Asymptotic Equivalence and Adaptive Estimation for Robust Nonparametric Regression

T. Tony Cai[1] and Harrison H. Zhou[2]

University of Pennsylvania and Yale University

Abstract

Asymptotic equivalence theory developed in the literature so far are only for bounded loss functions. This limits the potential applications of the theory because many commonly used loss functions in statistical inference are unbounded. In this paper we develop asymptotic equivalence results for robust nonparametric regression with unbounded loss functions. The results imply that all the Gaussian nonparametric regression procedures can be robustified in a unified way. A key step in our equivalence argument is to bin the data and then take the median of each bin.

The asymptotic equivalence results have significant practical implications. To illustrate the general principles of the equivalence argument we consider two important nonparametric inference problems: robust estimation of the regression function and the estimation of a quadratic functional. In both cases easily implementable procedures are constructed and are shown to enjoy simultaneously a high degree of robustness and adaptivity. Other problems such as construction of confidence sets and nonparametric hypothesis testing can be handled in a similar fashion.

Keywords: Adaptivity; Asymptotic equivalence; James-Stein estimator; moderate deviation; Nonparametric regression; Quantile coupling; Robust estimation; Unbounded loss function; Wavelets.

AMS 2000 Subject Classification: Primary 62G08, Secondary 62G20.

1 Introduction

The main goal of the asymptotic equivalence theory is to approximate general statistical models by simple ones. If a complex model is asymptotically equivalent to a simple model, then all asymptotically optimal procedures can be carried over from the simple model to the complex one for *bounded* loss functions and the study of the complex model is then essentially simplified. Early work on asymptotic equivalence theory was focused on the parametric models and the equivalence is local. See Le Cam (1986).

There have been important developments in the asymptotic equivalence theory for nonparametric models in the last decade or so. In particular, global asymptotic equivalence theory has been developed for nonparametric regression in Brown and Low (1996) and Brown, Cai, Low and Zhang (2002), nonparametric density estimation models in Nussbaum (1996) and Brown, Carter, Low and Zhang (2004), generalized linear models in Grama and Nussbaum (1998), nonparametric autoregression in Milstein and Nussbaum (1998), diffusion models in Delattre and Hoffmann (2002) and Genon-Catalot, Laredo and Nussbaum (2002), GARCH model in Wang (2002) and Brown, Wang and Zhao (2003), and spectral density estimation in Golubev, Nussbaum and Zhou (2005).

So far all the asymptotic equivalence results developed in the literature are only for bounded loss functions. However, for many statistical applications, asymptotic equivalence under bounded losses is not sufficient because many commonly used loss functions in statistical inference such as squared error loss are unbounded. As commented by Iain Johnstone (2002) on the asymptotic equivalence results, "Some cautions are in order when interpreting these results. .... Meaningful error measures ...... may not translate into, say, squared error loss in the Gaussian sequence model."

In this paper we develop asymptotic equivalence results for robust nonparametric regression with an unknown symmetric error distribution for *unbounded* loss functions which include for example the commonly used squared error and integrated squared error losses. Consider the nonparametric regression model

$$Y_i = f(\frac{i}{n}) + \xi_i, \quad i = 1, \ldots, n \tag{1}$$

where the errors ξ_i are independent and identically distributed with some density h. The error density h is assumed to be symmetric with median 0, but otherwise unknown. Note that for some heavy-tailed distributions such as Cauchy distribution the mean does not even exist. We thus do not assume the existence of the mean here. One is often interested in robustly estimating the regression function f or some functionals of f. These problems have been well studied in the case of Gaussian errors. In the present paper we introduce a unified approach to turn the general nonparametric regression model (1) into

a standard Gaussian regression model and then in principle any procedure for Gaussian nonparametric regression can be applied. More specifically, with properly chosen T and m, we propose to divide the observations Y_i into T bins of size m and then take the median X_j of the observations in the jth bin for $j = 1, ..., T$. The asymptotic equivalence results developed in Section 2 show that under mild regularity conditions, for a wide collection of error distributions the experiment of observing the medians $\{X_j : j = 1, ..., T\}$ is in fact asymptotically equivalent to the standard Gaussian nonparametric regression model

$$Y_i = f(\frac{i}{T}) + \frac{1}{2h(0)\sqrt{m}} z_i, \quad z_i \overset{iid}{\sim} N(0, 1), \quad i = 1, \ldots, T \tag{2}$$

for a large class of unbounded losses. Detailed arguments are given in Section 2.

We develop the asymptotic equivalence results for the general regression model (1) by first extending the classical formulation of asymptotic equivalence in Le Cam (1964) to accommodate unbounded losses. The asymptotic equivalence result has significant practical implications. It implies that all statistical procedures for any asymptotic decision problem in the setting of the Gaussian nonparametric regression can be carried over to solve problems in the general nonparametric regression model (1) for a class of unbounded loss functions. In other words, all the Gaussian nonparametric regression procedures can be robustified in a unified way. We illustrate the applications of the general principles in two important nonparametric inference problems under the model (1): robust estimation of the regression function f under integrated squared error loss and the estimation of the quadratic functional $Q(f) = \int f^2$ under squared error.

As we demonstrate in Sections 3 and 4 the key step in the asymptotic equivalence theory, binning and taking the medians, can be used to construct simple and easily implementable procedures for estimating the regression function f and the quadratic functional $\int f^2$. After obtaining the medians of the binned data, the general model (1) with an unknown symmetric error distribution is turned into a familiar Gaussian regression model, and then a Gaussian nonparametric regression procedure can be applied. In Section 3 we choose to employ a blockwise James-Stein wavelet estimator, BlockJS, for the Gaussian regression problem because of its desirable theoretical and numerical properties. See Cai (1999). The robust wavelet regression procedure has two main steps: 1. Binning and taking median of the bins; 2. Applying the BlockJS procedure to the medians. The procedure is shown to achieve four objectives simultaneously: robustness, global adaptivity, spatial adaptivity, and computational efficiency. Theoretical results in Section 3.2 show that the estimator achieves optimal global adaptation for a wide range of Besov balls as well as a large collection of error distributions. In addition, it attains the local adaptive minimax rate for estimating functions at a point. Figure 1 compares a direct wavelet

estimate with our robust estimate in the case of Cauchy noise. The example illustrates the fact that direct application of a wavelet regression procedure designed for Gaussian noise may not work at all when the noise is in fact heavy-tailed. On the other hand, our robust procedure performs well even in Cauchy noise.
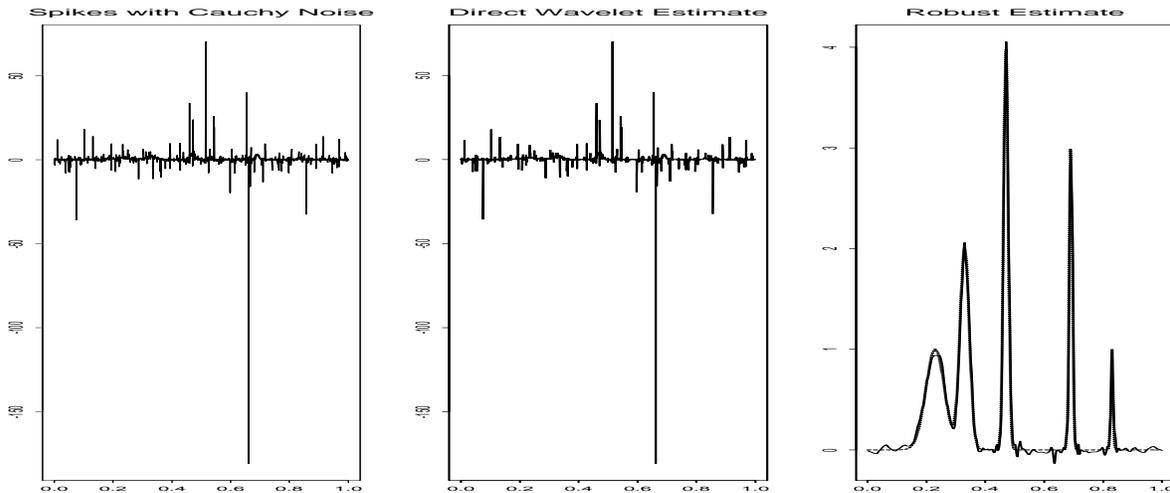

Figure 1: Left panel: Spikes signal with Cauchy noise; Middle panel: An estimate obtained by applying directly a wavelet procedure to the original noisy signal; Right panel: A robust estimate by apply a wavelet block thresholding procedure to the medians of the binned data. Sample size is 4096 and bin size is 8.

In Section 4 we construct a robust procedure for estimating the quadratic functional $Q(f) = \int f^2$ following the same general principles. Other problems such as construction of confidence sets and nonparametric hypothesis testing can be handled in a similar fashion.

Key technical tools used in our development are an improved moderate deviation result for the median statistic and a better quantile coupling inequality. Median coupling has been considered in Brown, Cai and Zhou (2008). For the asymptotic equivalence results given in Section 2 and the proofs of the theoretical results in Section 3 we need a more refined moderate deviation result for the median and an improved coupling inequality than those given in Brown, Cai and Zhou (2008). These improvements play a crucial role in this paper for establishing the asymptotic equivalence as well as robust and adaptive estimation results. The results may be of independent interest for other statistical applications.

The paper is organized as follows. Section 2 develops an asymptotic equivalence theory for unbounded loss functions. To illustrate the general principles of the asymptotic equivalence theory, we then consider robust estimation of the regression function f under integrated squared error in Section 3 and estimation of the quadratic functional $\int f^2$

4

under squared error in Section 4. The two estimators are easily implementable and are shown to enjoy desirable robustness and adaptivity properties. In Section 5 we derive a moderate deviation result for the medians and a quantile coupling inequality. The proofs are contained in Section 6.

2 Asymptotic equivalence

This section develops an asymptotic equivalence theory for unbounded loss functions. The results reduce the general nonparametric regression model (1) to a standard Gaussian regression model.

The Gaussian nonparametric regression has been well studied and it often serves as a prototypical model for more general nonparametric function estimation settings. A large body of literature has been developed for minimax and adaptive estimation in the Gaussian case. These results include optimal convergence rates and optimal constants. See, e.g., Pinsker (1980), Korostelev (1993), Donoho, Johnstone, Kerkyacharian, and Picard (1995), Johnstone (2002), Tsybakov(2004), Cai and Low (2005, 2006b) and references therein for various estimation problems under various loss functions. The asymptotic equivalence results established in this section can be used to robustify these procedures in a unified way to treat the general nonparametric regression model (1).

We begin with a brief review of the classical formulation of asymptotic equivalence and then generalize the classical formulation to accommodate unbounded losses.

2.1 Classical asymptotic equivalence theory

Lucien Le Cam (1986) developed a general theory for asymptotic decision problems. At the core of this theory is the concept of a distance between statistical models (or experiments), called Le Cam's deficiency distance. The goal is to approximate general statistical models by simple ones. If a complex model is close to a simple model in Le Cam's distance, then there is a mapping of solutions to decision theoretic problems from one model to the other for all *bounded* loss functions. Therefore the study of the complex model can be reduced to the one for the simple model.

A family of probability measures $E = \{P_\theta : \theta \in \Theta\}$ defined on the same σ-field of a sample space Ω is called a statistical model (or experiment). Le Cam (1964) defined a distance $\Delta(E, F)$ between E and another model $F = \{Q_\theta : \theta \in \Theta\}$ with the same parameter set Θ by the means of "randomizations". Suppose one would like to approximate E by a simpler model F. An observation x in E can be mapped into the sample space of F by generating an "observation" y according to a Markov kernel K_x, which is a probability

measure on the sample space of F. Suppose x is sampled from P_θ. Write KP_θ for the distribution of y with $KP_\theta(A) = \int K_x(A) \, dP_\theta$ for a measurable set A. The deficiency δ of E with respect to F is defined as the smallest possible value of the total variation distance between KP_θ and Q_θ among all possible choices of K, i.e.,

$$\delta(E, F) = \inf_K \sup_{\vartheta \in \Theta} |KP_\vartheta - Q_\vartheta|_{TV}.$$

See Le Cam (1986, page 3) for further details. The deficiency δ of E with respect to F can be explained in terms of risk comparison. If $\delta(E, F) \leq \varepsilon$ for some $\varepsilon > 0$, it is easy to see that for every procedure τ in F there exists a procedure ξ in E such that $R(\theta; \xi) \leq R(\theta; \tau) + 2\varepsilon$ for every $\theta \in \Theta$ and any loss function with values in the unit interval. The converse is also true. Symmetrically one may consider the deficiency of F with respect to E as well. The *Le Cam's deficiency distance* between the models E and F is then defined as

$$\Delta(E, F) = \max(\delta(E, F), \delta(F, E)). \tag{3}$$

For bounded loss functions, if $\Delta(E,F)$ is small, then to every statistical procedure for E there is a corresponding procedure for F with almost the same risk function and vice versa. Two sequences of experiments E_n and F_n are called *asymptotically equivalent*, if $\Delta(E_n, F_n) \to 0$ as $n \to \infty$. The significance of asymptotic equivalence is that all asymptotically optimal statistical procedures can be carried over from one experiment to the other for bounded loss functions.

