IMS Collections
Borrowing Strength: Theory Powering Applications –
A Festschrift for Lawrence D. Brown
Vol. 6 (2010) 199–215

DOI: 10.1214/10-IMSCOLL614

Nonparametric regression in natural exponential families

T. Toni Cai[1,*] **and Harrison H. Zhou**[2,†]

University of Pennsylvania and Yale University

Abstract: Theory and methodology for nonparametric regression have been particularly well developed in the case of additive homoscedastic Gaussian noise. Inspired by asymptotic equivalence theory, there have been ongoing efforts in recent years to construct explicit procedures that turn other function estimation problems into a standard nonparametric regression with Gaussian noise. Then in principle any good Gaussian nonparametric regression method can be used to solve those more complicated nonparametric models. In particular, Brown, Cai and Zhou [3] considered nonparametric regression in natural exponential families with a quadratic variance function.

In this paper we extend the scope of Brown, Cai and Zhou [3] to general natural exponential families by introducing a new explicit procedure that is based on the variance stabilizing transformation. The new approach significantly reduces the bias of the inverse transformation and as a consequence it enables the method to be applicable to a wider class of exponential families. Combining this procedure with a wavelet block thresholding estimator for Gaussian nonparametric regression, we show that the resulting estimator enjoys a high degree of adaptivity and spatial adaptivity with near-optimal asymptotic performance over a broad range of Besov spaces.

1. Introduction

The theory of asymptotic equivalence occupies an important position in statistical decision theory. The main goal is to approximate complex statistical models by simpler ones. If two models are asymptotically equivalent, then all asymptotically optimal procedures for the simpler model can be carried over to the complex one under all bounded losses. Asymptotic equivalence theory was pioneered by Lucien Le Cam and the early focus was on parametric models. See [16]. The first global asymptotic equivalence result for nonparametric function estimation models was developed in the seminal paper by Brown and Low [4] in the context of nonparametric regression and white noise with drift models. Since then there has been active research on asymptotic equivalence/nonequivalence among nonparametric function estimation models. Many important results have been developed in different contexts.

The main ideas behind asymptotic equivalence theory are very appealing, but the theory does have drawbacks. One is that full equivalence in Le Cam's sense is a very

[1]Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104

[2]Department of Statistics, Yale University, New Haven, CT 06511

*Supported in part by NSF FRG Grant DMS-0854973.

†Supported in part by NSF Career Award DMS-0645676 and NSF FRG Grant DMS-0854975.

Keywords and phrases: adaptivity, asymptotic equivalence, exponential family, James-Stein estimator, Gaussian nonparametric regression, quantile coupling, wavelets.

AMS 2000 subject classifications: Primary 62J08; secondary 62G20.

stringent goal and often the failures are caused by pathological cases which do not occur in many applications of interest. Another is that the equivalence mappings typically require randomizations and are thus not practical.

Inspired by the ideas from the asymptotic equivalence theory, there have been recent efforts to construct explicit and practical procedures to turn more complicated nonparametric function estimation problems into a standard nonparametric regression with homoscedastic Gaussian noise, which is relatively simple and has been particularly well studied in the literature. For example, explicit procedures based on binning and taking the median have been developed in [2] and [10] for nonparametric regression with general additive noise. Brown *et al.* [6] introduced a root-unroot transformation for density estimation. Brown, Cai and Zhou [3] considered nonparametric regression in natural exponential families with a quadratic variance function, which includes, for example, nonparametric Poisson regression, binomial regression, and Gamma regression as special cases.

A key tool in [3] is a mean-matching variance stabilizing transformation (VST) for natural exponential families. However, such a VST exists only for families with a quadratic variance function. The advantage of the mean-matching VST over the classical VST is that it reduces the bias due to the transformation up to a certain level while still stabilizing the variance. The bias reduction is a crucial property. Other methods for nonparametric regression in exponential families have been proposed and studied in the literature. The reader are referred to [3] for references and discussions.

In this paper we further extend the idea of Gaussianization given in [3] to cover nonparametric regression in general natural exponential families where the mean-matching VST may not exist. A new procedure is introduced to eliminate the transformation bias completely for every natural exponential family. The procedure has four steps: Binning, VST, Gaussian regression, and inverse VST. The main differences between the procedure proposed in the present paper and that in [3] are in the choices of the VST and the inverse VST as well as the selection of the bin size. Complete elimination of the transformation bias enables one to use much smaller bin size than that required in [3]. As a consequence the procedure can still perform well when the regression function is less smooth.

Our procedure begins by grouping the data into bins with size of order $(\log n)^{1+\nu}$ for some $\nu > 0$, where n is the sample size, and then a VST is applied to the binned data. In principle any good Gaussian regression procedure can then be applied to the transformed data. The final estimator of the regression function in the original problem is constructed by the inverse VST of the estimator obtained in the Gaussian regression problem. To illustrate our general methodology, we use a wavelet block thresholding procedure for Gaussian nonparametric regression in this paper. Wavelet thresholding methods have achieved considerable success in terms of spatial adaptivity and asymptotic optimality in such a setting. In particular, block thresholding rules have been shown to possess impressive properties. In the context of Gaussian nonparametric regression local block thresholding has been studied, for example, in [13, 7, 8, 9]. For concreteness, we shall use the BlockJS procedure proposed in [7] in the present paper.

Theoretical properties of our estimators are investigated. It is shown that the estimators enjoy excellent asymptotic adaptivity and spatial adaptivity. The procedure using BlockJS simultaneously attains the optimal rate of convergence under mean integrated squared error over a broader range of Besov classes than those in [3]. This is mainly due to the fact that a much smaller bin size is used in our procedure. The estimator also automatically adapts to the local smoothness of the

underlying function; it attains the local adaptive minimax rate for estimating functions at a point.

The paper is organized as follows. In Section 2, after the classical variance stabilizing transformation for natural exponential families is introduced, we present the procedure of using the VST to convert nonparametric regression in exponential families into a Gaussian nonparametric regression problem. Section 3 discusses in detail a particular estimation procedure based on the VST and wavelet block thresholding. Theoretical properties of the procedures are treated in Section 4. Technical proofs are given in Section 6.

2. Nonparametric regression in exponential families

For nonparametric regression in natural exponential families, the noise is not additive and non-Gaussian. Applying standard nonparametric regression methods directly to the data in general do not yield desirable results. Our strategy is to use a variance stabilizing transformation (VST) to turn this problem to a standard Gaussian regression problem. We begin by discussing the VST and then introduce our procedure for nonparametric regression in natural exponential families.

