A Framework For Estimation of Convex Functions

T. Tony Cai* and Mark G. Low
Department of Statistics
The Wharton School
University of Pennsylvania

Abstract

A general non-asymptotic framework, which evaluates the performance of any procedure at individual functions, is introduced in the context of estimating convex functions at a point. This framework, which is significantly different from the conventional minimax theory, is also applicable to other problems in shape constrained inference.

A benchmark is provided for the mean squared error of any estimate for each convex function in the same way that Fisher Information depends on the unknown parameter in a regular parametric model. A local modulus of continuity is introduced and is shown to capture the difficulty of estimating individual convex functions. A fully data-driven estimator is proposed and is shown to perform uniformly within a constant factor of the ideal benchmark for every convex function. Such an estimator is thus adaptive to every unknown function instead of to a collection of function classes as is typical in the nonparametric function estimation literature.

Keywords: Adaptive estimation, convex function, local modulus of continuity, minimax estimation, nonparametric regression, shape constrained inference, white noise model.

1 Introduction

The problem of estimating functions under assumptions of convexity or monotonicity has a long history dating back at least to Grenander (1956). The extensive literature on this topic has partly been motivated by specific applications but also by the fact that these shape constrained problems have features that are shared with regular parametric models and other features that are shared with nonparametric function estimation.

One connection with parametric models is that the least squares and maximum likelihood estimates perform well. On the other hand, as is typical in nonparametric function estimation, the rates of convergence in these models are slower than the usual root n rate. For example Mammen (1991) established rates of convergence of the least squares estimate for the value of a piecewise convex/concave function at a point. Precise analysis of the asymptotic distributions for estimating a convex function has also been given in Groeneboom, Jongbloed and Wellner

*The research of Tony Cai was supported in part by NSF Grant DMS-0604954 and NSF FRG Grant DMS-0854973.

(2001a, b) assuming that the function has a positive continuous second derivative at the point of interest. Attention has also focused on developing minimax theory under a variety of performance measures. For example Birgé (1989) studied the minimax risk of the Grenander estimator under an L_1 loss. For estimating at a point, Kiefer (1982) showed that adding conditions such as convexity to, for example, Lipschitz classes would not change the usual minimax rates of convergence found in unconstrained nonparametric function estimation.

For monotone functions Cator (2011) considers the problem of adaptation under probabilistic error adopting the point of view of Cai and Low (2006) but applied to shrinking neighborhoods of fixed monotone functions. It is shown that for shrinking neighborhoods the least squares estimate is rate optimal under probabilistic error. It should be noted that, in contrast to mean squared error, fully rate optimal adaptation under probabilistic error is possible over a wide range of Lipschitz classes even without shape constraints. See Cai and Low (2006).

In addition to local and global estimation there has also been considerable work on developing confidence bands. The main goal is the construction of confidence bands with guaranteed coverage probability that also adapt to the local smoothness of the underlying function. As shown in Low (1995) such confidence intervals often cannot be constructed, but assuming shape constraints Hengartner and Stark (1995) and Dümbgen (1998) give a variable width confidence band which does adapt to local smoothness while maintaining a given level of coverage probability. See also Cai, Low and Xia (2013) for the construction of adaptive confidence intervals rather than bands. This quick survey of a few results from the extensive literature is not meant to be exhaustive but to give a sense of the range of problems that have been considered.

In the present paper a new framework is introduced for estimating a shape constrained function at a point. This framework is non-asymptotic and focused on the performance at each individual function. Here we study the problem of estimating a convex function but the approach extends to estimating monotone functions as well as other classes of shape restrictions. A benchmark is provided for the mean squared error of any estimate which depends on the unknown convex function in a similar way that the Fisher Information bound depends on the unknown parameter in a regular parametric model. This approach should be contrasted to the minimax theory for nonparametric function estimation where only the maximum risk over a large parameter space is considered.

For ease of exposition these ideas will be explained in the context of a white noise with drift model

$$dY(t) = f(t)dt + n^{-1/2}dW(t) \quad -1/2 \le t \le 1/2 \tag{1}$$

where $W(t)$ is Brownian motion. The choice of the parameter n in this white noise with drift model is to make clear the connection to a nonparametric regression model with n equally spaced design points. See, for example, Brown and Low (1996a). We shall also consider convex regression in Section 5. We focus here on estimating f at 0 since estimation at other points away from the boundary is similar. For any estimator $\hat{T}_n$ of $f(0)$, write the risk under squared error loss as

$$R(\hat{T}_n, f) = E(\hat{T}_n - f(0))^2 \tag{2}$$

and denote by F_c the collection of convex functions on the interval $[-1/2, 1/2]$.

There are three main goals of the present paper. The first is to introduce a benchmark which measures the difficulty of estimating $f(0)$ for each convex function f as well as to explain why

this benchmark is natural in the present setting. The second is to provide a concise description of the benchmark in terms of an analytic quantity, the local modulus of continuity, and the third goal is to construct a data-driven estimator which has a risk uniformly within a constant factor of the benchmark for all convex functions and for all n. However, before explaining this approach in more detail, it is helpful to summarize a few key features and results from nonparametric function estimation theory, in particular those on estimating a function at a point.

1.1 Minimax Theory in Nonparametric Function Estimation

In classical parametric inference, optimality theory is well developed especially from an asymptotic point of view. Parameters are most often estimated at the root n rate and the efficiency of estimators can be evaluated by comparing the normalized asymptotic variance with the Fisher Information bound. The justification for this bound can be made precise either in terms of the Hájek-Le Cam convolution theorem or local asymptotic minimax theory. For example it is well known that under suitable regularity conditions the maximum likelihood estimator is efficient in terms of weak convergence and locally asymptotically minimax at each point in the parameter space.

In contrast, asymptotic optimality theory developed for nonparametric function estimation differs markedly from this classical theory. In particular rates of convergence are typically much slower than the root n rate and the precise rate of convergence depends on the parameter space. Another major difference, which is particularly important for the present paper, is that the rate of convergence is often described not in a local way but rather in terms of the maximum risk over an entire parameter space. See however Groeneboom, Jongbloed and Wellner (2001b) and Jongbloed (2000) for examples of where a more local analysis is given.

Consider for example the problem of focus in the present paper namely that of estimating the function f at the point $f(0)$. The minimax theory for estimating such linear functionals reached a mature form in the work of Ibragimov and Hasminskii (1984), Donoho and Liu (1991) and Donoho (1994). As a concrete example it is helpful to focus on one commonly used collection of parameter spaces the Lipschitz balls which are given by

$$\Lambda(\alpha, M) = \{f : |f(y) - f(x)| \le M|y - x|^\alpha \text{ for } x, y \in [-\tfrac{1}{2}, \tfrac{1}{2}]\}, \quad \text{if } 0 < \alpha \le 1$$

and for $\alpha > 1$

$$\Lambda(\alpha, M) \;=\; \{f : |f^{(\lfloor \alpha \rfloor)}(x) - f^{(\lfloor \alpha \rfloor)}(y)| \le M\,|x - y|^{\alpha'} \text{ for } x, y \in [-\tfrac{1}{2}, \tfrac{1}{2}]\},$$

where $\lfloor \alpha \rfloor$ is the largest integer less than α and $\alpha' = \alpha - \lfloor \alpha \rfloor$. In this setting the minimax risk summarized by

$$C_1 M^{\frac{2}{2\alpha+1}} n^{-\frac{2\alpha}{2\alpha+1}} \le \inf_{\hat{T}_n} \sup_{f \in \Lambda(\alpha, M)} E(\hat{T}_n - f(0))^2 \le C_2 M^{\frac{2}{2\alpha+1}} n^{-\frac{2\alpha}{2\alpha+1}} \tag{3}$$

for two constants $C_2 \ge C_1 > 0$ describes precisely the dependence of the minimax risk on α, M and n. Specific linear procedures and lower bound arguments found in the above mentioned papers narrow considerably the gap between the two constants C_1 and C_2.

One criticism of this theory is that the minimax benchmark is too conservative and the rate optimal estimate relies on the knowledge of the smoothness parameter α. Since a function can

belong to a range of these function classes, for which the minimax risk can be quite different, it is not immediately clear how well one should expect or hope to estimate individual functions. As a response to such concerns there has been a great effort over the last thirty years to develop adaptive procedures that are simultaneously minimax over a collection of parameter spaces. This point of view and history is particularly well explained in Donoho, Johnstone, Kerkyacharian and Picard (1995) in the context of global estimation under integrated mean squared error. For estimation at a point, Lepski (1990) considered adaptive estimation and showed that over a collection of Lipschitz balls it is not possible to simultaneously attain the minimax rate, as the parameter n tends to infinity, over every parameter space. It was also shown that the minimal cost of adaptation is a logarithmic factor of the parameter n. It should however be stressed again that the benchmark for adaptive estimators is still provided by considering the maximum risk over large parameter spaces and is not focused at the level of individual functions.

1.2 Local Framework

The decision theoretic framework introduced in the present paper is focused on the performance at every function. However in order to assess the difficulty of estimating a particular function one must at least consider an additional function since otherwise the problem is degenerate. For a given function $f \in F_c$ it is natural to choose the other convex function, say g, to be the one which is most difficult to distinguish from f in the mean squared error sense. The benchmark $R_n(f)$ can then be expressed as

$$R_n(f) = \sup_{g \in F_c} \inf_{\hat{T}_n} \max\{R(\hat{T}_n, f), R(\hat{T}_n, g)\}. \tag{4}$$

One of the major goals of this paper is to demonstrate why $R_n(f)$ is a useful benchmark in the present context of estimating convex functions. This shall be established by showing

1. $R_n(f)$ varies considerably over the collection of convex functions;
2. There is a procedure that has a risk uniformly within a constant factor of $R_n(f)$ for every convex function f and every n;
3. Outperforming the benchmark $R_n(f)$ at some convex function f leads to worse performance at other functions.

It is the combination of these three factors that make $R_n(f)$ a useful benchmark and establishing these three points in the context of estimating convex functions is the main purpose of the present paper.

In order to make progress towards these three goals it is important to develop technical tools for studying $R_n(f)$. First we show that $R_n(f)$ can be described in terms of an analytic quantity, a local modulus of continuity $\omega(\epsilon, f)$ defined by

$$\omega(\epsilon, f) = \sup_{g \in F_c} \{|g(0) - f(0)| : ||g - f||_2 \leq \epsilon\} \tag{5}$$

where $||f||_2$ is the usual L_2 norm of a function f. Note that this local modulus $\omega(\epsilon, f)$ clearly depends on f and can be regarded as an analogue of Fisher Information in regular parametric

models. The following bounds for $R_n(f)$, which are a direct consequence of Theorem 1, show that the local modulus ω captures the difficulty of estimating an individual convex function f,

$$\frac{1}{9}\omega^2(\frac{2}{\sqrt{n}}, f) \leq R_n(f) \leq \frac{1}{4}\omega^2(\frac{2}{\sqrt{n}}, f). \tag{6}$$

Thus this local modulus characterizes the difficulty of estimating a particular convex function in a similar manner that the usual global modulus of continuity describes the minimax risk as shown in Donoho and Liu (1991), replacing a statistical problem by an approximation theory problem.

Although these bounds help us towards the goal of showing that $R_n(f)$ varies considerably over the collection of convex functions the actual evaluation of these bounds for any given f still involves a supremum over all convex functions. For this reason it is also important to develop further technical tools for the study of $\omega(\epsilon, f)$ which will allow this supremum to be replaced by a quantity more specifically tied to the function f. In particular we show that the local modulus can be further expressed in terms of an easily computable function K which depends only on the convex function f, replacing a search over all convex functions and thus simplifiying considerably the study of $R_n(f)$. In particular it allows us to demonstrate that $R_n(f)$ varies considerably as f ranges over the class of convex functions.

