The Root-Unroot Algorithm for Density Estimation as Implemented via Wavelet Block Thresholding

Lawrence Brown, Tony Cai, Ren Zhang, Linda Zhao and Harrison Zhou

Abstract

We propose and implement a density estimation procedure which begins by turning density estimation into a nonparametric regression problem. This regression problem is created by binning the original observations into many small size bins, and by then applying a suitable form of root transformation to the binned data counts. In principle many common nonparametric regression estimators could then be applied to the transformed data. We propose use of a wavelet block thresholding estimator in this paper. Finally, the estimated regression function is un-rooted by squaring and normalizing.

The density estimation procedure achieves simultaneously three objectives: computational efficiency, adaptivity, and spatial adaptivity. A numerical example and a practical data example are discussed to illustrate and explain the use of this procedure. Theoretically it is shown that the estimator simultaneously attains the optimal rate of convergence over a wide range of the Besov classes. The estimator also automatically adapts to the local smoothness of the underlying function, and attains the local adaptive minimax rate for estimating functions at a point.

There are three key steps in the technical argument: Poissonization, quantile coupling, and oracle risk bound for block thresholding in the non-Gaussian setting. Some of the technical results may be of independent interest.

Keywords: Adaptation; Block thresholding; Coupling inequality; Density estimation; Nonparametric regression; Root-unroot transform; Wavelets.
AMS 2000 Subject Classification: Primary: 62G99; Secondary: 62F12, 62F35, 62M99.

1 Introduction

Density estimation and nonparametric regression are two fundamental nonparametric problems and have traditionally been treated separately in the literature. In this paper we describe a simple algorithm that allows density estimation to be treated as a nonparametric regression problem. We then show in detail how this algorithm can be used along with a wavelet regression estimator. The resulting procedure yields a convenient, effective density estimator that is adaptive and rate-optimal over a broad range of function classes.

Our basic algorithm can be termed a "root-unroot" procedure. It can easily be shown in special settings that the resulting density estimator shares analogous asymptotic optimality properties with the nonparametric regression estimator used in the algorithm. It is more complex to show this in broad adaptive settings. The current paper provides a complete proof of this in such a broad setting, and hence validates the root-unroot algorithm in this setting and provides strong evidence for the generality of the heuristic motivation underlying the algorithm.

As we describe in Section 3, the root-unroot procedure involves binning the observations and using a "mean-matching" square root of the bin counts. Virtually any reliable nonparametric regression estimator can be applied to these square rooted bin counts. The resulting regression estimator is then un-rooted and normalized in order to provide the final density estimator. Two key steps are the choice of bin-size and the use of the asymptotically "mean-matching" square root transformation. The algorithm is particularly convenient for the use of wavelet methods because with no difficulty it can provide the binary number of equally spaced regression observations for which a wavelet method is most suited.

There are two separate, though related, motivations for the root-unroot algorithm. First, recent results in asymptotic equivalence theory have shown that, under very mild regularity conditions, density estimation is asymptotically equivalent to nonparametric regression. For such equivalence results see Nussbaum (1996) and Brown, et al. (2004). Binning and taking the square-root of the bin counts lies at the heuristic heart of these equivalence results. It turns out that mean-matching allows simple and effective use of transformed bin-counts for the specific goal of density estimation without the necessity of implementing the much more complex equivalence mappings described in these papers.

A second motivation for the method involves the ideas of Poissonization and variance stabilization. Poissonization is discussed in several sources. See for example Le Cam (1974) and Low and Zhou (2007). The bin counts have a multinomial distribution and here Poissonization allows one to treat the bin counts as if they were independent Poisson variables. The variance stabilizing transformation for Poisson variables is any member of a family of square-root transformations. This family was discussed in Bartlett (1936) and Anscombe (1948). Anscombe described a particular member of this family that provides the greatest asymptotic control over the variance of the resulting transformed variables. However, for the present purposes it is more important (and often essential) to have better asymptotic control over the bias of the transformed variables, whereas optimal control of the variance term is not essential. The mean-matching transformation that we use provides the necessary degree of control over the bias of our resulting estimator.

The root transform turns the density estimation problem into a standard nonparametric regression problem. Virtually any good nonparametric regression procedure can then be applied. In this paper we shall use a wavelet estimator. Wavelet methodology has demonstrated considerable success in nonparametric regression in terms of spatial adaptivity and asymptotic optimality. In particular, block thresholding rules have been shown to possess impressive properties. The estimators make simultaneous decisions to retain or to discard all the coefficients within a block and increase estimation accuracy by utilizing information about neighboring coefficients. In the context of nonparametric regression local block thresholding has been studied, for example, in Hall, Kerkyacharian, and Picard (1998), Cai (1999, 2002) and Cai and Silverman (2001).

The wavelet regression estimator used in our implementation of the root-unroot algorithm is one such block thresholding procedure. It first divides the empirical coefficients at each resolution level into non-overlapping blocks and then simultaneously keeps or kills all the coefficients within a block, based on the sum of the squared empirical coefficients within that block. Motivated by the analysis of block thresholding rules for nonparametric regression in Cai (1999), the block size is chosen to be asymptotically $\log n$. It is shown that the estimator has a high degree of adaptivity. The root-unroot and block thresholding procedure is easy to implement and the procedure performs well for modest, realistic sample sizes and not only for sample sizes approaching infinity as is promised by asymptotic theory.

Theoretically we show that our density estimator possesses several desirable properties. It is shown that the estimator simultaneously attains the optimal rate of convergence under both the squared Hellinger distance loss and the integrated squared error over a wide range of the Besov classes. The estimator is also spatially adaptive: it attains the local adaptive minimax rate for estimating functions at a point. Implementation of our procedure is relatively straightforward, but the proof of the main theoretical results requires several steps. The first step is Poissonization. It is shown that the fixed sample size density problem is not essentially different from the density problem where the sample size is a Poisson random variable. The second step is the use of an appropriate version of the quantile coupling inequality of Komlós, Major and Tusnády (1975) to approximate the binned and root transformed data by independent normal variables. The third step is the derivation of a risk bound for block thresholding in the case where the noise is not necessarily Gaussian. Some of these technical results may be of independent interest.

It should be noted that density estimation has a long history and an extensive literature. See, e.g., Silverman (1986). The traditional estimators are typically linear and thus not spatially adaptive. A wavelet density estimator was first introduced by Donoho, et al. (1996). Minimax convergence rates over Besov classes were derived. It was shown that nonlinear thresholding approach can have significant advantage over the traditional linear methods. However, the wavelet density estimator introduced in that paper is not practical as the thresholds are not fully specified. The resulting density estimator is also not fully adaptive. In the case of term-by-term wavelet thresholding estimators, analysis of mean integrated squared error of single functions is also available. See Hall and Patil (1995).

We should also note that wavelet block thresholding has been used for density estimation in the literature. A local block thresholding density estimator was introduced in Hall, Kerkyacharian and Picard (1998). The estimator was shown to be globally rate optimal over a range of function classes of inhomogeneous smoothness under integrated squared error. However the estimator does not achieve the optimal local adaptivity under pointwise squared error. Chicken and Cai (2005) proposed a block thresholding density estimator which is adaptive under both the global and pointwise risk measures. However these estimators are not very practical as they are not easily implementable and require tuning parameters.

The paper is organized as follows. Section 2 discusses the mean-matching variance stabilizing root transform for a Poisson variable. We first discuss the general ideas for the root-unroot transform approach in Section 3 and then consider our specific wavelet block thresholding implementation of the general approach in Section 4. Theoretical properties of the root-unroot block thresholding density estimator are discussed in Section 5. In Section 6 we discuss the implementation of the estimator and application of the procedure to a call center data set. Technical proofs are given in Section 7.

2 Root transform

Variance stabilizing transformations, and closely related transformations to approximate normality, have been used in many statistical contexts. See Hoyle (1973) for a review of the extensive literature. See also Efron (1982) and Bar-Lev and Enis (1990). For Poisson distributions Bartlett (1936) was the first to propose the root transform $\sqrt{X}$ in a homoscedastic linear model where $X \sim \text{Poisson}(\lambda)$. Anscombe (1948) proposed improving the variance stabilizing properties by instead using $\sqrt{X + \frac{3}{8}}$. The constant $\frac{3}{8}$ is chosen to optimally stabilize the variance using the Taylor expansion. Anscombe's variance stabilizing transformation has also been briefly discussed in Donoho (1993) for density estimation.

In the context of nonparametric density estimation considered in the present paper, in comparison to variance stabilization, mean matching is more important. A mean-matching root transform is needed for minimizing the bias as well as stabilizing the variance. The goal of mean matching is to choose a constant c so that the mean of $\sqrt{X+c}$ is "closest" to $\sqrt{\lambda}$. The following lemma gives the expansions of the mean and variance of root transform of the form $\sqrt{X+c}$ where c is a constant. It can be seen easily that $c=\frac{1}{4}$ is the optimal choice for minimizing the bias $E(\sqrt{X+c})-\sqrt{\lambda}$ in the first order.

Lemma 1 *Let* $X \sim \text{Poisson}(\lambda)$ *with* $\lambda > 0$ *and let* $c \geq 0$ *be a constant. Then*

$$E(\sqrt{X+c}) = \lambda^{\frac{1}{2}} + \frac{4c-1}{8} \cdot \lambda^{-\frac{1}{2}} - \frac{16c^2-24c+7}{128} \cdot \lambda^{-\frac{3}{2}} + O(\lambda^{-\frac{5}{2}}) \tag{1}$$

$$\text{Var}(\sqrt{X+c}) = \frac{1}{4} + \frac{3-8c}{32} \cdot \lambda^{-1} + \frac{32c^2-52c+17}{128}\lambda^{-2} + O(\lambda^{-3}). \tag{2}$$

In particular, for $c = \frac{1}{4}$

$$E\left(\sqrt{X+\frac{1}{4}}\right) = \lambda^{\frac{1}{2}} - \frac{1}{64}\lambda^{-\frac{3}{2}} + O(\lambda^{-\frac{5}{2}}) \tag{3}$$

$$\text{Var}\left(\sqrt{X+\frac{1}{4}}\right) = \frac{1}{4} + \frac{1}{32}\lambda^{-1} + \frac{3}{64}\lambda^{-2} + O(\lambda^{-3}). \tag{4}$$

Lemma 1 shows that with $c=\frac{1}{4}$ the root transformed variable $\sqrt{X+c}$ has vanishing first order bias and almost constant variance. Lemma 1 follows from Taylor expansion and straightforward algebra. See also Anscombe (1948).

