Bernoulli **13**(1), 2007, 1–19
DOI: 10.3150/07-BEJ5001

Trade-offs between global and local risks in nonparametric function estimation

T. TONY CAI*, MARK G. LOW** and LINDA H. ZHAO†

*Department of Statistics, The Wharton School, University of Pennsylvania, 400 Jon M. Huntsman Hall, 3730 Walnut Street, Philadelphia PA 19104-6340, USA. E-mail: *tcai@wharton.upenn.edu; **lowm@wharton.upenn.edu; †lzhao@wharton.upenn.edu*

The problem of loss adaptation is investigated: given a fixed parameter, the goal is to construct an estimator that adapts to the loss function in the sense that the estimator is optimal both globally and locally at every point. Given the class of estimator sequences that achieve the minimax rate, over a fixed Besov space, for estimating the entire function a lower bound is given on the performance for estimating the function at each point. This bound is larger by a logarithmic factor than the usual minimax rate for estimation at a point when the global and local minimax rates of convergence differ. A lower bound for the maximum global risk is given for estimators that achieve optimal minimax rates of convergence at every point. An inequality concerning estimation in a two-parameter statistical problem plays a key role in the proof. It can be considered as a generalization of an inequality due to Brown and Low. This may be of independent interest. A particular wavelet estimator is constructed which is globally optimal and which attains the lower bound for the local risk provided by our inequality.

Keywords: Besov class; constrained risk inequality; loss adaptation; normal location–scale model; nonparametric function estimation; nonparametric regression; superefficiency; wavelets; white noise model

1. Introduction

There are two major formulations for the problem of estimating a function based on noisy data. One common approach is to fix attention on each point and use squared error loss to measure the accuracy of an estimate. Another approach based on integrated squared error loss gives a more global measure of accuracy.

In both the local and global approaches theory has been developed for the construction of minimax estimators over a given function space; see, for example, Pinsker (1980), Ibragimov and Hasminski (1984), Donoho and Liu (1991) and the references in Efromovich (1999).

Local and global theory have also focused on the construction of estimators which are simultaneously asymptotically (near) minimax over a collection of parameter spaces. Such estimators are called adaptive and, for a given loss function, are 'optimal' over a range of parameter spaces.

When attention is focused on adaptive estimation there are some striking differences between local and global theories. Under integrated squared error loss there are many situations where rate adaptive estimators can be constructed. In particular, Efromovich and Pinsker (1984) constructed fully adaptive estimators over a range of Sobolev spaces. Recent results focus on more general Besov spaces; see, for example, Donoho and Johnstone (1995), Cai (1999) and Härdle *et al.* (1998).

When attention is focused on estimating a function at a given point, rate-optimal adaptive procedures typically do not exist. A penalty, usually a logarithmic factor, must be paid for not knowing the smoothness. Important work in this area began with Lepski (1990) where attention focused on a collection of Lipschitz classes; see also Brown and Low (1996b) and Efromovich and Low (1994). Connections between local and global parameter space adaptation can be found in Cai (1999) and Efromovich (2002).

As just mentioned, for estimating a function at a point the payment of a logarithmic penalty can be viewed as a necessary price for adapting over different parameter spaces. Here we consider a complementary question: whether, given a fixed parameter space, we can construct an estimator that adapts to the loss function in the sense that the estimator is optimal both locally and globally. Specifically, for a given Besov space, we investigate whether an estimator can be constructed which is both minimax rate-optimal at each point under rth power loss ($1 \leqslant r < \infty$) and also minimax rate-optimal under integrated rth power loss.

To date, adaptation has mainly focused on different smoothness classes, although some recent nonparametric function estimation literature can also be viewed as falling within the area of loss adaptation. For example, Bickel and Ritov (2003) have studied nonparametric density estimators whose integrals provide efficient estimators of the corresponding functionals; Efromovich (2004) describes an estimator for censored data that achieves a similar multi-purpose goal. Cai (2002) is another example where a minimax rate estimator of the unknown nonparametric regression function is given such that the derivatives of this estimator attain the minimax rate for estimation of the derivatives of the function. These problems can all be viewed as one of finding a single estimator of the unknown function which are optimal for a variety of loss functions.

In the present setting we consider the class of estimator sequences that achieve the minimax rate, over a fixed Besov space, for estimating the entire function and establish a lower bound on the ratewise performance of any such sequence for estimation of the function at each point. This bound is larger by a logarithmic factor than the usual minimax rate for estimation at a point when the global minimax rate of convergence is faster than the local minimax rate of convergence. Hence, no estimator sequence can simultaneously have minimax rate both globally and at each point when the global and local minimax rates of convergence differ. We should emphasize that the loss of the logarithmic factor is not caused by trying to be adaptive over several function classes. We have a fixed function class but more than one loss function. We also consider estimators that achieve optimal minimax rates of convergence at every point and give a lower bound for the maximum global risk.

We then study a modification of the wavelet estimator of Delyon and Juditsky (1996). We show that this estimator attains the optimal global rate and also attains the rate given by that our lower bound for estimation locally at each point. This shows that our lower bound is sharp. These main results are stated in Sections 2–4; their proofs are completed in Section 8. An inequality concerning estimation in a two-parameter statistical problem plays a key role in our proof. An alternative proof can also be based on the two-point testing techniques used in Lepski (1990) and Tsybakov (1998). The inequality used in the present paper and presented in Section 6 can be considered as an ℓ^r-norm generalization of an inequality in Brown and Low (1996b). A generalization of the Hammersley–Chapman–Robbins inequality can also be deduced from our basic result. This generalization may be of independent interest and is included as an additional proposition in this section.

Our basic inequality also has some consequences relating to the possibility of superefficient parametric estimation under ℓ^r loss. These are described in Section 7.

2. Trade-offs between global and pointwise risks

Consider the white noise model in which we observe Gaussian processes $Y_n(t)$ governed by

$$\mathrm{d}Y_n(t) = f(t)\mathrm{d}t + n^{-1/2}\,\mathrm{d}W(t), \qquad 0 \leqslant t \leqslant 1, \tag{1}$$

where $W(t)$ is a standard Brownian motion and f is an unknown function of interest. This canonical white noise model is asymptotically equivalent to the conventional formulation of nonparametric regression; see Brown and Low (1996a) and Brown *et al.* (2002). There is also a slightly less direct equivalence to nonparametric density estimation; see Nussbaum (1996), Klemelä and Nussbaum (1999), and Brown *et al.* (2004).

The accuracy of an estimate $\hat{f}_n$ of f is measured both under the global L^r risk

$$R(\hat{f}, f) \equiv \mathrm{E}_f\|\hat{f}_n - f\|_r^r = \mathrm{E}_f\int_0^1 |\hat{f}_n(t) - f(t)|^r\,\mathrm{d}t, \tag{2}$$

as well as under the pointwise ℓ^r risk,

$$R(\hat{f}, f; x) \equiv \mathrm{E}_f|\hat{f}_n(x) - f(x)|^r, \tag{3}$$

for all $x \in (0, 1)$.

The benchmarks for the performance of an estimator $\hat{f}_n$ over a function class $\mathcal{F}$ are the minimax risks under the respective error measures. The global minimax L^r risk is

$$R_g(\mathcal{F}, n) = \inf_{\hat{f}_n}\sup_{\mathcal{F}}\mathrm{E}\|\hat{f}_n - f\|_r^r,$$

and the local minimax ℓ^r risk at a point x is

$$R_l(\mathcal{F}, x, n) = \inf_{\hat{f}_n}\sup_{\mathcal{F}}\mathrm{E}|\hat{f}_n(x) - f(x)|^r.$$

We investigate the possibility of having a loss adaptive estimator. We want an estimator which at every point attains the minimax pointwise rate while also attaining the minimax global rate.

We use Besov balls as examples of the function class $\mathcal{F}$. Besov spaces occur naturally in many areas of analysis. They include many traditional smoothness spaces such as Hölder and Sobolev spaces, as well as function spaces of significant spatial inhomogeneity such as the bump algebra and the bounded variation. Full details of Besov spaces are given, for example, in DeVore and Popov (1988).