There are two more steps to demonstrating that $R_n(f)$ is a useful benchmark. First, we show that $R_n(f)$ can be essentially attained at each convex function f. That is, we shall construct a data-driven estimator $\hat{T}_*$ which satisfies

$$\sup_{f \in F_c} \frac{R(\hat{T}_*, f)}{R_n(f)} \leq C \tag{7}$$

for some absolute constant $C > 0$ not depending on n. A procedure that satisfies (7) thus performs uniformly within a constant factor of the ideal benchmark $R_n(f)$. Such an estimator $\hat{T}_*$ is adaptive to every unknown function instead of to a collection of function classes as in the conventional nonparametric function estimation literature. A procedure satisfying (7) will be described in the next subsection.

Second, we show that there are consequences for outperforming the benchmark at a particular function. Specifically, we prove that for any estimator $\hat{T}_n$ if

$$R(\hat{T}_n, f) \leq cR_n(f), \tag{8}$$

then there exists another convex function h such that

$$R(\hat{T}_n, h) \geq d \ln(\frac{1}{c}) R_n(h) \tag{9}$$

for some absolute constant $d > 0$. This shows that if an estimator significantly outperforms the benchmark $R_n(f)$ at any particular convex function f there must be another convex function h at which the estimator performs poorly. This phenomenon is similar to superefficient estimators in classical parametric estimation problems. In this sense $R_n(f)$ is like the Fisher Information bound in regular parametric models although not at the level of constants.

1.3 Estimation Procedure

For the second goal we need to establish that the bound $R_n(f)$ is attainable. For this an estimator $\hat{T}_*$ is constructed and shown to perform well for every convex function f in the sense of (7). The procedure is particularly easy to describe and to implement.

Denote by $B(t)$ the symmetric box kernel $B(t) = I(|t| \le \frac{1}{2})$. Set $B_j(t) = 2^{j-1}B(2^{j-1}t)$ and let

$$\delta_j = \int B_j(t)dY(t), \quad j \ge 1 \tag{10}$$

be a sequence of local average estimators for $f(0)$. It can be shown that the estimators δ_j have nonnegative and monotonically decreasing biases. The key is to choose an estimator δ_j which optimally trades bias and variance. Set

$$\hat{j} = \inf_j\{j : \delta_j - \delta_{j+1} \le \lambda\frac{2^{(j-1)/2}}{\sqrt{n}}\} \tag{11}$$

where λ is a positive constant and define the estimator $\hat{T}_*$ of $f(0)$ by

$$\hat{T}_* = \hat{f}(0) = \delta_{\hat{j}}. \tag{12}$$

The motivation and analysis of this estimator will be given in Section 4. We should note here that many of the basic properties of this estimator hold for a large range of λ but in this paper we shall take λ to be equal to $\sqrt{2}$ and in this case it will be shown that the mean square error of this data-driven estimator is within a factor of 6 of an ideal local average oracle risk which itself is shown to be uniformly within a constant factor of the benchmark $R_n(f)$ for every convex function f and for all $n \ge 1$. These results together yield the uniform bound (7).

A key feature in the technical analysis of the performance of $\hat{T}_*$ is that for estimating convex functions the bias of estimators such as δ_j can be learned quite precisely which is not possible in general without shape constraints. This knowledge of the bias makes it possible to essentially mimic the performance of an oracle estimator for every convex function.

1.4 Organization of the Paper

The rest of the paper is organized as follows. Section 2 discusses in detail the benchmark using the hardest local alternative as well as other technical tools. Section 3 introduces a local average oracle risk which can also be described by the local modulus of continuity. This section also discusses superefficiency as measured by $R_n(f)$. Section 4 investigates the properties of the data-driven estimator (12) and shows that the estimator is within a factor of 6 of the ideal local average oracle risk for every convex function and consequently it is uniformly within a constant factor of the benchmark $R_n(f)$. The local non-asymptotic framework developed in this paper is applicable to shape constrained inference in general. Section 5 considers nonparametric convex regression under the same framework. Section 6 discusses extensions of the results and possible applications of the new framework to other related problems including estimation of monotone functions and construction of confidence sets. We also explain in this section why the framework given in this paper does not work for general nonparametric function estimation without shape constraints. All the technical proofs are given in Section 7.

2 Hardest Local Alternative and Oracle Benchmark

In this section we develop further the analysis of the new framework described in Section 1.2. The benchmark $R_n(f)$ introduced in Section 1.2 relies on the selection of a hardest local alternative to f. The consideration of using such local alternatives in the context of high dimensional statistical models can be viewed as a partial analogue of a semiparametric efficiency bound in terms of the Fisher Information as in Stein (1956) or Bickel, Klassen, Ritov and Wellner (1993) where a hardest one parameter family is considered.

2.1 Characterization of the Benchmark $R_n(f)$

In this section we show for a given convex function f, that effective upper and lower bounds for $R_n(f)$ can be given in terms of the local modulus of continuity $\omega(\epsilon, f)$ introduced in the introduction. These bounds are presented in equation (20) of Theorem 1. They are essentially all that is needed when a particular convex function f is studied. However, in order to understand how $R_n(f)$ depends on f it is useful to provide a more direct bound on $R_n(f)$ that does not rely on the computation of ω, a quantity that involves searching over all convex functions. For such an analysis it is convenient to replace the local modulus by a function that we call the K function which is easily computable. We first introduce some notations that are useful in our technical analysis.

For each convex function f define $f_s(t)$ by

$$f_s(t) = \begin{cases} \frac{f(t)+f(-t)}{2} - f(0) & \text{if } 0 \le t < \frac{1}{2} \\ \lim_{t\to\frac{1}{2}} \frac{f(t)+f(-t)}{2} - f(0) & \text{if } t = \frac{1}{2} \end{cases} . \tag{13}$$

So f_s is a symmetrized and centered version of the convex function f. It is easy to see that f_s is convex and nondecreasing on $[0, \frac{1}{2}]$. Let H be the function defined on the domain $[0, \frac{1}{2}]$ by

$$H(t) = \sqrt{t} f_s(t). \tag{14}$$

It can be easily checked that $H(t)$ is a continuous convex and nondecreasing function on $[0, 1/2]$. When $H(t) > 0$ it is strictly increasing. An inverse of this function with domain $[0, \infty)$ can be defined by

$$H^{-1}(x) = \sup\{t : 0 \le t \le \frac{1}{2}, H(t) \le x\}. \tag{15}$$

It is easy to check that for all $0 \le t < \infty$,

$$H(H^{-1}(t)) \le t \tag{16}$$

and that if $H^{-1}(t) < \frac{1}{2}$ then

$$H(H^{-1}(t)) = t. \tag{17}$$

Finally for $0 < t < \infty$ define the function K by

$$K(t) = \frac{t}{\sqrt{H^{-1}(t)}}. \tag{18}$$

It is clear that the function K depends only on the convex function f. This function can be viewed as a type of curvature that measures the rate of change of the function f near the origin.

It plays an essential role in our technical analysis. The following result shows that the local modulus of continuity can be bounded in terms of the function K.

Proposition 1 *For $\epsilon > 0$ and $f \in F_c$,*

$$K(\frac{2}{3}\epsilon) \leq \omega(\epsilon, f) \leq K(\sqrt{\frac{10}{3}}\epsilon). \tag{19}$$

It is important to see that the difficulty of estimating a convex function f at 0 can be captured by the local modulus $\omega(\frac{1}{\sqrt{n}}, f)$ or the function K.

Theorem 1 *The benchmark $R_n(f)$ satisfies*

$$\frac{1}{9}\omega^2(\frac{2}{\sqrt{n}}, f) \leq R_n(f) \leq \frac{1}{4}\omega^2(\frac{2\sqrt{2/e}}{\sqrt{n}}, f) \tag{20}$$

and also

$$\frac{1}{9}K^2(\frac{4}{3\sqrt{n}}) \leq R_n(f) \leq \frac{1}{4}K^2(\frac{4\sqrt{5}}{\sqrt{3e}\sqrt{n}}). \tag{21}$$

Note that in particular (20) implies (6) as $\omega(\epsilon, f)$ is strictly increasing so $\omega(\frac{2\sqrt{2/e}}{\sqrt{n}}, f) \leq \omega(\frac{2}{\sqrt{n}}, f)$. We should stress that the focus of these results is not on the specific constants provided, which can be improved with more refined calculations, but on the general nature of the results. An analysis of some specific examples will provide additional insights. Some of these are covered in Section 2.2.

Remark 1 Donoho and Liu (1991) used a global modulus of continuity to study minimax estimation of linear functionals. In the context of estimating a function at 0 it is defined as

$$\omega_*(\epsilon, \mathcal{F}) = \sup_{f,g \in \mathcal{F}} \{|g(0) - f(0)| : ||g - f||_2 \leq \epsilon\}.$$

Note that $\omega_*(\epsilon, \mathcal{F}) = \sup_{f \in \mathcal{F}} \omega(\epsilon, f)$. For a convex parameter space $\mathcal{F}$ Donoho and Liu (1991) and Donoho (1994) have shown that the minimax risk $R_*(n, \mathcal{F}) = \inf_{\hat{T}} \sup_{f \in \mathcal{F}} E(\hat{T} - f(0))^2$ can be bounded in terms of the global modulus of continuity,

$$\frac{1}{8}\omega_*^2(\frac{1}{\sqrt{n}}, \mathcal{F}) \leq R_*(n, \mathcal{F}) \leq \omega_*^2(\frac{1}{\sqrt{n}}, \mathcal{F}).$$

2.2 Examples

After the technical development given in the last section we now give a few examples which show that the benchmark $R_n(f)$ varies as f ranges over the class of convex functions. The results also show that $R_n(f)$ captures the difficulty of estimation in a much more precise way than the minimax risks can. For a given convex function we shall first evaluate the function K which can be easily done and which then yields immediately bounds for the local modulus $\omega(\epsilon, f)$ and hence also the benchmark $R_n(f)$.

Example 1 We begin with a simple example. Let $f(t) = ct$ where c is a constant. In this case $f_s(t) = 0$ and $H(t) = 0$, $0 < t \leq \frac{1}{2}$. So $H^{-1}(t) = 1/2$ and consequently $K(t) = \sqrt{2}t$. Hence for $f(t) = t$,

$$\frac{2\sqrt{2}}{3}\epsilon \leq \omega(\epsilon, f) \leq 2\sqrt{\frac{5}{3}}\epsilon$$

and

$$\frac{32}{81}n^{-1} \le R_n(f) \le \frac{40}{3e}n^{-1}.$$

In particular, the rate of convergence of $R_n(f)$ is parametric in this case.

Example 2 We now consider a symmetric function $f(t) = |t|^r$ with $r \ge 1$. In this case $f_s(t) = f(t) = |t|^r$ and $H(t) = t^{(2r+1)/2}$, $0 < t \le \frac{1}{2}$. So $H^{-1}(t) = t^{2/(2r+1)}$ and consequently $K(t) = t^{2r/(2r+1)}$. Hence for $f(t) = |t|^r$ with $r \ge 1$,

$$(\frac{2}{3})^{\frac{2r}{2r+1}}\epsilon^{\frac{2r}{2r+1}} \le \omega(\epsilon, f) \le (\frac{10}{3})^{\frac{r}{2r+1}}\epsilon^{\frac{2r}{2r+1}},$$

and

$$\frac{1}{9}(\frac{4}{3})^{\frac{4r}{2r+1}}n^{-\frac{2r}{2r+1}} \le R_n(f) \le \frac{1}{4}(\frac{80}{3e})^{\frac{2r}{2r+1}}n^{-\frac{2r}{2r+1}}.$$

In particular, for $f(t) = |t|$, the difficulty of estimating $f(0)$ as measured by the mean squared error is of order $n^{-2/3}$ and for $f(t) = t^2$, the difficulty of estimating $f(0)$ is of order $n^{-4/5}$. In fact for the case of $f(t) = |t|$ it is easy to check that for small ϵ, $\omega(\epsilon, f) = (\frac{2}{3})^{\frac{2r}{2r+1}}\epsilon^{\frac{2r}{2r+1}}$ and it is also possible to give an exact expression for $R_n(f)$.

