# Minimax Quadratic Estimation of a Quadratic Functional

By

David L. Donoho
University of California, Berkeley

Michael Nussbaum
Karl Weierstrass Inst. für Mathematik, Berlin

Technical Report No. 236
January 1990

This work was partially supported by NSF grants DMS 84-57153 and DMS 88-10192.

Department of Statistics
University of California
Berkeley, California 94720

# Minimax Quadratic Estimation of a Quadratic Functional

David L. Donoho
University of California, Berkeley

Michael Nussbaum
Karl Weierstrass Inst. für Mathematik, Berlin

January, 1990
Revised March, 1990
Revised May, 1990

**Abstract**

We study the problem of estimating the integrated squared derivative of a periodic function, when the observations are incomplete and noisy. We derive a simple quadratic estimate which is asymptotically minimax among quadratic estimates. Our estimate is also rate-optimal among all measurable estimates.

**Key Words and Phrases.** Modulus of Continuity. Optimal Rates of Convergence. Ellipsoids. Hyperrectangles. Estimation of Linear Functionals.

**Acknowledgements.** This work was partially supported by NSF grants DMS 84-57153 and DMS 88-10192. The authors would like to thank Jianqing Fan for helpful conversations.

# 1 Introduction

Suppose we observe noisy samples of a function $f$,

$$v_i = f(t_i) + z_i, \quad i = 1, \ldots, n,$$

where the $t_i = -\pi + 2\pi(i/n)$ are equispaced on $[-\pi, \pi]$ and the noise terms $z_i$ are independent and identically distributed (i.i.d.) according to a Gaussian $N(0, \sigma^2)$ distribution. We are interested in estimating the quadratic functional $Q(f) = \int_{-\pi}^{\pi}(f^{(k)}(t))^2 dt$, and we know, *a priori*, that $\int_{-\pi}^{\pi}(f^{(m)}(t))^2 dt \leq 1$, where $m > k$. Our aim is to find estimators $\hat{Q}(\mathbf{v})$ for $Q(f)$, and we evaluate performance according to the worst-case mean-squared error $\sup_f E(\hat{Q}(\mathbf{v}) - Q(f))^2$.

This is a problem of estimating a nonlinear functional of $f$ from incomplete, noisy data on $f$. Such problems have been addressed by Ibragimov, Nemirovskii, and Has'minskii (1987) and by Fan (1988). In general the study of such problems is just beginning, and precise optimality results are unavailable. For related literature, see Levit (1978), Hall and Marron (1987), and Ritov and Bickel (1988).

We will show in this note that if we restrict attention to the class $\mathcal{Q}_n$ of inhomogeneous *quadratic* estimators $\hat{Q}(\mathbf{v}) = c + \sum \hat{q}_{ij} v_i v_j$ of the quadratic functional $Q$, and if we restrict attention to *periodic* functions in

$$\begin{aligned} \mathcal{F}_m = \{f : \; & f, \ldots, f^{(m-1)} \text{ abs.cont.} \\ & f^{(l)}(\pi) = f^{(l)}(-\pi),\ l = 0, \ldots, m-1 \\ & \int_{-\pi}^{\pi} (f^{(m)}(t))^2 dt \leq 1\} \end{aligned}$$

then it is possible to derive precise asymptotic results on the *minimax quadratic risk*

$$R_Q^*(n) = \inf_{\mathcal{Q}_n} \sup_{\mathcal{F}_m} E(\hat{Q}(\mathbf{v}) - Q(f))^2,$$

and a simple, easily computable quadratic estimator. Specifically, we establish the following:

**Theorem 1** *Let* $r = \frac{4m-4k}{4m+1} < 1/2$ *and* $4m^2 > 1 + m$. *Put*

$$\beta = (2\pi)^r 2^{3r/2} r^{r/2} [2k + 2m + 1]^{-r/2} (\frac{\sigma}{\sqrt{n}})^{2r}.$$

*Let* $w_j = n^{-1} \sum_{u=1}^{n} v_u \exp\{2\pi \sqrt{-1}\, \frac{(u-1)(j-1)}{n}\}$ *denote the* $j$*-th (complex) Finite Fourier Coefficient of* $\mathbf{v}$, *and set* $W_j = 4\pi(|w_j|^2 - \sigma^2/n)$. *Then the estimator*

$$Q_0(\mathbf{v}) = \beta/2 + \sum_{j>0} j^{2k}(1 - \beta j^{2m-2k})_+ W_j$$

*is asymptotically minimax as* $n \to \infty$, *in the sense that*

$$\sup_{\mathcal{F}_m} E(Q_0(\mathbf{v}) - Q(f))^2 \sim R_Q^*(n).$$

*Moreover,*

$$R_Q^*(n) \sim A(k, m)\sigma^{4r} n^{-2r}$$

*where*

$$A(k, m) = (2\pi)^{2r} 2^{3r-2} (1 - r)_*^{-1} r^r [2k + 2m + 1]^{-r}.$$

Several remarks are in order. First, the estimator in question can be computed in order $O(n \log n)$ arithmetic operations, which serves as partial compensation for our decision to restrict attention to quadratic estimators. Second, we will show that even by employing arbitrary measurable functions of the data as estimators, the rate $(\frac{\sigma}{\sqrt{n}})^{4r}$ cannot be essentially improved. Third, the case $r \geq 1/2$ excluded by the above theorem corresponds to the case where estimates with rate of convergence $\frac{\sigma^2}{n}$ are available and classical methods are available.

An interesting aspect of our approach is the use of ideas from *linear* estimation to solve this problem. That is, we transform the problem to a problem of estimating a linear functional and use recent results of Donoho (1989) on minimax affine estimates of linear functionals to solve the problem.

In a final section we compare this result, which concerns optimal estimation in the presence of stochastic noise, with the problem of optimal estimation in the presence of deterministic noise.

## 2 Minimax Quadratic Estimation

In sections 2 and 3 we consider an apparently different estimation problem. We observe data $y_i = \theta_i + z_i$, $i = 1, 2, \ldots$, where the $z_i$ are i.i.d. $N(0, \epsilon^2)$. We know a priori that $\theta \in \Theta$, and we wish to estimate the quadratic functional $Q(\theta) = \sum_{i=1}^{\infty} q_i \theta_i^2$, (all $q_i \geq 0$), in such a way as to attain the minimax quadratic risk

$$R_Q^*(\epsilon) = \inf_{\mathcal{Q}} \sup_{\Theta} E(\hat{Q}(\mathbf{y}) - Q(\theta))^2$$

Here $\mathcal{Q}$ is the class of quadratic estimates, i.e. any rule of the form $\hat{Q}(\mathbf{y}) = \sum_{i,j} \hat{q}_{ij} y_i y_j + e$, where $(\hat{q}_{ij})$ and $e$ are constants. Fan (1988) has shown that for certain $\Theta$, we can specialize attention to the class $\mathcal{Q}_D$ of *diagonal* quadratic rules, which are of the form

$$\hat{Q}(\mathbf{y}) = \sum_i \hat{q}_i (y_i^2 - \epsilon^2) + e.$$

Say that $\Theta$ is *orthosymmetric* if, whenever $\theta \in \Theta$, then also $(\pm\theta_i) \in \Theta$ for all possible sequences of signs $(\pm)$. Fan's lemma says that, *if $\Theta$ is orthosymmetric, then the minimax risk over rules in $\mathcal{Q}$ is attained by rules in $\mathcal{Q}_D$*.

A further reduction is possible. A diagonal *shrinkage* rule is any diagonal rule with

$$0 \leq \hat{q}_i \leq q_i, \quad i = 1, 2, \ldots .$$

We denote the class of all such rules by $\mathcal{Q}_{DS}$. Like the other lemmas in this paper, the following is proved in the appendix.

**Lemma 1** *Let $\Theta$ be orthosymmetric. For estimating the orthosymmetric functional $Q(\theta) = \sum q_i \theta_i^2$, the minimax quadratic risk is attained within the class $\mathcal{Q}_{DS}$ of diagonal* **shrinkage** *rules.*

The reduction provided by this lemma is essential to our paper. Let us indicate why. We record that

$$\begin{aligned} E(y_i^2 - \epsilon^2) &= \theta_i^2 \\ Var(y_i^2 - \epsilon^2) &= 4\epsilon^2\theta_i^2 + 2\epsilon^4. \end{aligned}$$

For a diagonal rule, we get

$$E(\hat{Q}(\mathbf{y}) - Q(\theta))^2 = Bias^2(\hat{Q},\theta) + Var(\hat{Q},\theta),$$

where

$$Bias(\hat{Q},\theta) = \sum_i (\hat{q}_i - q_i)\theta_i^2 + \epsilon$$

and

$$Var(\hat{Q},\theta) = 4\epsilon^2 \sum_i \hat{q}_i^2\theta_i^2 + 2\epsilon^4 \sum_i \hat{q}_i^2.$$

Note that the variance of $\hat{Q}(\mathbf{y})$ is heterogeneous - it depends on $\theta$. However, it turns out that for shrinkage rules, the heterogeneity is asymptotically negligible, in a certain sense.

For example, suppose we are interested in the functional $Q(\theta) = \sum_i \theta_i^2$. Then shrinkage rules satisfy $0 \leq \hat{q}_i \leq 1$ for all $i$ and so for all such rules, the heterogeneous term satsifies

$$4\epsilon^2 \sum_i \hat{q}_i^2\theta_i^2 \leq 4\epsilon^2 \sum_i \theta_i^2.$$

If $\Theta$ is norm-bounded, so that $\sum_i \theta_i^2 \leq M$, say, then the heterogeneous term is uniformly of order $O(\epsilon^2)$. Consequently, in the cases where $R_Q^*(\epsilon) >> \epsilon^2$, heterogeneity is unimportant.

Define, then, the pseudo-risk

$$R(\hat{Q},\theta) = (\sum_i (\hat{q}_i - q_i)\theta^2 + \epsilon)^2 + 2\epsilon^4 \sum_i \hat{q}_i^2; \qquad (1)$$

this is the true risk minus the heterogeneous term of the variance. Consider the minimax pseudo-risk

$$R_Q(\epsilon) = \inf_{Q_{DS}} \sup_{\Theta} R(\hat{Q},\theta).$$

By the above comments, if $Q = \sum_i \theta_i^2$, and $\Theta$ is orthosymmetric and norm-bounded, then

$$R_Q^* \sim R_Q \text{ as } \epsilon \to 0; \qquad (2)$$

we call this the *homogeneous variance* approximation.

