Probab. Theory Relat. Fields 99, 277–303 (1994)

Probability Theory and Related Fields

Minimax risk over l_p-balls for l_q-error

David L. Donoho, Iain M. Johnstone

Department of Statistics, Stanford University, Stanford, CA 94305, USA

Received: 23 June 1992/In revised form: 13 November 1993

Summary. Consider estimating the mean vector θ from data $N_n(\theta, \sigma^2 I)$ with l_q norm loss, $q \geqq 1$, 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. We also give asymptotic evaluations of the minimax linear risk. Our results are basic to a theory of estimation in Besov spaces using wavelet bases (to appear elsewhere).

Mathematics Subject Classification (1985): 62C20, 62F12, 62G20

1 Introduction

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 $1 + \varepsilon$ for some (unknown) $\varepsilon > 0$ 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, et al. (1990) have shown that the Ibragimov–Hasminskii constant is not larger than 5/4. Moreover, Donoho and Liu (1991) have shown that even if Θ is convex but asymmetric, still $R_L^*/R_N^* < 5/4$–provided inhomogeneous linear etimators are allowed. It follows that for estimating a single linear functional, minimax linear estimates cannot be *dramatically* improved on in the worst case.

Some results for ℓ_2 error

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, et al. (1990) showed that if Θ is an l_p-body with $p \geqq 2$ then $R_L^*/R_N^* \leqq 5/4$, 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.* We begin by highlighting some conclusions for the case of ℓ_2-error, and give later a systematic description of more general results for ℓ_q-error. Let $\Theta_{p,n}$ denote the standard n-dimensional unit ball of l_p, i.e. $\Theta_{p,n} = \{\theta : \sum_1^n |\theta_i|^p \leqq 1\}$.

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

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

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 Sects. 7–9 in Donoho, et al. (1990), and the discussion below.

Our technique sheds some light on this phenomenon of Pinsker's. It shows

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

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

where $\lambda = \lambda(n, \sigma, p)$. Consider, for example, the case $p = 1$ and $\sigma = cn^{-1/2}$. Then, *on average* $|\theta_i| \leqq n^{-1}$. 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 estimate most coordinates as 0 and so 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* $0 < p < 2$, $n\sigma^p \to \infty$ *and* $\sigma^2 \log n\sigma^p \to 0$. *Let* $\lambda^2 = 2\log n\sigma^p$. *Then*

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

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

$$R_L^*/R_N^* = (1 + n\sigma^2)^{-1} n\sigma^p (2\log n\sigma^p)^{-1+p/2} \text{ as } n\sigma^p \to \infty . \tag{5}$$

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

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

The estimator $\theta_\lambda^{(s)}$ in (2) can be said to use *soft thresholding*, since it is continuous in y. An alternative *hard threshold* estimator is

$$\theta_{\lambda,i}^{(h)} = y_i I\{|y_i| > \lambda\} . \tag{6}$$

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 same* $\alpha > p - 1$.

General conclusions for l_q *loss functions*

Our general situation 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 $q \geqq 1$. 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 $q = \infty$. For $q = \infty$, see Korostelev (1991) and Donoho (1994).

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

$$R_N^* = R_{N,q}^*(\sigma; \Theta_{p,n}(r)) = \inf_{\hat{\theta}} \sup_{\Theta_{p,n}(r)} E_\theta \sum_1^n |\hat{\theta}_i - \theta_i|^q . \tag{7}$$

The subscript 'N' indicates that non-linear procedures $\hat{\theta}(y)$ are allowed in the infimum. Of course $\Theta_{\infty,n}(r)$ is the hypercube $\{\theta: |\theta_i| \leqq r \ \forall i\}$.

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 known 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 $\hat{\theta}(y) = y$. 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 $Z \sim N(0, 1)$.

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 is the average signal to noise ratio measured in the l_p-norm: $(n^{-1}\sum|\theta_i/\sigma|^p)^{1/p} \leqq n^{-1/p}(r/\sigma)$.

The asymptotics of R_N^* depend on a standard univariate Gaussian location problem in which $X \sim N(\mu, 1)$ and we estimate μ with 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 $\mathscr{F}_p(\eta)$ denote the class of probability measures $F(d\mu)$ satisfying the moment condition $\int |\mu|^p F(d\mu) \leqq \eta^p$ (for $p = \infty$, the support condition $\operatorname{supp} F \subset [-\eta, \eta]$). An important role is played by the largest Bayes risk over $\mathscr{F}_p$

$$\rho_{p,q}(\eta) = \sup_{\mathscr{F}_p} \rho_q(F) . \tag{8}$$

A distribution $F_{p,q} = F_{p,q}(\eta)$ maximising (8) 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 $\tilde{F}_{p,q,n} \in \mathscr{F}_p(\eta_n)$ such that $\rho_q(\tilde{F}_{p,q,n}) \sim \rho_{p,q}(\eta_n)$. We use ν_μ to denote Dirac measure at the point μ and $\hat{\theta}_N(y)$ to denote *an* asymptotically minimax rule, not necessarily unique.

Theorem 5 *Let $(p, q) \in (0, \infty] \times [1, \infty)$ and set $\eta_n = n^{-1/p}(r/\sigma)$. If either (i) $p \geqq q$ or (ii) $0 < p < q$ and $(\sigma/r)^2 \log n(\sigma/r)^p \to 0$, then*

$$R_N^* \sim n\sigma^q \rho_{p,q}(\eta_n) \quad \textit{as } n \to \infty . \tag{9}$$

In specific instances, more can be said:

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

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

3a. $\eta_n \to 0, p \geqq q$. $\quad R_N^* \sim n\sigma^q \eta_n^q$, $\quad \hat{\theta}_N(y) = 0$.

The two point distributions $\tilde{F}_n = (\nu_{-\eta_n} + \nu_{\eta_n})/2$ are asymptotically least favorable.

3b. $\eta_n \to 0, p < q$. *Let* $\lambda_n^2 = 2 \log n(\sigma/r)^p = 2 \log \eta_n^{-p}$.

$$\begin{aligned} R_N^* &\sim n\sigma^q \eta_n^p (2 \log \eta_n^{-p})^{(q-p)/2}, \quad \hat{\theta}_{N,i}(y) = \operatorname{sgn}(y_i)(|y_i| - \lambda_n \sigma)_+ \\ &\sim (2(\sigma/r)^2 \log n(\sigma/r)^p)^{(q-p)}. \end{aligned} \tag{10}$$

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

$$\varepsilon\mu^p = \eta_n^p \textit{ and } \phi(a_n + \mu) = \varepsilon\phi(a_n) \tag{11}$$

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

When $p = \infty$, the minimax risk over a hypercube separates into the product of n univariate minimax problems: $R_N^* = n\sigma^q \rho_{\infty,q}(1, \eta_n)$, $\eta_n = r/\sigma$. (see e.g. Donoho, et al. (1990)), and Theorem 5 follows from asymptotics for $\rho_{\infty,q}$ (Theorem 15).

For $p < \infty$, the proof of Theorem 5 is the subject of Sect. 2 through 5. For $p < q$, the asymptotically minimax estimators given in (10) 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 Sect. 4, these Bayes estimators approximately have the form

$\delta_{\tilde{F}_n}(x) = \mu_n \operatorname{sgn}(x) I\{|x| > \mu_n + a_n\}$, where $a_n = o(\mu_n)$ and $\mu_n^2 \sim \lambda_n^2 \sim 2\log \eta_n^{-p}$. It follows that $\tilde{\theta}_{N,i}(y) = \sigma\delta_{\tilde{F}_n}(\sigma^{-1} y_i)$ has approximately the same zero set as the simpler threshold rule $\hat{\theta}_{N,i}(y)$. Hard threshold rules of the form (6) are also asymptotically minimax in the setting (3b) 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 $\alpha > p - 1$. Note that all of these asymptotically minimax estimators act co-ordinatewise: the estimate of θ_i depends only on y_i.

The threshold estimators of the previous section have a more general asymptotic near optimality property that holds whenever (9) 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) $p \geqq q$ or (ii) $0 < p < q$ and $(\sigma/r)^2 \log n(\sigma/r)^p \to 0$, then*

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

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

The theorem is proved in Sect. 6, 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 $p \geqq 2$ and computational experiments indicate that $\Lambda_s(1, 2) \leqq 1.6$.