The VST for natural exponential families has been widely used in many contexts. See, for example, [14] for an extensive review. Note that the probability density/mass function of a distribution in a natural one-parameter exponential families can be written as

$$q(x|\eta) = e^{\eta x - \psi(\eta)} h(x),$$

where η is the natural parameter. The mean and variance are respectively

$$\mu(\eta) = \psi'(\eta), \text{ and } \sigma^2(\eta) = \psi''(\eta).$$

We shall denote the distribution by $NEF(\mu)$. Let $X_1, \ldots, X_m \overset{iid}{\sim} NEF(\mu)$ be a random sample and set $X = \sum_{i=1}^m X_i$. The Central Limit Theorem yields that

$$\sqrt{m}(X/m - \mu(\eta)) \xrightarrow{L} N(0, V(\mu(\eta))), \quad \text{as } m \to \infty.$$

A variance stabilizing transformation (VST) is a function $G : \mathbb{R} \to \mathbb{R}$ such that

$$G'(\mu) = V^{-\frac{1}{2}}(\mu). \tag{1}$$

The standard delta method then yields

$$\sqrt{m}\{G(X/m) - G(\mu(\eta))\} \xrightarrow{L} N(0, 1).$$

Since the natural exponential can be mean parameterized, we define

$$H_m(\mu) = \mathbb{E}G(X/m), \tag{2}$$

where H_m depends on m. For notational simplicity, we shall drop the subscript m hereafter.

Now consider nonparametric regression in natural exponential families. Suppose we observe

$$Y_i \overset{ind.}{\sim} NEF(f(t_i)), \quad i = 1, \ldots, n, \; t_i = \frac{i}{n} \tag{3}$$

and wish to estimate the mean function $f(t)$. As mentioned earlier, applying standard nonparametric regression methods directly to the data $\{Y_i\}$ in general do not yield desirable results. We shall turn this problem to a standard Gaussian regression problem based on a sample $\{\tilde{Y}_j : \ j = 1, \ldots, T\}$ where

$$\tilde{Y}_j \sim N\left(H\left(f\left(t_j\right)\right), m^{-1}\right), \quad t_j = j/T, \ j = 1, 2, \ldots, T.$$

Here H is defined as in (2), T is the number of bins, and m is the number of observations in each bin. Later we will discuss the specific choice of T and m in Section 4.

We begin by dividing the interval into T equi-length subintervals with $m = n/T$ observations in each subintervals. Let Q_j be the sum of observations on the j-th subinterval $I_j = (\frac{j-1}{T}, \frac{j}{T}]$, $j = 1, 2, \ldots, T$,

$$Q_j = \sum_{i=(j-1)m+1}^{jm} Y_i. \tag{4}$$

The sums $\{Q_j\}$ can be treated as observations for a Gaussian regression directly, but this in general leads to a heteroscedastic problem. Instead, we apply the VST and then treat $G(Q_j/m)$ as new observations in a homoscedastic Gaussian regression problem. To be more specific, let

$$Y_j^* = G\left(\frac{Q_j}{m}\right), \quad j = 1, \ldots, T. \tag{5}$$

The transformed data $Y^* = (Y_1^*, \ldots, Y_T^*)$ is then treated as the new equi-spaced sample for a Gaussian nonparametric regression problem.

We will first estimate $H(f(t_i))$, then take an inverse transformation H^{-1} of the estimator to estimate the mean function f. After the original regression problem is turned into a Gaussian regression problem through binning and the VST, in principle any good Gaussian nonparametric regression method can be applied to the transformed data $\{Y_j^*\}$ to construct an estimate of $H(f(\cdot))$. The general ideas for our approach can be summarized as follows.

1. **Binning:** Divide $\{Y_i\}$ into T equal length intervals between 0 and 1. Let $Q_1, Q_2, \ldots, Q_T$ be the sum of the observations in each of the intervals. Later results suggest a choice of T satisfying $T \asymp n/\log^{1+\nu} n$. See Section 4 for details.
2. **VST:** Let $Y_j^* = G(\frac{Q_j}{m})$, $j = 1, \ldots, T$, and treat $Y^* = (Y_1^*, Y_2^*, \ldots, Y_T^*)$ as the new equi-spaced sample for a Gaussian nonparametric regression problem.
3. **Gaussian regression:** Apply your favorite nonparametric regression procedure to the binned and transformed data Y^* to obtain an estimate $\widehat{H(f)}$ of $H(f)$.
4. **Inverse VST:** Estimate the mean function f by $\widehat{f} = H^{-1}(\widehat{H(f)})$. If $\widehat{H(f)}$ is not in the domain of H^{-1} which is an interval between a and b (a and b can be ∞), we set $H^{-1}(\widehat{H(f)}) = H^{-1}(a)$ if $\widehat{H(f)} < a$ and set $H^{-1}(\widehat{H(f)}) = H^{-1}(b)$ if $\widehat{H(f)} > b$. For example, $H^{-1}(a) = 0$ when $a < 0$ in the case of Negative Binomial and NEF-GHS distributions. Note that this step is different from the "Inverse VST" in [3].

2.1. Effects of binning and VST

As mentioned earlier, after binning and the VST, the transformed data $\{Y_j^*\}$ can be treated as if they were data from a homoscedastic Gaussian nonparametric regression problem. A key step in understanding why this procedure works is to understand the effects of binning and the VST. Quantile coupling provides an important technical tool to shed insights on the procedure.

The following result, which is a direct consequence of the quantile coupling inequality developed by Komlós, Major and Tusnády [15], shows that the binned and transformed data can be well approximated by independent normal variables. See also [24].

Lemma 1. *Let $X_i \overset{iid}{\sim} NEF(\mu)$ with variance V for $i = 1, \ldots, m$ and let $X = \sum_{i=1}^m X_i$. There exists a standard normal random variable $Z \sim N(0,1)$ and positive constants c_i, $i = 1, 2, 3, 4, 5$, not depending on m such that whenever the event $A = \{|X - m\mu| \le c_1 m\}$ occurs,*

$$|X - m\mu - \sqrt{mV}Z| < c_2 Z^2 + c_3 \tag{6}$$

and

$$\mathbb{P}\left(|X - m\mu - \sqrt{mV}Z| > a\right) \le c_4 \exp\left(-c_5 a\right)$$

uniformly over μ in a compact set in the interior of the natural parameter space.

Hence, for large m, X can be treated as a normal random variable with mean $m\mu$ and variance mV. Let $Y = G(X/m)$, and Z be a standard normal variable satisfying (6). Then Y can be written as

$$Y = H(\mu) + m^{-\frac{1}{2}} Z + \xi, \tag{7}$$

where

$$\xi = G\left(\frac{X}{m}\right) - H(\mu) - m^{-\frac{1}{2}} Z \tag{8}$$

is a zero mean and "stochastically small" random variable. The following result is proved in Section 6.1.

Lemma 2. *Let $X_i \overset{iid}{\sim} NEF(\mu)$ with variance V for $i = 1, \ldots, m$, and $X = \sum_{i=1}^m X_i$. Let Z be the standard normal variable given as in Lemma 1 and let ξ be given as in (8). Then for any integer $k \ge 1$ there exist positive constants $c_k > 0$ such that for all $a > 0$,*

$$\mathbb{P}(m|\xi| > a) \le c_1 \exp\left(-c_2 a\right) + c_3 \exp\left(-c_4 m\right). \tag{9}$$

The discussion so far has focused on the effects of the VST for i.i.d. observations. In the nonparametric function estimation problem mentioned earlier, observations in each bin are independent but not identically distributed since the mean function f is not a constant in general. However, through coupling, observations in each bin can in fact be treated as if they were i.i.d. random variables when the function f is smooth. Let $X_i \sim NEF(\mu_i)$, $i = 1, \ldots, m$, be independent. Here the means μ_i are "close" but not equal. Let μ be a value close to the μ_i's. The analysis given in Section 6.1 shows that X_i in each bin can in fact be coupled with i.i.d. random

variables $X_{i,c}$ with $X_{i,c} \overset{iid}{\sim} NEF(\mu_c^*)$, for some $\mu_c^* > 0$. See Lemma 4 in Section 6.1 for a precise statement.