2.2 Extension of the classical asymptotic equivalence formulation

For many statistical applications, asymptotic equivalence under bounded losses is not sufficient because many commonly used loss functions are unbounded. Let $E_n = \{P_{\theta,n} : \theta \in \Theta\}$ and $F_n = \{Q_{\theta,n} : \theta \in \Theta\}$ be two asymptotically equivalent models in Le Cam's sense. Suppose that the model F_n is simpler and well studied and a sequence of estimators $\hat{\theta}_n$ satisfy

$$\mathbb{E}_{Q_{\theta,n}} n^r d\left(\widehat{\theta}_n, \theta\right) \to c \quad \text{as } n \to \infty,$$

where d is a distance between $\widehat{\theta}$ and θ, and $r, c > 0$ are constants. This implies that θ can be estimated by $\hat{\theta}_n$ under the distance d with a rate n^{-r}. Examples include $\mathbb{E}_{Q_{\theta,n}} n\left(\widehat{\theta} - \theta\right)^2 \to c$ in many parametric estimation problems, and $\mathbb{E}_{Q_{f,n}} n^r \int (\widehat{f} - f)^2 d\mu \to c$, where f is an unknown function and $0 < r < 1$, in many nonparametric estimation problems. The asymptotic equivalence between E_n and F_n in the classical sense does not imply that there is an estimator $\hat{\theta}^*$ in E_n such that

$$\mathbb{E}_{P_{\theta,n}} n^r d\left(\hat{\theta}^*, \theta\right) \to c.$$

6

In this setting the loss function is actually $L(\vartheta, \theta) = n^r d(\vartheta, \theta)$ which grows as n increases, and is usually unbounded.

In this section we introduce a new asymptotic equivalence formulation to handle unbounded losses. Let Λ be a set of procedures, and Γ be a set of loss functions. We define the deficiency distance $\Delta(E, F; \Gamma, \Lambda)$ as follows.

Definition 1 *Define $\delta(E, F; \Gamma, \Lambda) \equiv \inf\{\epsilon \geq 0 : $ for every procedure $\tau \in \Lambda$ in F there exists a procedure $\xi \in \Lambda$ in E such that $R(\theta; \xi) \leq R(\theta; \tau) + 2\varepsilon$ for every $\theta \in \Theta$ for any loss function $L \in \Gamma\}$. Then the deficiency distance between models E and F for the loss class Γ and procedure class Λ is defined as $\Delta(E, F; \Gamma, \Lambda) = \max\{\delta(E, F; \Gamma, \Lambda), \delta(F, E; \Gamma, \Lambda)\}$.*

In other words, if the deficiency $\Delta(E, F; \Gamma, \Lambda)$ is small, then to every statistical procedure for one experiment there is a corresponding procedure for another experiment with almost the same risk function for losses $L \in \Gamma$ and procedures in Λ.

Definition 2 *Two sequences of experiments E_n and F_n are called asymptotically equivalent with respect to the set of procedures Λ_n and set of loss functions Γ_n if $\Delta(E_n, F_n; \Gamma_n, \Lambda_n) \to 0$ as $n \to \infty$.*

If E_n and F_n are asymptotically equivalent, then all asymptotically optimal statistical procedures in Λ_n can be carried over from one experiment to the other for loss functions $L \in \Gamma_n$ with essentially the same risk. The definitions here generalize the classical asymptotic equivalence formulation, which corresponds to the special case with Γ being the set of loss functions with values in the unit interval.

For most statistical applications the loss function is bounded by a certain power of n. We now give a sufficient condition for the asymptotic equivalence under such losses. Suppose that we estimate f or a functional of f under a loss L. Let $p_{f,n}$ and $q_{f,n}$ be the density functions respectively for E_n and F_n. Note that in the classical formulation of asymptotic equivalence for bounded losses, the deficiency of E_n with respect to F_n goes to zero if there is a Markov kernel K such that

$$\sup_f |KP_{f,n} - Q_{f,n}|_{TV} \to 0 \tag{4}$$

For unbounded losses the condition (4) is no longer sufficient to guarantee that the deficiency goes to zero. Let $p^*_{f,n}$ and $q_{f,n}$ be the density functions of $KP_{f,n}$ and $Q_{f,n}$ respectively. Let $\varphi(f)$ be an estimand, which can be f or a functional of f. Suppose that in F_n there is an estimator $\widehat{\varphi(f)}_q$ of $\varphi(f)$ such that

$$\int L(\widehat{\varphi(f)}_q, \varphi(f)) q_{f,n} \to c.$$

We would like to derive sufficient conditions under which there is an estimator $\widehat{\varphi(f)}_p$ in E_n such that

$$\int L(\widehat{\varphi(f)}_p, \varphi(f)) p_{f,n} \to c.$$

Note that if $\widehat{\varphi(f)}_p$ is constructed by mapping over $\widehat{\varphi(f)}_q$ via a Markov kernel T, then

$$\mathbb{E}L(\widehat{\varphi(f)}_p, \varphi(f)) = \int L(\widehat{\varphi(f)}_q, \varphi(f)) p_{f,n}^* \leq \int L(\widehat{\varphi(f)}_q, \varphi(f)) q_{f,n} + \int L(\widehat{\varphi(f)}_q, \varphi(f)) |p_{f,n}^* - q_{f,n}|$$

Let $A_{\varepsilon_n} = \left\{ |1 - p_{f,n}^*/q_{f,n}| < \varepsilon_n \right\}$ for some $\varepsilon_n \to 0$, and write

$$\begin{aligned}
\int L(\widehat{\varphi(f)}_q, \varphi(f)) |p_{f,n}^* - q_{f,n}| &= \int L(\widehat{\varphi(f)}_q, \varphi(f)) |p_{f,n}^* - q_{f,n}| [I(A) + I(A^c)] \\
&\leq \varepsilon_n \int L(\widehat{\varphi(f)}_q, \varphi(f)) q_{f,n} + \int L(\widehat{\varphi(f)}_q, \varphi(f)) q_{f,n} I(A_n^c).
\end{aligned}$$

If $Q_{f,n}(A_n^c)$ decays exponentially fast uniformly over $\mathcal{F}$ and L is bounded by a polynomial of n, this formula implies that $\int L(\widehat{\varphi(f)}_q, \varphi(f)) |p_{f,n}^* - q_{f,n}| = o(1)$.

Assumption (A0): For each estimand $\varphi(f)$, each estimator $\widehat{\varphi(f)} \in \Lambda_n$ and each $L \in \Gamma_n$, there is a constant $M > 0$, independent of the loss function and the procedure, such that $L(\widehat{\varphi(f)}, \varphi(f)) \leq M n^M$.

The following result summarizes the above discussion and gives a sufficient condition for the asymptotic equivalence for the set of procedures Λ_n and set of loss functions Γ_n.

Proposition 1 *Let $E_n = \{P_{\theta,n} : \theta \in \Theta\}$ and $F_n = \{Q_{\theta,n} : \theta \in \Theta\}$ be two models. Suppose there is a Markov kernel K such that $KP_{\theta,n}$ and $Q_{\theta,n}$ are defined on the same σ-field of a sample space. Let $p_{f,n}^*$ and $q_{f,n}$ be the density functions of $KP_{f,n}$ and $Q_{f,n}$ w.r.t. a dominating measure such that for a sequence $\varepsilon_n \to 0$*

$$\sup_f Q_{f,n}(|1 - p_{f,n}^*/q_{f,n}| \geq \varepsilon_n) \leq C_D n^{-D}$$

for all $D > 0$, then $\delta(E_n, F_n; \Gamma_n, \Lambda_n) \to 0$ as $n \to \infty$ under the Assumption (A0).

Examples of loss functions include

$$L(\hat{f}_n, f) = n^{2\alpha/(2\alpha+1)} \int (\hat{f}_n - f)^2 \quad \text{and} \quad L(\hat{f}_n, f) = n^{2\alpha/(2\alpha+1)} \int (\sqrt{\hat{f}_n} - \sqrt{f})^2$$

for estimating f and $L(\hat{f}_n, f) = n^{2\alpha/(2\alpha+1)} (\hat{f}_n(t_0) - f(t_0))^2$ for estimating f at a fixed point t_0 where α is the smoothness of f, as long as we require $\hat{f}_n$ to be bounded by a power of n. If the maximum of $\hat{f}_n$ or $\hat{f}_n(t_0)$ grows faster than a polynomial of n, we commonly obtain a better estimate by truncation, e.g., defining a new estimate $\min(\hat{f}_n, n^2)$.

8

2.3 Asymptotic equivalence for robust estimation under unbounded losses

We now return to the nonparametric regression model (1) and denote the model by E_n,

$$E_n : Y_i = f(i/n) + \xi_i, \; i = 1, \ldots, n.$$

An asymptotic equivalence theory for nonparametric regression with a known error distribution has been developed in Grama and Nussbaum (2002), but the Markov kernel (randomization) there was not given explicitly, and so it is not implementable. In this section we propose an explicit and easily implementable procedure to reduce the nonparametric regression with an unknown error distribution to a Gaussian regression. We begin by dividing the interval $[0, 1]$ into T equal-length subintervals. Without loss of generality we shall assume that n is divisible by T, and let $m = n/T$, the number of observations in each bin. We then take the median X_j of the observations in each bin, i.e.,

$$X_j = \text{median}\left\{ Y_i, (j-1)\, m + 1 \leq i < jm \right\},$$

and make statistical inferences based on the median statistics $\{X_j\}$. Let F_n be the experiment of observing $\{X_j, 1 \leq j \leq T\}$. In this section we shall show that F_n is in fact asymptotically equivalent to the following Gaussian experiment

$$G_n : X_j^{**} = f(j/T) + \frac{1}{2h(0)\sqrt{m}} Z_j, \; Z_j \overset{i.i.d.}{\sim} N(0,1), 1 \leq j \leq T$$

under mild regularity conditions. The asymptotic equivalence is established in two steps.

Suppose the function f is smooth. Then f is locally approximately constant. We define a new experiment to approximate E_n as follows

$$E_n^* : Y_i^* = f^*(i/n) + \xi_i, \; 1 \leq i \leq n,$$

where $f^*(i/n) = f\left(\frac{[iT/n]}{T}\right)$. For each of the T subintervals, there are m observations centered around the same mean.

For the experiment E_n^* we bin the observations Y_i^* and then take the medians in exactly the same way and let X_j^* be the median of the Y_i^*'s in the j-th subinterval. If E_n^* approximates E_n well, the statistical properties X_j^* are then similar to X_j. Let η_j be the median of corresponding errors ξ_i in the j-th bin. Note that the median of X_j^* has a very simple form,

$$F_n^* : X_j^* = f(j/T) + \eta_j, \; 1 \leq j \leq T.$$

Theorem 6 in Section 5 shows that η_j can be well approximated by a normal variable with mean 0 and variance $\frac{1}{4mh^2(0)}$, which suggests that F_n^* is close to the experiment G_n.

We formalize the above heuristics in the following theorems. We first introduce some conditions. We shall choose $T = n^{2/3}/\log n$ and assume that f is in a Hölder ball,

$$f \in \mathcal{F} = \{f : |f(y) - f(x)| \le M|x - y|^d\}, d > 3/4. \tag{5}$$

Assumption (A1): Let ξ be a random variable with density function h. Define $r_a(\xi) = \log \frac{h(\xi - a)}{h(\xi)}$ and $\mu(a) = \mathbb{E}r(\xi)$. Assume that

$$
\begin{aligned}
\mu(a) &\le Ca^2 & (6)\\
\mathbb{E}\exp[t(r_a(\xi) - \mu(a))] &\le \exp(Ct^2a^2) & (7)
\end{aligned}
$$

for $0 \le |a| < \varepsilon$ and $0 \le |ta| < \varepsilon$ for some $\varepsilon > 0$. Equation (7) is roughly equivalent to $\mathbb{V}ar(r_a(\xi)) \le Ca^2$. Assumption (A1) is satisfied by many distributions including Cauchy and Gaussian.

The following asymptotic equivalence result implies that any procedure based on X_j has exactly the same asymptotic risk as a similar procedure by just replacing X_j by X_j^*.

Theorem 1 *Under Assumptions (A0) and (A1) and the Hölder condition (5), the two experiments E_n and E_n^* are asymptotically equivalent with respect to the set of procedures Λ_n and set of loss functions Γ_n.*

The following asymptotic equivalence result implies that asymptotically there is no need to distinguish X_j^*'s from the Gaussian random variables X_j^{**}'s. We need the following assumptions on the density function $h(x)$ of ξ.

Assumption (A2): $\int_{-\infty}^0 h(x) = \frac{1}{2}$, $h(0) > 0$, and $|h(x) - h(0)| \le Cx^2$ in an open neighborhood of 0.

The last condition $|h(x) - h(0)| \le Cx^2$ is basically equivalent to $h'(0) = 0$. The assumption (A2) is satisfied when h is symmetric and h'' exists in a neighborhood of 0.

Theorem 2 *Under Assumptions $A(0)$ and (A2), the two experiments F_n^* and G_n are asymptotically equivalent with respect to the set of procedures Λ_n and set of loss functions Γ_n.*

These theorems imply under assumptions (A1) and (A2) and the Hölder condition (5), the experiment F_n is asymptotically equivalent to G_n with respect to the set of procedures Λ_n and set of loss functions Γ_n. So any statistical procedure δ in G_n can be carried over to the E_n (by treating X_j as if it were X_j^{**}) in the sense that the new procedure has the same asymptotic risk as δ for all loss functions bounded by a certain power of n.

2.4　Discussion

The asymptotic equivalence theory provides deep insight and useful guidance for the construction of practical procedures in a broad range of statistical inference problems under the nonparametric regression model (1) with an unknown symmetric error distribution. Interesting problems include robust and adaptive estimation of the regression function, estimation of linear or quadratic functionals, construction of confidence sets, nonparametric hypothesis testing, etc. There is a large body of literature on these nonparametric problems in the case of Gaussian errors. With the asymptotic equivalence theory developed in this section, many of these procedures and results can be extended and robustified to deal with the case of an unknown symmetric error distribution. For example, the SureShrink procedure of Donoho and Johnstone (1995), the empirical Bayes procedures of Johnstone and Silverman (2005) and Zhang (2005), and SureBlock in Cai and Zhou (2008a) can be carried over from the Gaussian regression to the general regression. Theoretical properties such as rates of convergence remain the same under the regression model (1) with suitable regularity conditions.