A Besov space $B^{\alpha}_{p,q}$ has three parameters: α measures the degree of smoothness, and p and q specify the type of norm used to measure the smoothness. For $f \in L^p[0, 1]$ and $h > 0$, denote the Kth difference by $\Delta_h^{(K)} f(t) = \sum_{k=0}^{K} (-1)^k f(t + kh)$. The modulus of smoothness of order K of f is $\omega_{K,p}(f, h) = \|\Delta_h^{(K)} f\|_{L^p[0,1-Kh]}$. The Besov norm of index (α, p, q) is defined for $K > \alpha$ by

$$\|f\|_{B^{\alpha}_{p,q}} = \begin{cases} \|f\|_p + \left(\displaystyle\int_0^1 [h^{-\alpha} \omega_{K,p}(f, h)]^q \frac{\mathrm{d}h}{h} \right)^{1/q}, & \text{for } q < \infty, \\ \|f\|_p + \|h^{-\alpha} \omega_{K,p}(f, h)\|_{\infty}, & \text{for } q = \infty. \end{cases} \tag{4}$$

The Besov class $B^{\alpha}_{p,q}(M)$ is a ball of radius M under the Besov norm $\|\cdot\|_{B^{\alpha}_{p,q}}$:

$$B^{\alpha}_{p,q}(M) = \{f : \|f\|_{B^{\alpha}_{p,q}} \leqslant M\}.$$

The estimation problems over the Besov class $B^{\alpha}_{p,q}(M)$ under the local risk (3) and the global risk (2) have important distinctions. Write $\nu = \alpha - 1/p$. We will assume $\alpha > 1/p$ so $\nu > 0$. It is shown in Cai (2003) that the minimax convergence rate over $B^{\alpha}_{p,q}(M)$ under the local risk (3) is $n^{l(\alpha,p,r)}$, where $l(\alpha, p, r) = r\nu/(1 + 2\nu)$. The minimax rate under the global L^r risk for $1 \leqslant r < \infty$ is $A_{\alpha,p,r}(n)$, where

$$A_{\alpha,p,r}(n) = \begin{cases} n^{r\alpha/(1+2\alpha)} & \text{when } r < p(1 + 2\alpha), \\ \left(\dfrac{n}{\log n}\right)^{(1+r\nu)/(1+2\nu)} (\log n)^{-(r/2-p/q)_+}, & \text{when } r = p(1 + 2\alpha), \\ \left(\dfrac{n}{\log n}\right)^{(1+r\nu)/(1+2\nu)} & \text{when } r > p(1 + 2\alpha); \end{cases} \tag{5}$$

see Donoho *et al.* (1995, 1996) and Delyon and Juditsky (1996). It will be convenient to denote the exponent of the algebraic term in the rate of convergence by $g(\alpha, p, r)$. So

$$g(\alpha, p, r) = \begin{cases} \dfrac{r\alpha}{1 + 2\alpha}, & \text{when } r < p(1 + 2\alpha), \\ \dfrac{1 + r\nu}{1 + 2\nu}, & \text{when } r \geqslant p(1 + 2\alpha). \end{cases} \tag{6}$$

Note that $g(\alpha, p, r) = l(\alpha, p, r)$ when $p = \infty$ and $g(\alpha, p, r) > l(\alpha, p, r)$ when $p < \infty$. Therefore the local rate is the same as the global rate for $p = \infty$ and is always slower than the global rate for $p < \infty$.

Theorem 1 below states that, when the global and local minimax rates are different, any estimator attaining the minimax rate at any fixed function f_0 under the global risk (2) must be suboptimal in terms of the maximum pointwise risk at 'most' points in (0, 1); the minimum penalty is a logarithmic factor.

Theorem 1. *Suppose* $\hat{f}_n$ *is an estimator based on* (1) *satisfying*

$$\overline{\lim_{n\to\infty}} A_{\alpha,p,r}(n) \cdot \mathrm{E}\|\hat{f}_n - f_0\|_r^r = K < \infty \tag{7}$$

at some $f_0 \in B^\alpha_{p,q}(M')$ *with* $p < \infty$ *and* $M' < M$. *Then, for any measurable set* $\Omega \subset (0, 1)$ *with the Lebesgue measure* $m(\Omega) > 0$ *and any* $0 < \epsilon < m(\Omega)$, *there exists a subset* $\Omega_0 \subseteq \Omega$ *with* $m(\Omega_0) \geqslant m(\Omega) - \epsilon$ *such that*

$$\overline{\lim_{n\to\infty}} \left(\frac{n}{\log n}\right)^{l(\alpha,p,r)} \inf_{x\in\Omega_0} \sup_{f\in B^\alpha_{p,q}(M)} \mathrm{E}|\hat{f}_n(x) - f(x)|^r > 0. \tag{8}$$

In particular, for any $\epsilon > 0$, *there exists a subset* $\Omega_0 \subset (0, 1)$ *with the Lebesgue measure* $m(\Omega_0) \geqslant 1 - \epsilon$ *such that*

$$\overline{\lim_{n\to\infty}} \left(\frac{n}{\log n}\right)^{l(\alpha,p,r)} \inf_{x\in\Omega_0} \sup_{f\in B^\alpha_{p,q}(M)} \mathrm{E}|\hat{f}_n(x) - f(x)|^r > 0. \tag{9}$$

One of the main tools for the proof of Theorem 1 is a general constrained risk inequality which may be of independent interest. This risk inequality can be used to derive a generalized version of the Hammersley–Chapman–Robbins inequality as well as results relating to the possibility of superefficient parametric estimation under ℓ^r loss. The constrained risk inequality and these applications are given in Sections 6 and 7. The proof of Theorem 1 using the constrained risk inequality will be postponed to Section 8. Alternatively, Theorem 1 can also be proved using the testing arguments given in Lepski (1990) and Tsybakov (1998).

A direct consequence of Theorem 1 is that if an estimator $\hat{f}_n$ attains the global minimax rate of convergence at some f, then the set of points at which the estimator $\hat{f}_n$ attains the local minimax rate of convergence has measure 0.

Corollary 1. *Suppose* $\hat{f}_n$ *is an estimator satisfying* (7). *Let* $p < \infty$ *and let* Λ *be the set of points at which* $\hat{f}_n$ *attains the pointwise minimax rate of convergence, that is,*

$$\Lambda = \left\{ x : \overline{\lim_{n\to\infty}}\, n^{l(\alpha,p,r)} \sup_{f\in B^\alpha_{p,q}(M)} \mathrm{E}|\hat{f}_n(x) - f(x)|^r < \infty \right\}.$$

Then the set Λ *has Lebesgue measure* 0.

Therefore it is impossible for any estimator to simultaneously attain the global minimax rate of convergence and the local minimax rate at every point when the global and local minimax rates are different.

Remark. It can be seen from the proof of Theorem 1 that the conclusions of Theorem 1 and Corollary 1 remain valid if condition (7) is replaced by

$$\overline{\lim_{n\to\infty}}\, n^{\rho}\cdot \mathrm{E}\|\hat{f}_n - f_0\|_r^r < \infty$$

for any $\rho > l(\alpha, p, r)$.

3. A rate-optimal wavelet estimator

Theorem 1 shows that the minimum penalty for a global rate-optimal estimator is a logarithmic factor in terms of the maximum pointwise risk over $B^{\alpha}_{p,q}(M)$. This lower bound can in fact be attained. Delyon and Juditsky (1996) propose a wavelet estimator and show the optimality of the estimator for density estimation and nonparametric regression over $B^{\alpha}_{p,q}(M)$ under the Sobolev norm loss. In the following we use a slightly modified version of the wavelet estimator given in Delyon and Juditsky (1996) for the white noise model (1) and show that the estimator attains the optimal global rate and the local rate of $(n/\log n)^{l(\alpha,p,r)}$ at every point $x \in (0, 1)$. In this sense, the lower bound given in Theorem 1 is sharp.

