Minimax Risk over l_p-Balls for l_q-error

David L. Donoho
Iain M. Johnstone
Department of Statistics
Stanford University
Stanford, CA 94305

July 20, 1994

Abstract

Consider estimating the mean vector θ from data $N_n(\theta, \sigma^2 I)$ with l_q norm loss, $q \geq 1$, when θ is known to lie in an n-dimensional l_p ball, $p \in (0, \infty)$. For large n, the ratio of minimax *linear* risk to minimax risk can be *arbitrarily large* if $p < q$. Obvious exceptions aside, the limiting ratio equals 1 only if $p = q = 2$. Our arguments are mostly indirect, involving a reduction to a univariate Bayes minimax problem. When $p < q$, simple non-linear co-ordinatewise threshold rules are asymptotically minimax at small signal-to-noise ratios, and within a bounded factor of asymptotic minimaxity in general. Our results are basic to a theory of estimation in Besov spaces using wavelet bases (to appear elsewhere).

Key Words. Minimax Decision Theory. Minimax Bayes Estimation. Fisher Information. Non-linear estimation. White noise model. Loss convexity. Estimating a bounded normal mean.

Running Title: Minimax risk over l_p-balls.

AMS 1980 Subject Classification (1985 Rev): Primary: 62C20, Secondary: 62F12, 62G20

1 Introduction: l_2 error

Suppose we observe $y = (y_i)_{i=1}^n$ with $y_i = \theta_i + z_i$, z_i i.i.d. $N(0, \sigma^2)$, with $\theta = (\theta_i)_{i=1}^n$ an unknown element of the convex set Θ. Sacks and Strawderman (1982) showed that, in some cases, the minimax linear estimator of a linear functional $L(\theta)$ could be improved on by a nonlinear estimator. Specifically, they showed that for squared error loss, the ratio R_L^*/R_N^* of minimax risk among linear estimates to minimax risk among all estimates exceeded $1+\epsilon$ for some (unknown) $\epsilon > 0$ depending on the problem. This raised the possibility that nonlinear estimators could dramatically improve on linear estimators in some cases.

However, Ibragimov and Hasminskii (1984) established a certain limitation on this possibility by showing that there is a positive finite constant bounding the ratio R_L^*/R_N^* for any problem where Θ is symmetric and convex. Donoho, Liu, and MacGibbon (1990) have shown that the Ibragimov-Hasminskii constant is not larger than 5/4. Moreover, Donoho and Liu (1988) have shown that even if Θ is convex but asymmetric, still $R_L^*/R_N^* < 5/4$ – provided inhomogeneous linear estimators are allowed. It follows that for estimating a single linear functional, minimax linear estimates cannot be *dramatically* improved on in the worst case.

For the problem of estimating the whole object θ, with squared l_2-loss $\| \hat{\theta} - \theta \|^2 = \sum(\hat{\theta}_i - \theta_i)^2$, one could ask again whether linear estimates are nearly minimax. Pinsker (1980) discovered that if Θ is an ellipsoid, then $R_L^*/R_N^* \to 1$ as $n \to \infty$. Donoho, Liu, and MacGibbon(1990) showed that if Θ is an l_p-body with $p \geq 2$ then $R_L^*/R_N^* \leq 5/4$, nonasymptotically. Thus there are again certain limits on the extent to which nonlinear estimates can improve on linear ones in the worst case.

However, these limits are less universal in the case of estimating the whole object than they are in the case of estimating a single linear functional. In this paper we show that *there are cases where the ratio R_L^*/R_N^* may be arbitrarily large.* Let $\Theta_{p,n}$ denote the standard n-dimensional unit ball of l_p, i.e. $\Theta_{p,n} = \{\theta : \sum_1^n |\theta_i|^p \leq 1\}$.

Theorem 1 *Let $n\sigma^2 =$ constant and $\Theta = \Theta_{p,n}$. Then as $n \to \infty$*

$$\frac{R_L^*}{R_N^*} \to \begin{cases} 1 & p \geq 2 \\ \infty & p < 2 \end{cases} \tag{1}$$

This reflects the phenomenon that in some function estimation problems of a linear nature, the optimal rate of convergence over certain convex function classes is not attained by any linear estimate (Kernel, Spline, ...). Compare also Sections 7, 8, and 9 in Donoho, Liu, and MacGibbon (1990), and the discussion in section 3 below.

Our technique sheds some light on Pinsker's phenomenon, as mentioned above. It establishes

Theorem 2 *Let p be fixed, and set $\Theta = \Theta_{p,n}$. Suppose that we can choose $\sigma^2 = \sigma^2(n)$ in such a way that $R_L^*/R_N^* \to 1$. There are 3 possibilities:*

a. $R_N^*/n\sigma^2 \to 1$ (Classical case).

b. $p = 2$ (Pinsker's case).

c. $R_L^*/n\sigma^2 \to 0$ *(trivial case).*

In words, if the minimax linear estimator is nearly minimax, then: either (case a) the raw data y is nearly minimax, or (case c) the trivial estimator 0 is nearly minimax, or else we are in the case $p = 2$ covered by Pinsker (1980). Put differently, Pinsker's phenomenon happens among l_p constraints only if $p = 2$.

Theorems 1 and 2 show that improvement on minimax linear estimation is possible without showing how (or by how much). A heuristic argument suggests that a non-linear estimator that is near optimal has the form

$$\hat{\theta}_i = sgn(y_i)(|y_i| - \lambda\sigma)_+ \tag{2}$$

where $\lambda = \lambda(n, \sigma, p)$. Consider, for example, the case $p = 1$ and $\sigma = cn^{-1/2}$. Then, *on average* $|\theta_i| \le n^{-1}$. Therefore most of the coordinates θ_i are of order n^{-1} in magnitude. But for Theorem 1, $\sigma = O(n^{-1/2})$. The nonlinear estimator with $\lambda = 5 \cdot \sigma$ will be wrong in most coordinates by only $O_P(n^{-1})$, and the case of the few others by only $O_P(n^{-1/2})$. As the minimax linear estimator is wrong in every coordinate by $O_P(n^{-1/2})$, the result (1) for $p < 2$ might not be surprising.

In fact, for an appropriate choice of λ, the estimator (2) is asymptotically minimax, and the improvement R_N^*/R_L^* can be calculated.

Theorem 3 *Assume* $0 < p < 2$, *with* $n\sigma^p \to \infty$ *and* $\sigma^2 \log n\sigma^p \to 0$. *Let* $\lambda^2 = 2 \log n\sigma^p$. *Then (i)*

$$R_N^* = \sup_{\theta \in \Theta_{p,n}} E_\theta \sum_1^n (\hat{\theta}_{i,\lambda} - \theta)^2 (1 + o(1)) \tag{3}$$

$$= \sigma^{2-p} (2 \log n\sigma^p)^{1-p/2} (1 + o(1)) \quad \text{as} \quad n\sigma^p \to \infty. \tag{4}$$

(ii) $\quad R_L^*/R_N^* = (1 + n\sigma^2)^{-1} n\sigma^p (2 \log n\sigma^p)^{-1+p/2} \quad \text{as} \quad n\sigma^p \to \infty$

Suppose for example that $\sigma = n^{-1/2}$. Then $n\sigma^p = n^{1-p/2}$, $R_L^* = 1/2$ and

$$R_N^* \sim (\frac{(2-p)\log n}{n})^{1-p/2}$$

Thus, in the special case $p = 1$, $R_N^* \sim (\log n/n)^{1/2}$.

The estimator (2) can be said to use *soft thresholding* , since it is continuous in y. An alternative *hard threshold* estimator is

$$\theta_{\lambda,i}^{(h)} = y_i I\{|y_i| > \lambda\}. \tag{5}$$

(When needed, we use the notation $\theta_\lambda^{(s)}$ to distinguish soft threshold estimators.)

Corollary 4 *Results (3), (4) hold also for hard threshold estimators so long as* $\lambda^2 = 2 \log n\sigma^p + \alpha \log(2 \log n\sigma^p)$ *for some* $\alpha > p - 1$.

These results will be derived as special cases of those for l_q losses, to which we now turn.

2 Results for l_q loss functions

The general situation we study has, as before, $y \sim N_n(\theta, \sigma^2 I)$, but with estimators evaluated according to l_q-loss $||\hat{\theta} - \theta||_q^q = \sum_1^n |\hat{\theta}_i - \theta_i|^q$. We need convexity of the loss function, and so require that $q \geq 1$. Thus the class of possible 'shapes' (p, q) for parameter space and loss function is given by $S = (0, \infty) \times [1, \infty)$. In applications, interest usually centers on p or $q = 1$, 2, or ∞, but for the theory it is instructive to study also intermediate cases. This is especially true here as we do not explicitly allow p or $q = \infty$ (though $p = \infty$ is an easy extension).

In addition, it is natural (and important for the applications in Donoho and Johnstone (1992)) to allow balls of arbitrary radius: $\Theta_{p,n}(r) = \{\theta : \sum |\theta_i|^p \leq r^p\}$. Consider therefore the minimax risk

$$R_N^* = R_{N,q}^*(\sigma; \Theta_{p,n}(r)) = \inf_{\hat{\theta}} \sup_{\Theta_{p,n}(r)} E_\theta \sum_1^n |\hat{\theta}_i - \theta_i|^q. \tag{6}$$

The subscript 'N' indicates that non-linear procedures $\hat{\theta}(y)$ are allowed in the infinum.

Our object is to study the asymptotic behavior of R_N^* as n, the number of unknown parameters, increases. We regard the noise level $\sigma = \sigma(n)$ and ball radius $r = r(n)$ as functions of n. This framework accommodates a common feature of statistical practice: as the amount of data increases (here thought of as a decreasing noise level σ per parameter), so too does the number of parameters that one may contemplate estimating.

If there were no prior constraints, $\Theta = R^n$, then the unmodified raw data would give a minimax estimator $\hat{\theta}(y) = y$. In that case, the unconstrained minimax risk equals $E_\theta |Y - \theta|^q = n\sigma^q c_q$, where $c_q = E|Z|^q = 2^{q/2}\pi^{-1/2}\Gamma((q+1)/2)$, and Z denotes a single standard Gaussian deviate.

Asymptotically, R_N^* depends on the size of $\Theta_{p,n}(r)$ through the dimension-normalized radius $\eta_n = n^{-1/p}(r/\sigma)$. This may be interpreted as the maximum scalar multiple in standard deviation units of the vector $(1, \ldots, 1)$ that lies within $\Theta_{p,n}(r)$. Alternatively, it can be thought of as the average signal to noise ratio measured in the l_p- norm: $(n^{-1}\sum(\theta_i/\sigma)^p)^{1/p} \leq n^{-1/p}(r/\sigma)$.

The asymptotic behavior of R_N^* will be expressed in terms of a standard univariate Gaussian location problem in which we observe $X \sim N(\mu, 1)$ and seek to estimate μ using loss function $|\delta(x) - \mu|^q$. Write $\delta_F(x)$ for the Bayes estimator corresponding to a prior distribution $F(d\mu)$, and $\rho_q(F) = \inf_{\delta(x)} \int E_\mu |\delta(x) - \mu|^q F(d\mu)$ for the Bayes risk. Let $\mathcal{F}_p(\eta)$ denote the class of probability measures $F(d\mu)$ satisfying the moment condition $\int |\mu|^p F(d\mu) \leq \eta^p$. An important role is played by the largest Bayes risk over $\mathcal{F}_p$

$$\rho_{p,q}(\eta) = \sup_{\mathcal{F}_p} \rho_q(F). \tag{7}$$

A distribution $F_{p,q} = F_{p,q}(\eta)$ maximising (7) will be called *least favorable.* Usually the least favorable distribution $F_{p,q}(\eta)$ cannot be described analytically, but when $\eta_n \to 0$, it is sometimes possible to find an *asymptotically least favorable* sequence of simple structure $\tilde{F}_{p,q,n} \in \mathcal{F}_p(\eta_n)$ such that $\rho_q(\tilde{F}_{p,q,n}) \sim \rho_{p,q}(\eta_n)$. We use $\hat{\theta}_N(y)$ to denote *an* asymptotically minimax rule, which of course need not be unique.

Theorem 5 *Let $(p,q) \in (0,\infty) \times [1,\infty)$. If either (i) $p \geq q$ or (ii) $0 < p < q$ and $(\sigma/r)^2 \log n(\sigma/r)^p \to 0$, then*

$$R_N^* \sim n\sigma^q \rho_{p,q}(\eta_n) \qquad \text{as } n \to \infty. \tag{8}$$

In specific instances, more can be said:

1. $\eta_n \to \infty$. $\quad R_N^* \sim n\sigma^q c_q, \qquad \hat{\theta}_N(y) = y.$

2. $\eta_n \to \eta \in (0,\infty)$. $\quad R_N^* \sim n\sigma^q \rho_{p,q}(\eta) \qquad \hat{\theta}_{N,i}(y) = \sigma \delta_{F_{p,q}}(\sigma^{-1} y_i).$

3a. $\eta_n \to 0, p \geq q$. $\quad R_N^* \sim n\sigma^q \eta_n^q, \qquad \hat{\theta}_N(y) = 0.$
The two point distributions $\tilde{F}_n = (\nu_{-\eta_n} + \nu_{\eta_n})/2$ are asymptotically least favorable.

3b. $\eta_n \to 0, p < q$, *and assume also that* $(\sigma/r)^2 \log n(\sigma/r)^p \to 0$. *Let* $\lambda_n^2 = 2\log n(\sigma/r)^p$.

$$\begin{aligned} R_N^* \quad &\sim n\sigma^q \eta_n^p (2\log \eta_n^{-p})^{(q-p)/2}, \quad \hat{\theta}_{N,i}(y) = \mathrm{sgn} y_i (|y_i| - \lambda_n \sigma/r)_+ \\ &\sim (2(\sigma/r)^2 \log n(\sigma/r)^p)^{(q-p)} \end{aligned} \tag{9}$$

The three point distributions $\tilde{F}_n = (1-\epsilon)\nu_0 + \epsilon(\nu_\mu + \nu_{-\mu})/2$ are asymptotically least favorable, where $\epsilon = \epsilon_n, \mu = \mu_n \sim (2\log \epsilon_n^{-1})^{1/2}$ are determined from the equations

$$\epsilon \mu^p = \eta_n^p \qquad \text{and} \qquad \phi(a_n + \mu) = \epsilon \phi(a_n) \tag{10}$$

where $a_n = a(\eta_n) \uparrow \infty$ but $a_n^2 = o(\log \eta_n^{-p})$.