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

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

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

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

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

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

where $\tilde{s}(\gamma)$ is the least concave majorant of $s(\gamma) = E|Z + \gamma^{1/p}|^q$ on $[0, \infty)$. (When $p = \infty$, replace $\tilde{s}(b^p\eta^p)$ by $E|Z + b\eta|^q$.) If $b_*(\eta)$ attains the minimum in $c_{p,q}(\eta)$, then

$\hat{\theta}_{L,i}(y) = y_i/(1 + b_*(\eta))$.

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 Sect. 7. In the special case of squared error loss, $q = 2$, we have the explicit evaluation

$$R_L^* = n\sigma^2 \bar{\eta}_n^2/(1 + \bar{\eta}_n^2) \,. \tag{12}$$

In conjunction with Theorem 5, this establishes Theorems 1 and 3.

The following corollary describes the possible limiting behaviors for R_L^*/R_N^*, and incidentally includes Theorem 2. 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*

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

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.

Discussion and remarks

1. Constraints on the l_p norm of θ arise in various scientific contexts. Hypercube constraints ($a \leqq \theta_i \leqq b$) 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: only a small number of components can be significantly non-zero. Formally

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

Donoho et al. (1992) use methods of this paper and the latter "nearly-black" condition to study behavior of non-linear estimation rules such as maximum entropy.

2. Some important earlier works exhibit function classes over which non-linear estimators have dramatically better worst-case performance than the best linear estimators in global norms. Nemirovskii et al. (1985) show that non-parametric

M-estimates (including constrained least squares) achieve faster rates 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 $1 \leqq p < 2$. For related results see van de Geer (1990) and Birgé, Massart (1991).

Our assumptions of highly symmetric parameter spaces and Gaussian white noise are very restrictive. However this symmetry permits reduction to simple one-dimensional estimation problems and 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).) Indeed, the basic dichotomy between $p < q$ and $p \geqq q$, as expressed in the loss-convexity condition, appears already in our very simple setting.

3. However, the idealised considerations of this paper lead to a theory of estimation over a wide class of (Besov) function spaces. These include 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, 1992)), 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. See especially Donoho, Johnstone (1995a, 1995b, 1993) and Donoho et al. (1995, 1993).

4. Johnstone (1994) describes analogues of the results of this paper for weak (or Marcinkiewicz) ℓ_p-balls $\Theta^*_{n,p}(r) = \{\theta : k^{1/p}|\theta|_{(k)} \leqq r, k = 1, \ldots, n\}$ where $|\theta|_{(1)} \geqq |\theta|_{(2)} \ldots \geqq |\theta|_{(n)}$ are order statistics of $|\theta_i|$. These sets arise naturally as approximation spaces for non-linear estimation methods such as piecewise polynomials and free dyadic splines.

5. An alternative approach to some of the results of this paper follows from the use of oracle inequalities – see Donoho and Johnstone (1993).

Outline of the paper

Section 2 introduces the main tool for evaluating R_N^*, namely a related Bayes-minimax risk R_B^* in which Θ is enlarged to a set of prior distributions satisfying the same moment constraint as members of Θ. Thus $R_N^* \leqq R_B^*$, but R_B^* can be reduced to the *univariate* minimax risk $\rho_{p,q}(\eta)$ in (8). Minimax rules for $\rho_{p,q}(\eta)$ do not have a simple explicit form, so Sect. 3 looks at how well they may be approximated by univariate forms of the soft and hard threshold estimators (2) and (6). In particular the minimax choice of threshold λ is found as $\eta \to 0$ and its corresponding risk evaluated (Proposition 13). Section 4 evaluates the small η behaviour of $\rho_{p,q}$, using two and three point approximately least favorable priors for lower bounds and the threshold results of Sect. 3 for upper bounds. This completes the evaluation of R_B^* and Sect. 5 finishes the proof of Theorem 5 by establishing conditions under which the upper bound $R_N^* \leqq R_B^*$ is asymptotically an equality. The brief Sect. 6 lifts the inefficiency bounds on univariate thresholds of Sect. 3 to the n-dimensional setting and establishes Theorem 6. Section 7 is devoted to linear minimax rules and Theorem 7, while the Appendix collects proof details.

2 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

$$R_B^*(\sigma, \Theta_{n,p}(r)) = \inf_{\hat{\theta}} \sup_{\pi} \left\{ E_\pi E_\theta \| \hat{\theta} - \theta \|_q^q, \text{ for } \pi\colon E_\pi \sum_1^n |\theta_i|^p \leqq r^p \right\} . \tag{13}$$

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

$$R_N^* \leqq R_B^* .$$

Further discussion of this Bayes-minimax approach is in Donoho, Johnstone (1995a) and Johnstone (1994).

In this section, we give a simpler description of R_B^* in terms of a univariate estimation problem. 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. In view of the permutation invariance of the problem, consider 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 $\{z_i\}$,

$$\begin{aligned} E_\pi E_\theta \| \delta^n - \theta \|_q^q &= \sum_i \int E_{\theta_i} |\delta(y_i) - \theta_i|^q \pi_i(d\theta_i) \\ &= \int E_{\theta_1} |\delta(y_1) - \theta_1|^q (\textstyle\sum \pi_i)(d\theta_1) \\ &= n E_{F_\pi} E_{\theta_1} |\delta(y_1) - \theta_1|^q \end{aligned} \tag{14}$$

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, as $E_{F_\pi} |\theta_1|^p \leqq n^{-1} r^p$, since

$$E_\pi \sum_i |\theta_i|^p = \sum_i \int |\theta_i|^p \pi_i(d\theta_i) = n \int |\theta_1|^p F_\pi(d\theta_1) . \tag{15}$$

Now define a univariate Bayes-minimax problem for data $y_1 \sim N(\theta_1, \sigma^2)$ with p^{th} moment constraint τ: let $\mathscr{F}_p(\tau)$ denote the collection of distributions F on $\mathscr{R}$ satisfying $\int |\mu|^p F(d\mu) \leqq \tau^p$, and set

$$\rho(\tau, \sigma) = \rho_{p,q}(\tau, \sigma) = \inf_\delta \sup_F \{ E_F E_{\theta_1} |\delta(y_1) - \theta_1|^q \colon F \in \mathscr{F}_p(\tau) \} . \tag{16}$$

The point is that the n-variate problem (13) is no harder than n copies of (16).

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

Proof. Let (F^0, δ^0) be a saddlepoint for the univariate problem (16): that is, δ^0 is a minimax rule, F^0 is a least favorable prior distribution and δ^0 is Bayes for F^0. Let

F^{0n} denote the n-fold cartesian product measure derived from F^0: from (15) and (14), it satisfies the moment constraint for $R_B^\ddagger$, and

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

We need to verify that (F^{0n}, δ^{0n}) is a saddlepoint for $R_B^\ddagger$, which amounts to showing

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

But (14) and (15) reduce this to the saddlepoint property of (F^0, δ^0). ▌

Properties of $\rho(\sigma, \tau)$

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 $\mathscr{F}_p(\eta)$. If $F_a(d\mu) = F(a^{-1}d\mu)$, then $\rho_q(F_a) \leqq a^q \rho_q(F)$ for $a \geqq 1$. (see Appendix, Sect 8.1) If we set $\rho(F, \delta) = E_F E_\mu | \delta(x) - \mu |^q$, the minimax theorem (e.g. Sion (1958, Theorem 4.2′), LeCam (1986, p. 16)) provides a minimax rule δ_τ such that

$$\rho(\tau, \sigma) = \sup_{\mathscr{F}_p(\tau)} \rho(F, \delta_\tau) = \inf_\delta \sup_{\mathscr{F}_p(\tau)} \rho(F, \delta) = \sup_{\mathscr{F}_p(\tau)} \rho(F) . \tag{17}$$

Let F_τ be a distribution maximizing $\rho(F)$ over $\mathscr{F}_p(\tau)$. Since $\rho(F_\tau, \delta_\tau) \leqq \rho(\tau, \sigma) = \rho(F_\tau, \delta_{F_\tau})$, it follows from the essential uniqueness of Bayes rules (see Appendix of Donoho and Johnstone 1992) that $\delta_\tau = \delta_{F_\tau}$ and hence that the pair (F_τ, δ_τ) is a saddlepoint for $\rho(F, \delta)$.