Figure 1 compares the mean and variance of three root transforms with $c=0$, $c=\frac{1}{4}$ and $c=\frac{3}{8}$. The left panel plots the bias $E_\lambda(\sqrt{X+c})-\sqrt{\lambda}$ as a function of λ for $c=0$, $c=\frac{1}{4}$ and $c=\frac{3}{8}$. It is clear from the plot that $c=\frac{1}{4}$ is the best choice among the three for matching the mean. For this value of c the bias is negligible for λ as small as 2. On the other hand, the root transform with $c=0$ yields significant negative bias and the transform with $c=\frac{3}{8}$ produces noticeable positive bias. The right panel plots the variance of $\sqrt{X+c}$ for $c=0$, $c=\frac{1}{4}$ and $c=\frac{3}{8}$. For variance $c=\frac{3}{8}$ is the best choice among the three when λ is not too small. The root transform with $c=\frac{1}{4}$ is slightly worse than but comparable to the case with $c=\frac{3}{8}$ and clearly $c=0$ is the worst choice of the three.

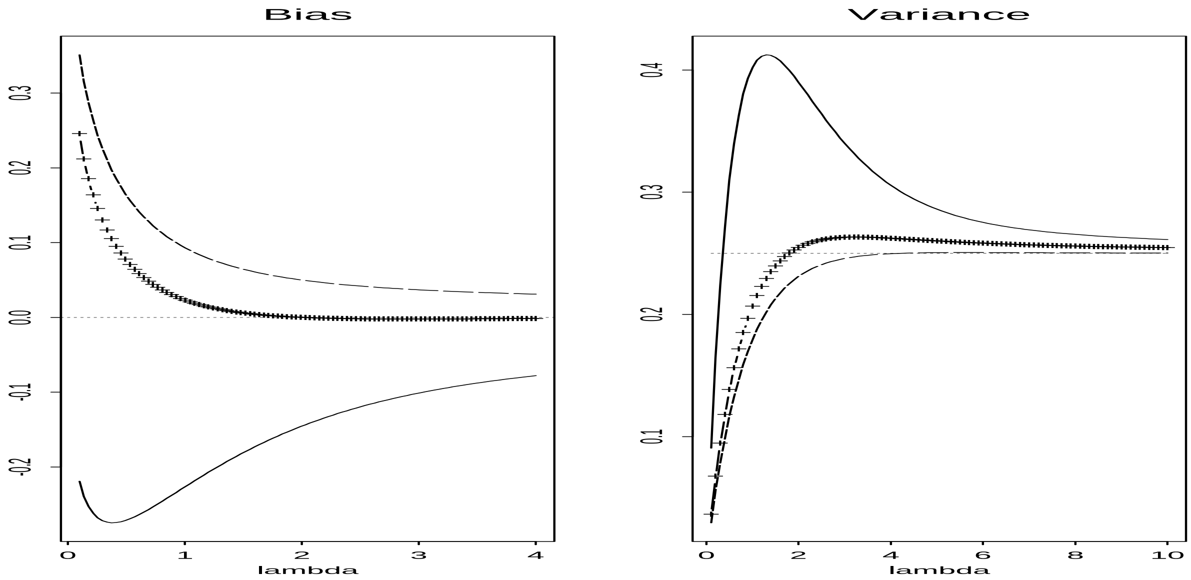

Figure 1: Comparison of the mean (left panel) and variance (right panel) of root transforms with $c = 0$ (solid line), $c = \frac{1}{4}$ (+ line) and $c = \frac{3}{8}$ (dashed line).

3 Density estimation through regression

We now consider density estimation, the main problem of interest in this paper. We shall discuss the general ideas for the root-unroot transform approach in this section and consider a specific wavelet block thresholding implementation of the general approach in Section 4.

Suppose that $\{X_1, ..., X_n\}$ is a random sample from a distribution with the density function f. We assume that the density function f is compactly supported on an interval, say the unit interval $[0, 1]$. Divide the interval into T equi-length subintervals and let Q_i be the number of observations on the i-th subinterval $I_i = [\frac{i-1}{T}, \frac{i}{T})$, $i = 1, 2, \ldots T$. Set $m = \frac{n}{T}$. The counts $\{Q_i\}$ can be treated as observations for a nonparametric regression directly, but this then becomes a heteroscedastic problem since the variance of Q_i is $mp_i(1-p_i/T)$ where $p_i = T \int_{\frac{i-1}{T}}^{\frac{i}{T}} f(x)dx$. Instead, we first apply the root transform discussed in Section 2, and treat $\{\sqrt{Q_i + \frac{1}{4}}\}$ as new regression observations. The constant $\frac{1}{4}$ is chosen to stabilize the variance and at the same time match the mean as discussed in Section 2. We will estimate $\sqrt{f}$ first, then square it back and normalize to get an estimator of f. After the density estimation problem is transferred into a regression problem, any nonparametric regression method can be applied. The general ideas for the root-unroot transform approach can be more formally explained as follows.

The first step of the procedure is binning. Let T be some positive integer (The choice of

T will be discussed later.) Divide $\{X_i\}$ into T equal length subintervals between 0 and 1. Let $Q_1, ..., Q_T$ be the number of observations in each of the subintervals. The Q_i's jointly have a multinomial distribution. Note that if the sample size is Poissonized, that is, it is not fixed but a Poisson random variable with mean n and independent of the X_i's, then the counts $\{Q_i : \ i = 1, ..., T\}$ are independent Poisson random variables with

$$Q_i \sim \text{Poisson}(mp_i) \quad \text{where} \quad p_i = T \int_{\frac{i-1}{T}}^{\frac{i}{T}} f(x)dx.$$

We then apply the mean-matching root transform discussed in Section 2. Set

$$Y_i = \sqrt{Q_i + \frac{1}{4}}, \quad \text{where } Q_i = \text{Card}(\{k : X_k \in I_i\}),\ i = 1, \cdots, T, \tag{5}$$

and treat $Y = (Y_1, Y_2, \ldots, Y_T)$ as the new equi-spaced sample for a nonparametric regression problem. Through binning and the root transform the density estimation problem has now been transferred to an equi-spaced, nearly constant variance nonparametric regression problem. Any good nonparametric regression procedure, such as a kernel, spline or wavelet procedure, can be applied to yield an estimator $\widehat{\sqrt{f}}$ of $\sqrt{f}$. The final density estimator can be obtained by normalizing the square of $\widehat{\sqrt{f}}$. Algorithmically, the root-unroot density estimation procedure can be summarized as follows.

1. **Binning:** Divide $\{X_i\}$ into T equal length intervals between 0 and 1. Let $Q_1, Q_2, ..., Q_T$ be the number of observations in each of the intervals.

2. **Root Transform:** Let $Y_i = \sqrt{Q_i + \frac{1}{4}}$, $i = 1, \cdots, T$, and treat $Y = (Y_1, Y_2, \ldots, Y_T)$ as the new equi-spaced sample for a nonparametric regression problem.

3. **Nonparametric Regression:** Apply your favorite nonparametric regression procedure to the binned and root transformed data Y to obtain an estimate $\widehat{\sqrt{f}}$ of $\sqrt{f}$.

4. **Unroot:** The density function f is estimated by $\widehat{f} = (\widehat{\sqrt{f}})^2$.

5. **Normalization:** The estimator $\hat{f}$ given in Step 4 may not integrate to 1. Set

$$\widetilde{f}(t) = \widehat{f}(t) / \int_0^1 \widehat{f}(t)dt$$

   and use $\widetilde{f}$ as the final estimator.

In this paper we combine the formal Root-Unroot procedure with the wavelet block thresholding method BlockJS given in Cai (1999). We will describe the BlockJS procedure in the next section and show in Section 5 that the resulting density estimator enjoys a high degree of adaptivity over a wide range of Besov classes. The numerical performance of this type of root-unroot procedure was investigated in Zhang (2002), using the VisuShrink wavelet estimator at Step 3.

Remark 1 An advantage of the root-unroot methodology is that it turns the density estimation problem to a standard homoscedastic nonparametric regression in which better-understood tools can then be used to construct confidence sets for the density, in addition to estimates. For the construction of confidence sets in regression setting, see, for example, Genovese and Wasserman (2005) and Cai and Low (2006).

4 Wavelets and block thresholding

Let $\{\phi, \psi\}$ be a pair of compactly supported father and mother wavelets with $\int \phi = 1$. 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

$$\phi^p_{j,k}(x) = \sum_{l=-\infty}^{\infty} \phi_{j,k}(x-l), \;\; \psi^p_{j,k}(x) = \sum_{l=-\infty}^{\infty} \psi_{j,k}(x-l), \qquad \text{for } x \in [0,1]$$

where $\phi_{j,k}(x) = 2^{j/2}\phi(2^j x - k)$ and $\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k)$. The collection $\{\phi^p_{j_0,k},\ k = 1, \ldots, 2^{j_0};\ \psi^p_{j,k},\ j \ge j_0 \ge 0, k = 1, ..., 2^j\}$ is then an orthonormal basis of $L^2[0,1]$, provided j_0 is large enough to ensure that the support of the wavelets at level j_0 is not the whole of $[0,1]$. The superscript "p" will be suppressed from the notation for convenience. A square-integrable function f on $[0,1]$ can be expanded into a wavelet series,

$$f(x) = \sum_{k=1}^{2^{j_0}} \xi_{j_0,k}\phi_{j_0,k}(x) + \sum_{j=j_0}^{\infty}\sum_{k=1}^{2^j} \theta_{j,k}\psi_{j,k}(x), \tag{6}$$

where $\xi_{j_0,k} = \langle f, \phi_{j_0,k}\rangle$ are the coefficients of the father wavelets at the coarsest level which represent the gross structure of the function f, and $\theta_{j,k} = \langle f, \psi_{j,k}\rangle$ are the wavelet coefficients which represent finer and finer structures as the resolution level j increases.