To illustrate the general ideas, we consider in the next two sections two important nonparametric problems under the model (1): adaptive estimation of the regression function f and robust estimation of the quadratic functional $Q(f) = \int f^2$. These examples show that for a given statistical problem it is easy to turn the case of nonparametric regression with general symmetric errors into the one with Gaussian noise and construct highly robust and adaptive procedures. Other robust inference problems can be handled in a similar fashion.

3　Robust wavelet regression

We consider in this section robust estimation of the regression function f under the model (1). Many estimation procedures have been developed in the literature for case where the errors ξ_i are assumed to be i.i.d. Gaussian. However, these procedures are not readily applicable when the noise distribution is unknown. In fact direct application of the procedures designed for the Gaussian case can fail badly if the noise is in fact heavy-tailed.

In this section we construct a robust procedure by following the general principles of the asymptotic equivalence theory. The estimator is robust, adaptive, and easily implementable. In particular, its performance is not sensitive to the error distribution.

3.1 Wavelet procedure for robust nonparametric regression

We begin with basic notation and definitions and then give a detailed description of our robust wavelet regression procedure.

Let $\{\phi,\ \psi\}$ be a pair of father and mother wavelets. The functions ϕ and ψ are assumed to be compactly supported and $\int \phi = 1$. Dilation and translation of ϕ and ψ generate an orthonormal wavelet basis. For simplicity in exposition, we work with periodized wavelet bases on $[0,1]$. Let

$$\phi_{j,k}^{p}(t) = \sum_{l=-\infty}^{\infty} \phi_{j,k}(t-l), \quad \psi_{j,k}^{p}(t) = \sum_{l=-\infty}^{\infty} \psi_{j,k}(t-l), \qquad \text{for } t \in [0,1]$$

where $\phi_{j,k}(t) = 2^{j/2}\phi(2^j t - k)$ and $\psi_{j,k}(t) = 2^{j/2}\psi(2^j t - k)$. The collection $\{\phi_{j_0,k}^{p},\ k = 1,\ldots,2^{j_0};\ \psi_{j,k}^{p},\ j \geq j_0 \geq 0, k = 1,...,2^j\}$ is then an orthonormal basis of $L^2[0,1]$, provided the primary resolution level j_0 is large enough to ensure that the support of the scaling functions and wavelets at level j_0 is not the whole of $[0,1]$. The superscript "p" will be suppressed from the notation for convenience. An orthonormal wavelet basis has an associated orthogonal Discrete Wavelet Transform (DWT) which transforms sampled data into the wavelet coefficients. See Daubechies (1992) and Strang (1992) for further details on wavelets and discrete wavelet transform. A square-integrable function f on $[0,1]$ can be expanded into a wavelet series:

$$f(t) = \sum_{k=1}^{2^{j_0}} \tilde{\theta}_{j_0,k}\phi_{j_0,k}(t) + \sum_{j=j_0}^{\infty}\sum_{k=1}^{2^j} \theta_{j,k}\psi_{j,k}(t) \tag{8}$$

where $\tilde{\theta}_{j,k} = \langle f, \phi_{j,k}\rangle$, $\theta_{j,k} = \langle f, \psi_{j,k}\rangle$ are the wavelet coefficients of f.

We now describe the robust regression procedure in detail. Let the sample $\{Y_i,\ i = 1,\ldots,n\}$ be given as in (1). Set $J = \left\lfloor \log_2 \frac{n}{\log^{1+b} n} \right\rfloor$ for some $b > 0$ and let $T = 2^J$. We first group the observations Y_i consecutively into T equi-length bins and then take the median of each bin. Denote the medians by $X = (X_1,\ldots,X_T)$. Apply the discrete wavelet transform to the binned medians X and let $U = T^{-\frac{1}{2}}WX$ be the empirical wavelet coefficients, where W is the discrete wavelet transformation matrix. Write

$$U = (\tilde{y}_{j_0,1}, \cdots, \tilde{y}_{j_0,2^{j_0}}, y_{j_0,1}, \cdots, y_{j_0,2^{j_0}}, \cdots, y_{J-1,1}, \cdots, y_{J-1,2^{J-1}})'. \tag{9}$$

Here $\tilde{y}_{j_0,k}$ are the gross structure terms at the lowest resolution level, and $y_{j,k}$ ($j = j_0, \cdots, J-1, k = 1, \cdots, 2^j$) are empirical wavelet coefficients at level j which represent fine structure at scale 2^j. Set

$$\sigma_n = \frac{1}{2h(0)\sqrt{n}}. \tag{10}$$

Then the empirical wavelet coefficients can be written as

$$y_{j,k} = \theta_{j,k} + \epsilon_{j,k} + \sigma_n z_{j,k} + \xi_{j,k}, \tag{11}$$

where $\theta_{j,k}$ are the true wavelet coefficients of f, $\epsilon_{j,k}$ are "small" deterministic approximation errors, $z_{j,k}$ are i.i.d. $N(0,1)$, and $\xi_{j,k}$ are some "small" stochastic errors. The asymptotic equivalence theory given in Section 2 indicates that both $\epsilon_{j,k}$ and $\xi_{j,k}$ are "negligible" and the calculations in Section 6 will show this is indeed the case. If these negligible errors are ignored then we have

$$y_{j,k} \approx \theta_{j,k} + \sigma_n z_{j,k} \quad \text{with} \quad z_{j,k} \overset{iid}{\sim} N(0,1) \tag{12}$$

which is the idealized Gaussian sequence model.

The BlockJS procedure introduced in Cai (1999) for Gaussian nonparametric regression is then applied to $y_{j,k}$ as if they are exactly distributed as in (12). More specifically, at each resolution level j, the empirical wavelet coefficients $y_{j,k}$ are grouped into nonoverlapping blocks of length L. Let $B_j^i = \{(j,k) : (i-1)L + 1 \le k \le iL\}$ and let $S_{j,i}^2 \equiv \sum_{(j,k)\in B_j^i} y_{j,k}^2$. Let $\hat{\sigma}_n^2$ be an estimator of σ_n^2 (see equation (16) for an estimator). Set $J_* = \left\lfloor \log_2 \frac{T}{\log^{1+b} n} \right\rfloor$. A modified James-Stein shrinkage rule is then applied to each block B_j^i with $j \le J_*$, i.e.,

$$\hat{\theta}_{j,k} = \left(1 - \frac{\lambda_* L \hat{\sigma}_n^2}{S_{j,i}^2} \right)_+ y_{j,k} \quad \text{for } (j,k) \in B_j^i, \tag{13}$$

where $\lambda_* = 4.50524$ is a constant satisfying $\lambda_* - \log \lambda_* = 3$. For the gross structure terms at the lowest resolution level j_0, we set $\hat{\tilde{\theta}}_{j_0,k} = \tilde{y}_{j_0,k}$. The estimate of f at the sample points $\{\frac{i}{T} : i = 1, \cdots, T\}$ is obtained by applying the inverse discrete wavelet transform (IDWT) to the denoised wavelet coefficients. That is, $\{f(\frac{i}{T}) : i = 1, \cdots, T\}$ is estimated by $\hat{f} = \{\widehat{f(\frac{i}{T})} : i = 1, \cdots, T\}$ with $\hat{f} = T^{\frac{1}{2}} W^{-1} \cdot \hat{\theta}$. The whole function f is estimated by

$$\hat{f}_n(t) = \sum_{k=1}^{2^{j_0}} \hat{\tilde{\theta}}_{j_0,k} \phi_{j_0,k}(t) + \sum_{j=j_0}^{J_*-1} \sum_{k=1}^{2^j} \hat{\theta}_{j,k} \psi_{j,k}(t). \tag{14}$$

Remark 1 An estimator of $h^{-2}(0)$ can be given by

$$\widehat{h}^{-2}(0) = \frac{8m}{T} \sum (X_{2k-1} - X_{2k})^2. \tag{15}$$

and the variance σ_n^2 is then estimated by

$$\hat{\sigma}_n^2 = \frac{1}{4\hat{h}^2(0)n} = \frac{2}{T^2} \sum (X_{2k-1} - X_{2k})^2. \tag{16}$$

It is shown in Section 6 that the estimator $\hat{\sigma}_n^2$ is an accurate estimate of σ_n^2.

13

The robust estimator $\hat{f}_n$ constructed above is easy to implement. Figure 2 below illustrate the main steps of the procedure. As a comparison, we also plotted the estimate obtained by applying directly the BlockJS procedure to the original noisy signal. It can be seen clearly that this wavelet procedure does not perform well in the case of heavy-tailed noise. Other standard wavelet procedures have similar performance qualitatively. On the other hand, the BlockJS procedure performs very well on the medians of the binned data.
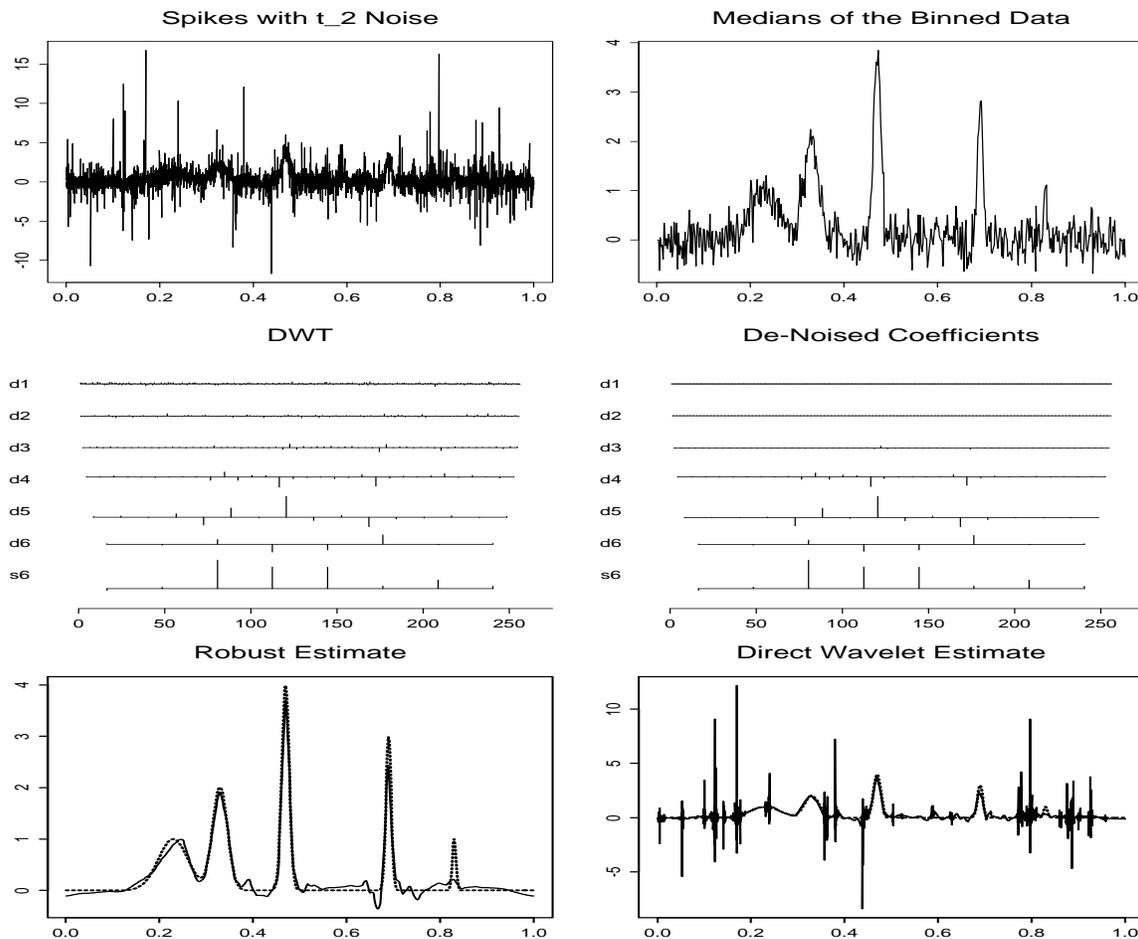

Figure 2: Top left panel: Noisy Spikes signal with sample size $n = 4096$ where the noise has t_2 distribution; Top right panel: The medians of the binned data with the bin size $m = 8$; Middle left panel: The discrete wavelet coefficients of the medians; Middle right panel: Blockwise thresholded wavelet coefficients of the medians; Bottom left panel: The robust estimate of the Spikes signal (dotted line is the true signal); Bottom right panel: The estimate obtained by applying directly the BlockJS procedure to the original noisy signal.

14

3.2 Adaptivity and Robustness of the Procedure

The robust regression procedure presented in Section 3.1 enjoys a high degree of adaptivity and robustness. We consider the theoretical properties of the procedure over the Besov spaces. For a given r-regular mother wavelet ψ with $r > \alpha$ and a fixed primary resolution level j_0, the Besov sequence norm $\| \cdot \|_{b_{p,q}^\alpha}$ of the wavelet coefficients of a function f is defined by

$$\|f\|_{b_{p,q}^\alpha} = \|\underline{\xi}_{j_0}\|_p + \left(\sum_{j=j_0}^{\infty} (2^{js} \|\underline{\theta}_j\|_p)^q \right)^{\frac{1}{q}} \tag{17}$$

where $\underline{\xi}_{j_0}$ is the vector of the father wavelet coefficients at the primary resolution level j_0, $\underline{\theta}_j$ is the vector of the wavelet coefficients at level j, and $s = \alpha + \frac{1}{2} - \frac{1}{p} > 0$. Note that the Besov function norm of index (α, p, q) of a function f is equivalent to the sequence norm (17) of the wavelet coefficients of the function. See Meyer (1992), Triebel (1983) and DeVore and Popov (1988) for further details on Besov spaces. We define

$$B_{p,q}^\alpha (M) = \left\{ f; \|f\|_{b_{p,q}^\alpha} \le M \right\}. \tag{18}$$

In the case of Gaussian noise the minimax rate of convergence for estimating f over the Besov body $B_{p,q}^\alpha(M)$ is $n^{-2\alpha/(1+2\alpha)}$. See Donoho and Johnstone (1998).