The proof of Theorem 5 is the subject of Sections 4 through 7. The asymptotically minimax estimators given in (9) are the same as in (2) except that now the choice of the threshold parameter is specified: it is noteworthy that this does not depend on the loss function. Another sequence of asymptotically minimax estimators in this case would, of course, be the Bayes estimators corresponding to an asymptotically least favorable sequence of distributions. In a sense made more precise in Section 6, these Bayes estimators approximately have the form $\delta_{\tilde{F}_n}(x) \doteq \mu_n \mathrm{sgn}(x) I\{|x| > \mu_n + a_n\}$. Since $a_n = o(\mu_n)$ and $\mu_n^2 \sim \lambda_n^2 \sim 2\log \eta_n^{-p}$ it follows that $\hat{\theta}_{N,i}(y) = \sigma \delta_{\tilde{F}_n}(\sigma^{-1} y_i)$ has approximately the same zero set as the simpler threshold rule $\hat{\theta}_{N,i}(y)$. Hard threshold rules of the form (5) are also asymptotically minimax in the setting *3.(b)* of Theorem 5, so long as λ^2 is chosen equal to $2\log n\sigma^p + \alpha \log(2\log n\sigma^p)$ for $\alpha > p - 1$.

The threshold estimators of the previous section have a more general asymptotic near-optimality property that holds whenever (8) is valid.

Theorem 6 *Let $(p,q) \in (0,\infty) \times [1,\infty)$. There exist constants $\Lambda_s(p,q), \Lambda_h(p,q) \in (1,\infty)$ such that if either (i) $p \geq q$ or (ii) $0 < p < q$ and $(\sigma/r)^2 \log n(\sigma/r)^p \to 0$, then*

$$\inf_{\lambda} \sup_{\Theta_{p,n}(r)} E_\theta ||\theta_\lambda^{(s)} - \theta||_q^q \leq \Lambda_s(p,q) R_N^*(\sigma, \Theta_{p,n}(r))(1 + o(1))$$

and the corresponding property holds for $\theta_\lambda^{(h)}$ (with bound $\Lambda_h(p,q)$).

The theorem is proved in Section 8, where definitions of $\Lambda(p,q)$ are given in terms of a univariate Bayes minimax estimation problem. In fact $\Lambda_s(p,2)$ and $\Lambda_h(p,2)$ are both smaller than 2.22 for all $p \geq 2$ and computational experiments indicate that $\Lambda_s(1,2) \leq 1.6$.

We turn now to the minimax *linear* risk $R_L^* = R_{L,q}^*(\sigma;\Theta_{p,n}(r))$, obtained by restricting attention to estimators that are linear in the data y. Because of the symmetry of Θ, this effectively means estimators of the form $\hat{\theta}(y) = ay$ for $a \in [0,1]$, or equivalently, of the form $y/(1+b)$ for $b \in [0,\infty]$.

Call a set Θ *loss-convex* if the set $\{(\theta_i^q); \theta \in \Theta\}$ is convex (cf. the notion of q-convexity in Lindenstrauss and Tzafiri (1979)). Clearly $\Theta_{p,n}$ is loss-convex exactly when $p \geq q$. If $p < q$ then the *loss-convexification* of $\Theta_{p,n}$, namely the smallest loss-convex set containing $\Theta_{p,n}$, is $\Theta_{q,n}$. The size of the loss-convexification of $\Theta_{p,n}$ turns out to determine minimax linear risk, and so in analogy with η_n we define $\bar{\eta}_n = n^{-1/p\vee q}(r/\sigma)$. Finally, we use $\hat{\theta}_L(y)$ to denote *an* asymptotically minimax linear rule, again not necessarily unique.

Theorem 7 *Let $(p,q) \in S = (0,\infty) \times [1,\infty)$. The limiting behavior of R_L^* depends on that of $\bar{\eta}_n = n^{-1/p\vee q}(r/\sigma)$ as follows.*

1. $\bar{\eta}_n \to \infty$. $\qquad R_L^* \sim n\sigma^q c_q, \qquad \hat{\theta}_{L,i}(y) = y_i.$

2. $\bar{\eta}_n \to \eta \in (0,\infty)$. $\qquad R_L^* \sim n\sigma^q c_{p,q}(\eta). \qquad \hat{\theta}_{L,i}(y) = a_* y_i.$

   (a) *If $p < q$ or $p = q \leq 2$, then*

   $$c_{p,q}(\eta) = \begin{cases} c_1 \wedge \eta & a_* = I\{\eta > c_1\} \quad q = 1 \\ c_q[1+b_*]^{-(q-1)} & b_* = c_q^{1/(q-1)}\eta^{-q'} \quad q > 1 \end{cases}$$

   where $(1+b_*)^{-1} = a_*$ and $1/q' + 1/q = 1$.

   (b) *If $p > q$ or $p = q \geq 2$, then*

   $$c_{p,q}(\eta) = \inf_{b \geq 0}(1+b)^{-q}\tilde{s}(b^p\eta^p),$$

   where $\tilde{s}(\gamma)$ is the least concave majorant of $s(\gamma) = E|Z + \gamma^{1/p}|^q$ on $[0,\infty)$. If b_* attains the minimum in $c_{p,q}(\eta)$, then $\hat{\theta}_{L,i}(y) = y_i/(1+b_*)$.

3. $\bar{\eta}_n \to 0$. $\qquad R_L^* \sim n\sigma^q\bar{\eta}_n^q = r^q n^{(1-q/p)_+}, \qquad \hat{\theta}_{L,i}(y) = 0.$

The proof appears in Section 9. The following corollary describes the possible limiting behaviors for R_L^*/R_N^*. Note that by passing to subsequences, we may always assume that η_n converges.

Corollary 8 *Suppose $(p,q) \in S$ and $\eta_n \to \eta \in [0,\infty]$. Then*

$$lim\frac{R_L^*}{R_N^*} = \begin{cases} 1 & \text{if } (i)\eta_n \to \infty, (ii)\eta_n \to 0, p \geq q, \text{ or } (iii)p = q = 2 \\ \in (1,\infty) & \text{if } \eta_n \to \eta \in (0,\infty),\ p,q \text{ not both equal to 2.} \\ \infty & \text{if } \eta_n \to 0, p < q \text{ and } (\sigma/r)^2 \log n(\sigma/r)^p \to \infty. \end{cases}$$

Thus, among l_p ball constraints and l_q losses, *exact* asymptotic optimality of linear estimators occurs in "non-trivial" cases only for Euclidean norm constraints and squared error loss. If Θ is loss-convex, the inefficiency of linear estimates is always bounded. If Θ is not loss-convex, and is asymptotically 'small' ($\eta_n \to 0$), then the inefficiency becomes infinite at a rate which can be explicitly read off from Theorems 5 and 7.

In summary, if Θ is large ($\eta_n \approx \infty$), then the prior information conferred by restriction to Θ is weak and the raw data is nearly minimax. On the other hand, if Θ is small ($\eta_n \approx 0$), then prior information is strong, but it is the *shape* of Θ that is decisive: if Θ is loss convex, then the trivial zero estimator is near minimax, whereas in the non loss convex cases, threshold rules successfully capture the few non-zero parameters and are near minimax. In the intermediate Θ cases ($\eta_n \approx \eta > 0$), one might say that prior information is partially decisive: linear rules are *rate* optimal, but *not* efficient, except for the isolated (but important!) case of Hilbertian norms on parameter space and loss function.

3 Discussion and Remarks

1. Constraints on the l_p norm of θ can arise naturally in various scientific contexts. Hypercube constraints ($a \leq \theta_i \leq b$) correspond to *a priori* pointwise bounds; l_2 constraints to energy bounds, and l_1 constraints to bounds on distribution of total mass. As $p \to 0$, the l_p balls become cusp-like; so that only a small number of components can be significantly non-zero. Formally

$$\lim_{p \to o} \Theta_{p,n}((n\epsilon)^{1/p}) = \Theta_{n,0}(\epsilon) = \{\theta : n^{-1} \sum I\{\theta_i \neq 0\} \leq \epsilon\}.$$

The latter "nearly-black" condition has been used by Donoho *et. al.* (1990) to study behavior of non-linear estimation rules such as maximum entropy, using the methods of this paper.

2. In addition to works already mentioned, there exist a number of papers that exhibit function classes over which non-linear estimators have dramatically better performance in the sense of worst case risk than the best linear estimators. Nemirovskii *et. al.* (1985) show that non-parametric M-estimates (including constrained least squares) achieve faster fates of mean-squared error convergence than best linear over function classes described by monotonicity or total-variation constraints (for which $p = 1$), or more generally, over norm-bounded sets in Sobolev spaces W_p^k for $1 \leq p < 2$. Related results are given by van de Geer (1990) and Birgé and Massart (1990).

This paper's consideration of such highly symmetric parameter spaces and Gaussian white noise amounts to imposing very restrictive assumptions. However this symmetry permits reduction to simple one-dimensional estimation problems and thus avoids the appeal to approximation theoretic properties of function classes that is useful in treating problems of more direct practical relevance. (see, for example, Ibragimov and Hasminskii, (1990), van de Geer (1990) and Donoho (1990).) The basic dichotomy between $p < q$ and $p \geq q$, as expressed in the loss-convexity condition, appears already in our very simple setting.

3. It turns out, however, that the idealised considerations of this paper can be made the basis of a discussion of estimation over a wide class of function spaces, namely the Besov family. This allows one to treat the familiar Hölder and Hilbertian Sobolev spaces in addition to other classes of scientific relevance, such as bounded total variation and the "bump algebra". On these latter spaces, non-linear methods and local bandwidth adaptivity are essential for optimal minimax estimation. The connection comes via orthonormal bases of compactly supported wavelets (e.g. Meyer, 1990, Daubechies, 1988), which permit an identification, in an appropriate sense, of estimation over Besov spaces with estimation over sequence spaces. The relevant least-favorable subsets in sequence space are given by cartesian products of l_p-balls corresponding to the various resolution levels of the wavelet expansion. A more complete account appears in Donoho and Johnstone (1992).

4 A Bayes-minimax Approximation

A standard way to study the minimax risk R_N^* is to use Bayes rules. By usual arguments based on the minimax theorem, $R_N^* = \sup_{\pi \in \Pi} \rho(\pi)$, where $\rho(\pi)$ denotes the Bayes risk $E_\pi E_\theta ||\hat{\theta}_\pi - \theta||_q^q$, with θ random, $\theta \sim \pi$; $\hat{\theta}_\pi$ denotes the Bayes estimator corresponding to prior π and l_q loss, and Π denotes the set of all priors supported on Θ.

To obtain an approximation to R_N^* with simpler structure, consider a Bayes-minimax problem in which θ is a random variable that is only required to belong to Θ *on average*. Define

$$R_B^*(\sigma, \Theta_{n,p}(r)) = \inf_{\hat{\theta}} \sup_{\pi} \left\{ E_\pi E_\theta ||\hat{\theta} - \theta||_q^q, \text{ for } \pi : E_\pi \sum_1^n |\theta_i|^p \le r^p \right\}.$$

Since degenerate prior distributions concentrated at points $\theta \in \Theta_{n,p}(r)$ trivially satisfy the moment constraint, the Bayes-minimax risk is an upper bound for the non-linear minimax risk

$$R_N^* \le R_B^*.$$

The moment constraint depends on π only through its univariate marginal distributions π_i. If $\hat{\theta}$ is a *co-ordinatewise* estimator, that is, one for which $\hat{\theta}_i$ depends only on y_i, then the integrated risk $E_\pi E_\theta ||\hat{\theta} - \theta||_q^q$ depends on π only through the marginals π_i. This leads to a description of R_B^* in terms of a simpler *univariate* Bayes-minimax estimation problem and will be the subject of this section.

In view of the permutation invariance of the problem, it is enough to use estimators $\delta^n(y) = (\delta(y_1), \ldots, \delta(y_n))$ constructed from a single univariate estimator δ. From the co-ordinatewise nature of δ^n, and the i.i.d. structure of the errors $\{z_i\}$,

$$\begin{aligned} E_\pi E_\theta ||\delta^n - \theta||_q^q &= \sum_i \int E_{\theta_i} |\delta(y_i) - \theta_i|^q \pi_i(d\theta_i) \\ &= \int E_{\theta_1} |\delta(y_1) - \theta_1|^q \left(\sum \pi_i \right) (d\theta_1) \\ &= n E_{F_\pi} E_{\theta_1} |\delta(y_1) - \theta_1|^q \end{aligned} \tag{11}$$

where $F_\pi(d\theta_1) = n^{-1} \sum \pi_i(d\theta_1)$ is a univariate prior. The moment condition on π can also be expressed in terms of F_π, since

$$E_\pi \sum_i |\theta_i|^p = \sum_i \int |\theta_i|^p \pi_i(d\theta_i) = n \int |\theta_1|^p F_\pi(d\theta_1). \tag{12}$$

Thus $E_{F_\pi} |\theta_1|^p \le n^{-1} r^p$. Define a univariate Bayes-minimax problem with p^{th} moment constraint τ and noise level σ:

$$\rho_{p,q}(\tau, \sigma) = \inf_{\delta} \sup_{F} \{ E_F E_{\theta_1} |\delta(y_1) - \theta_1|^q : E_F |\theta_1|^p \le \tau^p \}. \tag{13}$$

The dependence on p and q will usually not be shown explicitly.

Proposition 9 $\quad R_B^*(\sigma, \Theta_{n,p}(r)) = n \rho(r n^{-1/p}, \sigma).$

Proof. This is easily completed from (11) and (12). Indeed, let (F^0, δ^0) be a saddlepoint for the univariate problem (13): that is, δ^0 is a minimax rule, F^0 is a least favorable prior distribution and δ^0 is Bayes for F^0. (Existence is discussed in Section 4.1 below.) Let F^{0n} denote the n-fold cartesian product measure derived from F^0: from (12) and (11), it satisfies the moment constraint for R_B^*, and

$$E_{F^{0n}} E_\theta ||\delta^{0n} - \theta||_q^q = n\rho(rn^{-1/p}, \sigma).$$

To establish the Proposition, it is enough to verify that (F^{0n}, δ^{0n}) is a saddlepoint for R_B^*, which would follow from the inequality

$$E_\pi E_\theta ||\delta^{0n} - \theta||_q^q \le E_{F^{0n}} E_\theta ||\delta^{0n} - \theta||_q^q.$$

But (11) and (12) reduce this to the saddlepoint property of (F^0, δ^0). ∎

4.1 Properties of the univariate Bayes minimax problem

We focus now on the univariate Gaussian location problem implicit in (13) in which we observe $X \sim N(\mu, \sigma^2)$. Assume that μ is a random variable with distribution belonging to $\mathcal{F}_p(\tau)$, the collection of distributions F on R satisfying $\int |\mu|^p F(d\mu) \le \tau^p$.