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 about the wavelets and discrete wavelet transform. Note that one can also use boundary corrected wavelet bases, instead of periodized wavelet bases. See Cohen, et. al (1993) and Daubechies (1994) for more on boundary corrected wavelet bases.

4.1 Root-unroot and block thresholding for density estimation

We now return to the density estimation problem. Set $J = J_n = \lfloor \log_2 n^{3/4} \rfloor$ and let $T = 2^J$. Divide $\{X_i\}$ into T equal length subintervals between 0 and 1, $I_i = [\frac{i-1}{T}, \frac{i}{T})$ for $i = 1, ..., T$. Apply the discrete wavelet transform to the binned and root transformed data $Y = (Y_1, \ldots, Y_T)$ where Y_i are given as in (5), and let $U = n^{-\frac{1}{2}} WY$ be the empirical wavelet coefficients, where W is the discrete wavelet transformation matrix. Write

$$U = (\tilde{u}_{j_0,1}, \cdots, \tilde{u}_{j_0,2^{j_0}}, u_{j_0,1}, \cdots, u_{j_0,2^{j_0}}, \cdots, u_{J-1,1}, \cdots, u_{J-1,2^{J-1}})'. \tag{7}$$

Here $\tilde{u}_{j_0,k}$ are the gross structure terms at the lowest resolution level, and $u_{j,k}$ ($j = j_0, \cdots, J-1, k = 1, \cdots, 2^j$) are empirical wavelet coefficients at level j which represent detail structure at scale 2^j. It is important to note that the empirical wavelet coefficients can be written as

$$u_{j,k} = \theta_{j,k} + \epsilon_{j,k} + \frac{1}{2\sqrt{n}} z_{j,k} + \xi_{j,k} \tag{8}$$

where θ_{jk} are the true wavelet coefficients of $\sqrt{f}$, $\epsilon_{j,k}$ 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 theoretical calculations given in Section 7 will show that both the approximation errors $\epsilon_{j,k}$ and the stochastic errors $\xi_{j,k}$ are negligible in certain sense. If these negligible errors are ignored then we have an idealized sequence model with noise level $\sigma = \frac{1}{2\sqrt{n}}$,

$$u_{j,k} \approx \theta_{j,k} + \frac{1}{2\sqrt{n}} z_{j,k}, \quad \text{where } z_{j,k} \overset{iid}{\sim} N(0,1). \tag{9}$$

The BlockJS procedure was proposed in Cai (1999) for nonparametric regression and was shown to achieve simultaneously three objectives: adaptivity, spatial adaptivity, and computational efficiency. We now apply the BlockJS procedure to the empirical coefficients $u_{j,k}$ as if they are observed as in (9). More specifically, at each resolution level j, the empirical wavelet coefficients $u_{j,k}$ are grouped into nonoverlapping blocks of length $L = \log n$ (in the numerical implementation we use $L = 2^{\lfloor \log_2(\log n) \rfloor}$). Let B_j^i denote the set of indices of the coefficients in the i-th block at level j, i.e.

$$B_j^i = \{(j,k) : (i-1)L + 1 \le k \le iL\}.$$

Let $S_{j,i}^2 \equiv \sum_{(j,k) \in B_j^i} u_{j,k}^2$ denote the sum of squared empirical wavelet coefficients in the block B_j^i. The James-Stein shrinkage rule is then applied to each block B_j^i. For $(j,k) \in B_j^i$,

$$\hat{\theta}_{j,k} = (1 - \frac{\lambda_* L}{4nS_{j,i}^2})_+ u_{j,k} \tag{10}$$

where, as in Section 4, $\lambda_* = 4.50524$ is the solution to the equation $\lambda_* - \log \lambda_* = 3$ and $4n$ in the shrinkage factor of (10) is due to the fact that the noise level in (9) is $\sigma = \frac{1}{2\sqrt{n}}$. The block size $L = \log n$ and the threshold $\lambda_* = 4.50524$ are selected according to a block thresholding oracle inequality and a minimax criterion. See Cai (1999) for further details.

For the gross structure terms at the lowest resolution level j_0, we set $\hat{\hat{\theta}}_{j_0,k} = \tilde{u}_{j_0,k}$. The estimate of $\sqrt{f}$ at the equi-spaced sample points $\{\frac{i}{T} : i = 1, \cdots, T\}$ is then obtained by applying the inverse discrete wavelet transform (IDWT) to the denoised wavelet coefficients. That is, $\{\sqrt{f}(\frac{i}{T}) : i = 1, \cdots, T\}$ is estimated by $\widehat{\sqrt{f}} = \{\widehat{\sqrt{f}}(\frac{i}{T}) : i = 1, \cdots, T\}$ with $\widehat{\sqrt{f}} = T^{\frac{1}{2}} W^{-1} \cdot \hat{\theta}$. The estimate of the whole function $\sqrt{f}$ is given by

$$\widehat{\sqrt{f}}(t) = \sum_{k=1}^{2^{j_0}} \hat{\hat{\theta}}_{j_0,k} \phi_{j_0,k}(t) + \sum_{j=j_0}^{J-1} \sum_{k=1}^{2^j} \hat{\theta}_{j,k} \psi_{j,k}(t) \tag{11}$$

and the estimator of the density function f is given by the square of $\widehat{\sqrt{f}}$:

$$\widehat{f}(t) = \left(\sum_{k=1}^{2^{j_0}} \hat{\hat{\theta}}_{j_0,k} \phi_{j_0,k}(t) + \sum_{j=j_0}^{J-1} \sum_{k=1}^{2^j} \hat{\theta}_{j,k} \psi_{j,k}(t) \right)^2 . \tag{12}$$

By normalizing $\widehat{f}$ we obtain the final density estimator $\widetilde{f}$ where

$$\widetilde{f}(t) = \widehat{f}(t) / \int_0^1 \widehat{f}(t) dt. \tag{13}$$

This density estimation procedure is easily implementable and possesses desirable properties.

5 Theoretical properties

We turn in this section to the theoretical properties of the root-unroot BlockJS density estimators introduced in Sections 3 and 4. The asymptotic results show that the procedure enjoys a high degree of adaptivity and spatial adaptivity. Specifically, we consider adaptivity of the estimator over a wide range of Besov spaces under both the squared Hellinger distance loss $l_H(g, f) = \|\sqrt{g} - \sqrt{f}\|_2^2$ and the usual integrated squared error $l_2(g, f) = \|g - f\|_2^2$. We also consider spatial adaptivity under pointwise squared error.

Besov spaces contain a number of traditional smoothness spaces such as Hölder and Sobolev spaces as special cases and arise naturally in many fields of analysis. See Donoho and Johnstone (1998) for a discussion on the relevance of Besov spaces to scientific problems. A Besov space $B^\alpha_{p,q}$ has three parameters: α measures degree of smoothness, p and q specify the type of norm used to measure the smoothness. Besov spaces can be defined in several ways. For the present paper, we will use the Besov sequence norm based on the wavelet coefficients. Let $(\phi,\ \psi)$ be a pair of compactly supported father and mother wavelets. A mother 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 a function g is then defined by

$$\|g\|_{b^{\alpha}_{p,q}} = \|\xi_{j_0,k}\|_{\ell^p} + \left(\sum_{j=j_0}^{\infty}\left(2^{js}\left(\sum_k |\theta_{j,k}|^p\right)^{1/p}\right)^q\right)^{1/q} \tag{14}$$

where $s = \alpha + 1/2 - 1/p$, $\xi_{jk} = \int_0^1 g(t)\phi_{jk}(t)\,dt$ and $\theta_{jk} = \int_0^1 g(t)\psi_{jk}(t)\,dt$. The standard modification applies for the cases $p, q = \infty$. See Triebel (1992) and Meyer (1992) for further details on Besov spaces. We define $B^{\alpha}_{p,q}(M) = \left\{f; \|f\|_{b^{\alpha}_{p,q}} \le M\right\}$.

In the present paper we consider the risk of estimating the density function f over Besov balls,

$$F^{\alpha}_{p,q}(M,\epsilon) = \left\{f : f \in B^{\alpha}_{p,q}(M),\ \int_0^1 f(x)dx = 1,\ f(x) \ge \epsilon \text{ for all } x \in [0,1]\right\}.$$

Note that when f is bounded below from 0 and above from a constant, the condition $f \in B^{\alpha}_{p,q}(M)$ is equivalent to that there exists $M' > 0$ such that $\sqrt{f} \in B^{\alpha}_{p,q}(M')$. See Runst (1986).

Remark 2 The assumption $f(x) \ge \epsilon$ implies that the number of observations Q_i in each bin is large so that Y_i defined in equation (5) can be treated as if it were a normal random variable. See Lemmas 2 and 3 for more details. This assumption can be relaxed. For instance, the main results in this paper can be extended to the case that the density f is 0 at a fixed number of points so long as f' is not 0 at those points.

The two density estimators, $\hat{f}$ in (12) and the normalized version $\tilde{f}$ in (13), share the same asymptotic properties. To save space we shall use f_* to denote either $\hat{f}$ or $\tilde{f}$ in the theoretical results given below. The following results show that the estimators enjoy a high degree of adaptivity under the squared Hellinger distance loss.