We shall consider the following collection of error distributions. For $0 < \epsilon_1 < 1$, $\epsilon_i > 0$, $i = 2, 3, 4$, let

$$\mathcal{H}_{\epsilon_1, \epsilon_2} = \left\{ h : \int_{-\infty}^0 h(x) = \frac{1}{2}, \epsilon_1 \le h(0) \le \frac{1}{\epsilon_1}, |h(x) - h(0)| \le \frac{x^2}{\epsilon_1} \text{ for all } |x| < \epsilon_2 \right\} \tag{19}$$

and define $\mathcal{H} = \mathcal{H}(\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4)$ by

$$\mathcal{H} = \{ h \in \mathcal{H}_{\epsilon_1, \epsilon_2} : \int |x|^{\epsilon_3} h(x) dx < \epsilon_4, h(x) = h(-x), |h^{(3)}(x)| \le \epsilon_4 \text{ for } |x| \le \epsilon_3 \}. \tag{20}$$

The assumption $\int |x|^{\epsilon_3} h(x) dx < \epsilon_4$ guarantees that the moments of the median of the binned data are well approximated by those of the normal random variable. Note that this assumption is satisfied by a large collection of distributions including Cauchy distribution.

The following theorem shows that our estimator achieves optimal global adaptation for a wide range of Besov balls $B_{p,q}^\alpha (M)$ defined in (18) and uniformly over the family of error distributions given in (20).

Theorem 3 *Suppose the wavelet ψ is r-regular. Then the estimator $\hat{f}_n$ defined in (14) satisfies, for $p \ge 2$, $\alpha \le r$ and $\frac{2\alpha^2}{1+2\alpha} > \frac{1}{p}$*

$$\sup_{h \in \mathcal{H}} \sup_{f \in B_{p,q}^\alpha(M)} E\|\hat{f}_n - f\|_2^2 \le Cn^{-\frac{2\alpha}{1+2\alpha}},$$

and for $1 \leq p < 2$, $\alpha \leq r$ and $\frac{2\alpha^2}{1+2\alpha} > \frac{1}{p}$,

$$\sup_{h \in \mathcal{H}} \sup_{f \in B_{p,q}^\alpha(M)} E\|\widehat{f}_n - f\|_2^2 \leq Cn^{-\frac{2\alpha}{1+2\alpha}} (\log n)^{\frac{2-p}{p(1+2\alpha)}}.$$

In addition to global adaptivity, the estimator also enjoys a high degree of local spatial adaptivity. For a fixed point $t_0 \in (0,1)$ and $0 < \alpha \leq 1$, define the local Hölder class $\Lambda^\alpha(M, t_0, \delta)$ by

$$\Lambda^\alpha(M, t_0, \delta) = \{f : |f(t) - f(t_0)| \leq M |t - t_0|^\alpha, \text{ for } t \in (t_0 - \delta,\ t_0 + \delta)\}.$$

If $\alpha > 1$, then

$$\Lambda^\alpha(M, t_0, \delta) = \{f : |f^{(\lfloor \alpha \rfloor)}(t) - f^{(\lfloor \alpha \rfloor)}(t_0)| \leq M |t - t_0|^{\alpha'} \text{ for } t \in (t_0 - \delta,\ t_0 + \delta)\}$$

where $\lfloor \alpha \rfloor$ is the largest integer less than α and $\alpha' = \alpha - \lfloor \alpha \rfloor$.

In Gaussian nonparametric regression setting, it is well known that the optimal rate of convergence for estimating $f(t_0)$ over $\Lambda^\alpha(M, t_0, \delta)$ with α completely known is $n^{-2\alpha/(1+2\alpha)}$. On the other hand, when α is unknown, Lepski (1990) and Brown and Low (1996) showed that the local adaptive minimax rate over the Hölder class $\Lambda^\alpha(M, t_0, \delta)$ is $(\log n / n)^{2\alpha/(1+2\alpha)}$. So one has to pay at least a logarithmic factor for adaptation.

Theorem 4 below shows that our estimator achieves optimal local adaptation with the minimal cost uniformly over the family of noise distributions defined in (20).

Theorem 4 *Suppose the wavelet ψ is r-regular with $r \geq \alpha > 0$. Let $t_0 \in (0,1)$ be fixed. Then the estimator $\widehat{f}_n$ defined in (14) satisfies*

$$\sup_{h \in \mathcal{H}} \sup_{f \in \Lambda^\alpha(M, t_0, \delta)} E(\widehat{f}_n(t_0) - f(t_0))^2 \leq C \cdot \left(\frac{\log n}{n}\right)^{\frac{2\alpha}{1+2\alpha}}. \tag{21}$$

Remark 2 Note that in the general asymptotic equivalence theory given in Section 2 the bin size was chosen to be $n^{1/3} \log n$. However, for specific estimation problems such as robust estimation of f discussed in this section, the bin size can be chosen differently. Here we choose a small bin size $\log^{1+b} n$. There is a significant advantage in choosing such a small bin size in this problem. Note that the smoothness assumptions for α in Theorems 3 and 4 are different from those in Theorems 3 and 4 in Brown, Cai and Zhou (2008). For example, in Theorem 4 in Brown, Cai and Zhou (2008) it was assumed $\alpha > 1/6$, but now we need only $\alpha > 0$ due to the choice of the small bin size.

4 Robust estimation of the quadratic functional $\int f^2$

An important nonparametric estimation problem is that of estimating the quadratic functional $Q(f) = \int f^2$. This problem is interesting in its own right and closely related to the construction of confidence balls and nonparametric hypothesis testing in nonparametric function estimation. See for example Li (1989), Dümbgen (1998), Spokoiny (1998), Genovese and Wasserman (2005), and Cai and Low (2006a). In addition, as shown in Bickel and Ritov (1988), Donoho and Nussbaum (1990) and Fan (1991), this problem connects the nonparametric and semiparametric literatures.

This problem has been well studied in the Gaussian noise setting. See, for example, Donoho and Nussbaum (1990), Fan (1991). Efromovich and Low (1996), Spokoiny (1999), Laurent and Massart (2000), Klemelä (2002) and Cai and Low (2005 and 2006b). In this section we shall consider robust estimation of the quadratic functional $Q(f)$ under the regression model (1) with an unknown symmetric error distribution. We shall follow the same notation used in Section 3. Note that the orthonormality of the wavelet basis implies the isometry between the L_2 function norm and the ℓ_2 wavelet sequence norm which yields

$$Q(f) = \int f^2 = \sum_{k=1}^{2^{j_0}} \tilde{\theta}_{j_0,k}^2 + \sum_{j=j_0}^{\infty} \sum_{k=1}^{2^j} \theta_{j,k}^2.$$

The problem of estimating $Q(f)$ is then translated into estimating the squared coefficients.

We consider adaptively estimating $Q(f)$ over Besov balls $B_{p,q}^{\alpha}(M)$ with $\alpha > \frac{1}{p} + \frac{1}{2}$. We shall show that it is in fact possible to find a simple procedure which is asymptotically rate optimal simultaneously over a large collection of unknown symmetric error distributions. In this sense, the procedure is robust.

As in Section 3 we group the observations Y_i into T bins of size $\log^{1+b}(n)$ for some $b > 0$ and then take the median of each bin. Let $X = (X_1, ..., X_T)$ denote the binned medians and let $U = T^{-\frac{1}{2}}WX$ be the empirical wavelet coefficients, where W is the discrete wavelet transformation matrix. Write U as in (9). Then the empirical wavelet coefficients can be approximately decomposed as in (12):

$$\tilde{y}_{j_0,k} \approx \tilde{\theta}_{j_0,k} + \sigma_n \tilde{z}_{j_0,k} \quad \text{and} \quad y_{j,k} \approx \theta_{j,k} + \sigma_n z_{j,k} \tag{22}$$

where $\sigma_n = 1/(2h(0)\sqrt{n})$ and $\tilde{z}_{j_0,k}$ and $z_{j,k}$ are i.i.d. standard normal variables.

The quadratic functional $Q(f)$ can then be estimated as if we have exactly the idealized sequence model (22). More specifically, let $J_q = \lfloor \log_2 \sqrt{n} \rfloor$ and set

$$\hat{Q} = \sum_{k=1}^{2^{j_0}} (\tilde{y}_{j_0,k}^2 - \hat{\sigma}_n^2) + \sum_{j=j_0}^{J_q} \sum_{k=1}^{2^j} (y_{j,k}^2 - \hat{\sigma}_n^2). \tag{23}$$

The following theorem shows that this estimator is robust and rate-optimal for a large collection of symmetric error distributions and a wide range of Besov classes simultaneously.

Theorem 5 *For all Besov balls $B_{p,q}^{\alpha}(M)$ with $\alpha > \frac{1}{p} + \frac{1}{2}$, the estimator $\hat{Q}$ given in (23) satisfies*

$$\sup_{f \in B_{p,q}^{\alpha}(M)} E_f(\hat{Q} - Q(f))^2 \leq \frac{M^2}{h^2(0)} n^{-1}(1 + o(1)). \qquad (24)$$

Remark 3 We should note that there is a slight tradeoff between efficiency and robustness. When the error distribution is known to be Gaussian, it is possible to construct a simple procedure which is efficient, asymptotically attaining the exact minimax risk $4M^2 n^{-1}$. See for example, Cai and Low (2005). In the Gaussian case, the upper bound in (24) is $2\pi M^2 n^{-1}$ which is slightly larger than $4M^2 n^{-1}$. On the other hand, our procedure is robust over a large collection of unknown symmetric error distributions.

The examples of adaptive and robust estimation of the regression function and the quadratic functional given in the last and this sections illustrate the practical use of the general principles in the asymptotic equivalence theory given in Section 2. It is easy to see that other nonparametric inference problems such as the construction of confidence sets and nonparametric hypothesis testing under the general nonparametric regression model (1) can be handled in a similar way. Hence our approach can be viewed as a general method for robust nonparametric inference.

5   Technical tools: Moderate deviation and quantile coupling for median

Quantile coupling is an important technical tool in probability and statistics. For example, the celebrated KMT coupling results given in Komlós, Major and Tusnády (1975) plays a key role in the Hungarian construction in the asymptotic equivalence theory. See, e.g., Nussbaum (1996). Standard coupling inequalities are mostly focused on the coupling of the mean of i.i.d. random variables with a normal variable. Brown, Cai and Zhou (2008) studied the coupling of a median statistic with a normal variable. For the asymptotic equivalence theory given in Section 2 and the proofs of the theoretical results in Section 3 we need a more refined moderate deviation result for the median and an improved coupling inequality than those given in Brown, Cai and Zhou (2008). This improvement plays a crucial role in this paper. It is the main tool for reducing the problem of robust regression with unknown symmetric noise to a well studied and relatively simple problem

of Gaussian regression. The result here may be of independent interest because of the fundamental role played by the median in statistics.

Let X be a random variable with distribution G, and Y with a continuous distribution F. Define

$$\widetilde{X} = G^{-1}\left(F\left(Y\right)\right), \tag{25}$$

where $G^{-1}\left(x\right) = \inf\left\{u:\ G\left(u\right) \geq x\right\}$, then $\mathcal{L}\left(\widetilde{X}\right) = \mathcal{L}\left(X\right)$. Note that $\widetilde{X}$ and Y are now defined on the same probability space. This makes it possible to give a pointwise bound between $\widetilde{X}$ and Y. For example, one can couple Binomial$(m, 1/2)$ and $N(m/2, m/4)$ distributions. Let $X = 2\left(W - m/2\right)/\sqrt{m}$ with $W \sim$Binomial$(m, 1/2)$ and $Y \sim N(0, 1)$, and let $\widetilde{X}\left(Y\right)$ be defined as in equation (25). Komlós, Major and Tusnády (1975) showed that for some constant $C > 0$ and $\varepsilon > 0$, when $|\widetilde{X}| \leq \varepsilon\sqrt{m}$,

$$|\widetilde{X} - Y| \leq \frac{C}{\sqrt{m}} + \frac{C}{\sqrt{m}}|\widetilde{X}|^2. \tag{26}$$

Let $\xi_1, \ldots, \xi_m$ be i.i.d. random variables with density function h. Denote the sample median by ξ_{med}. The classical theory shows that the limiting distribution of $2h(0)\sqrt{m}\xi_{med}$ is $N(0, 1)$. We will construct a new random variable $\widetilde{\xi}_{med}$ by using quantile coupling in (25) such that $\mathcal{L}(\widetilde{\xi}_{med}) = \mathcal{L}\left(\xi_{med}\right)$ and show that $\widetilde{\xi}_{med}$ can be well approximated by a normal random variable as in equation (26). Denote the distribution and density function the sample median ξ_{med} by G and g respectively. We obtain an improved approximation of the density g by a normal density which leads to a better moderate deviation result for the distribution of sample median and consequently improve the classical KMT bound from the rate $1/\sqrt{m}$ to $1/m$. A general theory for improving the classical quantile coupling bound was given in Zhou (2006).

