn\}|x)$ is proportional to $\varepsilon^k\phi(x-kn)$ and, for n in the range of interest ($n\geq 2.5$), this is almost entirely concentrated on at most two points. So, introduce the change of variables $x = k_0 n + z$ and note that

$$G_\varepsilon(\{kn\}|x) \propto \begin{cases} \varepsilon^{-1}\phi(z+n), & \text{for } k = k_0 - 1,\\ \phi(z), & \text{for } k = k_0,\\ \varepsilon\phi(z-n), & \text{for } k = k_0+1. \end{cases}$$

In fact, for most z, a single value of k dominates, so that the Bayes estimator d_{G_ε}, being the mean of the posterior distribution, approximately equals kn. The contributions from the support points $(k_0-1)n$ and $k_0 n$ balance when $\varepsilon^{-1}\phi(z+n) = \phi(z)$, that is, when $\log\varepsilon^{-1} - nz - n^2/2 = 0$. The defining equation (7) shows that this occurs when $z = a_n$, that is, when $x = k_0 n + a_n$. One finds similarly that $k_0 n$ and $(k_0+1)n$ balance when $z = a_n + n$. Thus the posterior distribution is approximately periodic in x with period n, at least for x situated away from the left endpoint of the support of G_ε.

For x within $n/2$ standard deviations of $k_0 n + a_n$, we therefore approximate $d_{G_\varepsilon}(x)$ by the Bayes rule for a two-point prior, setting mass proportional to 1 at $k_0 n$ and to ε^{-1} at $(k_0-1)n$. The approximation is slightly modified for $x \leq n/2 + a$ to the Bayes rule for the two-point prior $\delta_0 + \varepsilon\delta_n$.

Explicitly, let $z \sim N(\zeta,1)$ and $\zeta \sim \delta_0 + \varepsilon^{-1}\delta_{-n}$. Then

$$d_0(z) = E(\zeta \mid z) = \frac{-ne^{-nz}}{e^{-na} + e^{-nz}}. \tag{21}$$

Let $k_0(x)$ denote the *positive* integer k for which $kn + a$ is closest to x, and set

$$d_\varepsilon(x) = nk_0(x) + d_0\big(x - nk_0(x)\big). \tag{22}$$

Estimator d_ε uniformly approximates d_{G_ε} both pointwise and in risk (see Section A3):

$$\big|d_{G_\varepsilon}(x) - d_\varepsilon(x)\big| \leq Mne^{-n^2/2}, \tag{23}$$

$$\big|R(\theta,d_{G_\varepsilon}) - R(\theta,d_\varepsilon)\big| \leq Mn^2e^{-n^2/2}. \tag{24}$$

(Here and in the sequel M denotes a generic constant, not necessarily the same at each appearance.)

Section A4 shows that any maxima of $\theta \to R(\theta, d_\varepsilon)$ with values greater than $(1-\delta)n^2$ occur only within intervals of the form $[kn - \beta n, kn + \beta n]$, for $k \geq 1$ and a small positive constant β. Set $\gamma = \frac{1}{2} - \beta$. Now replace consideration of estimator d_ε on each of these intervals by a single two-point prior Bayes rule: The inequality (see Section A5)

$$\text{(25)} \qquad \sup\left\{\left|R(\theta, d_\varepsilon) - R(\theta - k_0 n, d_0)\right| : \left|\theta - k_0 n\right| \leq \beta n\right\} \leq Mn\phi(\gamma n - a_n)$$

means we need only look at the maximum of the risk function

$$R_0(\zeta) = E_\zeta\left(d_0(z) - \zeta\right)^2$$

on the single interval $|\zeta| \leq \beta n$.

The logistic form (21) of d_0 still precludes explicit evaluation of $R_0(\zeta)$, so consider as a further approximation the step function

$$\text{(26)} \qquad d_{00}(z) = -nI(z < a_n),$$

with easily evaluated risk function and derivative

$$\text{(27)} \qquad R_{00}(\zeta) = E_\zeta\left(d_{00}(z) - \zeta\right)^2 = \left(n^2 + 2n\zeta\right)\Phi(a_n - \zeta) + \zeta^2,$$

$$\text{(28)} \qquad R'_{00}(\zeta) = 2n\Phi(a_n - \zeta) - \left(n^2 + 2n\zeta\right)\phi(a_n - \zeta) + 2\zeta.$$

For large n, Taylor expansions give the error in these approximations (see Section A6):

$$\text{(29)} \qquad R_0(\zeta) - R_{00}(\zeta) = -n\phi(a_n - \zeta)\left[1 + \frac{\pi^2}{6}\frac{a_n - \zeta}{n} + O\left(n^{-2}(a_n - \zeta)^2\right)\right],$$

$$\text{(30)} \qquad R'_0(\zeta) - R'_{00}(\zeta) = -n(a_n - \zeta)\phi(a_n - \zeta)\left[1 + O\left(n^{-1}|a_n - \zeta|\right)\right],$$

valid uniformly in $|\zeta| \leq$ const.