Given a sequence $(r_i)$ of positive entries, define the $l_p$-body $\Theta_p((r_i)) = \{\theta : \sum_i r_i|\theta_i|^p \leq 1\}$. This set is orthosymmetric, and if $p \geq 1$, convex, so Lemma 1 applies. The following result shows that the homogeneous variance approximation holds for a special class of $l_p$-bodies.

**Theorem 2** *(Risk Approximation) Let $\Theta = \Theta_p((r_i))$, with $r_{2j-1} = r_{2j} = j^{pm}$, $p \geq 2$, $m > 0$. Let $Q$ be orthosymmetric, with weights $q_{2j-1} = q_{2j} = j^{2k}$, with $0 \leq k < m + 1/p - 1/2$. If $R_Q^*(\epsilon) >> \epsilon^2$ then the homogeneous variance approximation (2) holds.*

The theorem is proved in the appendix. All other Theorems in this paper follow from Lemmas proved in the appendix, and arguments in the main body of the paper.

Consider the problem of estimating the linear functional $L(\mathbf{x}) = \sum_i l_i x_i$ from data $\mathbf{u} = (u_i)$, where $u_i = x_i + v_i$, and the $v_i$ are orthogonal random variables with zero mean and common variance $\eta^2$. We suppose that we know *a priori* that $\mathbf{x} \in \mathbf{X}$, a convex subset of $l_2$. We use estimates from the class $\mathcal{A}$ of *affine rules* $\hat{L}(\mathbf{u}) = \sum_i \hat{l}_i u_i + c$, and we wish to attain the minimax affine risk

$$\inf_{\mathcal{A}} \sup_{\mathbf{X}} E(\hat{L}(\mathbf{u}) - L(\mathbf{x}))^2. \tag{3}$$

The problem is of interest here because of the following remark.

$$E(\hat{L}(\mathbf{u}) - L(\mathbf{x}))^2 = (\sum_i(\hat{l}_i - l_i)x_i + c)^2 + \eta^2 \sum_i \hat{l}_i^2. \tag{4}$$

Comparing (4) with the definition (1) of the pseudo-risk for quadratic estimation, we see that under the correspondence

$$\begin{array}{ccc} \hat{l}_i & \leftrightarrow & \hat{q}_i \\ l_i & \leftrightarrow & q_i \\ \eta^2 & \leftrightarrow & 2\epsilon^4 \\ x_i & \leftrightarrow & \theta_i^2 \end{array}$$

we get the precise equality

$$E(\hat{L}(\mathbf{u}) - L(\mathbf{x}))^2 = R(\hat{Q}, \theta). \tag{5}$$

Let us define

$$\begin{aligned}\mathbf{X} &= \{(\theta_i^2) : \theta \in \Theta\} \\ \mathcal{A}_S &= \{\hat{L} : \hat{L}(\mathbf{x}) = \sum_i \hat{l}_i x_i + e \\ & \qquad 0 \le \hat{l}_i \le l_i\}.\end{aligned}$$

Then we have, under the correspondence above,

$$R_Q(\epsilon) = \inf_{Q_{DS}} \sup_{\Theta} R(\hat{Q}, \theta) = \inf_{\mathcal{A}_S} \sup_{\mathbf{X}} E(\hat{L}(\mathbf{u}) - L(\mathbf{x}))^2. \qquad (6)$$

In other words, the minimax risk $R_Q$ may be evaluated by solving for the minimax risk among affine shrinkage estimates in a certain linear problem.

Say that $\mathbf{X}$ is *contractive* if the mapping $C_k$ defined by $C_k(\mathbf{x}) = (x_1, \ldots, x_{k-1}, 0, x_{k+1}, \ldots)$ is a contraction of $\mathbf{X}$:

$$C_k(\mathbf{X}) \subset \mathbf{X}.$$

If $\Theta$ is orthosymmetric and convex, one easily sees that the corresponding $\mathbf{X}$ is contractive. For $\theta = (\theta_1, \ldots \theta_{k-1}, +\theta_k, \theta_{k+1}, \ldots)$ and $\theta' = (\theta_1, \ldots \theta_{k-1}, -\theta_k, \theta_{k+1}, \ldots)$ are both in $\Theta$, hence their average $(\theta+\theta')/2 = (\theta_1, \ldots \theta_{k-1}, 0, \theta_{k+1}, \ldots)$ is in $\Theta$ and the property is evident.

**Lemma 2** *Suppose $\mathbf{X}$ is a contractive subset of the nonnegative orthant. Then for estimating the positive linear functional $L(\mathbf{x}) = \sum_i l_i x_i$ with $l_i \ge 0$, the minimax affine risk (3) is attained within the class $\mathcal{A}_S$ of affine shrinkage rules.*

In symbols

$$\begin{aligned}\inf_{\mathcal{A}_S} \sup_{\mathbf{X}} E(\hat{L}(\mathbf{u}) - L(\mathbf{x}))^2 &= \inf_{\mathcal{A}} \sup_{\mathbf{X}} E(\hat{L}(\mathbf{u}) - L(\mathbf{x}))^2 \qquad (7) \\ &= R_A^*(\eta; L, \mathbf{X}) \text{ say.}\end{aligned}$$

Combining (6) with (7) gives:

**Risk Isomorphism.** *Let $\Theta$ be convex and orthosymmetric. Then*

$$R_Q(\epsilon; Q, \Theta) = R_A^*(\eta; L, \mathbf{X})$$

*with $L$, $\mathbf{X}$, etc. defined according to the correspondence above. Moreover, if $L_0(\mathbf{u}) = \sum_i l_i u_i + e$ is a minimax affine shrinkage rule for $L$, then $Q_0(\mathbf{y}) = \sum_i l_i(y_i^2 - \epsilon^2) + e$ is a minimax quadratic shrinkage rule for $Q$ under the pseudo risk $R$.*

A thorough study of affine minimax estimation is given in Donoho (1989). Define the modulus of continuity

$$\Omega(\delta) = \sup\{|L(\mathbf{x}_1) - L(\mathbf{x}_{-1})| : ||\mathbf{x}_1 - \mathbf{x}_{-1}|| \leq \delta, \ \mathbf{x}_i \in \mathbf{X}\}.$$

Then, if $\mathbf{X}$ is convex, and if $\Omega(\delta) \to 0$ as $\delta \to 0$, we have, from Theorem 2 of Donoho (1989)

$$\Omega^2(\eta)/4 \leq R_A^*(\eta) \leq \Omega^2(\eta),$$

so that if $\Omega(\delta) \asymp \delta^r$ then $R_A^* \asymp \eta^{2r}$. More precisely, if $\Omega(\delta) \sim A\delta^r$ as $\delta \to 0$ then

$$R_A^*(\eta) \sim 2^{2r-2} r^r (1-r)^{1-r} \Omega^2(\eta)$$

as $\eta \to 0$. Finally, we can characterize the Affine Minimax estimator as follows. Suppose that $\mathbf{X}$ is a norm-bounded and norm-closed convex subset of $l_2$. Then let $\delta_0$ be the(!) maximizer of

$$\sup_\delta \frac{\Omega^2(\delta)\eta^2}{4\eta^2 + \delta^2}.$$

Suppose the modulus $\Omega(\delta_0)$ is attained by a pair $(\mathbf{x}_{-1}, \mathbf{x}_1)$, so that $L(\mathbf{x}_1) - L(\mathbf{x}_{-1}) = \Omega(\delta_0)$, $||\mathbf{x}_1 - \mathbf{x}_{-1}|| \leq \delta$, and $\mathbf{x}_1, \mathbf{x}_{-1} \in \mathbf{X}$. Then putting $\mathbf{x}_0 = (\mathbf{x}_1 + \mathbf{x}_{-1})/2$, the estimator

$$L_0(\mathbf{u}) = L(\mathbf{x}_0) + c_0 \frac{\Omega(\delta_0)}{\delta_0^2} < \mathbf{x}_1 - \mathbf{x}_{-1}, \mathbf{u} - \mathbf{x}_0 >$$

is minimax among affine estimates of $L$. Here $c_0 = \delta_0^2/(\delta_0^2 + 4\eta^2)$.

Via the isomorphism above, these results all have implications for quadratic estimation. But to apply them, we need to take care that $\mathbf{X}$ is convex.

**Definition.** The orthosymmetric set $\Theta$ is *quadratically convex* if the set $\mathbf{X} = \{(\theta_i^2) : \theta \in \Theta\}$ is convex.

Note that $l_p$-bodies $\Theta_p$ are quadratically convex iff $p \geq 2$.

**Theorem 3** *Suppose that $\Theta$ is orthosymmetric, convex, and quadratically convex. Define the modulus of continuity*

$$\omega(\delta) = \sup\{|Q(\theta_1) - Q(\theta_{-1})| : \sum_i (\theta_{1,i}^2 - \theta_{-1,i}^2)^2 \leq \delta^2, \theta_1, \theta_{-1} \in \Theta\}.$$

*Then*

$$\omega^2(\sqrt{2}\epsilon^2)/4 \leq R_Q(\epsilon) \leq \omega^2(\sqrt{2}\epsilon^2).$$

*In fact*

$$R_Q(\epsilon) = \sup_{\delta > 0} \frac{\omega^2(\delta)\, 2\epsilon^4}{8\epsilon^4 + \delta^2}. \tag{8}$$

*Moreover, if $\omega(\delta) \sim A\delta^r$ as $\epsilon \to 0$, then*

$$R_Q(\epsilon) \sim 2^{3r-2} r^r (1-r)^{1-r} \omega^2(\epsilon^2)$$

*Finally, suppose that $\Theta$ is norm-bounded. Then the supremum in (8) is attained at $\delta_0 > 0$. The modulus of continuity $\omega(\delta_0)$ is attained by some pair $(\theta_{-1}, \theta_1)$, and, putting $\theta_{0,i} = \sqrt{(\theta_{1,i}^2 + \theta_{-1,i}^2)/2}$, a minimax quadratic estimator for the pseudo-risk $R$ is*

$$Q_0(\mathbf{y}) = e_0 + \sum_i \hat{q}_i (y_i^2 - \epsilon^2),$$

*where $e_0 = \sum_i (q_i - \hat{q}_i)\theta_{0,i}^2$,*

$$\hat{q}_i = c_0 \frac{\omega(\delta_0)}{\delta_0^2} (\theta_{1,i}^2 - \theta_{-1,i}^2),$$

*and $c_0 = \frac{\delta_0^2}{8\epsilon^4 + \delta_0^2}$.*

# 3 Minimax Weights for Ellipsoids

We now specialize. Let $\Theta = \Theta_2((r_i))$ with $r_i$ a sequence of positive constants tending to $\infty$. Geometrically, this is a compact ellipsoid.