In the two examples above, the function K are computed analytically and the bounds for ω and R_n are exact. However, this is not always possible in general. In the following two examples the function K is not given in an analytic form and the bounds for ω and R_n are first-order accurate.

Example 3 Consider the exponential function $f(t) = e^t$. In this case $f_s(t) = \cosh(t) - 1$ and $H(t) = \sqrt{t}(\cosh(t) - 1)$, $0 < t \le \frac{1}{2}$. Taylor expansion yields $K(t) = 2^{-1/5}t^{4/5}(1 + o(1))$ for small $t > 0$. Hence for $f(t) = e^t$,

$$2^{\frac{3}{5}}3^{-\frac{4}{5}}\epsilon^{\frac{4}{5}}(1 + o(1)) \le \omega(\epsilon, f) \le 2^{\frac{1}{5}}(\frac{5}{3})^{\frac{2}{5}}\epsilon^{\frac{4}{5}}(1 + o(1))$$

and

$$2^{\frac{14}{5}}3^{-\frac{18}{5}}n^{-\frac{4}{5}}(1 + o(1)) \le R_n(f) \le (\frac{10}{3e})^{\frac{4}{5}}n^{-\frac{4}{5}}(1 + o(1)).$$

Thus for $f(t) = e^t$, the difficulty of estimating $f(0)$ is of order $n^{-4/5}$.

Example 4 Consider the function $f(t) = -ctI(-\frac{1}{2} \le t \le 0) + t^r I(0 < t \le \frac{1}{2})$ where $r > 1$ and $c \ge 0$ are constants. Note that f is a nonsymmetric function and f' has a singularity at 0 when $c > 0$. It is easy to see that $f_s(t) = \frac{1}{2}(-ct + (-t)^r)I(-\frac{1}{2} \le t \le 0) + \frac{1}{2}(ct + t^r)I(0 < t \le \frac{1}{2})$ and $H(t) = \frac{c}{2}t^{3/2} + \frac{1}{2}t^{(2r+1)/2}$, $0 < t \le \frac{1}{2}$. Hence for small $t > 0$,

$$K(t) = \begin{cases} (\frac{c}{2})^{\frac{1}{3}}t^{\frac{2}{3}}(1 + o(1)) & \text{if } c > 0 \\ 2^{-\frac{1}{2r+1}}t^{\frac{2r}{2r+1}} & \text{if } c = 0 \end{cases}.$$

Consequently for $f(t) = -ctI(\frac{1}{2} \le t \le 0) + t^r I(0 < t \le \frac{1}{2})$,

$$(\frac{2c}{9})^{\frac{1}{3}}\epsilon^{\frac{2}{3}}(1 + o(1)) \le \omega(\epsilon, f) \le (\frac{5c}{3})^{\frac{1}{3}}\epsilon^{\frac{2}{3}}(1 + o(1)), \quad \text{if } c > 0$$

and

$$(\frac{2}{9})^{\frac{r}{2r+1}}\epsilon^{\frac{2r}{2r+1}} \le \omega(\epsilon, f) \le (\frac{5}{3})^{\frac{r}{2r+1}}\epsilon^{\frac{2r}{2r+1}}, \quad \text{if } c = 0.$$

Therefore $R_n(f)$ can be bounded as

$$(8c)^{\frac{2}{3}}3^{-\frac{10}{3}}n^{-\frac{2}{3}}(1+o(1)) \le R_n(f) \le (\frac{5c}{3e})^{\frac{2}{3}}n^{-\frac{2}{3}}(1+o(1)), \quad \text{if } c>0$$

and

$$2^{\frac{8r}{2r+1}}3^{-\frac{8r+2}{2r+1}}n^{-\frac{2r}{2r+1}} \le R_n(f) \le 2^{\frac{2r-2}{2r+1}}(\frac{5}{3e})^{\frac{2r}{2r+1}}n^{-\frac{2r}{2r+1}}, \quad \text{if } c=0.$$

It follows that the difficulty of estimating $f(0)$ is of order $n^{-2/3}$ when $c>0$ and of order $n^{-\frac{2r}{2r+1}}$ when $c=0$. The difference between the two cases is due to the singularity of f' at 0 when $c>0$.

These examples show that the difficulty of estimation as measured by $R_n(f)$ varies significantly from function to function under the convexity constraint depending on the local smoothness property of the convex function near the point of estimation. They also show that $R_n(f)$ can be easily bounded from below and above by computing the function K for any given $f \in F_c$.

3 An Oracle Risk and Superefficiency

Kernel estimators are one of the most widely used techniques in nonparametric function estimation. For such estimators the most important issue is that of bandwidth selection and the major difficulty is that the optimal bandwidth clearly depends on the unknown function f. In this paper we consider a particular kernel estimator namely a local average estimator for estimating the value $f(0)$ and in the present section we focus on understanding the smallest risk of such a procedure as the function varies over the class of convex functions.

This oracle risk is perhaps of interest in its own right as an alternative benchmark for particular procedures. However for us its importance lies as a way to connect the performance of the particular estimator introduced in Section 1.3 to the benchmarks of Section 2. This oracle risk is also instrumental in understanding the concept of superefficiency in the context of estimating a convex function. Although it is possible to beat the benchmarks for a particular convex function, such superefficient estimators must pay a penalty at other convex functions.

Consider the class of procedures $\hat{L}_a$ with

$$\hat{L}_a = \frac{1}{2a}\int_{-a}^{a} dY(t) \tag{22}$$

which correspond to the local averages over the subintervals $[-a,a]$ for $0 < a \le \frac{1}{2}$. Then it is easy to see that the mean squared error of $\hat{L}_a$ is given by

$$F_n(a) = (\frac{1}{a}\int_0^a f_s(t)dt)^2 + \frac{1}{2an}, \tag{23}$$

where again f_s, given in (13), is the symmetrized and centered version of the convex function f. Then the risk of the ideal local average is defined by

$$r_n(f) = \inf_{0<a\le\frac{1}{2}} F_n(a). \tag{24}$$

It is clear that $r_n(f)$ represents an ideal target for the class of local average estimators. For a given convex function f, $r_n(f)$ is easy to calculate exactly or first-order accurately. Let us consider again the examples discussed in Section 2.2.

Example 5 For $f(t) = ct$, it is easy to see that $r_n(f) = n^{-1}$. For $f(t) = |t|^r$ with $r \geq 1$, straightforward calculus yields that

$$r_n(f) = (2r+1)(r+1)^{-\frac{2}{2r+1}}(4r)^{-\frac{2r}{2r+1}}n^{-\frac{2r}{2r+1}}.$$

For $f(t) = e^t$, by expanding the function $f_s(t)$ it is not difficult to show that

$$r_n(f) = 2^{-\frac{14}{5}}3^{-\frac{2}{5}}5n^{-\frac{4}{5}}(1+o(1)).$$

Similarly it can be shown that for $f(t) = -ctI(\frac{1}{2} \leq t \leq 0) + t^r I(0 < t \leq \frac{1}{2})$ where $r > 1$ and $c \geq 0$ are constants,

$$r_n(f) = \begin{cases} 2^{-8/3}3c^{\frac{2}{3}}n^{-\frac{2}{3}}(1+o(1)) & \text{if } c > 0 \\ \frac{1}{4}(2r+1)(r+1)^{-\frac{2}{2r+1}}r^{-\frac{2r}{2r+1}}n^{-\frac{2r}{2r+1}} & \text{if } c = 0 \end{cases}.$$

The quantity $r_n(f)$ plays an important role in our analysis. It links the performance of the data-driven estimator $\hat{T}_*$ at a given convex function f with the benchmark $R_n(f)$. Similar to the local modulus of continuity $\omega(\epsilon, f)$, the oracle risk $r_n(f)$ can also be upper and lower bounded in terms of the function K.

Proposition 2 *For any convex function $f \in F_c$,*

$$\frac{9}{16}K^2(\frac{2}{3\sqrt{n}}) \leq r_n(f) \leq \frac{11}{8}K^2(\frac{2}{3\sqrt{n}}). \tag{25}$$

It is sometimes useful to rewrite the bounds in this lemma by using the inequality $C^{2/3}K(t) \leq K(Ct) \leq CK(t)$ given in Lemma 4 in Section 7 which then gives $\frac{1}{4}K^2(\frac{1}{\sqrt{n}}) \leq r_n(f) \leq \frac{11}{8}K^2(\frac{1}{\sqrt{n}})$ as a bound on $r_n(f)$. The bounds

$$\frac{1}{12}\omega^2(\frac{1}{\sqrt{n}}, f) \leq r_n(f) \leq \frac{99}{32}\omega^2(\frac{1}{\sqrt{n}}, f) \tag{26}$$

on the oracle risk $r_n(f)$ in terms of the local modulus and the bounds

$$\frac{32}{891}r_n(f) \leq R_n(f) \leq \frac{6}{e}r_n(f) \tag{27}$$

on the benchmark $R_n(f)$ in terms of the oracle risk then follow from Theorem 1 and Proposition 1.

The oracle risk $r_n(f)$ is also useful for understanding the third point made in section 1.2, namely that outperforming the benchmark $R_n(f)$ at some convex function must result in worse performance at other functions. It is convenient to do this in two steps. First the following result shows that there are consequences for outperforming the oracle risk $r_n(f)$ at a particular function.

Theorem 2 *Let $\hat{T}_n$ be an estimate of $f(0)$. Suppose that $R(\hat{T}_n, f) \leq cr_n(f)$ where $0 < c < 1$. Then there is another convex function h such that*

$$\frac{R(\hat{T}_n, h)}{r_n(h)} \geq \frac{8}{9}(1 - \frac{1}{0.88(\ln(1/c))^{1/3}})^2 \ln\frac{1}{c}. \tag{28}$$

This shows that if an estimator significantly outperforms the oracle risk $r_n(f)$ at any particular convex function f, then there must be another convex function h at which the estimator performs poorly. Then Theorem 2 and (27) together yield directly the following corollary which expresses the consequence of superefficiency in terms of the benchmark $R_n(f)$.

Corollary 1 *There is a constant $d > 0$ such that, if $\hat{T}_n$ is an estimate of $f(0)$ with $R(\hat{T}_n, f) \leq cR_n(f)$ for some convex function f where $0 < c < 1$, then there is another convex function h with*

$$\frac{R(\hat{T}_n, h)}{R_n(h)} \geq d \ln \frac{1}{c}. \tag{29}$$

4 Analysis of the Estimation Procedure

We now turn to an analysis of the estimator $\hat{T}_*$ given in (12). The estimate $\hat{T}_*$ is a local average estimate with a data-driven bandwidth selector. We shall show that the estimator $\hat{T}_*$ performs well for every convex function f in the sense that its mean squared error is uniformly within a constant factor of $R_n(f)$ for all $f \in F_c$ and all n. As a direct consequence a standard adaptive minimaxity result then follows.

The procedure $\hat{T}_*$ is strongly related to the oracle procedure of the last section. First a bandwidth is chosen from a dyadic sub-collection of these procedures and then a local average is performed using that bandwidth. The key is to find a data driven bandwidth that optimally trades bias and variance in the sense that a larger bandwidth will inflate the squared bias much more than it reduces the variance and a smaller bandwidth will increase the variance more than it reduces the squared bias. A critical step is to accurately learn the bias of the local average estimate for individual convex functions. The ability to learn the bias of an estimator is a key special feature in problems with shape constraints. This feature makes it possible to adapt to individual functions.

Denote the standard deviation of δ_j defined in (10) by σ_j. Then it is easy to see that

$$\sigma_j = \frac{2^{(j-1)/2}}{\sqrt{n}}. \tag{30}$$

Within the dyadic class of local average estimators $\{\delta_j\}$, one would like to use the ideal "estimator" δ_{j_*} where

$$j_* = \arg\min_j E(\delta_j - f(0))^2 = \arg\min_j \left\{ \text{Bias}^2(\delta_j) + \sigma_j^2 \right\}. \tag{31}$$

This dyadic class is rich in the sense that the mean squared error performance of δ_{j_*} is within a factor of 2 of the oracle risk $r_n(f)$ as given in the follow lemma.

