On Modulus of Continuity And Adaptability in Nonparametric Functional Estimation

T. Tony Cai and Mark G. Low
Department of Statistics
The Wharton School
University of Pennsylvania
Philadelphia, PA 19104

Abstract

We study adaptive estimation of linear functionals over a collection of finitely many parameter spaces.A between class modulus of continuity, a geometric quantity, is introduced and is shown to be instrumental in characterizing the degree of adaptability over two parameter spaces in the same way that the usual modulus of continuity captures the minimax difficulty of estimation over a single parameter space. The between class modulus of continuity is used to describe when full adaptation is possible. A sharp rate optimal lower bound on the cost of adaptation is derived. An ordered modulus of continuity is used to construct a procedure which which is within a constant factor of attaining these bounds. The results are complemented by several illustrative examples.

Keywords: Adaptability, adaptive estimation, between class modulus, constrained risk inequality, cost of adaptation, modulus of continuity, nonparametric functional estimation, white noise model.

AMS 2000 Subject Classification: Primary: 62G99; Secondary: 62F12, 62F35, 62M99.

1 Introduction

The problem of estimating linear functionals occupies an important position in the general theory of nonparametric function estimation. In particular the global recovery of an unknown function is sometimes best viewed as estimation at each point: an example of a problem of estimating a linear functional. When attention is focused on pointwise estimation a natural goal is the construction of estimators which adapt to the unknown local smoothness of the function. In this framework an adaptive estimator would be one which is simultaneously near minimax over a number of different local smoothness classes.

As a step towards this goal of adaptive estimation attention first focused on the more concrete goal of developing a minimax theory for estimating linear functionals over a fixed parameter space which can for example specify the smoothness of the function. This theory is now well developed particularly in the white noise with drift model:

$$dY(t) = f(t)dt + \frac{1}{\sqrt{n}}dW(t) \tag{1}$$

where $W(t)$ is a standard Brownian motion. Ibragimov and Hasminskii (1981) showed that this model captures many of the essential conceptual difficulties of other nonparametric models without some of the additional technical features. This model also arises as an approximation to many other nonparametric function models such as that of density estimation,nonparametric regression and spectrum estimation. See for example Low (1992), Brown and Low (1996a), Nussbaum (1996).

Based on white noise data a general theory for estimating a linear functional has been given assuming only that the parameter space is convex. The first general result of this type was given in Ibragimov and Hasminskii (1984) where the parameter space was also assumed to be symmetric. Ibragimov and Hasminskii (1984) constructed a linear estimator with the smallest maximum mean squared error. Donoho and Liu (1991) and Donoho (1994) extended this theory to general convex parameter spaces.

In the general case of a fixed convex parameter space $\mathcal{F}$, Donoho and Liu (1991) and Donoho (1994) showed that the minimax theory for estimating a linear functional Tf over $\mathcal{F}$ can be completely described by a modulus of continuity

$$\omega(\epsilon, \mathcal{F}) = \sup\{|Tg - Tf| : \|g - f\|_2 \leq \epsilon; f \in \mathcal{F}, g \in \mathcal{F}\}. \tag{2}$$

Affine estimators play a fundamental role in this theory. The minimax affine risk $R_A^*(n, \mathcal{F})$ and the minimax risk $R_N^*(n, \mathcal{F})$ of estimating Tf over a convex function class

$\mathcal{F}$ can be defined respectively by

$$R_A^*(n,\mathcal{F}) = \inf_{\hat{T} \text{ affine}} \sup_{f\in\mathcal{F}} E(\hat{T} - Tf)^2$$

and

$$R_N^*(n,\mathcal{F}) = \inf_{\hat{T}} \sup_{f\in\mathcal{F}} E(\hat{T} - Tf)^2.$$

Donoho and Liu (1991) and Donoho (1994) have shown that

$$\frac{1}{8}\omega^2(\frac{1}{\sqrt{n}},\mathcal{F}) \le R_A^*(n,\ \mathcal{F}) = \sup_{\epsilon>0} \omega^2(\epsilon,\ \mathcal{F})\frac{\frac{1}{4n}}{\frac{1}{n}+\frac{\epsilon^2}{4}} \le \omega^2(\frac{1}{\sqrt{n}},\mathcal{F}) \tag{3}$$

and that the modulus can be used to give a recipe for constructing an affine procedure which has the maximum mean squared error attaining the risk given in (3). In addition Donoho and Liu (1991) also showed that

$$\frac{R_A^*(n,\ \mathcal{F})}{R_N^*(n,\ \mathcal{F})} \le 1.25 \tag{4}$$

and so the maximum risk of the optimal affine procedure is within a small constant factor of the minimax risk when the parameter space is convex. It should be stressed that (3) and (4) taken together show that the minimax mean squared error for estimating a linear functional over a convex parameter space is always of order $\omega^2(\frac{1}{\sqrt{n}},\mathcal{F})$.

It should also be emphasized that the minimax theory described above is a theory for a given convex parameter space. The goal which we alluded to earlier of constructing estimators which adapt to the smoothness of the function is not covered by this theory. A natural way to extend the minimax theory to an adaptation theory is to construct estimators which are simultaneously minimax or near minimax over a collection of smoothness classes. In general however this goal cannot be realized.

In particular Lepski (1990) was the first to give examples which demonstrated that rate optimal adaptation over a collection of Lipschitz classes is not possible when estimating the function at a point. In this case a logarithmic penalty must be paid over all but one of the parameter spaces. Efromovich and Low (1994) showed that this phenomena is true in general whenever we try to adapt over a collection of nested symmetric sets where the minimax rates are algebraic of different orders. Klemelä and Tsybakov (2001) give sharp results for adaptive estimation over convex, symmetric renormalizable function classes. Lepski and Levit (1998) study adaptive estimation over infinitely differentiable functions.

On the other hand the goal of fully rate adaptive estimation of linear functionals can sometimes be realized. When the minimax rates over each parameter space is slower than any algebraic rate Cai and Low (2001) have given examples of nested symmetric sets where fully adaptive estimators can be constructed. In addition, when the sets are not symmetric there are also examples where rate adaptive estimators can be constructed. Such is the case for estimating monotone functions where an estimator can adapt over different Lipschitz classes. See Kang and Low (2002). Other recent results on adaptive estimation can be found in Butucea (2001), Efromovich (1997a,b, 2000) and Efromovich and Koltchinskii (2001).

Although the above mentioned examples show that there are cases where fully rate adaptive estimators exist and other cases where fully rate adaptive estimators do not exist, to date there is no general theory that characterizes exactly when adaptation is possible. In the present paper building upon the framework of Donoho and Liu (1991) and Donoho (1994) we provide a general adaptation theory for estimating linear functionals.

In Section 2 we extend the modulus of continuity defined by (2) to a between class modulus of continuity and show that the between class modulus can be used to characterize when adaptation is possible. This modulus captures the degree of adaptability over two parameter spaces in the same way that the usual modulus of continuity captures the minimax difficulty of estimation over a single parameter space. In Section 2 a sharp rate optimal lower bound on the cost of adaptation is given in terms of the between class modulus.

Another closely related modulus which we call an ordered modulus is introduced in Section 2 and used in Section 4 to construct adaptive estimators that are within a constant factor of this lower bound. The bounds and construction hold for an arbitrary pair of convex parameter spaces and any linear functional. In particular we do not assume that the parameter spaces are nested or symmetric. We also show that this general theory for two parameter spaces can be extended to any finite nested collection of convex parameter spaces and under mild regularity conditions on the modulus can be extended to finitely many non-nested convex parameter spaces. An estimator with minimum adaptation cost over a collection of parameter spaces is constructed in Section 5.

The theory also shows that there are three main cases in terms of the cost of adaptation. We shall call the first case the regular one where as in the case of estimating a function at a point over Lipschitz classes considered by Lepski (1990) the cost of adaptation is a logarithmic factor of the noise level. In the second case full adaptation is possible as in the examples considered in Lepski and Levit (1998) and Cai and Low (2001). More dramatically in the third case the cost of adaptation is much greater than in the regular case. The cost of adaptation in this case is a power of the noise level. Examples of all three cases are given in Section 3. In addition, we give in Section 5 an interesting example of adaptively estimating a function at a fixed point over a finite collection of parameter spaces where one loses a necessary logarithmic penalty on some parameter spaces and yet is fully adaptive over other spaces.

2 Lower bound on the cost of adaptation

In this section we focus on lower bounds for the cost of adaptation where it suffices to consider just two parameter spaces, say $\mathcal{F}_1$ and $\mathcal{F}_2$. The bounds are easily expressed in terms of a between class modulus of continuity. For convenience however we first define an ordered modulus of continuity $\omega(\epsilon,\ \mathcal{F}_1, \mathcal{F}_2)$ between two classes $\mathcal{F}_1$ and $\mathcal{F}_2$ as

$$\omega(\epsilon,\ \mathcal{F}_1, \mathcal{F}_2) = \sup\{Tg - Tf :\ \|g-f\|_2 \le \epsilon; f \in \mathcal{F}_1, g \in \mathcal{F}_2\}.$$

The ordered modulus of continuity will be useful in the construction of adaptive estimators given in Sections 4 and 5. This is a quantity derived from the geometry of the graph of the linear functional T between the regularity classes $\mathcal{F}_1$ and $\mathcal{F}_2$. Note that $\omega(\epsilon,\ \mathcal{F}_1, \mathcal{F}_2)$ does not necessarily equal $\omega(\epsilon,\ \mathcal{F}_2, \mathcal{F}_1)$. It is however clear that the modulus $\omega(\epsilon,\ \mathcal{F}_1, \mathcal{F}_2)$ is an increasing function of ϵ. Moreover if $\mathcal{F}_1$ and $\mathcal{F}_2$ are convex with $\mathcal{F}_1 \cap \mathcal{F}_2 \neq \emptyset$, then for a linear functional T the modulus $\omega(\epsilon,\ \mathcal{F}_1, \mathcal{F}_2)$ is also a concave function of ϵ. See Cai and Low (2002). In particular, for $D > 1$

$$\omega(D\epsilon,\ \mathcal{F}_1, \mathcal{F}_2) \le D\omega(\epsilon,\ \mathcal{F}_1, \mathcal{F}_2). \tag{5}$$

For the statement of lower bounds for the risk it is also convenient to define a between class modulus of continuity $\omega_+(\epsilon, \mathcal{F}_1, \mathcal{F}_2)$

$$\omega_+(\epsilon,\ \mathcal{F}_1, \mathcal{F}_2) = \sup\{|Tg - Tf| :\ \|g-f\|_2 \le \epsilon; f \in \mathcal{F}_1, g \in \mathcal{F}_2\}. \tag{6}$$

Clearly, $\omega_+(\epsilon,\ \mathcal{F}_1, \mathcal{F}_2) = \max(\omega(\epsilon,\ \mathcal{F}_1, \mathcal{F}_2), \omega(\epsilon,\ \mathcal{F}_2, \mathcal{F}_1))$. Note also that although ω_+ need not be concave it follows from (5) that

$$\omega_+(D\epsilon,\ \mathcal{F}_1, \mathcal{F}_2) \le D\omega_+(\epsilon,\ \mathcal{F}_1, \mathcal{F}_2). \tag{7}$$

When $\mathcal{F}_1 = \mathcal{F}_2 = \mathcal{F}$, $\omega(\epsilon, \mathcal{F}, \mathcal{F}) = \omega_+(\epsilon, \mathcal{F}, \mathcal{F})$ is the usual modulus of continuity over $\mathcal{F}$ and will be denoted by $\omega(\epsilon,\ \mathcal{F})$ as in (2).

In the minimax theory of estimating linear functionals Donoho and Liu (1991) showed that in many problems $\omega(\epsilon, \mathcal{F})$ is Hölderian, $\omega(\epsilon, \mathcal{F}) \sim C\epsilon^{q(\mathcal{F})}$ where $0 < q(\mathcal{F}) \leq 1$ and hence from (3) and (4) the minimax mean squared error rate is of order $n^{-q(\mathcal{F})}$. In related adaptive estimation problems the between class modulus of continuity $\omega_+(\epsilon,\ \mathcal{F}_1,\ \mathcal{F}_2)$ is also typically Hölderian $\omega_+(\epsilon,\ \mathcal{F}_1,\ \mathcal{F}_2) \sim C\epsilon^{q(\mathcal{F}_1, \mathcal{F}_2)}$. We shall show later in this section that the cost of adaptation can then easily be explained in terms of the relationship between $\min(q(\mathcal{F}_1), q(\mathcal{F}_2))$ and $q(\mathcal{F}_1, \mathcal{F}_2)$. See the following discussion and examples in Section 3.

The problem of adaptability is also easily explored when the modulus is ordinary.

Definition 1 *We shall call a between class modulus* $\omega_+(\epsilon, \mathcal{F}_1, \mathcal{F}_2)$ **ordinary** *if*

$$\overline{\lim_{D\to\infty}}\ \overline{\lim_{\epsilon\to 0}} \frac{\omega_+(\epsilon, \mathcal{F}_1, \mathcal{F}_2)}{\omega_+(D\epsilon, \mathcal{F}_1, \mathcal{F}_2)} = 0 \tag{8}$$

or equivalently if for some $D > 1$

$$\overline{\lim_{\epsilon\to 0}} \frac{\omega_+(\epsilon, \mathcal{F}_1, \mathcal{F}_2)}{\omega_+(D\epsilon, \mathcal{F}_1, \mathcal{F}_2)} < 1. \tag{9}$$

Note that if the between class modulus is Hölderian with $0 < q \leq 1$ then it is also ordinary.

The following results all give bounds for the risk under various assumptions on the modulus of continuity. Theorem 1 is the most general whereas Theorem 2 is useful for providing the lower bounds for the performance of the adaptive estimators given in Sections 4 and 5. Corollary 1 considers an important case for adaptability and Corollary 2 gives sharp lower bounds for the cost of adaptation in the case of Hölderian moduli when adaptation for free is impossible. Corollary 2 is also useful for the applications in Section 3. In the results that follow we write

$$R(\hat{T},\ Tf) = E_f(\hat{T} - Tf)^2$$

for the mean squared error of an estimator $\hat{T}$ based on the white noise data (1).