Lemmas 1, 2 and 4 together yield the following result which shows how far away are the transformed data $\{Y_j^*\}$ from the ideal Gaussian model.

Theorem 1. *Let $Y_j^* = G(\frac{Q_j}{m})$ be given as in (5). Assume that f is continuous, and for all $x \in [0,1]$, $f(x) \in [\varepsilon, v]$, a compact set in the interior of the mean parameter space of the natural exponential family. Then Y_j^* can be written as*

$$Y_j^* = H\left(f\left(\frac{j_*}{T}\right)\right) + m^{-\frac{1}{2}} Z_j + \xi_j, \quad j = 1, 2, \ldots, T, \tag{10}$$

where $jm + 1 \le j_* \le (j+1)m$, $Z_j \overset{i.i.d.}{\sim} N(0,1)$, and ξ_j are independent and "stochastically small" random variables satisfying that for any integer $k > 0$ and any constant $a > 0$

$$\mathbb{P}(m|\xi_j| > a) \le c_1 \exp(-c_2 a) + c_3 \exp(-c_4 m), \tag{11}$$

where $c_k > 0$.

Theorem 1 provides explicit bounds for both the deterministic and stochastic errors. This is an important technical result which serves as a major tool for the proof of the main results given in Section 4.

3. A wavelet procedure for generalized regression

One can apply any good Gaussian nonparametric regression procedure to the transformed data $\{Y_j^*\}$ to construct an estimator of the function f. To illustrate our general methodology, in the present paper we shall use wavelet block thresholding to construct the final estimators of the regression function. Before we can give a detailed description of our procedure, we need a brief review of basic notation and definitions.

Let $\{\phi, \psi\}$ be a pair of father and mother wavelets. The functions ϕ and ψ are assumed to be compactly supported and $\int \phi = 1$, and dilation and translation of ϕ and ψ generates an orthonormal wavelet basis. For simplicity in exposition, in the present paper 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 $\{\phi_{j_0,k}^p, k = 1, \ldots, 2^{j_0};\ \psi_{j,k}^p,\ j \ge j_0 \ge 0, k = 1, \ldots, 2^j\}$ 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 [11] and [22]. A square-integrable function f on $[0,1]$ can be expanded into a wavelet series:

$$f(t) = \sum_{k=1}^{2^{j_0}} \tilde{\theta}_{j_0,k} \phi_{j_0,k}(t) + \sum_{j=j_0}^{\infty} \sum_{k=1}^{2^j} \theta_{j,k} \psi_{j,k}(t), \tag{12}$$

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 give a detailed description of the wavelet thresholding procedure BlockJS in this section and study the properties of the resulting estimator in Section 4. We shall show that our estimator enjoys a high degree of adaptivity and spatial adaptivity and are easily implementable.

Apply the discrete wavelet transform to the binned and transformed data Y^* given in (5), and let $U = T^{-\frac{1}{2}}WY^*$ be the empirical wavelet coefficients, where W is the discrete wavelet transformation matrix. Write

$$U = (\widetilde{y}_{j_0,1}, \ldots, \widetilde{y}_{j_0,2^{j_0}}, y_{j_0,1}, \ldots, y_{j_0,2^{j_0}}, \ldots, y_{J-1,1}, \ldots, y_{J-1,2^{J-1}})'. \tag{13}$$

Here $\widetilde{y}_{j_0,k}$ are the gross structure terms at the lowest resolution level, and $y_{j,k}$ ($j = j_0, \ldots, J-1, k = 1, \ldots, 2^j$) are empirical wavelet coefficients at level j which represent fine structure at scale 2^j. The empirical wavelet coefficients can then be written as

$$y_{j,k} = \theta_{j,k} + \frac{1}{\sqrt{n}} z_{j,k} + \xi_{j,k}, \tag{14}$$

where $\theta_{j,k}$ are the true wavelet coefficients of $H(f)$, and $z_{j,k}$ are i.i.d. $N(0,1)$, and $\xi_{j,k}$ are some "small" stochastic errors. The theoretical calculations given in Section 6 will show that $\xi_{j,k}$ is negligible. If the error $\xi_{j,k}$ is ignored then we have

$$y_{j,k} \approx \theta_{j,k} + \frac{1}{\sqrt{n}} z_{j,k}, \tag{15}$$

which is the idealized Gaussian sequence model with noise level $\sigma = 1/\sqrt{n}$. The BlockJS [7] was originally developed for this ideal model. Here we shall apply block thresholding to the empirical coefficients $y_{j,k}$ as if they were observed as in (15).

At each resolution level j, the empirical wavelet coefficients $y_{j,k}$ are grouped into nonoverlapping blocks of length L. As in the sequence estimation setting let $B_j^i = \{(j,k) : (i-1)L+1 \le k \le iL\}$ and let $S_{j,i}^2 \equiv \sum_{(j,k)\in B_j^i} y_{j,k}^2$. Set $J_* = \lfloor \log_2 \frac{T}{\log^{1+\gamma} n} \rfloor$ with some $\gamma > 0$. At each resolution level $j \le J_*$, a modified James-Stein shrinkage rule is then applied to each block B_j^i, i.e.,

$$\hat{\theta}_{j,k} = \left(1 - \frac{\lambda_* L}{n S_{j,i}^2}\right)_+ y_{j,k} \quad \text{for } (j,k) \in B_j^i, \tag{16}$$

where $\lambda_* = 4.50524$ is the solution to the equation $\lambda_* - \log \lambda_* = 3$ (see [7], for details), and $\frac{1}{n}$ is approximately the variance of each $y_{j,k}$. 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 $H(f(\cdot))$ at the equally spaced sample points $\{\frac{i}{T} : i = 1, \ldots, T\}$ is then obtained by applying the inverse discrete wavelet transform (IDWT) to the denoised wavelet coefficients. That is, $\{H(f(\frac{i}{T})) : i = 1, \ldots, T\}$ is estimated by $\widehat{H(f)} = \{\widehat{H(f(\frac{i}{T}))} : i = 1, \ldots, T\}$ with $\widehat{H(f)} = T^{\frac{1}{2}} W^{-1} \cdot \hat{\theta}$. The estimate of the whole function $H(f)$ is given by

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

Once the estimator $\widehat{H(f)}$ is obtained, the mean function f is estimated by applying the inverse transformation H^{-1},

$$\widehat{f}_{BJS}(t) = H^{-1}(\widehat{H(f(t))}). \tag{17}$$

Remark. The value $J_* = \lfloor \log_2 \frac{T}{\log^{1+\gamma} n} \rfloor$ is chosen for a technical reason. In Section 6.2, it is shown that a good tail probability bound for $\xi_{j,k}$ (given in equation (31)) holds for all $j \le J_*$.