Theorem 1 *Let $x_1, x_2, \ldots, x_n$ be a random sample from a distribution with density function f. Suppose the wavelet ψ is r-regular. Let f_* be either $\hat{f}$ given in (12) or $\tilde{f}$ given in (13) with $m = Cn^{\frac{1}{4}}$. Then for $p \ge 2$, $\alpha \le r$ and $\frac{2\alpha^2-\alpha/3}{1+2\alpha} - \frac{1}{p} > 0$*

$$\sup_{f \in F^{\alpha}_{p,q}(M,\epsilon)} E\|\sqrt{f_*} - \sqrt{f}\|_2^2 \le Cn^{-\frac{2\alpha}{1+2\alpha}}, \tag{15}$$

and for $1 \le p < 2$, $\alpha \le r$ *and* $\frac{2\alpha^2-\alpha/3}{1+2\alpha} - \frac{1}{p} > 0$

$$\sup_{f \in F^{\alpha}_{p,q}(M,\epsilon)} E\|\sqrt{f_*} - \sqrt{f}\|_2^2 \le Cn^{-\frac{2\alpha}{1+2\alpha}} (\log n)^{\frac{2-p}{p(1+2\alpha)}}. \tag{16}$$

Remark 3 Note that the two density estimators depend on the number of bins T or equivalently the bin size m. Lemma 1 implies that the mean-matching variance stabilization transformation leads to a bias term of order $m^{-3/2}$. The cumulative contribution of the biases in the mean squared error of all T bins is then at a level of $\frac{1}{n}T\left(m^{-3/2}\right)^2 = m^{-4}$. To make this term negligible for all α, we set $m^{-4} = O(n^{-1})$, i.e., $m = Cn^{1/4}$, or equivalently $T = C^{-1}n^{3/4}$. In practice, we define $T = 2^{\lceil \log_2 n^{3/4} \rceil}$, where $\lceil \mathrm{a} \rceil$ *denotes the smallest integer greater than or equal to a, and consequently the average bin size m is* $n/T = n2^{-\lceil \log_2 n^{3/4} \rceil}$.

Theorem 1 together with the lower bound given in Theorem 2 below show that the estimators $\hat{f}$ and $\tilde{f}$ are adaptively minimax rate optimal over Besov balls with $p \ge 2$ for a large range of α, and at the same time is within a logarithmic factor of the minimax risk over Besov balls with $1 \le p < 2$ for a range of α.

Theorem 1 states the adaptivity results in the squared Hellinger distance error. Same results hold for the conventional integrated squared error.

Corollary 1 *Under the conditions of Theorem 1,*

$$\sup_{f \in F^{\alpha}_{p,q}(M,\epsilon)} E\|f_* - f\|_2^2 \le \begin{cases} Cn^{-\frac{2\alpha}{1+2\alpha}} & p \ge 2 \text{ and } \frac{2\alpha^2-\alpha/3}{1+2\alpha} - \frac{1}{p} > 0 \\ Cn^{-\frac{2\alpha}{1+2\alpha}} (\log n)^{\frac{2-p}{p(1+2\alpha)}} & 1 \le p < 2 \text{ and } \frac{2\alpha^2-\alpha/3}{1+2\alpha} - \frac{1}{p} > 0. \end{cases} \tag{17}$$

The following theorem gives lower bound for the minimax risk under the squared Hellinger distance loss.

Theorem 2 *Let* $x_1, x_2, \ldots, x_n$ *be a random sample from a distribution with the density function* f. *Then for* $p \ge 1$ *and* $\alpha + \frac{1}{2} - \frac{1}{p} > 0$ *there exists constants* c_1, $c_2 > 0$ *such that*

$$\inf_{\hat{f}} \sup_{f \in F^{\alpha}_{p,q}(M,\epsilon)} E\|\sqrt{\hat{f}} - \sqrt{f}\|_2^2 \ge c_1 n^{-\frac{2\alpha}{1+2\alpha}} \quad \text{and} \quad \inf_{\hat{f}} \sup_{f \in F^{\alpha}_{p,q}(M,\epsilon)} E\|\widehat{f} - f\|_2^2 \ge c_2 n^{-\frac{2\alpha}{1+2\alpha}}.$$

Theorem 2 follows from similar arguments used in the proof of Theorem 2 of Donoho, et al. (1996).

The upper bounds and the lower bounds given above together show that the density estimator enjoys a high degree of adaptivity over a wide range of the Besov classes under both the squared Hellinger distance loss and the integrated squared error. However, for functions of spatial inhomogeneity, the local smoothness of the functions varies significantly from point to point and global risk given in Theorem 1 cannot wholly reflect the performance of estimators at a point. We thus consider spatial adaptivity as measured by the local risk

$$R(\widehat{f}(t_0),\ f(t_0)) = E(\widehat{f}(t_0) - f(t_0))^2 \tag{18}$$

where $t_0 \in (0,1)$ is any given point. 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 $\alpha' = \alpha - \lfloor\alpha\rfloor$. In Gaussian nonparametric regression setting, it is well known that for local estimation, 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 (1990) and Brown and Low (1996) 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 density estimator automatically attains the local adaptive minimax rate for estimation at a point, without prior knowledge of the smoothness of the underlying functions.

Theorem 3 *Suppose the wavelet ψ is r-regular with $r \geq \alpha > 1/6$. Let $t_0 \in (0,1)$ be fixed. Then the estimator $\hat{f}_n$ defined in (12) satisfies*

$$\sup_{f \in \Lambda^\alpha(M, t_0, \delta)} E(\widehat{f_n}(t_0) - f(t_0))^2 \leq C \cdot (\frac{\log n}{n})^{\frac{2\alpha}{1+2\alpha}}. \tag{19}$$

5.1 A brief outline for the proof of Theorem 1

The proof of Theorem 1 is somewhat involved. There are three key steps in the proof.

The first step is Poissonization. It is shown that the fixed sample size density problem is not essentially different from the density problem where the sample size is a Poisson random variable. This step enables us to treat the counts on subintervals as independent Poisson variables as discussed briefly in Section 3.

The second step is the use of the quantile coupling inequality of Komlós, Major and Tusnády (1975) to approximate the binned and root transformed data by independent normal variables. In this step we shall give tight bounds for both the deterministic approximation errors $\epsilon_{j,k}$ and the stochastic errors $\xi_{j,k}$ in the decomposition of the empirical wavelet coefficients given in (8).

The third step is the derivation of a risk bound for block thresholding in the case where the noise is not necessarily Gaussian. This risk bound is useful in turning the analysis of the density estimator into the bias-variance trade-off calculation which is often used in more standard Gaussian nonparametric regression.

6 Numerical implementation and examples

The root-unroot approach is easy to implement if the nonparametric regression procedure in Step 3 is computationally efficient. In Section 4 we discuss in detail a wavelet block thresholding implementation of the root-unroot approach which is fast to compute. We implement the procedure in Splus. The following plot illustrates the steps in the root-unroot BlockJS procedure. A random sample is generated from a distribution with a multi-modal density function. The histogram of the data is given in the upper left panel and the binned and root transformed data is plotted in the upper right panel. The empirical wavelet coefficients and the BlockJS denoised coefficients are plotted respectively in the middle left and right panels. The estimate of the square root of the density function (solid line) is given in the lower left panel and the estimate of the density function is plotted in the lower right panel. The dotted lines in the lower panels are the true functions.

The following example taken from a practical data set illustrates the application of our root-unroot wavelet method. The data are the arrival times of calls to agents at an Israeli financial call-center. This data is a portion of that described in much more detail and analyzed from several related perspectives in Brown, et al. (2005). The values recorded are the times (in second) at which telephone calls seeking service from agents arrive to be served by the agent pool. In this example we only use calls received throughout the year on non-holiday Sundays. (Sunday is the first day of the regular work-week in Israel.) The data for current analysis contains about 55,000 call arrival times.

Features of arrival densities are of practical interest. The left panel of Figure 3 shows the histogram of this arrival time data. The right panel shows the density estimate produced by the root-unroot, wavelet block thresholding methodology. In this example we use $T = 512$, symmelet "s16" and the primary level $j_0 = 4$. Note the three modes in the density plot. These occur at roughly 10-11 am, 2-3 pm and approximately 10pm. The morning buildup and afternoon fall-off in call density is otherwise a fairly smooth curve. The dip in arrival