The aim now is to choose a_n so that the risk function $R_0(\zeta)$ attains its maximum at a support point of the prior, namely, 0. A sign change argument (see Section A7) shows that in fact $R_0(\zeta)$ has a single local maximum. Together, (28) and (30) show that $R'_0(0) \approx 0$ if a is required to satisfy the second defining equation (8). Further approximation (see Section A8), in (30) shows that, for this choice of a_n,

$$\text{(31)} \qquad R'_0\left(-n^{-1}\right) \sim na_n\phi(a_n) > 0, \qquad R'_0(0) \sim -\tfrac{1}{6}\pi^2 a_n^2 \phi(a_n) < 0.$$

Thus, for large n, the maximum of $R_0(\zeta)$ occurs at some point $\xi^* \in (-n^{-1}, 0)$. From (8) and (9) follows $n\phi(a_n) \sim 2$, and since $R_0(\zeta)$ turns out to be concave on $(-n^{-1}, 0)$,

$$\text{(32)} \qquad R_0(\xi^*) - R_0(0) \leq n^{-1}|R'_0(0)| = O(a_n^2/n^2),$$

which is the order of error claimed in Theorem 1.

Together, (24), (25) and (32) imply that

$$\sup R(\theta, d_{G_\varepsilon}) = R_0(0) + O\big(n^2 e^{-n^2/2} + n\phi(\gamma n - a_n) - a_n^2/n^2\big),$$
$$R(kn, d_{G_\varepsilon}) = R_0(0) + O\big(n^2 e^{-n^2/2} + n\phi(\gamma n - a_n)\big), \qquad k \ge 1.$$

Thus (19) and (20) show that, up to terms of order $O(\varepsilon a_n^2/n^2)$, d_{G_ε} is asymptotically minimax. This establishes (11), the second part of Theorem 1.

The approximation (10) to the minimax risk requires also the risk of d_{G_ε} at $\theta = 0$. Replacing $\mathcal{G}_\varepsilon$ by the two-point prior $\delta_0 + \varepsilon\delta_n$, yields an approximation (see Section A9),

$$R\big(0, d_{G_\varepsilon}\big) = \varepsilon n\phi(a_n)\big[1 + O(a_n^2/n^2)\big]. \tag{33}$$

Since $r(G_\varepsilon) \le m(\varepsilon) \le m(d_{G_\varepsilon}, \varepsilon)$, we now evaluate the minimax risk from (27), (29) and (33) as

$$m(\varepsilon) = (1-\varepsilon)\varepsilon n\phi(a_n) + \varepsilon\left\{ n^2\Phi(a_n) - n\phi(a_n)\left(1 + \frac{\pi^2}{6}\frac{a_n}{n}\right) + O\left(\varepsilon\frac{a_n^2}{n^2}\right)\right\},$$

which reduces to (10). □

REMARK. It does not seem likely that the geometric prior is exactly least favorable for $\varepsilon > 0$ in this setting—it would be necessary to choose the lattice spacing n so that the risk of d_{G_ε} was constant at $\theta = kn$, for $k = 1, 2, \ldots$. However, the geometric prior is probably asymptotically minimax to higher orders: This might be shown using a three-piece linear approximation to d_0, tangent to d_0 at 0, in place of (26).

PROOF of Theorem 2 (Asymptotics for l_1-rule). These are fairly straightforward, since the maximum risk of $\theta \to R(\theta, d_\lambda)$ on $[0, \infty)$ occurs (see Section A10) at ∞, for any $\lambda > 0$, and equals $1 + \lambda^2$. Evaluating also $R(0, d_\lambda)$, we obtain

$$m(d_{\lambda_\varepsilon}, \varepsilon) = \inf_\lambda H(\lambda) = \inf_\lambda (1-\varepsilon)\Big[(\lambda^2+1)\widetilde{\Phi}(\lambda) - \lambda\phi(\lambda)\Big] + \varepsilon(1+\lambda^2). \tag{34}$$

The derivatives of H are easily found and, in particular, show that H is convex. The form (16) of λ_ε is obtained (see Section A11) by a one-step approximation to the initial value $n_0^2 = 2\log\varepsilon^{-1}$ and substitution into H yields (17). □

4. Discussion. The problem (1) studied in this paper is essentially identical to Mallows' problem involving two-sided contamination. Problem (1) is technically simpler, since one does not have to deal with the effect of prior probability mass at negative atoms, but it seems likely that the techniques developed here would readily yield corresponding expansions and third-order results for Mallows' model.