**Theorem 4** *Let $q_i, r_i \geq 0$ with $q_i/r_i$ decreasing to 0. With respect to the pseudo-risk $R$, the minimax quadratic estimator for $Q = \sum_i q_i\theta_i^2$ over $\Theta = \Theta_2((r_i))$ has weights*

$$\hat{q}_i = a_0(q_i - b_0 r_i)_+, i = 1, 2, \ldots$$

*for $a_0$ and $b_0$ determined as follows: Put*

$$\begin{aligned} g_1(b) &= \sum_i q_i(q_i - br_i)_+ \\ g_2(b) &= \sum_i (q_i - br_i)_+^2 \\ g_3(b) &= \sum_i r_i(q_i - br_i)_+. \end{aligned}$$

*[For each $b > 0$, each of these sums has a finite number of nonzero terms.] Then $b_0$ is the maximizer of*

$$J(b) = \frac{(g_1(b)/g_3(b))^2\, 2\epsilon^4}{8\epsilon^4 + g_2(b)/g_3(b)^2}$$

*and*

$$a_0 = \frac{g_1(b_0)/g_3(b_0)^2}{8\epsilon^4 + g_2(b_0)/g_3(b_0)^2}.$$

*Moreover,*

$$R_Q(\epsilon) = J(b_0).$$

So the optimal estimator has $\hat{q}_i = q_i a_0 (1 - b_0 \frac{r_i}{q_i})_+$. The "minimax weights" $(1 - b_0\frac{r_i}{q_i})_+$ are similar in form to those derived by Pinsker (1980) in solving a certain optimal filtering problem. We also remark that the optimal constant term $e_0$ for the estimator in question is

$$e_0 = b_0/2 + \frac{(1 - a_0)g_2(b_0)}{2g_3(b_0)}.$$

The proof of the Theorem results from applying Theorem 3 above with the following

**Lemma 3** *Let $q_i, r_i \geq 0$ and let $\psi(x)$ be positive and monotone increasing in $x$. Then the optimization problem*

$$\sup \sum_i q_i(x_{1,i} - x_{-1,i}) \qquad \textit{subject to}$$

$$\textstyle\sum_i (x_{1,i} - x_{-1,i})^2 \leq \delta^2$$

$$\textstyle\sum_i r_i \psi(x_{\cdot,i}) \leq 1$$

$$0 \leq x_{\cdot,i}$$

*has a solution with $\mathbf{x}_{-1} = \mathbf{0}$ identically. If $\psi(x) = x$, $\mathbf{x}_1$ is given by*

$$x_{1,i} = a(q_i - br_i)_+ \tag{9}$$

*where $a$ and $b$ satisfy*

$$a^2 \sum_i (q_i - br_i)_+^2 = \delta^2 \tag{10}$$

$$a \sum_i r_i(q_i - br_i)_+ = 1 \tag{11}$$

# 4 The White Noise Model

To relate the results of the last two sections to the problem of the introduction we take one intermediate step. Consider (yet another!) estimation problem: we observe

$$Y(t) = \int_{-\pi}^{t} f(u)du + \epsilon W(t), \quad t \in [-\pi, \pi] \tag{12}$$

where $W$ is a Wiener Process, started at $W(-\pi) = 0$, i.e. a Gaussian process with $EW(t) = 0$, $Cov(W(t), W(s)) = \pi + min(t, s)$. We are again interested in estimating $Q(f) = \int_{-\pi}^{\pi} (f^{(m)}(t))^2 dt$ and we again know that $f \in \mathcal{F}_m$.

An isometry reduces this to a problem treated in sections 2 and 3. Define an orthonormal set of functions $(\varphi_i)_{i=1}^{\infty}$ in $L_2[-\pi, \pi]$ by the rules $\varphi_{2j-1}(t) = \frac{1}{\sqrt{\pi}} sin(jt)$ and $\varphi_{2j}(t) = \frac{1}{\sqrt{\pi}} cos(jt)$, for $j = 1, 2, \ldots$. With respect to this system, $f$ has the Fourier-Bessel coefficients $\theta_i(f)$, where

$$\theta_i(f) = \int_{-\pi}^{\pi} \varphi_i(t) f(t) dt.$$

Define the empirical Fourier-Bessel coefficients

$$y_i = \int_{-\pi}^{\pi} \varphi_i(t) Y(dt);$$

then $y_i = \theta_i + z_i$ with the $z_i$ i.i.d. $N(0, \epsilon^2)$. Now define $(q_i)$ by $q_{2j-1} = q_{2j} = j^{2k}$ and $r_{2j-1} = r_{2j} = j^{2m}$. One verifies that if $f \in \mathcal{F}_m$,

$$\int_{-\pi}^{\pi} (f^{(k)}(t))^2 dt = \sum_i q_i \theta_i^2;$$

and

$$\int_{-\pi}^{\pi} (f^{(m)}(t))^2 dt = \sum_i r_i \theta_i^2.$$

Hence the estimation problem of this section becomes a problem from section 2, with $Q(\theta) = \sum_i q_i \theta_i^2$, and $\Theta = \Theta_2((r_i))$.

Although the $\varphi_i$ do not make up a complete orthonormal system (they are missing the constant function), the coefficients $(y_i)$

are *sufficient*, in the measure theoretic sense, for the problem we consider. Hence, for our purposes, observing $Y$ is completely equivalent to observing the $\mathbf{y} = (y_i)$. It follows that the minimax risk for estimating $\int_{-\pi}^{\pi} (f^{(m)}(t))^2 dt$ using quadratic functions of $Y$ is equivalent to estimating $Q(\theta)$ using quadratic functions of $\mathbf{y}$, etc.

Applying Theorems 2,3, and 4, we easily get asymptotically minimax quadratic estimates of $Q$. To work out the asymptotics, note that

$$g_1(b) = 2\sum j^{2k}(j^{2k} - bj^{2m})_+ \sim 2\int_1^\infty \lambda^{2k}(\lambda^{2k} - b\lambda^{2m})_+ d\lambda$$
$$\sim b^{\frac{-4k-1}{2m-2k}} [\frac{2}{4k+1} - \frac{2}{2k+2m+1}]$$

as $b \to 0$, and, similarly,

$$g_2(b) \sim b^{\frac{-4k-1}{2m-2k}} [\frac{2}{4k+1} - \frac{4}{2k+2m+1} + \frac{2}{4m+1}]; \quad b \to 0 \quad (13)$$

$$g_3(b) \sim b^{\frac{-2k-2m-1}{2m-2k}} [\frac{2}{2k+2m+1} - \frac{2}{4m+1}]; \quad b \to 0. \quad (14)$$

These calculations, together with Theorem 4, give immediately that

$$R_Q(\epsilon) \sim 2^{3r-2}(1-r)^{r-1} r^r [2k+2m+1]^{-2r} \epsilon^{4r}. \quad (15)$$

We also get asymptotics for the optimal $a_0$ and $b_0$ in Theorem 4. As $\epsilon \to 0$,

$$a_0(\epsilon) \to 1,$$
$$b_0(\epsilon) \sim 2^{3r/2} r^{r/2} [2k+2m+1]^{-r/2} \epsilon^{2r}.$$

and, as $g_2(b_0)/g_3(b_0) = O(1)$ while $a_0 \sim 1$, the optimal centering constant $c_0 \sim b_0/2$. It turns out that, to get asymptotic minimaxity, it is enough to use the asymptotic forms in these relations. So define

$$\beta(\epsilon) = 2^{3r/2} r^{r/2} [2k+2m+1]^{-r/2} \epsilon^{2r} \quad (16)$$

and

$$\hat{q}_i = (q_i - \beta r_i)_+. \quad (17)$$

With extra calculations, which we omit, one sees that although the coefficients (17) are not exactly minimax for any $\epsilon > 0$, the excess risk is of smaller order than $R_Q$. This implies:

**Theorem 5** *Let* $r = \frac{4m-4k}{4m+1} < 1/2$. *Put* $y_i = \int_{-\pi}^{\pi} \varphi_i Y(dt)$, $i = 1, 2, \ldots$. *Let* $\beta$ *and* $\hat{q}_i$ *be defined as in (16) and (17). Then the estimator*

$$Q_0(\mathbf{y}) = \beta/2 + \sum_i \hat{q}_i (y_i^2 - \epsilon^2)$$

*is asymptotically minimax among quadratic estimates:*

$$\sup_{\mathcal{F}_m} E(\hat{Q}(\mathbf{y}) - Q(f))^2 \sim R_Q(\epsilon) \quad as\ \epsilon \to 0$$

*and the minimax pseudo-risk* $R_Q(\epsilon)$ *obeys (15).*

## 5 White Noise and sampled data

We are now in a position to solve the problem of the introduction. Note that observing $v_i = f(t_i) + z_i$ is the same as observing

$$Y_n(t) = \frac{2\pi}{n} \sum_{t_i \le t} f(t_i) + \frac{2\pi}{n} \sum_{t_i \le t} z_i, \quad t \in [-\pi, \pi]. \tag{18}$$

But, if we have $\frac{2\pi}{n} z_i = \epsilon(W(t_i) - W(t_{i-1}))$ this is visibly a Riemann sum approximation to the white noise equation (12). Hence, under the calibration

$$\epsilon = \sqrt{\frac{2\pi\sigma^2}{n}}. \tag{19}$$

we expect the sampling model and the white noise model to be essentially equivalent.

Let us be more precise. Define the empirical Fourier-Bessel coefficient

$$\tilde{y}_i = \int_{-\pi}^{\pi} \varphi_i(t) Y_n(dt).$$

We propose to act as if the data $\tilde{y}_i$ were equivalent to $y_i$. We will apply the estimate $Q_0$ designed for use with $(y_i)$ at this noise level (19) to the data $(\tilde{y}_i)$ instead.