Proposition 10 *The function $\rho(\tau, \sigma)$ is continuous, monotone increasing in τ, concave in τ^p and converges to $\sigma^q c_q$ as $\tau/\sigma \to \infty$. It satisfies*

$$\rho(\tau, \sigma) = \sigma^q \rho(\tau/\sigma, 1), \textit{(invariance)}$$

$$\rho(a\tau, \sigma) \leqq a^q \rho(\tau, \sigma), \; a \geqq 1 .$$

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 $t = \tau^p$, $\tilde{\mathscr{F}}(t) = \{F: \int |\mu|^p dF \leqq t\}$, and hence $\tilde{\rho}(t) = \rho(\tau) = \sup\{\rho(F): F \in \tilde{\mathscr{F}}_p(t)\}$. Concavity (and hence continuity) follows immediately, because $(1-\varepsilon)F_1 + \varepsilon F_2 \in \tilde{\mathscr{F}}((1-\varepsilon)t_1 + \varepsilon t_2)$ whenever $F_i \in \tilde{\mathscr{F}}_p(t_i)$ $i = 1, 2$.

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 $\lim_{\tau\to\infty} \rho(\tau) \geqq \lim_{\sigma\to\infty} \rho(\Phi_\sigma)$ 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 $q \geqq 1$, and a simple calculation using the formula (44) for linear rules in Sect. 7 shows that $\lim_\sigma \rho(\Phi_\sigma) = c_q$. ▌

ℓ_2 loss and optimization of Fisher information

If $I(G) = \int (g'(x))^2/g(x)dx$ denotes the Fisher information for a distribution with absolutely continuous density g, and Φ the standard Gaussian cumulative, then Brown's (1971) identity states

$$\rho_2(F) = 1 - I(F * \Phi) . \tag{18}$$

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

$$I_p(\tau) = \inf\{I(F * \Phi): F \in \mathscr{F}_p(\tau)\} . \tag{19}$$

Some additional properties of $\rho_{p,2}(\tau, \sigma)$ flow from (18). For example, since the density of $\Phi * F$ must be strictly positive on the whole real line, an argument of Huber (1964, 1974) shows the solution to (19) is unique. Call this (unique) least favorable distribution $F_{p,\tau}$.

Recall the well-known inequality $I(F)\mathrm{Var}(F) \geqq 1$, with equality only at the Gaussian. This implies that for $p = 2$ we have

$$I_2(\tau) = (1 + \tau^2)^{-1} \tag{20}$$

and that the solution $F_{2,\tau}$ is the Gaussian distribution $N(0, \tau^2)$. Indeed, it may further be shown that $F_{p,\tau}$ is Gaussian only if $p = 2$ (For integer p, see Feldman (1991)). Further, $\rho_{2,2}(\tau) = \tau^2(1 + \tau^2)^{-1}$, which equals the minimax linear risk $\inf_{a,b} \sup \{E_\theta(ax + b - \mu)^2 \colon |\theta| \leqq \tau\}$.

When $p \to \infty$, we get $I_\infty(\tau) = \inf\{I(\Phi * F)\colon \mathrm{supp}(F) \in [-\tau, \tau]\}$ which 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). From the latter paper follows

$$\rho_{2,2}(\tau)/\rho_{\infty,2}(\tau) \leqq \mu^* = 1.25 .$$

3 Univariate threshold rules

In this section, we study two families of threshold estimators 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 $E_F|\mu|^p \leqq \eta^p$. These threshold estimators 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:

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

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

$$\rho_s(\eta, \sigma) = \inf_\lambda \sup \{E_F E_\mu |\delta_\lambda^{(s)}(y) - \mu|^q \colon E_F|\mu|^p \leqq \eta^p\} \tag{21}$$

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

$$\rho_s(\eta, \sigma) = \sigma^q \rho_s(\eta/\sigma, 1), \quad \rho_h(\eta, \sigma) = \sigma^q \rho_h(\eta/\sigma, 1) \tag{22}$$

again ensure that it suffices to assume $\sigma = 1$. As 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.

To measure how much is lost relative to Bayes-minimax estimators, define $\Lambda_s(p, q)$ (and $\Lambda_h(p, q)$) by

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

Theorem 11 *For* $(p, q) \in (0, \infty) \times [1, \infty)$, $\Lambda_s(p, q) < \infty$ *and* $\Lambda_h(p, q) < \infty$.

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

Theorem 12 $\dfrac{\rho_s(\eta, 1)}{\rho(\eta, 1)}$ *and* $\dfrac{\rho_h(\eta, 1)}{\rho(\eta, 1)} \to 1$ *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 $\rho_h(\eta, 1) \leqq c_q = \lim \rho(\eta, 1)$. (Proposition 10).

For the limit as $\eta \to 0$, we compute upper bounds for threshold rules in this section. [Of course, these also provide bounds for $\rho(\eta, 1)$.] In the next section, separate arguments are used to provide lower bounds (Theorem 15) for $\rho(\eta, 1)$ that agree asymptotically with the upper bounds and so complete the proof of Theorem 12 and so of Theorem 11.

When $q \leqq p$, upper bounds are straightforward. Choosing $d = 0$ (i.e. $\lambda = \infty$) gives the upper bound $\rho_q(F) \leqq E_F|\mu|^q = |F|_q^q$. If $q \leqq p$, and $F \in \mathscr{F}_p(\eta)$, then $|F|_q \leqq |F|_p \leqq \eta$, and

$$\max\{\rho_s(\eta, 1), \rho_h(\eta, 1)\} \leqq \sup_{\mathscr{F}_p} E_F|\mu|^q \leqq \eta^q . \tag{23}$$

When $q > p$, the least favorable F in $\mathscr{F}_p(\eta)$ have the form $F_{\varepsilon,\mu} = (1 - \varepsilon)\nu_0 + \varepsilon(\nu_\mu + \nu_{-\mu})/2$ where $\mu \sim (2 \log \eta^{-p})^{1/2}$ and $\varepsilon \sim (\eta/\mu)^p$ (cf. Proposition 16). The minimax thresholds then are of order $\lambda(\eta) \sim (2 \log \eta^{-p})^{1/2}$.

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. (Sect. 8.2) 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 $1 + \lambda^2$, whereas the risk of $\delta_\lambda^{(h)}$ rises from $\mu = 0$ roughly like $\mu \to \mu^2$ to a maximum at $\lambda - o(\lambda)$ (as $\lambda \to \infty$) before decreasing (sharply) 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 $\mathscr{F}_p(\eta)$ is

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

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

Proposition 13 *Suppose that* $p < q$ *and let* $\lambda = \lambda(\eta)$ *be chosen such that*

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

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

$$\bar{r}(\delta_\lambda, \eta) \sim \eta^p \lambda^{q-p},$$

$$\rho_s(\eta, 1), \rho_h(\eta, 1) \leqq \eta^p (2 \log \eta^{-p})^{q-p}(1 + o(1)) \textit{ as } \eta \to 0 .$$

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 $n\varepsilon$, where $n\varepsilon(\lambda\sigma)^p = 1$. Thus we consider a univariate hypothesis testing problem with $Y \sim N(\theta, \sigma)$, with H_0: $\theta = 0$ and H_1: $\theta = \lambda\sigma$ and assign prior probabilities $\pi(H_0) = 1 - \varepsilon$ and $\pi(H_1) = \varepsilon$. The estimator θ_λ is related to the decision rule $\phi(y) = I\{|y| > \lambda\sigma\}$. Let P_π denote the joint distribution of θ and Y. Let us choose λ to minimise the maximum of the two error probabilities of the decision rule ϕ, namely $e_0 = P_\pi(H_0, |y| > \lambda\sigma)$, and $e_1 = P_\pi(H_1, |y| < \lambda\sigma)$. Since $e_0(\lambda)$ is decreasing and $e_1(\lambda)$ is increasing, the minimax choice of λ is found by equating the two; i.e. approximately by solving $(1-\varepsilon)(1-\Phi)(\lambda) = \varepsilon/2$. Since λ will turn out to be large, we may use the standard approximation $1 - \Phi(\lambda) \sim \phi(\lambda)/\lambda$, and solve the equation $\phi(\lambda) = \lambda\varepsilon/2$. If $p = 1$, the constraint $n\varepsilon\lambda\sigma = 1$, combined with the definition $\eta^{-1} = n\sigma$ implies that $\lambda^2 \approx 2\log\eta^{-1} + 2\log\sqrt{2/\pi}$. For general p, the equation for λ becomes approximately

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

where we have used the constraints $\varepsilon\lambda^p = n^{-1}\sigma^{-p} = \eta^p$. This has 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, $r(\delta_\lambda^{(h)}, 0) = 2[\lambda\phi(\lambda) + \tilde{\Phi}(\lambda)] \gg 4\lambda^{-3}\phi(\lambda) = r(\delta_\lambda^{(s)}, 0)$. 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 14 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 $\mathscr{F}_p^+(\eta)$ are two point distributions $F = (1-\varepsilon)\nu_{a_0} + \varepsilon\nu_{a_1}$ for which $(1-\varepsilon)a_0^p + \varepsilon a_1^q = \eta^p$. For such distributions,