Lemma 1 *The bias of the oracle estimator δ_{j_*} satisfies*

$$|E\delta_{j_*} - f(0)| \leq \frac{2}{\sqrt{3}} \sigma_{j_*} \tag{32}$$

and the risk of δ_{j_*} is within a factor of 2 of $r_n(f)$, i.e.,

$$\frac{E(\delta_{j_*} - f(0))^2}{r_n(f)} \leq 2. \tag{33}$$

Our goal is to select a data-driven estimator δ_j which mimics the oracle estimator δ_{j_*} and hence also that of the best local average estimator.

Recall that $B_j(t)$ and δ_j are respectively the scaled box kernel and local average estimator of $f(0)$ as defined in Section 1.3. Write $\bar{f}_j = \int f(t)B_j(t)dt$ for the average of the function f over the interval $[-2^{-j}, 2^{-j}]$. Then $E\delta_j = \bar{f}_j$. To learn the bias of δ_j, we introduce

$$T_j = \delta_j - \delta_{j+1}. \tag{34}$$

Note that the T_j's are independent normal variables with $Var(T_j) = \sigma_j^2 = \frac{2^{j-1}}{n}$. The following lemma summarizes important properties of δ_j and T_j.

Lemma 2 *For any convex function f,*

$$0 \le \text{Bias}(\delta_{j+1}) \le \frac{1}{2}\text{Bias}(\delta_j) \tag{35}$$

$$0 \le ET_{j+1} \le \frac{1}{2}ET_j \tag{36}$$

$$\frac{1}{2}\text{Bias}(\delta_j) \le ET_j \le \text{Bias}(\delta_j) \tag{37}$$

and the constant factor $\frac{1}{2}$ on the right hand sides of (35)-(36) is sharp in both cases.

Lemma 2 shows that the bias of the estimator δ_j is always nonnegative and as j increases the bias decreases at least by a factor of 2 each time while the standard deviation of δ_j increases by a factor of $\sqrt{2}$. Furthermore, (37) shows that T_j can be used to accurately learn the bias of δ_j.

Our goal is to choose δ_j which nearly optimally trades bias and variance for individual functions. This can be achieved by comparing the value of T_j with the standard deviation of δ_j. For the collection of those T_j which are within a small constant factor of its standard deviation it is most natural to choose the smallest as the corresponding estimator δ_j will have the smallest standard deviation within this collection. More specifically for some constant $\lambda > 0$, set

$$\hat{j} = \inf_k\{k : T_k \le \lambda\sigma_k\} \tag{38}$$

and then

$$\hat{T}_* = \hat{f}(0) = \delta_{\hat{j}}. \tag{39}$$

One of the main results of the present paper is that the mean squared error performance of the data-driven estimator $\hat{T}_*$ comes within a small constant factor of that of the ideal estimator δ_{j_*} and so consequently also within a constant factor of the benchmark $R_n(f)$ for all $f \in F_c$ and all n.

Theorem 3 *Let the estimator $\hat{T}_*$ be given as in (12) with $\lambda = \sqrt{2}$. Then*

$$E(\hat{T}_* - f(0))^2 \le 6 \inf_j E(\delta_j - f(0))^2 \tag{40}$$

for all convex function $f \in F_c$. Consequently,

$$E(\hat{T}_* - f(0))^2 \le 12r_n(f) \le CR_n(f) \tag{41}$$

for all convex function $f \in F_c$, where $C > 0$ is an absolute constant.

Remark 2 It is possible to improve the constants 6 and 12 in this Theorem with more refined calculations. Furthermore λ can be chosen to be any sufficiently large constant and the resulting estimate has the same properties but with different constants.

The adaptation of the estimator $\hat{T}_*$ defined in (12) to each convex function as measured by $R_n(f)$ differs from the usual minimax adaptive estimation statements usually made in the nonparametric function estimation literature. The more typical adaptive minimaxity results do however follow quite easily. Consider, for example, $F_c(\alpha, M) = F_c \cap \Lambda(\alpha, M)$. It can be shown that the estimator $\hat{T}_*$ adaptively attains within a constant factor of the minimax risk over each parameter space $F_c(\alpha, M)$ for $1 \leq \alpha \leq 2$.

Theorem 4 *Under the white noise model (1), the data-driven estimator $\hat{T}_*$ defined in (12) satisfies for some absolute constants $C, c > 0$*

$$\sup_{f \in F_c(\alpha,M)} E(\hat{T}_* - f(0))^2 \leq CM^{\frac{2}{2\alpha+1}} n^{-\frac{2\alpha}{2\alpha+1}} \leq cR_*(n, F_c(\alpha, M)) \tag{42}$$

for all $1 \leq \alpha \leq 2$, all $M \geq 1$ and all $n > 1$, where

$$R_*(n, F_c(\alpha, M)) = \inf_{\hat{T}_n} \sup_{f \in F_c(\alpha,M)} E(\hat{T}_n - f(0))^2$$

is the minimax risk over $F_c(\alpha, M)$.

Adaptation over the range of $1 \leq \alpha \leq 2$ is reminiscent of a similar adaptive rate result found in Dümbgen and Rufibach (2009) in the context of estimating a log-concave density. The restriction to the range $1 \leq \alpha \leq 2$ in Theorem 4 is actually necessary. Extensions to $\alpha > 2$ are ruled out by Theorem 2. More specifically, the following result holds.

Proposition 3 *If an estimator $\hat{T}_n$ attains the optimal rate of convergence over $F_c(\alpha, M)$ for some $\alpha > 2$ and $M > 0$, i.e.,*

$$\sup_{f \in F_c(\alpha,M)} E(\hat{T}_n - f(0))^2 \leq CM^{\frac{2}{2\alpha+1}} n^{-\frac{2\alpha}{2\alpha+1}} \tag{43}$$

for some constant $C > 0$, then there exists a constant $c > 0$ such that

$$\sup_{f \in F_c(2,M)} E(\hat{T}_n - f(0))^2 \geq cM^{\frac{2}{5}} \left(\frac{\log n}{n}\right)^{\frac{4}{5}}. \tag{44}$$

Hence, there is no estimator $\hat{T}_n$ such that for some constant $C > 0$ not depending on n and M

$$\sup_{f \in F_c(\alpha,M)} E(\hat{T}_n - f(0))^2 \leq CM^{\frac{2}{2\alpha+1}} n^{-\frac{2\alpha}{2\alpha+1}} \tag{45}$$

for all $\alpha \geq 1$, $M \geq 1$ and $n > 1$.

5 Convex Regression

We have so far focused on the white noise model. The framework introduced in Section 2 can also be applied to nonparametric regression where the regression function is assumed to be convex. In

this setting the estimation procedure only requires slight modifications. Consider the regression model

$$y_i = f(x_i) + \sigma z_i, \quad i = -n, -(n-1), ..., -1, 0, 1, ..., n \tag{46}$$

where $x_i = \frac{i}{2n}$ and $z_i \overset{iid}{\sim} N(0,1)$ and where for notational convenience we index the observations from $-n$ to n. The noise level σ can be accurately estimated easily and we shall assume it is known in this section. Under the assumption that f is convex, we wish to estimate $f(0)$.

Let J be the largest integer less than or equal to $\log_2 n$. For $1 \le j \le J$ define the local average estimators

$$\bar{\delta}_j = 2^{-j} \sum_{k=1}^{2^{j-1}} (y_{-k} + y_k). \tag{47}$$

As in the white noise model, we shall build a sequence of independent tests to empirically choose an optimal bandwidth using

$$T_j = \bar{\delta}_j - \bar{\delta}_{j-1}, \tag{48}$$

and then select the corresponding $\bar{\delta}_j$ as an estimator of $f(0)$. Note that the T_j's are independent normal variables with $\mathrm{Var}(T_j) = \mathrm{Var}(\bar{\delta}_j) = \sigma^2 2^{-j}$. Set

$$\hat{j} = \max_j \{j : T_j \le \sqrt{2}\sigma 2^{-j/2}\} \tag{49}$$

and define the estimator of $f(0)$ by

$$\hat{T}_* = \hat{f}(0) = \bar{\delta}_{\hat{j}}. \tag{50}$$

The properties of the data-driven estimator $\hat{T}_*$ can be analyzed in the same way as before. The major difference in the analysis for the regression case is in the details of the properties of the T_j and $\bar{\delta}_j$ as summarized in the following lemma.

Lemma 3 *For any convex function f,*

$$2ET_{j-1} \le ET_j \tag{51}$$

$$\mathrm{Bias}(\bar{\delta}_{j-1}) \le \frac{2^{j-2}+1}{2^{j-1}+1}\mathrm{Bias}(\bar{\delta}_j) \tag{52}$$

$$\frac{2^{j-2}}{2^{j-1}+1}\mathrm{Bias}(\bar{\delta}_j) \le ET_j \le \mathrm{Bias}(\bar{\delta}_j). \tag{53}$$

The remaining analysis easily yields the near-optimality of the corresponding estimator $\hat{T}_*$ for all convex functions.

Theorem 5 *For some constants $C_3 > C_2 > C_1 > 0$,*

$$E(\hat{T}_* - f(0))^2 \le C_1 \inf_j E(\delta_j - f(0))^2 \le C_2 r_n(f) \le C_3 R_n(f) \tag{54}$$

for all $f \in F_c$.

6 Discussion

The framework introduced in the present paper for estimating convex functions extends to other settings of shape constrained inference such as estimating monotone functions. A key to any such analysis is the construction of good estimates of bias which then allows for the selection of an estimator which optimally trades bias and variance for every given function with the shape constraint.

The theory can also be extended to the construction and evaluation of the performance of pointwise confidence intervals. For example, under the convexity constraint, a benchmark for the expected length of a confidence interval, similar to $R_n(f)$ for mean squared error, can be developed for each convex function f. Then a data-driven confidence interval can be constructed and shown to be adaptive to every convex function f in the sense that it has the shortest expected length (up to an absolute constant factor) for f among all confidence intervals which have prespecified level of coverage over the collection of convex functions. Such adaptivity is much stronger than the conventional adaptive minimaxity over a collection of smoothness classes as considered, for example, in Cai and Low (2004). We shall report the details of these results elsewhere.

Although this paper constructed a particular estimate which nearly attains the bound $R_n(f)$ for all convex functions, a major goal of the present work is to frame the discussion and create targets for the evaluation of other estimators such as least square or maximum likelihood estimators, a goal that lies outside the scope of the present work. We should however mention that the use of the simple box kernel in Section 4 for the construction of the estimation procedure can easily be extended to other kernels although the analysis is sometimes more involved. For example, similar results hold for the quadratic kernel

$$Q(t) = \frac{3}{4}(1 - t^2)I(|t| \le 1).$$

We should emphasize that the framework we developed in this paper does not work without shape constraints, although in principle $R_n(f)$ can still be evaluated in those settings. Under the usual smoothness conditions without shape constraints, the benchmark $R_n(f)$ is almost the same for all functions f and thus of the same order as the minimax risk. This makes the framework uninteresting and not useful in that setting. The reason is that the usual smoothness class is too "rich" which makes all the functions nearly equally difficult to estimate against their respective hardest local alternative. Consider, for example, the Lipschitz ball $\Lambda(\alpha, M)$. It is not difficult to check that for any $f \in \Lambda(\alpha, M_1)$ with $0 < M_1 < M$,

$$R_n(f) \asymp n^{-2\alpha/(2\alpha+1)}$$

which is of the same order as the minimax risk over $\Lambda(\alpha, M)$ given in (3).

Finally we should note that if the estimator developed in this paper is applied at each point the resulting estimate of the entire function need not be convex.

7 Proofs

In this section we shall first prove the propositions and lemmas given in the earlier sections before proving the main theorems. The proof of Theorem 5 is omitted as it is entirely analogous

to the proof of Theorem 3 by using Lemma 3. The proofs of some of the main results rely on a few additional technical results. We first collect these technical results in Section 7.1 and prove them in the supplement, Cai and Low (2013).