Theorem 1 *Consider two function classes* $\mathcal{F}_1$ *and* $\mathcal{F}_2$ *with* $\mathcal{F}_1 \cap \mathcal{F}_2 \neq \emptyset$. *Let* T *be a linear functional and suppose that*

$$\sup_{f\in\mathcal{F}_1} R(\hat{T},\ Tf) \leq \gamma^{-2}\omega_+^2(\frac{1}{\sqrt{n}},\ \mathcal{F}_1,\ \mathcal{F}_2) \tag{10}$$

for some $\gamma > 1$. Then for any $0 < \rho \le 1$

$$\sup_{f\in\mathcal{F}_2} (R(\hat{T},\ Tf))^{\frac{1}{2}} \ge \omega_+(\sqrt{\frac{\rho\ln\gamma^2}{n}}, \mathcal{F}_1,\ \mathcal{F}_2) - \gamma^{-(1-\rho)}\omega_+(\frac{1}{\sqrt{n}}, \mathcal{F}_1,\ \mathcal{F}_2). \tag{11}$$

Now suppose there exists a sequence $d_n > 0$ such that

$$\overline{\lim_{n\to\infty}}\, d_n^{-1} \cdot \sup_{f\in\mathcal{F}_1} R(\hat{T},\ Tf) \le \lambda \tag{12}$$

with $0 < \lambda < \infty$ and

$$h_n \equiv \frac{d_n}{\omega_+^2(\frac{1}{\sqrt{n}},\ \mathcal{F}_1, \mathcal{F}_2)} \to 0. \tag{13}$$

Then

$$\overline{\lim_{n\to\infty}} \frac{\sup_{f\in\mathcal{F}_2} R(\hat{T},\ Tf)}{\omega_+^2(\sqrt{\frac{A_n}{n}},\ \mathcal{F}_1, \mathcal{F}_2)} \ge 1 \tag{14}$$

with $A_n = -\frac{1}{2}\ln(\lambda h_n) \to \infty$. In particular if in addition the modulus $\omega_+(\epsilon, \mathcal{F}_1,\ \mathcal{F}_2)$ is ordinary as defined in (8) then

$$\overline{\lim_{n\to\infty}} \frac{\sup_{f\in\mathcal{F}_2} R(\hat{T},\ Tf)}{\omega_+^2(\frac{1}{\sqrt{n}},\ \mathcal{F}_1, \mathcal{F}_2)} = \infty. \tag{15}$$

Remark: Conditions (12) and (13) are satisfied if

$$\overline{\lim_{n\to\infty}} \frac{\sup_{f\in\mathcal{F}_1} R(\hat{T}, Tf)}{\omega_+^2(\frac{1}{\sqrt{n}}, \mathcal{F}_1, \mathcal{F}_2)} = 0.$$

Proof of Theorem 1: We shall only consider the case where $\mathcal{F}_1$ and $\mathcal{F}_2$ are closed and norm bounded. The general case is proved by taking limits of this case as in Section 14 of Donoho (1994).

For $0 < \rho \le 1$, choose $f_{1,n} \in \mathcal{F}_1$ and $f_{2,n} \in \mathcal{F}_2$ such that

$$\|f_{1,n} - f_{2,n}\|_2 \le \sqrt{\frac{\rho\ln\gamma^2}{n}}$$

and such that the modulus is attained at $\{f_{1,n},\ f_{2,n}\}$:

$$|Tf_{2,n} - Tf_{1,n}| = \omega_+(\sqrt{\frac{\rho\ln\gamma^2}{n}}, \mathcal{F}_1, \mathcal{F}_2).$$

Let $\theta_1 = Tf_{1,n}$, $\theta_2 = Tf_{2,n}$ and let $\beta_n = n\|f_{1,n} - f_{2,n}\|_2^2$. Then $\beta_n \le \rho\ln\gamma^2$.

Denote by $P_{i,n}$ the probability measure associated with the white noise process

$$dY(t) = f_{i,n}(t)dt + \frac{1}{\sqrt{n}}dW(t), \quad -\frac{1}{2} \le t \le \frac{1}{2}, \ \ i = 1, 2.$$

Then a sufficient statistic for the family of measures $\{P_{i,n} : \ i = 1, \ 2\}$ is given by $S_n = \log(dP_{2,n}/dP_{1,n})$ with

$$S_n \sim \begin{cases} N(-\frac{\beta_n}{2}, \ \beta_n) & \text{under } P_{1,n} \\ N(\frac{\beta_n}{2}, \ \beta_n) & \text{under } P_{2,n}. \end{cases}$$

Denote by $\theta_1 = Tf_{1,n}$, $\theta_2 = Tf_{2,n}$, and s_{θ_i} the density of S_n under $P_{i,n}$ $(i = 1, \ 2)$. Then,

$$I(\theta_1, \ \theta_2) = \int \frac{s^2_{\theta_2}(x)}{s_{\theta_1}(x)}\, dx = e^{\beta_n} \le \gamma^{2\rho}.$$

Applying the constrained risk inequality of Brown and Low (1996b) with $I(\theta_1, \ \theta_2) \le \gamma^{2\rho}$, $\epsilon \le \gamma^{-1}\omega_+(\frac{1}{\sqrt{n}}, \mathcal{F}_1, \mathcal{F}_2)$, and $\Delta = |\theta_1 - \theta_2| = \omega_+(\sqrt{\frac{\rho \ln \gamma^2}{n}}, \ \mathcal{F}_1, \mathcal{F}_2)$, we have

$$\begin{aligned} (R(\hat{T}, \ Tf_{2,n}))^{\frac{1}{2}} \ &\ge \ \omega_+(\sqrt{\frac{\rho \ln \gamma^2}{n}}, \ \mathcal{F}_1, \mathcal{F}_2) - \gamma^{-1}\omega_+(\frac{1}{\sqrt{n}}, \ \mathcal{F}_1, \ \mathcal{F}_2)\, \gamma^{\rho} \\ &= \ \omega_+(\sqrt{\frac{\rho \ln \gamma^2}{n}}, \ \mathcal{F}_1, \mathcal{F}_2) - \gamma^{-(1-\rho)}\omega_+(\frac{1}{\sqrt{n}}, \ \mathcal{F}_1, \ \mathcal{F}_2). \end{aligned}$$

Inequality (14) follows by evaluating (11) with $\gamma = (\lambda h_n)^{-\frac{1}{2}}$ and $\rho = \frac{1}{2}$ and from the assumption that $h_n \to 0$. Inequality (15) now follows from the assumption that the modulus $\omega_+(\epsilon, \mathcal{F}_1, \mathcal{F}_2)$ is ordinary, i.e.,

$$\overline{\lim_{D\to\infty}}\ \overline{\lim_{\epsilon\to 0}} \frac{\omega_+(\epsilon, \mathcal{F}_1, \mathcal{F}_2)}{\omega_+(D\epsilon, \mathcal{F}_1, \mathcal{F}_2)} = 0.$$

This completes the proof of Theorem 1. ▮

Remark: Using a two point risk inequality in Cai, Low and Zhao (2001), Theorem 1 can be further generalized to ℓ^p-loss for all $1 \le p < \infty$.

In most applications of interest the minimax rate of convergence differs on the parameter spaces $\mathcal{F}_1$ and $\mathcal{F}_2$. If the rate is faster over $\mathcal{F}_1$ then the following theorem can be used to give a bound on the maximum risk over $\mathcal{F}_2$ assuming that the estimator is minimax rate optimal over $\mathcal{F}_1$.

Theorem 2 *Let T be a linear functional and let $\mathcal{F}_1$ and $\mathcal{F}_2$ be parameter spaces with $\mathcal{F}_1 \cap \mathcal{F}_2 \neq \emptyset$ and $\omega(\epsilon, \mathcal{F}_1) \leq \omega(\epsilon, \mathcal{F}_2)$ for all sufficiently small $0 < \epsilon \leq \epsilon_0$. Suppose that $\hat{T}$ is an estimator of Tf based on the white noise data (1) satisfying*

$$\sup_{f\in\mathcal{F}_1} R(\hat{T},\ Tf) \leq c_*^2 \omega^2(\frac{1}{\sqrt{n}}, \mathcal{F}_1) \tag{16}$$

for some constant $c_* > 0$. Let

$$\gamma_n = \max\left\{ \frac{\omega_+(\frac{1}{\sqrt{n}}, \mathcal{F}_1, \mathcal{F}_2)}{c_*\omega(\frac{1}{\sqrt{n}}, \mathcal{F}_1)},\ \ e \right\}.$$

Then for all sufficiently large n

$$\sup_{f\in\mathcal{F}_2} R(\hat{T},\ Tf) \geq c\left\{ \omega_+^2\left(\sqrt{\frac{\ln\gamma_n}{n}},\ \mathcal{F}_1,\ \mathcal{F}_2\right) + \omega^2(\frac{1}{\sqrt{n}},\ \mathcal{F}_2) \right\} \tag{17}$$

where $c > 0$ is some fixed constant.

Proof: First note that standard two point testing arguments as for example contained in Donoho and Liu (1991) or Brown and Low (1996b) show that the minimax risk for estimating a linear functional Tf over $\mathcal{F}_1$ is bounded from below by

$$\inf_{\hat{T}} \sup_{f\in\mathcal{F}_2} E(\hat{T} - Tf)^2 \geq \frac{1}{8}\omega^2\left(\frac{1}{\sqrt{n}},\ \mathcal{F}_2\right). \tag{18}$$

Let n be sufficiently large so $\frac{1}{\sqrt{n}} \leq \epsilon_0$. If $\gamma_n = e$, then

$$\omega_+(\frac{1}{\sqrt{n}}, \mathcal{F}_1, \mathcal{F}_2) \leq ec_*\omega(\frac{1}{\sqrt{n}}, \mathcal{F}_1) \leq ec_*\omega(\frac{1}{\sqrt{n}}, \mathcal{F}_2).$$

Hence

$$\sup_{f\in\mathcal{F}_2} R(\hat{T},\ Tf) \geq \frac{1}{8}\omega^2\left(\frac{1}{\sqrt{n}},\ \mathcal{F}_2\right) \geq c\left\{ \omega_+^2\left(\sqrt{\frac{\ln\gamma_n}{n}},\ \mathcal{F}_1,\ \mathcal{F}_2\right) + \omega^2(\frac{1}{\sqrt{n}},\ \mathcal{F}_2) \right\}$$

with $c = \frac{1}{8(1+ec_*)}$.

Now assume that $\gamma_n > e$. Then it follows from (16) that

$$\sup_{f\in\mathcal{F}_1} R(\hat{T},\ Tf) \leq \gamma_n^{-2}\omega_+^2(\frac{1}{\sqrt{n}},\ \mathcal{F}_1,\ \mathcal{F}_2).$$

Applying Theorem 1 with $\gamma > e$ and $\rho = \frac{1}{2}$, we have

$$\sup_{f \in \mathcal{F}_2} R(\hat{T},\ Tf) \geq (1 - e^{-\frac{1}{2}})^2 \omega_+^2 (\sqrt{\frac{\ln \gamma_n}{n}},\ \mathcal{F}_1,\ \mathcal{F}_2).$$

This and equation (18) now yields

$$\sup_{f \in \mathcal{F}_2} R(\hat{T},\ Tf) \geq \frac{1}{16} \left\{ \omega_+^2 \left(\sqrt{\frac{\ln \gamma_n}{n}},\ \mathcal{F}_1,\ \mathcal{F}_2 \right) + \omega^2 (\frac{1}{\sqrt{n}},\ \mathcal{F}_2) \right\}. \quad \blacksquare$$

In light of the lower bound given in Theorem 2 we give the following definition.

Definition 2 *We shall call an estimator* $\hat{T}$ **optimally adaptive** *over* $\mathcal{F}_1$ *and* $\mathcal{F}_2$ *if it satisfies both (16) and (17).*

Notation: For two sequences a_n and b_n, we will denote by "$a_n \asymp b_n$" if as $n \to \infty$ the ratio of a_n and b_n is bounded away from 0 and ∞. Similar, for two functions $a(\epsilon)$ and $b(\epsilon)$, "$a(\epsilon) \asymp b(\epsilon)$" means $a(\epsilon)/b(\epsilon)$ is bounded away from 0 and ∞ as $\epsilon \to 0+$. Throughout of the paper, we denote by C a generic constant that may vary from place to place.

Although Theorem 2 holds whether or not $\mathcal{F}_1$ and $\mathcal{F}_2$ are convex, it should be noted that estimators satisfying equation (16) are only guaranteed when $\mathcal{F}_1$ is convex or a union of finitely many convex parameter spaces. As mentioned in the introduction when the parameter space $\mathcal{F}$ is convex, Donoho (1994) establishes that the minimax risk $R_N^*(n, \mathcal{F}) \asymp \omega^2(\frac{1}{\sqrt{n}},\ \mathcal{F})$, as $n \to \infty$. In Cai and Low (2002) it is shown that if $\mathcal{F}$ is the union of a fixed finite number of closed and convex sets, then this result still holds. Therefore, in both these cases conditions in Theorem 2 are satisfied with $\gamma_n \to \infty$ and in these cases the following corollary which shows when fully rate adaptive estimators do not exist immediately follows from Theorem 2.