First some preliminary observations. For any given $q \ge 1$ and prior distribution $F(d\mu)$, the Bayes estimator $\mu_F(x)$ is uniquely determined for Lebesgue-almost-all x. This follows easily from strict convexity of the loss function when $q > 1$ and also, with some extra argument, when $q = 1$ [1]. (Where indicated by footnotes, extra details are given in the Appendix.) The Bayes risk function, $F \to \rho_q(F)$ is concave and weakly upper semicontinuous, and hence attains a maximum on the weakly compact set $\mathcal{F}_p(\eta)$. [We conjecture that this maximum is unique, admittedly only with the case $q = 2$ discussed below as direct support.] If $F_a(d\mu) = F(a^{-1}d\mu)$, then $\rho_q(F_a) \le a^q \rho_q(F)$ for $a >= 1$ [2] .

Let us summarize now some properties of the minimax Bayes risk (13). Setting $\rho(F, \delta) = E_F E_\mu |\delta(x) - \mu|^q$, we have

$$\rho(\tau, \sigma) = \inf_\delta \sup_{\mathcal{F}_p(\tau)} \rho(F, \delta).$$

The minimax theorem guarantees that

$$\rho(\tau, \sigma) = \sup_{\mathcal{F}_p(\tau)} \rho(F) = \sup_{\mathcal{F}_p(\tau)} \inf_\delta \rho(F, \delta) \tag{14}$$

and the existence of a minimax rule δ_τ such that

$$\sup_F \rho(F, \delta_\tau) = \rho(\tau, \sigma).$$

Let F_τ be a distribution maximizing $\rho(F)$ over $\mathcal{F}_p(\tau)$. Since $\rho(F_\tau, \delta_\tau) \le \rho(\tau, \sigma) = \rho(F_\tau, \delta_{F_\tau})$, it follows from the essential uniqueness of Bayes rules that $\delta_\tau = \delta_{F_\tau}$ and hence that δ_τ is Bayes for the least favorable distribution F_τ. The pair (F_τ, δ_τ) is thus a saddlepoint for $\rho(F, \delta)$.

Proposition 10 *The function $\rho(\tau, \sigma)$ satisfies the invariance*

$$\rho(\tau, \sigma) = \sigma^q \rho(\tau/\sigma, 1),$$

and the inequality

$$\rho(a\tau, \sigma) \leq a^q \rho(\tau, \sigma), \qquad a \geq 1.$$

It is continuous, monotone increasing in τ, concave in τ^p and has $\rho(\tau, \sigma) \to \sigma^q c_q$ as $\tau/\sigma \to \infty$.

Proof The invariance follows by a simple rescaling, and thus all remaining properties may be derived by considering the reduced function $\rho(\tau) = \rho(\tau, 1)$. The inequality follows from the corresponding inequality for F_a noted above. Monotonicity is clear from the definition. For concavity, set $t = \tau^p$, $\widetilde{\mathcal{F}}(t) = \{F : \int |\mu|^p \, dF \leq t\}$, and $\tilde{\rho}(t) = \rho(\tau)$. Since $\widetilde{\mathcal{F}}(t) = \mathcal{F}_p(\tau)$, (14) shows that $\tilde{\rho}(t) = \sup\{\rho(F) : F \in \widetilde{\mathcal{F}}_p(t)\}$. Concavity (and hence continuity) follows immediately, because $(1-\epsilon)F_1 + \epsilon F_2 \in \widetilde{\mathcal{F}}((1-\epsilon)t_1 + \epsilon t_2)$ whenever $F_i \in \widetilde{\mathcal{F}}_p(t_i)$ $i = 1, 2$.

To show that $\rho(\tau) \nearrow c_q$ as $\tau \nearrow \infty$, we note that appropriately scaled zero mean Gaussian priors satisfy the moment constraints, so that $\lim_{\tau \to \infty} \rho(\tau) \geq \lim_{\sigma \to \infty} \rho(\Phi_\sigma)$ where Φ_σ denotes the $N(0, \sigma^2)$ distribution. Since the posterior is also Gaussian $\delta_{\Phi_\sigma}(x) = \sigma^2 x/(\sigma^2+1)$ for all $q \geq 1$, and a simple calculation using the formulas (41) and (41) for linear rules in Section 9 shows that $\lim_\sigma \rho(\Phi_\sigma) = c_q$. ∎

Let $t = \tau^p$, $s = \sigma^p$. For use in Donoho and Johnstone (1992) we record here some properties of the function

$$r(t, s) = \sup\{E_F E_\mu |\delta_t - \mu|^q : F \text{ s.t. } E_F |\mu|^p \leq s\}.$$

Proposition 11 *a)* $r(t, t) = \rho(\tau, \sigma)$,

b) $s \to r(t, s)$ *is concave,*

c) $r_1(s, s) = 0$.

Part a) simply restates that δ_t is minimax for $\widetilde{\mathcal{F}}_p(t)$. Part b) is proved in exactly the same manner as for concavity of $t \to \tilde{\rho}(t)$. If we assume that $t \to r(t, s)$ is differentiable, part c) is a simple consequence of the fact that δ_s is minimax for $\widetilde{\mathcal{F}}_p(s)$, and hence $t = s$ minimises $t \to r(t, s)$.

4.2 Squared error loss and optimization of Fisher information

An identity of Brown (1971) connecting Bayes risk with Fisher information simplifies the study of the Bayes-minimax risk (13) for Gaussian data with unit variance under squared error loss, $q = 2$. Indeed, if $I(G) = \int (g'(x))^2/g(x)dx$ denotes the Fisher information for a distribution with absolutely continuous density g, then

$$\rho_2(F) = 1 - I(F * \Phi), \tag{15}$$

where Φ denotes the standard Gaussian distribution function. The results below are not strictly necessary for the development in this paper, being special cases of the previous work (except for Lemma 13 below). We include them because of the importance of squared error loss and the connections to other work that (15) establishes.

In view of (15), the Bayes-minimax risk $\rho_{p,2}(\tau, 1) = 1 - I_p(\tau)$, where

$$I_p(\tau) = \inf\{I(F * \Phi) : F \in \mathcal{F}_p(\tau)\}. \tag{16}$$

As I is lower-semicontinuous with respect to vague convergence of distributions (Port and Stone, 1974), and $\mathcal{F}_p(\tau)$ is tight – hence vaguely compact – the infimum in (4.1) is attained for every $p \in (0, \infty)$ and every $\tau \in (0, \infty)$. In fact as the density of $\Phi * F$ must be strictly positive on the whole real line, an argument of Huber (1964) (see also Huber (1974)) shows the solution to (16) is unique. Call the solution $F_{p,\tau}$.

The quantity I_p is probably new (but see also Feldman (1991)), but existing work does supply some information about it. Recall the well-known inequality $I(F)Var(F) \geq 1$, with equality only at the Gaussian. This implies that for $p = 2$ we have

$$I_2(\tau) = (1 + \tau^2)^{-1} \tag{17}$$

and that the solution $F_{2,\tau}$ is the Gaussian distribution $N(0, \tau^2)$. The limiting case $p \to \infty$ is of interest; we get

$$I_\infty(\tau) = \inf\{I(\Phi * F) : supp(F) \in [-\tau, \tau]\}$$

This has arisen before in the study of estimating a single bounded normal mean (Casella and Strawderman (1981), Bickel (1981); see Donoho et al. (1990) for further references and information). The case $p \to 0$ may also be considered; for $\epsilon \in (0, 1)$ put

$$I_0(\epsilon) = \inf\{I(\Phi * F) : F(0) \geq 1 - \epsilon\}$$

then $I_p(\epsilon^{1/p}) \to I_0(\epsilon)$ as $p \to 0$. I_0 has arisen before in the study of robust estimation (Mallows, 1979), and also in the study of estimating a normal mean which is likely to be near zero (Bickel, 1983).

Lemma 12 *Let $p \in (0, \infty]$. Then $I_p(\tau)$ is continuous and monotone decreasing in τ and*

$$\lim_{\tau \to 0} I_p(\tau) = 1$$

$$\lim_{\tau \to \infty} I_p(\tau) = 0$$

$$0 < I_p(\tau) < 1, \quad \tau \in (0, \infty)$$

Lemma 13 *The unique solution $F_{p,\tau}$ to the optimization problem (16) is Gaussian if and only if $p = 2$.*

The proof of these lemmas is omitted. Lemma 12 is a special case of Proposition 10, while Lemma 13 is proved when p is an integer in Feldman (1991). We note an interesting consequence of (17): $\rho_2(\tau) = \tau^2(1+\tau^2)^{-1}$ equals the minimax linear risk $\inf_{a,b} \sup\{E_\theta(ax + b - \mu)^2 : |\theta| \leq \tau\}$. From Donoho, Liu and MacGibbon (1990) now follows

$$\rho_2(\tau)/\rho_\infty(\tau) \leq \mu^* \doteq 1.25.$$

5 Univariate threshold rules

In this section, we study two families of threshold estimates that offer simple, near-optimal alternatives to the minimax-Bayes estimator in the univariate model $y = \mu + z,\ z \sim N(0, \sigma^2)$ in which μ is known to satisfy $E_F|\mu|^p \le \eta^p$. These families are useful because explicit expressions for the minimax Bayes estimator are available only when $p = q = 2$.

We consider both 'soft' and 'hard' threshold rules:

$$\delta_\lambda^{(s)}(y) = \mathrm{sgn}(y)(|y| - \lambda)_+, \quad \delta_\lambda^{(h)}(y) = yI\{|y| > \lambda\} \qquad \lambda \in (0, \infty).$$

The 'hard' threshold is a discontinuous estimator of the 'pretest' type. The 'soft' threshold is continuous, and goes also by the names of Hodges-Lehmann, limited translation, or ℓ_1-estimator. The latter terminology arises because $\delta_\lambda^{(s)}(y)$ is the minimising value of μ in $(y - \mu)^2 + \lambda|\mu|$.

We shall be interested in how an optimally-chosen threshold rule performs in comparison with the Bayes-minimax rule. Define

$$\rho_s(\eta, \sigma) = \inf_\lambda \sup \left\{ E_F E_\mu |\delta_\lambda^{(s)}(y) - \mu|^q : E_F|\mu|^p \le \eta^p \right\} \tag{18}$$

with a corresponding quantity $\rho_h(\eta, \sigma)$ for the hard-threshold rules. We note the invariances

$$\rho_s(\eta, \sigma) = \sigma^q \rho_s(\eta/\sigma, 1) \quad , \quad \rho_h(\eta, \sigma) = \sigma^q \rho_h(\eta/\sigma, 1) \tag{19}$$

which again ensure that it suffices to assume $\sigma = 1$. As was shown in Proposition 10 for $\rho(\eta, 1)$, the functions $\rho_s(\eta, 1)$ and $\rho_h(\eta, 1)$ are continuous, monotonic in η and concave in η^p.

For the comparison with Bayes-minimax estimators, define

$$\Lambda_s(p, q) = \sup_{\eta, \sigma} \frac{\rho_s(\eta, \sigma)}{\rho(\eta, \sigma)} > 1,$$

and $\Lambda_h(p, q)$ similarly.

Theorem 14 *For $(p, q) \in (0, \infty) \times [1, \infty)$, $\Lambda_s(p, q) < \infty$ and $\Lambda_h(p, q) < \infty$.*

Because of the invariance (19) and since $\rho(\eta, 1) > 0$ and $\rho_s(\eta, 1)$ are continuous on $(0, \infty)$, it suffices to consider limiting behavior as $\eta \to 0$ and ∞. In fact we will show more, namely that optimally chosen threshold rules are *asymptotically Bayes minimax* in these limiting cases.

Theorem 15 $\frac{\rho_s(\eta,1)}{\rho(\eta,1)}$ *and* $\frac{\rho_h(\eta,1)}{\rho(\eta,1)} \to 1$ *as $\eta \to 0$ and ∞,*

The limits as $\eta \to \infty$ are trivial since both threshold families include $\delta(x) = x$ which has risk equal to c_q. Thus $\rho_s(\eta, 1)$ and $\rho_h(\eta, 1) \le c_q$, whereas Proposition 10 showed that $\rho(\eta, 1) \nearrow c_q$.

The argument for $\eta \to 0$ is broken into two steps. First Proposition 16 below provides upper bounds for $\rho_s(\eta, 1)$ and $\rho_h(\eta, 1)$ by specifying certain choices of λ. Of course, these

serve also as upper bounds for $\rho(\eta, 1)$. In the next section, separate arguments are used to provide lower bounds (Theorem 18) for $\rho(\eta, 1)$ that agree asymptotically with the upper bounds and so complete the proof of Theorem 15 and so of Theorem 14.

We first establish some notation for risk functions of estimators in the case $\sigma = 1$. Write x for an $N(\mu, 1)$ variate and $r(\delta, \mu) = E_\mu|\delta(x) - \mu|^q$. Explicit formulas for the risk functions of the thresholds $\delta_\lambda^{(s)}$ and $\delta_\lambda^{(h)}$ are given in the Appendix. [3] We note here only that both risk functions are symmetric about $\mu = 0$, and that $r(\delta_\lambda^{(s)}, \mu)$ increases monotonically on $[0, \infty)$ to a bounded limit, whereas the risk of $\delta_\lambda^{(h)}$ rises from $\mu = 0$ to a maximum at $\lambda - o(\lambda)$ (as $\lambda \to \infty$) before decreasing to c_q as $\mu \to \infty$.

The average risk of an estimator δ under prior F will be written $r(\delta, F) = \int r(\delta, \mu) F(d\mu)$, and the worst average risk over $\mathcal{F}_p(\eta)$ is

$$\bar{r}(\delta, \eta) = \sup\{r(\delta, F) : F \in \mathcal{F}_p(\eta)\}.$$

Thus $\rho_s(\eta, 1) = \inf_\lambda \bar{r}(\delta_\lambda^{(s)}, \eta)$ and similarly for $\rho_h(\eta, 1)$.