Let the functions ϕ and ψ be a pair of compactly supported father and mother wavelets with $\int \phi = 1$. We call a wavelet ψ *K-regular* if ψ has K vanishing moments and K continuous derivatives. Let $\phi_{j,k}(t) = 2^{j/2}\phi(2^j t - k)$ and $\psi_{j,k}(t) = 2^{j/2}\psi(2^j t - k)$. These functions, with appropriate treatments at the boundaries, form an orthonormal basis. For simplicity in exposition, we work with periodized wavelet bases on [0, 1], letting

$$\phi^p_{jk}(t) = \sum_{l\in\mathcal{Z}} \phi_{jk}(t-l), \quad \psi^p_{jk}(t) = \sum_{l\in\mathcal{Z}} \psi_{jk}(t-l), \qquad \text{for } t \in [0, 1].$$

Then the collection $\{\phi^p_{j_0k},\ k = 1, \ldots, 2^{j_0};\ \psi^p_{jk},\ j \geqslant j_0 \geqslant 0,\ k = 1, \ldots, 2^j\}$ is an orthonormal basis of $L^2[0, 1]$, provided the primary resolution level j_0 is large enough to ensure that the support of the scaling functions and wavelets at level j_0 is not the whole of [0, 1]. The superscript p will be suppressed from the notation for convenience. See Cohen *et al.* (1993), Daubechies (1994) and Meyer (1991) for other boundary-corrected wavelet bases on the unit interval [0, 1]. For wavelets on the line, see Daubechies (1992) and Meyer (1992).

For any $l \geqslant j_0$, a function f can be expanded into a wavelet series

$$f(t) = \sum_{k=1}^{2^l} \xi_{lk}\phi_{l,k}(t) + \sum_{j=l}^{\infty}\sum_{k=1}^{2^j} \theta_{jk}\psi_{j,k}(t), \tag{10}$$

where $\xi_{l,k} = \int_0^1 f(t)\phi_{l,k}(t)\,\mathrm{d}t$ and $\theta_{j,k} = \int_0^1 f(t)\psi_{j,k}(t)\,\mathrm{d}t$.

For the white noise model (1), let $\tilde{y}_{j,k} = \int \phi_{j,k}(t)\,\mathrm{d}Y_n(t)$, $y_{j,k} = \int \psi_{j,k}(t)\,\mathrm{d}Y_n(t)$, $\tilde{z}_{j,k} = \int \phi_{j,k}(t)\,\mathrm{d}W(t)$ and $z_{j,k} = \int \psi_{j,k}(t)\,\mathrm{d}W(t)$. The white noise model (1) is then equivalent to a sequence model in which one observes an empirical wavelet coefficient sequence:

$$\tilde{y}_{l,k} = \xi_{l,k} + n^{-1/2}\tilde{z}_{l,k}, \qquad k = 1, 2, \ldots, 2^l, \tag{11}$$

$$y_{j,k} = \theta_{j,k} + n^{-1/2}z_{j,k}, \qquad k = 1, 2, \ldots, 2^j,\ j \geqslant l, \tag{12}$$

where $l \geqslant j_0$ and $\tilde{z}_{l,k}$ and $z_{j,k}$ are independently and identically distributed as $N(0, 1)$.

Let J_0 and J be integers satisfying $n^{(1+2\alpha)} \leqslant 2^{J_0} < 2n^{(1+2\alpha)}$ and $n \leqslant 2^J < 2n$, respectively. For $j \geqslant J_0 + 1$, let

$$\lambda_j = \sqrt{rn^{-1}\log(2^{j-J_0})} \tag{13}$$

and let $\eta_\lambda(y) = \operatorname{sgn}(y)(|y| - \lambda)_+$ be the soft threshold function. We define the following wavelet estimator:

$$\hat{\xi}_{J_0,k} = \tilde{y}_{J_0,k} \quad \text{and} \quad \hat{\theta}_{j,k} = \begin{cases} \eta_{\lambda_j}(y_{j,k}), & \text{if } J_0 \leqslant j < J, \\ 0, & \text{if } j \geqslant J. \end{cases} \tag{14}$$

The estimator of f is given by

$$\hat{f}_n(t) = \sum_{k=1}^{2^{J_0}} \hat{\xi}_{J_0,k}\phi_{J_0,k}(t) + \sum_{j=J_0}^{J-1}\sum_{k=1}^{2^j} \hat{\theta}_{j,k}\psi_{j,k}(t). \tag{15}$$

The estimator given in (15) differs from the wavelet estimator in Delyon and Juditsky (1996) in the choice of upper and lower resolution levels J_0 and J as well as in the choice of the threshold λ_j.

Over the Besov class $B^\alpha_{p,q}(M)$, the wavelet estimator $\hat{f}_n$ is rate-optimal under the global L^r risk and at the same time is within a logarithmic factor of the minimax risk under the pointwise ℓ^r risk at every point in (0, 1).

Theorem 2. *Let $\hat{f}_n$ be the wavelet estimator given in* (15). *Let $B^\alpha_{p,q}(M)$ be a Besov class with $\alpha - 1/p > 0$, $0 < q \leqslant \infty$, and $M > 0$. Suppose the wavelet ψ is K-regular with $K > \alpha$. Then $\hat{f}$ is rate-optimal over the Besov class $B^\alpha_{p,q}(M)$ under the global L^r risk* (2) *and is within a logarithmic factor of the minimax risk under the local risk* (3) *for any $1 \leqslant r < \infty$. That is,*

$$\overline{\lim_{n\to\infty}} A_{\alpha,p,r}(n) \sup_{f\in B^\alpha_{p,q}(M)} \mathrm{E}\|\hat{f}_n - f\|_r^r < \infty \tag{16}$$

and

$$\overline{\lim_{n\to\infty}} \left(\frac{n}{\log n}\right)^{l(\alpha,p,r)} \sup_{f\in B^\alpha_{p,q}(M)} \mathrm{E}_f|\hat{f}_n(x) - f(x)|^r < \infty, \tag{17}$$

for every $x \in (0, 1)$.

The proof of Theorem 2 is given in Section 8.

Remark. The estimator $\hat{f}_n$ given in (15) depends on the loss function. It can be modified slightly so that a single estimator $\hat{f}^*_n$ satisfies (16) and (17) simultaneously for a range of loss functions. Let $r^* > 1$ be fixed and, for $j \geqslant J_0 + 1$, set $\lambda^*_j = \sqrt{r^* n^{-1}\log(2^{j-J_0})}$. Let the estimator $\hat{f}^*_n$ be defined as in (15) with λ_j replaced by λ^*_j. Then the estimator $\hat{f}^*_n$ satisfies (16) and (17) for all $1 \leqslant r \leqslant r^*$.

4. Local rate-optimal estimators

In the previous section, we considered the local performance of a global rate-optimal estimator. We now consider the global performance of an estimator which is rate-optimal at every point under the local risk (3).

Theorem 3. *Suppose* $\hat{f}_n$ *is an estimator based on* (1). *Let*

$$\Omega = \left\{ x : \overline{\lim_{n\to\infty}}\, n^{l(\alpha,p,r)} \sup_{f\in B^{\alpha}_{p,q}(M)} \mathrm{E}|\hat{f}_n(x) - f(x)|^r < \infty \right\}. \tag{18}$$

If the Lebesgue measure $m(\Omega) > 0$, *then for any f with* $\|f\|_{B^{\alpha}_{p,q}} < M$ *and any* $\rho > l(\alpha, p, r)$,

$$\underline{\lim}_{n\to\infty}\, n^{\rho} \cdot \mathrm{E}\|\hat{f}_n - f\|_r^r = \infty. \tag{19}$$

It follows from Theorem 3 that, when the global and local minimax rates are different (i.e. $p < \infty$), if an estimator is minimax rate-optimal at every point under the local risk (3), then it cannot attain the global minimax rate at *any* f in the interior of the parameter space $B^{\alpha}_{p,q}(M)$, and consequently the maximum global risk of the estimator over $B^{\alpha}_{p,q}(M)$ must be suboptimal. Moreover, it follows from Theorem 3 the penalty on the maximum global risk for being local rate-optimal at every point is a power of n.