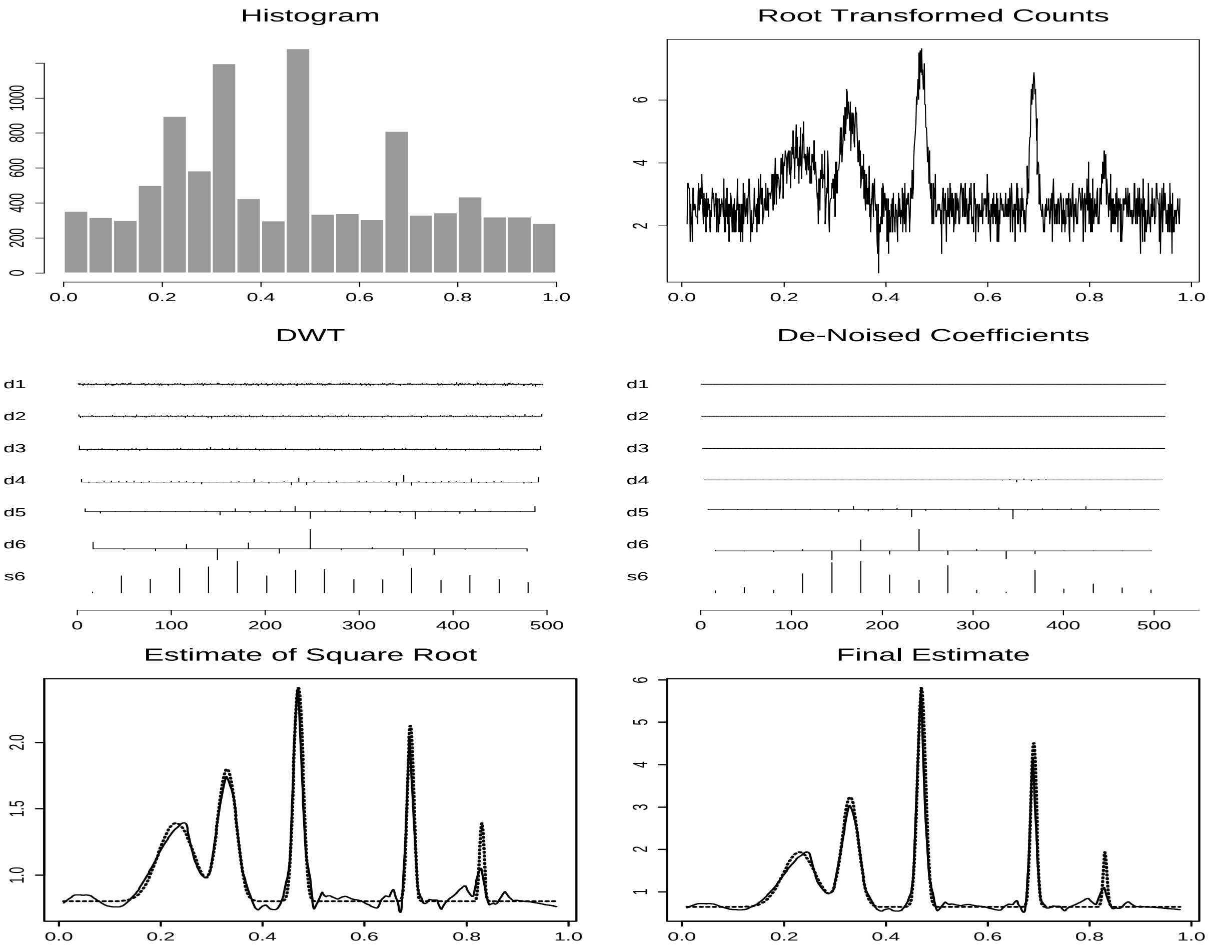

Figure 2: An example of the root-unroot BlockJS density estimator.

density between the first two modes is presumably connected to the time of lunch break, when Israelis seem less inclined to call their bank. The noticeable mode at around 10pm may be related to societal TV or bedtime habits or to phone rates in Israel, which change at 10pm. See Weinberg, Brown and Stroud (2007) for a more sophisticated analysis of a similar set of data from an American financial call center.

7 Proofs

We shall only give a complete proof for Theorem 1. Theorem 2 can be proved by using similar arguments given in the proof of Theorem 2 of Donoho, et al. (1996) and the proof of Theorem 3 is similar to that of Theorem 4 of Brown, Cai and Zhou (2008).

As outlined in Section 5.1, the proof of Theorem 1 contains three key steps: Poissoniza-

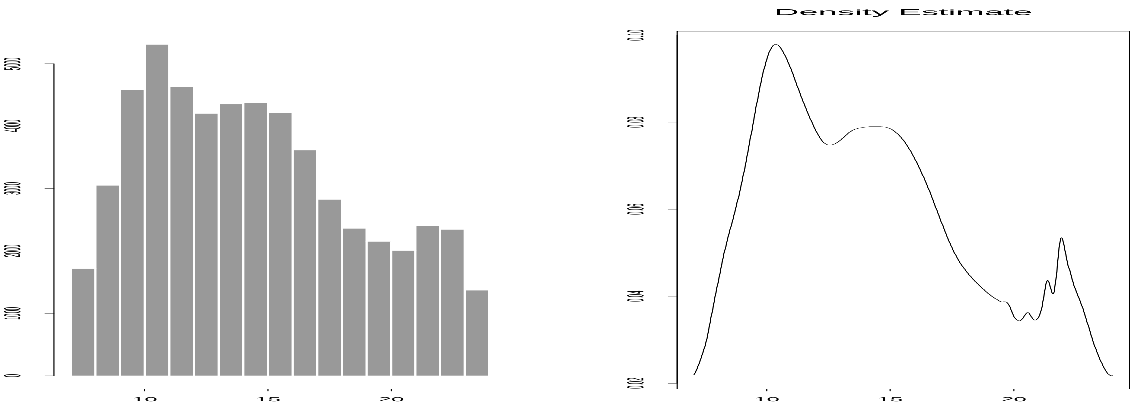

Figure 3: The histogram (left panel) and the density estimate (right panel)of the call center data.

tion, coupling, and bounding the risk of block thresholding estimators. We shall proceed according to these three steps. We first prove a Poissonized version of Theorem 1. The proof for squared Hellinger distance loss is given in Section 7.4 and the proof for integrated squared error is given in Section 7.5. Section 7.6 shows that the normalized estimator $\tilde{f}$ shares the same properties as the estimator $\hat{f}$. Finally we complete the proof of Theorem 1 in Section 7.7 by showing that the Poissonized version of Theorem 1 yields the corresponding results for density estimation.

7.1 Poissonized density estimation

We begin by introducing a Poissonized version of the density estimation problem. Let $N \sim \text{Poisson}(n)$ and let $x_1, x_2, \ldots, x_N$ be a random sample from a distribution with density function f. Suppose that x_i's and N are independent and that the density f is supported on the unit interval $[0, 1]$. Let Q_i be the number of observations on the interval $[(i-1)/T, i/T)$, $i = 1, 2, \ldots T$. Set $m = n/T$. Then $Q_i \sim \text{Poisson}(mp_i)$ where $p_i = T \int_{\frac{i-1}{T}}^{\frac{i}{T}} f(x)dx$. Set

$$Y_i = \sqrt{Q_i + \frac{1}{4}}, \quad i = 1, 2, \ldots, T \tag{20}$$

and let $\widehat{f}$ be given as in (12) and $\tilde{f}$ given in (13). We shall first prove a Poissonized version of Theorem 1 which shows that $\widehat{f}$ has the same rate of convergence when the sample size is a Poisson variable as when the sample size is fixed.

Theorem 4 *Let $x_1, x_2, \ldots, x_N \overset{i.i.d.}{\sim} f$, $N \sim Poisson(n)$. Suppose the wavelet ψ is r-regular* *and $\alpha \le r$. Let f_* be either $\hat f$ given in (12) or $\tilde f$ given in (13) with $m = Cn^{\frac{1}{4}}$. Then*

$$\sup_{f \in F_{p,q}^{\alpha}(M,\epsilon)} E\|\widehat{\sqrt{f_*}} - \sqrt{f}\|_2^2 \le \begin{cases} Cn^{-\frac{2\alpha}{1+2\alpha}} & p \ge 2\ ,\ \frac{2\alpha^2-\alpha/3}{1+2\alpha} - \frac{1}{p} > 0 \\ Cn^{-\frac{2\alpha}{1+2\alpha}}(\log n)^{\frac{2-p}{p(1+2\alpha)}} & 1 \le p < 2,\ \frac{2\alpha^2-\alpha/3}{1+2\alpha} - \frac{1}{p} > 0. \end{cases}$$

Corollary 2 *Under the conditions of Theorem 4,*

$$\sup_{f \in F_{p,q}^{\alpha}(M,\epsilon)} E\|f_* - f\|_2^2 \le \begin{cases} Cn^{-\frac{2\alpha}{1+2\alpha}} & p \ge 2 \text{ and } \frac{2\alpha^2-\alpha/3}{1+2\alpha} - \frac{1}{p} > 0 \\ Cn^{-\frac{2\alpha}{1+2\alpha}}(\log n)^{\frac{2-p}{p(1+2\alpha)}} & 1 \le p < 2 \text{ and } \frac{2\alpha^2-\alpha/3}{1+2\alpha} - \frac{1}{p} > 0. \end{cases}$$

We should note that the result in Theorem 4 holds in more general setting where one is interested in estimating the intensity function of an inhomogeneous Poisson process. A similar result has been used for this purpose in Zhang (2002) and Brown, et al. (2005).

The proof of Theorem 4 requires analysis of both the deterministic part (the mean) and the stochastic part of Y_i given in (20). We shall first collect in the next section technical lemmas that are needed for the proof of Theorem 4. In section 7.6, we show the risk difference between $\hat f$ and $\tilde f$ is negligible.

7.2 Coupling and preparatory results

We shall use the quantile coupling inequality of Komlós, Major and Tusnády (1975) to approximate the binned and root transformed data by independent normal variables. The following lemma is a direct consequence of the results given in Komlós, Major and Tusnády (1975) and Zhou (2006).

Lemma 2 *Let $\lambda > 0$ and let $X \sim \text{Poisson}(\lambda)$. There exists a standard normal random variable $Z \sim N(0,1)$ and constants $c_1, c_2, c_3 > 0$ not depending on λ such that whenever the event $A = \{|X - \lambda| \le c_1\lambda\}$ occurs,*

$$|X - \lambda - \sqrt{\lambda} Z| < c_2 Z^2 + c_3. \tag{21}$$

We shall develop tight bounds for both the deterministic approximation errors $\epsilon_{j,k}$ and the stochastic errors $\xi_{j,k}$ in the decomposition of the empirical wavelet coefficients given in (8). Let $X \sim \text{Poisson}(\lambda)$ and let $Y = \sqrt{X + \frac{1}{4}}$ and $\epsilon = EY - \sqrt{\lambda}$. Let Z be a standard normal variable satisfying (21). Then Y can be written as $Y = \sqrt{\lambda} + \epsilon + \frac{1}{2}Z + \xi$ where

$$\xi = \frac{X - \lambda}{\sqrt{X + \frac{1}{4}} + \sqrt{\lambda + \frac{1}{4}}} - \frac{1}{2}Z - E\left(\frac{X - \lambda}{\sqrt{X + \frac{1}{4}} + \sqrt{\lambda + \frac{1}{4}}}\right). \tag{22}$$

It follows from Lemma 1 that when λ is large, ϵ is "small", $|\epsilon| \le \frac{1}{64}\lambda^{-\frac{3}{2}}(1 + o(1))$. We shall show, using Lemma 2, that the random variable ξ is "stochastically small".