Most of the recent work on minimax properties of various models for sparsity (references below) has concentrated on first-order risk behavior, in part for reasons set out in Donoho, Johnstone, Kerkyacharian and Picard (1994). Two- and three-point asymptotically least favorable prior distributions arise commonly in this work. Thus, an additional contribution of this paper is to provide tools that might be adapted to study of higher-order risk properties in these related settings.

Sparsity and wavelet bases. As preparation for studying the beneficial properties of wavelet bases in function estimation, Donoho and Johnstone (1994a, b) have investigated minimax estimation in Gaussian white noise of a mean vector known to lie in a finite-dimensional l_p-ball $0 < p < \infty$. Since $\|\theta\|_{n,p}^p = \Sigma_1^n |\theta_i|^p \to \#\{i: |\theta_i| > 0\}$ as $p \to 0$, the "nearly black" conditions studied here may be regarded in some sense as l_0-ball constraints.

Indeed, a characteristic property of the wavelet transform is that the wavelet coefficients of smooth or piecewise smooth functions are *typically* sparse—at higher resolution levels, only those coefficients in the vicinity of a discontinuity of the function or its derivatives are significantly nonzero.

Not surprisingly, therefore, the ideas and methods of each paper are related. Thus, an asymptotically least favorable distribution over $\{\theta \geq 0: \|\theta\|_{n,p} \leq r\}$ [as $n \to \infty$ in model (2)] is $(1 - \varepsilon)\delta_0 + \varepsilon\delta_\mu$, where ε and μ are determined by $\varepsilon\mu^p = n^{-1}(r/\sigma)^p$ and equations (7) and (8) (with n replaced by μ). Indeed a two-point prior of this form is all that is needed to establish first-order asymptotic minimaxity in the nearly black setting in Donoho, Johnstone, Hoch and Stern (1992). The l_p-balls (and their weak analogs) with $0 < p < 1$ arise naturally in the study of optimal spatially adaptive function estimates [such as, e.g., variable kernel estimators—Donoho (1992) and Johnstone (1994)].

Related work. In problem (2) the large sample ($n \to \infty$) limit plus constraints led to a "restricted Bayes" minimax problem (1) for estimating a Gaussian mean subject to constraints on the class of prior distributions. [For more on the large sample limiting process, see Johnstone (1994).] A number of other such restricted minimax problems have been studied: As noted previously, Donoho and Johnstone (1994a) study moment constraints; see also Feldman (1991). The limit as $p \to \infty$ yields the case when the mean is known to be bounded in absolute value, by η, say. See, for example, Casella and Strawderman (1981) and Donoho, Liu and MacGibbon (1990). Casella and Strawderman gave the exact minimax value for $\eta < 1.05$, when a symmetric two-point prior $\frac{1}{2}\delta_\eta + \frac{1}{2}\delta_{-\eta}$ is exactly least favorable. Although this paper also studies situations in which "most of the mass is small," the situation here is of large, infrequent nonzero components, as compared to uniformly nonzero but small components. The relation between these "sparse" and "dense" settings becomes clearer in Donoho and Johnstone (1994a), where the L_p moment constraint is combined with an l_q loss function (here of course $q = 2$). As $\eta \to 0$, the "sparse" situation arises when $p < q$ (e.g., $p = 0$), and the "dense" case when $p \geq q$ (e.g., $p = \infty$).

APPENDIX

A1. Note on terminology. The use of the term "minimax risk" in the title of the paper derives from the motivating problem (4) of estimating a high-dimensional vector with few nonzero components. The reduced form (1) poses a minimax problem in which Nature chooses from a restricted class of prior distributions on a single Gaussian mean. After noting connections with the restricted Bayes ideas of Hodges and Lehmann (1952), Bickel (1983) refers to (1) as a restricted minimax problem. The fuller term "restricted Bayes minimax" emphasizes that the payoff function $r(G,d)$ involves expectation with respect to the prior distribution as well as the data distribution.

A2. The derivation of approximations (13) and (14) involves three steps. First, define $a^0 = (2\log c_0 n)^{1/2}$ [cf. (9)] and a one-step Newton approximation a^n to the solution of (8) starting from a^0. Second, express the right-hand side of (10) in terms of n and a^0. Finally, obtain an expression in terms of n_0 rather than n by exploiting (18) derived from (7). This results in a sequence of approximations, of which (13) and (14) are the crudest. The rather tedious details are omitted.