5. The case of $p = \infty$

For a Besov class $B^{\alpha}_{p,q}(M)$, when $p = \infty$, $g(\alpha, p, r) = l(\alpha, p, r)$ and the global and local minimax rates coincide. In this case the optimal global rate and the optimal local rate at every point in (0, 1) can be attained simultaneously. For the class of analytic functions, Efromovich (1999) showed that the global minimax rate under the mean integrated squared error, is the same as the local minimax rate under the mean squared error, and in this case a single estimator can attain simultaneously the global minimax rate and the local minimax rate at every point in (0, 1).

Over a Besov class $B^{\alpha}_{\infty,q}(M)$, it is not difficult to construct a wavelet estimator which is both global rate-optimal and local rate-optimal at every point in (0, 1). Using the same notation as in the previous section, let J_0 be an integer satisfying $n^{(1+2\alpha)} \leqslant 2^{J_0} < 2n^{(1+2\alpha)}$ and let the estimator of f be a projection (truncation) estimator,

$$\hat{f}_n(t) = \sum_{k=1}^{2^{J_0}} \tilde{y}_{J_0,k}\phi_{J_0 k}(t). \tag{20}$$

Then it is straightforward to verify that this wavelet estimator $\hat{f}_n$ is simultaneously rate-optimal under the global risk and the local risk at every point.

Furthermore, for any estimator $\hat{f}_n$ that is both global rate-optimal and local rate-optimal at every point, the estimator must attain the same global rate at every function f in the interior of $B^{\alpha}_{\infty,q}(M)$, that is,

$$0 < \varliminf_{n\to\infty} n^{r\alpha/(1+2\alpha)}\mathrm{E}\|\hat{f}_n - f\|_r^r \leqslant \varlimsup_{n\to\infty} n^{r\alpha/(1+2\alpha)}\mathrm{E}\|\hat{f}_n - f\|_r^r < \infty,$$

for all f with $\|f\|_{B^{\alpha}_{\infty,q}} < M$.

6. A general constrained risk inequality

Throughout this section, we let $1 \leqslant r < \infty$ and let w satisfy

$$\frac{1}{r} + \frac{1}{w} = 1.$$

Let X be a (vector-valued) random variable having distribution P_θ with density f_θ with respect to a measure λ. The parameter θ takes two possible values, θ_1 or θ_2. We wish to estimate θ based on X. Suppose the parameters $\theta_i = (\theta_{i,1}, \ldots, \theta_{i,K}) \in \mathbb{R}^K$ $(i = 1, 2)$. For any estimator δ based on X, its ℓ^r risk is defined by

$$R_r(\delta, \gamma) = \mathrm{E}\|\delta(X) - \gamma\|_{\ell^r}^r = \int \sum_{k=1}^{K} |\delta_k(x) - \gamma_k|^r f_\gamma(x)\lambda(\mathrm{d}x).$$

Denote by $s(x) = f_{\theta_2}(x)/f_{\theta_1}(x)$ the ratio of the two density functions. ($s(x) = \infty$ for some x is possible, with the obvious interpretation $s(x)f_{\theta_1}(x) = f_{\theta_2}(x)$.) Write

$$\Delta_r = \|\theta_2 - \theta_1\|_{\ell^r} = \left(\sum_{k=1}^{K} |\theta_{2,k} - \theta_{1,k}|^r\right)^{1/r}. \tag{21}$$

For $1 \leqslant w < \infty$, let

$$I_w = I_w(\theta_1, \theta_2) = (\mathrm{E}_{\theta_1}(s^w(X)))^{1/w}, \tag{22}$$

and let $I_\infty = I_\infty(\theta_1, \theta_2) = \|s\|_\infty$ where the supremum norm is taken with respect to P_{θ_1}.

When $I_w < \infty$ the following result gives a lower bound for the bias at θ_2 and, in particular, the risk $R(\delta, \theta_2)$ given an upper bound on the risk at θ_1.

Theorem 4. *Suppose $R_r(\delta, \theta_1) \leqslant \epsilon_r^r$ and $\Delta_r > \epsilon_r I_w$. Then*

$$\|\mathrm{E}_{\theta_2}\delta(X) - \theta_2\|_{\ell^r} \geqslant \Delta_r - \epsilon_r I_w \tag{23}$$

and, in particular,

$$R_r(\delta, \theta_2) \geqslant (\Delta_r - \epsilon_r I_w)^r. \tag{24}$$

Hence,

$$R_r(\delta, \theta_2) \geqslant \Delta_r^r\left(1 - \frac{r\epsilon_r I_w}{\Delta_r}\right). \tag{25}$$

Proof. The triangle inequality yields

$$\|\mathrm{E}_{\theta_2}\delta(X) - \theta_2\|_{\ell^r} \geqslant \|\theta_2 - \theta_1\|_{\ell^r} - \|\mathrm{E}_{\theta_2}\delta(X) - \theta_1\|_{\ell^r}.$$

It then follows from Hölder's inequality that

$$|\mathrm{E}_{\theta_2}\delta_k(X) - \theta_{1,k}| \leqslant (\mathrm{E}_{\theta_1}|\delta_k(X) - \theta_{1,k}|^r)^{1/r}(\mathrm{E}_{\theta_1}s^w(X))^{1/w} \leqslant (\mathrm{E}_{\theta_1}|\delta_k(X) - \theta_{1,k}|^r)^{1/r} I_w,$$

and so

$$\|\mathrm{E}_{\theta_2}\delta(X) - \theta_1\|_{\ell^r} \leqslant (\mathrm{E}_{\theta_1}\|\delta(X) - \theta_1\|^r_{\ell^r})^{1/r} \cdot I_w \leqslant \epsilon_r \; I_w.$$

Therefore,

$$\|\mathrm{E}_{\theta_2}\delta(X) - \theta_2\|_{\ell^r} \geqslant \Delta_r - \epsilon_r I_w.$$

The inequality in (25) follows from Jensen's inequality and the following elementary inequality:

$$(1 - x)^r \geqslant 1 - rx, \qquad \text{for } 0 \leqslant x \leqslant 1 \text{ and } r \geqslant 1.$$

□

The following result shows that the risk lower bound (24) in Theorem 4 is sharp.

Proposition 1. *Fix* $\Delta > 0$, $B > 1$ *and* $0 < \epsilon < \Delta B^{-1/r}$. *Let* f_θ *be the uniform distribution on* $(0, 1)$ *when* $\theta = 0$ *and the uniform distribution on* $(0, B^{-1})$ *when* $\theta = \Delta$. *Then*

$$\min_{\delta : R_r(0,\delta)\leqslant\epsilon} R(\delta, \Delta) = (\Delta - \epsilon I_w)^r. \tag{26}$$

Hence, the bound (24) *is attained.*

Proof. Following Brown and Low (1996b), it is easy to see that $I_w = B^{1/r}$. Let the estimator δ^* of θ be

$$\delta^*(x) = \begin{cases} 0, & \text{if } B^{-1} \leqslant x \leqslant 1, \\ \epsilon B^{1/r}, & \text{if } 0 < x \leqslant B^{-1}. \end{cases}$$

Then it is easy to verify that $R_r(\delta, 0) = \epsilon^r$ and $R_r(\delta, \Delta) = (\Delta - \epsilon I_w)^r$. In view of Theorem 4, this proves (26). □

Theorem 4 shows that estimators with 'small' risk at θ_1 must have 'large' bias at θ_2 when $I_w < \infty$. In many common problems of interest, including Gaussian models, $I_w < \infty$ for $1 \leqslant w < \infty$. However, in the case of mean absolute error, in most problems of interest $I_\infty = \infty$. Then it is easy to construct estimators which have arbitrarily small risk for θ_1 and zero bias at θ_2. When $I_\infty = \infty$ it is useful to focus on a subset where the likelihood ratio $s(x)$ is bounded under P_{θ_1} and the measure of this subset is positive under P_{θ_2}. The following result then gives a lower bound on the mean absolute error at θ_2 for all estimators with 'small' mean absolute error at θ_1.