Lemma 3 *Let $X \sim$ Poisson(λ) and let the standard normal variable Z be given as in Lemma 2. Let ξ be given as in (22). Then for any integer $i \geq 1$ there exists a constant $C_i > 0$ such that for all $\lambda \geq 1$ and all $a > 0$,*

$$E|\xi|^i \leq C_i \lambda^{-\frac{i}{2}} \quad and \quad P(|\xi| > a) \leq C_i (a^2\lambda)^{-\frac{i}{2}}. \tag{23}$$

<u>*Proof:*</u> First note that $E\xi = 0$. Set $\delta = E(\frac{X-\lambda}{\sqrt{X+\frac{1}{4}}+\sqrt{\lambda+\frac{1}{4}}})$. Then

$$\xi \;=\; \frac{X-\lambda}{\sqrt{X+\frac{1}{4}}+\sqrt{\lambda+\frac{1}{4}}} - \frac{1}{2}Z - \delta = \xi_1 + \xi_2 + \xi_3$$

where

$$\xi_1 \;=\; -\frac{(X-\lambda)^2}{2\left(\sqrt{X+\frac{1}{4}}+\sqrt{\lambda+\frac{1}{4}}\right)^2\sqrt{\lambda+\frac{1}{4}}} - \delta \tag{24}$$

$$\xi_2 \;=\; \frac{X-\lambda-\sqrt{\lambda}Z}{2\sqrt{\lambda+\frac{1}{4}}} \tag{25}$$

$$\xi_3 \;=\; -\frac{1}{8\lambda\sqrt{1+\frac{1}{4\lambda}}(1+\sqrt{1+\frac{1}{4\lambda}})}Z. \tag{26}$$

Note that $E\xi_l = 0$, $l = 1, 2, 3$ and $|\xi_2| \leq \lambda^{-\frac{1}{2}}(C_2 Z^2 + C_3)$ and $|\xi_3| \leq \frac{1}{16}\lambda^{-1}|Z|$ on $A = \{|X-\lambda| \leq c_1\lambda\}$ with $P(A^c) \leq \exp(-c\lambda)$ for some $c > 0$. Hence for any integer $i \geq 1$ the Cauchy-Schwarz inequality implies, for some constant $d_i > 0$,

$$E|\xi_2|^i \leq d_i\lambda^{-\frac{i}{2}} \quad \text{and} \quad E|\xi_3|^i \leq d_i\lambda^{-\frac{i}{2}}. \tag{27}$$

Note also that (24) yields $\delta = E(\frac{X-\lambda}{\sqrt{X+\frac{1}{4}}+\sqrt{\lambda+\frac{1}{4}}}) = -E\left(\frac{(X-\lambda)^2}{2\left(\sqrt{X+\frac{1}{4}}+\sqrt{\lambda+\frac{1}{4}}\right)^2\sqrt{\lambda+\frac{1}{4}}}\right)$. Hence $|\delta| \leq \frac{E(X-\lambda)^2}{2\lambda^{\frac{3}{2}}} = \frac{1}{2}\lambda^{-\frac{1}{2}}$. On the other hand, it follows directly from Lemma 4 below that for any integer $i \geq 1$ there exists a constant $c_i > 0$ such that $E(X-\lambda)^{2i} \leq c_i\lambda^i$. Note that for $i \geq 1$, $(a+b)^i \leq 2^{i-1}(|a|^i + |b|^i)$. It then follows that

$$E|\xi_1|^i \leq 2^{i-1}\left[\frac{E(X-\lambda)^{2i}}{2^i\lambda^{\frac{3i}{2}}} + |\delta|^i\right] \leq 2^{i-1}(\frac{c_i\lambda^i}{2^i\lambda^{\frac{3i}{2}}} + 2^{-i}\lambda^{-\frac{i}{2}}) = (\frac{1}{2}c_i + \frac{1}{2})\lambda^{-\frac{i}{2}}. \tag{28}$$

The first bound in (23) now follows by combining (27) and (28). The second bound in (23) is a direct consequence of the first one and Markov inequality. ▮

Lemmas 1, 2 and 3 together yield the following result.

Proposition 1 *Let $Y_i = \sqrt{Q_i + \frac{1}{4}}$ be given as in (20). Then Y_i can be written as*

$$Y_i = \sqrt{mp_i} + \epsilon_i + \frac{1}{2} Z_i + \xi_i, \quad i = 1, 2, \ldots, T, \tag{29}$$

where $Z_i \overset{i.i.d.}{\sim} N(0,1)$, ϵ_i are constants satisfying $|\epsilon_i| \le \frac{1}{64}(mp_i)^{-\frac{3}{2}}(1+o(1))$ and consequently for some constant $C > 0$

$$\frac{1}{n}\sum_{i=1}^{T} \epsilon_i^2 \le C \cdot m^{-4}, \tag{30}$$

and ξ_i are independent and "stochastically small" random variables satisfying

$$E|\xi_i|^l \le C_l (mp_i)^{-\frac{l}{2}} \quad \text{and} \quad P(|\xi_i| > a) \le C_l (a^2 mp_i)^{-\frac{l}{2}} \tag{31}$$

where $l > 0$, $a > 0$ and $C_l > 0$ is a constant depending on l only.

We need the following moment bounds for an orthogonal transform of independent variables.

Lemma 4 *Let $X_1, \ldots, X_n$ be independent variables with $E(X_i) = 0$ for $i = 1, \ldots, n$. Suppose that $E|X_i|^k < M_k$ for all i and all $k > 0$ with $M_k > 0$ some constant not depending on n. Let $Y = WX$ be an orthogonal transform of $X = (X_1, ..., X_n)'$. Then there exist constants M_k' not depending on n such that $E|Y_i|^k < M_k'$ for all $i = 1, \ldots, n$ and all $k > 0$.*

Proof: Let a_i, $i = 1, \cdots, n$ be constants such that $\sum_{i=1}^n a_i^2 = 1$. It suffices to show that for $U = \sum_{i=1}^n a_i X_i$ there exist constants M_k' not depending on n and $a = (a_1, \ldots, a_n)$ such that $E|U|^k < M_k'$ for all even positive integer k.

Let k be a fixed even integer. Then, since $E(X_i) = 0$ for $i = 1, \ldots, n$,

$$\begin{aligned} E|U|^k &= E(\sum_{i=1}^n a_i X_i)^k = \sum_{k_1+\cdots+k_n=k} \binom{k}{k_1, ..., k_n} a_1^{k_1} \cdots a_n^{k_n} EX_1^{k_1} \cdots EX_n^{k_n} \\ &= \sum_{(k_1,\ldots,k_n)\in S(k)} \binom{k}{k_1, ..., k_n} a_1^{k_1} \cdots a_n^{k_n} EX_1^{k_1} \cdots EX_n^{k_n}. \end{aligned}$$

where $S(k) = \{(k_1, \ldots, k_n) : k_i \text{ nonnegative integers}, \ k_i \ne 1 \text{ and } \sum_{i=1}^n k_i = k\}$. Set $A_k = (1 + M_1)(1 + M_2) \cdots (1 + M_k)$. Then, since $|a_i| \le 1$,

$$\begin{aligned} E|U|^k &\le k! A_k \sum_{(k_1,\ldots,k_n)\in S(k)} |a_1|^{k_1} \cdots |a_n|^{k_n} \le k! A_k \sum_{(k_1,\ldots,k_n)\in S(k)} |a_1|^{2[\frac{k_1}{2}]} \cdots |a_n|^{2[\frac{k_n}{2}]} \\ &\le k! A_k \sum_{k'=1}^{k/2} \sum_{(k_1',\ldots,k_n')\in S(k')} (a_1^2)^{k_1'} \cdots (a_n^2)^{k_n'}. \end{aligned}$$

Since $\sum_{(k'_1,\ldots,k'_n)\in S(k')}(a_1^2)^{k'_1}\cdots(a_n^2)^{k'_n} \le \left(\sum_{i=1}^n a_i^2\right)^{k'} = 1$, $E|U|^k \le k! A_k \frac{k}{4}(\frac{k}{2}+1)$. The lemma is proved by taking $M'_k = k! A_k \frac{k}{4}(\frac{k}{2}+1)$. ▮

From (29) in Proposition 1 we can write $\frac{1}{\sqrt{n}}Y_i = \frac{\sqrt{p_i}}{\sqrt{T}} + \frac{\epsilon_i}{\sqrt{n}} + \frac{Z_i}{2\sqrt{n}} + \frac{\xi_i}{\sqrt{n}}$. Let $(u_{j,k}) = n^{-\frac{1}{2}}W\cdot Y$ be the discrete wavelet transform of the binned and root transformed data. Then one may write

$$u_{j,k} = \theta'_{j,k} + \epsilon_{j,k} + \frac{1}{2\sqrt{n}}z_{j,k} + \xi_{j,k} \tag{32}$$

where θ'_{jk} are the discrete wavelet transform of $(\frac{\sqrt{p_i}}{\sqrt{T}})$ which are approximately equal to the true wavelet coefficients of $\sqrt{f}$, $z_{j,k}$ 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 $(\frac{\xi_i}{\sqrt{n}})$. Then it follows from Proposition 1 that

$$\sum_j\sum_k \epsilon_{j,k}^2 = \frac{1}{n}\sum_i \epsilon_i^2 \le Cm^{-4}. \tag{33}$$

It now follows from Lemma 4 and Proposition 1 that for all $i > 0$ and $a > 0$

$$E|\xi_{j,k}|^i \le C'_i(mn)^{-\frac{i}{2}} \quad \text{and} \quad P(|\xi_{j,k}| > a) \le C'_i(a^2mn)^{-\frac{i}{2}}. \tag{34}$$

7.3 Risk bound for a single block

Oracle inequalities for block thresholding estimators were derived in Cai (1999) in the case when the noise is i.i.d. normal. In the present paper we need the following risk bound for block thresholding estimators without the normality assumption.