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

The first term turns out to be negligible, regardless of the choice of ε and a_0 (see Lemma 14(a)), so we study the function

$$s(\mu) = \left(\frac{\eta}{\mu}\right)^p r(\mu) \quad \mu \geqq \eta\,.$$

A simpler approximation to $r(\mu)$ which is adequate for calculation (see Lemma 14(b)) is given by 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 $|\lambda^2 - 2\log\eta^{-p}| \leqq c_0$ for some $c_0 > 0$. By comparing coefficients of λ^q in $\mu^{p+1}\eta^{-p}s'_+(\mu)$, it turns out that $s'_+(\mu)$ has a zero at approximately $\mu_{pq} = \lambda + \tilde{\Phi}^{-1}(p/q) = \lambda + z_{pq}$, say. Calculation shows that

$$s(\lambda + z_{pq}) \sim \eta^p\lambda^{q-p} \text{ as } \eta \to 0\,. \tag{24}$$

For hard thresholds, one finds that the zero of $s_+(\mu)$ occurs at approximately $\mu_{pq} = \lambda - (2\log\lambda c_0^{-1})^{1/2}$ with $c_0 = (q-p)\sqrt{2\pi}$, and (24) remains true for $s(\mu_{pq})$.

To complete the outline for Proposition 13, we now collect the steps required to show that (24) maximises $s(\mu)$. The proof is derived by detailed calculus and analysis from the risk formulas (60) and (62) in the Appendix. Details are left to the determined reader.

Lemma 14 *Suppose that $p < q$ and that $\lambda = \lambda(\eta)$ is chosen as in Proposition 13.*

(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 $p < q$ and $0 \leqq a_0 \leqq \eta$, and $\lambda = \lambda(\eta)$ as specified in Proposition 13, then $r(\delta_\lambda, a_0) \leqq r(\delta_\lambda, \eta) \sim r(\delta_\lambda, 0) = o(\eta^p \lambda^{q-p})$. 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* $[0, \infty)$, $0 \leqq \delta(\mu) \leqq \delta(0) = r_+(0) = r(0)/2 = o(\eta^p \lambda^{q-p})$.

(c) *For sufficiently large $d_0 > |z_{pq}|$ (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 $[\lambda - (2\log \lambda c_1^{-1})^{1/2}, \lambda - (2\log \lambda c_2^{-1})^{1/2}]$).*

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

4 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 $p \geqq q$, 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 $p \geqq q$. 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 $\pm\, \mu(\eta)$ defined following Proposition 16 below. In addition, $\rho_{p,q}(\eta)$ has a slower rate of convergence.

Theorem 15 *As $\eta \to 0$*

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

Proof. Upper Bounds. The Bayes risk $\rho_q(F)$ is the minimal value of $E_F|d(x) - \mu|^q$ over all estimators d. When $p \geqq q$, choosing $d = 0$ and arguing as for (23) gives $\rho_{p,q}(\eta) \leqq \eta^q$.

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 13, we obtain

$$\rho_{p,q}(\eta) = \sup_{\mathscr{F}_p(\eta)} \inf_{\delta} R(\delta, F) \leqq \sup_{\mathscr{F}_p(\eta)} R(\delta_\lambda, F) = \eta^p(2\log \eta^{-p})^{(q-p)/2}(1 + o(1)).$$

Lower bounds. It suffices to evaluate $\rho_q(F)$ for distributions F approximately least favorable for $\mathscr{F}_p(\eta)$. 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

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

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

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

Proposition 16 *Let* $F_{\varepsilon,\mu} = (1-\varepsilon)\nu_0 + \varepsilon(\nu_\mu + \nu_{-\mu})/2$. *Fix* $a > 0$, *and for all sufficiently small* ε, *define* $\mu = \mu(\varepsilon)$ *by*

$$\phi(a+\mu) = \varepsilon\phi(a) \tag{25}$$

Then

$$\rho_q(F_{\varepsilon,\mu}) \sim \varepsilon\mu^q\Phi(a) \quad as\ \varepsilon \to 0 . \tag{26}$$

Before proving Proposition 16, we use it to complete the proof of Theorem 15. Clearly $|F_{\varepsilon,\mu}|_p^p = \varepsilon\mu^p$, while from (25) it follows that $\mu(\varepsilon) \sim (2\log\varepsilon^{-1})^{1/2}$. If we connect η and ε by the relation $\eta^p = \varepsilon\mu^p$, then $F_{\varepsilon,\mu}$ belongs to $\mathscr{F}_p(\eta)$ and so from (26)

$$\rho_{p,q}(\eta) \geqq \rho_q(F_{\varepsilon,\mu}) \sim \varepsilon\mu^p.\mu^{q-p}\Phi(a) \sim \eta^p(2\log\eta^{-p})^{(q-p)/2}\Phi(a) \quad \text{as } \eta \to 0 . \tag{27}$$

The lower bound needed for Theorem 15 follows by taking a large.

Proof of Proposition 16. Let $d_F(x)$ denote the Bayes rule for estimation of τ from data $x \sim N(\tau, 1)$ and prior distribution $F_{\varepsilon,\mu}(d\tau)$. Since the posterior distribution of τ given x is concentrated on $\{0, \pm\mu\}$, we may write $d_F(x) = \mu e_{q,\varepsilon}(x)$, where $|e_{q,\varepsilon}(x)| \leqq 1$ and in addition $e_{q,\varepsilon}(x)$ is an odd function of x. Thus the Bayes risk

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

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

$$\int_0^\infty |e_{q,\varepsilon}(x)|^q\phi(x)dx = o(\varepsilon), \text{ and} \tag{28}$$

$$e_{q,\varepsilon}(\mu + z) \to I\{z > a\} . \tag{29}$$

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

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

Substituting definition (25) for ε, we find that $x_0 = a + \mu + \mu^{-1}\log 2(1-\varepsilon) + o(1)$. The integral in (28) is thus bounded by $\tilde{\Phi}(x_0) \leqq \tilde{\Phi}(a+\mu) \leqq \phi(a+\mu)/(a+\mu) = o(\varepsilon)$ from definition (25). Relation (29) 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 $0 \leqq d_F(x) \leqq \mu$, and differentiation shows that $d_F(x)$ is the solution of the equation

$$\varepsilon(\mu - a)^{q-1}p_+ = 2(1-\varepsilon)a^{q-1}p_0 + \varepsilon(\mu + a)^{q-1}p_- \tag{30}$$

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

Using (25) one verifies that for $x > 0$ and μ large, $\varepsilon(\mu + a)^{q-1}p_- < \varepsilon a^{q-1}p_0$ and hence that $d_F(x) \in [d_{\varepsilon/2,\varepsilon}(x), d_{\varepsilon,\varepsilon}(x)]$, where $d_{\delta,\varepsilon}(x)$ is the solution in a of the simpler equation

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

Using (25), $p_0/\varepsilon p_+ = \phi(x)/\varepsilon\phi(x-\mu) = e^{-\mu(x-\mu-a)}$, and thus

$$d_{\delta,\varepsilon}(x) = \mu[1 + 2^p(1-\delta)^\beta e^{-\mu\beta(x-\mu-a)}]^{-1}, \; \beta = 1/(q-1) \, . \tag{31}$$

Substituting $z = x - \mu - a$ and using (25), the integral in (28) is bounded above by

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

–consider separately positive and negative z. Finally, (29) follows from (31). ▌

5 Asymptotic sharpness of Bayes-minimax bound

This section shows that the bound $R_N^* \leqq R_B^*$ of Sect. 2 is often asymptotically an *equality*: nothing is lost by replacing the n-variate problem by n univariate problems.