Proposition 16 *Let $\lambda = \lambda(\eta)$ be chosen such that*

a) for soft thresholds, $\lambda^2 = 2\log \eta^{-p} + \alpha$ for $|\alpha| \le c_0$,

b) for hard thresholds $\lambda^2 = 2\log \eta^{-p} + \alpha \log(2\log \eta^{-p})$ for $\alpha > p - 1$. Then

$$\bar{r}(\delta_\lambda, \eta) \sim \eta^p \lambda^{q-p} \qquad \text{as } \eta \to 0.$$

Remarks. 1. An heuristic argument for the choice of λ goes as follows. The estimator $\hat{\theta}_\lambda$ is clearly related to the problem of deciding whether $|\theta_i|$ is larger than $\lambda\sigma$. The parameter space constraint limits the number of θ_i that can equal $\lambda\sigma$ to at most $n\epsilon$, where $n\epsilon(\lambda\sigma)^p = 1$. Consider the hypothesis testing problem with $Y \sim N(\theta, \sigma)$, with $H_0 : \theta = 0$ and $H_1 : \theta = \lambda\sigma$ and assign prior probabilities $\pi(H_0) = 1 - \epsilon$ and $\pi(H_1) = \epsilon$. The estimator θ_λ is related to the decision rule $\phi(y) = I\{|y| > \lambda\sigma\}$. The value of λ which makes the error probabilities approximately equal solves

$$P(\Theta = 0, Y > \lambda\sigma) = P(\Theta = \lambda\sigma, Y < \lambda\sigma),$$

$$\text{i.e. } (1 - \epsilon)\tilde{\Phi}(\lambda) \cong \epsilon/2. \tag{20}$$

If $p = 1$, the constraint $n\epsilon\lambda\sigma = 1$, together with the approximation $\tilde{\Phi}(\lambda) \sim \phi(\lambda)/\lambda$ and the definition $\eta^{-p} = n\sigma$ implies that $\lambda^2 \approx 2\log \eta^{-1} + 2\log\sqrt{2/\pi}$. For general p, equation (20) becomes approximately

$$\phi(\lambda) = \frac{\epsilon}{2}\lambda^p\lambda^{1-p} = \eta^p\lambda^{1-p}/2$$

with solution $\lambda^2 \approx 2\log \eta^{-p} + (p-1)\log(2\log \eta^{-p} + c) - \log(\pi/2) \approx 2\log \eta^{-p}$.

2. The optimal hard thresholds are slightly larger than the corresponding optimal soft cutoffs. One reason for this is seen by considering behavior of the risk functions near $\mu = 0$ in the squared error case $q = 2$. Indeed for fixed λ, $r(\delta_\lambda^{(h)}, 0) = 2[\lambda\phi(\lambda) + \tilde{\Phi}(\lambda)] \gg$

$2\lambda^{-1}\phi(\lambda) = r(\delta_\lambda^{(s)}, 0)$. The risk of the hard threshold is larger because of the discontinuity at λ, and can only be reduced by increasing λ.

Proof. We give only an outline, spelling out the extra details needed in Lemma 17 below. Let $r(\mu)$ denote either $r(\delta_\lambda^{(h)}, \mu)$ or $r(\delta_\lambda^{(s)}, \mu)$. Since $r(\mu)$ is increasing on $[0, \infty)$ (for $\delta_\lambda^{(s)}$) and on $[0, \mu_0(\lambda)]$ (for $\delta_\lambda^{(h)}$, with $\mu_0(\lambda) \uparrow$), it follows that for sufficiently small η, the relevant extreme points of $\mathcal{F}_p^+(\eta)$ are two point distributions $F = (1-\epsilon)\nu_{a_0} + \epsilon\nu_{a_1}$ for which $(1-\epsilon)a_o^p + \epsilon a_1^p = \eta^p$. For such distributions,

$$r(\delta_\lambda, F) = (1-\epsilon)r(a_0) + \epsilon r(a_1).$$

The first term turns out to be negligible, regardless of the choice of ϵ and a_0, so we are led to study the function

$$s(\mu) = (\frac{\eta}{\mu})^p r(\mu) \quad \mu \geq \eta.$$

A simpler approximation to $r(\mu)$ which is adequate for calculation (see Lemma 17 below) is provided by using the risk function $r_+(\mu) = E_\mu|\delta_\lambda^+(X) - \mu|^q$ of the one sided rules $\delta_\lambda^{s,+}(x) = (x-\lambda)_+$ and $\delta_\lambda^{h,+} = xI\{x > \lambda\}$ in the soft and hard threshold cases respectively.

In the case of soft thresholds, choose λ so that $|\lambda^2 - 2\log\eta^{-p}| \leq c_0$ for some $c_0 > 0$. By comparing coefficients of λ^q in $\mu^{p+1}\eta^{-p}s_+'(\mu)$, it transpires that $s_+'(\mu)$ has a zero at approximately $\mu_p = \lambda + \tilde{\Phi}^{-1}(p/2) = \lambda + z_p$, say. Calculation shows that

$$s(\lambda + z_p) \sim \eta^p \lambda^{q-p} \quad \text{as} \quad \eta \quad \to \quad 0. \tag{21}$$

For hard thresholds, one proceeds similarly to find that the zero of $s_+^{(\mu)}$ occurs at approximately $\mu_{pq} = \lambda - (2\log\lambda c_1^{-1})^{1/2}$ with $c_1 = (q-p)\sqrt{2\pi}$, and (21) remains true for $s(\mu_{pq})$.

To complete the outline for Proposition 16, we now collect the steps required to show that (21) maximises $s(\mu)$.

Lemma 17 *(a). The risk function $\mu \to r(\delta_\lambda, \mu)$ is increasing in $\mu \in [0, \infty)$ (resp for μ in a fixed neighborhood of zero for sufficiently large λ.) If $0 \leq a_0 \leq \eta$, and $\lambda = \lambda(\eta)$ as specified in Proposition 16, then $r(\delta_\lambda, a_0) \leq r(\delta_\lambda, \eta) \sim r(\delta_\lambda, 0) = o(\eta^p\lambda^{q-p})$. Indeed $r(\delta_\lambda^{(s)}, 0) \sim 2\Gamma(q+1)\lambda^{-q-1}\phi(\lambda)$, while $r(\delta_\lambda^{(h)}, 0) \sim 2\lambda^{q-1}\phi(\lambda)$).*

(b). Let $\delta(\mu) = r(\mu) - r_+(\mu)$. On $[0, \infty)$, $0 \leq \delta(\mu) \leq \delta(0) = r_+(0) = r(0)/2 = o(\eta^p\lambda^{q-p})$.

(c). For sufficiently large $d_0 > |z_p|$ (resp. sufficiently small $c_1 > 0$ and large $c_2 > 0$) and sufficiently small η, $s(\mu)$ has a unique global maximum on $[\eta, \infty)$, which is contained in $[\lambda - d_0, \lambda + d_0]$ (resp $[\lambda - \sqrt{2\log\lambda c_1^{-1}}, \lambda - \sqrt{2\log\lambda c_2^{-1}}]$).

(d). $s(\mu) \sim \eta^p\lambda^{q-p}$ uniformly in $[\lambda - d_0, \lambda + d_0]$, (resp in $[\lambda - \sqrt{2\log\lambda c_1^{-1}}, \lambda - \sqrt{2\log\lambda c_2^{-1}}]$.

6 Asymptotics for $\rho_{p,q}(\eta)$ for small η

This section is devoted to obtaining the exact rates (and constants) at which the univariate Bayes-minimax risk $\rho_{p,q}(\eta)$ decays as $\eta \to 0$. A basic dichotomy emerges: when $p \geq q$, the asymptotically least favorable distributions put all their mass at $\pm\eta$ and $\rho_{p,q}(\eta)$ decays like η^q. This rate is independent of the particular value of $p \geq q$. When $p < q$, the priors may have fewer moments than the order of the loss function. In this case, the asymptotically least favorable distributions are "nearly black", and put most mass at 0, with a vanishing fraction of mass at two large values $\pm\mu(\eta)$. In addition, $\rho_{p,q}(\eta)$ has a slower rate of convergence.

Theorem 18 *As* $\eta \to 0$

$$\rho_{p,q}(\eta) \sim \begin{cases} \eta^q & p \geq q \\ \eta^p (2\log \eta^{-p})^{(q-p)/2} & 0 < p < q \end{cases}$$

Proof. *Upper Bounds.* The Bayes risk $\rho_q(F)$ is the minimal value of $E_F|d(x)-\mu|^q$ over all estimators d. Choosing $d = 0$ gives the upper bound $\rho_q(F) \leq E_F|\mu|^q = |F|_q^q$. If $q \leq p$, and $F \in \mathcal{F}_p(\eta)$, then $|F|_q \leq |F|_p \leq \eta$, and so

$$\rho_{p,q}(\eta) = \inf_{\mathcal{F}_p} \rho_q(F) \leq \inf_{\mathcal{F}_p} E_F|\mu|^q \leq \eta^q. \tag{22}$$

For $0 < p < q$, we use the bounds derived for threshold rules in the previous section. Indeed, from the minimax theorem, and choosing λ as in Proposition 16, we obtain

$$\rho_{p,q}(\eta) = \sup_{\mathcal{F}_p(\eta)} \inf_{\delta} R(\delta, F) = \inf_{\delta} \sup_{\mathcal{F}_p(\eta)} R(\delta, F)$$

$$\leq \sup_{\mathcal{F}_p(\eta)} R(\delta_\lambda, F) = \eta^p (2\log \eta^{-p})^{(q-p)/2}(1+o(1)).$$

Lower bounds. It suffices to evaluate $\rho_q(F)$ for distributions F approximately least favorable for $\mathcal{F}_p(\eta)$. As $\eta \to 0$, discrete priors supported on two or three points are enough.

Consider first two point priors $F_\eta = (\nu_\eta + \nu_{-\eta})/2$. By symmetry, $d_F(-x) = -d_F(x)$, and if we write $d_F(x) = \eta e_{q,\eta}(x)$, then

$$\rho_q(F_\eta) = E_F|d_F(x)-\mu|^q = \eta^q \int |1 - e_{q,\eta}(x)|^q \phi(x-\eta)dx.$$

By minimising the posterior risk, the Bayes rule is found to be

$$d_F(x) = \begin{cases} \eta \tanh \eta x/(q-1) & q > 1 \\ \eta \mathrm{sign}(x) & q = 1 \end{cases}$$

from which it follows that $\rho_q(F_\eta) \sim \eta^q$ for $q \geq 1$. Since $|F_\eta|_p = \eta$ for all p, this asymptotic lower bound for $\rho_{p,q}(\eta)$ establishes the theorem for $p \geq q$.

When $0 < p < q$, we employ three point priors putting most mass at zero and a small fraction vanishing at ∞.

Proposition 19 *Let $F_{\epsilon,\mu} = (1-\epsilon)\nu_0 + \epsilon(\nu_\mu + \nu_{-\mu})/2$. Fix $a > 0$, and for all sufficiently small ϵ, define $\mu = \mu(\epsilon)$ by*

$$\phi(a+\mu) = \epsilon\phi(a) \tag{23}$$

Then

$$\rho_q(F_{\epsilon,\mu}) \sim \epsilon\mu^q\Phi(a) \quad \text{as} \quad \epsilon \to 0. \tag{24}$$

Before proving Proposition 19, we use it to complete the proof of Theorem 18. Clearly $|F_{\epsilon,\mu}|_p^p = \epsilon\mu^p$, while from (23) it follows that $\mu(\epsilon) \sim (2\log\epsilon^{-1})^{1/2}$. If we connect η and ϵ by the relation $\eta^p = \epsilon\mu^p$, then $F_{\epsilon,\mu}$ belongs to $\mathcal{F}_p(\eta)$ and so from (24)

$$\rho_{p,q}(\eta) \geq \rho_q(F_{\epsilon,\mu}) \sim \epsilon\mu^p.\mu^{q-p}\Phi(a) \sim \eta^p(2\log\eta^{-p})^{(q-p)/2}\Phi(a) \quad \text{as} \quad \eta \to 0. \tag{25}$$

The lower bound needed for Theorem 18 follows by taking a large.

Proof of Proposition 19 . Let $d_F(x)$ denote the Bayes rule for estimation of τ from data $x \sim N(\tau,1)$ and prior distribution $F_{\epsilon,\mu}(d\tau)$. Since the posterior distribution of τ given x is concentrated on $\{0, \pm\mu\}$, we may write $d_F(x) = \mu e_{q,\epsilon}(x)$, where $|e_{q,\epsilon}(x)| \leq 1$ and in addition $e_{q,\epsilon}(x)$ is an odd function of x. Thus the Bayes risk

$$\rho_q(F_{\epsilon,\mu}) = 2(1-\epsilon)\mu^q\int_0^\infty |e_{q,\epsilon}(x)|^q\phi(x)dx + \epsilon\mu^q\int |1-e_{q,\epsilon}(x)|^q\phi(x-\mu)dx.$$

We complete the proof by showing, separately for $q > 1$ and $q = 1$, that as $\epsilon \to 0$,

$$\int_0^\infty |e_{q,\epsilon}(x)|^q\phi(x)dx = o(\epsilon), \qquad \text{and} \tag{26}$$

$$e_{q,\epsilon}(\mu+z) \to I\{z > a\}. \tag{27}$$

First, for $q = 1$, $d_F(x)$ is the posterior median, and thus for positive x, $e_{q,\epsilon}(x) = I\{x \geq x_0\}$, where x_0 solves $p(\mu|x) = 1/2$. Thus, x_0 solves

$$\epsilon\phi(x-\mu) = 2(1-\epsilon)\phi(x) + \epsilon\phi(x+\mu).$$

Substituting definition (23) for ϵ, we find that $x_0 = a + \mu + \mu^{-1}\log 2(1-\epsilon) + o(1)$. The integral in (26) is thus bounded by $\tilde{\Phi}(x_0) \leq \tilde{\Phi}(a+\mu) \leq \phi(a+\mu)/(a+\mu) = o(\epsilon)$ from definition (23). Relation (27) is immediate from the form of x_0.

For $q > 1$, $d_F(x)$ is the minimiser of $a \to E[|a-\mu|^q|x]$. If $x > 0$, then $0 \leq d_F(x) \leq \mu$, and differentiation shows that $d_F(x)$ is the solution of the equation

$$\epsilon(\mu-a)^{q-1}p_+ = 2(1-\epsilon)a^{q-1}p_0 + \epsilon(\mu+a)^{q-1}p_- \tag{28}$$

where $p_\pm = \phi(x \mp \mu)$ and $p_0 = \phi(x)$.