A3. Proof of (23) and (24). Denote the unnormalized posterior $\varepsilon^k\phi(x-kn)$ by $\phi_k = \phi_k(x)$. Fix $k_0 \geq 1$ and consider $x \in [(k_0-1)n + a + n/2,\ k_0 n + a + n/2]$. In this interval, setting $z = x - k_0 n$, we have

$$n^{-1}\left[d_{G_\varepsilon}(x) - nk_0\right] = \frac{\sum_{-k_0}^{\infty} j\phi_j}{\sum_{-k_0}^{\infty}\phi_j},$$
$$n^{-1}\left[d_\varepsilon(x) - nk_0\right] = \frac{-\phi_{-1}}{\phi_0 + \phi_{-1}}.$$

We apply the equality

$$\frac{a+\delta_1}{b+\delta_2} = \frac{a}{b} + \left(\frac{\delta_1}{b} - \frac{a}{b}\frac{\delta_2}{b}\right)\left(1 + \frac{\delta_2}{b}\right)^{-1}, \tag{35}$$

for

$$a = -\phi_{-1}, \qquad b = \phi_0 + \phi_1,$$
$$\delta_1 = \sum_{j\neq 0,-1} j\phi_j, \qquad \delta_2 = \sum_{j\neq 0,-1}\phi_j.$$

For $|z-a| \leq n/2$, the rapid decay of Gaussian tails ensures that δ_1 and δ_2 are at most a constant multiple of the leading terms $\delta_3 = \phi_{-2} + \phi_1$. Add the fact that $|a/b| \leq 1$, and it will be enough to show that δ_3/b is $O(e^{-n^2/2})$. However, simple algebra using (7), namely, $\log\varepsilon = -n^2/2 - an$, yields

$$\log\phi_1/\phi_0 = (z-a)n - n^2 \leq -n^2/2$$
$$\log\phi_{-2}/\phi_{-1} = -(z-a)n - n^2 \leq -n^2/2.$$

For $k = 0$ (so that $x = z$) and $x \leq n/2 + a$, apply a similar argument using $a = \phi_1$ and $b = \phi_0 + \phi_1$. In this case,

$$\log \phi_2/\phi_0 = 2n(z - a_n) - 3n^2 \leq -2n^2;$$

so one obtains the stronger estimate

$$d_{G_\varepsilon}(x) = d_\varepsilon(x) + O\left(ne^{-2n^2}\right). \tag{36}$$

To establish (23), note the global bounds

$$|d_{G_\varepsilon}(x) - x_+| \leq n, \qquad |d_\varepsilon(x) - x_+| \leq n, \tag{37}$$

where $x_+ = x \vee 0$. From (23),

$$\begin{aligned} R(\theta, d_{G_\varepsilon}) - R(\theta, d_\varepsilon) &\leq E_\theta \left|d_{G_\varepsilon} - d_\varepsilon\right| \left|d_{G_\varepsilon} + d_\varepsilon - 2\theta\right| \\ &\leq O(ne^{-n^2/2}) E_\theta \left[2n + 2|x_+ - \theta|\right] \\ &= O\left(n^2 e^{-n^2/2}\right). \end{aligned}$$

A4. Maxima of $R(\theta, d_\varepsilon)$. The key to showing that $R(\theta, d_\varepsilon)$ has maxima only within intervals $[kn - \beta n, kn + \beta n]$ is to show, for $k_0 \geq 1$ and $\theta - k_0 n \in [\beta n, (1-\beta)n]$, that $d_\varepsilon(x) - k_0 n$ is essentially zero over a somewhat larger range of x. More precisely, one can pick $\gamma < \beta/2$ and note that $d_\varepsilon(x) - k_0 n$ is an odd function about $x = k_0 n + a + n/2$, to verify that

$$\sup\left\{|d_\varepsilon(x) - k_0 n| \colon x - k_0 n - a \in \left[\gamma n, (1-\gamma)n\right]\right\} \leq e^{-\gamma n^2}. \tag{38}$$

Using (38) on the set $\{z = x - k_0 n \in [a + \gamma n, a + (1-\gamma)n]\}$ and (37) plus bounds on Gaussian tails on the complement of the set leads to

$$E_\theta[d_\varepsilon(x) - \theta]^2 \leq n^2\left[(1-\beta)^2 + M\widetilde{\Phi}(\beta n/4)\right]. \tag{39}$$

To establish the claim, it remains to establish (39) also for $\theta \leq n - \beta n$. This is done analogously, but now using the inequality $d_\varepsilon(x) \leq n[1 + \exp(n^2\gamma)]^{-1}$ valid for $x \leq a + (1-\gamma)n$.