Theorem 6 *Let $Z \sim N(0, 1)$ and let $\xi_1, \ldots, \xi_m$ be i.i.d. with density function h where $m = 2k + 1$ for some integer $k \geq 1$. Let Assumption (A2) hold. Then for $|x| \leq \varepsilon$*

$$g\left(x\right) = \frac{\sqrt{8k}f\left(0\right)}{\sqrt{2\pi}}\exp\left(-8kh^2\left(0\right)x^2/2 + O\left(kx^4 + k^{-1}\right)\right), \tag{27}$$

and for $0 < x < \varepsilon$

$$G\left(-x\right) = \Phi\left(-x\right)\exp\left(O\left(kx^4 + k^{-1}\right)\right), \ \text{and} \ \overline{G}\left(x\right) = \overline{\Phi}\left(x\right)\exp\left(O\left(kx^4 + k^{-1}\right)\right) \tag{28}$$

where $\overline{G}\left(x\right) = 1 - G\left(x\right)$, and $\overline{\Phi}\left(x\right) = 1 - \Phi\left(x\right)$. Consequently for every m there is a mapping $\widetilde{\xi}_{med}\left(Z\right): \mathbb{R} \mapsto \mathbb{R}$ such that $\mathcal{L}\left(\widetilde{\xi}_{med}\left(Z\right)\right) = \mathcal{L}\left(\xi_{med}\right)$ and

$$|2h(0)\sqrt{m}\widetilde{\xi}_{med} - Z| \leq \frac{C}{m} + \frac{C}{m}|2h(0)\sqrt{m}\widetilde{\xi}_{med}|^3, \ \text{when} \ \left|\widetilde{\xi}_{med}\right| \leq \varepsilon \tag{29}$$

19

and

$$|2h(0)\sqrt{m}\widetilde{\xi}_{med} - Z| \leq \frac{C}{m}\left(1 + |Z|^3\right), \text{ when } |Z| \leq \varepsilon\sqrt{m} \qquad (30)$$

where $C, \varepsilon > 0$ *depend on* h *but not on* m.

Remark 4 In Brown, Cai and Zhou (2008) the density g of the sample median was approximated by a normal density as

$$g(x) = \frac{\sqrt{8k}h(0)}{\sqrt{2\pi}}\exp\left(-8kh^2(0)x^2/2 + O\left(k|x|^3 + |x| + k^{-1}\right)\right), \text{ for } |x| \leq \varepsilon.$$

Since $\xi_{med} = O_p(1/\sqrt{m})$, the approximation error $O\left(k|x|^3 + |x| + k^{-1}\right)$ is at the level of $1/\sqrt{m}$. In comparison, the approximation error $O\left(kx^4 + k^{-1}\right)$ in equation (27) is at the level of $1/m$. This improvement is necessary for establishing (36) in the proof of Theorem 2, and leads to an improved quantile coupling bound (30) over the bound obtained in Brown, Cai and Zhou (2008)

$$|2h(0)\sqrt{m}\widetilde{\xi}_{med} - Z| \leq \frac{C}{\sqrt{m}} + \frac{C}{\sqrt{m}}Z^2, \text{ when } \left|\widetilde{\xi}_{med}\right| \leq \varepsilon.$$

Since Z is at a constant level, we improve the bound from a classical rate $1/\sqrt{m}$ to $1/m$.

Although the result is only given to m odd, it can be easily extended to the even case as discussed in Remark 1 of Brown, Cai and Zhou (2008). The coupling result given in Theorem 6 in fact holds uniformly for the whole family of $h \in \mathcal{H}_{\epsilon_1,\epsilon_2}$.

Theorem 7 *Let* $\xi_1, \ldots, \xi_m$ *be i.i.d. with density* $h \in \mathcal{H}_{\epsilon_1,\epsilon_2}$ *in equation (19). For every* $m = 2k + 1$ *with integer* $k \geq 1$, *there is a mapping* $\widetilde{\xi}_{med}(Z) : \mathbb{R} \mapsto \mathbb{R}$ *such that* $\mathcal{L}\left(\widetilde{\xi}_{med}(Z)\right) = \mathcal{L}(\xi_{med})$ *and for two constants* $C_{\epsilon_1,\epsilon_2}$, $\varepsilon_{\epsilon_1,\epsilon_2} > 0$ *depending only on* ϵ_1 *and* ϵ_2

$$|2h(0)\sqrt{m}\widetilde{\xi}_{med} - Z| \leq \frac{C_{\epsilon_1,\epsilon_2}}{m} + \frac{C_{\epsilon_1,\epsilon_2}}{m}|2h(0)\sqrt{m}\widetilde{\xi}_{med}|^3, \text{ when } \left|\widetilde{\xi}_{med}\right| \leq \varepsilon_{\epsilon_1,\epsilon_2} \qquad (31)$$

and

$$|2h(0)\sqrt{m}\widetilde{\xi}_{med} - Z| \leq \frac{C_{\epsilon_1,\epsilon_2}}{m} + \frac{C_{\epsilon_1,\epsilon_2}}{m}|Z|^3, \text{ when } |Z| \leq \varepsilon_{\epsilon_1,\epsilon_2}\sqrt{m}$$

uniformly over all $h \in \mathcal{H}_{\epsilon_1,\epsilon_2}$.

6 Proofs

We shall prove the main results in the order of Theorems 6 and 7, Theorems 1 and 2, Theorems 3, and then Theorem 5. Theorems 6 and 7 provide important technical tools for the proof of the rest of the theorems.

In this section C denotes a positive constant not depending on n that may vary from place to place and we set $d \equiv \min\left(\alpha - \frac{1}{p}, 1\right)$.

6.1 Proof of Theorems 6 And 7

We only prove (27) and (28). it follows from Zhou (2006) that the moderate deviation bound (28) implies the coupling bounds (29) and (30). Let $H(x)$ be the distribution function of ξ_1. The density of the median $\xi_{(k+1)}$ is

$$g(x) = \frac{(2k+1)!}{(k!)^2} H^k(x)(1-H(x))^k h(x).$$

Stirling formula, $j! = \sqrt{2\pi} j^{j+1/2} \exp(-j + \epsilon_j)$ with $\epsilon_j = O(1/j)$, gives

$$
\begin{aligned}
g(x) &= \frac{(2k+1)!}{4^k (k!)^2} [4H(x)(1-H(x))]^k h(x) \\
&= \frac{2\sqrt{2k+1}}{e\sqrt{2\pi}} \left(\frac{2k+1}{2k}\right)^{2k+1} [4H(x)(1-H(x))]^k h(x) \exp\left(O\left(\frac{1}{k}\right)\right).
\end{aligned}
$$

It is easy to see $\left|\sqrt{2k+1}/\sqrt{2k} - 1\right| \le k^{-1}$, and

$$\left(\frac{2k+1}{2k}\right)^{2k+1} = \exp\left(-(2k+1)\log\left(1 - \frac{1}{2k+1}\right)\right) = \exp\left(1 + O\left(\frac{1}{k}\right)\right).$$

Then we have, when $0 < H(x) < 1$,

$$g(x) = \frac{\sqrt{8k}}{\sqrt{2\pi}} [4H(x)(1-H(x))]^k h(x) \exp\left(O\left(\frac{1}{k}\right)\right). \tag{32}$$

From the assumption in the theorem, Taylor expansion gives

$$
\begin{aligned}
4H(x)(1-H(x)) &= 1 - 4(H(x) - H(0))^2 \\
&= 1 - 4\left[\int_0^x (h(t) - h(0))\,dt + h(0)x\right]^2 \\
&= 1 - 4\left(h(0)x + O\left(|x|^3\right)\right)^2
\end{aligned}
$$

for $0 \le |x| \le \varepsilon$, i.e., $\log(4H(x)(1-H(x))) = -4h^2(0)x^2 + O(x^4)$ when $|x| \le 2\varepsilon$ for some $\varepsilon > 0$. Here ε is chosen sufficiently small so that $h(x) > 0$ for $|x| \le 2\varepsilon$. Assumption (A2) also implies $\frac{h(x)}{h(0)} = 1 + O(x^2) = \exp(O(x^2))$ for $|x| \le 2\varepsilon$. Thus for $|x| \le 2\varepsilon$

$$
\begin{aligned}
g(x) &= \frac{\sqrt{8k}h(0)}{\sqrt{2\pi}} \exp\left(-8kh^2(0)x^2/2 + O\left(kx^4 + x^2 + k^{-1}\right)\right) \\
&= \frac{\sqrt{8k}h(0)}{\sqrt{2\pi}} \exp\left(-8kh^2(0)x^2/2 + O\left(kx^4 + k^{-1}\right)\right).
\end{aligned}
$$

Now we approximate the distribution function of ξ_{med} by a normal distribution. Without loss of generality we assume $h(0) = 1$. We write

$$g(x) = \frac{\sqrt{8k}}{\sqrt{2\pi}} \exp\left(-8kx^2/2 + O\left(kx^4 + k^{-1}\right)\right), \text{ for } |x| \le 2\varepsilon.$$

Now we use this approximation of density functions to give the desired approximation of distribution functions. Specifically we shall show

$$G(x) = \int_{-\infty}^{x} g(t)\, dt \leq \Phi\left(\sqrt{8k}x\right) \exp\left(C\left(kx^4 + k^{-1}\right)\right) \tag{33}$$

and

$$G(x) \geq \Phi\left(\sqrt{8k}x\right) \exp\left(-C\left(kx^4 + k^{-1}\right)\right) \tag{34}$$

for all $-\varepsilon \leq x \leq 0$ and some $C > 0$. The proof for $0 \leq x \leq \varepsilon$ is similar. We now prove the inequality (33). Note that

$$\left(\Phi\left(\sqrt{8k}x\right)\exp\left(C\left(kx^4 + k^{-1}\right)\right)\right)' \tag{35}$$
$$= \sqrt{8k}\varphi\left(\sqrt{8k}x\right)\exp\left(C\left(kx^4 + k^{-1}\right)\right) + \Phi\left(\sqrt{8k}x\right)4kCx^3\exp\left(C\left(kx^4 + k^{-1}\right)\right).$$

From the Mill's ratio inequality we have $\Phi\left(\sqrt{8k}x\right)\left(-\sqrt{8k}x\right) < \varphi\left(\sqrt{8k}x\right)$ and hence

$$\Phi\left(\sqrt{8k}x\right)\left(4Ckx^3\right)\exp\left(C\left(kx^4 + k^{-1}\right)\right) \geq \sqrt{8k}\varphi\left(\sqrt{8k}x\right)\left(-\frac{C}{2}x^2\right)\exp\left(C\left(kx^4 + k^{-1}\right)\right).$$

This and (35) yield

$$\left(\Phi\left(\sqrt{8k}x\right)\exp\left(C\left(kx^4 + k^{-1}\right)\right)\right)' \geq \sqrt{8k}\varphi\left(\sqrt{8k}x\right)\left(1 - \frac{C}{2}x^2\right)\exp\left(C\left(kx^4 + k^{-1}\right)\right)$$
$$\geq \sqrt{8k}\varphi\left(\sqrt{8k}x\right)\exp\left(-Cx^2\right)\exp\left(C\left(kx^4 + k^{-1}\right)\right)$$
$$\geq \sqrt{8k}\varphi\left(\sqrt{8k}x\right)\exp\left(C\left(kx^4 + k^{-1}\right)/4\right).$$

Here in the second inequality we apply $1 - t/2 \geq \exp(-t)$ when $0 < t < 1/2$. Thus we have

$$\left(\Phi\left(\sqrt{8k}x\right)\exp\left(C\left(kx^4 + k^{-1}\right)\right)\right)' \geq \sqrt{8k}\varphi\left(\sqrt{8k}x\right)\exp\left(C\left(kx^4 + k^{-1}\right)\right)$$

for C sufficiently large and for $-2\varepsilon \leq x \leq 0$. Then

$$\int_{-2\varepsilon}^{x} g(t)\, dt \leq \int_{-2\varepsilon}^{x}\left(\Phi\left(\sqrt{8k}t\right)\exp\left(C\left(kx^4 + k^{-1}\right)\right)\right)'$$
$$= \left[\begin{array}{c} \Phi\left(\sqrt{8k}x\right)\exp\left(C\left(kx^4 + k^{-1}\right)\right) \\ -\Phi\left(\sqrt{8k}\cdot(2\varepsilon)\right)\exp\left(C\left(k\left(2\varepsilon\right)^4 + k^{-1}\right)\right) \end{array}\right]$$
$$\leq \Phi\left(\sqrt{8k}x\right)\exp\left(C\left(kx^4 + k^{-1}\right)\right).$$

In (32) we see

$$\int_{-\infty}^{-2\varepsilon} g(t)\, dt = \int_{-\infty}^{-2\varepsilon} \frac{(2k+1)!}{(k!)^2} H^k(t)(1-H(t))^k h(t)\, dt = \int_{0}^{H(-2\varepsilon)} \frac{(2k+1)!}{(k!)^2} u^k(1-u)^k\, du$$

$$= o\left(k^{-1}\right) \int_{H(-3\varepsilon/2)}^{H(-\varepsilon)} \frac{(2k+1)!}{(k!)^2} u^k(1-u)^k\, du \leq o\left(k^{-1}\right) \int_{H(-2\varepsilon)}^{H(x)} \frac{(2k+1)!}{(k!)^2} u^k(1-u)^k\, du$$

$$= o\left(k^{-1}\right) \int_{-2\varepsilon}^{x} g(t)\, dt$$

where the third equality is from the fact that $u_1^k(1-u_1)^k = o\left(k^{-1}\right) u_2^k(1-u_2)^k$ uniformly for $u_1 \in [0, H(-2\varepsilon)]$ and $u_2 \in [H(-3\varepsilon/2), H(-\varepsilon)]$. Thus we have

$$G(x) \leq \Phi\left(\sqrt{8k}x\right) \exp\left(Ckx^4 + Ck^{-1}\right).$$

which is Equation (33). Equation (34) can be established in a similar way. ∎

Remark: Note that in the proof of Theorem 6 it can be seen easily that constants C and ϵ in equation (29) depends only on the ranges of $h(0)$ and the bound of Lipschitz constants of h at a fixed open neighborhood of 0. Theorem 7 then follows from the proof of Theorem 6 together with this observation.