Corollary 1 *Let* T *be a linear functional and let* $\mathcal{F}_1$ *and* $\mathcal{F}_2$ *be parameter spaces with* $\mathcal{F}_1 \cap \mathcal{F}_2 \neq \emptyset$. *Suppose the minimax risks* $R_N^*(n, \mathcal{F}_i) \asymp \omega^2(\frac{1}{\sqrt{n}},\ \mathcal{F}_i)$ *for* $i = 1,\ 2$ *and suppose that the modulus* $\omega_+(\epsilon, \mathcal{F}_1, \mathcal{F}_2)$ *is ordinary. If*

$$\overline{\lim_{\epsilon \to 0}} \frac{\omega(\epsilon,\ \mathcal{F}_1)}{\omega(\epsilon,\ \mathcal{F}_2)} = 0 \quad \textit{and} \quad \omega_+(\epsilon,\ \mathcal{F}_1, \mathcal{F}_2) \geq C\omega(\epsilon,\ \mathcal{F}_2) \tag{19}$$

for some $C > 0$ *and all* $0 < \epsilon \leq \epsilon_0$, *then for any estimator* $\hat{T}$,

$$\max_{i=1,2} \overline{\lim_{n \to \infty}} \frac{\sup_{f \in \mathcal{F}_i} R(\hat{T};\ Tf)}{\omega^2(\frac{1}{\sqrt{n}},\ \mathcal{F}_i)} = \infty.$$

That is, adaptation over $\mathcal{F}_1$ *and* $\mathcal{F}_2$ *without cost is impossible.*

Condition (19) holds for a wide range of functionals and function classes of interest. In particular it holds whenever $\mathcal{F}_1$ and $\mathcal{F}_2$ are symmetric and the minimax rate of convergence over the convex parameter spaces differ. Thus Corollary 1 shows in much generality that the problem of estimating a functional Tf lacks adaptability.

Stronger statements can be made when the between class modulus is Hölderian, i.e.

$$\omega_+(\epsilon, \mathcal{F}_i, \mathcal{F}_j) = C\epsilon^{q(\mathcal{F}_i, \mathcal{F}_j)}(1 + o(1))$$

for some constant $C > 0$. Such will be the case in the examples considered in Section 3. The results in the following corollary give lower bounds for the cost of adaptation.

Corollary 2 *Let T be a linear functional and let $\mathcal{F}_1$ and $\mathcal{F}_2$ be parameter spaces with $\mathcal{F}_1 \cap \mathcal{F}_2 \neq \emptyset$. Suppose the minimax risks for estimating Tf over the function classes $\mathcal{F}_1$ and $\mathcal{F}_2$ satisfy $R_N^*(n, \mathcal{F}_i) \asymp n^{-r_i}$ for $i = 1,\ 2$. Suppose the between class modulus of continuity $\omega_+(\epsilon, \mathcal{F}_1, \mathcal{F}_2)$ is Hölderian with exponent $q = q(\mathcal{F}_1, \mathcal{F}_2)$.*

If $q = r_2 < r_1$ or $q < r_2 \leq r_1$, then for any estimator $\hat{T}$ attaining a rate of convergence n^r over $\mathcal{F}_1$ with $r > q$,

$$\overline{\lim_{n\to\infty}} \left(\frac{n}{\log n}\right)^q \sup_{f\in\mathcal{F}_2} R(\hat{T},\ Tf) > 0. \tag{20}$$

Proof: Let $d_n = n^{-r}$ and $h_n = n^{q-r}$. Then (20) follows from (14) in Theorem 1 and the condition that $\omega_+(\epsilon,\ \mathcal{F}_1,\ \mathcal{F}_2) \asymp \epsilon^q$. ∎

Remark: In the first case when $q = r_2 < r_1$ then (20) shows that the maximum risk over $\mathcal{F}_2$ is elevated by at least a logarithmic factor. In the second case when $q < r_2 \leq r_1$ the maximum risk over $\mathcal{F}_2$ must be increased by at least a power of the sample size.

3 Examples

Section 2 gave lower bounds on the cost of adaptation based on the between class modulus. In this section we give examples of these lower bounds. In Section 4 we show that the lower bound is in fact rate sharp.

As mentioned in Section 2 in most common cases when estimating a linear functional over convex parameter spaces the modulus is Hölderian

$$\omega_+(\epsilon,\ \mathcal{F}_i,\ \mathcal{F}_j) = C_{i,j}\epsilon^{q(\mathcal{F}_i,\mathcal{F}_j)}(1 + o(1)) \tag{21}$$

where we write $q(\mathcal{F}_i)$ for $q(\mathcal{F}_i, \mathcal{F}_i)$. In such cases the most easily computed bounds in Section 2 are given in Corollary 2 which can then be used to classify the problem of adaptation over convex parameter spaces into three cases:

- **Case 1:** $q(\mathcal{F}_1, \mathcal{F}_2) = \min(q(\mathcal{F}_1), (\mathcal{F}_2)) < \max(q(\mathcal{F}_1), q(\mathcal{F}_2))$. This is the "regular case" which holds for many linear functionals and common function classes of interest. In this case, one must lose a logarithmic factor as the minimum cost for adaptation. A common example of such a case is estimating a function or a derivative at a point, i.e., $Tf = f^{(s)}(t_0)$ for some $s \geq 0$ when the parameter spaces are assumed to be Lipschitz. See Example 2 below, Lepski (1990), Brown and Low (1996b) and Efromovich and Low (1994).

  Besides the regular case, there are two extreme cases.

- **Case 2:** $q(\mathcal{F}_1, \mathcal{F}_2) > \min(q(\mathcal{F}_1), q(\mathcal{F}_2))$ or $q(\mathcal{F}_1, \mathcal{F}_2) = q(\mathcal{F}_1) = q(\mathcal{F}_2)$. This is a case which is not covered in Corollary 2. We shall show in Section 4 that in this case adaptation for free is always possible. That is, one can attain the optimal rate of convergence over $\mathcal{F}_1$ and $\mathcal{F}_2$ simultaneously. An example of this case is estimating a function at a point over two monotone Lipschitz classes. See Examples 1 and 3 below and Kang and Low (2002).

- **Case 3:** $q(\mathcal{F}_1, \mathcal{F}_2) < \min(q(\mathcal{F}_1), q(\mathcal{F}_2))$. In this case the cost of adaptation is significant, much more than the usual logarithmic penalty in the regular case. If f is known to be in $\mathcal{F}_1$, one can attain the rate of $n^{q(\mathcal{F}_1)}$; and if one knows that f is in $\mathcal{F}_2$, the rate of convergence $n^{q(\mathcal{F}_2)}$ can be achieved. Without the information, however, one can only achieve the rate of $(n/\log n)^{q(\mathcal{F}_1, \mathcal{F}_2)}$ at best. So the cost of adaptation is a power of n rather than the logarithmic factor as in the regular case. See Example 2 below.

Note that if the the parameter spaces $\mathcal{F}_1$ and $\mathcal{F}_2$ are nested, then only Cases 1 and 2 are possible and Case 3 does not arise.

We now consider a few examples below to illustrate the three different cases. Examples 1 and 3 covers Case 2 in which full adaptation is possible. Example 2 covers both Case 1 and Case 3 with different choices of parameters. In each of these examples we need to calculate the between class modulus of continuity. The basic idea behind these calculations is contained in Donoho and Liu (1991) and consists of finding extremal functions. For the calculation of extremal functions it is useful to introduce the inverse $\tilde{\omega}_+(a, \mathcal{F}_1, \mathcal{F}_2)$ of the ordered modulus

$$\tilde{\omega}(a, \mathcal{F}_1, \mathcal{F}_2) = \inf\{\|g - f\|_2 : \ |Tg - Tf| = a, \ f \in \mathcal{F}_1, g \in \mathcal{F}_2\}. \tag{22}$$

We shall also write $\tilde{\omega}(a, \mathcal{F})$ for $\tilde{\omega}(a, \mathcal{F}, \mathcal{F})$.

Example 1: In this example we shall have $0 < q(\mathcal{F}_2) < q(\mathcal{F}_1, \mathcal{F}_2) = q(\mathcal{F}_1) < 1$ and $\omega(\epsilon, \ \mathcal{F}_1, \mathcal{F}_2) = \omega(\epsilon, \ \mathcal{F}_2, \mathcal{F}_1)$. In this case full mean squared error adaptation is possible.

For $0 < \alpha \leq 1$, let $F(\alpha, M)$ be the collection of functions $f : [-\frac{1}{2}, \frac{1}{2}] \rightarrow I\!R$ such that

$$|f(x) - f(y)| \leq M|x - y|^{\alpha}.$$

Let D be the set of all decreasing functions and let $F_D(\alpha, M) = F(\alpha, M) \cap D$ be the set of decreasing functions which are also members of $F(\alpha, M)$. Let $Tf = f(0)$ and assume that $0 < \alpha_2 < \alpha_1 \leq 1$. Let $\mathcal{F}_1 = F_D(\alpha_1, M_1)$ and $\mathcal{F}_2 = F_D(\alpha_2, M_2)$. Then for these parameter spaces and the linear functional $Tf = f(0)$

$$f_1(x) = \begin{cases} \min(a, M_1|x|^{\alpha_1}), & \text{when } x \leq 0 \\ a, & \text{when } x \geq 0 \end{cases}$$

and

$$f_2(x) = \begin{cases} a, & \text{when } x \leq 0 \\ (a - M_2 x^{\alpha_2})_+, & \text{when } x \geq 0 \end{cases}$$

are extremal and it follows that for sufficiently small $a > 0$

$$\tilde{\omega}^2(a, \mathcal{F}_1, \mathcal{F}_2) = \frac{a^{\frac{2\alpha_1+1}{\alpha_1}}}{M_1^{\frac{1}{\alpha_1}}(2\alpha_1 + 1)} + \frac{a^{\frac{2\alpha_2+1}{\alpha_2}}}{M_2^{\frac{1}{\alpha_2}}(2\alpha_2 + 1)}$$

and hence as $\epsilon \rightarrow 0$

$$\omega^2(\epsilon, \ \mathcal{F}_1, \mathcal{F}_2) = (2\alpha_1 + 1)^{\frac{\alpha_1}{2\alpha_1+1}} M_1^{\frac{1}{2\alpha_1+1}} \epsilon^{\frac{2\alpha_1}{2\alpha_1+1}} (1 + o(1)). \tag{23}$$

Similar arguments yield $\omega(\epsilon, \mathcal{F}_2, \mathcal{F}_1) = \omega(\epsilon, \mathcal{F}_1, \mathcal{F}_2)$ and

$$\omega^2(\epsilon, \ \mathcal{F}_1) = (\alpha_1 + \frac{1}{2})^{\frac{\alpha_1}{2\alpha_1+1}} M_1^{\frac{1}{2\alpha_1+1}} \epsilon^{\frac{2\alpha_1}{2\alpha_1+1}} (1 + o(1)) \tag{24}$$

and

$$\omega^2(\epsilon, \ \mathcal{F}_2) = (\alpha_2 + \frac{1}{2})^{\frac{\alpha_2}{2\alpha_2+1}} M_2^{\frac{1}{2\alpha_2+1}} \epsilon^{\frac{2\alpha_2}{2\alpha_2+1}} (1 + o(1)). \tag{25}$$

In this case $q(\mathcal{F}_1, \mathcal{F}_2) = \max(q(\mathcal{F}_1), q(\mathcal{F}_2)) > \min(q(\mathcal{F}_1), q(\mathcal{F}_2))$ and hence adaptation for free can be achieved.

Example 2: This example shows that sometimes we must lose more than a logarithmic factor when we try to adapt.

Let $F^R(\alpha, M)$ be the collection of functions $f : [-\frac{1}{2}, \frac{1}{2}] \to I\!R$ such that

$$|f^{(s)}(x) - f^{(s)}(y)| \leq M|x - y|^{\alpha - s} \quad 0 \leq x \leq y \leq \frac{1}{2}$$

where s is the largest integer less than α. Similarly, let $F^L(\alpha, M)$ be the collection of functions $f : [-\frac{1}{2}, \frac{1}{2}] \to I\!R$ such that

$$|f^{(s)}(x) - f^{(s)}(y)| \leq M|x - y|^{\alpha - s} \quad -\frac{1}{2} \leq x \leq y \leq 0.$$

Finally let $F(\alpha_1, M_1, \alpha_2, M_2) = F^L(\alpha_1, M_1) \cap F^R(\alpha_2, M_2)$.

Note that for the linear functional $Tf = f(0)$ and the (ordered) parameter spaces $\mathcal{F}_1 = F(\alpha_1, M_1, \alpha_2, M_2)$ and $\mathcal{F}_2 = F(\beta_1, N_1, \beta_2, N_2)$ the functions

$$f_1(x) = \begin{cases} c_1, & \text{when } x \leq b_1 \\ M_1|x|^{\alpha_1}, & \text{when } b_1 \leq x \leq 0 \\ M_2|x|^{\alpha_2}, & \text{when } 0 \leq x \leq b_2 \\ c_2, & \text{when } x \geq b_2 \end{cases}$$

and

$$f_2(x) = \begin{cases} c_1, & \text{when } x \leq b_1 \\ a - N_1|x|^{\beta_1}, & \text{when } b_1 \leq x \leq 0 \\ a - N_2 x^{\beta_2}, & \text{when } 0 \leq x \leq b_2 \\ c_2, & \text{when } x \geq b_2 \end{cases}$$

are extremal for sufficiently small $a > 0$ where $b_1 < 0$ and $b_2 > 0$ are determined by the constraints

$$a - N_1|b_1|^{\beta_1} = M_1|b_1|^{\alpha_1} \quad \text{and} \quad a - N_2 b_2^{\beta_2} = M_2 b_1^{\alpha_2},$$

and c_1 and c_2 are constants chosen to make the functions continuous.

It is then easy to check that

$$\omega^2(\epsilon,\ \mathcal{F}_1) = C(\alpha_1, M_1, \alpha_2, M_2)\epsilon^{\frac{2\delta}{2\delta+1}}(1 + o(1)) \tag{26}$$

where $\delta = \max(\alpha_1, \alpha_2)$. And similarly

$$\omega^2(\epsilon,\ \mathcal{F}_2) = C(\beta_1, N_1, \beta_2, N_2)\epsilon^{\frac{2\rho}{2\rho+1}}(1 + o(1)) \tag{27}$$

where $\rho = \max(\beta_1, \beta_2)$.