Using (23) one verifies that for $x > 0$ and μ large, $\epsilon(\mu+a)^{q-1}p_- < \epsilon a^{q-1}p_0$ and hence that $d_F(x) \in [d_{\epsilon/2,\epsilon}(x), d_{\epsilon,\epsilon}(x)]$, where $d_{\delta,\epsilon}(x)$ is the solution of the simpler equation

$$\epsilon(\mu-a)^{q-1}p_+ = 2(1-\delta)a^{q-1}p_0.$$

Using (23), $p_0/\epsilon p_+ = \phi(x)/\epsilon\phi(x-\mu) = e^{-\mu(x-\mu-a)}$, and thus

$$d_{\delta,\epsilon}(x) = \mu[1 + 2^p(1-\delta)^\beta e^{-\mu\beta(x-\mu-a)}]^{-1}, \qquad \beta = 1/(q-1). \tag{29}$$

Making the substitution $z = x - \mu - a$ and using (23), the integral in (26) is bounded above by

$$\epsilon\phi(a)\int_{-\infty}^{\infty}[1 + e^{-\mu\beta z}]^{-q}e^{-z\mu - za - z^2/2}dz = o(\epsilon),$$

as may be seen by arguing separately for positive and negative z. The convergence required for (27) follows readily from the representation (29). ∎

7 Asymptotic sharpness of the Bayes-minimax risk bound

The purpose of this section is to show that the upper bound $R_N^* \leq R_B^*$ is often asymptotically an *equality*: nothing is lost by replacing the n-variate problem by n univariate problems.

Theorem 20 *If either (i) $p \geq q$, or (ii) $0 < p < q$ and $(\sigma/r)^2 \log n(\sigma/r)^p \to 0$, then*

$$R_N^*(\sigma; \Theta_{p,n}(r)) = R_B^*(\sigma; \Theta_{p,n}(r))(1 + o(1)). \tag{30}$$

In case (ii), the condition that $(\sigma/r)^2 \log n(\sigma/r)^p \to 0$ cannot be completely removed: if, for example, $\sigma/r = n^\alpha$, $\alpha > 0$ and (30) holds, then in combination with $R_L^* \sim 1$ (Theorem 7, part 3) and Theorem 18 we would conclude that $R_L^*/R_N^* \to 0$ which is absurd.

The approach is to show that certain nearly least favorable priors on $\Theta_{p,n}(r)$ can be approximated by i.i.d priors. Recall from Proposition 9 that $R_B^* = n\sigma^q \rho(\eta_n)$. Let F_n be a sequence of prior distributions on $\mu \in R^1$, to be chosen so that

$$r_n(F_n) = \rho(F_n)/\rho(\eta_n)$$

is close to 1. Denote by P_n the prior on θ which makes θ_i/σ, $i = 1, ..., n$, i.i.d. F_n. The i.i.d. structure implies that

$$\rho(P_n) = n\sigma^q \rho(F_n).$$

Thus $r_n(F_n) = \rho(P_n)/R_B^*$ also. Now let π_n be the conditional distribution of P_n restricted to $\Theta_n \stackrel{def}{=} \Theta_{p,n}(r)$: thus $\pi_n(A) = P_n(A|\theta \in \Theta_n)$. Clearly,

$$\frac{R_N^*}{R_B^*} \geq \frac{\rho(\pi_n)}{\rho(P_n)} r_n(F_n), \tag{31}$$

and the idea is to show that $\rho(\pi_n)/\rho(P_n) \geq 1 + o(1)$ for the sequence $\{F_n\}$.

Given a prior $\pi(d\theta)$ and estimator $\hat{\theta}(x)$, we denote the integrated risk of $\hat{\theta}$ over the joint distribution of (θ, x) by $\mathcal{E}_\pi |\hat{\theta} - \theta|^q$: of course for fixed π, the minimum over $\hat{\theta}$ is $\rho(\pi)$, which is attained by the Bayes rule $\hat{\theta}_\pi$. From the definition of π_n , we obtain

$$\begin{aligned}
\rho(P_n) &\leq \mathcal{E}_{P_n} |\hat{\theta}_{\pi_n} - \theta|^q && (32)\\
&= \mathcal{E}_{P_n} \{|\hat{\theta}_{\pi_n} - \theta|^q \mid \Theta_n\} P_n(\Theta_n) + \mathcal{E}_{P_n} \{|\hat{\theta}_{\pi_n} - \theta|^q, \Theta_n^c\} && (33)\\
&\leq \rho(\pi_n) P_n(\Theta_n) + 2^q \mathcal{E}_{P_n} \{|\hat{\theta}_{\pi_n}|^q + |\theta|^q; \Theta_n^c\}. && (34)
\end{aligned}$$

The argument now splits into cases according as $\eta_n \to \eta \in (0, \infty]$ or $\eta_n \to 0$. [Of course, by passing to subsequences, we may assume that such a limit exists.] In the latter case, the manner in which the approximately least favorable distributions F_n converge to 0 depends on whether Θ is loss-convex. What remains to be shown follows the same pattern in each situation: Choose F_n so that (i) $r_n(F_n)$ is close to 1, (ii) $P_n(\Theta_n) \to 1$ and (iii) that the final term in (34) is negligible relative to $\rho(P_n)$.

Case (a). Assume first that $\eta_n \to \eta \in (0,\infty]$. Choose $\epsilon > 0$ and a sequence of distributions $F_{(k)}(d\mu) \in \mathcal{F}_p(\eta - \epsilon)$ such that $\rho(F_{(k)}) \to \rho(\eta - \epsilon)$ and $\mathrm{supp} F_{(k)} \subset [-k,k]$. Now fix k and let $F_n = F_{(k)}$ for all n. The event $\{\theta \in \Theta_n\} = \{n^{-1}\sum_1^n |\mu_i|^p \le \eta_n^p\}$ has probability approaching 1 as $n \to \infty$ since $E|\mu|^p \le (\eta-\epsilon)^p < \eta^p = \lim \eta_n^p$. Since $\mathrm{supp} F_{(k)} \subset [-k,k]$,

$$\mathcal{E}_{P_n}\{|\hat\theta_{\pi_n}|^q + |\theta|^q; \Theta_n^c] \le 2n\sigma^q k^q P(\Theta_n^c).$$

which is therefore asymptotically negligible relative to $\rho(P_n) = n\sigma^q \rho(F_{(k)})$. Thus $\rho(\pi_n)/\rho(P_n) \ge 1 + o(1)$, and $r(F_n) = r(F_{(k)}) \sim \rho(F_{(k)})/\rho(\eta)$. The proof is completed by taking ϵ small and k large.

Case (b). Now assume that $\eta_n \to 0$ and $p \ge q$. The priors $F_{\eta_n} = (\nu_{\eta_n} + \nu_{-\eta_n})/2$ are asymptotically least favorable (Section 6), and thus $r_n(F_{\eta_n}) \to 1$. In addition, $P_n(\sum_1^n |\theta_i|^p \le r^p) = 1$, since $\sum_1^n |\theta_i|^p \equiv n\sigma^p\eta^p = r^p$, so that in this case $\pi_n = P_n$ and the equivalence (30) follows immediately from (34).

Case (c). Finally, if $p < q$, and $\eta_n \to 0$, we use the symmetric three point priors $F_{\epsilon,\mu}$ studied in Proposition 19. Fix $\delta, a > 0$, and define $\epsilon = \epsilon_n$ implicitly by the relation

$$\epsilon\mu^p = (1-\delta)\eta_n^p = (1-\delta)n^{-1}(r/\sigma)^p. \tag{35}$$

($\mu = \mu(\epsilon, a)$ is already defined by equation (23)). Let $F_n = F_{\epsilon_n,\mu_n}$. From Proposition 19 and (25),

$$\rho(F_n) \sim \epsilon\mu^q\Phi(a) \sim (1-\delta)\eta_n^p(2\log\eta_n^{-p})^{(q-p)/2}\Phi(a), \tag{36}$$

while from Theorem 18, $\rho_{p,q}(\eta) \sim \eta^p(2\log\eta^{-p})^{(q-p)/2}$. Thus $r(F_n) \sim (1-\delta)\Phi(a)$.

Let N_n count the number of non-zero μ_i: clearly N_n is distributed as Binomial(n,ϵ) and (35) implies that $EN_n = n\epsilon = (1-\delta)(r/\sigma)^p\mu^{-p}$. The event Θ_n equals

$$\{\sum |\mu_i|^p \le (r/\sigma)^p\} = \{N_n \le (r/\sigma)^p\mu^{-p} = EN_n/(1-\delta)\}.$$

In view of (35), $EN_n = n\epsilon \to \infty$ exactly when $\mu^p(\sigma/r)^p \to 0$. But $(\sigma/r)^2\mu^2 \sim 2(\sigma/r)^2\log\epsilon^{-1} \sim 2(\sigma/r)^2\log n(\sigma/r)^p \to 0$ by the hypothesis of case (ii) of the theorem. Now apply Chebychev's inequality to get

$$P_n(\Theta_n^c) = P\{(N_n - EN_n)/EN_n \ge \delta/(1-\delta)\} \le \delta^{-2}(1-\delta)^2/n\epsilon \to 0.$$

Similarly, $E_{P_n}|N_n - EN_n|/EN_n \to 0$.

The Bayes estimator may be bounded [4] in terms of the posterior moment:

$$|\hat\theta_{\pi_n}|^q \le 2^q E_{\pi_n}(|\theta|^q|x). \tag{37}$$

Since π_n is concentrated on Θ_n, the corresponding bound on the posterior law of N_n implies that

$$E_{\pi_n}(|\theta|^q|x) = \sigma^q\mu^q E_{\pi_n}(N_n|x) \le \sigma^q\mu^q EN_n/(1-\delta).$$

Thus,

$$\mathcal{E}_{P_n}[|\hat\theta_{\pi_n}|^q + |\theta|^q, \Theta_n^c] \le 2^{q+1}\sigma^q\mu^q\mathcal{E}_{P_n}\{EN_n + N_n, \Theta_n^c\}$$

while from (36)

$$\rho(P_n) \sim n\sigma^q \epsilon \mu^q \Phi(a) = \sigma^q \mu^q \Phi(a) E N_n.$$

The ratio of these two expressions is bounded by a constant multiple of

$$P(\Theta_n^c) + \mathcal{E}_{P_n}(\frac{N_n}{EN_n}, \Theta_n^c) \leq 2P(\Theta_n^c) + E\frac{|N_n - EN_n|}{EN_n} \to 0.$$

Returning to (34), we find therefore that $\rho(P_n) \leq \rho(\pi_n)(1 + o(1))$. Since δ and a may be chosen arbitrarily small and large respectively, this establishes the lower bound part of (23) in this case. ∎

8 Threshold rules over l_p balls

In this section we use the Bayes-minimax approach and the results for univariate threshold rules of the previous section to prove Theorem 6 on the asymptotic near-optimality of threshold rules.

Define a Bayes minimax quantity analogous to R_B^* except that attention is restricted to threshold rules:

$$R_s^*(\sigma;\Theta_{p,n}(r)) = \inf_{\lambda}\{\sup_{\pi} E_\pi E_\theta ||\hat{\theta}_\lambda^{(s)} - \theta||_q^q : E_\pi \sum_1^n |\theta_i|^p \le r^p\}. \tag{38}$$

(with a similar definition of R_h^* for hard thresholds). Clearly

$$\inf_{\lambda} \sup_{\theta\in\Theta_{p,n}(r)} E_\theta ||\hat{\theta}_\lambda^{(s)} - \theta||_q^q \le R_s^*.$$

We may relate R_s^* to the univariate threshold problem (18) of Section 5 exactly as Proposition 9 did for R_B^*.

Lemma 21 $\quad R_s^*(\sigma,\Theta_{p,n}(r)) = n\rho_s(n^{-1/p}r,\sigma).$

The proof is entirely analogous: if (F^o,δ_λ^o) is a saddlepoint for problem (18), then $(F^{on},\delta_\lambda^{on})$ is a saddlepoint for problem (38). Theorem 6 now follows from the bounded inefficiency of $\rho_s(\tau,\sigma)$ relative to $\rho(\tau,\sigma)$ (Theorem 14), and from Theorem 20:

$$\begin{aligned} R_s^*(\sigma,\Theta_{p,n}(r)) &= n\rho_s(n^{-1/p}r,\sigma) \\ &\le \Lambda_s(p,q) n\rho(n^{-1/p}r,\sigma) \\ &= \Lambda_s(p,q) R_B^*(\sigma,\Theta_{p,n}(r)) \\ &\le \Lambda_s(p,q) R_N^*(1+o(1)). \end{aligned}$$

Note also that

$$\frac{R_s^*}{R_B^*} = \frac{\rho_s(n^{-1/p}r,\sigma)}{\rho(n^{-1/p}r,\sigma)} = \frac{\rho_s(\eta_n,1)}{\rho(\eta_n,1)}$$

where $\eta_n = n^{-1/p}(r/\sigma)$, so that if $\eta_n \to 0$,

$$R_s^* \sim R_N^*$$

and threshold rules are asymptotically *efficient* .