Lemma 5 *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*

$$E\|\hat\theta - \theta\|_2^2 \le \|\theta\|_2^2 \wedge 4\lambda L + 4E\left[\|z\|_2^2 I(\|z\|_2^2 > \lambda L)\right]. \tag{35}$$

<u>*Proof:*</u> It is easy to verify that $\|\hat\theta - y\|_2^2 \le \lambda L$. Hence

$$\begin{aligned} E\left[\|\hat\theta-\theta\|_2^2 I(\|z\|_2^2 > \lambda L)\right] &\le 2E\left[\|\hat\theta - y\|_2^2 I(\|z\|_2^2 > \lambda L)\right] + 2E\left[\|y-\theta\|_2^2 I(\|z\|_2^2 > \lambda L)\right] \\ &\le 2\lambda L P(\|z\|_2^2 > \lambda L) + 2E\left[\|z\|_2^2 I(\|z\|_2^2 > \lambda L)\right] \\ &\le 4E\left[\|z\|_2^2 I(\|z\|_2^2 > \lambda L)\right]. \end{aligned} \tag{36}$$

On the other hand,

$$E\left[\|\hat\theta-\theta\|_2^2 I(\|z\|_2^2 \le \lambda L)\right] \le E\left[(2\|\hat\theta - y\|_2^2 + 2\|y-\theta\|_2^2) I(\|z\|_2^2 \le \lambda L)\right] \le 4\lambda L. \tag{37}$$

Note that when $S^2 \le \lambda L$, $\hat{\theta} = 0$ and hence $\|\hat{\theta} - \theta\|_2^2 = \|\theta\|_2^2$. When $\|z\|_2^2 \le \lambda L$ and $S^2 > \lambda L$,

$$\begin{aligned}
\|\hat{\theta} - \theta\|_2^2 &= \sum_i [(1 - \frac{\lambda L}{S^2}) y_i - \theta_i]^2 = (1 - \frac{\lambda L}{S^2})[S^2 - \lambda L - 2\sum_i \theta_i y_i] + \|\theta\|_2^2 \\
&= (1 - \frac{\lambda L}{S^2})[\sum (\theta_i + z_i)^2 - \lambda L - 2\sum_i \theta_i(\theta_i + z_i)] + \|\theta\|_2^2 \\
&= (1 - \frac{\lambda L}{S^2})(\|z\|_2^2 - \lambda L - \|\theta\|_2^2) + \|\theta\|_2^2 \le \|\theta\|_2^2.
\end{aligned}$$

Hence $E\left[\|\hat{\theta} - \theta\|_2^2 I(\|z\|_2^2 \le \lambda L)\right] \le \|\theta\|_2^2$ and (35) follows by combining this with (36) and (37). ▮

We also need the following bound on the tail probability of a central chi-square distribution (see Cai (2002)).

Lemma 6 *Let $X \sim \chi_L^2$ and $\lambda > 1$. Then*

$$P(X \ge \lambda L) \le e^{-\frac{L}{2}(\lambda - \log\lambda - 1)} \quad \text{and} \quad EXI(X \ge \lambda L) \le \lambda L e^{-\frac{L}{2}(\lambda - \log\lambda - 1)}. \tag{38}$$

Proposition 2 *Let the empirical wavelet coefficients $u_{j,k} = \theta'_{j,k} + \epsilon_{j,k} + \frac{1}{2\sqrt{n}} z_{j,k} + \xi_{j,k}$ be given as in (32) and let the block thresholding estimator $\hat{\theta}_{j,k}$ be defined as in (10). Then for some constant $C > 0$*

$$E \sum_{(j,k)\in B_j^i} (\hat{\theta}_{j,k} - \theta'_{j,k})^2 \le \min\left\{4 \sum_{(j,k)\in B_j^i} (\theta'_{j,k})^2,\ 8\lambda_* L n^{-1}\right\} + 6 \sum_{(j,k)\in B_j^i} \epsilon_{j,k}^2 + CLn^{-2}. \tag{39}$$

<u>*Proof:*</u> It follows from Lemma 5 that

$$\begin{aligned}
E \sum_{(j,k)\in B_j^i} (\hat{\theta}_{j,k} - \theta'_{j,k})^2 &\le 2E \sum_{(j,k)\in B_j^i} [\hat{\theta}_{j,k} - (\theta'_{j,k} + \epsilon_{j,k})]^2 + 2 \sum_{(j,k)\in B_j^i} \epsilon_{j,k}^2 \\
&\le \min\left\{4 \sum_{(j,k)\in B_j^i} (\theta'_{j,k})^2,\ 8\lambda_* L n^{-1}\right\} + 6 \sum_{(j,k)\in B_j^i} \epsilon_{j,k}^2 \\
&\quad + 2n^{-1} E \sum_{(j,k)\in B_j^i} (z_{j,k} + 2\sqrt{n}\xi_{j,k})^2 I\left(\sum_{(j,k)\in B_j^i} (z_{j,k} + 2\sqrt{n}\xi_{j,k})^2 > \lambda_* L\right).
\end{aligned}$$

Define the event A by $A = \{|2\sqrt{n}\xi_{j,k}| \le L^{-1} \quad \text{for all } (j,k) \in B_j^i\}$. Then it follows from (34) that for any $i \ge 1$

$$P(A^c) \le \sum_{(j,k)\in B_j^i} P(|2\sqrt{n}\xi_{j,k}| > L^{-1}) \le C_i'(L^{-2}m)^{-\frac{i}{2}}. \tag{40}$$

Note that

$$\begin{aligned} D &= E\sum_{(j,k)\in B_j^i}(z_{j,k}+2\sqrt{n}\xi_{j,k})^2 I\left(\sum_{(j,k)\in B_j^i}(z_{j,k}+2\sqrt{n}\xi_{j,k})^2>\lambda_* L\right)\\ &= E\sum_{(j,k)\in B_j^i}(z_{j,k}+2\sqrt{n}\xi_{j,k})^2 I\left(A\cap\sum_{(j,k)\in B_j^i}(z_{j,k}+2\sqrt{n}\xi_{j,k})^2>\lambda_* L\right)\\ &\quad+E\sum_{(j,k)\in B_j^i}(z_{j,k}+2\sqrt{n}\xi_{j,k})^2 I\left(A^c\cap\sum_{(j,k)\in B_j^i}(z_{j,k}+2\sqrt{n}\xi_{j,k})^2>\lambda_* L\right)\\ &\equiv D_1+D_2. \end{aligned}$$

Note that for any $L>1$, $(x+y)^2\le\frac{L}{L-1}x^2+Ly^2$ for all x and y. It then follows from Lemma 6 and Hölder's Inequality that

$$\begin{aligned} D_1 &= E\sum_{(j,k)\in B_j^i}(z_{j,k}+2\sqrt{n}\xi_{j,k})^2 I\left(A\cap\sum_{(j,k)\in B_j^i}(z_{j,k}+2\sqrt{n}\xi_{j,k})^2>\lambda_* L\right)\\ &\le 2E\sum_{(j,k)\in B_j^i}z_{j,k}^2 I\left(\sum_{(j,k)\in B_j^i}z_{j,k}^2>\lambda_* L-\lambda_*-1\right)\\ &\quad+8nE\sum_{(j,k)\in B_j^i}\xi_{j,k}^2 I\left(\sum_{(j,k)\in B_j^i}z_{j,k}^2>\lambda_* L-\lambda_*-1\right)\\ &\le 2(\lambda_* L-\lambda_*-1)e^{-\frac{L}{2}(\lambda_*-(\lambda_*+1)L^{-1}-\log(\lambda_*-(\lambda_*+1)L^{-1})-1)}\\ &\quad+8n\sum_{(j,k)\in B_j^i}(E\xi_{j,k}^{2r})^{\frac{1}{r}}\left(P(\sum_{(j,k)\in B_j^i}z_{j,k}^2>\lambda_* L-\lambda_*-1)\right)^{\frac{1}{w}} \end{aligned}$$

where $r,w>1$ and $\frac{1}{r}+\frac{1}{w}=1$. For $m=n^\epsilon$ we take $\frac{1}{w}=1-\epsilon$. Then it follows from Lemma 6 and (34) that

$$D_1\le\lambda_* e^{\frac{\lambda_*+1}{2}}Ln^{-1}+CLm^{-1}n^{-1-\epsilon}=CLn^{-1}.$$

On the other hand, it follows from (34) and (40) (by taking $i = 10$) that

$$\begin{aligned} D_2 &= E \sum_{(j,k)\in B_j^i} (z_{j,k} + 2\sqrt{n}\xi_{j,k})^2 I\left(A^c \cap \sum_{(j,k)\in B_j^i} (z_{j,k} + 2\sqrt{n}\xi_{j,k})^2 > \lambda_* L\right) \\ &\le E \sum_{(j,k)\in B_j^i} (2z_{j,k}^2 + 8n\xi_{j,k}^2) I(A^c) \le \sum_{(j,k)\in B_j^i} [2(Ez_{j,k}^4)^{\frac{1}{2}} + 8n(E\xi_{j,k}^4)^{\frac{1}{2}}] \cdot (P(A^c))^{\frac{1}{2}} \\ &\le CL(L^{-2}m)^{-5} \le n^{-1}. \end{aligned}$$