We now assume that $0 < \alpha_2 \le \alpha_1 \le 1$ and $0 < \beta_1 \le \beta_2 \le 1$. Then

$$q(\mathcal{F}_1) = \frac{2\alpha_1}{2\alpha_1 + 1} \quad \text{and} \quad q(\mathcal{F}_2) = \frac{2\beta_2}{2\beta_2 + 1}.$$

The extremal functions given above also lead to the between class modulus

$$\omega^2(\epsilon, \mathcal{F}_1, \mathcal{F}_2) = C(M_1, \alpha_1, M_2, \alpha_2, N_1, \beta_1, N_2, \beta_2)\epsilon^{\frac{2\gamma}{2\gamma+1}}(1 + o(1)) \tag{28}$$

where $\gamma = \max(\min(\alpha_1, \beta_1), \min(\alpha_2, \beta_2))$.

Two interesting cases may arise, depending on the relationship among α_1, α_2, β_1 and β_2.

- $\beta_2 > \beta_1 \ge \alpha_1 > \alpha_2$. Then the quantity γ in equation (28) is $\gamma = \alpha_1$ and so $q(\mathcal{F}_1, \mathcal{F}_2) = \frac{2\alpha_1}{2\alpha_1+1}$. Hence in this case

  $$q(\mathcal{F}_1, \mathcal{F}_2) = \min(q(\mathcal{F}_1), q(\mathcal{F}_2)) < \max(q(\mathcal{F}_1), q(\mathcal{F}_2)).$$

  This is a standard case where a logarithmic penalty term must be paid for adaptation.

- $\alpha_1 \ge \beta_2 > \beta_1 \ge \alpha_2$. In this case, the quantity γ in equation (28) is $\gamma = \beta_1$ and hence $q(\mathcal{F}_1, \mathcal{F}_2) = \frac{2\beta_1}{2\beta_1+1}$. Therefore in this case

  $$q(\mathcal{F}_1, \mathcal{F}_2) < \min(q(\mathcal{F}_1), q(\mathcal{F}_2)).$$

  Consequently the cost of adaption between $\mathcal{F}_1$ and $\mathcal{F}_2$ is much more than a logarithmic penalty. The maximum risk over the two different parameter spaces is of the order $n^{-\frac{2\beta_1}{2\beta_1+1}}$.

  A particularly interesting case is when $\alpha_1 = \beta_2 > \beta_1 \ge \alpha_2$. In this case, the minimax rates of convergence over $\mathcal{F}_1$ and $\mathcal{F}_2$ are the same, both equals $n^{-\frac{2\beta_2}{2\beta_2+1}}$. Yet it is impossible to achieve this optimal rate adaptively over the two parameter spaces, in fact the cost of adaption in this case is substantial.

Example 3: This will give an example where $0 < q(\mathcal{F}_1) < q(\mathcal{F}_1, \mathcal{F}_2) < q(\mathcal{F}_2) < 1$. It will also yield an example where $\omega(\epsilon,\ \mathcal{F}_1, \mathcal{F}_2) \ne \omega(\epsilon,\ \mathcal{F}_2, \mathcal{F}_1)$. In this case full mean squared error adaptation can be achieved. Let $Tf = f(0)$. Now let

$$F_D(\alpha_1, M_1, \alpha_2, M_2) = F(\alpha_1, M_1, \alpha_2, M_2) \cap D$$

where $F(\alpha_1, M_1, \alpha_2, M_2)$ is defined as in Example 2.

It is easy to see that for the (ordered) parameter spaces $\mathcal{F}_1 = F_D(\alpha_1, M_1, \alpha_2, M_2)$ and $\mathcal{F}_2 = F_D(\beta_1, N_1, \beta_2, N_2)$ the extremal functions are

$$f_1(x) = \begin{cases} \min(a, M_1|x|^{\alpha_1}), & \text{when } x \le 0 \\ a, & \text{when } x \ge 0 \end{cases} \tag{29}$$

and

$$f_2(x) = \begin{cases} a, & \text{when } x \le 0 \\ (a - N_2 x^{\beta_2})_+, & \text{when } x \ge 0 \end{cases} \tag{30}$$

and it follows as in the first example that

$$\tilde{\omega}^2(a, \mathcal{F}_1, \mathcal{F}_2) = \frac{a^{\frac{2\alpha_1+1}{\alpha_1}}}{M_1^{\frac{1}{\alpha_1}}(2\alpha_1+1)} + \frac{a^{\frac{2\beta_2+1}{\beta_2}}}{N_2^{\frac{1}{\beta_2}}(2\beta_2+1)}.$$

For now let $\beta_1 > \beta_2 > \alpha_1 > \alpha_2$. Then it is easy to check that

$$\omega^2(\epsilon,\ \mathcal{F}_1) = C\epsilon^{\frac{2\alpha_1}{2\alpha_1+1}}(1 + o(1))$$

and

$$\omega^2(\epsilon,\ \mathcal{F}_2) = C\epsilon^{\frac{2\beta_1}{2\beta_1+1}}(1 + o(1))$$

and also

$$\omega^2(\epsilon, \mathcal{F}_1, \mathcal{F}_2) = C\epsilon^{\frac{2\beta_2}{2\beta_2+1}}(1 + o(1)). \tag{31}$$

Now extremal functions for the ordered parameter spaces $\mathcal{F}_2$ and $\mathcal{F}_1$ analogous to (29) and (30) also easily yield

$$\omega^2(\epsilon, \mathcal{F}_2, \mathcal{F}_1) = C\epsilon^{\frac{2\beta_1}{2\beta_1+1}}(1 + o(1)). \tag{32}$$

Hence this is an example where $\omega(\epsilon, \mathcal{F}_1, \mathcal{F}_2) \ne \omega(\epsilon, \mathcal{F}_2, \mathcal{F}_1)(1+o(1))$. Note that $\beta_1 > \beta_2$, it then follows from (31) and (32) that $q(\mathcal{F}_1, \mathcal{F}_2) = \frac{2\beta_2}{2\beta_2+1}$. Hence this is an example where

$$0 < q(\mathcal{F}_1) < q(\mathcal{F}_1, \mathcal{F}_2) < q(\mathcal{F}_2) < 1.$$

In particular, $q(\mathcal{F}_1, \mathcal{F}_2) > \min(q(\mathcal{F}_1),\ q(\mathcal{F}_2))$ so it is also an example where full mean squared error adaptation is possible.

4 Adaptive estimators and upper bound for the cost of adaptation

In this section we consider adaptation over two parameter spaces which are not necessarily nested. The results in this section also shows that the lower bound on the cost of adaptation given in Theorem 2 is rate sharp.

Let $\mathcal{F}_1$ and $\mathcal{F}_2$ be two closed, convex parameter spaces with $\mathcal{F}_1 \cap \mathcal{F}_2 \neq \emptyset$. Denote by $\mathcal{G} = \mathcal{F}_1 \cup \mathcal{F}_2$. In the case of non-nested convex parameter spaces adaptive estimation theory requires a minimax analysis for sets which are not convex. The reason is that we need to know the minimax risk and minimax rate optimal procedure over the union $\mathcal{G} = \mathcal{F}_1 \cup \mathcal{F}_2$, which is in general nonconvex. The extension of the minimax theory for estimating linear functionals over nonconvex parameter spaces has been considered in Cai and Low (2002). In particular, it is shown that if $\mathcal{G}$ is a union of a finite number of closed convex parameter spaces, the minimax risk is still of the order $\omega^2(\frac{1}{\sqrt{n}}, \mathcal{G})$. However, in general rate optimal estimators are necessarily nonlinear. Explicit rate optimal procedures are constructed.

The following theorem is from Cai and Low (2002).

Theorem 3 *Suppose the white noise model (1) is observed. Let $\mathcal{G} = \cup_{i=1}^{k} \mathcal{F}_i$ where $\mathcal{F}_i$ are closed convex parameter spaces with nonempty intersections and k is fixed. The minimax risk for estimating the linear functional Tf over $\mathcal{G}$ satisfies*

$$C_1 \omega^2\left(\frac{1}{\sqrt{n}},\ \mathcal{G}\right) \leq \inf_{\hat{T}} \sup_{f \in \mathcal{G}} E(\hat{T} - Tf)^2 \leq C_2 \omega^2\left(\frac{1}{\sqrt{n}},\ \mathcal{G}\right) \tag{33}$$

for some constants $0 < C_1 \leq C_2 < \infty$.

Furthermore, there exists a rate optimal estimator $\hat{T}^*$ of the linear functional Tf satisfying

$$\sup_{f \in \mathcal{G}} E|\hat{T}^* - Tf|^p \leq C(k,p)\omega^p\left(\frac{1}{\sqrt{n}}, \mathcal{G}\right) \tag{34}$$

for all $1 \leq p < \infty$, where the constant $C(k,p)$ depends only on k and p.

In Section 2 we briefly introduced an ordered modulus of continuity although the bounds given in that section are based on a between class modulus. In the construction of adaptive estimators it is more convenient to work directly with the ordered modulus. In Cai and Low (2002) it was shown that for any linear functional the ordered modulus of continuity can be used to give a linear procedure which has upper bounds for the bias over one parameter space and lower bounds for the bias over the other parameter space. More specifically, for two convex sets $\mathcal{F}$ and $\mathcal{H}$ with $\mathcal{F} \cap \mathcal{H} \neq \emptyset$, it was shown how to construct a linear estimator $\hat{T}$ which has variance and bias satisfying

$$\begin{aligned} \mathrm{Var}(\hat{T}) &= E(\hat{T} - E\hat{T})^2 \leq V, & (35) \\ \sup_{f \in \mathcal{F}} (E\hat{T} - Tf) &\leq \frac{1}{2} \sup_{\epsilon > 0} \left(\omega(\epsilon,\ \mathcal{F}, \mathcal{H}) - \sqrt{nV}\,\epsilon \right) & (36) \end{aligned}$$

and

$$\inf_{f \in \mathcal{H}} (E\hat{T} - Tf) \geq -\frac{1}{2} \sup_{\epsilon > 0} \left(\omega(\epsilon,\ \mathcal{F}, \mathcal{H}) - \sqrt{nV}\,\epsilon \right). \qquad (37)$$

Such linear estimators are useful because they trade bias and variance in a precise manner. See Cai and Low (2002) for details of the construction.

4.1 Construction of the adaptive procedure

In this section we shall use the bias and variance properties given in (35) - (37) to construct a test between $\mathcal{F}_1$ and $\mathcal{F}_2$ which will be used in the construction of our adaptive estimator.

Let $\mathcal{F}_1$ and $\mathcal{F}_2$ be convex parameter spaces with $\mathcal{F}_1 \cap \mathcal{F}_2 \neq \emptyset$ and for $i = 1$ and 2, let $\hat{T}_i$ be linear estimators satisfying

$$\sup_{f \in \mathcal{F}_i} E(\hat{T}_i - Tf)^2 \leq \omega^2(\frac{1}{\sqrt{n}}, \mathcal{F}_i).$$

By (3) these estimators exist and are minimax rate optimal over $\mathcal{F}_i$. In our problem here, $\mathcal{G} = \mathcal{F}_1 \cup \mathcal{F}_2$ is the union of two closed convex sets. Denote by $\hat{T}_2^*$ the estimator given in Theorem 3 satisfying (34). In particular, $\hat{T}_2^*$ satisfies

$$\sup_{f \in \mathcal{G}} E(\hat{T}_2^* - Tf)^2 \leq C\omega^2(\frac{1}{\sqrt{n}}, \mathcal{G}) \qquad (38)$$

and

$$\sup_{f \in \mathcal{G}} E|\hat{T}_2^* - Tf|^4 \leq C\omega^4\left(\frac{1}{\sqrt{n}},\ \mathcal{G}\right). \qquad (39)$$

We will now construct two linear estimators $\hat{T}_{1,2}$ and $\hat{T}_{2,1}$ which have bias and variance properties over $\mathcal{F}_1$ and $\mathcal{F}_2$ which allow for the construction of a test tailored to the construction of an adaptive estimator over $\mathcal{F}_1$ and $\mathcal{F}_2$. For $1 \le i \ne j \le 2$ let

$$\gamma_{i,j} = 1 + \frac{\omega(\frac{1}{\sqrt{n}}, \mathcal{F}_i, \mathcal{F}_j)}{\omega(\frac{1}{\sqrt{n}}, \mathcal{F}_1)} \tag{40}$$

and

$$\gamma_+ = \max(\gamma_{1,2},\ \gamma_{2,1}) = 1 + \frac{\omega_+(\frac{1}{\sqrt{n}}, \mathcal{F}_1, \mathcal{F}_2)}{\omega(\frac{1}{\sqrt{n}}, \mathcal{F}_1)}. \tag{41}$$

Note that

$$\gamma_{1,2} \ge \frac{\omega(\frac{1}{\sqrt{n}}, \mathcal{F}_1, \mathcal{G})}{\omega(\frac{1}{\sqrt{n}}, \mathcal{F}_1)}, \quad \gamma_{2,1} \ge \frac{\omega(\frac{1}{\sqrt{n}}, \mathcal{G}, \mathcal{F}_1)}{\omega(\frac{1}{\sqrt{n}}, \mathcal{F}_1)} \quad \text{and} \quad \gamma_+ \ge \frac{\omega_+(\frac{1}{\sqrt{n}}, \mathcal{F}_1, \mathcal{G})}{\omega(\frac{1}{\sqrt{n}}, \mathcal{F}_1)}.$$

For $i \ne j$ let

$$V_{i,j} = \frac{1}{\ln \gamma_{i,j}} \omega^2(\sqrt{\frac{\ln \gamma_{i,j}}{n}}, \mathcal{F}_i, \mathcal{F}_j).$$