A5. Proof of (25). The bound (25) is easily established after noting that $d_\varepsilon(k_0 n + z) = d_0(z)$, for $|z - a| \leq n/2$, so that (if $\zeta = \theta - k_0 n \in [-\beta n, \beta n]$, $z = x - k_0 n$),

$$\begin{aligned} \left|R(\theta, d_\varepsilon) - R(\zeta, d_0)\right| &\leq E_\theta\left\{\left(d_\varepsilon(x) - \theta\right)^2 + \left(d_0(x) - \theta\right)^2, |z - a| > n/2\right\} \\ &\leq 2E_\zeta\left[n^2 + (z - \zeta)^2, |z - a| > n/2\right] \\ &\leq Mn\phi(\gamma n - a). \end{aligned}$$

A6. Proof of (29) and (30). Suppose first that γ is an even function and set $\gamma_k = \int_{-\infty}^{\infty} v^k \gamma(v)\,dv$. Then if $z \sim N(\zeta, 1)$,

$$\begin{aligned} E_\zeta \gamma\big[n(z - a_n)\big] &= n^{-1} \int_{-\infty}^{\infty} \gamma(v)\phi\big(a_n - \zeta + n^{-1}v\big)\,dv \\ &= \gamma_0 n^{-1}\phi(a_n - \zeta) + \tfrac{1}{2}\gamma_2 n^{-3}\phi''(a_n - \zeta) + \cdots. \end{aligned}$$

Using the derivatives $\phi''(x) = (x^2 - 1)\phi(x)$ and $\phi'''(x) = x(x^2 - 3)\phi(x)$, and arguing analogously for odd functions, we obtain the following lemma.

LEMMA 4. *Suppose $\int |v|^k \gamma(v)\,dv < \infty$, for $k \le 4$, and that $\gamma_n(w) = c_n\gamma(nw)$. Accordingly as γ is even or odd,*

$$E_\zeta \gamma_n(z - a_n) = c_n n^{-1}\phi(a_n - \zeta)\Big[\gamma_0 + O\Big(n^{-2}\big[(a_n - \zeta)^2 - 1\big]\Big)\Big],$$

or

$$E_\zeta \gamma_n(z - a_n) = -c_n n^{-2}(a_n - \zeta)\phi(a_n - \zeta)\Big[\gamma_1 + O\Big(n^{-2}\big[(a_n - \zeta)^2 - 3\big]\Big)\Big].$$

Before applying the lemma, we note the following relations between estimators d_0 and d_{00} and their risk functions:

$$d_0(z) = d_{00}(z) + \psi(z), \qquad \psi(z) = n\tilde{\psi}(nw), \qquad w = z - a_n,$$

where $\tilde{\psi}(w)$ is an *odd* function defined for $w > 0$ by

$$\tilde{\psi}(w) = \frac{-e^{-w}}{1 + e^{-w}}.$$

Using the risk (27) and its derivative (28) of d_{00} and using the identity $\partial E_\zeta g(z)/\partial\zeta = E_\zeta\,(z - \zeta)g(z)$,

$$R_0(\zeta) = R_{00}(\zeta) + E_\zeta\Big[\psi^2 + 2\psi(d_{00} - \zeta)\Big], \tag{40}$$

$$R_0'(\zeta) = R_{00}'(\zeta) + E_\zeta(z - \zeta)\Big[\psi^2 + 2\psi(d_{00} - \zeta)\Big] - 2E_\zeta\psi. \tag{41}$$

Applying Lemma 3 to the various components of (40) and (41) yields [showing leading terms only, and setting $r = \phi(a_n - \zeta)$]

$$\begin{aligned} r^{-1}E\psi^2 &= n\gamma_{a,0} + \cdots, \\ r^{-1}2E\psi d_{00} &= -n\gamma_{b,0} + (\zeta - a_n)\gamma_{b,1} + \cdots, \\ r^{-1}E\psi &= n^{-1}(\zeta - a_n)\gamma_{c,1} + \cdots, \\ r^{-1}E(z - a_n)\psi^2 &= n^{-1}(\zeta - a_n)\gamma_{d,1} + \cdots, \\ r^{-1}2E(z - a_n)\psi d_{00} &= -\gamma_{e,0} - n^{-1}(\zeta - a_n)\gamma_{ee,1} + \cdots, \\ r^{-1}E(z - a_n)\psi &= n^{-1}\gamma_{f,0} + \cdots, \end{aligned} \tag{42}$$