6.2 Proof of Theorems 1 And 2

Proof of Theorem 1: Let ε_n be a sequence approaching to 0 slowly, for example $\varepsilon_n = 1/\log\log n$. Let $p_{f,n}$ be the joint density of $Y_i's$ and $p_{f,n}^*$ be the joint density of $Y_i^{*'}s$. And let $P_{f,n}$ be the joint distribution of $Y_i's$ and $P_{f^*,n}$ be the joint distribution of $Y_i^{*'}s$. We want to show that

$$\max\left\{P_{f,n}\left(|1 - p_{f^*,n}/p_{f,n}| \geq \varepsilon_n\right), P_{f^*,n}\left(|1 - p_{f,n}/p_{f^*,n}| \geq \varepsilon_n\right)\right\}$$

decays exponentially fast uniformly over the function space.

Note that $P_{f,n}\left(|1 - p_{f^*,n}/p_{f,n}| \geq \varepsilon_n\right) = P_{0,n}\left(|1 - p_{f^*-f,n}/p_{0,n}| \geq \varepsilon_n\right)$. It suffices to show that $P_{0,n}\left(|\log\left(p_{f^*-f,n}/p_{0,n}\right)| \geq \varepsilon_n\right)$ decays exponentially fast. Write

$$\log\left(p_{f^*-f,n}/p_{0,n}\right) = \sum_{i=1}^{n} \log\frac{h(\xi_i - a_i)}{h(\xi_i)}$$

with $a_i = f^*(i/n) - f(i/n)$, where ξ_i has density $h(x)$. Under Assumption (A1), we have $\mathbb{E}r_{a_i}(\xi_i) \leq Ca_i^2$ and $\mathbb{E}\exp\left[t\left(r_{a_i}(\xi_i) - \mu(a_i)\right)\right] \leq \exp\left(Ct^2 a_i^2\right)$ which imply

$$P_{0,n}\left(\exp\left[t\sum_{i=1}^{n} r_{a_i}(\xi_i) - \mu(a_i)\right] \geq \exp\left(t\varepsilon_n\right)\right) \leq \exp\left(Ct^2 \sum_{i=1}^{n} a_i^2 - t\varepsilon_n\right).$$

Since

$$\sum_{i=1}^{n} a_i^2 \le C_1 n \cdot \left(\frac{n^{4/3}}{\log^2 n}\right)^{-d} = C_1 n^{1-4d/3} \log^{2d} n,$$

which goes to zero for $d > 3/4$, by setting $t = n^{(4d/3-1)/2}$ the Markov inequality implies that $P_{0,n}\left(|\log\left(p_{f^*-f,n}/p_{0,n}\right)| \ge \varepsilon_n\right)$ decays exponentially fast. ∎

Proof of Theorem 2: Let $g_{f,n}$ be the joint density of X_j^*'s and $q_{f,n}$ be the joint density of X_j^{**}'s. And let $G_{0,n}$ be the joint distribution of η_j's and $Q_{0,n}$ be the joint distribution of Z_j's. Theorem 6 yields

$$g\left(x\right) = \frac{\sqrt{4mh}\left(0\right)}{\sqrt{2\pi}} \exp\left(-4mh^2\left(0\right)x^2/2 + O\left(mx^4 + m^{-1}\right)\right)$$

for $|x| \le m^{-1/3}$. Since $G_{0,n}\left(|\eta_j| > m^{-1/3}\right)$ and $Q_{0,n}\left(|Z_j| > m^{-1/3}\right)$ decay exponentially fast, it suffices to study

$$\sum_{i=1}^{T} \log \frac{g\left(Z_j\right)}{\phi_{\sigma_m}\left(Z_j\right)} I\left(|Z_j| \le m^{-1/3}\right).$$

Let

$$l\left(Z_j\right) = \log \frac{g\left(Z_j\right)}{\phi_{\sigma_m}\left(Z_j\right)} I\left(|Z_j| \le m^{-1/3}\right)$$

with Z_j normally distributed with density $\phi_{\sigma_m}\left(x\right)$. It can be easily shown that

$$\mathbb{E}l\left(Z_j\right) \le CQ_{0,n}\left(1 - \frac{g\left(Z_j\right)}{\phi_{\sigma_m}\left(Z_j\right)}\right)^2 \le C_1 m^{-2}$$

and

$$Var\left(l\left(Z_j\right)\right) \le Cm^{-2}.$$

Since $|Z_j| \le Cm^{-1/3}$, then $|l\left(Z_j\right)| \le Cm^{-1/3}$. Taylor expansion gives

$$\mathbb{E}\exp\left[t\left(l\left(Z_j\right) - \mathbb{E}l\left(Z_j\right)\right)\right] \le \exp\left(Ct^2 m^{-2}\right)$$

for $t = \log^{3/2} n$, then similar to the Proof of Theorems 1 we have

$$Q_{f,n}\left(|\log\left(g_{f,n}/q_{f,n}\right)| \ge \varepsilon_n\right) \le \exp\left(Ct^2 Tm^{-2} - t\varepsilon_n\right). \tag{36}$$

Since $Tm^{-2} = 1/\log^3 n \to 0$, it decays faster than any polynomial of n. ∎

6.3 Proof of Theorems 3 and 4

In the proofs of Theorems 3, 4 and 5, we shall replace $\hat{\sigma}_n^2$ by σ_n^2. We assume that $h(0)$ is known and equal to 1 without loss of generality, since it can be shown easily that the estimator $\hat{h}(0)$ given in (15) satisfies

$$P\left\{\left|\hat{h}^{-2}(0) - h^{-2}(0)\right| > n^{-\delta}\right\} \le c_l n^{-l} \tag{37}$$

for some $\delta > 0$ and all constants $l \ge 1$. Note that $E\frac{m}{T/2}\sum(X_{2k-1}-X_{2k})^2 = \frac{1}{4}h^{-2}(0) + O\left(\sqrt{m}T^{-d}\right)$, and it is easy to show

$$E\left|\frac{8m}{T}\sum(X_{2k-1}-X_{2k})^2 - h^{-2}(0)\right|^l \le C_l\left(\sqrt{m}T^{-d}\right)^l$$

where $\sqrt{m}T^{-d} = n^{-\delta}$ with $\delta > 0$ in our assumption. Then equation (37) holds by Chebyshev inequality. It is very important to see that the asymptotic risk properties of our estimators for f in (13) and $Q(f)$ in (23) do not change when replacing σ_n^2 by $\sigma_n^2(1 + O(n^{-\delta}))$, thus in our analysis we may just assume that $h(0)$ is known without loss of generality.

For simplicity we shall assume that n is divisible by T in the proof. The coupling inequality and the fact that a Besov ball $B_{p,q}^\alpha(M)$ can be embedded into a Hölder ball with smoothness $d = \min\left(\alpha - \frac{1}{p}, 1\right) > 0$ (see Meyer (1992)) enable us to precisely control of the errors. Proposition 2 gives the bounds for both the deterministic and stochastic errors.

Proposition 2 *Let X_j be given as in our procedure and let $f \in B_{p,q}^\alpha(M)$. Then X_j can be written as*

$$\sqrt{m}X_j = \sqrt{m}f\left(\frac{j}{T}\right) + \frac{1}{2}Z_j + \epsilon_j + \zeta_j \tag{38}$$

where

(i). $Z_j \overset{i.i.d.}{\sim} N\left(0, \frac{1}{h^2(0)}\right)$;

(ii). ϵ_j *are constants satisfying* $|\epsilon_j| \le C\sqrt{m}T^{-d}$ *and so* $\frac{1}{n}\sum_{i=1}^T \epsilon_j^2 \le CT^{-2d}$;

(iii). ζ_j *are independent and "stochastically small" random variables satisfying with* $E\zeta_j = 0$, *and can be written as*

$$\zeta_j = \zeta_{j1} + \zeta_{j2} + \zeta_{j3}$$

with

$$
\begin{aligned}
|\zeta_{j1}| &\le C\sqrt{m}T^{-d} \\
E\zeta_{j2} &= 0 \text{ and } |\zeta_{2j}| \le \frac{C}{m}\left(1 + |Z_j|^3\right) \\
P(\zeta_{j3} = 0) &\ge 1 - C\exp(-\varepsilon m) \text{ and } E|\zeta_{j3}|^D \text{ exists}
\end{aligned}
$$

25

for some $\varepsilon > 0$ and $C > 0$, and all $D > 0$.

Remark 5 Equation (38) is different from Proposition 1 in Brown, Cai and Zhou (2008), where there is an additional bias term $\sqrt{m}b_m$. Lemma 5 in Brown, Cai and Zhou (2008) showed that the bias b_m can be estimated with a rate $\max\left\{T^{-2d}, m^{-4}\right\}$. Therefore in that paper we need to choose the bin size $m = n^{1/4}$ such that $m^{-4} = o\left(n^{-2\alpha/(2\alpha+1)}\right)$ is negligible relative to the minimax risk. In the present paper we can choose $m = \log^{1+b} n$ because there is no bias term and as a result the condition on the smoothness is relaxed.

$\underline{\text{Proof of Proposition 2:}}$ Let $\eta_j = \text{median}(\{\xi_i : (j-1)m+1 \le i \le jm\})$. We define $Z_j = \frac{1}{h(0)}\Phi^{-1}\left(G\left(\eta_j\right)\right)$ where G is the distribution of η_j. It follows from Theorem 6 that $\sqrt{4m}\eta_j$ is well approximated by Z_j whose distribution is $N\left(0, \frac{1}{h^2(0)}\right)$. Set

$$\epsilon_j = \sqrt{m}EX_j - \sqrt{m}f(\frac{j}{T}) = E\left\{\sqrt{m}X_j - \sqrt{m}f(\frac{j}{T}) - \sqrt{m}\eta_j\right\}.$$

This is the deterministic component of the approximation error due to binning. It is easy to see that

$$\min_{(j-1)m+1 \le i \le jm}\left[f(\frac{i}{n}) - f(\frac{j}{T})\right] \le X_j - \eta_j - f(\frac{j}{T}) \le \max_{(j-1)m+1 \le i \le jm}\left[f(\frac{i}{n}) - f(\frac{j}{T})\right]. \quad (39)$$

Since f is in a Hölder ball with smoothness $d = \min(\alpha - \frac{1}{p}, 1)$, then equation (39) implies

$$|\epsilon_j| \le \sqrt{m}E|X_j - f(\frac{j}{T}) - \eta_j| \le \sqrt{m}\max_{(j-1)m+1 \le i \le jm}|f(\frac{i}{n}) - f(\frac{j}{T})| \le C\sqrt{m}T^{-d}. \quad (40)$$

Set

$$\zeta_j = \sqrt{m}X_j - \sqrt{m}f(\frac{j}{T}) - \sqrt{m}b_m - \epsilon_j - \frac{1}{2}Z_j.$$

Then $E\zeta_j = 0$ and $\sqrt{m}X_j = \sqrt{m}f(j/T) + \epsilon_j + \frac{1}{2}Z_j + \zeta_j$.

The random error ζ_j is the sum of three terms,

$$\zeta_{1j} = \sqrt{m}X_j - \sqrt{m}f(j/T) - \sqrt{m}\eta_j - \epsilon_j + E\left(\sqrt{m}\eta_j - \frac{1}{2}Z_j\right)\{|Z_j| \le \varepsilon\sqrt{m}\}$$

$$\zeta_{2j} = \left(\sqrt{m}\eta_j - \frac{1}{2}Z_j\right)\{|Z_j| \le \varepsilon\sqrt{m}\} - E\left(\sqrt{m}\eta_j - \frac{1}{2}Z_j\right)\{|Z_j| \le \varepsilon\sqrt{m}\}$$

$$\zeta_{3j} = \left(\sqrt{m}\eta_j - \frac{1}{2}Z_j\right)\{|Z_j| > \varepsilon\sqrt{m}\}.$$

Note that

$$E\left(\sqrt{m}\eta_j - \frac{1}{2}Z_j\right)\{|Z_j| \le \varepsilon\sqrt{m}\} = E\left(\sqrt{m}\eta_j - \frac{1}{2}Z_j\right)\{|Z_j| > \varepsilon\sqrt{m}\} \le C\exp\left(-\varepsilon m/2\right).$$

From equation (39) we have $|\zeta_{1j}| \le C\sqrt{m}T^{-d} < a/2$. A bound for the approximation error ζ_{2j} is given in equation (30),

$$|\zeta_{2j}| \le \frac{C}{m}\left(1 + |Z_j|^3\right) \quad \blacksquare \tag{41}$$

We now consider the wavelet transform of the medians of the binned data. From Proposition 2 we may write

$$\frac{1}{\sqrt{T}}X_i = \frac{f(\frac{i}{T})}{\sqrt{T}} + \frac{\epsilon_i}{\sqrt{n}} + \frac{Z_i}{2\sqrt{n}} + \frac{\zeta_i}{\sqrt{n}}.$$

Let $(y_{j,k}) = T^{-\frac{1}{2}}W \cdot X$ be the discrete wavelet transform of the binned data. Then one may write

$$y_{j,k} = \theta'_{j,k} + \epsilon_{j,k} + \frac{1}{2\sqrt{n}}z_{j,k} + \xi_{j,k} \tag{42}$$

where $\theta'_{j,k}$ are the discrete wavelet transform of $(f\left(\frac{i}{T}\right))_{1\le i\le T}$, $z_{j,k}$ are the transform of the Z_i's and so are i.i.d. $N(0,1)$ and $\epsilon_{j,k}$ and $\xi_{j,k}$ are respectively the transforms of $(\frac{\epsilon_i}{\sqrt{n}})$ and $(\frac{\zeta_i}{\sqrt{n}})$. The following proposition gives the risk bounds of the block thresholding estimator in a single block. These risk bounds are similar to results for the Gaussian case given in Cai (1999). But in the current setting the error terms $\epsilon_{j,k}$ and $\xi_{j,k}$ make the problem more complicated.