9 Linear Minimax Risk

We now turn to the minimax risk amongst linear estimators of the form $\hat{\theta}(y) = Ay + c$ for $n \times n$ matrix A, and $n \times 1$ vector c. As noted earlier, the estimation problem is invariant under the action of the group G corresponding to permutation of indices. It follows then (using convexity of the loss functions $l_q, q \ge 1$) that the minimax linear estimator is itself invariant : $\hat{\theta}(gy) = g\hat{\theta}(y)$ for $g \in G$. Thus $\hat{\theta}$ has the form $\hat{\theta}_{abc,i}(x) = ax_i + b(\sum_{j\ne i} x_j) + c$. A

further convexity argument [5] using orthosymmetry of $\Theta = \Theta_{p,n}(r)$ shows that $\hat{\theta}_{a00}(x) = ax$ has smaller maximum risk over $\Theta_{p,n}(r)$ than $\hat{\theta}_{abc}$. Finally, $\hat{\theta}_{|a|}$ dominates $\hat{\theta}_a$ for a negative, and $\hat{\theta}_1$ dominates $\hat{\theta}_a$ for $a > 1$. Thus

$$R_L^*(\sigma; \Theta_{p,n}(r)) = \inf_{0 \le a \le 1} \sup_{\Theta_{p,n}(r)} E_\theta ||aY - \theta||_q^q. \tag{39}$$

Converting to variables $X_i = Y_i/\sigma$ and $\mu_i = \theta_i/\sigma$, and recalling that $\eta_n^p = n^{-1}(r/\sigma)^p$, we obtain

$$\sup_{\Theta_{p,n}(r)} E_\theta ||aY - \theta||_q^q = n\sigma^q \sup\{n^{-1} \sum_1^n E_{\mu_i} |aX_i - \mu_i|^q : n^{-1} \sum |\mu_i|^p \le \eta_n^p\}. \tag{40}$$

The risk function in the univariate location problem that appears on the right side of (40) can be expressed in terms of a single standard Gaussian deviate Z:

$$\begin{aligned} r_q(a, \mu) &= E_\mu |aX - \mu|^q \quad = a^q E|Z + b\mu|^q \\ &= a^q s(b^p |\mu|^p), \end{aligned}$$

where $b = a^{-1} - 1 \in [0, \infty)$, and we have introduced the function

$$s(\gamma) = E|Z + \gamma^{1/p}|^q, \qquad \gamma \in [0, \infty).$$

Since $s(\gamma)$ is increasing in γ, there is no harm in replacing the inequality in the supremum in (40) by equality. We obtain

$$\begin{aligned} R_L^* &= n\sigma^q \inf_a a^q \sup\{n^{-1} \sum s(\gamma_i) : n^{-1} \sum \gamma_i = b^p \eta^p, \gamma_i \ge 0\} & (41) \\ &= n\sigma^q \inf_{b \ge 0} (1 + b)^{-q} s_n^*(b^p \eta^p). & (42) \end{aligned}$$

Remark. The function s_n^* implicitly defined in (42) is closely related to the *concave majorant* of s, the smallest concave function pointwise larger than s. The empirical distribution of a vector $(\gamma_1, \ldots, \gamma_n)$ with $\gamma_i \ge 0, n^{-1} \sum \gamma_i = \tau$ belongs to the class $\mathcal{F}_1^{(n)}$ of probability measures supported on $[0, n\tau]$ with mean equal to τ. Thus

$$s_n^*(\tau) \le \tilde{s}_n(\tau) \stackrel{\text{def}}{=} \sup\{\int s(\gamma) F(d\gamma), F \in \mathcal{F}_1^{(n)}\}. \tag{43}$$

The extreme points of the convex set $\mathcal{F}_1^{(n)}$ are two point distributions with mean τ, so that

$$\begin{aligned} \tilde{s}_n(\tau) &= \sup\{\alpha s(\gamma_1) + (1 - \alpha) s(\gamma_2) : \\ &\qquad \alpha\gamma_1 + (1 - \alpha)\gamma_2 = \tau, 0 \le \alpha \le 1, 0 \le \gamma_i \le n\tau\} \end{aligned}$$

which shows that $\tilde{s}_n$ is indeed the concave majorant of s on the interval $[0, n\tau]$ (e.g. Rockafellar, 1970, Corollary 17.1.5).

To evaluate (41), we first study the convexity properties of $s(\gamma) = E|Z + \gamma^{1/p}|^q$, chiefly using sign change arguments. Let $c = \gamma^{1/p}$ and v denote an $N(c,1)$ variate, so that $s(\gamma) = E_c|v|^q$. Some calculus shows that

$$q^{-1}p\gamma^{1-1/p}s'(\gamma) = E_c v|v|^{q-2}, \tag{44}$$

and, more importantly, that

$$\begin{aligned} q^{-1}p^2\gamma^{2-1/p}s''(\gamma) &\stackrel{\text{def}}{=} F(c;p,q) && (45)\\ &= \begin{cases} (q-1)E_c c|v|^{q-2} - (p-1)E_c v|v|^{q-2} & q>1 \\ 2c\phi(c) - (p-1)[2\Phi(c)-1] & q=1 \end{cases} && (46)\end{aligned}$$

A useful representation [6] is

$$e^{c^2/2}F(c) = 2\int_0^\infty g(v)v^{q-3}\phi(v)\sinh cv\,dv \qquad q>1, \tag{47}$$

where $g(v) = (q-p)v^2 - (q-1)(q-2)$ has at most one sign change on $[0,\infty)$. The kernel $(c,v) \to \sinh cv$ is totally positive of order 2 on $[0,\infty)$, and so, according to the variation diminishing property of totally positive kernels, $F(c)$ has no more sign changes than $g(v)$. By examining particular cases, we are led to a partition of S according to the convexity behavior of $s(\gamma)$. Formally [7] ,

$$\begin{aligned} X &= S\cap\{p\le q, p\le 2\} = \{(p,q): s \text{ is convex on } [0,\infty)\} \\ V &= S\cap\{p\ge q, p\ge 2\} = \{(p,q): s \text{ is concave on } [0,\infty)\} \\ XV &= S\cap\{p> q, p< 2\} = \{(p,q): s \text{ is convex on } [0,\gamma_0],\ \text{concave on } [\gamma_0,\infty]\} \\ VX &= S\cap\{p< q, p> 2\} = \{(p,q): s \text{ is concave on } [0,\gamma_0],\ \text{convex on } [\gamma_0,\infty]\} \end{aligned}$$

In the last two cases $\gamma_0 = \gamma_0(p,q)$ satisfies $0 < \gamma_0 < \infty$.

Lower bound based on a spike image While this argument is generally valid for $(p,q) \in S$, it is most useful when $p \le q$. Fix $\theta = (r,0,\ldots,0)$, which corresponds to $\mu = (r\sigma^{-1},\ldots,0)$, to obtain the lower bound

$$\begin{aligned} R_L^* &\ge n\sigma^q \inf_a\ (1-n^{-1})E_0|aX|^q + n^{-1}E_{r\sigma^{-1}}|aX - r\sigma^{-1}|^q && (48)\\ &= n\sigma^q\ \inf_a\ (1-n^{-1})a^q c_q + \tilde\eta_n^q(1-a)^q t(a,\sigma), && (49)\end{aligned}$$

where we have introduced the abbreviations $\tilde\eta_n^q = n^{-1}(r/\sigma)^q$ and $t(a,\sigma) = E|a(1-a)^{-1}\sigma Z - 1|^q$. Note that when $q \ge p$, $\tilde\eta_n = \bar\eta_n = n^{-1/(p\vee q)}r\sigma^{-1}$. Consider now the function

$$f(a;\eta) = a^q c_q + \eta^q(1-a)^q, \qquad q\ge 1, \eta\in(0,\infty).$$

For $q > 1$, $f(\cdot;\eta)$ has unique minimizer and minimum given by

$$a_*(\eta) = (1 + b_q\eta^{q'})^{-1}, \qquad f(a_*;\eta) = c_q a_*^{q-1}(\eta)$$

where $q' = q/(q-1)$ is the conjugate exponent to q, and $b_q = c_q^{1/(q-1)}$. When $q = 1$, $f(\cdot, \eta)$ is linear and the corresponding values are

$$a_* = I\{c_1 < \eta\} \qquad f(a_*; \eta) = c_1 \wedge \eta.$$

Some technical work shows that

$$\inf_a (1 - n^{-1})a^q c_q + \tilde{\eta}_n^q (1-a)^q t(a, \sigma) \sim f(a_*(\tilde{\eta}_n), \tilde{\eta}_n) \quad \text{as} \quad n \to \infty. \tag{50}$$

Combining these results with the lower bound in (49) yields

$$R_L^* \geq (1 + o(1)) \begin{cases} n\sigma^q c_q & \tilde{\eta}_n \to \infty \quad (a) \\ n\sigma^q f(a_*(\eta); \eta) & \tilde{\eta}_n \to \eta \in (0, \infty) \quad (b) \\ r^q & \tilde{\eta}_n \to 0. \quad (c) \end{cases} \tag{51}$$

Upper bounds. It is now necessary to consider the various cases in the decomposition of Figure 1 in turn. We begin by making various choices of a in (39). The choice $a = 1$ leads to

$$R_L^* \leq \sup_{\Theta_{p,n}(r)} E_\theta |Y - \theta|^q = n\sigma^q c_q,$$

which is sharp when $\tilde{\eta}_n \to \infty$ (cf. (51a)), and so establishes case 1 of Theorem 7. The choice $a = 0$ gives

$$\begin{aligned} R_L^* &\leq n\sigma^q \sup_{n^{-1}\sum \mu_i^p = n^{-1} r^p \sigma^{-p}} n^{-1} \sum_1^n |\mu_i|^q \\ &= n\sigma^q \sup_{n^{-1}\sum \gamma_i = \eta_n^p} n^{-1} \sum_1^n \gamma_i^{q/p}, \end{aligned}$$

after setting $\gamma_i = \mu_i^p$. When $q \geq p$, the function $\gamma \to \gamma^{q/p}$ is convex on $[0, \infty)$, so the least favorable configuration of γ_i is $(n\eta_n^p, 0, \ldots, 0)$ which implies that

$$R_L^* \leq n\sigma^q \; n^{-1} (n\eta_n^p)^{q/p} = r^q,$$

which is in turn sharp when $\tilde{\eta_n} \to 0$. (cf. (51c)).

Consider now the case $\tilde{\eta}_n \to \eta \in (0, \infty)$. When $q \geq p$ *and* $p \leq 2$ (i.e. $(p, q) \in X$), $s(\gamma)$ is convex and $\mu = (r\sigma^{-1}, 0, \ldots, 0)$ is a least favorable configuration. Consequently, *equality* holds in (49). When combined with (50), this shows that (51b) is sharp.

When $q > p$ and $p > 2$ (i.e. $(p, q) \in VX$), $s(\gamma)$ is concave near 0 but convex for large γ. For fixed n, the configuration $\mu = (r\sigma^{-1}, 0, \ldots, 0)$ is not exactly least favorable, but it is *asymptotically* least favorable, and so again (51b) is asymptotically sharp [8] . This completes the proof of Theorem 7 for the sets X and VX.

Let us now assume that $s(\gamma)$ is concave, i.e. that $(p, q) \in V = S \cap \{p \geq q, p \geq 2\}$. In this case, the vector $\mu = \eta_n(1, \ldots, 1)$ is least favorable, and from (41), we obtain

$$R_L^* = n\sigma^q \inf_{b \geq 0} \frac{s(b^p \eta_n^p)}{(1+b)^q}. \tag{52}$$

It turns out [9] that there is a unique minimax linear estimator $\mu(x) = x/(1+b_*)$, where $b_* = b_*(q,\eta_n) \in (0,\infty)$ if $\eta_n \in (0,\infty)$. If $\eta_n \to \infty$, then $b_*(\eta_n) \sim \eta_n^{-2}$ and $R_L^* \sim n\sigma^q c_q$. On the other hand, if $\eta_n \to 0$, then $b_* \sim (q-1)\eta_n^{-2}$ and $R_L^* \sim n\sigma^q \eta_n^q$. We believe that $b_*(\eta)$ decreases monotonically from ∞ to 0 as η increases from 0 to ∞, but have only verified this for loss functions with $q = 1, 2$ and 4.

We turn finally to the exceptional case in which $s(\gamma)$ is convex-concave, i.e. when $(p,q) \in XV = S \cap \{p > q, p < 2\}$. Consider first the simple case in which $\eta_n \to 0$. The right side of (52) is still a valid lower bound for R_L^*, and so from the discussion above, we conclude that $R_L^* \geq n\sigma^q \eta_n^q (1+o(1))$. On the other hand, a natural upper bound is obtained from the estimator with $a = 0$:

$$R_L^* \leq n\sigma^q \sup_{n^{-1}\sum \gamma_i = \eta_n^p} n^{-1} \sum_{i=1}^n \gamma_i^{q/p} = n\sigma^q \eta_n^q,$$

since $\gamma \to \gamma^{q/p}$ is concave. This establishes that $R_L^* \sim n\sigma^q \eta_n^q = n^{1-q/p} r^q$ when $\eta_n \to 0$.

Now suppose that $\eta_n \to \eta \in (0,\infty)$. An upper bound is derived form (42) and (43):

$$R_L^* \leq n\sigma^q \inf_b (1+b)^{-q} \tilde{s}(b^p \eta_n^p) \tag{53}$$

where the least concave majorant $\tilde{s}$ has the form

$$\tilde{s}(\gamma) = \begin{cases} c_q + R\gamma & \gamma \leq \gamma_0 \\ s(\gamma) & \gamma \geq \gamma_0 \end{cases} \tag{54}$$

where $R = [s(\gamma_0) - s(0)]/\gamma_0$ and $\gamma_0 = \gamma_0(p,q) \in (0,\infty)$ is the solution to the equation

$$s'(\gamma_0) = \frac{s(\gamma_0) - s(0)}{\gamma_0}.$$

As is shown in the Appendix [10], the error involoved in the upper bound (53) is $0(n\sigma^q n^{-1})$. Since this is negligible relative to the maximum value, the bound may be treated as an asymptotic equality.

Again, it turns out [11] that there is a unique value $b_* = b_*(p,q,\eta_n)$ optimizing the right side of (53). If the corresponding value of $\gamma_*(= b_*^p \eta_n^p)$ exceeds γ_0, then the least favorable configuration $\mu_* = \eta_n(1,\ldots,1)$ as in the concave case. However, if $\gamma_* < \gamma_0$, then the least favorable distribution (in (43)) has the form $(1-\epsilon)\delta_0 + \epsilon\delta_{\gamma_o}$, where $\epsilon\gamma_0 = \gamma_*$. It turns out that $\gamma_* < \gamma_0$ exactly when

$$\eta p(s(\gamma_o) - c_q) + [p(s(\gamma_0) - c_q) - qs(\gamma_0)]\gamma_0^{1/p} > 0, \tag{55}$$

which can always be ensured by taking η sufficiently large. Thus the set XV provides examples where the least favorable configuration is neither a spike image nor uniformly grey.

Acknowledgements. Donoho was supported at U.C. Berkeley by NSF DMS 84-51753, 88-10192 and Schlumberger-Doll. Johnstone was supported in part by NSF DMS 84-51750, 86-00235 NIH PHS GM21215-12, SERC and the Sloan Foundation. This is an expanded version of Technical Report 322, "Minimax risk over l_p-balls", Statistics Dept., Stanford University, May 1989.

10 Appendix

1. *Uniqueness and properties of the Bayes estimator.* $\mu_F(x) = \min_a^{-1} E_F[|a-\mu|^q \mid x]$. When $q > 1$, this follows from strict convexity of $a \to |a-\mu|^q$ (e.g., Lehmann, 1983, p. 240). When $q = 1$, a standard argument shows that $\mu_F(x)$ may be taken as any posterior median of μ, i.e., as any member of the interval

$$I(x) = \{a : \int_{(-\infty,a)} \phi(x-\mu)F(d\mu) = \int_{(a,\infty)} \phi(x-\mu)F(d\mu)\},$$

and let $\overline{I(x)} = [a_1(x), a_2(x)]$. Let $\lambda(A)$ denote the Lebesque measure of a set $A \subset I\!R$. If $\mu_F(x)$ is not unique for Lebesque almost all x, then

$$0 < \int \lambda(I(x))\, dx = \int d\mu \int dx I\{a_1(x) < \mu_0 < a_2(x)\},$$

so for some μ_0, the set $A_{\mu_0} = \{x : a_1(x) < \mu_0 < a_2(x)\}$ has positive Lebesque measure and hence positive probability under each distribution $N(\mu, 1)$. It would then follow that $\tilde{\mu}(x) = [a_1(x) + a_2(x)]/2$ has strictly smaller Bayes risk than either $a_1(x)$ or $a_2(x)$, which would contradict their definition.