Proposition 2. *Suppose* $R_1(\delta, \theta_1) \leqslant \epsilon_1$ *and that there exists a measurable set* Λ_0 *such that*

$P_{\theta_2}(\Lambda_0) \geqslant \rho > 0$ and $\|s(x)\, I(x \in \Lambda_0)\|_\infty \leqslant I^0_\infty < \infty$, where the supremum norm is taken with respect to P_{θ_1}. Suppose $\Delta_1 > \epsilon_1 I^0_\infty/\rho$. Then

$$R_1(\delta, \theta_2) \geqslant \rho\,\Delta_1\left(1 - \frac{\epsilon_1 I^0_\infty}{\rho\,\Delta_1}\right). \tag{27}$$

Proof. The inequality in (27) follows from Jensen's inequality and the triangle inequality

$$R_1(\delta, \theta_2) \geqslant \mathrm{E}_{\theta_2}\{\|\delta(X) - \theta_2\|_{\ell^1} I(X \in \Lambda_0)\} \geqslant \rho\|\theta_2 - \theta_1\|_{\ell^1} - \epsilon_1 I^0_\infty.$$

□

6.1. Discussion

The Hammersley–Chapman–Robbins inequality gives a lower bound for the variance of unbiased estimators. This is a classical inequality in mathematical statistics; see Hammersley (1950), Chapman and Robbins (1951) and Lehmann (1983). The constrained inequality given in Theorem 4 yields a generalization for other loss functions.

Proposition 3. *Suppose δ is an estimator of θ which is unbiased at $\theta = \theta_1$ and $\theta = \theta_2$. Then, for $r > 1$,*

$$\mathrm{E}_{\theta_1}|\delta - \theta_1|^r \geqslant \sup_{c \in \mathbb{R}} \frac{|\theta_2 - \theta_1|^r}{(\mathrm{E}_{\theta_1}|f_{\theta_2}(X)/f_{\theta_1}(X) - c|^w)^{r/w}}. \tag{28}$$

When $r = 2$, the left-hand side is $\mathrm{var}_{\theta_1}(\delta)$ and the right-hand side is maximized at $c = 1$ and inequality (28) becomes the Hammersley–Chapman–Robbins inequality.

The constrained inequalities given in Theorem 4 are useful for providing lower bounds in nonparametric function estimation problems such as those given in Section 2. See also Cai and Low (2006) for an application to estimating a nonlinear functional.

Besides the applications in nonparametric estimation problems, the constrained risk inequality given in Theorem 4 is also useful for more standard parametric problems such as estimating a bounded normal mean. In the next section we consider super efficiency in the classical normal location–scale model.

7. Superefficiency in the normal location–scale model

Let $X_1, X_2, \ldots, X_n$ be a random sample from $N(\mu, \nu)$ with both the mean μ and the variance ν unknown. Write $\omega = (\mu, \nu)$. We wish to estimate $\theta \equiv T(\omega) = T(\mu, \nu)$, where T is a continuously differentiable function. Then the minimax rate of convergence for estimating θ under ℓ^r loss is $n^{r/2}$.

Theorem 5. *Suppose $A_n \to \infty$, and $n/\log A_n \to \infty$. Let $\omega_1 = (\mu_1, \nu_1)$ and $\theta_1 = T(\omega_1)$. Suppose $\partial T(\omega_1)/\partial\mu$ and $\partial T(\omega_1)/\partial\nu$ are not both zero. If*

$$\overline{\lim_{n\to\infty}}\, n^{r/2} A_n \mathrm{E}_{\omega_1} |\delta - \theta_1|^r < \infty$$

and the parameter set Ω *contains an ellipsoid centred at* ω_1,

$$B(\omega_1, b_n) = \{\omega = (\mu, \nu) : |\mu - \mu_1|^2 + \nu_1^{-1}|\nu - \nu_1|^2 \leqslant b_n^2\}$$

with $b_n^2 = r(\log A_n)/n$, *then*

$$\overline{\lim_{n\to\infty}} \left(\frac{n}{\log A_n}\right)^{r/2} \sup_{\omega\in\Omega} \mathrm{E}_\omega |\delta - \theta|^r > 0. \tag{29}$$

Proof. Denote the sample mean by $Y_1 = \bar{X}$ and the sum of squares by $Y_2 = \sum(X_i - \bar{X})^2$. Then the statistics (Y_1, Y_2) are sufficient and have joint density

$$f(y_1, y_2) = \frac{n^{1/2}}{\nu^{n/2}\pi^{1/2}2^{n/2}\Gamma((n-1)/2)} \mathrm{e}^{-n/(2\nu)(y_1-\mu)^2} y_2^{(n-1)/2-1} \mathrm{e}^{-y_2/(2\nu)}.$$

We first consider the case $r > 1$. Assume that $\partial T(\omega_1)/\partial\nu \neq 0$. Let $\mu_2 = \mu_1$ and $\nu_2 = (1 - ((\log A_n)/wn)^{1/2})\nu_1$. Then $\omega_2 = (\mu_2, \nu_2) \in \Omega$. Standard calculations show that

$$I_w^w(\omega_1, \omega_2) = \iint \frac{f_{\omega_2}^w(y_1, y_2)}{f_{\omega_1}^{w-1}(y_1, y_2)}\, \mathrm{d}y_1\, \mathrm{d}y_2 = \left[w\left(\frac{\nu_2}{\nu_1}\right)^{w-1} - (w-1)\left(\frac{\nu_2}{\nu_1}\right)^w \right]^{-n/2}$$

$$\to \mathrm{e}^{(w-1)\log A_n} = A_n^{w/r}.$$

Therefore, for sufficiently large n,

$$I_w(\omega_1, \omega_2) \leqslant 2A_n^{1/r}.$$

Using the Taylor expansion, one obtains

$$T(\omega_2) = T(\omega_1) + \frac{\partial T}{\partial\nu}(\omega_1)\,(\nu_2 - \nu_1) + o(|\nu_2 - \nu_1|).$$

So, when n is sufficiently large,

$$\Delta_r = |T(\omega_2) - T(\omega_1)| \geqslant c_1\left(\frac{n}{\log A_n}\right)^{-1/2},$$

where the constant $c_1 = w^{-1/2}\nu_1|\partial T(\omega_1)/\partial\nu| > 0$.

By assumption, there exists a constant $c_2 > 0$ such that for sufficiently large n,

$$\epsilon_r = (\mathrm{E}_{\omega_1}|\delta - \theta|^r)^{1/r} \leqslant c_2 n^{-1/2} A_n^{-1/r}.$$

It now follows from (24) that

$$R(\delta, \theta_2) \geqslant c_1^r\left(\frac{n}{\log A_n}\right)^{-r/2}\left[1 - \frac{rc_2 n^{-1/2} A_n^{-1/r} 2A_n^{1/r}}{c_1 n^{-1/2}(\log A_n)^{1/2}}\right].$$

The second term inside the bracket tends to zero. Therefore,

$$\overline{\lim_{n\to\infty}}\left(\frac{n}{\log A_n}\right)^{r/2}\mathrm{E}_{\omega_1}|\delta-\theta|^r > c_1^r > 0.$$

Now consider the case of $r = 1$. Again assume that $\partial T(\omega_1)/\partial\nu \neq 0$. Let

$$\omega_2 = (\mu_2, \nu_2) \equiv \left(\mu_1, \left(1 - \left(\frac{\log A_n}{n}\right)^{1/2}\right)\nu_1\right).$$