Then for $i \ne j$

$$\begin{aligned} B_{i,j} &= \frac{1}{2} \sup_{\epsilon > 0} (\omega(\epsilon, \mathcal{F}_i, \mathcal{F}_j) - \epsilon \sqrt{n V_{i,j}}) \\ &= \frac{1}{2} \sup_{\epsilon \le \sqrt{\frac{\ln \gamma_{i,j}}{n}}} (\omega(\epsilon,\ \mathcal{F}_i, \mathcal{F}_j) - \epsilon \sqrt{\frac{n}{\ln \gamma_{i,j}}} \omega(\sqrt{\frac{\ln \gamma_{i,j}}{n}}, \mathcal{F}_i, \mathcal{F}_j)) \\ &\le \frac{1}{2} \omega(\sqrt{\frac{\ln \gamma_{i,j}}{n}}, \mathcal{F}_i, \mathcal{F}_j). \end{aligned}$$

For $i \ne j$ let $\hat{T}_{i,j}$ be the linear estimator satisfying (35) - (37) with $\mathcal{F} = \mathcal{F}_i$, $\mathcal{H} = \mathcal{F}_j$ and $V = V_{i,j}$. Then if $f \in \mathcal{F}_1$,

$$\begin{aligned} E(\hat{T}_1 - \hat{T}_{1,2}) &= E(\hat{T}_1 - Tf) - E(\hat{T}_{1,2} - Tf) \\ &\ge -\omega(\frac{1}{\sqrt{n}}, \mathcal{F}_1) - \omega(\sqrt{\frac{\ln \gamma_{1,2}}{n}}, \mathcal{F}_1, \mathcal{F}_2) \\ &= -b_{1,2} \end{aligned} \tag{42}$$

and

$$\begin{aligned} E(\hat{T}_1 - \hat{T}_{2,1}) &= E(\hat{T}_1 - Tf) - E(\hat{T}_{2,1} - Tf) \\ &\le \omega(\frac{1}{\sqrt{n}}, \mathcal{F}_1) + \omega(\sqrt{\frac{\ln \gamma_{2,1}}{n}}, \mathcal{F}_2, \mathcal{F}_1) \\ &= b_{2,1}. \end{aligned} \tag{43}$$

Note also that

$$\begin{aligned} Var(\hat{T}_1 - \hat{T}_{1,2}) &\leq 2(\omega^2(\frac{1}{\sqrt{n}}, \mathcal{F}_1) + \frac{1}{\ln \gamma_{1,2}}\omega^2(\sqrt{\frac{\ln \gamma_{1,2}}{n}}, \mathcal{F}_1, \mathcal{F}_2)) \\ &= v_{1,2} \end{aligned} \tag{44}$$

and

$$\begin{aligned} Var(\hat{T}_1 - \hat{T}_{2,1}) &\leq 2(\omega^2(\frac{1}{\sqrt{n}}, \mathcal{F}_1) + \frac{1}{\ln \gamma_{2,1}}\omega^2(\sqrt{\frac{\ln \gamma_{2,1}}{n}}, \mathcal{F}_2, \mathcal{F}_1)) \\ &= v_{2,1}. \end{aligned} \tag{45}$$

The estimators $\hat{T}_1$, $\hat{T}_{1,2}$ and $\hat{T}_{2,1}$ can be used as a test between $\mathcal{F}_1$ and $\mathcal{F}_2$. Let

$$I_n = 1(\hat{T}_{1,2} - 5b_{1,2} - 3\omega(\frac{1}{\sqrt{n}}, \mathcal{G}) \leq \hat{T}_1 \leq \hat{T}_{2,1} + 5b_{2,1} + 3\omega(\frac{1}{\sqrt{n}}, \mathcal{G}))$$

Finally let

$$\hat{T} = I_n \hat{T}_1 + (1 - I_n)\hat{T}_2^* \tag{46}$$

Remark : The estimator $\hat{T}$ uses $\hat{T}_1$, which is rate optimal over $\mathcal{F}_1$, when the test I_n is in favor of $\mathcal{F}_1$; the estimator $\hat{T}$ uses $\hat{T}_2^*$, which is rate optimal over $\mathcal{G} = \mathcal{F}_1 \cup \mathcal{F}_2$, when the test I_n is in favor of $\mathcal{F}_2$. It is important in this case to use $\hat{T}_2^*$ instead of a rate optimal estimator over $\mathcal{F}_2$.

4.2 Adaptivity of the procedure

Theorem 4 below shows that the estimator $\hat{T}^*$ constructed above is adaptively rate optimal and that the lower bound for adaptation between $\mathcal{F}_1$ and $\mathcal{F}_2$ as given in Theorem 2 is sharp.

Theorem 4 *Suppose $\mathcal{F}_1$ and $\mathcal{F}_2$ are two closed convex parameter spaces with $\mathcal{F}_1 \cap \mathcal{F}_2 \neq \emptyset$ and $\omega(\epsilon, \mathcal{F}_1) \leq \omega(\epsilon, \mathcal{F}_2)$. The estimator $\hat{T}$ defined in (46) is adaptively rate optimal over $\mathcal{F}_1$ and $\mathcal{F}_2$. That is, $\hat{T}$ attains the exact minimax rate of convergence over $\mathcal{F}_1$ and attains the lower bound on adaptation over $\mathcal{F}_2$ as given in Theorem 2. Hence,*

$$\sup_{f \in \mathcal{F}_1} E(\hat{T} - Tf)^2 \leq C\omega^2(\frac{1}{\sqrt{n}}, \mathcal{F}_1) \tag{47}$$

and

$$\sup_{f \in \mathcal{F}_2} E(\hat{T} - Tf)^2 \leq C\left\{\omega_+^2(\sqrt{\frac{\ln \gamma_+}{n}}, \mathcal{F}_1, \mathcal{F}_2) + \omega^2(\frac{1}{\sqrt{n}}, \mathcal{F}_2)\right\} \tag{48}$$

where γ_+ is defined in equation (41).

Remark: The estimator $\hat{T}$ defined in (46) is also adaptive between $\mathcal{F}_1$ and $\mathcal{G} = \mathcal{F}_1 \cup \mathcal{F}_2$. Note that (48) is equivalent to

$$\sup_{f \in \mathcal{G}} E(\hat{T} - Tf)^2 \le C \left\{ \omega_+^2(\sqrt{\frac{\ln \gamma_+^*}{n}}, \mathcal{F}_1, \mathcal{G}) + \omega^2(\frac{1}{\sqrt{n}}, \mathcal{G}) \right\}. \tag{49}$$

where

$$\gamma_+^* = 1 + \frac{\omega_+(\frac{1}{\sqrt{n}}, \mathcal{F}_1, \mathcal{G})}{\omega(\frac{1}{\sqrt{n}}, \mathcal{F}_1)}. \tag{50}$$

Therefore $\hat{T}$ attains the exact minimax rate of convergence over $\mathcal{F}_1$ and attains the lower bound on adaptation over $\mathcal{G}$ as given in Theorem 2.

The proof of Theorem 4 is facilitated by the following lemmas.

Lemma 1 *If $f \in \mathcal{F}_1$, then*

$$(P(I_n = 0))^{\frac{1}{2}} \le \frac{\omega^2(\frac{1}{\sqrt{n}}, \mathcal{F}_1)}{\omega^2(\frac{1}{\sqrt{n}}, \mathcal{G})}. \tag{51}$$

Proof: First note that for a standard normal random variable Z,

$$P(Z \ge \lambda) \le \exp(-\frac{\lambda^2}{2})$$

holds for all $\lambda \ge 0$. It then follows from (42) - (45) that

$$\begin{aligned} P(I_n = 0) &\le P(\hat{T}_1 - \hat{T}_{1,2} \le -5b_{1,2} - 3\omega(\frac{1}{\sqrt{n}}, \mathcal{G})) \\ &\quad + P(\hat{T}_1 - \hat{T}_{2,1} \ge 5b_{2,1} + 3\omega(\frac{1}{\sqrt{n}}, \mathcal{G})) \\ &\le \exp\left(-\frac{(4b_{1,2} + 3\omega(\frac{1}{\sqrt{n}}, \mathcal{G}))^2}{2v_{1,2}}\right) + \exp\left(-\frac{(4b_{2,1} + 3\omega(\frac{1}{\sqrt{n}}, \mathcal{G}))^2}{2v_{2,1}}\right) \end{aligned}$$

If $\omega^2(\frac{1}{\sqrt{n}}, \mathcal{F}_1) \ge \frac{1}{\ln \gamma_{1,2}} \omega^2(\sqrt{\frac{\ln \gamma_{1,2}}{n}}, \mathcal{F}_1, \mathcal{F}_2)$, noting that $e^{-2x} > \frac{1}{2}x^{-2}$ for $x > 0$, then

$$\exp\left(-\frac{(4b_{1,2} + 3\omega(\frac{1}{\sqrt{n}}, \mathcal{G}))^2}{2v_{1,2}}\right) \le \exp\left(-2\frac{\omega^2(\frac{1}{\sqrt{n}}, \mathcal{G})}{\omega^2(\frac{1}{\sqrt{n}}, \mathcal{F}_1)}\right) \le \frac{1}{2}\frac{\omega^4(\frac{1}{\sqrt{n}}, \mathcal{F}_1)}{\omega^4(\frac{1}{\sqrt{n}}, \mathcal{G})}. \tag{52}$$

If $\omega^2(\frac{1}{\sqrt{n}}, \mathcal{F}_1) < \frac{1}{\ln \gamma_{1,2}} \omega^2(\sqrt{\frac{\ln \gamma_{1,2}}{n}}, \mathcal{F}_1, \mathcal{F}_2)$, then

$$\begin{aligned}
&\exp\left(-\frac{(4b_{1,2} + 3\omega(\frac{1}{\sqrt{n}}, \mathcal{G}))^2}{2v_{1,2}}\right) \le \exp\left(-\frac{16\omega^2(\sqrt{\frac{\ln \gamma_{1,2}}{n}}, \mathcal{F}_1, \mathcal{F}_2) + 9\omega^2(\frac{1}{\sqrt{n}}, \mathcal{G})}{\frac{4}{\ln \gamma_{1,2}} \omega^2(\sqrt{\frac{\ln \gamma_{1,2}}{n}}, \mathcal{F}_1, \mathcal{F}_2)}\right) \\
&\le \ \exp\left(-(4\ln \gamma_{1,2} + 2\frac{\omega^2(\frac{1}{\sqrt{n}}, \mathcal{G})}{\omega^2(\frac{1}{\sqrt{n}}, \mathcal{F}_1, \mathcal{F}_2)})\right) \le \ \exp\left(-(4\ln \gamma_{1,2} + 2\frac{\omega^2(\frac{1}{\sqrt{n}}, \mathcal{G})}{\omega^2(\frac{1}{\sqrt{n}}, \mathcal{F}_1, \mathcal{G})})\right) \\
&\le \ \frac{\omega^4(\frac{1}{\sqrt{n}}, \mathcal{F}_1)}{\omega^4(\frac{1}{\sqrt{n}}, \mathcal{F}_1, \mathcal{G})} \cdot \frac{1}{2} \frac{\omega^4(\frac{1}{\sqrt{n}}, \mathcal{F}_1, \mathcal{G})}{\omega^4(\frac{1}{\sqrt{n}}, \mathcal{G})} = \frac{1}{2} \frac{\omega^4(\frac{1}{\sqrt{n}}, \mathcal{F}_1)}{\omega^4(\frac{1}{\sqrt{n}}, \mathcal{G})}. \qquad (53)
\end{aligned}$$

Combining (52) and (53) yields that

$$\exp\left(-\frac{(4b_{1,2} + 3\omega(\frac{1}{\sqrt{n}}, \mathcal{G}))^2}{2v_{1,2}}\right) \le \frac{1}{2} \frac{\omega^4(\frac{1}{\sqrt{n}}, \mathcal{F}_1)}{\omega^4(\frac{1}{\sqrt{n}}, \mathcal{G})}.$$

Similarly,

$$\exp\left(-\frac{(4b_{2,1} + 3\omega(\frac{1}{\sqrt{n}}, \mathcal{G}))^2}{2v_{2,1}}\right) \le \frac{1}{2} \frac{\omega^4(\frac{1}{\sqrt{n}}, \mathcal{F}_1)}{\omega^4(\frac{1}{\sqrt{n}}, \mathcal{G})}.$$

Therefore,

$$(P(I_n = 0))^{\frac{1}{2}} \le \frac{\omega^2(\frac{1}{\sqrt{n}}, \mathcal{F}_1)}{\omega^2(\frac{1}{\sqrt{n}}, \mathcal{G})}. \quad \blacksquare$$

Now consider the case $f \in \mathcal{F}_2$.

Lemma 2 *If $f \in \mathcal{F}_2$ and $|E\hat{T}_1 - Tf| \ge \lambda(b_{1,2} + b_{2,1} + \omega(\frac{1}{\sqrt{n}}, \mathcal{G}))$ for some $\lambda > 6$, then*

$$P(I_n = 1) \le e^{-\frac{(\lambda-6)^2}{4}}. \qquad (54)$$

Proof: We shall only give details of the proof when

$$E\hat{T}_1 - Tf \ge \lambda(b_{1,2} + b_{2,1} + \omega(\frac{1}{\sqrt{n}}, \mathcal{G})).$$

The case of $E\hat{T}_1 - Tf \le -\lambda(b_{1,2} + b_{2,1} + \omega(\frac{1}{\sqrt{n}}, \mathcal{G}))$ can be handled similarly.