The reader will note that the resulting estimate $Q_0(\tilde{\mathbf{y}})$ is precisely the estimate mentioned in the introduction. To see this, compare Theorem 5 with Theorem 1, keeping in mind (19); and also the analytic fact that the complex Fourier transform is related to our real orthogonal transform via

$$2\sqrt{\pi} w_j = \tilde{y}_{2j} + \sqrt{-1}\, \tilde{y}_{2j-1}, \quad j = 1, 2, \ldots.$$

Theorem 1 follows from Theorem 5 by an approximation argument. The argument has two halves. First, we show that under the calibration (19), $Q_0(\tilde{\mathbf{y}})$ is asymptotically equivalent to $Q_0(\mathbf{y})$. Second, we show that no estimator based on $Y_n$ is better than $Q_0(\tilde{\mathbf{y}})$.

Putting

$$\tilde{\theta}_i(f) = \frac{2\pi}{n} \sum_{u=1}^{n} \varphi_i(t_u) f(t_u)$$

we have $\tilde{y}_i = \tilde{\theta}_i + z_i$, where the $z_i$ are i.i.d. $N(0, \epsilon^2)$ for $1 \le i < n$. For a given function $f$, compare the empirical coefficients $(\tilde{y}_i)$ with $(y_i)$. Both have noise which is i.i.d. $N(0, \epsilon^2)$ for $1 \le i < n$. On the other hand, the difference between the signal terms $\tilde{\theta}_i(f)$ and $\theta_i(f)$ is not large either.

**Lemma 4** *If $f$ is a real trigonometric polynomial of degree $< n/2$, so that $\theta_i(f) = 0$ for $i \ge n-1$, then*

$$\tilde{\theta}_{i,n}(f) = \theta_i(f), \quad i = 1, 2, \ldots n-1. \tag{20}$$

*Also, if $f \in \mathcal{F}_m$ then*

$$|\tilde{\theta}_{i,n}(f) - \theta_i(f)|^2 \le \gamma_m n^{-2m}, \quad i = 1, \ldots, n-1$$

*for a numerical constant $\gamma_m$ which is finite if $m > 1/2$.*

This leads to

**Lemma 5** *Let $r < 1/2$ and $4m^2 > m + 1$. With $\epsilon$ as in (19), on an appropriate probability space,*

$$\sup_{\mathcal{F}_m} E(Q_0(\mathbf{y}) - Q_0(\tilde{\mathbf{y}}))^2 = o(n^{-2r})$$

From these lemmas, it follows immediately that

$$\sup_{\mathcal{F}_m} E(Q_0(\tilde{\mathbf{y}}) - Q(f))^2 \le R_Q(\epsilon)(1 + o(1)).$$

This completes the first half of the proof of Theorem 1. For the second half, we argue that actually the reverse holds as well: for all sufficiently large $n$,

$$\inf_{Q_n} \sup_{\mathcal{F}_m} E(\hat{Q}(\mathbf{v}) - Q(f))^2 \ge R_Q(\epsilon). \tag{21}$$

Hence the estimator $Q_0(\tilde{\mathbf{y}})$ is *asymptotically minimax* among quadratic estimates based on $Y_n$.

For fixed $k$, define $\mathcal{F}_{m,k}$ to be the subset of $\mathcal{F}_m$ consisting of those $f$ with $\theta_i(f) = 0$ for $i \ge k$. If $k < n - 1$, we may apply the quadrature formula (20) to get that, on an appropriate probability space, the first $n-1$ empirical Fourier-Bessel Coefficients are identical in the two different models:

$$\tilde{y}_i = y_i, \quad i = 1, \ldots, n-1. \tag{22}$$

Define

$$\Theta_m = \{\theta = (\theta_i(f)) : f \in \mathcal{F}_m\}$$

and

$$\Theta_{m,k} = \{\theta = (\theta_i(f)) : f \in \mathcal{F}_{m,k}\}.$$

For $k < n-1$, we have

$$\inf_{Q} \sup_{\Theta_{m,k}} E(\hat{Q}(\tilde{\mathbf{y}}) - Q(\theta))^2 = \inf_{Q} \sup_{\Theta_{m,k}} E(\hat{Q}(\mathbf{y}) - Q(\theta))^2. \tag{23}$$

We need the following:

**Lemma 6** *Define $(q_i)$ by $q_{2j-1} = q_{2j} = j^{2k}$ and $(r_i)$ by $r_{2j-1} = r_{2j} = j^{2m}$. Then the solution vector $\mathbf{x}_1$ provided by Lemma 3 has $x_{1,i} = 0$ for $i > n_0(\delta)$, say. If $m > 1/4$,*

$$n_0(\delta) = o(\delta^{-1}) \tag{24}$$

Its implications can be summarized as follows. Given $\tau = (\tau_i)$, define the hyperrectangle

$$\Theta(\tau) = \{\theta : \theta_i^2 \le \tau_i,\ i = 1, 2, \ldots\}.$$

If $\Theta(\tau) \subset \Theta$ we call the problem of minimax estimation of $Q$ over $\Theta(\tau)$ a rectangular subproblem of $\Theta$.

**Lemma 7** *Let $r_{2j-1} = r_{2j} = j^{2m}$ and $q_{2j-1} = q_{2j} = j^{2k}$, where $k < m$. For the pseudo-risk $R$, there is a rectangular subproblem of $\Theta_m$ which is equally as hard as the full problem, i.e. a sequence $\tau = (\tau_i)$ such that*

$$\inf_{Q} \sup_{\Theta} R(\hat{Q}, \theta) = \inf_{Q} \sup_{\Theta(\tau)} R(\hat{Q}, \theta). \tag{25}$$

*Moreover, the subproblem is $n_0(\delta_0)$-dimensional,*

$$\Theta(\tau) \subset \Theta_{m,n_0} \tag{26}$$

*where $n_0(\delta)$ was defined in Lemma 6; and $\delta_0 \asymp \epsilon^2$.*

Combining these facts, and noting that (24) together with (19) give $n_0(\delta_0) = o(n)$, we get that for all sufficiently large $n$, $n_0(\delta_0) < n - 1$ and, as $m > 1/4$,

$$\begin{aligned}
\inf_{Q_n} \sup_{\mathcal{F}_m} E(\hat{Q}(\mathrm{v}) - Q(f))^2 &\geq \inf_{Q_n} \sup_{\mathcal{F}_{m,n_0}} E(\hat{Q}(\mathrm{v}) - Q(f))^2 \\
&= \inf_{Q} \sup_{\Theta_{m,n_0}} E(\hat{Q}(\tilde{y}) - Q(\theta))^2 \\
&= \inf_{Q} \sup_{\Theta_{m,n_0}} E(\hat{Q}(\mathrm{y}) - Q(\theta))^2 \quad \text{[by (23)]} \\
&\geq \inf_{Q} \sup_{\Theta_{m,n_0}} R(\hat{Q}, \theta) \\
&= \inf_{Q} \sup_{\Theta_m} R(\hat{Q}, \theta) \quad \text{[by (25)-(26)]} \\
&= R_Q(\epsilon).
\end{aligned}$$

and (21) is proven.

Our approach to Theorem 1 may be summarized as follows. We solved the problem for the white noise observations (12). Then we showed that sampled data (18) are in some sense equivalent to white noise observations (12). See Nussbaum (1985) for a specific instance, and Low(1988), Donoho and Low (1989), for some general results on "white noise approximation" in linear problems. The notion of rectangular subproblems which are equally as hard as the full problem arises, in a different context, in Donoho, Liu, and MacGibbon (1989).

# 6 Rate Optimality

We now turn to the optimality, as regards rate, of our proposed estimate, not just among quadratic estimates, but among *all* estimates. To discuss this fully, we first consider the white noise model of section 2. As we saw there $R_Q^* \asymp \omega^2(\epsilon^2)$. We now show that the rate $\omega^2(\epsilon^2)$ cannot be exceeded even by using arbitrary measurable estimates.

**Theorem 6** *Let $\Theta$ be quadratically convex and orthosymmetric set, bounded in $l_2$-norm, so that $\sum \theta_i^2 \leq M < \infty$. Let $Q(\theta) = \sum q_i \theta_i^2$ with $q_i \geq 0$. For small $c > 0$, there exists $\alpha = \alpha(c, M) > 0$ so that*

$$\inf_{\hat{Q}} \sup_{\Theta} E(\hat{Q}(\mathbf{y}) - Q(\theta))^2 \geq \omega^2(c\epsilon^2)\alpha(c, M).$$

The proof is based on the following Bayesian hypothesis testing argument. Let $y_i = \tilde{\theta}_i + z_i$, where as before the $z_i$ are i.i.d. $N(0, \epsilon^2)$, but now $\tilde{\theta}$ is a *random variable*, and $\tilde{\theta}$ is independent of $(z_i)$. Consider two different probability distributions $\mu_0$, $\mu_1$ for $\tilde{\theta}$ with the properties

$$Q(\tilde{\theta}) = q_0 \ a.s. \ [\mu_0] \tag{27}$$

and

$$Q(\tilde{\theta}) = q_1 \ a.s. \ [\mu_1]. \tag{28}$$

Thus $\mu_0$ and $\mu_1$ concentrate on certain level sets of the functional in question. In addition, suppose that

$$Supp(\mu_i) \subset \Theta. \tag{29}$$

Let $P_{0,\epsilon}$ and $P_{1,\epsilon}$ denote the marginal distributions of $\mathbf{y}$ that result.

Now estimating $Q$ to within a precision finer than $(q_1 - q_0)/2$ is no easier than performing a hypothesis test between $H_0 : \theta \sim \mu_0$ and $H_1 : \theta \sim \mu_1$. Indeed, given an estimator $\hat{Q}$, we can always test hypotheses by saying: accept $H_0$ if $\hat{Q} \leq (q_1 + q_0)/2$ and reject otherwise. Therefore, if we can show that no hypothesis test performs very well in this setting then no estimator can consistently have errors smaller than $(q_1 - q_0)/2$.