where

$$\gamma_{a,0} = 2\int_0^\infty [e^{-x}/(1+e^{-x})]^2\,dx = 2\log 2 - 1,$$
$$\gamma_{b,0} = 2\int_0^\infty e^{-x}/(1+e^{-x})\,dx = 2\log 2,$$
$$\gamma_{b,1} = 2\int_0^\infty xe^{-x}/(1+e^{-x})\,dx = \pi^2/6,$$

and it is unnecessary to compute the remaining constants. Combining the terms in (42) in accordance with (40) and (41) and tracking the error terms provided by Lemma 3 leads to (29) and (30).

A7. Single maximum for $R_0(\zeta)$. Using Stein's unbiased estimate for risk [Stein (1981)], we may write [setting $z = x - a$, $\zeta = \eta - a$ and $\gamma(x) = -n(1+e^{nx})^{-1}$],

$$\begin{aligned} R_0(\zeta) - c &= E_\eta\big(\gamma(x) + \eta - a\big)^2 - c \\ &= E_\eta\Big[\big(\gamma(x) + x - a\big)^2 - 2\gamma'(x) - (1+c)\Big]. \end{aligned}$$

For large values of n and $a = a(n)$, the integrand turns out to have at most four sign changes. Since the Gaussian kernel is totally positive of all orders, $R_0(\zeta) - c$ can have at most four sign changes. Since $R_0(\zeta) \nearrow \infty$ as $|\zeta| \nearrow \infty$ and R_0 dips down to values that are at most $O(a_n^2)$ near $\zeta = a + c\log n/n$ and $\zeta = -n + a$, it follows that R_0 can have at most one maximum of order n^2.

A8. Proof of (31). Here it is necessary to retain one extra term in the approximation to $R_0'(\zeta)$ derived from (41) and (42):

$$\begin{aligned} R_0'(\zeta) = &-\big[n^2 + 2n\zeta + n(a_n - \zeta) + (a_n - \zeta)^2\gamma_{b,1} + \gamma_{e,0}\big]\phi(a_n - \zeta) \\ &+ 2n\Phi(a_n - \zeta) + 2\zeta + o\big((a_n - \zeta)^2\phi(a_n - \zeta)\big). \end{aligned}$$

This expansion is valid at least for $|\zeta| \le$ constant, and so (31) follows by substitution. Differentiation shows that

$$R''(\zeta) \sim -n^2 a_n \phi(a_n - \zeta) < 0$$

on $[-n^{-1}, 0]$, so that $R_0(\zeta)$ lies below its tangent at $\zeta = 0$. This establishes the bound (32).

A9. Proof of (33). Let $d_1(x) = n\varepsilon\phi(x-n)/[\phi(x) + \varepsilon\phi(x-n)]$ be the Bayes rule for the two-point prior $\delta_0 + \varepsilon\delta_n$. Set $\Delta = d_{G_\varepsilon}^2 - d_1^2 = (d_{G_\varepsilon} - d_1)(d_{G_\varepsilon} + d_1)$. Let I_0, I_1 and I_2 be the events that x lies in $(-\infty, n/2 + a)$, $[n/2 + a, 3n/2 + a]$ and $(3n/2 + a, \infty)$, respectively. Using (23), (36), (37), bounds on the tail of the

Gaussian distribution and, finally, that $|d_{G_\varepsilon}(x)| \le |x| + n$, $|d_1(x)| \le n$,

$$E_0[\Delta, I_0] \le Mne^{-2n^2} E_0\big[|x| + n, I_0\big] \le Mn^2 e^{-2n^2},$$
$$E_0[\Delta, I_1] \le Mn^2 e^{-n^2/2} E_0\big[|x| + n, I_0^c\big] \le Mn^2 e^{-n^2/2}\phi(n/2),$$
$$E_0[\Delta, I_2] \le ME_0\big[|x|^2 + n^2, I_2\big] \le Mn\phi(3n/2).$$

Using equation (7), $\varepsilon = \exp(-n^2/2 - na_n)$, we have

$$\begin{aligned} E_0 d_1^2(x) &= \int_{-\infty}^{\infty} \frac{n^2}{[1 + e^{n(n+a-x)}]^2} \phi(x)\,dx \\ &= n\varepsilon\phi(a_n)\left[\int \frac{e^w\,dw}{(1+e^w)^2} + \frac{a_n}{n}\int \frac{we^w}{(1+e^w)^2}\,dw + O\Big(\frac{a_n}{n}\Big)^2\right], \end{aligned}$$

after making the substitution $w = n(n + a - x)$. This establishes (33) since the first integral equals 1 and the transformation $x = e^w[1 + e^w]^{-1}$ reduces the second integral to $\int_0^1 w(x)\,dx$, which vanishes since $w(x) = \log x/(1-x)$ is odd about $x = \frac{1}{2}$.