4. Asymptotic optimalities

In this section we investigate the theoretical properties of the procedures proposed in Section 2. The asymptotic performance of our procedures is considered over the Besov spaces. This is by now the standard analysis for wavelet regression methods. Besov spaces are a very rich class of function spaces and contain as special cases many traditional smoothness spaces such as Hölder and Sobolev Spaces. Roughly speaking, the Besov space $B^\alpha_{p,q}$ contains functions having α bounded derivatives in L^p norm, the third parameter q gives a finer gradation of smoothness. Full details of Besov spaces are given, for example, in [23] and [12]. A wavelet ψ is called *r-regular* if ψ has r vanishing moments and r continuous derivatives. For a given r-regular mother wavelet ψ with $r > \alpha$ and a fixed primary resolution level j_0, the Besov sequence norm $\|\cdot\|_{b^\alpha_{p,q}}$ of the wavelet coefficients of a function f is then defined by

$$\|f\|_{b^\alpha_{p,q}} = \|\tilde{\underline{\theta}}_{j_0}\|_p + \left(\sum_{j=j_0}^{\infty} (2^{js} \|\underline{\theta}_j\|_p)^q \right)^{\frac{1}{q}}, \tag{18}$$

where $\tilde{\underline{\theta}}_{j_0}$ 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 (18) of the wavelet coefficients of the function. See [19]. Define the Besov ball

$$B^\alpha_{p,q}(M) = \left\{ f; \|f\|_{b^\alpha_{p,q}} \le M \right\} \tag{19}$$

and set

$$F^\alpha_{p,q}(M, \varepsilon, v) = \{ f : f \in B^\alpha_{p,q}(M), f(t) \in [\varepsilon, v] \text{ for all } t \in [0,1] \}, \tag{20}$$

where $[\varepsilon, v]$ with $\epsilon < v$ is a compact set in the interior of the mean parameter space of the natural exponential family.

We first show that the center of the approximate Gaussian regression problem remains in the Besov with the same index (α, p, q).

Lemma 3. *For H_m defined in (2) there exists a constant $M' > 0$ such that*

$$H_m(f) \in B^\alpha_{p,q}(M')$$

for all f in $F^\alpha_{p,q}(M, \varepsilon, v)$ defined in (20).

The following theorem shows that our estimators achieve near optimal global adaptation under mean integrated squared error for a wide range of Besov balls.

Theorem 2. *Suppose the wavelet ψ is r-regular. Let $X_i \sim NEF(f(t_i))$, $i = 1, \ldots, n$, $t_i = \frac{i}{n}$. Let $T = \frac{n}{\log^{1+\nu} n}$ with $\nu > 0$. Then the estimator $\hat{f}_{BJS}$ defined in*

(17) *satisfies*

$$\sup_{f\in F^{\alpha}_{p,q}(M,\varepsilon,v)} \mathbb{E}\|\hat{f}_{BJS}-f\|_2^2 \leq \begin{cases} Cn^{-\frac{2\alpha}{1+2\alpha}}, & p\geq 2,\ \alpha\leq r,\ \textit{and}\ \frac{2\alpha^2}{1+2\alpha}>\frac{1}{p}, \\ Cn^{-\frac{2\alpha}{1+2\alpha}}(\log n)^{\frac{2-p}{p(1+2\alpha)}}, & 1\leq p<2,\ \alpha\leq r,\ \textit{and}\ \frac{2\alpha^2}{1+2\alpha}>\frac{1}{p}. \end{cases}$$

For functions of spatial inhomogeneity, the local smoothness of the functions varies significantly from point to point and global risk given in Theorem 2 cannot wholly reflect the performance of estimators at a point. For local performance and spatial adaptivity, it is more appropriate to use the pointwise mean squared error

$$R(\widehat{f}(t_0),\ f(t_0)) = \mathbb{E}(\widehat{f}(t_0)-f(t_0))^2. \tag{21}$$

The local smoothness of a function can be measured by its local Hölder smoothness index. For a fixed point $t_0\in(0,1)$ and $0<\alpha\leq 1$, define the local Hölder class $\Lambda^{\alpha}(M,t_0,\delta)$ as follows:

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

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 $\alpha'=\alpha-\lfloor\alpha\rfloor$. Define

$$F^{\alpha}(M,t_0,\delta,\varepsilon,v)=\{f: f\in\Lambda^{\alpha}(M,t_0,\delta), f(x)\in[\varepsilon,v] \text{ for all } x\in[0,1]\}.$$

In Gaussian nonparametric regression setting, it is a well known fact that for estimation at a point, one must pay a price for adaptation. The optimal rate of convergence for estimating $f(t_0)$ over function class $\Lambda^{\alpha}(M,t_0,\delta)$ with α completely known is $n^{-2\alpha/(1+2\alpha)}$. Lepski [18] and Brown and Low [5] showed that one has to pay a price for adaptation of at least a logarithmic factor. It is shown that the local adaptive minimax rate over the Hölder class $\Lambda^{\alpha}(M,t_0,\delta)$ is $(\log n/n)^{2\alpha/(1+2\alpha)}$.

The following theorem shows that our estimators achieve the optimal local adaptation with the minimal cost.

Theorem 3. *Suppose the wavelet ψ is r-regular with $0<\alpha\leq r$. Let $t_0\in(0,1)$ be fixed. Let $X_i\sim NEF(f(t_i))$, $i=1,\ldots,n$, $t_i=\frac{i}{n}$. Let $T=\frac{n}{\log^{1+\nu}n}$ with $\nu>0$. Then for $\hat{f}=\hat{f}_{BJS}$*

$$\sup_{F^{\alpha}(M,t_0,\delta,\varepsilon,v)} \mathbb{E}(\widehat{f}(t_0)-f(t_0))^2 \leq C\cdot\left(\frac{\log n}{n}\right)^{\frac{2\alpha}{1+2\alpha}}. \tag{22}$$

Theorem 3 shows that the BlockJS estimator is spatially adaptive, without prior knowledge of the smoothness of the underlying functions.

5. Discussions

The general principle of turning complicated statistical models into simpler ones is practically important and appealing. We developed in this paper an explicit and practical procedure which turns nonparametric regression in natural exponential

families into a standard Gaussian nonparametric regression. The method and the results extend the scope of those introduced in [3]. A key component of the procedure in [3] is the use of a mean-matching VST and its inverse. The mean-matching VST only exists in natural exponential families with a quadratic variance function. This thus limits the applicability of the method. In addition, although the use of the mean-matching VST and its inverse reduces the bias due to the transformation, it does not completely eliminate the transformation bias. As a result the bin size needs to be a power of n and as a consequence the regression function f is required to be smoother than that needed in the present paper.

In our setting the bin size m is logarithmic in n, smaller than the choice $m \asymp n^{1/4}$ in [3]. As a result, the discretization error is smaller in our analysis. Consequently, the procedure proposed in this paper attains the optimal rates of convergence over a wider range of Besov classes as shown in Section 4. In Theorem 2 we require $\frac{2\alpha^2}{1+2\alpha} > \frac{1}{p}$, i.e., $\alpha - \frac{1}{p} > \frac{2\alpha}{1+2\alpha}$ which is weaker than the condition $\frac{3}{2}(\alpha - \frac{1}{p}) > \frac{2\alpha}{1+2\alpha}$ in Theorem 1 of [3]. Similarly the local adaptation optimality in Theorem 3 is attained for $\alpha > 0$, while $\alpha > 1/6$ was assumed in [3].