We shall use λ to denote the constant used in the test given in (11) that selects $\hat{j}$, but we always take λ to be $\sqrt{2}$ throughout this section.

7.1 Preparatory Technical Results

We state in this section the additional technical results that are used in the proofs of the main results. The proofs of these lemmas are given in the supplement, Cai and Low (2013).

The following lemma gives further characterizations of the functions H and K introduced in Section 2.1.

Lemma 4 *The function H^{-1} defined in (15) is concave and nondecreasing. It is strictly increasing for all x where $H^{-1}(x) < \frac{1}{2}$. Moreover for $C \geq 1$ it satisfies*

$$H^{-1}(Ct) \leq C^{\frac{2}{3}} H^{-1}(t). \tag{55}$$

The function K defined in (18) is also increasing and satisfies for $C \geq 1$

$$C^{\frac{2}{3}} K(t) \leq K(Ct) \leq CK(t). \tag{56}$$

We shall use the next two lemmas to study the properties of the local modulus of continuity $\omega(\epsilon, f)$ and the local average oracle risk $r_n(f)$.

Lemma 5 *Let f be a nonnegative convex function on $[-\frac{1}{2}, \frac{1}{2}]$. For $d > 0$ let t_* be the supremum over all y with $f_s(y) \leq d$ where f_s defined in (13) is the symmetrized and centered version of f. Then there is a convex function g with $g(0) - f(0) = d$ and for which*

$$\int_{-\frac{1}{2}}^{\frac{1}{2}} (g(x) - f(x))^2 dx \leq \frac{9}{4} d^2 t_*. \tag{57}$$

It follows that for each $0 \leq u \leq \frac{1}{2}$ there is a convex function g with $g(0) - f(0) = f_s(u)$ such that

$$\int_{-\frac{1}{2}}^{\frac{1}{2}} (g(x) - f(x))^2 dx \leq \frac{9}{4} H^2(u) \tag{58}$$

where the function H is defined in (14). Moreover for any convex h with $h(0) - f(0) = d > 0$

$$\int_{-\frac{1}{2}}^{\frac{1}{2}} (h(x) - f(x))^2 dx \geq \frac{2}{3} d^2 t_*. \tag{59}$$

Remark: The constants $\frac{9}{4}$ and $\frac{2}{3}$ in (57) and (59) are sharp.

Lemma 6 *Let f and g be convex functions with $f(0) - g(0) = a > 0$. Let t be the supremum of all y for which $f_s(y) \leq a$. Then*

$$\int_{-1/2}^{1/2} (f(x) - g(x))^2 dx \geq 0.3ta^2. \tag{60}$$

Lemma 7 below provides the bounds for the bias of δ_j and the mean of T_j for $j \le j_*$. The lemma is useful for the proof of Theorem 3.

Lemma 7 *Set* $\sigma_{j_*} = \frac{2^{(j_*-1)/2}}{\sqrt{n}}$. *Let* j_* *be defined as in (31), then*

$$ET_{j_*} \le \min(E\delta_{j_*} - f(0), \sigma_{j_*}). \tag{61}$$

For $k \ge 1$,

$$E\delta_{j_*-k} - f(0) \ge 2^{k-\frac{3}{2}}\sigma_{j_*} \tag{62}$$

and

$$ET_{j_*-k} \ge 2^{k-1}\frac{1}{\sqrt{6}}\sigma_{j_*}. \tag{63}$$

We shall use Lemma 8 in the proof of Theorem 2 on the consequence of superefficiency.

Lemma 8 *For* $b > 0$ *let* t_b *be the supremum over all* t *where* $f_s(t) \le br_n^{\frac{1}{2}}(f)$. *Then*

$$t_b \le \frac{2}{4-b^2}\frac{1}{nr_n(f)} \tag{64}$$

and for $b \ge \frac{2}{\sqrt{3}}$,

$$t_b \le \frac{b3\sqrt{3}}{8nr_n(f)}. \tag{65}$$

The final two lemmas are purely on some numerical results. They are not of direct interest but are useful in the proof of Theorem 3.

Lemma 9 *Let* $\lambda = \sqrt{2}$ *and let* $h(x)$ *be the function given by*

$$\begin{aligned} h(x) &= P(Z \le \lambda - \frac{x}{2}) + \frac{0.649}{1+x^2} + \frac{1}{4}\frac{x^2}{1+x^2} \\ &\quad + \frac{1}{1+x^2}\sum_{m=1}^{\infty}(2^m\sqrt{3} + 2^{-m/2}2x)\left(P(Z \le \lambda)\prod_{l=0}^{m-1}P(Z > \lambda - 2^{-3l/2}\min(x,1))\right)^{1/2}. \end{aligned}$$

Then

$$\sup_{0 \le x \le \frac{2}{\sqrt{3}}} h(x) \le 4.7.$$

Lemma 10 *Let* $g_m(x,y) = (x^2 + 2^{-m})P(Z \le \lambda - 2^{m/2}(x-y))$. *Then for* $m \ge 2$ *and* $y \ge 2^{m-3/2}$

$$\sup_{x \ge 2y} g_m(x,y) = (4y^2 + 2^{-m})P(Z \le \lambda - 2^{m/2}y) \tag{66}$$

$$\sup_{x \ge 2y, y \ge 2^{m-3/2}} g_m(x,y) = (2^{2m-1} + 2^{-m})P(Z \ge 2^{3(m-1)/2} - \lambda). \tag{67}$$

Moreover

$$\sup_{x \ge 2y, y \ge \sqrt{2}} g_2(x,y) \le 0.649 \tag{68}$$

$$\sup_{x \ge \max(\frac{1}{\sqrt{2}}, 2y), y \ge 0} \frac{g_1(x,y)}{1+y^2} \le 1.2. \tag{69}$$

7.2 Proof of Propositions and Lemmas

With the additional technical results given in Section 7.1, we are now ready to prove the results given in the previous sections. We shall first prove the propositions and lemmas as some of these are needed in the proofs of the main theorems.

7.2.1 Proof of Proposition 1

We shall first prove $\omega(\epsilon, f) \geq K(\frac{2}{3}\epsilon)$ by considering two cases, $H^{-1}(\frac{2}{3}\epsilon) = \frac{1}{2}$ and $H^{-1}(\frac{2}{3}\epsilon) < \frac{1}{2}$. When $H^{-1}(\frac{2}{3}\epsilon) = \frac{1}{2}$ it follows that $K(\frac{2}{3}\epsilon) = \frac{2\sqrt{2}}{3}\epsilon$. Since for any convex function f, the function $g = f + \epsilon$ is also convex with $g(0) - f(0) = \epsilon$ and $\int_{-1/2}^{1/2}(g(x) - f(x))^2 dx = \epsilon^2$. It immediately follows that $\omega(\epsilon, f) \geq \epsilon$ and since $\epsilon > \frac{2\sqrt{2}}{3}\epsilon$ the inequality $\omega(\epsilon, f) \geq K(\frac{2}{3}\epsilon)$ clearly holds in this case.

In the other case $H^{-1}(\frac{2}{3}\epsilon) < 1/2$, then we have from (17) that $H(H^{-1}(\frac{2}{3}\epsilon)) = \frac{2}{3}\epsilon$ and hence from the definition of K in (18) that $f_s(H^{-1}(\frac{2}{3}\epsilon)) = K(\frac{2}{3}\epsilon)$. It then follows from equation (58) of Lemma 5 that there is a convex function g with $g(0) - f(0) = K(\frac{2}{3}\epsilon)$ such that

$$\int_{-1/2}^{1/2}(g(x) - f(x))^2 dx \leq \frac{9}{4}K^2(\frac{2}{3}\epsilon)H^{-1}(\frac{2}{3}\epsilon) = \epsilon^2.$$

Once again it follows that $\omega(\epsilon, f) \geq K(\frac{2}{3}\epsilon)$.

We now turn to the proof of $\omega(\epsilon, f) \leq K(\sqrt{\frac{10}{3}}\epsilon)$. Let g be any convex function such that $|g(0) - f(0)| = K(\sqrt{\frac{10}{3}}\epsilon)$ and let $t_1 = H^{-1}(\sqrt{\frac{10}{3}}\epsilon)$. Then by equation (59) of Lemma 5 and equation (60) of Lemma 6 it follows that

$$\int_{-1/2}^{1/2}(g(x) - f(x))^2 dx \geq \frac{3}{10}K^2(\sqrt{\frac{10}{3}}\epsilon)H^{-1}(\sqrt{\frac{10}{3}}\epsilon) = \epsilon^2. \quad \blacksquare$$

7.2.2 Proof of Proposition 2

Note that for any convex function f, $f_s(x) = \frac{f(x)+f(-x)}{2} - f(0)$ is convex with $f_s(0) = 0$. It follows that for any $0 < t_1 \leq \frac{1}{2}$, $0 \leq f_s(x) \leq x\frac{f_s(t_1)}{t_1}$ for $0 \leq x \leq t_1$ and hence

$$(\frac{1}{t_1}\int_0^{t_1} f_s(x)dx)^2 \leq (\frac{1}{t_1}\int_0^{t_1} x\frac{f_s(t_1)}{t_1}dx)^2 = \frac{f_s^2(t_1)}{4}.$$

Hence from (23) and (24) it follows that for any $0 < t_1 \leq \frac{1}{2}$ we have $r_n(f) \leq \frac{f_s^2(t_1)}{4} + \frac{1}{2t_1 n}$. Now setting $t_1 = H^{-1}(\frac{2}{3\sqrt{n}})$ and noting that $\sqrt{t_1}f_s(t_1) = H(t_1) \leq \frac{2}{3\sqrt{n}}$ yields

$$r_n(f) \leq \frac{11}{18nH^{-1}(\frac{2}{3\sqrt{n}})} = \frac{11}{8}K^2(\frac{2}{3\sqrt{n}})$$

and thus the right hand side of the inequality in equation (25) holds.

In order to show that the left hand side of equation (25) holds it is convenient to consider separately the cases $H^{-1}(\frac{2}{3\sqrt{n}}) \geq \frac{1}{4}$ and the case $H^{-1}(\frac{2}{3\sqrt{n}}) < \frac{1}{4}$. In the case of $H^{-1}(\frac{2}{3\sqrt{n}}) \geq \frac{1}{4}$ note that $\frac{1}{4nH^{-1}(\frac{2}{3\sqrt{n}})} \leq \frac{1}{n}$ and since $r_n(f) \geq \frac{1}{n}$ the left hand side of equation (25) clearly holds in this case.

Hence we may now assume that $H^{-1}(\frac{2}{3\sqrt{n}}) < \frac{1}{4}$. In this case it is helpful to note that $(\frac{1}{a}\int_0^a f_s(t)dt)^2$ is a nondecreasing function of a and that $\frac{1}{2an}$ is strictly decreasing in a. It then follows from (23) and (24) that for any $0 < x_1 \leq \frac{1}{2}$,

$$r_n(f) \geq \min\{\frac{1}{2nx_1}, (\frac{1}{x_1}\int_0^{x_1} f_s(t)dt)^2\}. \tag{70}$$

Now let $x_1 = 2H^{-1}(\frac{2}{3\sqrt{n}})$ and note that by assumption $x_1 \leq \frac{1}{2}$. Note also that

$$\frac{1}{2nx_1} = \frac{1}{4nH^{-1}(\frac{2}{3\sqrt{n}})} = \frac{9}{16}K^2(\frac{2}{3\sqrt{n}}) \tag{71}$$

Note that $f_s(\frac{x_1}{2}) = f_s(H^{-1}(\frac{2}{3\sqrt{n}})) = K(\frac{2}{3\sqrt{n}})$. Note for $\frac{x_1}{2} \leq t \leq x_1$ it follows from the convexity of f_s that $f_s(t) \geq f_s(\frac{x_1}{2})\frac{2t}{x_1} = K(\frac{2}{3\sqrt{n}})\frac{2t}{x_1}$. It follows that

$$(\frac{1}{x_1}\int_0^{x_1} f_s(t)dt)^2 \geq (\frac{1}{x_1}\int_{\frac{x_1}{2}}^{x_1} \frac{2t}{x_1} f_s(\frac{x_1}{2})dt)^2 = \frac{9}{16}K^2(\frac{2}{3\sqrt{n}}) \tag{72}$$

Now taken together (70), (71) and (72) yield $r_n(f) \geq \frac{9}{16}K^2(\frac{2}{3\sqrt{n}})$ showing that the left hand side of the inequality in (25) also holds in this case. ∎

We shall now prove Lemma 2 before we prove Lemma 1 as the proof of this later lemma depends on the results in Lemma 2.