We note also that if the posterior distribution of μ is symmetric about some point, then that point equals the posterior mean of μ and $\mu_F(x) = E_F[\mu|x]$ for all $q \geq 1$. In general, of course, the value of $\mu_F(x)$ will depend on q.

2. *Properties of $F \to B_q(F)$: Upper semi-continuity.* Set $R(\mu, \hat{\mu}) = E_\mu|\hat{\mu}(X) - \mu|^q$ and $B(F, \hat{\mu}) = \int R(\mu, \hat{\mu})\, dF(\mu)$: since $\mu \to R(\mu, \hat{\mu})$ is continuous, $F \to B(F, \hat{\mu})$ is weakly continuous on $\mathcal{F}_p(\eta)$ when the risk function is *also* bounded. We recall that

$$B_q(F) = \inf_{\hat{\mu}} B(F, \hat{\mu}).$$

Since the infinum includes estimators with unbounded risk function we define an increasing family of subclasses of estimators $\mathcal{D}_m = \{\hat{\mu} : \hat{\mu}(x) = x \text{ for } |x| > m\}$, and let

$$B_{qm}(F) = \inf_{\hat{\mu} \in \mathcal{D}_m} B(F, \hat{\mu}).$$

Since each estimator in $\mathcal{D}_m$ has bounded risk, $F \to B_{qm}(F)$ is weakly upper semi-continuous (usc). Since $B_{qm}(F)$ decreases as $m \nearrow \infty$ it has a limit, $\tilde{B}_q(F)$ say, and if we assume for the moment that

$$B_q(F) = \tilde{B}_q(F) \tag{56}$$

then $B_q(F)$ is the decreasing limit of a family of usc functions and is hence also usc.

To verify (56), note first that trivially $B_q(F) \leq \tilde{B}_q(F)$. For the reverse inequality, observe that for any estimator $\hat{\mu}$ with finite integrated risk,

$$\int R(\hat{\mu}_m, \mu) dF \to \int R(\hat{\mu}, \mu) dF, \quad m \to \infty$$

where $\hat{\mu}_m \in \mathcal{D}_m$ is defined by

$$\hat{\mu}_m(x) = \begin{cases} \mu(x) & |x| \leq m, \\ x & |x| > m \end{cases}$$

[because $R(\hat{\mu}_m, \mu) \to R(\hat{\mu}, \mu)$ uniformly on compact intervals]. This establishes upper semicontinuity. Since $B_q(F)$ is the pointwise infinim of linear functions, it is concave, and $\mathcal{F}_p(\eta)$ is weakly compact because of the moment condition.

To verify that $\rho_q(F_{1+c}) \le (1+c)^q \rho(F)$, first let $\hat{\mu}_F$ denote the Bayes estimator of μ for prior F. Let $\phi = (1+c)\mu$, and suppose that $y|\phi \sim N(\phi, 1)$. We define a randomized estimator $\tilde{\phi}(y, z)$ based on y and an independent variate $Z \sim N(0, (2c+c^2)/(1+c^2))$:

$$\tilde{\phi}(y, z) = (1+c)\hat{\mu}_F((1+c)^{-1}y + z).$$

By construction, $W = (1+c)^{-1}Y + Z \sim N(\mu, 1)$, and so

$$\begin{aligned} r(\phi, \tilde{\phi}) &= E_\phi |\tilde{\phi}(Y, Z) - \phi|^q \\ &= (1+c)^q E|\hat{\mu}_F(W) - \mu|^q \\ &= (1+c)^q r(\mu, \hat{\mu}_F). \end{aligned}$$

By averaging over $M \sim F$, we obtain, as required,

$$\rho_q(F) \le Er(\Phi, \tilde{\phi}) = (1+c)^q Er(M, \hat{\mu}_F) = (1+c)^q \rho_q(F).$$

3. *Risk functions for soft and hard threshold rules.* For reference, we record explicit formulas for the risks of $\delta_\lambda^{(s)}$ and $\delta_\lambda^{(h)}$ when $\sigma = 1$. Write x for an $N(\mu, 1)$ variate and $r(\delta, \mu) = E_\mu|\delta(x) - \mu|^q$. Then for $\mu \ge 0$

$$\begin{aligned} r\left(\delta_\lambda^{(s)}, \mu\right) &= E_\mu|(x-\lambda)_+ + (x+\lambda)_- - \mu|^q \\ &= \int_{-\infty}^{-\lambda-\mu} |w+\lambda|^q \phi(w)\, dw + \mu^q \int_{-\lambda-\mu}^{\lambda-\mu} \phi(w)\, dw + \int_{\lambda-\mu}^{\infty} |w-\lambda|^q \phi(w)\, dw \end{aligned}$$

and

$$\frac{\partial}{\partial \mu} r(\delta_\lambda^{(s)}, \mu) = q\mu^{q-1}\left[\Phi(\lambda-\mu) - \Phi(-\lambda-\mu)\right] \ge 0,$$

so that the risk function increases monotonically on $[0, \infty)$ to a bounded limit. For hard thresholds,

$$r(\delta_\lambda^{(h)}, \mu) = \int_{-\infty}^{-\lambda-\mu} |w|^q \phi(w)\, dw + \mu^q \int_{-\lambda-\mu}^{\lambda-\mu} \phi(w)\, dw + \int_{\lambda-\mu}^{\infty} |w|^q \phi(w)\, dw,$$

but is no longer monotonic: indeed the risk function rises from $\mu = 0$ to a maximum at $\lambda - o(\lambda)$ (as $\lambda \nearrow \infty$) before decreasing to c_q as $\mu \nearrow \infty$.

For squared error loss ($q = 2$) more explicit expressions are available:

$$\begin{aligned} r(\delta_\lambda^{(s)}, \mu) &= 1 + \lambda^2 + (\mu^2 - \lambda^2 - 1)\left[\Phi(\lambda-\mu) - \Phi(-\lambda-\mu)\right] - (\lambda-\mu)\phi(\lambda+\mu) \\ &\quad - (\lambda+\mu)\phi(\lambda-\mu) \\ r(\delta_\lambda^{(h)}, \mu) &= 1 + (\mu^2 - 1)\left[\Phi(\lambda-\mu) - \Phi(-\lambda-\mu)\right] + (\lambda+\mu)\phi(\lambda+\mu) + (\lambda-\mu)\phi(\lambda-\mu). \end{aligned}$$

4. *Moment Inequality.* Let $F(dx)$ be a probability distribution on R and for $q \geq 1$, define $\mu_q(F)$, the *q-mean* of F, as any minimizer of $\int |x-\mu|^q F(dx)$. We have the inequality

$$|\mu_q|^q \leq 2^{q-1}[E|X|^q + E|\mu_q - X|^q] \leq 2^q E|X|^q. \tag{57}$$

Equation (37) follows by taking for F the posterior distribution of θ_i given x under the prior π_n. [A refined version of (57) appears in Johnstone (1991).]

5. *Structure of the linear minimax rule.*

Lemma 22 *Suppose that $\Theta \subset R^r$ is orthosymmetric. Let $\hat{\theta}_{abc,i}(x) = ax_i + bx_i' + c$, where $x_i' = \sum_{j \neq i} x_j$. Then*

$$\sup_{\Theta} R(\theta, \hat{\theta}_{abd}) \geq \sup_{\Theta} R(\theta, \hat{\theta}_{a00}).$$

where $R(\theta, \hat{\theta}) = \sum_\theta \sum_{i=1}^p |\hat{\theta}_i - \theta_i|^q$.

Proof. Consider first a single component and a fixed constant d. Convexity of the function $y \to |x+y|^q$ implies

$$\begin{aligned} 2|aX_1 - \theta_1|^q &\leq |aX_1 - \theta_1 + d|^q + |aX_1 - \theta_1 - d|^q \\ &= |aX_1 + d - \theta_1|^q + |a(-X_1) + d + \theta_1|^q. \end{aligned}$$

Let $\sigma = (\sigma_1, \ldots, \sigma_p)$ belong to $\{\pm 1\}^p \equiv \mathcal{Z}_2^p$. Since the components of X are independent, one can apply this argument conditionally on x' to obtain

$$2E|aX_1 - \theta_1|^q \leq \sum_{\sigma_1} E|a\sigma_1 X_1 + d - \sigma_1\theta_1 + b\sum_{j \geq 2} \sigma_j X_j|^q$$

Now let $\sigma' = (\sigma_2, \ldots, \sigma_p)$. Representing the random variables X_j explicitly in terms of the constituent errors ϵ_j and exploiting symmetry leads to

$$\begin{aligned} 2^p E|aX_1 - \theta_1|^q &\leq \sum_{\sigma_1}\sum_{\sigma'} E|a(\sigma_1\theta_1 + \epsilon_1) + d - \sigma_1\theta_1 + b\sum_{j\geq 2}(\sigma_j\theta_j + \epsilon_j)|^q \\ &= \sum_{\sigma} E_{\sigma\theta}|aX_1 + d + bX_1' - \sigma_1\theta_1|^q \\ &= \sum_{\sigma} E_{\sigma\theta}|\hat{\theta}_{abd,1} - \sigma_1\theta_1|^q. \end{aligned}$$

Now add over i to get

$$2^p R(\theta, \hat{\theta}_{a00}) \leq \sum_{\sigma} R(\sigma\theta, \hat{\theta}_{abd})$$

and so, using $\bar{R}$ to denote maximum risk over Θ,

$$2^p \bar{R}(\hat{\theta}_{a00}) \leq \sum_{\sigma} \bar{R}(\hat{\theta}_{abd}) = 2^p \bar{R}(\hat{\theta}_{abd}),$$

which completes the proof. ∎

6. *Convexity decompositions of loss functions and parameter spaces.*

Lemma 23 *Fix $(p,q) \in S$. The function $c \to F(c;p,q)$ of (46) has no more sign changes than $g(v)$ on $[0,\infty)$.*

We remark that sign changes are counted in the weak sense; whenever $F(c) = 0$, it is assigned a sign in such a way as to minimise the total number of sign changes (cf. Karlin, 1968 or Brown, Johnstone and McGibbon, 1981).

We first note that for $0 \le c < \infty$, the mapping $(p,q) \to F(c;p,q)$ is continuous on S. This is clear from (46), except possibly for $q \searrow 1$. That continuity holds here also is evident from the representations

$$\begin{aligned}(q-1)E_c|v|^{q-2} &= \int_0^\infty \frac{d}{dv}(v^{q-1})\left[\phi(v-c)+\phi(v+c)\right]dv \\ &= -\int_0^\infty v^{q-1}\frac{d}{dv}\left[\phi(v-c)+\phi(v+c)\right]dv \to 2\phi(c)\end{aligned}$$

and

$$E_c v|v|^{q-2} = \int_0^\infty v^{q-1}\left[\phi(v-c)-\phi(v+c)\right]dv \to 2\Phi(c)-1$$

as $q \searrow 1$. This continuity implies that we need only establish the sign behavior of $F(c;p,q)$ on the *interior* of S; in particular, we will assume that $q > 1$ henceforth.

The representation (47) is obtained by combining the following identities in accordance with (46).

$$\begin{aligned}E_c v|v|^{q-2} &= 2e^{-c^2/2}\int_0^\infty v^{q-1}\phi(v)\sinh cv\,dv, \\ cE_c|v|^{q-2} &= 2e^{-c^2/2}\int_0^\infty v^{q-2}\phi(v)c\cosh cv\,dv, \\ &= 2e^{-c^2/2}\int_0^\infty \left[v^{q-1}-(q-2)v^{q-3}\right]\phi(v)\sinh cv\,dv.\end{aligned}$$

Total positivity (or order 2) of $\sinh cv$ follows from the relation

$$\begin{vmatrix} \sinh cv & \frac{\partial}{\partial c}\sinh cv \\ \frac{\partial}{\partial v}\sinh cv & \frac{\partial^2}{\partial c\partial v}\sinh cv \end{vmatrix} = \frac{1}{2}\left[\sinh(2cv)-2cv\right] \ge 0$$

(cf. Karlin, 1968). In turn, it follows that the kernel $(c,v) \to v^{q-3}\phi(v)\sinh cv$ is TP_2 and the lemma is established by appealing to the variation diminishing property of totally positive kernels (cf. Karlin or Brown, Johnstone, McGibbon op. cit.).

7. To classify the sign change behavior of $g(v)$ for $(p,q) \in (0,\infty)\times(1,\infty)$ and $v \in [0,\infty)$ we find the following cases.

a) $p \le q \le 2$. g has no sign changes and is non-negative, so $s(\gamma)$ is convex.

b) $p \ge q \ge 2$. g has no sign changes and is non-positive, so $s(\gamma)$ is concave.

In the remaining cases, g has exactly one sign change, so the sign change behavior of $F(c)$ is determined by its limits at 0 and ∞. From (46), one sees that $F(c) \sim (q-p)c^{q-1}$ as $c \nearrow \infty$. To determine behavior at 0, we note that for $q > -1$

$$c_q = E|Z|^q = \frac{2^{q/2}}{\sqrt{\pi}}\Gamma\left(\frac{q+1}{2}\right), \text{ and } (q-1)c_{q-2} = c_q. \tag{58}$$

Substituting the expansion $\sinh cv = cv + (cv)^3/6 + \ldots$ into (47) yields

$$e^{c^2/2}F(c) \sim c\,[(q-p)c_q - (q-1)(q-2)c_{q-2}] + \tag{59}$$
$$+\frac{c^3}{6}\,[(q-p)c_{q+2} - (q-1)(q-2)c_q] + o(c^3)$$
$$\sim \begin{cases} (2-p)c_q\, c & p \neq 2 \\ (q-2)\frac{c_q}{3}c^3 & p = 2, q \neq 2 \end{cases} \tag{60}$$

[Of course, $F(c) \equiv 0$ if $p = q = 2$.]

c) $q > p$, $q > 2$. Here $F(\infty) > 0$. If $p \leq 2$, then $F(0+) > 0$ so that $s(\gamma)$ is convex on $[0, \infty)$. However, if $p > 2$, then $F(0+) < 0$, and so there exists a value $\gamma_0 = c_0^p = c_0^p(p,q)$ such that s is concave on $[0, \gamma_0]$ and convex on $[\gamma_0, \infty)$.

d) $q < p$, $q < 2$. Now $F(\infty) < 0$. If $p \geq 2$, then $F(0+) < 0$ also, so that $s(\gamma)$ is concave on $[0, \infty)$. However, if $p < 2$, then $F(0+) > 0$ and there exists γ_0 such that s in convex on $[0, \gamma_0]$ and concave on $[\gamma_0, \infty)$.