Technical analyses are more challenging in this paper. Since $m \asymp \log^{1+\nu} n$, it is easy to see m^{-D} for each finite $D > 0$ is no longer $o(n^{-2\alpha/(2\alpha+1)})$. Actually we have $m^{-D} \gg n^{-2\alpha/(2\alpha+1)}$, thus the polynomial tail bounds obtained in [3] are not negligible anymore. In this paper, we provide a finer analysis with exponential tail bounds. Note that $\exp(-c\log^{1+\nu_1} n) = o(n^{-2\alpha/(2\alpha+1)})$ for any $c > 0$ and $\nu_1 > 0$.

6. Proofs

In this section we give proofs for Theorems 1 and 2. Theorem 3 can be proved in a similar way as Theorem 4 in [2] by applying Proposition 1 in Section 6.2. We begin by proving Lemmas 2, 3 and 4. Lemmas 2 and 4 are needed to establish Theorem 1 in which an approximation bound between our model and a Gaussian regression model is given explicitly. Finally we apply Theorem 1 and risk bounds for block thresholding estimators in Proposition 1 to prove Theorem 2.

6.1. Proof of preparatory technical results

Proof of Lemma 2. Write

$$\begin{aligned} & G\left(\frac{X}{m}\right) - G_m(\mu) - \frac{1}{\sqrt{m}}Z \\ &= \left[G\left(\frac{X}{m}\right) - G\left(\mu + \sqrt{\frac{V}{m}}Z\right)\right] + \left[G\left(\mu + \sqrt{\frac{V}{m}}Z\right) - G(\mu) - \frac{1}{\sqrt{m}}Z\right] \\ &\quad + \left[G(\mu) - G_m(\mu)\right]. \end{aligned}$$

Taylor expansion yields

$$\begin{aligned} G\left(\frac{X}{m}\right) - G_m(\mu) - \frac{1}{\sqrt{m}}Z = G'(\mu_1^*)\left(\frac{X}{m} - \mu - \sqrt{\frac{V}{m}}Z\right) \\ + G''(\mu_2^*)\frac{V}{m}Z^2 + G(\mu) - G_m(\mu). \end{aligned}$$

From Lemma 1 we have

$$\mathbb{P}\left(m\left|\frac{X}{m} - \mu - \sqrt{\frac{V}{m}}Z\right| > a\right) \le c_1 \exp(-c_2 a),$$

Since Z is standard normal, Z^2 has an exponential tail

$$\mathbb{P}\left(m\left|\frac{V}{m}Z^2\right| > a\right) \leq c_3 \exp\left(-c_4 a\right).$$

It is easy to see that

$$|G\left(\mu\right) - G_m\left(\mu\right)| = O\left(\frac{1}{m}\right).$$

(cf. [3], Lemma 1). Note that for any $\varepsilon > 0$,

$$\mathbb{P}(|\mu_1^*| > \mu + \varepsilon) \leq c_5 \exp\left(-c_6 \varepsilon^2 m\right), \text{and } \mathbb{P}(|\mu_2^*| > \mu + \varepsilon) \leq c_7 \exp\left(-c_8 \varepsilon^2 m\right),$$

Thus we have

$$\mathbb{P}(m|\xi_j| > a) \leq c_9 \exp\left(-c_{10} a\right) + c_{11} \exp\left(-c_{12} m\right). \qquad \square$$

Proof of Lemma 3. For $f(t) \in [\varepsilon, v]$ uniformly over all t and $f \in B_{p,q}^{\alpha}(M)$, if $c_l = \sup_{y\in[\varepsilon,v]} |H^{(l)}(y)|$, $l = 0, \ldots, \lfloor\alpha\rfloor + 1$, are finite constants independent of m, then we have $H(f) \in B_{p,q}^{\alpha}(M')$ where

$$M' = c_0 + cM\left[\sum_{l=1}^{\lfloor\alpha\rfloor+1} c_l v^{l-1} + c_{\lfloor\alpha\rfloor+1}\right]$$

for some $c > 0$, according to Theorem 3 on page 344 and Remark 3 on page 345 of [21]. Now we show c_l are finite constants independent of m. Recall that $H(\mu) = EG[(X_1 + \cdots + X_m)/m]$. For $K = \lfloor\alpha\rfloor + 1 > 0$, Taylor expansion yields

$$\begin{aligned} H\left(\mu\right) - G\left(\mu\right) &= E\left[G\left(\bar{X}\right) - G\left(\mu\right)\right] \\ &= E\left\{\sum_{k=1}^{K} \frac{G^{(k)}\left(\mu\right)}{k!}\left(\bar{X} - \mu\right)^k + \int_{\mu}^{\bar{X}} \frac{G^{(K)}\left(t\right)}{K!}\left(\bar{X} - t\right)^K dt\right\} \\ &= R_1\left(\mu\right) + R_2\left(\mu\right). \end{aligned}$$

It is easy to see $\sup_{\mu\in[\varepsilon,v]} |R_1^{(l)}(\mu)|$ is bounded, since $[E(X_i - \mu)^k]^{(l)}$ and $G^{(k+l)}(\mu)$ are bounded over $\mu \in [\varepsilon, v]$, and $k \leq K$ finite. The distribution of $\bar{X}$ is from a natural exponential family

$$q(x|\theta) = e^{m\theta x - m\psi(\theta)} h_m(x)$$

(cf. [1]). Note that $\sup_{\mu\in[\varepsilon,v]} |R_2^{(l)}(\mu)|$ is bounded by $\sup_{\mu\in[\varepsilon,v]} |\frac{d^l}{d\theta^l} R_2(\mu(\theta))|$ up to a constant factor. That fact together with some straightforward calculations gives, for $1 \leq l \leq K$,

$$\begin{aligned} \sup_{\mu\in[\varepsilon,v]} \left|R_2^{(l)}\left(\mu\right)\right| &\leq C_K \sup_{\mu\in[\varepsilon,v]} E\left[\sum_{k=1}^{l} \left|\bar{X} - \mu\right|^{K-k+1} \cdot \left(\left|m\bar{X} - m\mu\right|^{l-k} + m^{(l-k)/2}\right)\right] \\ &= C_K \sup_{\mu\in[\varepsilon,v]} E\left[\sum_{k=1}^{l} m^{l/2-K/2-1/2}\left(\left|\sqrt{m}\left(\bar{X} - \mu\right)\right|^{K+l-2k+1} + 1\right)\right] \end{aligned}$$

which is at an order of $m^{-1/2}$, and thus bounded. $\square$

The variance stabilizing transformation considered in Equation (1) is for i.i.d. observations. In the function estimation procedure, observations in each bin are independent but not identically distributed. However, observations in each bin can be treated as i.i.d. random variables through coupling. Let $X_i \sim NEF(\mu_i)$, $i = 1, \ldots, m$, be independent. Here the means μ_i are "close" but not equal. Let $X_{i,c}$ be a set of i.i.d. random variables with $X_{i,c} \sim NEF(\mu_c)$. We define

$$D = G\left(\frac{\sum_{i=1}^m X_i}{m}\right) - G\left(\frac{\sum_{i=1}^m X_{i,c}}{m}\right).$$

If $\mu_c = \max_i \mu_i$, it is easy to see $\mathbb{E}D \leq 0$ since $X_{i,c}$ is stochastically larger than X_i for all i [see e.g. 17]. Similarly $\mathbb{E}D \geq 0$ when $\mu_c = \min_i \mu_i$. We will select a

$$\mu_c^* \in \left[\min_i \mu_i, \max_i \mu_i\right] \tag{23}$$

such that $\mathbb{E}D = 0$, which is possible by the intermediate value theorem. In the following lemma we construct i.i.d. random variables $X_{i,c} \sim NEF(\mu_c^*)$ on the sample space of X_i such that D is very small and has negligible contribution to the final risk bounds in Theorems 2 and 3.