Formalize this (compare Donoho and Liu (1988)). Define the testing affinity

$$\pi(P_0, P_1) = \inf_{\psi \text{ mble}} E_{P_0}\psi + E_{P_1}(1 - \psi).$$

which measures the sum of type I and type II errors of the best test between $P_0$ and $P_1$. Then for any measurable function of the data,

$$(E_{P_{0,\epsilon}}(\hat{Q}(\mathbf{y}) - q_0)^2 + E_{P_{1,\epsilon}}(\hat{Q}(\mathbf{y}) - q_1)^2) \geq [(q_1 - q_0)/2]^2 \, \pi(P_{1,\epsilon}, P_{0,\epsilon})$$

Now, as Max $\geq$ Sum/2, we get

$$\inf_{\hat{Q}} \sup_{Supp(\mu_0) \cup Supp(\mu_1)} E(\hat{Q}(\mathbf{y}) - Q(\theta))^2 \geq [(q_1 - q_0)^2/8] \, \pi(P_{1,\epsilon}, P_{0,\epsilon})$$

and as $Supp(\mu_0) \cup Supp(\mu_1) \subset \Theta$

$$\inf_{\hat{Q}} \sup_{\Theta} E(\hat{Q}(\mathbf{y}) - Q(\theta))^2 \geq [(q_1 - q_0)^2/8] \, \pi(P_{1,\epsilon}, P_{0,\epsilon}).$$

Let us assume for the moment that $\{(\theta_i^2)\}$ is closed in the $l_2$ norm. Under this assumption, we will now exhibit a pair $\mu_0$, $\mu_1$ so that

$$q_1 - q_0 = \omega(c\epsilon^2) \tag{30}$$
$$\pi(P_{1,\epsilon}, P_{0,\epsilon}) \geq 8\alpha(c, M) > 0. \tag{31}$$

and the theorem will follow.

We describe the construction. Let $\delta = c\epsilon^2$. Applying Lemma 2 of Donoho (1989), the convexity, $l_2$ norm-closure and norm-boundedness of $\{(\theta_i^2)\}$ imply the modulus $\omega(\delta)$ is attained by some pair $(\theta_1, \theta_{-1})$. Let $s_i$ be an i.i.d. sequence of random variables taking just the values $+1$ and $-1$, each with probability $1/2$. Define $\mu_0$ by

$$\tilde{\theta} = (s_i \theta_{-1,i})$$

and $\mu_1$ by

$$\tilde{\theta} = (s_i \theta_{1,i}).$$

Thus, in each case, $\tilde{\theta}$ amounts to randomly changing the signs on a constant vector. Geometrically, $\mu_0$ and $\mu_1$ concentrate on the vertices of two hyperrectangles. Note that as $\Theta$ is orthosymmetric, the condition (29) is met. Also, putting $q_0 = Q(\theta_{-1})$, $q_1 = Q(\theta_1)$, conditions (27), (28) hold. Now by construction, $Q(\theta_1) - Q(\theta_{-1}) = \omega(c\epsilon^2)$. Thus (30) holds, and the theorem reduces to verifying (31). This follows from:

**Lemma 8** *Let $\{z_i\}$ be i.i.d. $N(0, \epsilon^2)$, $s_i$ be i.i.d. with $P(s_i = 1) = 1/2$, and $P(s_i = -1) = 1/2$, and $(s_i)$ independent of the $(z_i)$. Let $(\theta_{1,i})$ and $(\theta_{-1,i})$ be sequences of positive constants each satisfying $\sum \theta_{\cdot,i}^2 \le M$. Let $P_{1,\epsilon}$ denote the probability law of $(y_i)$ when*

$$y_i = s_i \theta_{1,i} + z_i, i = 1, 2, \ldots.$$

*Let $P_{0,\epsilon}$ denote the probability law of $(y_i)$ when*

$$y_i = s_i \theta_{-1,i} + z_i, i = 1, 2, \ldots.$$

*For all sufficiently small $c$, there exists an absolute constant $\alpha(c, M)$ 0 so that*

$$\sum_i (\theta_{1,i}^2 - \theta_{-1,i}^2)^2 \le c^2 \epsilon^4$$

*implies*

$$\pi(P_{1,\epsilon}, P_{0,\epsilon}) \ge 8\alpha(c, M)$$

The proof of this lemma is given in the appendix. Modulo the assumption of closedness, the theorem is proven. Now even without closedness, we may for each $n > 0$ find a pair $\mu_0$, $\mu_1$ attaining (30) to within $1/n$ and also satisfying (31). The theorem therefore follows in the more general case as well.

For many applications of this theorems, it would suffice to use the approach of Fan (1988). Our approach may be compared to Fan's by saying that we test between vertices of two (possibly infinite-dimensional) hyperrectangles while he tests between vertices of two finite-dimensional hypercubes.

What seems innovative in our approach is that the hyperrectangles we use are automatically derived for us from the modulus of continuity, rather than being found by trial and error. Our approach has a practical advantage, in that it is nonasymptotic, and a conceptual advantage as well. It directly shows the (unusual) "distance"

$$\sum_i (\theta_{1,i}^2 - \theta_{-1,i}^2)^2$$

to be the key quantity in deriving a lower bound for nonregular quadratic functionals.

In any event, let us apply this to the model of the introduction.

**Lemma 9** *Let $r_{2j-1} = r_{2j} = j^{2m}$ and $q_{2j-1} = q_{2j} = j^{2k}$, where $k < m$. Let $\eta > 0$. There is a rectangular subproblem $\Theta(\tau)$ of $\Theta_m$ with*

$$\inf_{\hat{Q}} \sup_{\Theta(\tau)} E(\hat{Q}(\mathbf{y}) - Q(\theta))^2 \geq \omega^2(c\epsilon^2)\alpha(c,1) \tag{32}$$

*where the infimum is over all measurable estimates. Moreover, the subproblem is $n_0(c\epsilon^2)$-dimensional,*

$$\Theta(\tau) \subset \Theta_{m,n_0} \tag{33}$$

*where $n_0$ was defined in Lemma 6.*

Applying again (22), together with $n_0(c\epsilon^2) = o(n)$, we get that for all sufficiently large $n$, $n_0(c\epsilon^2) < n - 1$, and so

$$\begin{aligned} \inf_{\hat{Q}} \sup_{\mathcal{F}_m} E(\hat{Q}(\mathbf{v}) - Q(f))^2 &\geq \inf_{\hat{Q}} \sup_{\mathcal{F}_{m,n_0}} E(\hat{Q}(\mathbf{v}) - Q(f))^2 \\ &= \inf_{\hat{Q}} \sup_{\Theta_{m,n_0}} E(\hat{Q}(\mathbf{y}) - Q(\theta))^2 \\ &\geq \omega^2(c\epsilon^2)\alpha(c,1) \quad \text{[by (32)]} \end{aligned}$$

But $\omega^2(c\epsilon^2)$ goes to zero at the same rate as $R_Q(\epsilon)$. Hence quadratic estimates are rate-optimal.

## 7 Comparison: Optimal Recovery

Suppose we observe

$$u_i = x_i + v_i$$

where we know $\mathbf{x} \in \mathbf{X}$ a priori, and the noise $\mathbf{v} = (v_i)$ is now *deterministic*, satisfying $\|\mathbf{v}\|_2 \leq \eta$. We wish to estimate $J$, a general, nonlinear functional. We evaluate performance of an estimator $\hat{J}$ by the worst-case error

$$Err(\hat{J}, \mathbf{x}) = \sup_{\|\mathbf{v}\| \leq \eta} |\hat{J}(\mathbf{u}) - J(\mathbf{x})|,$$

and we are interested in procedures attaining the minimax error

$$E^*(\eta; J, \mathbf{X}) = \inf_{\hat{J}} \sup_{\mathbf{x} \in \mathbf{X}} Err(\hat{J}, \mathbf{x}).$$

This is the standard problem of *optimal recovery* of a functional $J$ from noisy data (see e.g. Micchelli and Rivlin, 1977). Define the modulus of continuity of $J$,

$$\Omega(\delta; J, \mathbf{X}) = \sup\{|J(\mathbf{x}) - J(\mathbf{y})| : \|\mathbf{x} - \mathbf{y}\| \leq \delta, \mathbf{x}, \mathbf{y} \in \mathbf{X}\}.$$

If $\mathbf{X}$ is convex, then the "central algorithm" (Traub, Wasilkowski, Woźniakowski, 1983, 1988) is minimax, with error

$$E^*(\eta; J, \mathbf{X}) = \Omega(2\eta)/2.$$

The results of this paper, where the noise is assumed random, make for an interesting comparison. As discussed in section 2, if $J$ is an *affine* functional, and the observations are contaminated with random white noise of variance $\eta^2$, the minimax root mean squared error is between $\Omega(\eta)/2$ and $\Omega(\eta)$. Hence estimating an affine functional $J$, with a priori information $\mathbf{X}$, leads to essentially the same difficulty of estimation, whether the noise is deterministic and chosen by an adversary, subject to the constraint that the norm of the noise vector be no larger than $\eta$, or whether the noise is random and of variance $\eta^2$. Formally,

$$(R^*(\eta; J, \mathbf{X}))^{1/2} \asymp E^*(\eta; J, \mathbf{X}), \quad \eta \to 0\,.$$

In fact there exist estimators which perform very well in both problems. Donoho (1989) has a fuller discussion of the correspondence between the statistical estimation problem and the optimal recovery problem in the case of estimating affine functionals.

The results of this paper show that for estimating quadratic functionals, the correspondence no longer holds. In the statistical problem, the modulus $\omega$, rather than $\Omega$, controls the difficulty of estimation. Not only are the two moduli defined differently, they can have completely different asymptotics. For example, consider the functional $Q(\theta) = \sum_i \theta_i^2$. Consider the ellipsoidal class $\Theta_m$ defined by the constraint that $\sum_i i^{2m}\theta_i^2 \leq 1$. Then $Q \leq 1$ on this class. It follows that $\Omega(\delta; Q, \Theta_m) \leq 2\delta$, for every $m > 0$. On the other hand, with $r = 4m/(4m+1)$, then $\omega(\delta; Q, \Theta_m) \asymp \delta^r$. Hence, if $r < 1/2$, there is no longer a comparability between deterministic noise of size $\eta$ and statistical noise of variance $\eta^2$. The statistical problem is harder, in the sense that

$$(R^*(\eta; Q, \Theta))^{1/2} \asymp \eta^{2r} >> E^*(\eta; Q, \Theta) \asymp \eta, \quad \eta \to 0.$$

# 8 Appendix: Proofs

## 8.1 Proof of Lemma 1

Consider the operation $(\hat{q}_i) \mapsto (\tilde{q}_i)$ defined by

$$\tilde{q}_i = \max(0, \min(q_i, \hat{q}_i)),\ i = 1, \ldots.$$

We will show that with an optimal choice of constant $\tilde{a}$, the induced estimator $\tilde{Q}(y) = \tilde{a} + \sum_i \tilde{q}_i(y_i^2 - \epsilon^2)$ has, better worst case, MSE than does $\hat{Q}(y)$. Write $MSE = Bias^2 + Var$, where

$$Bias(\hat{Q}, \theta) = \sum_i (\hat{q}_i - q_i)\theta_i^2 + \hat{a}$$

$$Var(\hat{Q}, \theta) = 4\epsilon^2 \sum_i \hat{q}_i^2 \theta_i^2 + 2\epsilon^4 \sum_i \hat{q}_i^2.$$

As $\tilde{q}_i^2 \leq \hat{q}_i^2$ for all $i$,

$$Var(\tilde{Q}, \theta) \leq Var(\hat{Q}, \theta),\ \theta \in \Theta \tag{34}$$

Define

$$B_+(\hat{q}) = \sup_\Theta \sum_i (\hat{q}_i - q_i)\theta_i^2$$

$$B_-(\hat{q}) = \inf_\Theta \sum_i (\hat{q}_i - q_i)\theta_i^2$$

Then

$$\inf_a \sup_\Theta Bias^2(a + \sum_i \hat{q}_i(y_i^2 - \epsilon^2), \theta) = (B_+(\hat{q}) - B_-(\hat{q}))^2/4 \tag{35}$$

and the optimal choice of $a$ is

$$\hat{a} = -(B_+(\hat{q}) + B_-(\hat{q}))/2.$$

Similar formulas hold for $\tilde{q}$.