Theorem 17 *If either (i)* $p \geqq q$, *or (ii)* $0 < p < q$ *and* $(\sigma/r)^2 \log n(\sigma/r)^p \to 0$, *then*

$$R_N^*(\sigma; \Theta_{p,n}(r)) = R_B^*(\sigma; \Theta_{p,n}(r))(1 + o(1)) \, . \tag{32}$$

Proof. 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, $i = 1, \ldots, n$, 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 $\Theta_n \overset{\text{def}}{=} \Theta_{p,n}(r)$: thus $\pi_n(A) = P_n(A \mid \theta \in \Theta_n)$. Clearly,

$$\frac{R_N^*}{R_B^*} \geq \frac{\rho(\pi_n)}{\rho(P_n)} r_n(F_n) \, , \tag{33}$$

and the idea is to show that $\rho(\pi_n)/\rho(P_n) \geqq 1 + o(1)$ 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 $\mathscr{E}_\pi|\hat{\theta} - \theta|^q$: 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

$$\rho(P_n) \leqq \mathscr{E}_{P_n}|\hat{\theta}_{\pi_n} - \theta|^q \tag{34}$$

$$= \mathscr{E}_{P_n}\{|\hat{\theta}_{\pi_n} - \theta|^q \mid \Theta_n\} P_n(\Theta_n) + \mathscr{E}_{P_n}\{|\hat{\theta}_{\pi_n} - \theta|^q, \Theta_n^c\} \tag{35}$$

$$\leqq \rho(\pi_n)P_n(\Theta_n) + 2^q\mathscr{E}_{P_n}\{|\hat{\theta}_{\pi_n}|^q + |\theta|^q; \Theta_n^c\} \, . \tag{36}$$

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 (36) is negligible relative to $\rho(P_n)$.

Case (*a*). Assume first that $\eta_n \to \eta \in (0, \infty]$. Choose $\varepsilon > 0$ and a sequence of distributions $F_{(k)}(d\mu) \in \mathscr{F}_p(\eta - \varepsilon)$ such that $\rho(F_{(k)}) \to \rho(\eta - \varepsilon)$ and $\operatorname{supp} F_{(k)} \subset [-k, k]$. Now fix k and let $F_n = F_{(k)}$ for all n. Now $P_n\{\theta \in \Theta_n\} = P_n\{n^{-1}\sum_1^n |\mu_i|^p \leqq \eta_n^p\} \to 1$ since $E|\mu|^p \leqq (\eta - \varepsilon)^p < \eta^p = \lim \eta_n^p$. Since $\operatorname{supp} F_{(k)} \subset [-k, k]$,

$$\mathscr{E}_{P_n}\{|\hat{\theta}_{\pi_n}|^q + |\theta|^q; \Theta_n^c\} \leqq 2n\sigma^q k^q P(\Theta_n^c) = o(\rho(P_n))\,,$$

since $\rho(P_n) = n\sigma^q \rho(F_{(k)})$. Thus $\rho(\pi_n)/\rho(P_n) \geqq 1 + o(1)$, 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 $p \geqq q$. The priors $F_{\eta_n} = (\nu_{\eta_n} + \nu_{-\eta_n})/2$ are asymptotically least favorable (Sect. 4), so $r_n(F_{\eta_n}) \to 1$. In addition, P_n is already supported on $\Theta_{p,n}(r)$ so $\pi_n = P_n$ and the equivalence (32) follows immediately from (36).

Case (*c*). Finally, if $p < q$, and $\eta_n \to 0$, we use the symmetric three point priors $F_{\varepsilon,\mu}$ studied in Proposition 16. Fix $\delta, a > 0$, and define $\varepsilon = \varepsilon_n$ implicitly by the relation

$$\varepsilon\mu^p = (1 - \delta)\eta_n^p = (1 - \delta)n^{-1}(r/\sigma)^p\,. \tag{37}$$

($\mu = \mu(\varepsilon, a)$ is already defined by Eq. (25)). Let $F_n = F_{\varepsilon_n, \mu_n}$. From Proposition 16 and (27),

$$\rho(F_n) \sim \varepsilon\mu^q\Phi(a) \sim (1 - \delta)\eta_n^p(2\log\eta_n^{-p})^{(q-p)/2}\Phi(a)\,, \tag{38}$$

while from Theorem 15, $\rho_{p,q}(\eta) \sim \eta^p(2\log\eta^{-p})^{(q-p)/2}$. Thus $r(F_n) \sim (1 - \delta)\Phi(a)$. Let $N_n \sim$ Binomial (n, ε) count the number of non-zero μ_i: (37) implies that $EN_n = n\varepsilon = (1 - \delta)(r/\sigma)^p\mu^{-p}$. Thus

$$\Theta = \{\textstyle\sum|\mu_i|^p \leqq (r/\sigma)^p\} = \{N_n \leqq (r/\sigma)^p\mu^{-p} = EN_n/(1 - \delta)\}\,. \tag{39}$$

In view of (37), $EN_n = n\varepsilon \to \infty$ iff $\mu^p(\sigma/r)^p \to 0$. But $(\sigma/r)^2\mu^2 \sim 2(\sigma/r)^2\log\varepsilon^{-1} \sim 2(\sigma/r)^2\log n(\sigma/r)^p \to 0$ by the hypothesis of case (ii) of the theorem. Now apply Chebychev's inequality to get

$$P_n(\Theta_n^c) = P\{(N_n - EN_n)/EN_n \geqq \delta/(1 - \delta)\} \leqq \delta^{-2}(1 - \delta)^2/\eta\varepsilon \to 0\,.$$

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

The Bayes estimator may be bounded (Sect. 8.3) in terms of the posterior moment:

$$\begin{aligned}|\hat{\theta}_{\pi_n}|^q &\leqq 2^q E_{\pi_n}(|\theta|^q|x)\\ &\leqq 2^q\sigma^q\mu^q E_{\pi_n}(N_n|x) \leqq 2^q\sigma^q\mu^q EN_n/(1 - \delta)\,,\end{aligned}$$

since the posterior is concentrated on Θ_n (cf. (39)). Thus,

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

while from (38)

$$\rho(P_n) \sim n\sigma^q\varepsilon\mu^q\Phi(a) = \sigma^q\mu^q\Phi(a)EN_n\,.$$

Asymptotic neglibility of the ratio of these two least expressions follows from that of $P_n(\Theta_n^c)$ and $E_{P_n}|N - EN_n|/EN_n$. In (36), therefore, $\rho(P_n) \leqq \rho(\pi_n)(1 + o(1))$. The proof is completed by taking δ small and a large. ▌

Remark. In case (ii), the condition that $(\sigma/r)^2\log n(\sigma/r)^p \to 0$ cannot be completely removed. For example, in a very low signal to noise case such as $r = 1$,

$\sigma = n^{\alpha}, \alpha > 0$, then $R_L^* \sim 1$ (Theorem 7, Part 3). However, $\eta_n = n^{-1/p-\alpha}$ and from Theorem 15, $R_B^* \asymp n^{1+\alpha q}(\log n)^{(q-p)/2} \gg 1$, whereas $R_N^* \sim 1$.

6 Threshold rules over l_p balls

We use the Bayes-minimax approach and the univariate threshold results 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:

$$R_S^*(\sigma; \Theta_{p,n}(r)) = \inf_{\lambda} \sup_{\pi} \{E_\pi E_\theta \| \hat{\theta}_\lambda^{(s)} - \theta \|_q^q : E_\pi \sum_1^n |\theta_i|^p \leqq r^p\} . \tag{40}$$

(with a similar definition of R_h^* for hard thresholds). Just as was argued in Sect. 3,

$$\inf_{\lambda} \sup_{\theta \in \Theta_{p,n}(r)} E_\theta \| \hat{\theta}_\lambda^{(s)} - \theta \|_q^q \leqq R_s^* = n\rho_s(n^{-1/p}r, \sigma) . \tag{41}$$

The proof is entirely analogous: if (F^0, δ_λ^0) is a saddlepoint for problem (21), then $(F^{0n}, \delta_\lambda^{0n})$ is a saddlepoint for problem (40). Theorem 6 now follows from (41), the bounded inefficiency of $\rho_s(\tau, \sigma)$ relative to $\rho(\tau, \sigma)$ (Theorem 11), and from Theorem 17:

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

In fact, when $\eta_n \to 0$ under the conditions of Theorem 6, threshold rules are asymptotically *efficient*. Indeed, from Theorems 13 and 15,

$$\frac{R_s^*}{R_B^*} = \frac{\rho_s(n^{-1/p}r, \sigma)}{\rho(n^{-1/p}r, \sigma)} = \frac{\rho_s(\eta_n, 1)}{\rho(\eta_n, 1)} \to 1 \quad \text{as } \eta_n \to 0 .$$

7 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, $q \geqq 1$) 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 $\hat{\theta}_{abc,i}(x) = ax_i + b(\sum_{j \neq i} x_j) + c$. A further convexity argument (Sect. 8.4) 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

$$R_L^*(\sigma; \Theta_{p,n}(r)) = \inf_{0 \leqq a \leqq 1} \sup_{\Theta_{p,n}(r)} E_\theta \| aY - \theta \|_q^q . \tag{42}$$