A10. Maximum risk of l_1-rule. The l_1-rule $d_\lambda(x) = (x-\lambda)_+$ may be written in the form $x + \psi(x)$ with $\psi(x) = -xI\{x \le \lambda\} - \lambda I\{x > \lambda\}$. Applying Stein's (1981) unbiased estimate of risk,

$$\begin{aligned} R(\theta, d_\lambda) &= 1 + E_\theta\big[2\psi'(x) + \psi^2(x)\big] \\ &= 1 + E_\theta\big[-2I(x < \lambda) + x^2 I(x \le \lambda) + \lambda^2 I(x > \lambda)\big]. \end{aligned} \tag{43}$$

The integrand in (43) crosses any horizontal line at most twice, so it follows from the variation diminishing property of the Gaussian kernel that $R(\theta, d_\lambda)$ attains its maximum on $[0, \infty)$ at either 0 or $+\infty$.

A11. Minimum of $H(\lambda)$. We record the following derivatives and approximations:

$$H'(\lambda) = 2(1-\varepsilon)\big[\lambda\tilde{\Phi}(\lambda) - \phi(\lambda)\big] + 2\lambda\varepsilon, \tag{44}$$

$$H''(\lambda) = 2(1-\varepsilon)\tilde{\Phi}(\lambda) + 2\varepsilon, \tag{45}$$

$$\begin{aligned} \lambda\tilde{\Phi}(\lambda) - \phi(\lambda) &= -\lambda\int_\lambda^\infty x^{-2}\phi(x)\,dx \\ &= \lambda^{-2}\phi(\lambda)\big[1 + O(\lambda^{-2})\big], \end{aligned} \tag{46}$$

$$\begin{aligned} (\lambda^2+1)\tilde{\Phi}(\lambda) - \lambda\phi(\lambda) &= \int_\lambda^\infty x^{-2}(x^2 - \lambda^2)\phi(x)\,dx \\ &= 2\lambda^{-3}\phi(\lambda)\big[1 + O(\lambda^{-2})\big]. \end{aligned} \tag{47}$$

To locate a zero of (44) approximately, set $\lambda^2(a) = n_0^2 - 3\log a n_0^2$, where $n_0^2 = 2\log\varepsilon^{-1}$. The choices $a_0 = (2\pi)^{1/3}$ and $a_1 = a_0(1 - cn_0^{-2}\log n_0^2)$, for $c > 0$

large, lead to

$$H'\big(\lambda(a_0)\big) \sim -9\varepsilon\lambda(a_0)n_0^{-2}\log a_0 n_0^2 < 0,$$
$$H'\big(\lambda(a_1)\big) \sim (c-9)\varepsilon n_0^{-2}\log a_0 n_0^2 > 0.$$

From (45), H is convex, and so λ_ε is bracketed between $\lambda(a_0)$ and $\lambda(a_1)$. Since $\lambda^2(a_1) - \lambda^2(a_0) = 3\log(a_0/a_1) = O(n_0^{-2}\log n_0)$, (16) follows. From (45), H is convex, and since $H'(\lambda(a_0)) < 0$,

$$\begin{aligned} 0 \le H\big(\lambda(a_0)\big) - H(\lambda_\varepsilon) &\le \big(\lambda(a_0) - \lambda_\varepsilon\big)H'\big(\lambda(a_0)\big) \\ &\le |\lambda(a_0) - \lambda(a_1)|\ |H'(\lambda(a_0))| \\ &= O\big(\varepsilon n_0^{-4}\log^2 n_0\big). \end{aligned} \tag{48}$$

Finally, using (34) and (47) and setting $\lambda_1 = \lambda(a_0)$,

$$\begin{aligned} H(\lambda_1) &= (1-\varepsilon)\Big\{2\lambda_1^{-3}\phi(\lambda_1)\big[1 + O\big(\lambda_1^{-2}\big)\big]\Big\} + \varepsilon\big(1+\lambda_1^2\big) \\ &= \varepsilon\big[\lambda_1^2 + 3 + 9n_0^{-2}\log(2\pi)^{1/3}n_0^2 + O\big(\lambda_1^{-2}\big)\big]. \end{aligned}$$

In view of error bound (48), this establishes (17) and completes the proof of Theorem 2. In fact, the bound (48) suggests that (17) could easily be improved by adding the next term in expansion (47).