Now we invoke orthosymmetry and convexity. This implies, as in section 2, that if $\mathcal{I}$ is any set of subscripts, and $\theta \in \Theta$, then $\tau = (\tau_i)$ defined by

$$\tau_i = \theta_i 1_{i \in \mathcal{I}}$$

is also an element of $\Theta$.

Now let $\mathcal{I}_+ = \{i : \hat{q}_i > q_i\}$. Defining $\tau$ as above, we get

$$\sum_i \tau_i^2(\hat{q}_i - q_i) \geq 0.$$

and so

$$B_+(\hat{q}) \geq 0.$$

By the same token, $B_+(\tilde{q}) \leq 0$.

On the other hand, let $\mathcal{I}_- = \{i : \hat{q}_i \leq q_i\}$. Given an arbitrary $\theta \in \Theta$, define $\tau$ using $\mathcal{I} = \mathcal{I}_-$. Now

$$\sum_i (\hat{q}_i - q_i)\theta_i^2 \geq \sum_i (\hat{q}_i - q_i)\tau_i^2 = \sum_{i \in \mathcal{I}_-} (\hat{q}_i - q_i)\theta_i^2$$

and so

$$B_-(\hat{q}) = \inf_\Theta \sum_{i \in \mathcal{I}_-} (\hat{q}_i - q_i)\theta_i^2.$$

On the other hand, for $i \in \mathcal{I}_-$,

$$(\hat{q}_i - q_i)\theta_i^2 \leq (\tilde{q}_i - q_i)\theta_i^2$$

termwise, for each $\theta$. Therefore,

$$B_-(\hat{q}) \leq B_-(\tilde{q})$$

We conclude from the above that

$$B_+(\hat{q}) - B_-(\hat{q}) \geq B_+(\tilde{q}) - B_-(\tilde{q})$$

which implies, via (34)-(35), that $\tilde{Q}$ has smaller worst-case MSE than $\hat{Q}$.

## 8.2 Proof of Theorem 2

If $2k < m + 1/p - 1/2$ then $Q(\theta)$ is bounded on $\theta$, and so the simple arguments given in the paragraphs before the statement of the theorem suffice to establish the conclusion. So suppose that $2k \geq m + 1/p - 1/p$.

By the same type of asymptotics as in Section 4, putting

$$r = \frac{m + 1/p - k - 1/2}{m + 1/p - 1/4}, \tag{36}$$

then if $r < 1/2$,

$$R_Q(\epsilon) \asymp \epsilon^{4r}, \quad \epsilon \to 0$$

Hence the result is equivalent to

$$\sup_{\Theta} \epsilon^2 \sum \hat{q}_i^2 \theta_i^2 << \epsilon^{4r}. \tag{37}$$

We will show below that if we know that

$$\sum \hat{q}_i^2 \leq M \tag{38}$$

and also $0 \leq \hat{q}_i \leq q_i$ then

$$\sup\{\sum \hat{q}_i^2 \theta_i^2 : \theta \in \Theta_p((r_i))\} \leq C_1 M^{\frac{4k-2m+1/s}{4k+1}}, \quad M \to \infty \tag{39}$$

where $s = (1 - 2/p)^{-1}$. On the other hand, if $\hat{Q}$ is minimax quadratic for the pseudo-risk $R$, then from

$$R(Q,\theta) \geq \dot{Var}(\hat{Q},\theta) = 2\epsilon^4 \sum \hat{q}_i^2$$

we have

$$R_Q(\epsilon) \geq 2\epsilon^4 \sum \hat{q}_i^2.$$

We may therefore take

$$M = R_Q(\epsilon)\epsilon^{-4} \sim C_2 \epsilon^{4(r-1)}, \quad \epsilon \to 0$$

in (38) and so, by (39) and some calculation

$$\epsilon^2 \sum \hat{q}_i^2 \theta_i^2 \leq C_3 \epsilon^{4r+1/(2m+2/p-1/2)} = o(\epsilon^{4r}), \quad \epsilon \to 0.$$

It remains only to prove (39). We wish to evaluate

$$\sup\{\sum \hat{q}_i^2 \theta_i^2 : \quad \begin{array}{l} \sum r_i |\theta_i|^p \leq 1 \\ 0 \leq \hat{q}_i \leq q_i \\ \sum \hat{q}_i^2 \leq M\} \end{array}$$

By Hölder this is

$$\sup\{(\sum \hat{q}_i^{2s}/r_i^{2s/p})^{1/s} : \quad 0 \le \hat{q}_i \le q_i, \quad \sum \hat{q}_i^2 \le M\}.$$

Now as $r_i^{-2s/p}$ is decreasing in $i$, the answer is to set $\hat{q}_i = q_i$ for $i < m_0$ and $\hat{q}_i = 0$ for $i > m_0$, for $m_0 = \inf\{i : \sum_{j=1}^{i} q_i^2 \ge M\}$.

Thus the value of the problem is bounded above by

$$\begin{aligned}(\sum_{i=1}^{m_0} q_i^{2s}/r_i^{2s/p})^{1/s} &\asymp (m_0^{2s(2k-m)+1})^{1/s} \\ &= m_0^{2(2k-m)+1/s}\end{aligned}$$

Now $m_0 \asymp M^{\frac{1}{4k+1}}$, which, together with the last display, gives (39).

## 8.3 Proof of Lemma 2

Note that

$$\inf_a \sup_{\mathbf{X}} (\hat{L}(\mathbf{x}) - L(\mathbf{x}))^2 + \eta^2 \sum \hat{l}_i^2 = (B_+(\hat{l}) - B_-(\hat{l}))^2/4 + \eta^2 \sum \hat{l}_i^2.$$

where

$$B_+(\hat{l}) = \sup_{\mathbf{X}} \sum (\hat{l}_i - l_i) x_i$$

etc. From this point on, the argument is similar to that for Lemma 1.

## 8.4 Proof of Lemma 3

**First, we show that we may take $x_{1,i} \ge x_{-1,i}$ for $i = 1, \dots,$. Indeed, if $(\mathbf{x}_1, \mathbf{x}_{-1})$ is a candidate for solution, then by relabelling if necessary, we may suppose that $\sum q_i x_{1,i} > \sum q_i x_{-1,i}$. Define**

$$\tilde{x}_{-1,i} = \min(x_{-1,i}, x_{1,i})$$

$$\tilde{x}_{1,i} = x_{1,i}.$$

The new pair is at least as good as the original one: it makes

$$\sum q_i(\tilde{x}_{1,i} - \tilde{x}_{-1,i}) \geq \sum q_i(x_{1,i} - x_{-1,i})$$

and so the new objective value is at least as good; but on the other hand we have

$$\sum r_i\psi(\tilde{x}_{-1,i}) \leq \sum r_i\psi(x_{-1,i})$$

and

$$\sum(\tilde{x}_{1,i} - \tilde{x}_{-1,i})^2 \leq \sum(x_{1,i} - x_{-1,i})^2 \leq \delta^2$$

and so the new pair is still feasible.

Second, we show that among pairs with $x_{1,i} \geq x_{-1,i}$ element-wise, we may take $\mathbf{x}_{-1} = 0$ identically. Indeed, define

$$\tilde{x}_{1,i} = x_{1,i} - x_{-1,i}$$

$$\tilde{x}_{-1,i} = 0.$$

Then

$$\sum q_i(\tilde{x}_{1,i} - \tilde{x}_{-1,i}) = \sum q_i(x_{1,i} - x_{-1,i})$$

and

$$\sum(\tilde{x}_{1,i} - \tilde{x}_{-1,i})^2 = \sum(x_{1,i} - x_{-1,i})^2 \leq \delta^2$$

but, as $\psi$ is monotone

$$\sum r_i\psi(\tilde{x}_{1,i}) \leq \sum r_i\psi(x_{1,i})$$

etc. Hence the new pair is still feasible, and delivers the same objective value.

Third, we note that $\mathbf{x}_1$ may therefore be taken as the solution to

$$\sup \sum_i q_i x_i \quad \text{subject to} \quad \begin{array}{c} \sum_i x_i^2 \leq \delta^2 \\ \sum_i r_i\psi(x_i) \leq 1 \\ 0 \leq x_i \end{array}$$

Fourth, we check that with $\psi(x) = x$ the solution has the indicated form. Suppose that $\mathbf{y} = (y_i)$ is an alternative vector satisfying the same feasibility conditions. Then with $h_i = y_i - x_{1,i}$

$$0 \geq \sum x_{1,i} h_i \tag{40}$$
$$0 \geq \sum r_i h_i \tag{41}$$

Let $\mathcal{I} = \{i : x_{1,i} > 0\}$. On $\mathcal{I}^c$, by definition,

$$q_i < br_i, \; i \in \mathcal{I}^c. \tag{42}$$

Hence

$$\begin{aligned}
\sum q_i h_i &< \sum_{i \in \mathcal{I}} q_i h_i + b \sum_{i \notin \mathcal{I}} r_i h_i \quad \text{[by (42)]} \\
&= \sum_{\mathcal{I}} (x_{1,i}/a + br_i) h_i + b \sum_{\mathcal{I}^c} r_i h_i \quad \text{[by (9)]} \\
&= a^{-1} \sum x_{1,i} h_i + b \sum r_i h_i \\
&\leq 0 \quad \text{[by (40),(41)]}
\end{aligned}$$

and the proof is complete.