Then, exactly as before, when n is sufficiently large,

$$\Delta_1 = |T(\omega_2) - T(\omega_1)| \geqslant c_3\left(\frac{n}{\log A_n}\right)^{-1/2}$$

for some constant $c_3 > 0$. Note that

$$s(y_1, y_2) = \frac{f_{\omega_2}(y_1, y_2)}{f_{\omega_1}(y_1, y_2)} = \left(\frac{\nu_2}{\nu_1}\right)^{-n/2}\exp\left[-\frac{n}{2}\left(\frac{1}{\nu_2}-\frac{1}{\nu_1}\right)(y_1-\mu)^2\right]\exp\left[-\left(\frac{1}{\nu_2}-\frac{1}{\nu_1}\right)\frac{y_2}{2}\right]$$

$$\leqslant \exp\left[-\frac{n}{2}(\log\nu_2 - \log\nu_1) - \frac{n}{2\nu_1}\left(\frac{\nu_1}{\nu_2}-1\right)(y_1-\mu)^2\right].$$

Now suppose $(y_1 - \mu_1)^2 \geqslant \nu_1$, then

$$s(y_1, y_2) \leqslant \mathrm{e}^{-[\log\nu_2 - \log\nu_1 + \nu_1/\nu_2 - 1]\, n/2} \leqslant \mathrm{e}^{-(\log A_n)/4}(1 + o(1)) \to 0.$$

Let $\Lambda_0 = \{(y_1, y_2) : (y_1 - \mu_1)^2 \geqslant \nu_1\}$. Then

$$P_{\omega_2}(\Lambda_0) = P_{\omega_2}(|y_1 - \mu_1| \geqslant \nu_1^{1/2}) = 2\Phi(-(\nu_1/\nu_2)^{1/2}) \to 2\Phi(-1) = 0.3174.$$

So, when n is sufficiently large,

$$\sup[s(y_1, y_2)\cdot I((y_1, y_2)\in\Lambda_0)] \leqslant 1 \quad \text{and} \quad P_{\omega_2}(\Lambda_0) \geqslant 0.3.$$

Now applying inequality (27) with $\rho = 0.3$ and $I_\infty^0 = 1$, we have

$$R(\delta, \theta_2) \geqslant \rho c_3\left(\frac{n}{\log A_n}\right)^{-1/2}\left[1 - \frac{c_2 n^{-1/2} A_n^{-1}}{\rho c_3 n^{-1/2}(\log A_n)^{1/2}}\right].$$

Again, the second term inside the brackets tends to zero. So (29) holds.

The case of $\partial T(\omega_1)/\partial\nu = 0$ and $\partial T(\omega_1)/\partial\mu \neq 0$ are similar for both $r > 1$ and $r = 1$. It is in fact slightly simpler. We omit the proof here. □

Remarks. (i) It suffices to assume that there exists a non-zero directional derivative of T at $\omega_1 = (\mu_1, \nu_1)$.

(ii) The result (29) holds if, for $r > 1$, $(\mu_1 + ((r-1)n^{-1}\log A_n)^{1/2}, \nu_1)$ and $(\mu_1, (1 - (w^{-1}n^{-1}\log A_n)^{1/2})\nu_1)$ are in Ω, and for $r = 1$, $(\mu_1 + (n^{-1}\log A_n)^{1/2}, \nu_1)$ and $(\mu_1, (1 - (n^{-1}\log A_n)^{1/2})\nu_1)$ are in Ω.

(iii) The result can be naturally extended to the case where X_i has a multivariate normal distribution.

8. Proof of Theorems 1, 2 and 3

Proof of Theorem 1. We will use the constrained risk inequality given in Theorem 4 to prove Theorem 1.

Suppose that

$$\overline{\lim_{n\to\infty}}\, A_{\alpha,p,r}(n)\mathrm{E}\int_0^1 |\hat{f}_n(t) - f_0(t)|^r \,\mathrm{d}t = K < \infty.$$

By Fubini's theorem,

$$\mathrm{E}\int_0^1 |\hat{f}_n(t) - f_0(t)|^r \,\mathrm{d}t = \int_0^1 \mathrm{E}|\hat{f}_n(t) - f_0(t)|^r \,\mathrm{d}t.$$

So, for sufficiently large n, say $n \geqslant 2^{k_1}$,

$$\int_0^1 \mathrm{E}|\hat{f}_n(t) - f_0(t)|^r \,\mathrm{d}t \leqslant 2KA^{-1}_{\alpha,p,r}(n). \tag{30}$$

Let $0 < \gamma < \frac{1}{2}[g(\alpha, p, r) - l(\alpha, p, r)]$ and write

$$S_n = \{x : \mathrm{E}|\hat{f}_n(x) - f_0(x)|^r \geqslant n^{-l(\alpha,p,r)}n^{-\gamma}\}.$$

Then equation (30) yields that

$$\begin{aligned} 2KA^{-1}_{\alpha,p,r}(n) &\geqslant \int_0^1 \mathrm{E}|\hat{f}_n(t) - f_0(t)|^r \,\mathrm{d}t \\ &\geqslant \int_{S_n} \mathrm{E}|\hat{f}_n(t) - f_0(t)|^r \,\mathrm{d}t \\ &\geqslant n^{-l(\alpha,p,r)}n^{-\gamma} \cdot m(S_n). \end{aligned}$$

Since $l(\alpha, p, r) + \gamma - g(\alpha, p, r) < -\gamma$, the Lebesgue measure of S_n, $m(S_n)$, satisfies

$$m(S_n) \leqslant 2Kn^{l(\alpha,p,r)+\gamma}A^{-1}_{\alpha,p,r}(n) \leqslant n^{-\gamma}$$

for all sufficiently large n, say $n \geqslant 2^{k_2}$. Let $n_k = 2^k$ and let k_0 be the smallest integer satisfying $2^{-\gamma k_0} \leqslant (1 - 2^{-\gamma})\epsilon$. Let $k_* = \max\{k_0, k_1, k_2\}$ and set

$$S = \bigcup_{k=k_*}^{\infty} S_{n_k} \quad \text{and} \quad \Omega_0 = \Omega \cap S^c.$$

It is easy to see that $m(S) \leqslant \sum_{k=k_*}^{\infty} m(S_{n_k}) \leqslant \sum_{k=k_0}^{\infty} 2^{-\gamma k_0} \leqslant \epsilon$. Hence,

$$m(\Omega_0) \geqslant m(\Omega) - m(S) \geqslant m(\Omega) - \epsilon.$$

Note that for all $x \in \Omega_0$,

$$\mathrm{E}|\hat{f}_{n_k}(x) - f_0(x)|^r < n_k^{-l(\alpha,p,r)}n_k^{-\gamma}, \qquad \text{for all } k \geqslant k_*. \tag{31}$$

We now show, using the constrained risk inequality, that for any fixed $x \in \Omega_0$ there exists $f_1 \in B^{\alpha}_{p,q}(M)$ such that for $k > k_*$,

$$\mathrm{E}|\hat{f}_{n_k}(x) - f_1(x)|^r \geqslant C\left(\frac{\log n_k}{n_k}\right)^{l(\alpha,p,r)}$$

for some constant $C > 0$ not depending on x. We give only the proof for $r > 1$. The case of $r = 1$ is similar.

Let g be a compactly supported function satisfying the following conditions:

$$g(0) > 0, \qquad \|g\|_2^2 > 0, \qquad g \in B^{\alpha}_{p,q}(M - M').$$

Such a function can be easily constructed either directly or by using wavelets. Write $b = 2(1 - 1/r)\|g\|_2^{-2}$ and let

$$\gamma_{n_k} = \left(\frac{n_k}{b\gamma \log n_k}\right)^{\nu/(1+2\nu)} \quad \text{and} \quad \beta_{n_k} = \left(\frac{n_k}{b\gamma \log n_k}\right)^{1/(1+2\nu)}.$$

Then

$$\gamma_{n_k}^2 \beta_{n_k} = \frac{n_k}{b\gamma \log n_k} \quad \text{and} \quad \gamma_{n_k}^{-1}\beta_{n_k}^{\nu} = 1.$$

Let

$$f_1(x) = \gamma_{n_k}^{-1} g(\beta_{n_k}(t - x)) + f_0(x). \tag{32}$$

It is straightforward to check that $f_1 \in B^{\alpha}_{p,q}(M)$.