7.2.3 Proof of Lemma 2

As in the previous lemma for any convex function f we set $f_s(t) = \frac{f(t)+f(-t)}{2} - f(0)$. Now note that $f_s(tx)$ is convex in x for all $0 \leq t \leq 1$. Hence $g(x) = \int_0^1 f_s(tx)dt$ is also convex with $g(0) = 0$. For $x > 0$ set $z = xt$ and it follows that $g(x) = \frac{1}{x}\int_0^x f_s(z)dz = \frac{1}{2x}\int_{-x}^x (f(z) - f(0))dz$. Equations (35), (36) and (37) follow immediately from the convexity of g. For example (35) is equivalent to $g(x) \leq \frac{1}{2}g(2x)$ for $x = 2^{-(j+1)}$. ∎

7.2.4 Proof of Lemma 1

First note that from equation (35) of Lemma 2, $(E\delta_{j+1} - f(0))^2 \leq \frac{1}{4}(E\delta_j - f(0))^2$ and hence

$$E(\delta_{j+1} - f(0))^2 - E(\delta_j - f(0))^2 \leq -\frac{3}{4}(E\delta_j - f(0))^2 + \frac{2^{j-1}}{n} < 0$$

whenever $(E\delta_j - f(0))^2 > \frac{4}{3}\frac{2^{j-1}}{n}$. Equation (32) then immediately follows.

We now turn to a proof of (33).

Recall that $F_n(a)$ defined by equation (23) gives the risk of a local average estimator. It can be written as $F_n(a) = g^2(a) + \frac{1}{2an}$ where $g(x) = \frac{1}{2x}\int_{-x}^x f_s(t)dt$. In the proof of lemma 2 we showed that g is convex and hence g^2 is convex as well. Moreover F_n is convex since it is a sum of two convex functions. Now since F_n is convex on $(0, \frac{1}{2}]$ and since $\lim_{x\to 0^+} F_n(x) = \infty$ it follows that there is a point a_* with $0 < a_* \leq \frac{1}{2}$ such that $F_n(a_*) = r_n(f)$ and for this a_* $E(\hat{L}_{a_*} - f(0))^2 = r_n(f)$. It also follows that for $0 < x \leq a_*$ the function F_n is nonincreasing and for $a_* \leq x \leq \frac{1}{2}$ the function F_n is nondecreasing. Hence either $2^{-j_*} \leq a_* \leq 2^{-(j_*-1)}$ or

$2^{-(j_*+1)} \le a_* \le 2^{-j_*}$. If $2^{-j_*} \le a_* \le 2^{-(j_*-1)}$ then $(E\hat{L}_{a_*} - f(0))^2 \ge (E\delta_{j_*} - f(0))^2$ whereas $Var(\delta_{j_*}) \le 2Var(\hat{L}_{a_*})$. It immediately follows that

$$\frac{E(\delta_{j_*} - f(0))^2}{r_n(f)} \le 2$$

and in this case (33) follows.

On the other hand if $2^{-(j_*+1)} \le a_* \le 2^{-j_*}$ it follows that

$$(E\hat{L}_{a_*} - f(0))^2 \ge (E\delta_{j_*+1} - f(0))^2$$

whereas $Var(\delta_{j_*+1}) \le 2Var(\hat{L}_{a_*})$. It immediately follows that $\frac{E(\delta_{j_*+1}-f(0))^2}{r_n(f)} \le 2$ and hence also that

$$\frac{E(\delta_{j_*} - f(0))^2}{r_n(f)} \le 2$$

and in this case (33) also follows. ∎

7.3 Proof of Proposition 3

This proposition is a consequence of a constrained risk inequality introduced in Brown and Low (1996b) applied to two carefully chosen functions. Adopting the notation of Brown and Low (1996b), the chi-square distance between two white noise with drift models as given in (1), one with drift f and the other with drift g is given by $I_n(f,g) - 1$ where

$$I_n(f,g) = \exp(n\int_{-1/2}^{1/2}(g(t) - f(t))^2 dt). \tag{73}$$

Without loss of generality take $M = 2$ and take $f(t) = \frac{t^2}{2}$ as one of the two functions. Clearly $f \in F_c(\alpha, 2)$ for all $\alpha \ge 2$. We shall suppose that for this function equation (43) holds. In addition consider the one parameter family of functions

$$h_a(t) = \begin{cases} t^2 - \frac{a^2}{4} & \text{if } |t| \le \frac{a}{2} \\ at - \frac{a^2}{2} & \text{if } \frac{a}{2} \le |t| \le a \\ \frac{t^2}{2} & \text{if } a < |t| \le \frac{1}{2} \end{cases}.$$

Note that $h_a \in F_c(2,2)$ and that the L_2 distance between h_a and f is given by

$$\int_{-1/2}^{1/2}(h_a(t) - f(t))^2 dt = 2\int_0^{a/2}(\frac{a^2}{4} - \frac{t^2}{2})^2 dt + 2\int_{a/2}^{a}(\frac{t^2}{2} - at + \frac{a^2}{2})^2 dt = \frac{23}{480}a^5.$$

Now take $a = a_n = (d\frac{\ln n}{n})^{1/5}$ for some constant $d > 0$. Then $I_n(f, h_a) = n^{\frac{23}{480}d}$. If equation (43) holds for $\alpha = \alpha_0 > 2$, i.e.,

$$E_f(\hat{T}_n - f(0))^2 \le Cn^{-\frac{2\alpha_0}{2\alpha_0+1}},$$

then it follows from equation (2.3) in Brown and Low (1996b) that

$$E_{h_{a_n}}(\hat{T}_n - h_{a_n})^2 \ge (\max\{a_n^2 - \sqrt{C}n^{-\frac{\alpha_0}{2\alpha_0+1}}n^{\frac{23}{960}d}, 0\})^2 \ge \frac{1}{2}(d\frac{\ln n}{n})^{4/5}$$

as long as d is sufficiently small and n is large. Equation (44) then clearly follows. Equation (45) is an immediate consequence of (44). ∎

7.3.1 Proof of Lemma 3

As before for any convex function f, let $f_s(x) = \frac{1}{2}(f(x) + f(-x)) - f(0)$. Then $f_s(x)$ is convex, increasing in $|x|$ and $f_s(0) = 0$. Convexity of f_s yields that for $0 < x \leq y$

$$\frac{f_s(x)}{x} \leq \frac{f_s(y)}{y}. \tag{74}$$

Note that $E\delta_j = 2^{-(j-1)} \sum_{k=1}^{2^{j-1}} f_s(\frac{k}{n})$ and

$$ET_j = 2^{-(j-1)} \left\{ \sum_{k=2^{j-2}+1}^{2^{j-1}} f_s(\frac{k}{n}) - \sum_{k=1}^{2^{j-2}} f_s(\frac{k}{n}) \right\}.$$

So $ET_j \geq 2ET_{j-1}$ is equivalent to

$$\sum_{k=2^{j-2}+1}^{2^{j-1}} f_s(\frac{k}{n}) - \sum_{k=1}^{2^{j-2}} f_s(\frac{k}{n}) \geq 4 \sum_{k=2^{j-3}+1}^{2^{j-2}} f_s(\frac{k}{n}) - 4 \sum_{k=1}^{2^{j-3}} f_s(\frac{k}{n})$$

which is the same as

$$\sum_{k=2^{j-2}+1}^{2^{j-1}} f_s(\frac{k}{n}) + 3 \sum_{k=1}^{2^{j-3}} f_s(\frac{k}{n}) \geq 5 \sum_{k=2^{j-3}+1}^{2^{j-2}} f_s(\frac{k}{n}). \tag{75}$$

Now note that for $x \geq 0$ and $u \geq 0$,

$$f_s(x) + f_s(x+3u) \geq f_s(x+u) + f_s(x+2u) \quad \text{and} \quad f_s(x) + f_s(x+2u) \geq 2f_s(x+u)$$

and consequently $f_s(x+3u) + f_s(x+2u) + 3f_s(x) \geq 5f_s(x+u)$. Then (75) follows by taking $u = \frac{2^{j-3}}{n}$ and $x = \frac{k}{n}$ and then summing over $k = 1, ..., 2^{j-3}$.

Denote the bias of $\bar{\delta}_j$ by $\bar{b}_j = E\bar{\delta}_j - f(0)$. Then

$$\bar{b}_j = 2^{-(j-1)} \sum_{k=1}^{2^{j-1}} f_s(\frac{k}{n}) = 2^{-(j-1)} \left\{ \sum_{k=2^{j-2}+1}^{2^{j-1}} f_s(\frac{k}{n}) + \sum_{k=1}^{2^{j-2}} f_s(\frac{k}{n}) \right\}.$$

It follows from (74) that for $k > 2^{j-2}$, $f_s(\frac{k}{n}) \geq \frac{k}{2^{j-2}} f_s(\frac{2^{j-2}}{n})$, and for $k \leq 2^{j-2}$, $f_s(\frac{k}{n}) \leq \frac{k}{2^{j-2}} f_s(\frac{2^{j-2}}{n})$. Hence

$$\begin{aligned} \sum_{k=2^{j-2}+1}^{2^{j-1}} f_s(\frac{k}{n}) &\geq \sum_{k=2^{j-2}+1}^{2^{j-1}} \frac{k}{2^{j-2}} \cdot f_s(\frac{2^{j-2}}{n}) \geq \frac{\sum_{k=2^{j-2}+1}^{2^{j-1}} \frac{k}{2^{j-2}}}{\sum_{k=1}^{2^{j-2}} \frac{k}{2^{j-2}}} \sum_{k=1}^{2^{j-2}} f_s(\frac{k}{n}) \\ &= \frac{3 \cdot 2^{j-2} + 1}{2^{j-2} + 1} \sum_{k=1}^{2^{j-2}} f_s(\frac{k}{n}). \end{aligned}$$

Hence,

$$\bar{b}_j \geq 2^{-(j-1)} \cdot (\frac{3 \cdot 2^{j-2} + 1}{2^{j-2} + 1} + 1) \sum_{k=1}^{2^{j-2}} f_s(\frac{k}{n}) = \frac{2^{j-1} + 1}{2^{j-2} + 1} \bar{b}_{j-1}. \quad \blacksquare$$

7.4 Proof of Theorem 1

Consider the white noise model

$$dX(t) = h_\theta(t)dt + \frac{1}{\sqrt{n}}dW(t)$$

where $\theta = \pm 1$. The goal is to estimate $Th_\theta = h_\theta(0)$.

Let $\gamma_n = \frac{1}{2}\sqrt{n}(\int_{-1/2}^{1/2}(h_1(t) - h_{-1}(t))^2 dt)^{1/2}$ and let

$$W = \frac{n}{2\gamma_n}(\int_{-1/2}^{1/2}(h_1(t) - h_{-1}(t))dX(t) - \frac{1}{2}\int_{-1/2}^{1/2}(h_1^2(t) - h_{-1}^2(t))dt).$$

Then W is a sufficient statistic with $W \sim N(\gamma_n\theta,\ 1)$ where $\theta = \pm 1$.