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

$$\sup_{\Theta_{p,n}(r)} E_\theta \| aY - \theta \|_q^q = n\sigma^q \sup \{ n^{-1} \sum_1^n E_{\mu_i} |aX_i - \mu_i|^q : n^{-1} \sum |\mu_i|^p \leqq \eta_n^p \} . \quad (43)$$

The risk function in the univariate location problem that appears on the right side of (43) can be expressed in terms of a single standard Gaussian deviate Z:

$$r_q(a, \mu) = E_\mu |aX - \mu|^q = a^q E|Z + b\mu|^q = a^q s(b^p|\mu|^p) , \quad (44)$$

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

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

Since $s(\gamma)$ is increasing in γ, there is no harm in replacing the inequality in the supremum in (43) by equality. We obtain

$$R_L^* = n\sigma^q \inf_a a^q \sup \{ n^{-1} \sum s(\gamma_i) : n^{-1} \sum \gamma_i = b^p \eta^p, \gamma_i \geqq 0 \} \quad (45)$$

$$= n\sigma^q \inf_{b \geqq 0} (1 + b)^{-q} s_n^*(b^p \eta^p) . \quad (46)$$

Remark. The function s_n^* implicitly defined in (46) 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 $\gamma_i \geqq 0, n^{-1}\sum \gamma_i = \tau$ belongs to the class $\mathscr{F}_1^{(n)}$ of probability measures supported on $[0, n\tau]$ with mean equal to τ. Thus

$$s_n^*(\tau) \leqq \tilde{s}_n(\tau) \overset{\text{def}}{=} \sup \left\{ \int s(\gamma) F(d\gamma), F \in \mathscr{F}_1^{(n)} \right\} . \quad (47)$$

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

$$\tilde{s}_n(\tau) = \sup \{ \alpha s(\gamma_1) + (1 - \alpha) s(\gamma_2) : \alpha\gamma_1 + (1 - \alpha)\gamma_2 = \tau, 0 \leqq \alpha \leqq 1, 0 \leqq \gamma_i \leqq n\tau \}$$

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 (45), 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

$$q^{-1} p \gamma^{1 - 1/p} s'(\gamma) = E_c v |v|^{q-2}, \quad (48)$$

and, more importantly, that

$$q^{-1} p^2 \gamma^{2 - 1/p} s''(\gamma) \overset{\text{def}}{=} F(c; p, q) \quad (49)$$

$$= \begin{cases} (q - 1) E_c c |v|^{q-2} - (p - 1) E_c v |v|^{q-2} & q > 1 \\ 2c\phi(c) - (p - 1)[2\Phi(c) - 1] & q = 1 \end{cases} \quad (50)$$

A useful representation (Sect. 8.5) is

$$e^{c^2/2} F(c) = 2 \int_0^\infty g(v) v^{q-3} \phi(v) \sinh cv dv \quad q > 1, \quad (51)$$

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 (Sect. 8.6),

$$X = S \cap \{p \leqq q, p \leqq 2\} = \{(p, q)\text{: } s \text{ is convex on } [0, \infty)\}$$

$$V = S \cap \{p \geqq q, p \geqq 2\} = \{(p, q)\text{: } s \text{ is concave on } [0, \infty)\}$$

$$XV = S \cap \{p > q, p < 2\} = \{(p, q)\text{: } s \text{ is convex on } [0, \gamma_0], \text{ concave on } [\gamma_0, \infty)\}$$

$$VX = S \cap \{p < q, p > 2\} = \{(p, q)\text{: } s \text{ is concave on } [0, \gamma_0], \text{ convex on } [\gamma_0, \infty)\}$$

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

We evaluate R_L^* in turn for the sets in the partition. First on $X \cup VX$, where s is convex at least for large γ, we construct lower bounds using 'spikes'. Fix $\theta = (r, 0, \ldots, 0)$, which corresponds to $\mu = (r\sigma^{-1}, \ldots, 0)$, to obtain from (45) the lower bound

$$R_L^* \geqq n\sigma^q \inf_a (1 - n^{-1})E_0|aX|^q + n^{-1}E_{r\sigma^{-1}}|aX - r\sigma^{-1}|^q \tag{52}$$

$$= n\sigma^q \inf_a (1 - n^{-1})a^q c_q + \tilde{\eta}_n^q (1-a)^q t(a, \sigma r^{-1}), \tag{53}$$

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

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

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

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

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

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

Some technical work shows that

$$\inf_a (1 - n^{-1})a^q c_q + \tilde{\eta}_n^q (1-a)^q t(a, \sigma r^{-1}) \sim f(a_*(\tilde{\eta}_n), \tilde{\eta}_n) \text{ as } n \to \infty. \tag{54}$$

Combining these results with the lower bound in (53) yields

$$R_L^* \geqq (1 + o(1)) \cdot \begin{cases} n\sigma^q c_q & \tilde{\eta}_n \to \infty & \text{(a)} \\ n\sigma^q f(a_*(\eta); \eta) & \tilde{\eta}_n \to \eta \in (0, \infty) & \text{(b)} \\ r^q & \tilde{\eta}_n \to 0. & \text{(c)} \end{cases} \tag{55}$$

For upper bounds on $X \cup VX$, make various choices of a in (42). For $a = 1$,

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

which is sharp when $\tilde{\eta}_n \to \infty$ (cf. (55a)), and so establishes Case 1 of Theorem 7. The choice $a = 0$ gives

$$R_L^* \leqq n\sigma^q \sup\left\{n^{-1}\sum_1^n \gamma_i^{q/p} : n^{-1}\sum \gamma_i = \eta_n^p\right\},$$

after setting $\gamma_i = \mu_i^p$. When $q \geqq p$, 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

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

which is in turn sharp when $\tilde{\eta}_n \to 0$. (cf. (55c)).

Consider now the case $\tilde{\eta}_n \to \eta \in (0, \infty)$. When $q \geqq p$ and $p \leqq 2$ (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 (53). When combined with (54), this shows that (55b) 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 (55b) is asymptotically sharp (Sect. 8.7). 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 $(p, q) \in V = S \cap \{p \geqq q, p \geqq 2\}$. In this case, the vector $\mu = \eta_n(1, \ldots, 1)$ is least favorable, and from (45), (or, when $p = \infty$, directly from (42)) we obtain

$$R_L^* = n\sigma^q \inf\{w(b; \eta_n) : b \geqq 0\}\ , \tag{56}$$

where $w(b; \eta) = (1 + b)^{-q} E|Z + b\eta|^q$ for $Z \sim N(0, 1)$, and does not depend on p. It turns out (Sect. 8.8) that there is a unique minimax linear estimator $\mu(x) = x/(1 + b_*)$, not depending on p, 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 pause to be more explicit in the case $q = 2$. If $p \leqq 2$, then $s(\gamma)$ is convex, $\mu = (r\sigma^{-1}, 0, \ldots, 0)$ is least favorable and there is equality in (52) and (53) which reduces to (12). If $p \geqq 2$ then $s(\gamma)$ is concave and we arrive at (12) via (56).

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 (56) is still a valid lower bound for R_L^*, and so from the discussion above, we conclude that $R_L^* \geqq n\sigma^q \eta_n^q(1 + o(1))$. On the other hand, a natural upper bound is obtained from the estimator with $a = 0$:

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

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 from (46) and (47):

$$R_L^* \leqq n\sigma^q \inf_b (1 + b)^{-q} \tilde{s}(b^p \eta_n^p) \tag{57}$$

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

$$\tilde{s}(\gamma) = \begin{cases} c_q + R\gamma & \gamma \leqq \gamma_0 \\ s(\gamma) & \gamma \geqq \gamma_0 \end{cases} \tag{58}$$

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 $s'(\gamma_0) = [s(\gamma_0) - s(0)]/\gamma_0$. As is shown in Donoho and Johnstone (1992, p. 32), the error involved in the upper bound (57) 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 (Sect. 8.9) that there is a unique value $b_* = b_*(p, q, \eta_n)$ optimizing the right side of (57). 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 (47)) has the form $(1-\varepsilon)\nu_0 + \varepsilon\nu_{\gamma_0}$, where $\varepsilon\gamma_0 = \gamma_*$. It turns out that $\gamma_* < \gamma_0$ exactly when

$$\eta p(s(\gamma_0) - c_q) + [p(s(\gamma_0) - c_q) - qs(\gamma_0)]\gamma_0^{1/p} > 0\,, \tag{59}$$

which occurs when η is sufficiently large. Thus the set XV provides examples where the least favorable configuration is neither a spike nor uniformly grey.