Write P_0^n and P_1^n for the probability measure associated with the process (1) with $f = f_0$ and $f = f_1$, respectively. Then a sufficient statistic for the family of measures $\{P_0^n, P_1^n\}$ is given by $T_n = \log(\mathrm{d}P_1^n/\mathrm{d}P_0^n)$ with

$$T_n \sim \begin{cases} N(-\rho_n/2, \rho_n), & \text{under } P_0^n, \\ N(\rho_n/2, \rho_n), & \text{under } P_1^n, \end{cases}$$

where

$$\rho_n = n\|f_1 - f_0\|_2^2 = n\gamma_n^{-2}\beta_n^{-1}\|g\|_2^2 = 2\left(1 - \frac{1}{r}\right)\gamma \log n.$$

Write $\delta_n = f_n(x)$, $\theta_0 = f_0(x)$ and $\theta_1 = f_1(x)$. Then (31) can be rewritten as

$$\mathrm{E}_{\theta_0}|\delta_{n_k} - \theta_0|^r \leqslant n_k^{-l(\alpha,p,r)}\; n_k^{-\gamma}.$$

Since T_n is sufficient for $\{P_0^n, P_1^n\}$ we may apply Theorem 4(i) along the subsequence $\{n_k\}$. Let w satisfy $1/r + 1/w = 1$. Noting that $(r-1)(w-1) = 1$, we have

$$I_w(\theta_0, \theta_1) = \mathrm{e}^{\rho_{n_k}\cdot(w-1)/2} = \mathrm{e}^{2(1-1/r)\gamma(\log n_k)(w-1)/2} = n_k^{\gamma/r}.$$

It follows from Theorem 4(i), after some algebra, that for $k \geqslant k_*$,

$$
\begin{aligned}
\mathrm{E}_{\theta_1}|\delta_{n_k} - \theta_1|^r &\geqslant \left(\frac{g(0)}{\gamma_{n_k}}\right)^r \left(1 - rn_k^{-\nu/(1+2\nu)}\; n_k^{-\gamma/r} \cdot n_k^{\gamma/r} \cdot (g(0))^{-1}\gamma_{n_k}\right) \\
&= (b\gamma g(0))^{l(\alpha,p,r)} \left(\frac{\log n_k}{n_k}\right)^{l(\alpha,p,r)} (1 + o(1)).
\end{aligned}
$$

Hence, for $k \geqslant k_*$,

$$
\sup_{f \in B^\alpha_{p,q}(M)} \mathrm{E}|f_{n_k}(x) - f(x)|^r \geqslant (b\gamma g(0))^{l(\alpha,p,r)} \left(\frac{\log n_k}{n_k}\right)^{l(\alpha,p,r)}.
$$

□

Proof of Theorem 2. The global optimality of the estimator (15) can be shown using the same proof as given in Delyon and Juditsky (1996). In fact the proof can be slightly simpler in the white noise model. We will prove the near-optimality (17) under the pointwise risk.

We need the following risk bound which can be derived from the general ℓ^r oracle inequality given in Cai (2003: Theorem 7).

Lemma 1. *Let $y \sim N(\theta, \sigma^2)$ and let $\hat\theta = \eta_{\lambda\sigma}(y) = \mathrm{sgn}(y)(|y| - \lambda\sigma)_+$ be a soft threshold estimator of θ with $\lambda \geqslant 1$. Then, for any $1 \leqslant r < \infty$,*

$$
\mathrm{E}|\hat\theta - \theta|^r \leqslant \min(|\theta|^r, 2^r\lambda^r\sigma^r) + C(r)\lambda^r \mathrm{e}^{-\lambda^2/2}\sigma^r, \tag{33}
$$

where $C(r) > 0$ is a constant depending on r only.

We now recall the Minkowski inequality. Let X_i be random variables, $i = 1, \ldots, n$. Then, for $1 \leqslant r < \infty$,

$$
\mathrm{E}\left|\sum_{i=1}^n X_i\right|^r \leqslant \left(\sum_{i=1}^n (\mathrm{E}|X_i|^r)^{1/r}\right)^r. \tag{34}
$$

Applying the Minkowski inequality (34), we have

$$
\begin{aligned}
\mathrm{E}|\hat f_n(x) - f(x)|^r &= \mathrm{E}\left|\sum_{k=1}^{2^{J_0}} (\hat\xi_{J_0k} - \xi_{J_0k})\phi_{J_0k}(x) + \sum_{j=J_0}^{\infty}\sum_{k=1}^{2^j} (\hat\theta_{jk} - \theta_{jk})\psi_{jk}(x)\right|^r \\
&\leqslant \left[\sum_{k=1}^{2^{J_0}} |\phi_{J_0k}(x)|(\mathrm{E}|\hat\xi_{J_0k} - \xi_{J_0k}|^r)^{1/r} + \sum_{k=1}^{2^{J_0}} |\psi_{J_0k}(x)|(\mathrm{E}|\hat\theta_{J_0k} - \theta_{J_0k}|^r)^{1/r}\right. \\
&\qquad \left. + \sum_{j=J_0+1}^{J-1}\sum_{k=1}^{2^j} |\psi_{jk}(x)|(\mathrm{E}|\hat\theta_{jk} - \theta_{jk}|^r)^{1/r} + \sum_{j=J}^{\infty}\sum_{k=1}^{2^j} |\theta_{jk}\psi_{jk}(x)|\right]^r \\
&\equiv (T_1 + T_2 + T_3 + T_4)^r.
\end{aligned}
$$

It is easy to see that both T_1 and T_2 are small:

$$T_1 = O(n^{-\alpha/(1+2\alpha)}) \quad \text{and} \quad T_2 = O(n^{-\alpha/(1+2\alpha)}). \tag{35}$$

Write $A_j(x) = \{k : \psi_{j,k}(x) \neq 0\}$. Then $\text{card}(A_j(x)) \leqslant L$, where L is the support length of ψ. For all $f \in B^\alpha_{p,q}(M)$,

$$|\theta_{j,k}| \leqslant C2^{-j(\alpha+1/2-1/p)} \tag{36}$$

for all (j, k), where C is a constant. It is then easy to verify that T_4 is also small:

$$T_4 = O(n^{-\nu}). \tag{37}$$

We now consider the term T_3. With $\sigma = n^{-1/2}$ and $\lambda = (r\log(2^{j-J_0}))^{1/2}$, the oracle inequality (33) and the bound on the coefficient (36) yield that for $J_0 < j < J$,

$$\mathrm{E}|\hat{\theta}_{jk} - \theta_{jk}|^r \leqslant C\min(2^{-jr(\alpha+1/2-1/p)}, (j-J_0)^{r/2}n^{-r/2}) + C(j-J_0)^{r/2}2^{-r(j-J_0)/2}n^{-r/2}. \tag{38}$$

Let J_2 be an integer satisfying

$$\left(\frac{n}{\log n}\right)^{1/(1+2\nu)} \leqslant 2^{J_2} < 2\left(\frac{n}{\log n}\right)^{1/(1+2\nu)}.$$