Proposition 3 *Let $y_{j,k}$ be given as in (42) and let the block thresholding estimator $\hat{\theta}_{j,k}$ be defined as in (13). Then*

(i). for some constant $C > 0$

$$E \sum_{(j,k)\in B^i_j} (\hat{\theta}_{j,k} - \theta'_{j,k})^2 \le \min\{4 \sum_{(j,k)\in B^i_j} (\theta'_{j,k})^2, \ 8\lambda_* Ln^{-1}\} + 6 \sum_{(j,k)\in B^i_j} \epsilon^2_{j,k} + CLn^{-2}; \tag{43}$$

(ii). for any $0 < \tau < 1$, there exists a constant $C_\tau > 0$ depending on τ only such that for all $(j,k) \in B^i_j$

$$E(\hat{\theta}_{j,k} - \theta'_{j,k})^2 \le C_\tau \cdot \min\left\{ \max_{(j,k)\in B^i_j}\{(\theta'_{j,k} + \epsilon_{j,k})^2\}, \ Ln^{-1}\right\} + n^{-2+\tau}. \tag{44}$$

(iii) For $j \le J_*$ and $\epsilon_n > 1/\log n$, $P\left(\sqrt{n}\,|\xi_{j,k}| \ge \varepsilon_n\right) \le C\exp\left(-\varepsilon_n m\right)$.

Proof of Proposition 3: We only give the proof for (i). From Proposition 2, we have $|\epsilon_j| \le C\sqrt{m}T^{-d}$ and $\epsilon_{j,k} = \sum_i \frac{\epsilon_i}{\sqrt{n}}\int \phi_{J,i}\psi_{j,k}$. Hence

$$|\epsilon_{j,k}| \le \sup_x \sum_i |\frac{\epsilon_i}{\sqrt{n}}\phi_{J,i}(x)| \cdot \int |\psi_{j,k}(x)|dx \le CT^{-d}2^{-\frac{j}{2}}. \tag{45}$$

This, as well as Proposition 2, yields that

$$\sum_j \sum_k \epsilon_{j,k}^2 = \frac{1}{n} \sum_i \epsilon_i^2 \leq CT^{-2d}. \tag{46}$$

It follows from Lemma 1 that

$$E \sum_{(j,k) \in B_j^i} (\hat{\theta}_{j,k} - \theta'_{j,k})^2 \leq 2E \sum_{(j,k) \in B_j^i} [\hat{\theta}_{j,k} - (\theta'_{j,k} + \epsilon_{j,k})]^2 + 2 \sum_{(j,k) \in B_j^i} \epsilon_{j,k}^2$$

$$\leq 2 \min \left\{ \sum_{(j,k) \in B_j^i} (\theta'_{j,k} + \epsilon_{j,k})^2, \ 4\lambda_* L n^{-1} \right\} + 2 \sum_{(j,k) \in B_j^i} \epsilon_{j,k}^2$$

$$+ 8E \sum_{(j,k) \in B_j^i} \left(\frac{1}{2\sqrt{n}} z_{j,k} + \xi_{j,k} \right)^2 I \left(\sum_{(j,k) \in B_j^i} (\frac{1}{2\sqrt{n}} z_{j,k} + \xi_{j,k})^2 > \frac{\lambda_* L}{4n} \right)$$

$$\leq \min \left\{ 4 \sum_{(j,k) \in B_j^i} (\theta'_{j,k})^2, \ 8\lambda_* L n^{-1} \right\} + 6 \sum_{(j,k) \in B_j^i} \epsilon_{j,k}^2$$

$$+ 2n^{-1} E \sum_{(j,k) \in B_j^i} (z_{j,k} + 2\sqrt{n}\xi_{j,k})^2 I \left(\sum_{(j,k) \in B_j^i} (z_{j,k} + 2\sqrt{n}\xi_{j,k})^2 > \lambda_* L \right).$$

It can be shown that

$$D = E \sum_{(j,k) \in B_j^i} (z_{j,k} + 2\sqrt{n}\xi_{j,k})^2 I \left(\sum_{(j,k) \in B_j^i} (z_{j,k} + 2\sqrt{n}\xi_{j,k})^2 > \lambda_* L \right) \leq CLn^{-1}.$$

Note that

$$\sum_{(j,k) \in B_j^i} (z_{j,k} + 2\sqrt{n}\xi_{j,k})^2 \leq \left[\sqrt{\sum_{(j,k) \in B_j^i} z_{j,k}^2} + \sqrt{\sum_{(j,k) \in B_j^i} (2\sqrt{n}\xi_{j,k})^2} \right]^2$$

Let $\varepsilon_n = 1/\log\log n$

$$\mathbb{P} \left(\sum_{(j,k) \in B_j^i} (2\sqrt{n}\xi_{j,k})^2 \geq \varepsilon_n L \right) \leq L\mathbb{P} \left(2\sqrt{n}\xi_{j,k} \geq \sqrt{\varepsilon_n} \right) \leq C_D n^{-D} \text{ for all } D > 0$$

and

$$E \sum_{(j,k) \in B_j^i} z_{j,k}^2 \left\{ \sum_{(j,k) \in B_j^i} z_{j,k}^2 > \left(\sqrt{\lambda_*} - \sqrt{\varepsilon_n} \right)^2 L \right\} \leq CLn^{-1}$$

$$P \left\{ \sum_{(j,k) \in B_j^i} z_{j,k}^2 > \left(\sqrt{\lambda_*} - \sqrt{\varepsilon_n} \right)^2 L \right\} \leq Cn^{-1}$$

then

$$\begin{aligned} D &\leq E \sum_{(j,k) \in B_j^i} (z_{j,k} + 2\sqrt{n}\xi_{j,k})^2 \left\{ \sum_{(j,k) \in B_j^i} z_{j,k}^2 > \left(\sqrt{\lambda_*} - \sqrt{\varepsilon_n} \right)^2 L \right\} + CLn^{-1} \\ &\leq CE \left(\sum_{(j,k) \in B_j^i} z_{j,k}^2 + \varepsilon_n L \right) \left\{ \sum_{(j,k) \in B_j^i} z_{j,k}^2 > \left(\sqrt{\lambda_*} - \sqrt{\varepsilon_n} \right)^2 L \right\} + C_1 Ln^{-1} \\ &\leq C_2 Ln^{-1}. \end{aligned}$$

Then

$$E \sum_{(j,k) \in B_j^i} (\hat{\theta}_{j,k} - \theta'_{j,k})^2 \leq \min \left\{ 4 \sum_{(j,k) \in B_j^i} (\theta'_{j,k})^2, \ 8\lambda_* Ln^{-1} \right\} + 6 \sum_{(j,k) \in B_j^i} \epsilon_{j,k}^2 + CLn^{-1}$$

for some constant $C > 0$.

The third part follows from Lemma 3 in Cai and Wang (2008) which gives a concentration inequality for wavelet coefficients at a given resolution. ∎

For reasons of space we omit the proof of Proposition 3 here. See Cai and Zhou (2008) for a complete proof. We also need the following lemmas for the proof of Theorems 3 and 4. The proof of these lemmas is relatively straightforward and is thus omitted.

Lemma 1 *Suppose $y_i = \theta_i + z_i$, $i = 1, ..., L$, where θ_i are constants and z_i are random variables. Let $S^2 = \sum_{i=1}^L y_i^2$ and let $\hat{\theta}_i = (1 - \frac{\lambda L}{S^2})_+ y_i$. Then*

$$E\|\hat{\theta} - \theta\|_2^2 \leq \|\theta\|_2^2 \wedge 4\lambda L + 4E \left[\|z\|_2^2 I(\|z\|_2^2 > \lambda L) \right]. \tag{47}$$

Lemma 2 *Let $X \sim \chi_L^2$ and $\lambda > 1$. Then*

$$P(X \geq \lambda L) \leq e^{-\frac{L}{2}(\lambda - \log \lambda - 1)} \quad and \quad EXI(X \geq \lambda L) \leq \lambda L e^{-\frac{L}{2}(\lambda - \log \lambda - 1)}. \tag{48}$$

Lemma 3 *Let $T = 2^J$ and let $f_J(x) = \sum_{k=1}^{T} \frac{1}{\sqrt{T}} f(\frac{k}{T})\phi_{J,k}(x)$. Then*

$$\sup_{f \in B^\alpha_{p,q}(M)} \|f_J - f\|_2^2 \le CT^{-2d} \quad \text{where } d = \min(\alpha - 1/p, 1).$$

Let $\{\theta'_{j,k}\}$ be the discrete wavelet transform of $\{f(\frac{i}{T}), 1 \le i \le T\}$ and let $\{\theta_{j,k}\}$ be the true wavelet coefficients of f. Then $|\theta'_{j,k} - \theta_{j,k}| \le CT^{-d}2^{-j/2}$ and consequently $\sum_{j=j_0}^{J-1} \sum_k (\theta'_{j,k} - \theta_{j,k})^2 \le CT^{-2d}$.

6.3.1   Global Adaptation: Proof of Theorem 3

Decompose $E\|\hat{f}_n - f\|_2^2$ into three terms as follows,

$$
\begin{aligned}
E\|\hat{g}_n - g\|_2^2 &= \sum_k E(\hat{\tilde{\theta}}_{j_0,k} - \tilde{\theta}_{j,k})^2 + \sum_{j=j_0}^{J_*-1} \sum_k E(\hat{\theta}_{j,k} - \theta_{j,k})^2 + \sum_{j=J_*}^{\infty} \sum_k \theta_{j,k}^2 \\
&\equiv S_1 + S_2 + S_3.
\end{aligned}
\tag{49}
$$

It is easy to see that the first term S_1 and the third term S_3 are small.

$$S_1 = 2^{j_0} n^{-1} \epsilon^2 = o(n^{-2\alpha/(1+2\alpha)}). \tag{50}$$

Note that for $x \in I\!\!R^m$ and $0 < p_1 \le p_2 \le \infty$,

$$\|x\|_{p_2} \le \|x\|_{p_1} \le m^{\frac{1}{p_1} - \frac{1}{p_2}} \|x\|_{p_2}. \tag{51}$$

Since $f \in B^\alpha_{p,q}(M)$, so $2^{js}(\sum_{k=1}^{2^j} |\theta_{j,k}|^p)^{1/p} \le M$. Now (51) yields that

$$S_3 = \sum_{j=J_*}^{\infty} \sum_k \theta_{j,k}^2 \le C 2^{-2J_*(\alpha \wedge (\alpha + \frac{1}{2} - \frac{1}{p}))}. \tag{52}$$

Proposition 3, Lemma 3 and Equation (46) yield that

$$
\begin{aligned}
S_2 &\le 2 \sum_{j=j_0}^{J_*-1} \sum_k E(\hat{\theta}_{j,k} - \theta'_{j,k})^2 + 2 \sum_{j=j_0}^{J_*-1} \sum_k (\theta'_{j,k} - \theta_{j,k})^2 \\
&\le \sum_{j=j_0}^{J_*-1} \sum_{i=1}^{2^j/L} \min\left\{ 8 \sum_{(j,k)\in B^i_j} \theta_{j,k}^2, \; 8\lambda_* L n^{-1} \right\} + 6\sum_{j=j_0}^{J_*-1}\sum_k \epsilon_{j,k}^2 + Cn^{-1} + 10\sum_{j=j_0}^{J_*-1}\sum_k (\theta'_{j,k} - \theta_{j,k})^2 \\
&\le \sum_{j=j_0}^{J_*-1} \sum_{i=1}^{2^j/L} \min\left\{ 8 \sum_{(j,k)\in B^i_j} \theta_{j,k}^2, \; 8\lambda_* L n^{-1} \right\} + Cn^{-1} + CT^{-2d}.
\end{aligned}
\tag{53}
$$

We now divide into two cases. First consider the case $p \geq 2$. Let $J_1 = [\frac{1}{1+2\alpha} \log_2 n]$. So, $2^{J_1} \approx n^{1/(1+2\alpha)}$. Then (53) and (51) yield

$$S_2 \leq 8\lambda_* \sum_{j=j_0}^{J_1-1} \sum_{i=1}^{2^j/L} Ln^{-1} + 8 \sum_{j=J_1}^{J_*-1} \sum_k \theta_{j,k}^2 + Cn^{-1} + CT^{-2d} \leq Cn^{-2\alpha/(1+2\alpha)}.$$

By combining this with (50) and (52), we have $E\|\hat{f}_n - f\|_2^2 \leq Cn^{-2\alpha/(1+2\alpha)}$ for $p \geq 2$. Now let us consider the case $p < 2$. First we state the following lemma without proof.