Lemma 4. *Let $X_i \sim NEF(\mu_i)$, $i = 1, \ldots, m$, be independent with $\mu_i \in [\varepsilon, v]$, a compact subset in the interior of the mean parameter space of the natural exponential family. Assume that $|\min_i \mu_i - \max_i \mu_i| \leq C\delta$. Then there are i.i.d. random variables $X_{i,c}$ where $X_{i,c} \sim NEF(\mu_c^*)$ with $\mu_c^* \in [\min_i \mu_i, \max_i \mu_i]$ such that $\mathbb{E}D = 0$ and*

(i)

$$\mathbb{P}\left(\{X_i \neq X_{i,c}\}\right) \leq C\delta, \tag{24}$$

(ii) *and for any fixed integer $k \geq 1$ there exists a constant $C_k > 0$ such that for all $a > 0$,*

$$\mathbb{P}(|D| > a) \leq c_1 \exp\left(-c_2 a^2 m\right) + c_3 \exp\left(-c_4 m\right). \tag{25}$$

Proof of Lemma 4. (i). There is a classical coupling identity for the Total variation distance. Let P and Q be distributions of two random variables X and Y on the same sample space respectively, then there is a random variable Y_c with distribution Q such that $\mathbb{P}(X \neq Y_c) = |P - Q|_{TV}$. See, for example, page 256 in [20]. The proof for the inequality (24) follows from that identity and the inequality that $|NEF(\mu_i) - NEF(\mu_c^*)|_{TV} \leq C|\mu_i - \mu_c^*|$ for some $C > 0$ which only depends on the family of the distribution of X_i and $[\varepsilon, v]$.

(ii) Using Taylor expansion we can rewrite D as $D = G'(\zeta)\frac{\sum_{i=1}^m (X_i - X_{i,c})}{m}$ for some ζ in between $\frac{\sum_{i=1}^m X_i}{m}$ and $\frac{\sum_{i=1}^m X_{i,c}}{m}$. Write

$$\frac{\sum_{i=1}^m (X_i - X_{i,c})}{m} = \frac{\sum_{i=1}^m (X_i - EX_i)}{m} - \frac{\sum_{i=1}^m (X_{i,c} - EX_{i,c})}{m} + \frac{\sum_{i=1}^m (EX_i - EX_{i,c})}{m}.$$

Since the distributions X_i and $X_{i,c}$ are in exponential family, we have

$$\mathbb{P}\left(\left|\frac{\sum_{i=1}^m (X_i - EX_i)}{m}\right| \geq a\right) \leq c_1 \exp\left(-c_2 a^2 m\right),$$
$$\mathbb{P}\left(\left|\frac{\sum_{i=1}^m (X_{i,c} - EX_{i,c})}{m}\right| \geq a\right) \leq c_3 \exp\left(-c_4 a^2 m\right).$$

(cf. [15]). Note that

$$\left|\frac{\sum_{i=1}^{m}(EX_i - EX_{i,c})}{m}\right| \leq c_5\delta.$$

Thus we have

$$\mathbb{P}\left(\left|\frac{\sum_{i=1}^{m}(X_i - X_{i,c})}{m}\right| > a\right) \leq c_6 \exp\left(-c_7 a^2 m\right)$$

when $a > 2c_5\delta$. The equation is apparently true when $a \leq 2C\delta$. Since $X_i - X_{i,c}$ are independent, it can be shown that Note that Thus the first inequality in (25) follows immediately by observing that $G'(\zeta)$ is bounded with a probability approaching to 1 exponentially fast, since for any $\varepsilon > 0$,

$$\mathbb{P}(|\zeta| > \mu + \varepsilon) \leq c_8 \exp\left(-c_9\varepsilon^2 m\right). \qquad \square$$

6.2. Proof of Theorem 1

From Lemma 4, there exist $Y_{j,c}^*$ where $X_{i,c} \sim NEF(f_j^*)$ with

$$f_{j,c}^* \in \left[\min_{jm+1\leq i\leq(j+1)m} f\left(\frac{i}{n}\right), \max_{jm+1\leq i\leq(j+1)m} f\left(\frac{i}{n}\right)\right]$$

as in equation (23) such that

$$\mathbb{E}\left[Y_j^* - Y_{j,c}^*\right] = 0, \tag{26}$$

$$\mathbb{P}(|Y_j^* - Y_{j,c}^*| > a) \leq c_1 \exp\left(-c_2 a^2 m\right) + c_3 \exp\left(-c_4 m\right). \tag{27}$$

Let $f_{j,c}^* = f(j_*/T)$, where $jm + 1 \leq j_* \leq (j+1)m$, by the intermediate value theorem. Lemmas 1 and 2 together yield

$$Y_j^* = H\left(f\left(\frac{j_*}{T}\right)\right) + m^{-\frac{1}{2}} Z_j + \xi_j, \quad j = 1, 2, \ldots, T, \tag{28}$$

and

$$\mathbb{P}(|\xi_j| > a) \leq c_1 \exp\left(-c_2 a^2 m\right) + c_3 \exp\left(-c_4 m\right). \tag{29}$$

Theorem 1 then follows immediately by combining equations (26) – (28). $\square$

6.3. Proof of Theorem 2

We first collect a few technical lemmas.

From (10) in Theorem 1 we can write $\frac{1}{\sqrt{T}} Y_i^* = \frac{H(f(\frac{j_*}{T}))}{\sqrt{T}} + \frac{Z_i}{\sqrt{n}} + \frac{\xi_i}{\sqrt{T}}$. Let $(u_{j,k}) = T^{-\frac{1}{2}} W \cdot Y^*$ be the discrete wavelet transform of the binned and transformed data. Then one may write

$$u_{j,k} = \theta'_{j,k} + \frac{1}{\sqrt{n}} z_{j,k} + \xi_{j,k}, \tag{30}$$

where θ'_{jk} are the discrete wavelet transform of $(H(f(\frac{j_*}{T})))/\sqrt{T})$ which are approximately equal to the true wavelet coefficients of $H(f)$, $z_{j,k}$ are the transform of the Z_i's and so are i.i.d. $N(0,1)$ and $\xi_{j,k}$ are respectively the transforms of $(\frac{\xi_i}{\sqrt{T}})$. We may obtain the following result on the risk bound for a single block. Its proof is close to that of Proposition 2 in [6].