8 Appendix

1 Properties of $F \to \rho_q(F)$: Upper semi-continuity

Set $R(\mu, \hat{\mu}) = E_\mu|\hat{\mu}(X) - \mu|^q$ and $\rho(F, \hat{\mu}) = \int R(\mu, \hat{\mu})dF(\mu)$: since $\mu \to R(\mu, \hat{\mu})$ is continuous, $F \to \rho(F, \hat{\mu})$ is weakly continuous on $\mathscr{F}_p(\eta)$ when the risk function is *also* bounded. Since the infimum defining $\rho_q(F)$ includes estimators with unbounded risk function we define an increasing family of subclasses of estimators $\mathscr{D}_m = \{\hat{\mu} : \hat{\mu}(x) = x \text{ for } |x| > m\}$, and let

$$\rho_{qm}(F) = \inf_{\hat{\mu} \in \mathscr{D}_m} \rho(F, \hat{\mu})\,.$$

Since each estimator in $\mathscr{D}_m$ has bounded risk, $F \to \rho_{qm}(F)$ is weakly upper semicontinuous (usc). Since $\rho_{qm}(F)$ decreases as $m \nearrow \infty$ it has a limit, $\tilde{\rho}_q(F)$ say, and if we assume for the moment that $\rho_q(F) = \tilde{\rho}_q(F)$ then $\rho_q(F)$ is the decreasing limit of a family of usc functions and is hence also usc.

To verify that $\rho_q(F) = \tilde{\rho}_q(F)$, note first that trivially $\rho_q(F) \leqq \tilde{\rho}_q(F)$. 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 $\hat{\mu}_m \in \mathscr{D}_m$ is defined by $\hat{\mu}_m(x) = \mu(x)I\{|x| \leqq m\} + xI\{|x| > m\}$ [because $R(\hat{\mu}_m, \mu) \to R(\hat{\mu}, u)$ uniformly on compact intervals]. This establishes upper semicontinuity. Since $\rho_q(F)$ is the pointwise infimum of linear functions, it is concave, and $\mathscr{F}_p(\eta)$ is weakly compact because of the moment condition.

To verify that $\rho_q(F_{1+c}) \leqq (1+c)^q \rho(F)$, first let $\hat{\mu}_F$ denote the Bayes estimator of μ for prior F. Let $\phi = (1+c)\mu$, and suppose that $y|\phi \sim N(\phi, 1)$. 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,

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

2 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 $\mu \geqq 0$

$$r(\delta_\lambda^{(s)}, \mu) = E_\mu|(x-\lambda)_+ + (x+\lambda)_- - \mu|^q \tag{60}$$

$$= \int_{-\infty}^{-\lambda-\mu} |w+\lambda|^q \phi(w)dw + \mu^q \int_{-\lambda-\mu}^{\lambda-\mu} \phi(w)dw + \int_{\lambda-\mu}^{\infty} |w-\lambda|^q \phi(w)dw \tag{61}$$

and

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

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

$$r(\delta_\lambda^{(h)}, \mu) = \int_{-\infty}^{-\lambda-\mu} |w|^q \phi(w)dw + \mu^q \int_{-\lambda-\mu}^{\lambda-\mu} \phi(w)dw + \int_{\lambda-\mu}^{\infty} |w|^q \phi(w)dw , \tag{62}$$

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:

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

3 Moment inequality

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

$$|\mu_q|^q \leqq 2^{q-1}[E|X|^q + E|\mu_q - X|^q] \leqq 2^q E|X|^q . \tag{63}$$

Equation (40) follows by taking for F the posterior distribution of θ_i given x under the prior π_n. [A refined version of (63) appears in Johnstone (1991)].

4 *Structure of the linear minimax rule*

Suppose that $\Theta \subset R^r$ is orthosymmetric. Let $\hat{\theta}_{abc,i}(x) = ax_i + bx_i' + c$, where $x_i' = \sum_{j \neq i} x_j$. Let $R(\theta, \hat{\theta}) = E_\theta \sum_{i=1}^p |\hat{\theta}_i - \theta_i|^q$. Then we show that

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

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

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

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

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

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

$$\begin{aligned} 2^p E|aX_1 - \theta_1|^q &\leqq \sum_{\sigma_1} \sum_{\sigma'} E|a(\sigma_1\theta_1 + \varepsilon_1) + d - \sigma_1\theta_1 + b \sum_{j \geqq 2} (\sigma_j\theta_j + \varepsilon_j)|^q \\ &= \sum_\sigma E_{\sigma\theta}|aX_1 + d + bX_1' - \sigma_1\theta_1|^q \\ &= \sum_\sigma E_{\sigma\theta}|\hat{\theta}_{abd,1} - \sigma_1\theta_1|^q . \end{aligned}$$

Now add over i to get

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

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

$$2^p \bar{R}(\hat{\theta}_{a00}) \leqq \sum_\sigma \bar{R}(\hat{\theta}_{abd}) = 2^p \bar{R}(\hat{\theta}_{abd}). \quad \blacksquare$$

5 *Convexity decompositions of loss functions and parameter spaces*

We first note that for $0 \leqq c < \infty$, the mapping $(p, q) \to F(c; p, q)$ is continuous on S. This is clear from (50), except possibly for $q \searrow 1$. That continuity holds here also is evident from the representations

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

and

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

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.

To obtain representation (51) combine the following identities and (50):

$$E_c v|v|^{q-2} = 2e^{-c^2/2} \int_0^\infty v^{q-1}\phi(v)\sinh cv\,dv\,,$$

$$cE_c|v|^{q-2} = 2e^{-c^2/2} \int_0^\infty v^{q-2}\phi(v)c\cosh cv\,dv\,,$$

$$= 2e^{-c^2/2} \int_0^\infty [v^{q-1} - (q-2)v^{q-3}]\phi(v)\sinh cv\,dv\,.$$

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

$$\begin{vmatrix} \sinh cv & \dfrac{\partial}{\partial c}\sinh cv \\ \dfrac{\partial}{\partial v}\sinh cv & \dfrac{\partial^2}{\partial c\partial v}\sinh cv \end{vmatrix} = \frac{1}{2}[\sinh(2cv) - 2cv] \geqq 0\,.$$

In turn, it follows that the kernel $(c, v) \to v^{q-3}\phi(v)\sinh cv$ is TP_2 and has the variation diminishing property. 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, et al. (1981)).

6 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 from which the decomposition of S follows.

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

b) $p \geqq q \geqq 2$. 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 (50), one sees that $F(\infty) = (q-p)c^{q-1}$. To determine behavior at 0, we note that for $q > -1$

$$c_q = E|Z|^q = \frac{2^{q/2}}{\sqrt{\pi}}\Gamma\left(\frac{q+1}{2}\right), \text{ and } (q-1)c_{q-2} = c_q\,. \tag{64}$$

Substituting the expansion $\sinh cv = cv + (cv)^3/6 + \ldots$ into (51) yields

$$e^{c^2/2}F(c) \sim c[(q-p)c_q - (q-1)(q-2)c_{q-2}] \tag{65}$$

$$+ \frac{c^3}{6}[(q-p)c_{q+2} - (q-1)(q-2)c_q] + o(c^3)$$

$$\sim \begin{cases} (2-p)c_q c & p \neq 2 \\ (q-2)\dfrac{c_q}{3}c^3 & p = 2, q \neq 2. \end{cases} \tag{66}$$

c) $q > p$, $q > 2$. Here $F(\infty) > 0$. If $p \leqq 2$, 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 $p \geqq 2$, 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 is convex on $[0, \gamma_0]$ and concave on $[\gamma_0, \infty)$.

7 *Sharpness of* (55b) *when* $\tilde{\eta}_n \to \eta \in (0, \infty)$; $q > p > 2$

Combining the equality (46) with the upper bound (47) we obtain

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

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 $\gamma \leqq \gamma_0$ and convex for $\gamma \geqq \gamma_0$, it follows that $\tilde{s}_n(\bar{\gamma}_n) = (1 - \varepsilon_n)s(\gamma_n) + \varepsilon_n s(n\bar{\gamma}_n)$ where ε_n and γ_n are determined by the equations

$$(1 - \varepsilon_n)\gamma_n + \varepsilon_n n\bar{\gamma}_n = \bar{\gamma}_n \tag{67}$$

$$s'(\gamma_n) = [s(n\bar{\gamma}_n) - s(\gamma_n)]/[n\bar{\gamma}_n - \gamma_n] . \tag{68}$$

[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

$$\tilde{s}_n(\bar{\gamma}_n) = (1 - \varepsilon_n)s(\gamma_n) + \varepsilon_n s(n\bar{\gamma}_n) \sim (1 - n^{-1})s(0) + n^{-1}s(n\bar{\gamma}_n) , \tag{69}$$

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

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

as is shown following (52).