Acknowledgments. My thanks go to David Donoho and Boris Levit for some helpful conversations and to the referees and an Associate Editor for their suggestions, which improved an earlier version of the manuscript.

REFERENCES

BERGER, J. (1982). Estimation in continuous exponential families: Bayesian estimation subject to risk restrictions and inadmissibility results. In *Statistical Decision Theory and Related Topics III*, Vol. I (S. S. Gupta and J. Berger, eds.) 109–141. Academic, New York.

BICKEL, P. J. (1983). Minimax estimation of the mean of a normal distribution subject to doing well at a point. In *Recent Advances in Statistics: Papers in Honor of Herman Chernoff on His Sixtieth Birthday* (M. H. Rizvi, J. S. Rustagi and D. Siegmund, eds.) 511–528. Academic, New York.

BICKEL P. J. (1984). Parametric robustness: Small biases can be worthwhile. *Ann. Statist.* **12** 864–879.

BROWN, L. D. (1971). Admissible estimators, recurrent diffusions and insoluble boundary value problems. *Ann. Math. Statist.* **42** 855–903. (Correction: *Ann. Statist.* **1** 594–596.)

CASELLA, G. and STRAWDERMAN, W. E. (1981). Estimating a bounded normal mean. *Ann. Statist.* **9** 870–878.

DONOHO, D. L. (1992). De-noising via soft-thresholding. Technical report, Dept. Statistics, Stanford Univ.

DONOHO, D. L. and JOHNSTONE, I. M. (1994a). Minimax risk over l_p balls for l_q error. *Probab. Theory Related Fields*. To appear.

DONOHO, D. L. and JOHNSTONE, I. M. (1994b). Minimax estimation via wavelet shrinkage. Technical report, Dept. Statistics, Stanford Univ.

DONOHO, D. L. and JOHNSTONE, I. M. (1994c). Ideal spatial adaptation via wavelet shrinkage. *Biometrika*. To appear.

DONOHO, D. L., JOHNSTONE, I. M., HOCH, J. C. and STERN, A. S. (1992). Maximum entropy and the nearly black object (with discussion). *J. Roy. Statist. Soc. Ser. B* **54** 41–81.

DONOHO, D. L., JOHNSTONE, I. M., KERKYACHARIAN, G. and PICARD, D. (1994). Wavelet shrinkage: Asymptopia? (with discussion). *J. Roy. Statist. Soc. Ser. B*, to appear.

DONOHO, D. L., LIU, R. C. and MAC GIBBON, K. B. (1990). Minimax risk over hyperrectangles, and implications. *Ann. Statist.* **18** 1416–1437.

EFRON, B. and MORRIS, C. (1971). Limiting the risk of Bayes and empirical Bayes estimates: Part I. The Bayes case. *J. Amer. Statist. Assoc.* **66** 807–815.

FELDMAN, I. (1991). Constrained minimax estimation of the mean of the normal distribution. *Ann. Statist.* **19** 2259–2265.

FRIEDEN, B. R. (1972). Restoring with maximum entropy II. Superresolution of photographs with diffraction-blurred impulses. *J. Opt. Soc. Amer.* **62** 1202–1210.

GULL, S. F. and DANIELL G. J. (1978). Image reconstruction from incomplete and noisy data. *Nature* **272** 686–690.

HODGES, J. L., JR. and LEHMANN, E. L. (1952). The use of previous experience in reaching statistical decisions. *Ann. Math. Statist.* **23** 396–407.

HUBER, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.* **35** 73–101.

HUBER, P. J. (1981). *Robust Statistics.* Wiley, New York.

JOHNSTONE, I. M. (1994). Minimax Bayes, asymptotic minimax and sparse wavelet priors. In *Statistical Decision Theory and Related Topics V* (S. S. Gupta and J. O. Berger, eds.) 303–326. Springer, New York.

LEVIT, B. YA. (1987). On bounds for the minimax risk. In *Probability Theory and Mathematical Statistics* (Yu. V. Prohorov, V. A. Statulevicius, V. V. Sazonov and B. Grigelionis, eds.) **2** 203–216. VNU Science Press, The Netherlands.

MALLOWS, C. L. (1978). Problem 78-4. *SIAM Rev.* **20** 183.

MARAZZI, A. (1980). Robust Bayesian estimation for the linear model. Technical Report 27, E.T.H. Zürich.

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

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305-4065