## 8.5 Proof of Lemma 4

$$\begin{aligned}
\tilde{\theta}_{i,n}(f) &= \frac{2\pi}{n} \sum \phi_i(t_u) f(t_u) \\
&= \frac{2\pi}{n} \sum \phi_i(t_u) \sum_j \theta_j \phi_j(t_u) \\
&= \sum_j \theta_j \Gamma_{ij}.
\end{aligned}$$

We recall the discrete orthogonality relations for sin and cos. Suppose that $i$ and $j$ are not both equal to 0 mod $n$ or to $n/2$ mod $n$.

$$\sum_{u=1}^{n} \cos(i \frac{2\pi u}{n}) \sin(j \frac{2\pi u}{n}) = 0$$

$$\sum_{u=1}^{n} \cos(i\frac{2\pi u}{n})\cos(j\frac{2\pi u}{n}) = \begin{cases} \frac{n}{2} & i = \pm j \bmod n \\ 0 & \text{else} \end{cases}$$
$$\sum_{u=1}^{n} \sin(i\frac{2\pi u}{n})\sin(j\frac{2\pi u}{n}) = \begin{cases} \frac{n}{2} & i = \pm j \bmod n \\ 0 & \text{else} \end{cases}$$

It follows that $\Gamma_{2i,2j} = 1$ if $i = j \bmod n$, $\Gamma_{2i-1,2j-1} = 1$ if $i = j \bmod n$, that $\Gamma_{2i,2n-2i+2kn} = 1$, that $\Gamma_{2i-1,2n-1-2i+2kn} = 1$ and that $\Gamma_{ij} = 0$ otherwise. On the other hand, if $i$ and $j$ are both equal to $n/2 \bmod n$, then $\Gamma_{2i,2j} = 2$; $\Gamma_{2i-1,2j-1} = 0$ if $n$ is even, but $\Gamma_{2i-1,2j-1} = 1$ if $n$ is odd. The first conclusion of the lemma follows.

For the final conclusion, note that if $\theta \in \Theta_m$, the formulas above give (with $1 \le i < n$ and $i = 2j - 1$; the case $i = 2j$ is analogous)

$$\begin{aligned} |\tilde{\theta}_{i,n} - \theta_i| &= |\sum_{k\ge 1} \theta_{i+2kn} + \sum_{k\ge 0} \theta_{2n-i-2+2kn}| \\ &\le (\sum_{k\ge 1}(r_{i+2kn})^{-1} + \sum_{k\ge 0}(r_{2n-i-2+2kn})^{-1})^{1/2} \\ &= (\sum_{k\ge 1}(j+kn)^{-2m} + \sum_{k\ge 0}((k+1)n-j)^{-2m})^{1/2} \\ &\le n^{-m}\gamma_m^{1/2} \end{aligned}$$

with $\gamma_m = 2^{2m+2}\sum_{k>0} k^{-2m}$.

## 8.6 Proof of Lemma 5

We define $Y_n$ and $Y$ as processes on a common probability space, as follows. Let $W$ be a Wiener process, started at $-\pi$. Let $\phi_0(t) = 1/\sqrt{2\pi}$ and define

$$w_i = \int \phi_i(t)W(dt) \quad i = 0, \ldots .$$

The $w_i$ are i.i.d. $N(0, \epsilon^2)$. Define

$$z_u = \sum_{i=0}^{n-1} \phi_i(t_u)w_i \quad u = 1, 2, \ldots, n.$$

Using a transposed form of the orthogonality relations of the previous lemma, the $z_u$ are i.i.d $N(0, \sigma^2)$. Moreover, by those same relations

$$w_i = \frac{2\pi}{n} \sum z_u \phi_i(t_u) \quad i = 1, \ldots, n-1.$$

It follows that with these $z_i$,

$$\begin{aligned} \tilde{y}_i &= \tilde{\theta}_{i,n}(f) + w_i, \quad i = 1, \ldots, n-1 \\ y_i &= \theta_i(f) + w_i. \end{aligned}$$

Hence

$$\tilde{y}_i - y_i = \tilde{\theta}_{i,n}(f) - \theta_i(f) \quad i = 1, \ldots, n-1.$$

Now consider

$$\hat{Q}_0(\tilde{\mathbf{y}}) - \hat{Q}_0(\mathbf{y}) = 2 \sum \hat{q}_i y_i \delta_i + \sum \hat{q}_i \delta_i^2.$$

where $\delta_i = \tilde{y}_i - y_i$. Note that $\hat{q}_i > 0$ only for $i < n_0$, where $n_0 = o(n)$, (see Lemma 6 below). Therefore, for all $i$ appearing in these sums, we have

$$\delta_i^2 \leq \gamma_m n^{-2m}.$$

Put $\Delta = \hat{Q}_0(\tilde{\mathbf{y}}) - \hat{Q}_0(\mathbf{y})$, and use $E\Delta^2 = (E\Delta)^2 + Var(\Delta)$.

$$E\Delta = 2 \sum \hat{q}_i \theta_i \delta_i + \sum \hat{q}_i \delta_i^2 = 2 \cdot I + II,$$

say. But

$$\begin{aligned} |I| &\leq \gamma_m^{1/2} n^{-m} \sum \hat{q}_i |\theta_i| \\ &\leq \gamma_m^{1/2} n^{-m} \sup\{\sum \hat{q}_i |\theta_i| : \sum r_i \theta_i^2 \leq 1\} \\ &= \gamma_m^{1/2} n^{-m} (\sum \frac{\hat{q}_i^2}{r_i})^{1/2} \\ &\asymp n^{-m} n_0^{2k-m+1/2} \asymp n^{-m+\frac{4k-2m+1}{4m+1}} = o(n^{-r}) \end{aligned}$$

as $4m^2 - m - 1 > 0$, while

$$II \leq \gamma_m n^{-2m} \sum_{i=1}^{n_0} \hat{q}_i$$

$$\leq \gamma_m n^{-2m} \sum_{i=1}^{n_0} q_i$$
$$\asymp n^{-2m} n_0^{2k+1}$$
$$\asymp n^{-2m+\frac{8k+2}{4m+1}} = o(n^{-r}).$$

Hence $(E\Delta)^2 = o(n^{-2r})$ uniformly in $\Theta$.

Now

$$Var(\Delta) = 4\epsilon^2 \sum_i \hat{q}_i^2 \delta_i^2 \leq 4\gamma_m n^{-2m} \epsilon^2 \sum \hat{q}_i^2$$

Now the weights $\hat{q}_i$ are minimax; so $R_Q(\epsilon) \geq 2\epsilon^4 \sum \hat{q}_i^2$. Hence

$$2\epsilon^2 \sum \hat{q}_i^2 \leq \epsilon^{-2} R_Q(\epsilon)$$

and so

$$Var(\Delta) \leq 2\gamma_m n^{-2m} \epsilon^{-2} R_Q(\epsilon).$$

As $m > 1/2$ and $\epsilon^{-2} = O(n)$, $Var(\Delta) = o(R_Q(\epsilon)) = o(n^{-2r})$.

## 8.7 Proof of Lemma 6

Let $b(\delta)$ be the solution to eqs. (10)-(11) of lemma 3. Then

$$\frac{g_2(b)}{g_3(b)^2} = \delta^2.$$

Using the asymptotics (13)-(14) for $g_2$ and $g_3$ in section 4, we have

$$b(\delta) \sim B(k,m)\delta^r, \quad \delta \to 0$$

with $B(k,m) = (1-r)^{r/2}[2k+2m+1]^{-r/2}$. As $x_{1,2j} = x_{1,2j-1} = 0$ if $bj^{2m-2k} \geq 1$, we have

$$n_0 \sim b^{\frac{-1}{2m-2k}}$$

substituting in functions of $\delta$ for functions of $b$ we get

$$n_0(\delta) \asymp \delta^{\frac{-2}{4m+1}}$$

and so, if $m > 1/4$, then $n_0(\delta) = o(\delta^{-1})$.

## 8.8 Proof of Lemma 7

Recall the risk isometry of section 2.

$$\inf_{Q_{DS}} \sup_{\Theta_m} R(\hat{Q}, \theta) = \inf_{\mathcal{A}} \sup_{\mathbf{X}} E(\hat{L}(\mathbf{u}) - L(\mathbf{x}))^2 \tag{43}$$

Consider applying Theorem 3 to the functional $Q$ over the set $\Theta_m$. Let $\epsilon_0$ be as in that Theorem, and let $\theta_1, \theta_{-1}$ be the pair mentioned there. Let $(x_{1,i})$ and $(x_{-1,i})$ be the corresponding sequences defined by our isomorphism. By Lemma 3, in solving the constrained optimization problem with $\delta_0 = \sqrt{2}\epsilon_0^2$, we may take $x_{-1,i} = 0$ identically and $x_{1,i} = 0$ for $i > n_0$, for a certain $n_0 = n_0(\delta_0)$. By Theorem 1 of Donoho (1989),

$$\inf_{\mathcal{A}} \sup_{\mathbf{X}} E(\hat{L}(\mathbf{u}) - L(\mathbf{x}))^2 = \inf_{\mathcal{A}} \sup_{[\mathbf{x}_{-1}, \mathbf{x}_1]} E(\hat{L}(\mathbf{u}) - L(\mathbf{x}))^2 \tag{44}$$

where $[\mathbf{x}_{-1}, \mathbf{x}_1]$ denotes the line segment joining $\mathbf{x}_{-1}$ to $\mathbf{x}_1$. Let $\Theta((x_{1,i})) = \{\theta : \theta_i^2 \leq x_{1,i}\}$. Then, one checks,

$$\inf_{\mathcal{A}} \sup_{[\mathbf{x}_{-1}, \mathbf{x}_1]} E(\hat{L}(\mathbf{u}) - L(\mathbf{x}))^2 = \inf_{Q_{DS}} \sup_{\Theta((x_{1,i}))} R(\hat{Q}, \theta). \tag{45}$$

Moreover $\Theta((x_{1,i})) \subset \Theta_{m,n_0}$. Hence

$$\inf_{Q_{DS}} \sup_{\Theta((x_{1,i}))} R(\hat{Q}, \theta) \leq \inf_{Q_{DS}} \sup_{\Theta_{m,n_0}} R(\hat{Q}, \theta) \tag{46}$$

Combining (43)-(46), we get

$$R_Q(\epsilon) = \inf_{Q_{DS}} \sup_{\Theta_{m,n_0}} R(\hat{Q}, \theta)$$

Now in the corresponding linear problem, $\delta_0 \asymp \eta$, see Donoho (1989). It follows that here $\epsilon_0 \asymp \epsilon$. The proof is complete.