Now every estimate $\hat{\theta}$ of θ gives an estimate of Tf_θ via

$$\hat{T} = \frac{Th_1 + Th_{-1}}{2} + \hat{\theta}\frac{Th_1 - Th_{-1}}{2}.$$

Now take the case where $\gamma_n = 1$. Then $W \sim N(\theta, 1)$where $\theta = \pm 1$. Then $\inf_{\hat{\theta}} \sup_{\theta=\pm 1} E(\hat{\theta} - \theta)^2 = \rho_N(1)$ where $\rho_N(1)$ is the minimax risk for the bounded normal mean problem with the absolute value bounded by 1 since in this case the least favorable prior is a two point prior supported on the end points as shown in Casella and Strawderman (1981) where it is also shown that $\rho_N(1) = 0.450$. Then

$$\inf_{\hat{T}} \sup_{\theta=\pm 1} E(\hat{T} - Th_\theta)^2 \geq \rho_N(1)(\frac{Th_1 - Th_{-1}}{2})^2.$$

Now take $h_{-1} = f$ and take a supremum of $|Th_1 - Tf|$ subject to $n\int(h_1(t) - f(t))^2 \leq 4$. This yields

$$R_n(f) \geq \frac{\rho_N(1)}{4}\omega^2(\frac{2}{\sqrt{n}}, f) \geq \frac{1}{9}\omega^2(\frac{2}{\sqrt{n}}, f) \geq \frac{1}{9}K^2(\frac{4}{3\sqrt{n}})$$

establishing the inequalities on the left hand side of equations (20) and (21).

We now turn to an upper bound to $R_n(f)$. Now for the pair h_1 and h_{-1} do not assume that $\gamma_n = 1$. Also note that every estimate of θ gives an estimate of $\gamma_n\theta$ and vice versa. In particular $\inf_{\hat{\theta}} \sup_{\theta=\pm 1} E(\hat{\theta} - \theta)^2 = \gamma_n^{-2}\rho(\gamma_n)$ and

$$\inf_{\hat{T}} \sup_{\theta=\pm 1} E(\hat{T} - Th_\theta)^2 = (\frac{Th_1 - Th_{-1}}{2})^2\gamma_n^{-2}\rho(\gamma_n)$$

where $\rho(\tau)$ is given by $\rho(\tau) = \inf_{\hat{\theta}} \sup_{\theta=\pm\tau} E(\hat{\theta}(X) - \theta)^2$ and $X \sim N(\theta, 1)$. By once again setting $f = h_{-1}$ and taking a supremum over all h_1 with $n\int_{-1/2}^{1/2}(h_1(x) - f(x))^2 = 4\gamma_n^2$ and then taking another supremum over all γ_n it follows that

$$R_n(f) = \sup_{\gamma_n} \frac{1}{4}\omega^2(\frac{2\gamma_n}{\sqrt{n}})\rho(\gamma_n).$$

The bound $\rho(\tau) \leq \tau^2 e^{-\frac{\tau^2}{2}}$ given in Donoho (1994) then immediately yields

$$R_n(f) \leq \sup_{\gamma_n} \frac{1}{4}\omega^2(\frac{2\gamma_n}{\sqrt{n}})e^{-\frac{\gamma_n^2}{2}}.$$

Now for any fixed $d > 0$ this last inequality can be written as

$$R_n(f) \le \max\left\{ \sup_{\gamma_n \le d} \frac{1}{4}\omega^2(\frac{2\gamma_n}{\sqrt{n}}, f)e^{-\frac{\gamma_n^2}{2}}, \ \sup_{\gamma_n \ge d} \frac{1}{4}\omega^2(\frac{2\gamma_n}{\sqrt{n}}, f)e^{-\frac{\gamma_n^2}{2}} \right\} \tag{76}$$

breaking the supremum into a maximum of two supremum. For the first supremum note that

$$\sup_{\gamma_n \le d} \omega^2(\frac{2\gamma_n}{\sqrt{n}}, f)e^{-\frac{\gamma_n^2}{2}} \le \omega^2(\frac{2d}{\sqrt{n}}, f). \tag{77}$$

For the second supremum note that since ω is concave it follows that

$$\sup_{\gamma_n \ge d} \frac{1}{4}\omega^2(\frac{2\gamma_n}{\sqrt{n}}, f)e^{-\frac{\gamma_n^2}{2}} = \sup_{C \ge 1} \omega^2(\frac{2Cd}{\sqrt{n}}, f)e^{-\frac{C^2d^2}{2}} \le \sup_{C \ge 1} C^2\omega^2(\frac{2d}{\sqrt{n}}, f)e^{-\frac{C^2d^2}{2}}.$$

The fact that xe^{-x} is maximized at $x = 1$ then implies that

$$\sup_{C \ge 1} \omega^2(\frac{Cd}{\sqrt{n}}, f)e^{-\frac{C^2d^2}{2}} \le \frac{2}{ed^2}\omega^2(\frac{2d}{\sqrt{n}}, f). \tag{78}$$

We set $d^2 = \frac{2}{e}$ to match the right hand sides of (77) and (78). It then follows from (76)-(78) and (19) that

$$R_n(f) \le \frac{1}{4}\omega^2(\frac{\sqrt{8}}{\sqrt{en}}, f) \le \frac{1}{4}K^2(\frac{4\sqrt{5}}{\sqrt{3e}\sqrt{n}})$$

establishing the inequalities on the right hand side of (20) and (21). ∎

7.5 Proof of Theorem 2

For $d \ge 1$ let t_d be the supremum of all t where $f_s(t) \le dr_n^{\frac{1}{2}}(f)$. Note that from Lemma 8 it follows that $t_d \le \frac{3\sqrt{3}d}{8nr_n(f)}$. By equation (57) of Lemma 5 there is a convex function g with $g(0) = f(0) + dr_n^{1/2}(f)$ satisfying

$$\int_{-1/2}^{1/2} (g(t) - f(t))^2 dt \le \frac{9}{4} t_d d^2 r_n(f)$$

and by the construction given in the proof of that lemma the function g is linear on $[-t_d, t_d]$.

Now choose d such that $\frac{9}{4}nt_d d^2 r_n(f) = \ln\frac{1}{c}$ from which it immediately follows that $\ln\frac{1}{c} \le \frac{9}{4}\frac{3\sqrt{3}}{8}d^3$. In this case we have $d \ge 0.88(\ln(1/c))^{1/3}$. As in the proof of proposition 3 we apply the constrained risk inequality of Brown and Low (1996b). As noted in the proof of Proposition 3 the chi-square distance between two white noise with drift models as given in (1), one with drift f and the other with drift g is given by $I_n(f, g) - 1$ where $I_n(f, g)$ is defined in (73). Hence for the f and g just given $I_n(f, g) \le \frac{1}{c}$. The constrained risk inequality given in Brown and Low (1996b) shows that for any estimator $\hat{T}$ for which $R(\hat{T}_n, f) \le cr_n(f)$ it follows that

$$R(\hat{T}_n, h) \ge (dr_n^{\frac{1}{2}}(f) - c^{\frac{1}{2}}r_n^{\frac{1}{2}}(f)\frac{1}{c^{\frac{1}{2}}})^2 = (d-1)^2 r_n(f).$$

Since the function g is linear on the interval $[-t_d, t_d]$ it follows that $r_n(h) \le \frac{1}{2t_d n}$. Hence

$$\frac{R(\hat{T}_n, h))}{r_n(h)} \ge (d-1)^2 r_n(f) 2t_d n = \frac{8}{9}\frac{(d-1)^2}{d^2}\frac{9}{4}d^2 n t_d r_n(f) = \frac{8}{9}\frac{(d-1)^2}{d^2}\ln\frac{1}{c},$$

and consequently

$$\frac{R(\hat{T}_n, h)}{r_n(h)} \ge \frac{8}{9}(1 - \frac{1}{0.88(\ln(1/c))^{1/3}})^2 \ln\frac{1}{c}. \quad ∎$$

7.6 Proof of Theorem 3

Recall $T_j = \delta_j - \delta_{j+1}$, $\hat{j} = \inf_j\{j : T_j \leq \lambda\sigma_j\}$, $j_* = \arg\min_j E(\delta_j - f(0))^2$ and $\lambda = \sqrt{2}$. Note that the T_i's are independent and that for any given $j \geq 1$, T_j is independent of $\delta_1, \ldots, \delta_j$. We shall denote the bias of δ_j by $b_j = E\delta_j - f(0)$. Note that the variances of δ_j and T_j are equal to $\sigma_j^2 = \frac{2^{j-1}}{n}$. Let $F_j = 1(T_j \leq \lambda\sigma_j)$ and note that F_j is independent of δ_j. The estimator $\hat{T}_*$ can be written as $\hat{T}_* = \sum_{j=1}^{\infty} \delta_j I(\hat{j} = j)$ and hence

$$E(\hat{T}_* - f(0))^2 = \sum_{j=1}^{\infty} E\left((\delta_j - f(0))^2 I(\hat{j} = j)\right).$$

Note that $1(\hat{j} = j) \leq F_j$ and hence

$$\begin{aligned} E(\hat{T}_* - f(0))^2 &\leq \sum_{j=1}^{j_*-3} E\left((\delta_j - f(0))^2 F_j\right) + \sum_{j=j_*-2}^{j_*} E\left((\delta_j - f(0))^2 F_j\right) \\ &+ \sum_{j=j_*+1}^{\infty} E\left((\delta_j - f(0))^2 I(\hat{j} = j)\right) \\ &= R_1 + R_2 + R_3. \end{aligned}$$

These three terms will be bounded separately. We begin with R_1. Now

$$R_1 = \sum_{j=1}^{j_*-3} (b_j^2 + \sigma_j^2) P(F_j = 1) = \sum_{m=3}^{j_*-1} (b_{j_*-m}^2 + \sigma_{j_*-m}^2) P(F_{j_*-m} = 1).$$

Set $\gamma_m = b_{j_*-m}/\sigma_{j_*}$. Then $ET_{j_*-m} = (\gamma_m - \gamma_{m-1})\sigma_{j_*} = (\gamma_m - \gamma_{m-1})2^{m/2}\sigma_{j_*-m}$. Hence

$$P(F_{j_*-m} = 1) = P(T_{j_*-m} \leq \lambda\sigma_{j_*-m}) = P(Z \leq \lambda - 2^{m/2}(\gamma_m - \gamma_{m-1})). \tag{79}$$

Note that for $m \geq 1$ equation (63) of Lemma 7 yields $\gamma_m - \gamma_{m-1} \geq \frac{2^{m-1}}{\sqrt{6}}$. It then follows that for $m \geq 1$, $P(F_{j_*-m} = 1) \leq P(Z \leq \lambda - \frac{2^{3m/2-1}}{\sqrt{6}})$ and consequently

$$\begin{aligned} R_1 &= \sigma_{j_*}^2 \left(\sum_{m=3}^{j_*-1} (\gamma_m^2 + 2^{-m}) P(Z \leq \lambda - 2^{m/2}(\gamma_m - \gamma_{m-1})) \right) \\ &\leq \sigma_{j_*}^2 \left(\sum_{m=3}^{j_*-1} (2^{2m-3} + 2^{-m}) P(Z \leq \lambda - \frac{2^{3m/2-1}}{\sqrt{6}} \right). \end{aligned}$$

Direct numerical calculations yield $\sum_{m=3}^{\infty} (2^{2m-3} + 2^{-m}) P(Z \leq \lambda - \frac{2^{3m/2-1}}{\sqrt{6}}) \leq 0.04$ and hence

$$R_1 \leq 0.04\sigma_{j_*}^2. \tag{80}$$

We now turn to a bound of R_3. The Cauchy-Swartz Inequality yields

$$\begin{aligned} R_3 &= \sum_{j=j_*+1}^{\infty} E\left(((\delta_j - f(0))^2 I(\hat{j} = j)\right) = \sum_{j=j_*+1}^{\infty} E\left((\delta_j - E\delta_j + b_j)^2 I(\hat{j} = j)\right). \\ &\leq \sqrt{3} \sum_{j=j_*+1}^{\infty} \sigma_j^2 (P(\hat{j} = j))^{1/2} + 2 \sum_{j=j_*+1}^{\infty} \sigma_j b_j (P(\hat{j} = j))^{1/2} + \sum_{j=j_*+1}^{\infty} b_j^2 P(\hat{j} = j) \\ &\equiv R_{31} + R_{32} + R_{33}. \end{aligned}$$