Applying (38) together with the elementary inequality $(a+b)^{1/r} \leqslant a^{1/r} + b^{1/r}$ for $a, b \geqslant 0$, we have

$$\begin{aligned}
T_3 &\leqslant \sum_{j=J_0+1}^{J-1} \sum_{k\in A_j(x)} 2^{j/2}\|\psi\|_\infty(\mathrm{E}|\hat{\theta}_{jk} - \theta_{jk}|^r)^{1/r} \\
&\leqslant C\sum_{j=J_0+1}^{J-1} 2^{j/2}\min(2^{-j(\alpha+1/2-1/p)}, (j-J_0)^{1/2}n^{-1/2}) + C\sum_{j=J_0+1}^{J-1} 2^{j/2}(j-J_0)^{1/2}2^{-(j-J_0)/2}n^{-1/2} \\
&\leqslant C\sum_{j=J_0+1}^{J_2-1} 2^{j/2}j^{1/2}n^{-1/2} + C\sum_{j=J_2}^{J-1} 2^{j/2}2^{-j(\alpha+1/2-1/p)} + C(\log n)^{3/2}n^{-\alpha/(1+2\alpha)} \\
&\leqslant C\left(\frac{n}{\log n}\right)^{-\nu/(1+2\nu)}(1+o(1)).
\end{aligned} \tag{39}$$

Combining (35), (37), (39), we have

$$\mathrm{E}|\hat{f}_n(x) - f(x)|^r \leqslant C\left(\frac{n}{\log n}\right)^{-r\nu/(1+2\nu)}(1+o(1)).$$

□

Proof of Theorem 3. Suppose, for some $\rho > l(\alpha, p, r)$ and some f with $\|f\|_{B^\alpha_{p,q}} < M$, that

$$\underline{\lim}_{n\to\infty} n^\rho \cdot \mathrm{E}\|f_n - f\|_r^r < \infty.$$

Then there exists a subsequence n_i such that

$$\overline{\lim_{i\to\infty}}\, n_i^{\rho}\cdot \mathrm{E}\|\hat{f}_{n_i}-f\|_r^r<\infty.$$

It then follows from the same proof as that of Theorem 1 with $A_{\alpha,p,r}(n)$ replaced by n^{ρ} and n replaced by n_i that there exists a subset $\Omega_0\subseteq\Omega\subset(0,\,1)$ with the Lebesgue measure $m(\Omega_0)>0$ such that for all $x\in\Omega_0$,

$$\overline{\lim_{i\to\infty}}\left(\frac{n_i}{\log n_i}\right)^{l(\alpha,p,r)}\sup_{f\in B^{\alpha}_{p,q}(M)}\mathrm{E}|\hat{f}_{n_i}(x)-f(x)|^r>0.$$

This contradicts the assumption that

$$\overline{\lim_{n\to\infty}}\, n^{l(\alpha,p,r)}\sup_{f\in B^{\alpha}_{p,q}(M)}\mathrm{E}|\hat{f}_n(x)-f(x)|^r<\infty$$

for every fixed $x\in\Omega$. □

References

Bickel, P.J. and Ritov, Y. (2003) Nonparametric estimators which can be ‘plugged-in’. *Ann. Statist.*, **31**, 1033–1053.

Brown, L.D. and Low, M.G. (1996a) Asymptotic equivalence of nonparametric regression and white noise. *Ann. Statist.*, **24**, 2384–2398.

Brown, L.D. and Low, M.G. (1996b) A constrained risk inequality with applications to nonparametric functional estimations. *Ann. Statist.*, **24**, 2524–2535.

Brown, L.D., Cai, T., Low, M.G. and Zhang, C. (2002) On asymptotic equivalence of white noise model and nonparametric regression with random designs. *Ann. Statist.*, **30**, 688–707.

Brown, L.D., Carter, A.V., Low, M.G. and Zhang, C. (2004) Equivalence theory for density estimation, Poisson processes and Gaussian white noise with drift. *Ann. Statist.*, **32**, 2074–2097.

Cai, T. (1999) Adaptive wavelet estimation: A block thresholding and oracle inequality approach. *Ann. Statist.*, **27**, 898–924.

Cai, T. (2002) On adaptive estimation of a derivative and other related linear inverse problems. *J. Statist. Plann. Inference*, **108**, 329–349.

Cai, T. (2003) Rates of convergence and adaptation over Besov spaces under pointwise risk. *Statist. Sinica*, **13**, 881–902.

Cai, T. and Low, M. (2006) Optimal adaptive estimation of a quadratic functional. *Ann. Statist*., **34**, 2298–2325.

Chapman, D.G. and Robbins, H. (1951) Minimum variance estimation without regularity assumptions. *Ann. Math. Statist.*, **22**, 581–586.

Cohen, A., Daubechies, I., Jawerth, B. and Vial, P. (1993) Multiresolution analysis, wavelets, and fast algorithms on an interval. *C. R. Acad. Sci. Paris Sér I*, **316**, 417–421.

Daubechies, I. (1992) *Ten Lectures on Wavelets*. Philadelphia: SIAM.

Daubechies, I. (1994) Two recent results on wavelets: wavelet bases for the interval, and biorthogonal wavelets diagonalizing the derivative operator. In L.L. Schumaker and G. Webb (eds), *Recent Advances in Wavelet Analysis*. pp. 237–258. Boston: Academic Press.

Delyon, B. and Juditsky, A. (1996) On minimax wavelet estimators. *Appl. Comput. Harmon. Anal.*, **3**, 215–228.

DeVore, R. and Popov, V. (1988) Interpolation of Besov spaces. *Trans. Amer. Math. Soc.*, **305**, 397–414.

Donoho, D.L. and Johnstone, I.M. (1995) Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.*, **90**, 1200–1224.

Donoho, D.L. and Liu, R.C. (1991) Geometrizing rates of convergence III. *Ann. Statist.*, **19**, 668–701.

Donoho, D.L., Johnstone, I.M., Kerkyacharian, G. and Picard, D. (1995) Wavelet shrinkage: asymptopia? *J. Roy. Statist. Soc. Ser. B*, **57**, 301–369.

Donoho, D.L., Johnstone, I.M., Kerkyacharian, G. and Picard, D. (1996) Density estimation by wavelet thresholding. *Ann. Statist.*, **24**, 508–539.

Efromovich, S.Y. (1999) *Nonparametric Curve Estimation.* New York: Springer-Verlag.

Efromovich, S.Y. (2002) On blockwise shrinkage estimation. Technical report.

Efromovich, S. (2004) Density estimation for biased data. *Ann. Statist.*, **32**, 1137–1161.

Efromovich, S. and Low, M.G. (1994) Adaptive estimates of linear functionals. *Probab. Theory Related Fields*, **98**, 261–275.

Efromovich, S.Y. and Pinsker, M.S. (1984) An adaptive algorithm of nonparametric filtering. *Autom. Remote Control*, **11**, 58–65.

Hammersley, J.M. (1950) On estimating restricted parameters. *J. Roy. Statist. Soc. Ser. B*, **12**, 192–240.

Härdle, W., Kerkyacharian, G., Picard, D. and Tsybakov, A. (1998) *Wavelets, Approximation, and Statistical Applications*. New York: Springer-Verlag.

Ibragimov, I.A. and Hasminskii, R.Z. (1984) Nonparametric estimation of the values of a linear functional in Gaussian white noise. *Theory Probab. Appl.*, **29**, 18–32.

Klemelä, J. and Nussbaum, M. (1999) Constructive asymptotic equivalence of density estimation and Gaussian white noise. Discussion Paper No. 53, Sonderforschungsbereich 373, Humboldt University, Berlin.

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

Lepski, O.V. (1990) On a problem of adaptive estimation on white Gaussian noise. *Theory Probab. Appl.*, **35**, 454–466.

Meyer, Y. (1991) Ondelettes sur l’intervalle. *Rev. Mat. Iberoamericana*, **7**, 115–133.

Meyer, Y. (1992) *Wavelets and Operators*. Cambridge: Cambridge University Press.

Nussbaum, M. (1996) Asymptotic equivalence of density estimation and Gaussian white noise. *Ann. Statist.*, **24**, 2399–2430.

Pinsker, M.S. (1980) Optimal filtering of square integrable signals in Gaussian white noise. *Problems Inform. Transmission*, **16**, 120–133.

Tsybakov, A.B. (1998) Pointwise and sup-norm sharp adaptive estimation of functions on the Sobolev classes. *Ann. Statist.*, **26**, 2420–2469.

Received January 2005 and revised June 2006