## 8.9 Proof of Lemma 8

We use, without comment, terminology and notation associated with the Hellinger distance between probability measures; compare, for example Donoho and Liu (1988), Le Cam (1986). Let

$\mu = P_{1,\epsilon} + P_{2,\epsilon}$ and define $g_1 = \frac{dP_{1,\epsilon}}{d\mu}$, $g_0 = \frac{dP_{0,\epsilon}}{d\mu}$. Let $\rho$ denote the Hellinger affinity

$$\rho = \int \sqrt{g_1}\sqrt{g_0}d\mu.$$

Then the quantity of interest is bounded by

$$\pi(P_{1,\epsilon}, P_{0,\epsilon}) \geq \frac{1}{2}\rho^2.$$

Let $\phi_\epsilon(y)$ denote the probability density of $N(0, \epsilon^2)$ Then, putting $g_{1,i}(y) = (\phi_\epsilon(y - \theta_{1,i}) + \phi_\epsilon(y + \theta_{1,i}))/2$ and $g_{0,i}(y) = (\phi_\epsilon(y - \theta_{0,i}) + \phi_\epsilon(y + \theta_{0,i}))/2$, we have that

$$\rho = \Pi_{i=1}^{\infty}\rho_i$$

where

$$\rho_i = \int \sqrt{g_{1,i}}\sqrt{g_{0,i}}dy.$$

Now in terms of (squared) Hellinger distance, $h_i^2 = \int(\sqrt{g_{1,i}} - \sqrt{g_{0,i}})^2 dy$,

$$\rho^2 = \exp\{2\sum \log(1 - h_i^2/2)\}.$$

Define $\xi(x) = |\log(1 - x/2)|$; then $\psi(x) = \xi(x)/x$ is increasing and

$$|\log \rho| = \sum \xi(h_i^2) \leq \psi(\sup_i h_i^2)\sum h_i^2.$$

We now use two facts. First, by the Lemma 10 below

$$\sup_i h_i^2 \leq 2(1 - \exp(-c/8))$$

Second, as $\sum_i \theta_{\cdot,i}^2 \leq M$, Markov's inequality gives, for each $a \in (0, 1]$,

$$\#\{i : \theta_{\cdot,i}^2 > a\} \leq \frac{M}{a}$$

Hence putting $\mathcal{I}_a = \{i : \theta_{0,i}^2 \leq a, \theta_{1,i}^2 \leq a\}$,

$$\begin{aligned} \sum h_i^2 &= \sum_{i \in \mathcal{I}_a} h_i^2 + \sum_{i \notin \mathcal{I}_a} h_i^2 \\ &\leq \sum_{i \in \mathcal{I}_a} h_i^2 + \frac{4M}{a} \end{aligned}$$

where we used $h_i^2 \le 2$. Invoking Lemma 11 below,

$$\sum_{i \in I_a} h_i^2 \le C(a) \cdot \sum_i (\theta_{0,i}^2 - \theta_{1,i}^2)^2.$$

Suppose that $\epsilon = 1$; then for each $a \in (0,1]$,

$$|\log \rho| \le \psi(2(1-\exp(-c/8)))\,(C^2(a)c^2 + \frac{4M}{a}) \equiv r(c,a,M),$$

say, and so the lemma holds in this special case, with

$$\alpha(c,M) = (16)^{-1}\exp(-2\,r(c,a,M)).$$

The general case then follows from this special case, with the same $\alpha$, by a certain scale invariance.

**Lemma 10** *Let* $\sum_i (\theta_{0,i}^2 - \theta_{1,i}^2)^2 \le c^2$ *and* $\epsilon = 1$. *Then*

$$\sup_i h_i^2 \le 2(1-\exp(-c/8))$$

**Proof.** By convexity of squared Hellinger distance, the definition of $g_{1,i}$ etc.,

$$\begin{aligned} h_i^2 &\le \frac{1}{2}H^2(\phi(\cdot - \theta_{0,i}),\phi(\cdot - \theta_{1,i})) + \frac{1}{2}H^2(\phi(\cdot + \theta_{0,i}),\phi(\cdot + \theta_{1,i})) \\ &= H^2(\phi,\phi(\cdot - (\theta_{0,i} - \theta_{1,i}))) \end{aligned}$$

But

$$\begin{aligned} c^2 &\ge \sup_i (\theta_{0,i}^2 - \theta_{1,i}^2)^2 \\ &\ge \sup_i (\theta_{0,i} - \theta_{1,i})^4. \end{aligned}$$

The lemma now follows from the formula $H^2(\phi,\phi(\cdot - \eta)) = 2(1 - \exp(-\eta^2/8))$.

**Lemma 11** *Suppose that* $1 > a \ge \theta_{1,i}, \theta_{0,i} \ge 0$. *Then*

$$h_i \le C(a)\,|\theta_{0,i}^2 - \theta_{1,i}^2|$$

*with*

$$C(a) = 1 + \int |x|^4 cosh^2(ax)\phi(x)dx$$

**Proof.**

$$h_i^2 = \int (\sqrt{e_1 cosh(\theta_1 x)} - \sqrt{e_0 cosh(\theta_0 x)})^2 \phi(x) dx,$$

where $e_1 = \exp(-\theta_1^2/2)$, $e_0 = \exp(-\theta_0^2/2)$. Hence

$$\begin{aligned} h_i &\leq (\int (\sqrt{e_1 cosh(\theta_1 x)} - \sqrt{e_0 cosh(\theta_1 x)})^2 \phi(x) dx)^{1/2} \\ &+ (\int (\sqrt{e_0 cosh(\theta_1 x)} - \sqrt{e_0 cosh(\theta_0 x)})^2 \phi(x) dx)^{1/2} \\ &= I + II, \quad \text{say.} \end{aligned}$$

Now, by a calculation,

$$I^2 = (\exp(-\theta_1^2/4) - \exp(-\theta_0^2/4))^2 \exp(\theta_1^2) \leq (\theta_1^2 - \theta_0^2)^2,$$

and

$$II^2 = e_0^2 \int (\sqrt{cosh(\theta_0 x)} - \sqrt{cosh(\theta_1 x)})^2 \phi(x) dx.$$

Put $\psi(t) = \sqrt{cosh(t)}$. Then, by a calculation,

$$\psi(\theta_0 x) - \psi(\theta_1 x) \leq (\theta_1^2 - \theta_0^2)/2 \; |x|^2 cosh(\theta_1 x)$$

so

$$II^2 \leq (\theta_1^2 - \theta_0^2)^2/4 \int |x|^4 cosh^2(\theta_1 x) \phi(x) dx$$

and the lemma follows.

## 8.10 Proof of Lemma 9

The argument is similar to that for Lemma 7. Put $\delta = c\epsilon^2$, and use Lemma 3 with that $\delta$. This gives sequences $\mathbf{x}_{-1} = \mathbf{0}$ and $\mathbf{x}_1$. Defining $\theta_{1,i} = \sqrt{x_{1,i}}$ etc., we get a pair $(\theta_1, \theta_{-1})$ to which Lemma 8 applies. Indeed, the random variable $(s, \theta_{1,i})$ is $\Theta((x_{1,i}))$-valued. We conclude, as in the argument for Theorem 6, that

$$\inf_{\hat{Q}} \sup_{\Theta((x_{1,i}))} E(\hat{Q}(\mathbf{y}) - Q(\theta))^2 \geq \omega(c\epsilon^2)\alpha(c)$$

Now, by construction $x_{1,i} = 0$ for $i > n_0$. This completes the proof.

## References


[1] Donoho, D.L. (1989) Statistical Estimation and Optimal recovery. Technical Report 214, Department of Statistics, U. C. Berkeley.

[2] Donoho, D. L., and Low, M. (1990) White Noise Approximation and Minimax Risk. Technical Report. Department of Statistics, U. C. Berkeley.

[3] Donoho, D. L. and Liu, R. C. (1988) Geometrizing rates of convergence II. Technical Report No. 120, Department of Statistics, U. C. Berkeley.

[4] Fan, J. (1988) Nonparametric estimation of quadratic functionals in Gaussian white noise. Technical Report 166, Deprtament of Statistics, U.C. Berkeley. (To appear, *Annals of Statistics*).

[5] Hall, P. and Marron, J. S.(1987) "Estimation of integrated squared density derivatives" *Statistics and Probability Letters* **6**, 109-115.

[6] Ibragimov, I.A., Nemirovskii, A.S., and Has'minskii, R.Z. (1987) Some problems on nonparametric estimation in Gaussian white noise (in Russian). *Teoria Verojatnostei i Primenenia*, **32**.

[7] Le Cam, L. (1986) *Asymptotic Methods in Statistical Decision Theory*. Springer-Verlag.

[8] Levit, B.Y. (1978) Asymptotically Efficient Estimation of Nonlinear Functionals. *Problemy Peredachi Informatsii* **14**, 65-72

[9] Low, M.G. (1988) Towards a unified theory of asymptotic minimax estimation. Ph. D. Thesis, Cornell Univ.

[10] Micchelli, C. A. (1975) Optimal estimation of linear functionals. *IBM Research Report* 5729.

[11] Micchelli, C. A. and Rivlin, T. J. (1977) A survey of optimal recovery. *Optimal Estimation in Approximation Theory*. Micchelli and Rivlin, eds. Plenum Press, New York. pp 1-54.

[12] Nussbaum, M. (1985) Spline Smoothing and Asymptotic Efficiency in $L_2$. *Ann. Statist.* **13** 984-987.

[13] Pinsker, M. S. (1980) Optimal filtering of square integrable signals in Gaussian white noise. *Problems of Info. Trans.* **16**, 2, 52-68.

[14] Ritov, Y. and Bickel, P.J. (1988) Estimating Integrated Squared Density Derivatives: sharp best order of convergence. *Sankhya* **50** 381-393.

[15] Traub, J. F., Wasilkowski G.W., and Woźniakowski, H. (1983) *Information, Uncertainty, Complexity* Addison-Wesley Pub Co.

[16] Traub, J. F., Wasilkowski G.W., and Woźniakowski, H. (1988) *Information-Based Complexity*. Addison-Wesley Pub Co.