Lemma 4 *Let $0 < p < 1$ and $S = \{x \in I\!\!R^k : \sum_{i=1}^k x_i^p \leq B, \ x_i \geq 0, \ i = 1, \cdots, k\}$. Then for $A > 0$, $\sup_{x \in S} \sum_{i=1}^k (x_i \wedge A) \leq B \cdot A^{1-p}$.*

Let J_2 be an integer satisfying $2^{J_2} \asymp n^{1/(1+2\alpha)} (\log n)^{(2-p)/p(1+2\alpha)}$. Note that

$$\sum_{i=1}^{2^j/L} \left(\sum_{(j,k) \in B_j^i} \theta_{j,k}^2 \right)^{\frac{p}{2}} \leq \sum_{k=1}^{2^j} (\theta_{j,k}^2)^{\frac{p}{2}} \leq M2^{-jsp}.$$

It then follows from Lemma 4 that

$$\sum_{j=J_2}^{J_*-1} \sum_{i=1}^{2^j/L} \min \left\{ 8 \sum_{(j,k) \in B_j^i} \theta_{j,k}^2, \ 8\lambda_* Ln^{-1} \right\} \leq Cn^{-\frac{2\alpha}{1+2\alpha}} (\log n)^{\frac{2-p}{p(1+2\alpha)}}. \tag{54}$$

On the other hand,

$$\sum_{j=j_0}^{J_2-1} \sum_{i=1}^{2^j/L} \min\{8 \sum_{(j,k) \in B_j^i} \theta_{j,k}^2, \ 8\lambda_* Ln^{-1}\} \leq \sum_{j=j_0}^{J_2-1} \sum_b 8\lambda_* Ln^{-1} \leq Cn^{-\frac{2\alpha}{1+2\alpha}} (\log n)^{\frac{2-p}{p(1+2\alpha)}}. \tag{55}$$

We finish the proof for the case $p < 2$ by putting (50), (52), (54) and (55) together,

$$E\|\hat{f}_n - f\|_2^2 \leq Cn^{-\frac{2\alpha}{1+2\alpha}} (\log n)^{\frac{2-p}{p(1+2\alpha)}}. \quad \blacksquare$$

6.3.2 Local Adaptation: Proof of Theorem 4

For simplicity, we give the proof for Hölder classes $\Lambda^\alpha(M)$ instead of local Hölder classes $\Lambda^\alpha(M, t_0, \delta)$. Note that for all $f \in \Lambda^\alpha(M)$, $|\theta_{j,k}| = |\langle f, \psi_{j,k} \rangle| \leq C2^{-j(1/2+\alpha)}$ for some constant $C > 0$ not depending on f. Note also that for any random variables X_i, $i =$

$1, \cdots, n$, $E(\sum_{i=1}^{n} X_i)^2 \le (\sum_{i=1}^{n} (EX_i^2)^{1/2})^2$. It then follows that

$$
\begin{aligned}
E(\hat{f}_n(t_0) - f(t_0))^2 &= E\left[\sum_{k=1}^{2^{j_0}} (\hat{\tilde{\theta}}_{j_0,k} - \tilde{\theta}_{j_0,k})\phi_{j_0,k}(t_0) + \sum_{j=j_0}^{\infty} \sum_{k=1}^{2^j} (\hat{\theta}_{j,k} - \theta_{j,k})\psi_{j,k}(t_0)\right]^2 \\
&\le \left[+ \sum_{k=1}^{2^{j_0}} (E(\hat{\tilde{\theta}}_{j_0,k} - \tilde{\theta}_{j_0,k})^2 \phi_{j_0,k}^2(t_0))^{\frac{1}{2}} \right. \\
&\qquad\left. + \sum_{j=j_0}^{J_*-1} \sum_{k=1}^{2^j} (E(\hat{\theta}_{j,k} - \theta_{j,k})^2 \psi_{j,k}^2(t_0))^{\frac{1}{2}} + \sum_{j=J}^{\infty} \sum_{k=1}^{2^j} |\theta_{j,k}\psi_{j,k}(t_0)|\right]^2 \\
&\equiv (Q_1 + Q_2 + Q_3)^2.
\end{aligned}
$$

Since the wavelet ψ is compactly supported, so there are at most N basis functions $\psi_{j,k}$ at each resolution level j that are nonvanishing at t_0, where N is the length of the support of ψ. Denote $K(t_0, j) = \{k : \psi_{j,k}(t_0) \ne 0\}$. Then $|K(t_0, j)| \le N$. It is easy to see that both Q_{12} and Q_3 are small:

$$
Q_1 = \sum_{k=1}^{2^{j_0}} (E(\hat{\tilde{\theta}}_{j_0,k} - \tilde{\theta}_{j_0,k})^2)^{\frac{1}{2}} |\phi_{j_0,k}(t_0)| = O(n^{-1}), \tag{56}
$$

$$
Q_3 = \sum_{j=J}^{\infty} \sum_{k=1}^{2^j} |\theta_{j,k}||\psi_{j,k}(t_0)| \le \sum_{j=J}^{\infty} N\|\psi\|_{\infty} 2^{\frac{j}{2}} C 2^{-j(\frac{1}{2}+\alpha)} \le CT^{-\alpha}. \tag{57}
$$

We now consider the third term Q_2. Applying the bound (44) in Proposition 3 with $\tau < 1/(1+2\alpha)$ together with Lemma 3 and the bound for $\epsilon_{j,k}$ given in (45), we have

$$
\begin{aligned}
Q_2 &\le \sum_{j=j_0}^{J_*-1} \sum_{k \in K(t_0, j)} 2^{\frac{j}{2}} \|\psi\|_{\infty} (E(\hat{\theta}_{j,k} - \theta_{j,k})^2)^{\frac{1}{2}} \\
&\le C \sum_{j=j_0}^{J_*-1} 2^{\frac{j}{2}} [\min(2^{-j(1+2\alpha)} + T^{-2(\alpha\wedge1)}2^{-j}, Ln^{-1}) + n^{-2+\tau}]^{\frac{1}{2}} \\
&\le C(\frac{\log n}{n})^{\frac{\alpha}{1+2\alpha}}. \tag{58}
\end{aligned}
$$

Combining Equations (56)-(58), we have $E(\hat{f}_n(t_0) - f(t_0))^2 \le C(\log n/n)^{2\alpha/(1+2\alpha)}$. ∎

6.4 Proof of Theorem 5

Recall that

$$
\hat{Q} = \sum_{k=1}^{2^{j_0}} (\tilde{y}_{j_0,k}^2 - \hat{\sigma}_n^2) + \sum_{j=j_0}^{J_q} \sum_{k=1}^{2^j} (y_{j,k}^2 - \hat{\sigma}_n^2)
$$

and note that the empirical wavelet coefficients can be written as

$$y_{j,k} = \theta_{j,k} + \epsilon_{j,k} + \sigma_n z_{j,k} + \xi_{j,k}.$$

Since $\left(\sum_{j>J_1} \theta_{j,k}^2\right)^2 \le C \left[2^{-2J_1(\alpha-1/p)}\right]^2 = o\left(\frac{1}{n}\right)$, as in Cai and Low (2005) it is easy to show that

$$E_f \left\{\sum_{k=1}^{2^{j_0}} \left[\left(\tilde{\theta}_{j_0,k} + \sigma_n \tilde{z}_{j_0,k}\right)^2 - \sigma_n^2\right] + \sum_{j=j_0}^{J_q} \sum_{k=1}^{2^j} \left(\left[(\theta_{j,k} + \sigma_n z_{j,k})^2 - \sigma_n^2\right] - Q(f)\right)\right\}^2 \le 4\sigma_n^2 M^2 (1 + o(1)).$$

The theorem then follows easily from the facts below

$$\sum_{k=1}^{2^{j_0}} \tilde{\epsilon}_{j,k}^2 + \sum_{j=j_0}^{J_q} \sum_{k=1}^{2^j} \epsilon_{j,k}^2 \le CT^{-2(\alpha-1/p)} = o\left(\frac{1}{n}\right)$$

$$E\left(\sum_{k=1}^{2^{j_0}} \xi_{j,k}^2 + \sum_{j=j_0}^{J_q} \sum_{k=1}^{2^j} \xi_{j,k}^2\right)^2 \le C\frac{1}{m^2 n} = o\left(\frac{1}{n}\right)$$

$$E\left[\sqrt{n}\left(\hat{\sigma}_n^2 - \sigma_n^2\right)\right]^2 = o\left(\frac{1}{n}\right). \quad \blacksquare$$

References

[1] Averkamp, R. and Houdré, C. (2003). Wavelet thresholding for non-necessarily Gaussian noise: Idealism. *Ann. Statist.* **31**, 110-151.

[2] Averkamp, R. and Houdré, C. (2005). Wavelet thresholding for non-necessarily Gaussian noise: Functionality. *Ann. Statist.* **33**, 2164-2193.

[3] Bickel, P. J. and Ritov, Y. (1988). Estimating integrated squared density derivatives: Sharp best order of convergence estimates. *Sankhya Ser. A* **50,** 381–393.

[4] Brown, L. D., Cai, T. T., Low, M. G. & Zhang, C. (2002). Asymptotic equivalence theory for nonparametric regression with random design. *Ann. Statist.* **30**, 688-707.

[5] Brown, L. D., Cai, T. T., Zhang, R., Zhao, L. H. and Zhou, H. H. (2006). The Root-unroot algorithm for density estimation as implemented via wavelet block thresholding. Manuscript.

[6] Brown, L. D., Cai, T. T. & Zhou, H. H. (2008). Robust nonparametric estimation via wavelet median regression. *Ann. Statist.*, to appear.

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

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

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

[10] Brown, L. D., Wang, Y. and Zhao, L. (2003) Statistical equivalence at suitable frequencies of GARCH and stochastic volatility models with the corresponding diffusion model. *Statistica Sinica* **13**, 993-1013.

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

[12] Cai, T. and Low, M. (2005). Non-quadratic estimators of a quadratic functional. *Ann. Statist.* **33**, 2930-2956.

[13] Cai, T. and Low, M. (2006a). Adaptive confidence balls. *Ann. Statist.* **34**, 202-228.

[14] Cai, T. and Low, M. (2006b). Optimal adaptive estimation of a quadratic functional. *Ann. Statist.* **34**, 2298-2325.

[15] Cai, T. and Wang, L. (2008). Adaptive variance function estimation in heteroscedastic nonparametric regression. *Ann. Statist.*, to appear.

[16] Cai, T. T. and Zhou, H. H. (2008). A data-driven block thresholding approach to wavelet estimation. *Ann. Statist.*, to appear.

[17] Daubechies, I. (1992). *Ten Lectures on Wavelets*. SIAM, Philadelphia.

[18] Delattre, S. and Hoffmann, M. (2002). Asymptotic equivalence for a null recurrent diffusion. *Bernoulli* **8**, 139-174.

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

[20] Donoho, D. L. and Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.* **90**, 1200-24.

[21] Donoho, D.L. and Johnstone, I.M. (1998). Minimax estimation via wavelet shrinkage. *Ann. Statist.* **26**, 879-921.

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

[23] Donoho, D. L. and Nussbaum, M. (1990). Minimax quadratic estimation of a quadratic functional. *J. Complexity* **6,** 290–323.

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

[25] Efromovich, S. Y. and Low, M. (1996). On optimal adaptive estimation of a quadratic functional. *Ann. Statist.* **24,** 1106–1125.

[26] Fan, J. (1991). On the estimation of quadratic functionals. *Ann. Statist.* **19,** 1273–1294.

[27] Genon-Catalot, V., Laredo, C. and Nussbaum, M. (2002). Asymptotic equivalence of estimating a Poisson intensity and a positive diffusion drift. *Ann. Statist.* **30**, 731-753.

[28] Genovese, C. R. and Wasserman, L. (2005). Confidence sets for nonparametric wavelet regression. *Ann. Statist.* **33,** 698–729.

[29] Grama, I. and Nussbaum, M. (1998). Asymptotic equivalence for nonparametric generalized linear models. *Probab. Theory Relat. Fields* **111**, 167-214.

[30] Grama, I. and Nussbaum, M. (2002). Asymptotic equivalence for nonparametric regression. *Math. Methods Statist.* **11** (1) 1-36

[31] Golubev, G. K., Nussbaum, M. and Zhou, H. H. (2005). Asymptotic equivalence of spectral density estimation and Gaussian white noise. Submitted, available at http://www.stat.yale.edu/~hz68.

[32] Johnstone, I. M. (2002). *Function Estimation and Gaussian Sequence Models.* Manuscript.

[33] Johnstone, I. M. and Silverman, B. W. (2005). Empirical Bayes selection of wavelet thresholds. *Ann. Statist.* **33**, 1700-1752.

[34] Klemelä, J. (2006). Sharp adaptive estimation of quadratic functionals. *Probab. Theory Related Fields* **134,** 539–564.

[35] Komlós, J., Major, P. and Tusnády, G. (1975). An approximation of partial sums of independent rv's and the sample df. I Z. *Wahrsch. verw. Gebiete* **32**, 111-131.

[36] Kovac, A. and Silverman, B.W. (2000). Extending the Scope of Wavelet Regression Methods by Coefficient-dependent Thresholding. *J. Amer. Statist. Assoc.* **95**, 172-183.

[37] Korostelev, A. P. (1993). Exact asymptotically minimax estimator for nonparametric regression in uniform norm. *Theory Probab. Appl.* **38**, no. 4, 775-782.

[38] Laurent, B. and Massart, P. (2000). Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.* **28,** 1302–1338.

[39] Le Cam, L. (1964). Sufficiency and approximate sufficiency. *Ann. Math. Statist.* **35** 1419–1455.

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

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

[42] Li, K.-C. (1989). Honest confidence regions for nonparametric regression. *Ann. Statist.* **17,** 1001–1008.

[43] Massart, P. (2002). Tusnády's lemma, 24 years later. *Ann. Inst. H. Poincaré Probab. Statist.* **38**, 991-1007.

[44] Meyer, Y. (1992). *Wavelets and Operators.* Cambridge University Press, Cambridge.

[45] Milstein, G. and Nussbaum, M.(1998). Diffusion approximation for nonparametric autoregression. *Prob. Theor. Rel. Fields* **112**, 535-543.

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

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

[48] Pollard, D. P. (2001). *A User's Guide to Measure Theoretic Probability.* Cambridge University Press.

[49] Spokoiny, V. (1996) Adaptive hypothesis testing using wavelets. *Ann. Statist.* **24**, 2477-2498.

[50] Spokoiny, V.G. (1998). Adaptive and spatially adaptive testing of nonparametric hypothesis. *Math. Methods Statist.*, **7**, 245-273.

[51] Strang, G. (1992). Wavelet and dilation equations: a brief introduction. *SIAM Review* **31**, 614-627.

[52] Tsybakov, A. B. (2004) *Introduction a l'estimation non-paramétrique.* Springer.

[53] Triebel, H. (1992). *Theory of Function Spaces II.* Birkhäuser Verlag, Basel.

[54] Wang, Y.Z. (2002). Asymptotic nonequivalence of GARCH models and diffusions. *Ann. Statist.* **30**, 754-783.

[55] Zhang, C.-H. (2005). General empirical Bayes wavelet methods and exactly adaptive minimax estimation. *Ann. Statist.* **33**, 54-100.

[56] Zhou, H. H. (2006). A note on quantile coupling inequalities and their applications. Submitted. Available from www.stat.yale.edu/~hz68 .