To establish (69), one sees from (67) 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. (48) and an argument similar to (66)) recasts (68) as the equation

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

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

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

8 *Minimax estimation in the concave case*

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

$$r(b) \overset{\text{def}}{=} q^{-1}(1 + b)^{q+1}w'(b) = E_c k(v) \quad k(v) = |v|^{q-2}\{\eta v - (q - 1)\}. \tag{70}$$

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 (70), we note from the series $\sinh cv = cv + (cv)^3/6 + \ldots$ that

$$E_c v|v|^{q-2} = 2e^{-c^2/2} \int_0^\infty v^{q-1}\phi(v) \sinh cv \, dv$$
$$= e^{-c^2/2}[c_q c + c_{q+2}c^3/6 + \cdots] \text{ and}$$
$$E_c c|v|^{q-2} = e^{-c^2/2}[c_{q-2}c + c_q c^3/2 + \cdots] .$$

Substituting leading terms into (70) and using the identity (64) 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 zero 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 $w(b_*) = (1 + (q-1)\eta^{-2})^{-q}E|Z + (q-1)\eta^{-1}|^q \sim \eta^q$.

9 Convex-concave case

Put $\gamma = b^p\eta^p$: the right side of (57) becomes

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

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

We now verify that $\rho(\gamma)$ has exactly one sign change on $[0, \infty)$ For $\gamma \leqq \gamma_0$, substitution from (58) yields

$$\rho(\gamma) = pR + \frac{p-q}{\eta}R\gamma^{1/p} - \frac{q}{\eta}c_q\gamma^{1/p-1} \tag{71}$$

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 writing $R = [s(\gamma_0) - c_q]/\gamma_0$ into (71), one verifies that this occurs as described in (59).

Acknowledgements. This is an expanded version of a technical report first written in May 1989. We are grateful to two referees for comments on the July 1992 version. 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, 92-09130, NIH PHS GM21215-12, SERC and the Sloan Foundation. I.J. thanks the Australian National University for hospitality during two visits.

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. Birgé, L., Massart, P.: Rates of convergence for minimum contrast estimators. Technical Report Université Paris VI. Probab. Theory Relat. Fields **97**, 113–150 (1993)
3. Brown, L.D.: Admissible estimators, recurrent diffusions, and insoluble boundary value problems. Ann. Math. Stat. **42**, 855–903 (1971)
4. Brown, L.D., Johnstone, I.M., MacGibbon, K.B.: Variation Diminishing Transformations: A direct approach to total positivity and its statistical applications. J. Am. Stat. Assoc. **76**, 824–832 (1981)
5. Casella, G., Strawderman, W.E.: Estimating a bounded normal mean. Ann. Stat. **9**, 870–878 (1981)
6. Daubechies, I.: Orthonormal bases of compactly supported wavelets. Commun. Pure Appl. Math. **41**, 909–996 (1988)
7. Daubechies, I.: Ten Lectures on Wavelets SIAM: Philadelphia (1992)
8. Donoho, D.L.: Gelfand n-widths and the method of least squares. (Preprint, 1990)

9. Donoho, D.L.: Asymptotic minimax risk for sup-norm loss: solution via optimal recovery. Probab. Theory Relat. Fields (to appear, 1994)
10. Donoho, D.L., Johnstone, I.M.: Minimax risk over l_p-balls. (Technical Report No. 322) Department of Statistics. Stanford: Stanford University 1989
11. Donoho, D.L., Johnstone, I.M.: Minimax risk over l_p-balls for l_q-error (Technical Report) Department of Statistics. Stanford: Stanford University 1992
12. Donoho, D.L., Johnstone, I.M.: Minimax estimation via wavelet shrinkage. (Technical Report). Ann. Stat. (to appear, 1995a)
13. Donoho, D.L., Johnstone, I.M.: Adapting to unknown smoothness via Wavelet shrinkage. J. Am. Stat. Assoc. (to appear, 1995b)
14. Donoho, D.L., Johnstone, I.M.: Non-classical Minimax Theorems, Thresholding, and Adaptation. Manuscript (1993)
15. Donoho, D.L., Johnstone, I.M.: Ideal Spatial Adaptation via Wavelet Shrinkage. Biometrika (to appear, 1994)
16. Donoho, D.L., Johnstone, I.M., Hoch, J.C., Stern A.S.: Maximum Entropy and the Nearly Black Image. J. R. Stat. Soc. Ser. B **54**, 41–81 (1992) with discussion
17. Donoho, D.L., Johnstone, I.M., Kerkyacharian, G., Picard, D.: Wavelet Shrinkage: Asymptopia?
18. Donoho, D.J., Johnstone, I.M., Kerkyacharian, G., Picard, D.: Density estimation by wavelet thresholding. (Technical Report) Department of Statistics. Stanford: Stanford University 1993
19. Donoho, D.L., Liu, R.C., MacGibbon, K.B.: Minimax risk over hyperrectangles, and implications. Ann. Stat. **18**, 1416–1437 (1990)
20. Donoho, D.L., Liu, R.C.: Geometrizing Rates of Convergence, III. Ann. Stat. **19**, 668–701 (1991)
21. Feldman, I.: Constrained minimax estimation of the mean of the normal distribution with known variance. Ann. Stat. **19**, 2259–2265 (1991)
22. Huber, P.J.: Robust Estimation of a Location Parameter. Ann. Math. Stat. **35**, 73–101 (1964)
23. Huber, P.J.: Fisher Information and Spline Interpolation. Ann. Stat. **2**, 1029–1034 (1974)
24. Ibragimov, I.A., Khasminskii, R.Z.: Nonparametric estimation of the value of a linear functional in a Gaussian white noise. Theory Probab. Appl. **29**, 1–32 (1984)
25. Ibragimov, I.A., Hasminskii, R.Z.: On density estimation in the view of Kolmogorov's ideas in approximation theory. Ann. Stat. **18**, 999–1010 (1990)
26. Johnstone, I.M.: A moment inequality for L_q estimation. Stat. Probab. Lett. **12**, 289–290 (1991)
27. Johnstone, I.M.: Minimax Bayes, asymptotic minimax and sparse wavelet priors. Statistical Decision Theory and Related Topics V (pp. 303–326). Berlin Heidelberg New York: Springer 1994
28. Karlin, S.: Total Positivity. Stanford: Stanford University Press 1968
29. Korostelev, A.P.: Asymptotic minimax estimation of regression function in the uniform norm. Teor. Veoryatn. Primen. **37** (in russian). Theory Probab. Appl. **37** (1993) (in english)
30. LeCam, L.: Asymptotic methods in statistical decision theory. Berlin Heidelberg New York: Springer 1986
31. Lehmann, E.L.: Theory of Point Estimation. New York: Wiley 1983
32. Lindenstrauss, J., Tzafiri, L.: The Classical Banach Spaces, II. Berlin Heidelberg New York: Springer 1979
33. Meyer, Y.: Ondelettes. Hermann: Paris 1990; Cambridge University Press 1993 (English translation)
34. Nemirovskii, A.S., Polyak, B.T., Tsybakov, A.B.: Rate of convergence of non-parametric estimates of maximum-likelihood type. Prob. Inf. Transm. **21**, 258–272 (1985)
35. Pinsker, M.S.: Optimal Filtration of square-integrable signals in Gaussian White Noise. Prob. Inf. Transm. **16**, 120–133 (1980)
36. Rockafellar, R.T.: Convex Analysis. Princeton: Princeton University Press 1970
37. Sacks, J., Strawderman, W.E.: Improving on linear minimax estimates. (Stat. Decis. Theory Relat. Topics III, vol. 2 pp. 287–304) J.R. Stat. Soc., Ser. B (with discussion) (to appear, 1995) Berlin Heidelberg New York: Springer 1982
38. Sion, M.: On general minimax theorems. Pac. J. Math. **8**, 171–176 (1958)
39. Van de Geer, S.: Estimating a regression function. Ann. Stat. **18**, 907–924 (1990)