Proposition 1. *Let the empirical wavelet coefficients* $u_{j,k} = \theta'_{j,k} + \frac{1}{\sqrt{n}} z_{j,k} + \xi_{j,k}$ *be given as in (30) and let the block thresholding estimator* $\hat{\theta}_{j,k}$ *be defined as in (16). Then for all* $j \le J_*$ *we have*

(i) for $\varepsilon_n = o\left(1/\sqrt{m}\right)$ *and some constant* $C > 0$,

$$\mathbb{P}\left(\sqrt{n}\,|\xi_{j,k}| \ge \varepsilon_n\right) \le C\left[\exp\left(-\frac{1}{C}\varepsilon_n^2 m\right) + \exp\left(-\frac{1}{C}\log^{1+\gamma} n\right)\right], \tag{31}$$

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

$$\mathbb{E}\sum_{(j,k)\in B_j^i}(\hat{\theta}_{j,k} - \theta'_{j,k})^2 \le \min\left\{4\sum_{(j,k)\in B_j^i}(\theta'_{j,k})^2,\ 8\lambda_* L n^{-1}\right\} + L n^{-2+\tau}; \tag{32}$$

(iii) for any $0 < \tau < 1$, *there exists a constant* $C_\tau > 0$ *depending on* τ *only such that for all* $(j,k) \in B_j^i$

$$\mathbb{E}(\hat{\theta}_{j,k} - \theta'_{j,k})^2 \le C_\tau \cdot \min\left\{\max_{(j,k)\in B_j^i}\{\left(\theta'_{j,k}\right)^2\},\ L n^{-1}\right\} + n^{-2+\tau}. \tag{33}$$

The first part plays an important role to prove the last two parts. It follows from classical concentration inequalities for sum of independent random variables with exponential moments. We use the second part to prove Theorem 2. The third part is needed to prove Theorem 3.

For $0 < d \le 1$, define the Lipschitz class $\Lambda^d(M)$ by

$$\Lambda^d(M) \ = \ \{f : |f(t_1) - f(t_2)| \le M\,|t_1 - t_2|^d \ \ 0 \le t_1,\ t_2 \le 1\}.$$

and

$$F^d(M, \varepsilon, v) = \{f : f \in \Lambda^d(M), f(t) \in [\varepsilon, v]\,, \text{ for all } t \in [0,1]\},$$

where $[\varepsilon, v]$ with $\epsilon < v$ is a compact set in the interior of the mean parameter space of the natural exponential family.

The following is a standard bound for wavelet approximation error. It follows directly from Lemma 1 in [8].

Lemma 5. *Let* $T_* = 2^{J_*}$ *and* $d = \min(\alpha - \frac{1}{p}, 1)$. *Set* $\bar{g}_{J_*}(x) = \sum_{k=1}^{T_*} \frac{1}{\sqrt{T_*}} \times H(f(k/n))\phi_{J_*,k}(x)$. *Then for some constant* $C > 0$

$$\sup_{g \in F^{\alpha}_{p,q}(M,\varepsilon)} \|\bar{g}_{J_*} - H\,(f)\,\|_2^2 \le C T_*^{-2d}. \tag{34}$$

Let $\widetilde{H(f)} = \max\{\widehat{H(f)}, 0\}$. We have

$$\begin{aligned}\mathbb{E}\|\widehat{f} - f\|_2^2 &= \mathbb{E}\|H^{-1}[\widetilde{H\,(f)}] - H^{-1}[H\,(f)]\|_2^2 = \mathbb{E}\|(H^{-1})'\,(g)\,[\widetilde{H\,(f)} - H\,(f)]\|_2^2 \\ &\le C\mathbb{E}\int V\left(H^{-1}\,(g)\right)[\widehat{H\,(f)} - H\,(f)]^2\,dt,\end{aligned}$$

where g is a function in between $\widetilde{H(f)}$ and $H(f)$. We will first give a lemma which implies $V(H^{-1}(g))$ is bounded with high probability, then prove Theorem 2 by establishing a risk bound for estimating $H(f)$. See [3] for a proof.

Lemma 6. *Let $\widehat{H(f)}$ be the BlockJS estimator of $H(f)$ defined in Section 2. Then there exists a constant $C > 0$ such that*

$$\sup_{f \in F^{\alpha}_{p,q}(M,\varepsilon,v)} \mathbb{P}\left\{ \left\| \widehat{H(f)} \right\|_{\infty} > C \right\} \le C_l n^{-l}$$

for any $l > 1$, where C_l is a constant depending on l.

Now we are ready to prove Theorem 2. Note that H^{-1} is an increasing and nonnegative function. Lemma 6 implies that there exists a constant C such that

$$\sup_{f \in F^{\alpha}_{p,q}(M,\varepsilon,v)} \mathbb{P}\left\{ \left\| V\left(H^{-1}(g)\right) \right\|_{\infty} > C \right\} \le C_l n^{-l}$$

for any $l > 1$. Thus it is enough to show $\sup_{f \in F^{\alpha}_{p,q}(M,\varepsilon,v)} \mathbb{E}\|\widehat{H(f)} - H(f)\|_2^2 \le Cn^{-\frac{2\alpha}{1+2\alpha}}$ for $p \ge 2$ and $Cn^{-\frac{2\alpha}{1+2\alpha}} (\log n)^{\frac{2-p}{p(1+2\alpha)}}$ for $1 \le p < 2$ under assumptions in Theorem 2.

We are now ready to prove our main results.

Proof of Theorem 2. Let Y and $\hat{\theta}$ be given as in (3) and (16) respectively. Then,

$$\begin{aligned} \mathbb{E}\|\widehat{H(f)} - H(f)\|_2^2 &= \sum_k \mathbb{E}(\hat{\tilde{\theta}}_{j_0,k} - \tilde{\theta}_{j,k})^2 + \sum_{j=j_0}^{J_*-1} \sum_k \mathbb{E}(\hat{\theta}_{j,k} - \theta_{j,k})^2 + \sum_{j=J_*}^{\infty} \sum_k \theta_{j,k}^2 \\ &\equiv S_1 + S_2 + S_3. \end{aligned} \tag{35}$$

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

$$S_1 = 2^{j_0} n^{-1} \epsilon^2 = o(n^{-2\alpha/(1+2\alpha)}) \tag{36}$$

Note that for $x \in \mathbb{R}^m$ and $0 < p_1 \le p_2 \le \infty$,

$$\|x\|_{p_2} \le \|x\|_{p_1} \le m^{\frac{1}{p_1} - \frac{1}{p_2}} \|x\|_{p_2}. \tag{37}$$

Since $f \in B^{\alpha}_{p,q}(M)$, so $2^{js}(\sum_{k=1}^{2^j} |\theta_{jk}|^p)^{1/p} \le M$. Now (37) yields that

$$S_3 = \sum_{j=J_*}^{\infty} \sum_k \theta_{j,k}^2 \le C 2^{-2J_*(\alpha \wedge (\alpha + \frac{1}{2} - \frac{1}{p}))}. \tag{38}$$