Equation (61) of Lemma 7 gives $ET_{j_*} \leq \min(1,\gamma_0)\sigma_{j_*}$ where $\gamma_0 = b_{j_*}/\sigma_{j_*}$. Hence it follows from equation (36) of Lemma 2 that for $m \geq 0$, $ET_{j_*+m} \leq 2^{-m}\min(1,\gamma_0)\sigma_{j_*}$ and consequently

$$\begin{aligned} P(F_{j_{*+m}} = 0) &= P(T_{j_{*+m}} > \lambda\sigma_{j_*+m}) \leq P(Z > \lambda - 2^{-m/2}2^{-m}\min(1,\gamma_0)) \\ &= P(Z > \lambda - 2^{-3m/2}\min(1,\gamma_0)). \end{aligned}$$

On the other hand, since $ET_{j_*+m} \geq 0$, $P(F_{j_{*+m}} = 1) = P(T_{j_{*+m}} \leq \lambda\sigma_{j_*+m}) \leq P(Z \leq \lambda)$. So for $m \geq 1$ it follows that

$$P(\hat{j} = j_{*+m}) \leq P(F_{j_*+m} = 1)\prod_{l=0}^{m-1} P(F_{j_*+l} = 0) \leq P(Z \leq \lambda)\prod_{l=0}^{m-1} P(Z > \lambda - 2^{-3l/2}\min(\gamma_0, 1)).$$

Note also that for $m \geq 0$, equation (35) of Lemma 2 yields $b_{j_*+m} \leq 2^{-m}b_{j_*} = \gamma_0 2^{-m}\sigma_{j_*}$ We then have

$$\begin{aligned} R_{31} &= \sqrt{3}\sum_{j=j_*+1}^{\infty}\sigma_j^2(P(\hat{j} = j))^{1/2} \\ &\leq \sqrt{3}\sigma_{j_*}^2\sum_{m=1}^{\infty}2^m\left(P(Z \leq \lambda)\prod_{l=0}^{m-1}P(Z > \lambda - 2^{-3l/2}\min(\gamma_0,1))\right)^{1/2} \end{aligned}$$

and

$$\begin{aligned} R_{32} &= 2\sum_{j=j_*+1}^{\infty}\frac{2^{(j-1)/2}}{\sqrt{n}}b_j(P(\hat{j} = j))^{1/2} \\ &\leq 2\sigma_{j_*}\sum_{m=1}^{\infty}2^{-m/2}\gamma_0\left(P(Z \leq \lambda)\prod_{l=0}^{m-1}P(Z > \lambda - 2^{-3l/2}\min(\gamma_0,1))\right)^{1/2}. \end{aligned}$$

Finally note that

$$R_{33} = \sum_{j=j_*+1}^{\infty}(E\delta_j - f(0))^2 P(\hat{j} = j) \leq B_{j_*+1}^2 \leq \frac{1}{4}\gamma_0^2\sigma_{j_*}^2$$

and hence it follows that

$$\frac{R_3}{\sigma_{j_*}^2} \leq (\frac{1}{4}\gamma_0^2 + \sum_{m=1}^{\infty}(2^m\sqrt{3} + 2^{-m/2}2\gamma_0)\left(P(Z \leq \lambda)\prod_{l=0}^{m-1}P(Z > \lambda - 2^{-3l/2}\min(\gamma_0,1))\right)^{1/2}.$$

We now bound R_2. Recall (79) and note that

$$\begin{aligned} R_2 &= \sum_{j=j_*-2}^{j_*}(b_j^2 + \sigma_j^2)P(F_j = 1) \leq \sigma_{j_*}^2(2^{-2} + \gamma_2^2)P(Z \leq \lambda - 2(\gamma_2 - \gamma_1)) \\ &\quad + \sigma_{j_*}^2(2^{-1} + \gamma_1^2)P(Z \leq \lambda - 2^{1/2}(\gamma_1 - \gamma_0)) + \sigma_{j_*}^2(1 + \gamma_0^2)P(Z \leq \lambda - \frac{\gamma_0}{2}) \end{aligned}$$

where from (62) of lemma 7, $\gamma_1 \geq \frac{1}{\sqrt{2}}$ and from (35) of lemma 2, $\gamma_0 \leq \frac{\gamma_1}{2}$. Since $\gamma_1 \geq \frac{1}{\sqrt{2}}$, Lemma 10 now yields $(2^{-2} + \gamma_2^2)P(Z \leq \lambda - 2(\gamma_2 - \gamma_1)) \leq 0.649$ Then we have

$$R_2 \leq \sigma_{j_*}^2\left\{0.649 + (2^{-1} + \gamma_1^2)P(Z \leq \lambda - 2^{1/2}(\gamma_1 - \gamma_0)) + (1 + \gamma_0^2)P(Z \leq \lambda - \frac{\gamma_0}{2})\right\}.$$

Hence

$$\frac{R_2+R_3}{\sigma_{j_*}^2(1+\gamma_0^2)} \le P(Z \le \lambda - \frac{\gamma_0}{2}) + \frac{0.649}{1+\gamma_0^2} + \frac{(2^{-1}+\gamma_1^2)P(Z \le \lambda - 2^{1/2}(\gamma_1-\gamma_0))}{1+\gamma_0^2}$$
$$+ \frac{1}{1+\gamma_0^2}(\frac{1}{4}\gamma_0^2 + \sum_{m=1}^{\infty}(2^m\sqrt{3}+2^{-m/2}2\gamma_0)\left(P(Z \le \lambda)\prod_{l=0}^{m-1}P(Z > \lambda - 2^{-3l/2}\min(\gamma_0,1))\right)^{1/2}.$$

The right hand side is a function of two variables γ_0 and γ_1. The numerical results given in Lemmas 9 and 10 and equation (80) now yield that

$$\frac{R_1+R_2+R_3}{\sigma_*^2(1+\gamma_0^2)} \le 0.04 + 4.7 + 1.2 \le 6. \quad \blacksquare$$

7.7 Proof of Theorem 4

Although this theorem may be proved directly it is interesting to see how it easily follows from Theorem 3. For the class $F_c(\alpha, M)$ let $\hat{L}_a$ be the local linear estimate defined in (22) where $a = M^{-2/2\alpha+1}n^{-1/2\alpha+1}$. It is then easy to check that

$$E(\hat{L}_a - f(0))^2 \le (\frac{1}{2} + \frac{1}{(\alpha+1)^2})M^{\frac{2}{2\alpha+1}}n^{-\frac{2\alpha}{2\alpha+1}} \le \frac{3}{4}M^{\frac{2}{2\alpha+1}}n^{-\frac{2\alpha}{2\alpha+1}}$$

and hence in particular that $r_n(f) \le \frac{3}{4}M^{\frac{2}{2\alpha+1}}n^{-\frac{2\alpha}{2\alpha+1}}$ for all $f \in F_c(\alpha, M)$. The inequality in (42) with, for example, $C = 9$ then immediately follows from (41).

To complete the proof we need also to demonstrate that $R_*(n, F_c(\alpha, M)) \ge dM^{\frac{2}{2\alpha+1}}n^{-\frac{2\alpha}{2\alpha+1}}$ for some $d > 0$. For this, fix α and M where $1 \le \alpha \le 2$ and let f and g be two convex functions defined on the whole real line which are also both in $\Lambda(\alpha, 1)$, such that $g(0) - f(0) > 0$ with $\int_{-\infty}^{\infty}(g(t) - f(t))^2dt \le 4$. Such convex functions f and g can be easily constructed. Now let f_n and g_n be the two functions with domain $[-1/2, 1/2]$ which are given by $f_n(t) = M^{\frac{1}{2\alpha+1}}n^{-\frac{\alpha}{2\alpha+1}}f(M^{\frac{2}{2\alpha+1}}n^{\frac{1}{2\alpha+1}}t)$ and $g_n(t) = M^{\frac{1}{2\alpha+1}}n^{-\frac{\alpha}{2\alpha+1}}g(M^{\frac{2}{2\alpha+1}}n^{\frac{1}{2\alpha+1}}t)$ for $-1/2 \le t \le 1/2$. Note that both f_n and g_n belong to $F_c(\alpha, M)$. It is also easy to check that $\int_{-1/2}^{1/2}(f_n(t) - g_n(t))^2dt \le \frac{4}{n}$.

We now follow the same argument as given for the proof of Theorem 1 where we take $h_1 = g_n$ and $h_{-1} = f_n$. Following that argument it is easy to check that $\gamma_n \le 1$ and hence

$$R_*(n, F_c(\alpha, M)) \ge \rho_N(1)(\frac{Th_1 - Th_{-1}}{2})^2 = (\frac{g(0)-f(0)}{2})^2\rho_N(1)M^{\frac{2}{2\alpha+1}}n^{-\frac{2\alpha}{2\alpha+1}}. \quad \blacksquare$$

Acknowledgment

We wish to thank the referees and the associate editor for their very thorough and useful comments to an earlier version of this manuscript which resulted in many improvements to the original work.

References

[1] Bickel, P. J., Klassen, A. J., Ritov, Y. and Wellner, J. A. (1993). *Efficient and Adaptive Inference in Semiparametric Models.* Johns Hopkins University Press, Baltimore.

[2] Birge, L. (1989). The Grenander estimator: a nonasymptotic approach. *Ann. Statist.* **17**, 1532-1549.

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

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

[5] Cai, T. T. and Low, M. G. (2004). An adaptation theory for nonparametric confidence intervals. *Ann. Statist.* **32**, 1805-1840.

[6] Cai, T. T. and Low, M. G. (2006). Adaptation under probabilistic error. *J. Multivariate Analysis* **97**, 231-245.

[7] Cai, T. T., Low, M. G. and Xia, Y. (2013). Adaptive confidence intervals for regression functions under shape constraints. *Ann. Statist.* **42**, 722-750.

[8] Cai, T. T. and Low, M. G. (2013). Supplement to "A Framework For Estimation of Convex Functions". Technical report.

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

[10] Cator, E. (2011). Adaptivity and optimality of the monotone least-squares estimator. *Bernoulli* **17**, 714-735.

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

[12] Donoho, D. L., Johnstone, I.M., Kerkyacharian, G. and Picard, D. (1995). Wavelet Shrinkage: Asymptopia. *J. Roy. Statist. Soc.* **57**, 301-369.

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

[14] Dümbgen, L. (1998). New goodness-of-fit tests and their applications to nonparametric confidence sets. *Ann. Statist.* **26**, 288-314.

[15] Dümbgen, L. and Rufibach, K. (2009). Maximum likelihood estimation of a log-concave density and its distribution function: Basic properties and uniform consistency. *Bernoulli* **15**, 40-68.

[16] Grenander, U. (1956). On the theory of mortality measurement. II. *Skand. Akt. Tid.* **39**, 125-153.

[17] Groeneboom, P., Jongbloed, G. and Wellner, J.A. (2001a). A canonical process for estimation of convex functions: the "invelope" of integrated Brownian motion $+t^4$. *Ann. Statist.* **29**, 1620-1652.

[18] Groeneboom, P., Jongbloed, G. and Wellner, J.A. (2001b). Estimation of a convex function: characterizations and asymptotic theory. *Ann. Statist.* **29**, 1653-1698.

[19] Hengartner, N. W. and Stark, P. B. (1995). Finite-sample confidence envelopes for shape-restricted densities. *Ann. Statist.* **23**, 525-550.

[20] 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.

[21] Jongbloed, G. (2000). Minimax lower bounds and moduli of continuity. *Statist. and Prob. Lett.* **50**. 279-284.

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

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

[24] Low, M. G. (1997). On nonparametric confidence intervals. *Ann. Statist.* **25**, 2547-2554.

[25] Mammen, E. (1991). Nonparametric regression under qualitative smoothness assumptions. *Ann. Statist.* **19**, 741-759.

[26] Stein, C. (1956). Efficient nonparametric testing and estimation. Proceedings of the Third Berkeley Symposium in Mathematical Statistics and Probability, vol 1, pp. 187-196. Berkeley, University of California Press.