Let $f \in \mathcal{F}_2$. Then

$$P(I_n = 1) \le P(\hat{T}_1 - \hat{T}_{2,1} \le 5b_{2,1} + 3\omega(\frac{1}{\sqrt{n}}, \mathcal{G})).$$

Note that

$$\begin{aligned}
& E(\hat{T}_1 - \hat{T}_{2,1} - 5b_{2,1} - 3\omega(\frac{1}{\sqrt{n}}, \mathcal{G})) \\
= \;& E(\hat{T}_1 - Tf) - E(\hat{T}_{2,1} - Tf) - 5b_{2,1} - 3\omega(\frac{1}{\sqrt{n}}, \mathcal{G})) \\
\geq \;& \lambda b_{2,1} + \lambda\omega(\frac{1}{\sqrt{n}}, \mathcal{G}) - \frac{1}{2}\omega(\sqrt{\frac{\ln\gamma_{2,1}}{n}}, \mathcal{F}_2, \mathcal{F}_1) - 5b_{2,1} - 3\omega(\frac{1}{\sqrt{n}}, \mathcal{G}) \\
\geq \;& (\lambda - 6)\left(\omega(\sqrt{\frac{\ln\gamma_{2,1}}{n}}, \mathcal{F}_2, \mathcal{F}_1) + \omega(\frac{1}{\sqrt{n}}, \mathcal{G})\right).
\end{aligned}$$

Noting that $Var(\hat{T}_1 - \hat{T}_{2,1}) \leq v_{2,1} = 2(\omega^2(\frac{1}{\sqrt{n}}, \mathcal{F}_1) + \frac{1}{\ln\gamma_{2,1}}\omega^2(\sqrt{\frac{\ln\gamma_{2,1}}{n}}, \mathcal{F}_2, \mathcal{F}_1))$, it follows that

$$\begin{aligned}
P(I_n = 1) \;&\leq\; \exp\left(-\frac{(\lambda-6)^2}{2} \cdot \frac{(\omega(\sqrt{\frac{\ln\gamma_{2,1}}{n}}, \mathcal{F}_2, \mathcal{F}_1) + \omega(\frac{1}{\sqrt{n}}, \mathcal{G}))^2}{v_{2,1}}\right) \\
&\leq\; \exp(-\frac{(\lambda-6)^2}{4}). \quad \blacksquare
\end{aligned}$$

<u>*Proof of Theorem 4:*</u> The minimax rate optimality of $\hat{T}$ over $\mathcal{F}_1$ now follows directly from Lemma 1 and Theorem 3.

$$\begin{aligned}
\sup_{f\in\mathcal{F}_1} E(\hat{T} - Tf)^2 \;&\leq\; \sup_{f\in\mathcal{F}_1} E(\hat{T}_1 - Tf)^2 + \sup_{f\in\mathcal{F}_1}(E|\hat{T}_2^* - Tf|^4)^{\frac{1}{2}} \cdot (P(I_n = 0))^{\frac{1}{2}} \\
&\leq\; \omega^2(\frac{1}{\sqrt{n}}, \mathcal{F}_1) + C\omega^2(\frac{1}{\sqrt{n}},\; \mathcal{G}) \cdot \frac{\omega^2(\frac{1}{\sqrt{n}}, \mathcal{F}_1)}{\omega^2(\frac{1}{\sqrt{n}}, \mathcal{G})} \\
&=\; C\omega^2(\frac{1}{\sqrt{n}}, \mathcal{F}_1). \qquad (55)
\end{aligned}$$

Now consider the case $f \in \mathcal{F}_2$. If $f \in \mathcal{F}_2$ and $|E\hat{T}_1 - Tf| \leq 6(b_{1,2} + b_{2,1} + \omega(\frac{1}{\sqrt{n}}, \mathcal{G}))$, then

$$\begin{aligned}
E(\hat{T} - Tf)^2 \;&\leq\; E(\hat{T}_1 - Tf)^2 + E(\hat{T}_2^* - Tf)^2 \\
&\leq\; \omega^2(\frac{1}{\sqrt{n}}, \mathcal{F}_1) + 36(b_{1,2} + b_{2,1} + \omega(\frac{1}{\sqrt{n}}, \mathcal{G}))^2 + C\omega^2(\frac{1}{\sqrt{n}}, \mathcal{G}) \\
&\leq\; C\omega^2(\sqrt{\frac{\ln\gamma_{1,2}}{n}}, \mathcal{F}_1, \mathcal{F}_2) + C\omega^2(\sqrt{\frac{\ln\gamma_{2,1}}{n}}, \mathcal{F}_2, \mathcal{F}_1) + C\omega^2(\frac{1}{\sqrt{n}}, \mathcal{G}) \qquad (56)
\end{aligned}$$

where C is a constant not depending on f.

Now note that if X has a normal distribution with mean μ and variance σ^2 then

$$(EX^4)^{\frac{1}{2}} \le 3(\mu^2 + \sigma^2). \tag{57}$$

If $f \in \mathcal{F}_2$ and $|E\hat{T}_1 - Tf| \ge \lambda(b_{1,2} + b_{2,1} + \omega(\frac{1}{\sqrt{n}}, \mathcal{G}))$ for some $\lambda > 6$, it then follows from Lemma 2 and inequality (57) that

$$\begin{aligned} E(\hat{T} - Tf)^2 &\le (E|\hat{T}_1 - Tf|^4)^{\frac{1}{2}} \cdot (P(I_n = 1))^{\frac{1}{2}} + E(\hat{T}_2^* - Tf)^2 \\ &\le \left(3Var(\hat{T}_1) + 3\lambda^2(b_{1,2} + b_{2,1} + \omega(\frac{1}{\sqrt{n}}, \mathcal{G}))^2\right) \cdot e^{-\frac{(\lambda-6)^2}{8}} + C\omega^2(\frac{1}{\sqrt{n}}, \mathcal{G}) \\ &\le C\omega^2(\sqrt{\frac{\ln \gamma_{1,2}}{n}}, \mathcal{F}_1, \mathcal{F}_2) + C\omega^2(\sqrt{\frac{\ln \gamma_{2,1}}{n}}, \mathcal{F}_2, \mathcal{F}_1) + C\omega^2(\frac{1}{\sqrt{n}}, \mathcal{G}). \end{aligned} \tag{58}$$

Again, the constant C does not depend on f. It then follows by combining (56) and (58) that

$$\begin{aligned} \sup_{f \in \mathcal{F}_2} E(\hat{T} - Tf)^2 &\le C\omega^2(\sqrt{\frac{\ln \gamma_{1,2}}{n}}, \mathcal{F}_1, \mathcal{F}_2) + C\omega^2(\sqrt{\frac{\ln \gamma_{2,1}}{n}}, \mathcal{F}_2, \mathcal{F}_1) + C\omega^2(\frac{1}{\sqrt{n}}, \mathcal{G}) \\ &\le C\left\{\omega_+^2(\sqrt{\frac{\ln \gamma_+}{n}}, \mathcal{F}_1, \mathcal{F}_2) + \omega^2(\frac{1}{\sqrt{n}}, \mathcal{F}_2)\right\}. \end{aligned} \tag{59}$$

Now (48) follows from (55) and (59). ∎

5 Adaptation over many parameter spaces

In Section 4 we showed that the lower bound on the cost of adaptation given in Section 2 is rate sharp. A general construction of adaptive estimators over two convex parameter spaces was given. However in many situations one is interested in adaptation over more than two parameter spaces. In this section we consider adaptation over a collection of parameter spaces. The construction of the adaptive estimator is based on the construction of tests between pairs of parameter spaces. The adaptive estimator shows that the lower bound on the cost of adaptation given in Section 2 can still be attained over finitely many convex parameter spaces which satisfy certain regularity conditions on the moduli. These conditions are always satisfied when the parameter spaces are nested.

5.1 Adaptation over nested parameter spaces

We begin by considering the case of adaptation over a collection of nested convex sets. Let $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \cdots \subset \mathcal{F}_k$ be closed convex parameter spaces. For convenience we will set $\mathcal{F}_0 = \emptyset$.

As in Section 4.1 let $\hat{T}_i$ be linear estimators which satisfy

$$\sup_{f \in \mathcal{F}_i} E(\hat{T}_i - Tf)^2 \le \omega^2(\frac{1}{\sqrt{n}}, \mathcal{F}_i).$$

For $i \neq j$, define the quantity $\gamma_{i,j} > 0$ as follows. If $i \wedge j = \min(i,j) = 1$, let

$$\gamma_{i,j}^2 = 1 + \frac{\omega^2(\frac{1}{\sqrt{n}}, \mathcal{F}_i, \mathcal{F}_j)}{\omega^2(\frac{1}{\sqrt{n}}, \mathcal{F}_1)} \quad \text{and} \quad \gamma_{i,j,+}^2 = 1 + \frac{\omega_+^2(\frac{1}{\sqrt{n}}, \mathcal{F}_i, \mathcal{F}_j)}{\omega^2(\frac{1}{\sqrt{n}}, \mathcal{F}_1)}. \tag{60}$$

If $i \wedge j \ge 2$, define $\gamma_{i,j}$ and $\gamma_{i,j,+}$ recursively by

$$\gamma_{i,j}^2 = 1 + \frac{\omega^2(\frac{1}{\sqrt{n}}, \mathcal{F}_i, \mathcal{F}_j)}{\max\limits_{1 \le m \le i \wedge j - 1} \{\omega_+^2(\sqrt{\frac{\ln \gamma_{m, i\wedge j,+}}{n}}, \mathcal{F}_m, \mathcal{F}_{i \wedge j})\} + \omega^2(\frac{1}{\sqrt{n}}, \mathcal{F}_{i\wedge j})}. \tag{61}$$

and

$$\gamma_{i,j,+}^2 = \max(\gamma_{i,j},\ \gamma_{j,i}) = 1 + \frac{\omega_+^2(\frac{1}{\sqrt{n}}, \mathcal{F}_i, \mathcal{F}_j)}{\max\limits_{1 \le m \le i \wedge j - 1} \{\omega_+^2(\sqrt{\frac{\ln \gamma_{m, i\wedge j,+}}{n}}, \mathcal{F}_m, \mathcal{F}_{i \wedge j})\} + \omega^2(\frac{1}{\sqrt{n}}, \mathcal{F}_{i\wedge j})}. \tag{62}$$

Let $A_i \ge 0$ be defined by $A_1^2 = \omega^2(\frac{1}{\sqrt{n}}, \mathcal{F}_1)$ and

$$A_i^2(n) = \max_{1 \le m \le i-1} \left\{ \omega_+^2(\sqrt{\frac{\ln \gamma_{m,i,+}}{n}}, \mathcal{F}_m, \mathcal{F}_i) \right\} + \omega^2(\frac{1}{\sqrt{n}}, \mathcal{F}_i). \tag{63}$$

for $2 \le i \le k$. Then

$$\gamma_{i,j}^2 = 1 + \frac{\omega^2(\frac{1}{\sqrt{n}}, \mathcal{F}_i, \mathcal{F}_j)}{A_{i\wedge j}^2(n)} \quad \text{and} \quad \gamma_{i,j,+}^2 = 1 + \frac{\omega_+^2(\frac{1}{\sqrt{n}}, \mathcal{F}_i, \mathcal{F}_j)}{A_{i\wedge j}^2(n)}.$$

Let

$$V_{i,j} = \frac{1}{\ln \gamma_{i,j}} \omega^2(\sqrt{\frac{\ln \gamma_{i,j}}{n}}, \mathcal{F}_i, \mathcal{F}_j).$$

Then for $i \neq j$

$$\begin{aligned} B_{i,j} &= \frac{1}{2}\sup_{\epsilon>0}(\omega(\epsilon, \mathcal{F}_i, \mathcal{F}_j) - \epsilon\sqrt{nV_{i,j}}) \\ &= \frac{1}{2}\sup_{\epsilon \leq \sqrt{\frac{\ln\gamma_{i,j}}{n}}}(\omega(\epsilon,\ \mathcal{F}_i, \mathcal{F}_j) - \epsilon\sqrt{\frac{n}{\ln\gamma_{i,j}}}\omega(\sqrt{\frac{\ln\gamma_{i,j}}{n}}, \mathcal{F}_i, \mathcal{F}_j)) \\ &\leq \frac{1}{2}\omega(\sqrt{\frac{\ln\gamma_{i,j}}{n}}, \mathcal{F}_i, \mathcal{F}_j). \end{aligned}$$

For $i \neq j$ let $\hat{T}_{i,j}$ be the estimator satisfying (35) - (37) with $\mathcal{F} = \mathcal{F}_i$, $\mathcal{H} = \mathcal{F}_j$ and $V = V_{i,j}$.

Let $1 \leq i < j \leq k$. Then if $f \in \mathcal{F}_i$,

$$\begin{aligned} E(\hat{T}_i - \hat{T}_{i,j}) &= E(\hat{T}_i - Tf) - E(\hat{T}_{i,j} - Tf) \\ &\geq -\omega(\frac{1}{\sqrt{n}}, \mathcal{F}_i) - \omega(\sqrt{\frac{\ln\gamma_{i,j}}{n}}, \mathcal{F}_i, \mathcal{F}_j) \\ &= -b_{i,j} \qquad (64) \end{aligned}$$

and

$$\begin{aligned} E(\hat{T}_i - \hat{T}_{j,i}) &= E(\hat{T}_i - Tf) - E(\hat{T}_{j,i} - Tf) \\ &\leq \omega(\frac{1}{\sqrt{n}}, \mathcal{F}_i) + \omega(\sqrt{\frac{\ln\gamma_{i,j}}{n}}, \mathcal{F}_j, \mathcal{F}_i) \\ &= b_{j,i}. \qquad (65) \end{aligned}$$

Note also that

$$\begin{aligned} Var(\hat{T}_i - \hat{T}_{i,j}) &\leq 2(\omega^2(\frac{1}{\sqrt{n}}, \mathcal{F}_i) + \frac{1}{\ln\gamma_{i,j}}\omega^2(\sqrt{\frac{\ln\gamma_{i,j}}{n}}, \mathcal{F}_i, \mathcal{F}_j)) \\ &= v_{i,j} \qquad (66) \end{aligned}$$

and

$$\begin{aligned} Var(\hat{T}_i - \hat{T}_{j,i}) &\leq 2(\omega^2(\frac{1}{\sqrt{n}}, \mathcal{F}_i) + \frac{1}{\ln\gamma_{i,j}}\omega^2(\sqrt{\frac{\ln\gamma_{i,j}}{n}}, \mathcal{F}_j, \mathcal{F}_i)) \\ &= v_{j,i}. \qquad (67) \end{aligned}$$

For $i < j$ define

$$\begin{aligned} I_{i,j} &= 1\Big(\hat{T}_{i,j} - (4k^{\frac{1}{2}} + 1)b_{i,j} - (8k)^{\frac{1}{2}}A_j(n) \leq \hat{T}_i \\ &\qquad \leq \hat{T}_{j,i} + (4k^{\frac{1}{2}} + 1)b_{j,i} + (8k)^{\frac{1}{2}}A_j(n)\Big). \qquad (68) \end{aligned}$$

The quantity $I_{i,j}$ defines a test between $\mathcal{F}_i$ and $\mathcal{F}_j$. The test is in favor of $\mathcal{F}_i$ if $I_{i,j} = 1$. Our adaptive estimation procedure is defined in terms of the tests $I_{i,j}$ and the minimax rate optimal estimator over $\mathcal{F}_i$, $\hat{T}_i$. The procedure can be described as follows.