Note that

$$\left| H\left(f\left(\frac{j_*}{T_*}\right)\right) - H\left(f\left(\frac{j}{T_*}\right)\right) \right| \le C T_*^{-d}. \tag{39}$$

Proposition 1, Lemma 5 and equation (39) yield that

$$\begin{aligned} S_2 &\le 2 \sum_{j=j_0}^{J_*-1} \sum_k \mathbb{E}(\hat{\theta}_{j,k} - \theta'_{j,k})^2 + 2 \sum_{j=j_0}^{J_*-1} \sum_k (\theta'_{j,k} - \theta_{j,k})^2 \\ &\le \sum_{j=j_0}^{J_*-1} \sum_{i=1}^{2^j/L} \min\left\{ 8 \sum_{(j,k) \in B_j^i} \theta_{j,k}^2,\ 8\lambda_* L n^{-1} \right\} + 10 \sum_{j=j_0}^{J_*-1} \sum_k (\theta'_{j,k} - \theta_{j,k})^2 \\ &\le \sum_{j=j_0}^{J_*-1} \sum_{i=1}^{2^j/L} \min\left\{ 8 \sum_{(j,k) \in B_j^i} \theta_{j,k}^2,\ 8\lambda_* L n^{-1} \right\} + C\,(T_*)^{-2d} \end{aligned} \tag{40}$$

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 (40) and (37) yield

$$S_2 \leq 8\lambda_* \sum_{j=j_0}^{J_1-1} \sum_{i=1}^{2^j/L} Ln^{-1} + 8 \sum_{j=J_1}^{J_*-1} \sum_{k} \theta_{j,k}^2 + C \left(\frac{T}{\log^{1+\gamma} n} \right)^{-2d} \leq Cn^{-2\alpha/(1+2\alpha)}. \tag{41}$$

By combining (41) with (36) and (38), we have $\mathbb{E}\|\hat{\theta} - \theta\|_2^2 \leq Cn^{-2\alpha/(1+2\alpha)}$, for $p \geq 2$.

Now let us consider the case $p < 2$. First we state the following lemma without proof.

Lemma 7. *Let $0 < p < 1$ and $S = \{x \in \mathbb{R}^k : \sum_{i=1}^k x_i^p \leq B,\ x_i \geq 0,\ i = 1, \ldots, k\}$. Then $\sup_{x \in S} \sum_{i=1}^k (x_i \wedge A) \leq B \cdot A^{1-p}$ for all $A > 0$.*

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

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

It then follows from Lemma 7 that

$$\sum_{j=J_2}^{J_*-1} \sum_{i=1}^{2^j/L} \min \left\{ 8 \sum_{(j,k) \in B_j^i} \theta_{j,k}^2,\ 8\lambda_* Ln^{-1} \right\} \leq Cn^{-\frac{2\alpha}{1+2\alpha}} (\log n)^{\frac{2-p}{p(1+2\alpha)}}. \tag{42}$$

On the other hand,

$$\begin{aligned} &\sum_{j=j_0}^{J_2-1} \sum_{i=1}^{2^j/L} \min \left\{ 8 \sum_{(j,k) \in B_j^i} \theta_{j,k}^2,\ 8\lambda_* Ln^{-1} \right\} \\ &\quad \leq \sum_{j=j_0}^{J_2-1} \sum_{b} 8\lambda_* Ln^{-1} \leq Cn^{-\frac{2\alpha}{1+2\alpha}} (\log n)^{\frac{2-p}{p(1+2\alpha)}}. \end{aligned} \tag{43}$$

Putting (36), (38), (42) and (43) together yields $\mathbb{E}\|\hat{\theta} - \theta\|_2^2 \leq Cn^{-\frac{2\alpha}{1+2\alpha}} (\log n)^{\frac{2-p}{p(1+2\alpha)}}$. □

References

[1] Brown, L. D. (1986). *Fundamentals of Statistical Exponential Families: With Applications in Statistical Decision Theory*. IMS, Hayworth, CA, USA.

[2] Brown, L. D., Cai, T. T. and Zhou, H. H. (2008). Robust nonparametric estimation via wavelet median regression. *Ann. Statist.* **36** 2055–2084.

[3] Brown, L. D., Cai, T. T. and Zhou, H. H. (2010). Nonparametric regression in exponential families. *Ann. Statist.* To appear.

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

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

[6] Brown, L., Cai, T., Zhang, R., Zhao, L. and Zhou, H. (2010). The root-unroot algorithm for density estimation as implemented via wavelet block thresholding. *Probab. Theory Related Fields* **146** 401–433.
[7] Cai, T. T. (1999). Adaptive wavelet estimation: A block thresholding and oracle inequality approach. *Ann. Statist.* **27** 898–924.
[8] Cai, T. T. (2002). On block thresholding in wavelet regression: Adaptivity, block size, and threshold level. *Statist. Sinica* **12** 1241–1273.
[9] Cai, T. T. and Silverman, B. W. (2001). Incorporating Information on Neighboring Coefficients into Wavelet Estimation. *Sankhya Ser. B* **63** 127–148.
[10] Cai, T. T. and Zhou, H. H. (2009). Asymptotic equivalence and adaptive estimation for robust nonparametric regression. *Ann. Statist.* **37** 3204–3235.
[11] Daubechies, I. (1992). *Ten Lectures on Wavelets. CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA.
[12] DeVore, R. and Popov, V. (1988). Interpolation of Besov spaces. *Trans. Amer. Math. Soc.* **305** 397 – 414.
[13] Hall, P., Kerkyacharian, G. and Picard, D. (1998). Block threshold rules for curve estimation using kernel and wavelet methods. *Ann. Statist.* **26** 922–942.
[14] Hoyle, M. H. (1973). Transformations–an introduction and a bibliography. *Internat. Statist. Rev.* **41** 203–223.
[15] 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* 111–131.
[16] Le Cam, L. (1986). *Asymptotic Methods in Statistical Decision Theory. Springer Series in Statistics*. Springer, New York.
[17] Lehmann, E. L. and Romano, J. P. (2005). *Testing Statistical Hypotheses*, 3rd ed. *Springer Texts in Statistics Series*. Springer, New York.
[18] Lepskii, O. V. (1990). On a problem of adaptive estimation in Gaussian white noise. *Theory Probab. Appl.* **35** 454–466.
[19] Meyer, Y. (1992). *Wavelets and Operators. Cambridge Studies in Advanced Mathematics* **37**. Cambridge Univ. Press, Cambridge.
[20] Pollard, D. (2002). *A user's guide to measure theoretic probability. Cambridge Series in Statistical and Probabilistic Mathematics* **8**. Cambridge Univ. Press, Cambridge.
[21] Runst, T. (1986). Mapping properties of non-linear operators in spaces of Triebel-Lizorkin and Besov type. *Anal. Math.* **12** 313–346.
[22] Strang, G. (1989). Wavelets and dilation equations: A brief introduction. *SIAM Rev.* **31** 614–627.
[23] Triebel, H. (1992). *Theory of Function Spaces. II. Monographs in Mathematics* **84**. Birkhäuser, Basel.
[24] Zhou, H. (2006). A note on quantile coupling inequalities and their applications. Technical report.