Putting a)–d) together yields the decomposition of Section 9.

8. *Sharpness of (51b) when* $\tilde{\eta}_n \to \eta \in (0, \infty)$; $\quad q > p > 2$.
Combining the equality (42) with the upper bound (43) we obtain

$$R_L^* \leq n\sigma^q (1 + b_*)^{-q} \tilde{s}_n(b_*^p \eta_n^p)$$

Since $\tilde{\eta}_n^q = n^{-1}(r/\sigma)^q \to \eta$ and $q > p$, it follows that $r/\sigma \to \infty$ and hence $\eta_n^p = n^{-1}(r/\sigma)^p \to 0$. Let $\bar{\gamma}_n = b_*^p \eta_n^p$. Since $s(\gamma)$ is concave for $\gamma \leq \gamma_0$ and convex for $\gamma \geq \gamma_0$, it follows that $\tilde{s}_n(\bar{\gamma}_n) = (1 - \epsilon_n)s(\gamma_n) + \epsilon_n s(n\bar{\gamma}_n)$ where ϵ_n and γ_n are determined by the equations

$$(1 - \epsilon_n)\gamma_n + \epsilon_n n\bar{\gamma}_n = \bar{\gamma}_n \tag{61}$$

$$s'(\gamma_n) = [s(n\bar{\gamma}_n) - s(\gamma_n)] \,/\, [n\bar{\gamma}_n - \gamma_n]\,. \tag{62}$$

[For these equations to be valid, we must have $\gamma_n < \bar{\gamma}_n$, but this is established below.]

Our goal is to show that

$$\tilde{s}_n(\bar{\gamma}_n) = (1 - \epsilon_n)s(\gamma_n) + \epsilon_n s(n\bar{\gamma}_n) \sim (1 - n^{-1})s(0) + n^{-1}s(n\bar{\gamma}_n), \tag{63}$$

for this would imply that $\mu = (r\sigma^{-1}, 0, \ldots, 0)$ is an asymptotically least favorable configuration. In turn, this implies that

$$R_L^* \leq n\sigma^q \left[(1 - n^{-1})E_0|a_*X|^q + n^{-1}E_{r\sigma^{-1}}|a_*X - r\sigma^{-1}|^q\right](1 + o(1))$$
$$\sim n\sigma^q f(a_*(\eta), \eta)$$

as is shown following (48).

To establish (63), one sees from (61) that it really suffices to show that $\gamma_n/\bar{\gamma}_n \to 0$, since $n\bar{\gamma}_n = b_*^p(r/\sigma)^p \to \infty$ and $s(\gamma) \sim \gamma^{q/p}$ as $\gamma \to \infty$. This last relation, together with

the approximation $s'(\gamma) \sim k_{p,q}\gamma^{2/p-1}$ as $\gamma \to 0$ (c.f. (44) and an argument similar to (60)) recasts (62) as the equation

$$k_1 \gamma_n^{\frac{2-p}{p}} = (n\bar{\gamma}_n)^{\frac{q-p}{p}}.$$

Expressing n and $n\bar{\gamma}_n$ in terms of σ/r, this leads to

$$\gamma_n/\bar{\gamma}_n = k_2(\sigma/r)^{2(q-p)/(p-2)} \to 0$$

as required. [k_i are constants whose values are unimportant here.]

9. *Minimax estimation in the concave case.*

According to (52),

$$R_L^* = n\sigma^q \inf_b w(b;\eta_n),$$

where $w(b) = w(b;\eta) = (1+b)^{-q}E|Z+b\eta|^q$ for a standard Gaussian variate Z. We verify that $b \to w(b)$ has a unique minimum on $[0,\infty)$. Introducing variables $c = b\eta$ and $v \sim N(c,1)$, one can verify that

$$r(b) \stackrel{\text{def}}{=} q^{-1}(1+b)^{q+1}w'(b) = E_c k(v) \qquad k(v) = |v|^{q-2}\{\eta v - (q-1)\}. \tag{64}$$

The function $k(v)$ has at most one (weak) sign change on $(-\infty,\infty)$. Since the Gaussian location family is TP_2, it follows that $w'(b)$ has at most a single sign change. However, since $r(0+) < 0$ and $r(\infty-) > 0$, we deduce that $w(b)$ has exactly *one* minimum $b_*(\eta)$, located in the *interior* of $[0,\infty)$.

Consider now the behavior of $b_*(\eta)$ as $\eta \to \infty$. To study the asymptotic behavior of equation (64), we note that

$$\begin{aligned} E_c v|v|^{q-2} &= 2e^{-c^2/2}\int_0^\infty v^{q-1}\phi(v)\sinh cv\, dv \\ &= 2e^{-c^2/2}\int_0^\infty v^{q-1}\phi(v)\left[cv + (cv)^3/6 + \cdots\right] dv \\ &= e^{-c^2/2}[c_q c + c_{q+2}c^3/6 + \cdots] \end{aligned}$$

and similarly

$$E_c c|v|^{q-2} = e^{-c^2/2}[c_{q-2}c + c_q c^3/2 + \ldots].$$

Substituting leading terms into (64) and using the identity (58) yields $c_*(\eta) \sim \eta^{-1}$ and hence $b_*(\eta) \sim \eta^{-2}$. Consequently $w(b_*(\eta),\eta) = (1+\eta^{-2})^{-q}E|Z+\eta^{-1}|^q \sim c_q$ as $\eta \nearrow \infty$.

Turn now to the contrary case in which $\eta \to 0$. Now for c large, $E_c k(v) \sim \eta c^{q-1} - (q-1)c^{q-2}$, and the unique root of the right side occurs at $c_* = (q-1)\eta^{-1}$. Thus, for small η, the unique minimum of $w(b)$ is to be found at $b_*(\eta) \sim (q-1)\eta^{-2}$, and $w(b_*) = (1+(q-1)\eta^{-2})^{-q}E|Z+(q-1)\eta^{-1}|^q \sim \eta^q$.

10. *Discreteness error in (53).*

Lemma 24 *Suppose that $s(\gamma)$ is non-negative and convex-concave on $[0,\infty)$. Let*

$$\begin{aligned} \tilde{s}(\tilde{\gamma}) &= \sup\{(1-\epsilon)s(0) + \epsilon s(\gamma) : \epsilon\gamma = \tilde{\gamma}\} \\ &= (1-\epsilon_0)s(0) + \epsilon_0 s(\gamma_0), \qquad \text{and} \\ \tilde{s}_n(\tilde{\gamma}) &= \sup\{(1-\epsilon)s(0) + \epsilon s(\gamma) : \epsilon\gamma = \tilde{\gamma}, n\epsilon \in N\} \end{aligned}$$

Then $0 \le \tilde{s}(\tilde{\gamma}) - \tilde{s}_n(\tilde{\gamma}) \le Cn^{-1}$, $C = s(\gamma_0) + \|s'\|_\infty \gamma_0$.

Proof. Let $\epsilon_1 = n^{-1}\lceil n\epsilon_0 \rceil$ and $\gamma_1 = \tilde{\gamma}/\epsilon_1 = \tilde{\gamma} n/\lceil n\epsilon_0 \rceil$. Then

$$\begin{aligned} \Delta = \tilde{s}(\tilde{\gamma}) - \tilde{s}_n(\tilde{\gamma}) &\le [(1-\epsilon_0)s(0) + \epsilon_0 s(\gamma_0)] - [(1-\epsilon_1)s(0) + \epsilon_1 s(\gamma_1)] \\ &= (\epsilon_1 - \epsilon_0)\,[s(0) - s(\gamma_0)] + \epsilon_1\,[s(\gamma_0) - s(\gamma_1)]\,. \end{aligned}$$

We note that $\epsilon_1 - \epsilon_0 < n^{-1}$, $s(0) < s(\gamma_0)$,

$$\epsilon_1(\gamma_0 - \gamma_1) \le \epsilon_1 \left[\frac{\tilde{\gamma}}{\epsilon_0} - \frac{n\tilde{\gamma}}{\lceil n\epsilon_0 \rceil}\right] \le \epsilon_1 \frac{\tilde{\gamma}}{\epsilon_0 \lceil n\epsilon_0 \rceil} = \frac{\gamma_0}{n},$$

so that

$$\Delta \le n^{-1} s(\gamma_0) + \|s'\|_\infty n^{-1} \gamma_0$$

as required.

11. *Convex-concave case.*
If we change variables to $\gamma = b^p \eta^p$, the function on the right side of (53) becomes

$$r(\gamma) = (1 + \eta^{-1}\gamma^{1/p})^{-q} \tilde{s}(\gamma),$$

and one calculates that

$$\rho(\gamma) \stackrel{\text{def}}{=} p(1 + \eta^{-1}\gamma^{1/p})^{q+1} r'(\gamma) = p(1 + \eta^{-1}\gamma^{1/p})\tilde{s}'(\gamma) - q\eta^{-1}\gamma^{1/p-1}\tilde{s}(\gamma).$$

We now verify that $\rho(\gamma)$ has exactly one sign change on $[0, \infty)$. For $\gamma \le \gamma_0$, substitution from (54) yields

$$\rho(\gamma) = pR + \frac{p-q}{\eta} R\gamma^{1/p} - \frac{q}{\eta} c_q \gamma^{1/p-1} \tag{65}$$

and in particular, $\rho(0) = -\infty$ and $\rho'(\gamma) > 0$ on $[0, \gamma_0]$. It follows from the discussion of the concave case (i.e., $(p,q) \in V$), that $\rho(\gamma)$ has at most one sign change on $[\gamma_0, \infty)$, and if a sign change occurs, then it is from negative to positive. Putting these observations together with the analyticity of $s(\gamma)$ on $(0, \infty)$, we conclude that $\rho(\gamma)$ has an isolated zero $\gamma_* = \gamma_*(p, q, \eta) \in (1, \infty)$. This zero $\gamma_* < \gamma_0$ exactly when $\rho(\gamma_0) > 0$, and by substituting the definition $R = [s(\gamma_0) - c_q]/\gamma_0$ into (65), one verifies that this occurs as described in (55).

References

[1] Bickel, P.J. Minimax estimation of the mean of a normal distribution when the parameter space is restricted. *Ann. Stat.* **9**, 1301-1309. (1981).

[2] Bickel, P. J. Minimax estimation of a normal mean subject to doing well at a point. In *Recent Advances in Statistics* (M. H.Rizvi, J. S. Rustagi, and D. Siegmund, eds.), Academic Press, New York, 511–528. (1983).

[3] Birgé, L. and Massart, P. Rates of convergence for minimum contrast estimators. Technical Report Université Paris VI. (1991)

[4] Brown, L.D. Admissible estimators, recurrent diffusions, and insoluble boundary value problems. *Annals of Mathematical Statistics* **42**, 855-903. (1971).

[5] Brown, L.D., Johnstone, I.M. and MacGibbon, K.B. Variation Diminishing Transformations: A direct approach to total positivity and its statistical applications. *J. Amer. Statist. Assoc.* **76** 824–832. (1981).

[6] Casella, G. and Strawderman, W.E. Estimating a bounded normal mean. *Ann. Stat.* **9**, 870-878. (1981).

[7] Daubechies, I. Orthonormal bases of compactly supported wavelets. *Comunications in Pure and Applied Mathematics*, **41**, 909–996. (1988).

[8] D. L. Donoho, Johnstone I.M., Hoch, J.C. and Stern A.S. Maximum Entropy and the Nearly Black Image. *J. Roy. Statist. Soc. B.* **54** 41 – 81 (with discussion) (1992).

[9] Donoho, D.L. and Johnstone, I.M. Minimax estimation via wavelets shrinkage. Technical Report, Department of Statistics, Stanford University. (1992).

[10] Donoho, D. L., Liu, R. C. and MacGibbon, K. B. Minimax risk over hyperrectangles, and implications. *Ann. Statist.*, **18**, 1416–1437. (1990).

[11] Donoho, D.L. and Liu, R.C Geometrizing Rates of Convergence, III. *Ann. Statist.* **19**, 668-701. (1991).

[12] Donoho, D.L. Gelfand n-widths and the method of least squares. Preprint. (1990).

[13] Feldman, I. Constrained minimax estimation of the mean of the normal distribution with known variance. *Ann. Statist.* **19**, 2259–2265. (1991).

[14] Huber, P.J. Robust Estimation of a Location Parameter. *Ann. Math. Statist.* **35**, 73-101. (1964).

[15] Huber, P.J. Fisher Information and Spline Interpolation. *Ann. Statist.* **2** 1029-1034. (1974).

[16] Ibragimov, I.A. and Hasminskii, R.Z. Nonparametric estimation of the value of a linear functional in a Gaussian white noise. *Theory of Probability and its Applications* **29**, 1-32. (1984)

[17] Ibragimov, I.A. and Hasminskii, R.Z. On density estimation in the view of Kolmogorov's ideas in approximation theory. Ann. Statist. **18**, 999-1010. (1990).

[18] Johnstone, I.M. A moment inequality for L_q estimation. *Statistics and Probability Letters* **12,** 289–290 (1991).

[19] Karlin, S. *Total Positivity.* Stanford University Press, Stanford. (1968).

[20] Lehmann, E.L. *Theory of Point Estimation.* Wiley, New York. (1983).

[21] Lindenstrauss, J. and Tzafiri, L. The Classical Banach Spaces, II. Springer, New York, (1979).

[22] Mallows, C.L. SIAM Problem 78.4. *SIAM Review* (1978).

[23] Meyer, Y. *Ondelettes.* Paris: Hermann. (1990).

[24] Nemirovskii, A.S., Polyak, B.T. and Tsybakov, A.B. Rate of convergence of non-paramtric estimates of maximum-likelihood type. Problems of Information Transmission, **21**, 258-272, (1985).

[25] Pinsker, M.S. Optimal Filtration of square-integrable signals in Gaussian White Noise. *Problems of Information Transmission* 120-133. (1980)

[26] Port, S. and Stone, C.J. Fisher Information and the Pitman estimation of a location parameter. *Annals of Statistics* **2**, 225-247. (1974).

[27] Rockafellar, R. T. *Convex Analysis.* Princeton University Press. Princeton, N.J. (1970).

[28] Sacks, J. and Strawderman. W.E. Improving on linear minimax estimates. *Statistical Decision Theory and Related Topics III* **2**, 287-304. Academic Press. (1982)

[29] Van de Geer, S. Estimating a regression function. Ann. Statist., **18**, 907-924.