1. Test between $\mathcal{F}_1$ and $\mathcal{F}_i$ based on $I_{1,i}$ for all $1 < i \le k$.

2. If all the test $I_{1,i}$ are in favor of $\mathcal{F}_1$, use $\hat{T}_1$ as the estimate of Tf.

3. Otherwise, delete $\mathcal{F}_1$ and repeat Steps 1 and 2.

Formally the estimator $\hat{T}^*$ can be written as

$$\hat{T}^* = \sum_{i=1}^{k}(1 - \prod_{m<i}\prod_{j>m} I_{m,j})(\prod_{j>i} I_{i,j})\hat{T}_i. \tag{69}$$

Theorem 5 *Let $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \cdots \subset \mathcal{F}_k$ be closed convex parameter spaces. The estimator $\hat{T}^*$ defined in (69) is adaptively rate optimal over $\mathcal{F}_i$ for $i = 1, ..., k$. That is, $\hat{T}^*$ attains the exact minimax rate of convergence over $\mathcal{F}_1$ and attains the lower bound on adaptation over $\mathcal{F}_i$ for $2 \le i \le k$ as given in Theorem 2. More specifically,*

$$\sup_{f\in\mathcal{F}_1} E(\hat{T}^* - Tf)^2 \le C\omega^2(\frac{1}{\sqrt{n}}, \mathcal{F}_1) \tag{70}$$

and for $2 \le i \le k$

$$\sup_{f\in\mathcal{F}_i} E(\hat{T}^* - Tf)^2 \le C\left(\max_{1\le m\le i-1}\left\{\omega_+^2(\sqrt{\frac{\ln\gamma_{m,i,+}}{n}}, \mathcal{F}_m, \mathcal{F}_i)\right\} + \omega^2(\frac{1}{\sqrt{n}}, \mathcal{F}_i)\right). \tag{71}$$

Remark: In light of the lower bound given in Theorem 2, it is easy to see that for any $1 \le i \le k$ the upper bound given in (71) for the maximum mean squared error over $\mathcal{F}_i$ is rate optimal given the performance of the estimator $\hat{T}^*$ over $\mathcal{F}_m$ for $m = 1, 2, ..., i-1$.

The basic ideas for the proof of Theorem 5 is similar to that of Theorem 4, but the calculations involved are more complicated.

Lemma 3 *The quantities $A_i(n)$ defined in (63) are nondecreasing in i for $1 \le i \le k$.*

Proof: Note that for $j < m \le i$, $\gamma_{j,m,+} \le \gamma_{j,i,+}$ and consequently

$$\omega_+(\sqrt{\frac{\ln\gamma_{j,m,+}}{n}}, \mathcal{F}_j, \mathcal{F}_m) \le \omega_+(\sqrt{\frac{\ln\gamma_{j,i,+}}{n}}, \mathcal{F}_j, \mathcal{F}_i)$$

It then follows that $A_i(n)$ are nondecreasing in i. ▮

Lemma 4 *If $f \in \mathcal{F}_i$, then for $j > i$,*

$$P(\hat{T}^* = \hat{T}_j) \leq k\frac{A_i^4(n)}{A_j^4(n)}. \tag{72}$$

where $A_i(n)$ is defined as in (63). In particular, (72) holds for $f \in \mathcal{F}_i \setminus \mathcal{F}_{i-1}$.

<u>*Proof:*</u> It follows from (64) - (67) that for $f \in \mathcal{F}_i$ and $i < m \leq k$

$$\begin{aligned} P(I_{i,m} = 0) &\leq P(\hat{T}_i - \hat{T}_{i,m} \leq -(4k^{\frac{1}{2}} + 1)b_{i,m} - (8k)^{\frac{1}{2}}A_m(n)) \\ &\quad + P(\hat{T}_i - \hat{T}_{m,i} \geq (4k^{\frac{1}{2}} + 1)b_{m,i} + (8k)^{\frac{1}{2}}A_m(n)) \\ &\leq \exp\left(-\frac{(4k^{\frac{1}{2}}b_{i,m} + (8k)^{\frac{1}{2}}A_m(n))^2}{2v_{i,m}}\right) \\ &\quad + \exp\left(-\frac{(4k^{\frac{1}{2}}b_{m,i} + (8k)^{\frac{1}{2}}A_m(n))^2}{2v_{m,i}}\right). \end{aligned}$$

If

$$\omega^2(\frac{1}{\sqrt{n}}, \mathcal{F}_i) \geq \frac{1}{\ln\gamma_{i,m}}\omega^2(\sqrt{\frac{\ln\gamma_{i,m}}{n}}, \mathcal{F}_i, \mathcal{F}_m),$$

then $v_{i,m} \leq 2A_i^2(n)$. Noting that $e^{-2kx} > \frac{1}{2}x^{-2k}$ for $x > 0$ and $k \geq 1$, so

$$\begin{aligned} \exp\left(-\frac{(4k^{\frac{1}{2}}b_{i,m} + (8k)^{\frac{1}{2}}A_m(n))^2}{2v_{i,m}}\right) &\leq \exp\left(-2k\frac{A_m^2(n)}{A_i^2(n)}\right) \\ &\leq \frac{1}{2}\frac{A_i^{4k}(n)}{A_m^{4k}(n)}. \end{aligned} \tag{73}$$

If $\omega^2(\frac{1}{\sqrt{n}}, \mathcal{F}_i) < \frac{1}{\ln\gamma_{i,m}}\omega^2(\sqrt{\frac{\ln\gamma_{i,m}}{n}}, \mathcal{F}_i, \mathcal{F}_m)$, then

$$\begin{aligned} \exp\left(-\frac{(4k^{\frac{1}{2}}b_{i,m} + (8k)^{\frac{1}{2}}A_m(n))^2}{2v_{i,m}}\right) &\leq \exp\left(-\frac{16k\omega^2(\sqrt{\frac{\ln\gamma_{i,m}}{n}}, \mathcal{F}_i, \mathcal{F}_m) + 8kA_m^2(n)}{\frac{4}{\ln\gamma_{i,m}}\omega^2(\sqrt{\frac{\ln\gamma_{i,m}}{n}}, \mathcal{F}_i, \mathcal{F}_m)}\right) \\ &\leq \exp\left(-(4k\ln\gamma_{i,m} + 2k\frac{A_m^2(n)}{\omega^2(\frac{1}{\sqrt{n}}, \mathcal{F}_i, \mathcal{F}_m)})\right) \\ &\leq \frac{A_i^{4k}(n)}{\omega^{4k}(\frac{1}{\sqrt{n}}, \mathcal{F}_i, \mathcal{F}_m)} \cdot \frac{1}{2}\frac{\omega^{4k}(\frac{1}{\sqrt{n}}, \mathcal{F}_i, \mathcal{F}_m)}{A_m^{4k}(n)} \\ &= \frac{1}{2}\frac{A_i^{4k}(n)}{A_m^{4k}(n)}. \end{aligned} \tag{74}$$

Combining (73) and (74) yields that

$$\exp\left(-\frac{(4k^{\frac{1}{2}}b_{i,m}+(8k)^{\frac{1}{2}}A_m(n))^2}{2v_{i,m}}\right)\le\frac{1}{2}\frac{A_i^{4k}(n)}{A_m^{4k}(n)}.$$

Similarly,

$$\exp\left(-\frac{(4k^{\frac{1}{2}}b_{i,m}+(8k)^{\frac{1}{2}}A_m(n))^2}{2v_{i,m}}\right)\le\frac{1}{2}\frac{A_i^{4k}(n)}{A_m^{4k}(n)}.$$

Therefore,

$$P(I_{i,m}=0)\le\frac{A_i^{4k}(n)}{A_m^{4k}(n)}\quad\text{for } 1\le i<m\le k. \tag{75}$$

Denote by $I_i=\prod_{j>i}I_{i,j}$. Then

$$\begin{aligned}P(\hat{T}^*=\hat{T}_j) &\le P(I_m=0\ \text{ for all }\ m=i,...,j-1)\\ &\le \min_{i\le m\le j-1}P(I_m=0)\\ &\le \left(\prod_{m=i}^{j-1}P(I_m=0)\right)^{\frac{1}{j-i-1}}.\end{aligned} \tag{76}$$

Noting that $A_i(n)$ is nondecreasing in i, it then follows from (75) that

$$P(I_m=0)\le\sum_{l=m+1}^{k}P(I_{m,l}=0)\le\sum_{l=m+1}^{k}\frac{A_m^{4k}(n)}{A_l^{4k}(n)}\le k\frac{A_m^{4k}(n)}{A_{m+1}^{4k}(n)}. \tag{77}$$

By combining (76) and (77), it follows

$$P(\hat{T}^*=\hat{T}_j)\le k\frac{A_i^4(n)}{A_j^4(n)}.$$

Lemma 5 *Suppose $f\in\mathcal{F}_i\setminus\mathcal{F}_{i-1}$ and $j<i$. If $|E\hat{T}_j-Tf|\ge\lambda(b_{j,i}+b_{i,j}+A_i(n))$ for some $\lambda>4k^{1/2}-2$, then*

$$P(\hat{T}^*=\hat{T}_j)\le\exp\left(-\frac{(\lambda-4k^{\frac{1}{2}}-2)^2}{4}\right). \tag{78}$$

Proof: We shall only give details of the proof when

$$E\hat{T}_j-Tf\ge\lambda(b_{j,i}+b_{i,j}+A_i(n)).$$

The case of $E\hat{T}_j - Tf \leq -\lambda(b_{j,i} + b_{i,j} + A_i(n))$ can be handled similarly.

Let $f \in \mathcal{F}_i \setminus \mathcal{F}_{i-1}$. Then

$$P(\hat{T}^* = T_j) \leq P(I_{j,i} = 1) \leq P(\hat{T}_j - \hat{T}_{i,j} \leq (4k^{\frac{1}{2}} + 1)b_{i,j} + (8k)^{\frac{1}{2}} A_i(n)).$$

Note that

$$\begin{aligned} & E(\hat{T}_j - \hat{T}_{i,j} - (4k^{\frac{1}{2}} + 1)b_{i,j} - (8k)^{\frac{1}{2}} A_i(n)) \\ = \ & E(\hat{T}_j - Tf) - E(\hat{T}_{i,j} - Tf) - (4k^{\frac{1}{2}} + 1)b_{i,j} - (8k)^{\frac{1}{2}} A_i(n) \\ \geq \ & \lambda b_{i,j} + \lambda A_i(n) - \frac{1}{2}\omega(\sqrt{\frac{\ln \gamma_{i,j}}{n}}, \mathcal{F}_i, \mathcal{F}_j) - (4k^{\frac{1}{2}} + 1)b_{i,j} - (8k)^{\frac{1}{2}} A_i(n) \\ \geq \ & (\lambda - 4k^{\frac{1}{2}} - 2)\left(\omega(\sqrt{\frac{\ln \gamma_{i,j}}{n}}, \mathcal{F}_i, \mathcal{F}_j) + A_i(n)\right). \end{aligned}$$

Note that

$$Var(\hat{T}_j - \hat{T}_{i,j}) \leq v_{i,j} = 2(\omega^2(\frac{1}{\sqrt{n}}, \mathcal{F}_j) + \frac{1}{\ln \gamma_{i,j}}\omega^2(\sqrt{\frac{\ln \gamma_{i,j}}{n}}, \mathcal{F}_i, \mathcal{F}_j)),$$

it follows that

$$\begin{aligned} P(\hat{T}^* = T_j) \ & \leq \ \exp\left(-\frac{(\lambda - 4k^{\frac{1}{2}} - 2)^2}{2} \cdot \frac{(\omega(\sqrt{\frac{\ln \gamma_{i,j}}{n}}, \mathcal{F}_i, \mathcal{F}_j) + A_i(n))^2}{v_{i,j}}\right) \\ & \leq \ \exp\left(-\frac{(\lambda - 4k^{\frac{1}{2}} - 2)^2}{4}\right). \quad \blacksquare \end{aligned}$$

Proof of Theorem 5: The minimax rate optimality of $\hat{T}$ over $\mathcal{F}_1$ now follows directly from Lemma 4.

$$\begin{aligned} \sup_{f \in \mathcal{F}_1} E(\hat{T}^* - Tf)^2 \ & \leq \ \sup_{f \in \mathcal{F}_1} E(\hat{T}_1 - Tf)^2 + \sum_{j=2}^{k} \sup_{f \in \mathcal{F}_1} (E|\hat{T}_j - Tf|^4)^{\frac{1}{2}} \cdot (P(\hat{T}^* = \hat{T}_j))^{\frac{1}{2}} \\ & \leq \ \omega^2(\frac{1}{\sqrt{n}}, \mathcal{F}_1) + \sum_{j=2}^{k} \omega^2(\frac{1}{\sqrt{n}}, \mathcal{F}_j) \cdot \frac{A_1^2(n)}{A_j^2(n)} \\ & \leq \ k\omega^2(\frac{1}{\sqrt{n}}, \mathcal{F}_1). \end{aligned} \tag{79}$$

Now consider the case $f \in \mathcal{F}_i \setminus \mathcal{F}_{i-1}$ for some $i > 1$. Denote

$$\begin{aligned} \mathcal{J}_1 \ & = \ \{j < i : \ |E\hat{T}_j - Tf| \leq (4k^{\frac{1}{2}} + 2)(b_{j,i} + b_{i,j} + A_i(n))\} \\ \mathcal{J}_2 \ & = \ \{j < i : \ |E\hat{T}_j - Tf| > (4k^{\frac{1}{2}} + 2)(b_{j,i} + b_{i,j} + A_i(n))\}. \end{aligned}$$

Then for $j \in \mathcal{J}_1$,

$$E(\hat{T}_j - Tf)^2 \le \omega^2(\frac{1}{\sqrt{n}}, \mathcal{F}_j) + (4k^{\frac{1}{2}} + 2)^2 (b_{j,i} + b_{i,j} + A_i(n))^2$$

If $j \in \mathcal{J}_2$, then $|E\hat{T}_j - Tf| = \lambda(b_{j,i} + b_{i,j} + A_i(n))$ for some $\lambda > 4k^{\frac{1}{2}} + 2$. So

$$\begin{aligned} (E|\hat{T}_j - Tf|^4)^{\frac{1}{2}} &\le 3Var(\hat{T}_j) + 3\lambda^2 (b_{j,i} + b_{i,j} + A_i(n))^2 \\ &\le 4\lambda^2 (b_{j,i} + b_{i,j} + A_i(n))^2. \end{aligned}$$

And it follows from Lemma 5 that

$$(P(\hat{T}^* = T_j))^{\frac{1}{2}} \le \exp\left(-\frac{(\lambda - 4k^{\frac{1}{2}} - 2)^2}{8}\right).$$

Hence, for $f \in \mathcal{F}_i \setminus \mathcal{F}_{i-1}$ with $i > 1$

$$\begin{aligned} E(\hat{T}^* - Tf)^2 &= \sum_{j=i}^{k} E\{(\hat{T}_j - Tf)^2 1(\hat{T}^* = \hat{T}_j)\} \\ &\le \sum_{j\in\mathcal{J}_1} E(\hat{T}_j - Tf)^2 + \sum_{j\in\mathcal{J}_2} (E|\hat{T}_j - Tf|^4)^{\frac{1}{2}} (P(\hat{T}^* = \hat{T}_j))^{\frac{1}{2}} \\ &\quad + E(\hat{T}_i - Tf)^2 + \sum_{j=i+1}^{k} (E|\hat{T}_j - Tf|^4)^{\frac{1}{2}} (P(\hat{T}^* = \hat{T}_j))^{\frac{1}{2}} \\ &\le \sum_{j\in\mathcal{J}_1} \{\omega^2(\frac{1}{\sqrt{n}}, \mathcal{F}_j) + (4k^{\frac{1}{2}} + 2)^2 (b_{j,i} + b_{i,j} + A_i(n))^2\} \\ &\quad + \sum_{j\in\mathcal{J}_2} 4\lambda^2 (b_{j,i} + b_{i,j} + A_i(n))^2 \cdot \exp\left(-\frac{(\lambda - 4k^{\frac{1}{2}} - 2)^2}{4}\right) \\ &\quad + \omega^2(\frac{1}{\sqrt{n}}, \mathcal{F}_i) + \sum_{j=i+1}^{k} 6\, \omega^2(\frac{1}{\sqrt{n}}, \mathcal{F}_j) \cdot k^{\frac{1}{2}} \frac{A_i^2(n)}{A_j^2(n)} \\ &\le CA_i^2(n) \end{aligned}$$

where C is a constant not depending on f. Note that in the last inequality we use the fact that $\lambda^2 \exp\left(-\frac{(\lambda - 4k^{1/2} - 2)^2}{4}\right)$ are bounded as a function of λ.

Hence,

$$\begin{aligned} \sup_{f\in\mathcal{F}_i} E(\hat{T}^* - Tf)^2 &= \max_{1\le m\le i} \left\{ \sup_{f\in\mathcal{F}_m \setminus \mathcal{F}_{m-1}} E(\hat{T}^* - Tf)^2 \right\} \\ &\le C \max_{1\le m\le i} \{A_m^2(n)\} = CA_i^2(n) \\ &= C\left(\max_{1\le m\le i-1} \left\{ \omega_+^2(\sqrt{\frac{\ln \gamma_{m,i,+}}{n}}, \mathcal{F}_m, \mathcal{F}_i) \right\} + \omega^2(\frac{1}{\sqrt{n}}, \mathcal{F}_i) \right). \quad \blacksquare \end{aligned}$$

5.2 Adaptation over non-nested parameter spaces

Many common parameter spaces of interest such as Lipschitz spaces and Besov spaces are not nested. However, they often have nested structure in terms of modulus of continuity. Theorem 5 can be generalized to such non-nested parameter spaces.

Let $\mathcal{F}_i,\ i = 1, ..., k$ be closed, convex parameter spaces which are not necessarily nested. For any parameter set $\mathcal{F}$, let $C.Hull(\mathcal{F})$ denote the convex hull of $\mathcal{F}$. Suppose the parameter spaces $\mathcal{F}_i$ satisfy the following conditions on the modulus of continuity.

1. For $l \le i$ and $m \le j$,
$$\omega(\epsilon, \mathcal{F}_l, \mathcal{F}_m) \le C\omega(\epsilon, \mathcal{F}_i, \mathcal{F}_j)$$
for some constant $C > 0$.

2. For $2 \le i \le k$,
$$\omega(\epsilon, \mathcal{G}_i) \asymp \omega(\epsilon, C.Hull(\mathcal{G}_i))$$
where $\mathcal{G}_i = \cup_{m=1}^{i} \mathcal{F}_m$.

Note that conditions 1 and 2 are trivially satisfied if $\mathcal{F}_i$ are nested.

As shown in Cai and Low (2002), the minimax linear rate of convergence for estimating a linear functional Tf over a parameter set $\mathcal{F}$ is determined by the modulus over its convex hull, $\omega(\frac{1}{\sqrt{n}}, C.Hull(\mathcal{F}))$. Conditions 1 and 2 together yield

$$\omega(\epsilon, \mathcal{F}_i) \asymp \omega(\epsilon, C.Hull(\mathcal{G}_i))$$

and this consequently implies that for $1 \le i \le k$ there exists a rate optimal linear estimator $\hat{T}_i$ over $\mathcal{F}_i$ such that

$$\sup_{f \in \cup_{m=1}^{i} \mathcal{F}_m} E(\hat{T}_i - Tf)^2 \le C\omega(\frac{1}{\sqrt{n}}, \mathcal{F}_i). \tag{80}$$

Now define the quantities $\gamma_{i,j}$ and $\gamma_{i,j,+}$ as in (60), (61) and (62). Let the estimator $\hat{T}^*$ be defined same as in Section 5.1 with the minimax rate optimal linear estimator $\hat{T}_i$ over $\mathcal{F}_i$ satisfying (80). Under the conditions 1 and 2 above, the estimator $\hat{T}^*$ then achieves adaptation over the parameter spaces $\mathcal{F}_i$ with minimum cost. More precisely, we have the following.

Theorem 6 *Let $\mathcal{F}_i$, $i = 1, ..., k$ be closed convex parameter spaces satisfying conditions 1 and 2 above and let the estimator $\hat{T}^*$ be given as above. Then $\hat{T}^*$ attains optimal adaptive rate of convergence over $\mathcal{F}_i$ for $i = 1, ..., k$. That is,*

$$\sup_{f\in\mathcal{F}_1} E(\hat{T}^* - Tf)^2 \le C\omega^2(\frac{1}{\sqrt{n}}, \mathcal{F}_1) \tag{81}$$

and for $2 \le i \le k$

$$\sup_{f\in\mathcal{F}_i} E(\hat{T}^* - Tf)^2 \le C\left(\max_{1\le m\le i-1}\left\{\omega_+^2(\sqrt{\frac{\ln\gamma_{m,i,+}}{n}}, \mathcal{F}_m, \mathcal{F}_i)\right\} + \omega^2(\frac{1}{\sqrt{n}}, \mathcal{F}_i)\right). \tag{82}$$

The proof of Theorem 6 is essentially the same as that of Theorem 5. We omit the proof for reasons of space.

5.3 An example of adaptation over non-nested spaces

The problem of estimating decreasing Lipschitz functions has been considered in Donoho and Liu (1991). See also Kiefer (1982). A fully rate adaptive procedure over the decreasing Lipschitz classes has been introduced in Kang and Low (2002). The estimator however performs poorly over general Lipschitz classes.

We now consider adaptive estimation over a mixed collection of decreasing and arbitrary Lipschitz classes. For simplicity consider adaptive estimation of $Tf = f(0)$ over four parameter spaces, $\mathcal{F}_i$, $i = 1, 2, 3$, and 4 where

$$\mathcal{F}_1 = F_D(1, M), \quad \mathcal{F}_2 = F(\frac{2}{3}, M), \quad \mathcal{F}_3 = F_D(\frac{1}{2}, M), \quad \text{and} \quad \mathcal{F}_4 = F(\frac{1}{2}, M).$$

Note that these parameter spaces are not nested. Standard arguments similar to those found in Section 3 show that for $i = 1, 2, 3, 4$

$$\omega_+(\epsilon, \mathcal{F}_i, \mathcal{F}_4) \asymp \omega(\epsilon, \mathcal{F}_4) = C\epsilon^{\frac{1}{2}}(1 + o(1)),$$

and

$$\begin{aligned}
\omega_+(\epsilon, \mathcal{F}_1, \mathcal{F}_3) &\asymp \omega(\epsilon, \mathcal{F}_1) = C\epsilon^{\frac{2}{3}}(1 + o(1)), \\
\omega_+(\epsilon, \mathcal{F}_2, \mathcal{F}_3) &\asymp \omega(\epsilon, \mathcal{F}_3) = C\epsilon^{\frac{1}{2}}(1 + o(1)), \\
\omega_+(\epsilon, \mathcal{F}_1, \mathcal{F}_2) &\asymp \omega(\epsilon, \mathcal{F}_2) = C\epsilon^{\frac{4}{7}}(1 + o(1)).
\end{aligned}$$

It is then easy to see that conditions 1 and 2 of Section 5.2 are satisfied.

It then follows that the estimator given in Theorem 6 is fully rate adaptive in the sense that it attains the minimax rate over the two decreasing parameter spaces $F_D(1, M)$ and $F_D(\frac{1}{2}, M)$ and only loses a logarithmic penalty over $F(1, M)$ and $F(\frac{1}{2}, M)$. Note that the loss of the logarithmic factor over both $F(1, M)$ and $F(\frac{1}{2}, M)$ is in fact necessary once you attain the optimal rate over $F_D(1, M)$. We believe that this is the first example of such a phenomena where one loses a logarithmic factor on some classes yet is fully adaptive over other function classes.

This example can be generalized to any finite number of Lipschitz classes, decreasing Lipschitz classes and increasing Lipschitz classes with smoothness index $0 < \alpha \leq 1$. Hence although an adaptive estimator must pay a logarithmic penalty in terms of maximum risk over Lipschitz classes if the function is either monotonely increasing or monotonely decreasing in a fixed interval of the point of interest then a penalty need not be paid. This would of course be the case for most functions of interest.

References

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

[2] Brown, L. D. & Low, M. G. (1996b). A constrained risk inequality with applications to nonparametric functional estimation. *Ann. Statist.* **24**, 2524 - 2535.

[3] Butucea, C. (2001). Exact adaptive pointwise estimation on Sobolev classes of densities. *ESAIM: Prob. Statist.* **5**, 1-31.

[4] Cai, T. & Low. M. (2001). On nonparametric estimation of linear functionals. Technical Report, Department of Statistics, University of Pennsylvania.

[5] Cai, T. & Low. M. (2002). Minimax estimation of linear functionals over nonconvex parameter spaces. Technical Report, Department of Statistics, University of Pennsylvania.

[6] Cai, T., Low, M., & Zhao, L. (2001). Tradeoffs between global and local risks in nonparametric function estimation. Technical Report, Department of Statistics, University of Pennsylvania.

[7] Donoho, D.L. (1994). Statistical estimation and optimal recovery. *Ann. Statist.* **22**, 238-270.

[8] Donoho, D. L. & Liu, R. G. (1991). Geometrizing rates of convergence III. *Ann. Statist.* **19**, 668-701.

[9] Efromovich, S. (2000). Sharp adaptive estimation of multivariate curves. *Math. Meth. Statist.* **9**, 117-139.

[10] Efromovich, S. & Low, M. G. (1994). Adaptive estimates of linear functionals. *Probab. Theory Rel. Fields* **98**, 261-275.

[11] Ibragimov, I. A. & Hasminskii, R. Z. (1981). *Statistical estimation: asymptotic theory*, Springer, New York.

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

[13] Kang, Y.-G. & Low, M. G. (2002). Estimating monotone functions. *Statist. Prob. Letters*, in press.

[14] Kiefer, J. (1982). Optimum rates for non-parametric density and regression estimates, under order restrictions. In Kallianpur, G., Krishnaiah, P.R. and Ghosh, J.K. (Eds.), *Statistics and Probability: Essays in honor of C.R. Rao,* North-Holland, 419-428.

[15] Klemelä, J. & Tsybakov, A. B. (2001). Sharp adaptive estimation of linear functionals. *Ann. Statist.* **29**, 1567-1600.

[16] Lepski, O. V. (1990). On a problem of adaptive estimation in Gaussian white noise. *Theory Probab. Appl.* **35**, 454-466.

[17] Lepski, O. V. & Levit, B. Y. (1998). Adaptive minimax estimation of infinitely differentiable functions. *Math. Methods Statist.* **7**, 123-156.

[18] Low, M. G. (1992). Renormalization and white noise approximation for nonparametric functional estimation problems. *Ann. Statist.* **20**, 545-554.

[19] Low, M. G. (1995). Bias-variance tradeoffs in functional estimation problems. *Ann. Statist.* **23**, 824-835.

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