Hence, $D = D_1 + D_2 \le CLn^{-1}$ and consequently, for some constant $C > 0$,

$$E \sum_{(j,k)\in B_j^i} (\hat{\theta}_{j,k} - \theta'_{j,k})^2 \;\le\; \min\left\{4 \sum_{(j,k)\in B_j^i} (\theta'_{j,k})^2,\ 8\lambda_* L n^{-1}\right\} + 6 \sum_{(j,k)\in B_j^i} \epsilon_{j,k}^2 + CLn^{-2}. \quad ∎$$

Lemma 7 *Let $T = 2^J$ and $d = \min(\alpha - \frac{1}{p}, 1)$. Set $p_i = T\int_{(i-1)/T}^{i/T} g^2(x)dx$ and $\bar{g}_J(x) = \sum_{k=1}^{T} \frac{1}{\sqrt{T}}\sqrt{p_k}\phi_{J,k}(x)$. Then for some constant $C > 0$*

$$\sup_{g\in F_{p,q}^{\alpha}(M,\epsilon)} \|\bar{g}_J - g\|_2^2 \le CT^{-2d}. \tag{41}$$

Proof: Note that it follows from embedding theorem of Besov spaces that for some constant $M' > 0$ $B_{p,q}^{\alpha}(M) \subseteq B_{\infty,\infty}^{d}(M')$. Hence for all $g \in B_{p,q}^{\alpha}(M)$ there exists a constant $C > 0$ such that $|\beta_{J,k} - \frac{1}{\sqrt{T}}g(\frac{k}{T})| \le C2^{-J(d+\frac{1}{2})}$. Let $\tilde{g}_J(x) = \sum_{k=1}^{T} \frac{1}{\sqrt{T}} g(\frac{k}{T})\phi_{J,k}(x)$. Then

$$\|\tilde{g}_J - g\|_2^2 = \sum_k (\beta_{J,k} - \frac{1}{\sqrt{T}}g(\frac{k}{T}))^2 + \sum_{j\ge J}\sum_k \theta_{J,k}^2 \le C^2 2^{-2dJ} + C2^{-2J(\alpha\wedge(\alpha+\frac{1}{2}-\frac{1}{p}))} \le CT^{-2d}.$$

Since $\epsilon \le g \le C_0$ for some $C_0 > 0$,

$$|\sqrt{p_k} - g(\frac{k}{T})| = \frac{|T\int_{(i-1)/T}^{i/T}(g^2(x) - g^2(\frac{k}{T}))dx|}{\sqrt{T\int_{(i-1)/T}^{i/T} g^2(x)dx} + g(\frac{k}{T})} \le \frac{2C_0T\int_{(i-1)/T}^{i/T}|g(x) - g(\frac{k}{T})|dx}{2\epsilon} \le CT^{-d}.$$

Hence $\|\tilde{g}_J - \bar{g}_J\|_2^2 = \frac{1}{T}\sum_k(\sqrt{p_k} - g(\frac{k}{T}))^2 \le CT^{-2d}$ and consequently

$$\sup_{g\in F_{p,q}^{\alpha}(M,\epsilon)} \|\bar{g}_J - g\|_2^2 \le \sup_{g\in F_{p,q}^{\alpha}(M,\epsilon)} (2\|\tilde{g}_J - g\|_2^2 + 2\|\tilde{g}_J - \bar{g}_J\|_2^2) \le CT^{-2d}. \quad ∎$$

7.4 Proof of Theorem 4

In this section we show the result holds for $\hat{f}$ given in (12). In Section 7.6, we will see the difference of the risk between $\hat{f}$ and $\tilde{f}$ is $o\left(n^{-2\alpha/(2\alpha+1)}\right)$ which is negligible.

Let Y and $\hat{\theta}$ be given as in (20) and (10) respectively. Then,

$$\begin{aligned} E\|\widehat{\sqrt{f}} - \sqrt{f}\|_2^2 &= \sum_k E(\hat{\tilde{\theta}}_{j_0,k} - \tilde{\theta}_{j,k})^2 + \sum_{j=j_0}^{J-1}\sum_k E(\hat{\theta}_{j,k} - \theta_{j,k})^2 + \sum_{j=J}^{\infty}\sum_k \theta_{j,k}^2 \\ &\equiv S_1 + S_2 + S_3 \end{aligned} \tag{42}$$

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

$$S_1 = 2^{j_0} n^{-1}\epsilon^2 = o(n^{-2\alpha/(1+2\alpha)}) \tag{43}$$

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

$$\|x\|_{p_2} \le \|x\|_{p_1} \le m^{\frac{1}{p_1} - \frac{1}{p_2}} \|x\|_{p_2} \tag{44}$$

Since $f \in B^{\alpha}_{p,q}(M)$, so $2^{js}(\sum_{k=1}^{2^j} |\theta_{jk}|^p)^{1/p} \le M$. Now (44) yields that

$$S_3 = \sum_{j=J}^{\infty}\sum_k \theta_{j,k}^2 \le C 2^{-2J(\alpha \wedge (\alpha + \frac{1}{2} - \frac{1}{p}))}. \tag{45}$$

Proposition 2, Lemma 7 and Equation (33) yield that

$$\begin{aligned} S_2 &\le 2\sum_{j=j_0}^{J-1}\sum_k E(\hat{\theta}_{j,k} - \theta'_{j,k})^2 + 2\sum_{j=j_0}^{J-1}\sum_k (\theta'_{j,k} - \theta_{j,k})^2 \\ &\le \sum_{j=j_0}^{J-1}\sum_{i=1}^{2^j/L} \min\left\{8 \sum_{(j,k)\in B_j^i} \theta_{j,k}^2,\ 8\lambda_* L n^{-1}\right\} + 6\sum_{j=j_0}^{J-1}\sum_k \epsilon_{j,k}^2 + Cn^{-1} + 10\sum_{j=j_0}^{J-1}\sum_k (\theta'_{j,k} - \theta_{j,k})^2 \\ &\le \sum_{j=j_0}^{J-1}\sum_{i=1}^{2^j/L} \min\left\{8 \sum_{(j,k)\in B_j^i} \theta_{j,k}^2,\ 8\lambda_* L n^{-1}\right\} + Cm^{-4} + Cn^{-1} + CT^{-2d} \end{aligned} \tag{46}$$

We now divide into two cases. First consider the case $p \ge 2$. Let $J_1 = [\frac{1}{1+2\alpha}\log_2 n]$. So, $2^{J_1} \approx n^{1/(1+2\alpha)}$. Then (46) and (44) yield

$$S_2 \le 8\lambda_* \sum_{j=j_0}^{J_1-1}\sum_{i=1}^{2^j/L} L n^{-1} + 8\sum_{j=J_1}^{J-1}\sum_k \theta_{j,k}^2 + Cn^{-1} + CT^{-2d} \le Cn^{-2\alpha/(1+2\alpha)} \tag{47}$$

By combining (47) with (43) and (45), we have $E\|\hat{\theta} - \theta\|_2^2 \le Cn^{-2\alpha/(1+2\alpha)}$, for $p \ge 2$.

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

Lemma 8 *Let $0 < p < 1$ and $S = \{x \in \mathbb{R}^k : \sum_{i=1}^k x_i^p \le B,\ x_i \ge 0,\ i = 1, \cdots, k\}$. Then $\sup_{x\in S} \sum_{i=1}^k (x_i \wedge A) \le B \cdot A^{1-p}$ 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

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

It then follows from Lemma 8 that

$$\sum_{j=J_2}^{J-1}\sum_{i=1}^{2^j/L}\min\left\{8\sum_{(j,k)\in B_j^i}\theta_{j,k}^2,\ 8\lambda_* Ln^{-1}\right\} \leq Cn^{-\frac{2\alpha}{1+2\alpha}}(\log n)^{\frac{2-p}{p(1+2\alpha)}}. \tag{48}$$

On the other hand,

$$\sum_{j=j_0}^{J_2-1}\sum_{i=1}^{2^j/L}\min\left\{8\sum_{(j,k)\in B_j^i}\theta_{j,k}^2,\ 8\lambda_* Ln^{-1}\right\} \leq \sum_{j=j_0}^{J_2-1}\sum_{b}8\lambda_* Ln^{-1} \leq Cn^{-\frac{2\alpha}{1+2\alpha}}(\log n)^{\frac{2-p}{p(1+2\alpha)}}. \tag{49}$$

Putting (43), (45), (48) and (49) together yields $E\|\hat{\theta}-\theta\|_2^2 \leq Cn^{-\frac{2\alpha}{1+2\alpha}}(\log n)^{\frac{2-p}{p(1+2\alpha)}}$. ▮

Remark 4 The condition $\frac{2\alpha^2-\alpha/3}{1+2\alpha} > \frac{1}{p}$ is purely due to approximation error over Besov spaces. To make the other terms negligible (or at least not dominant) for all α, we need to have $m^{-4} = O(n^{-\frac{2\alpha}{1+2\alpha}})$ and $T^{-2((\alpha-\frac{1}{p})\wedge 1)} = O(n^{-\frac{2\alpha}{1+2\alpha}})$. This condition puts constraints on both m and α (and p). We choose $m = n^{\frac{1}{4}}$ and so $T = n^{\frac{3}{4}}$. Then we need $\frac{3}{2}(\alpha-\frac{1}{p}) > \frac{2\alpha}{1+2\alpha}$ or equivalently $\frac{2\alpha^2-\alpha/3}{1+2\alpha} > \frac{1}{p}$. The other condition, $m \geq n^{\frac{1}{4}}$, is needed for bounding the stochastic error.

7.5 Asymptotic optimality under L_2 Loss

Proof of Corollary 2: In this section, we only give a proof for $\hat{f}$ given in (12). See the next section for $\tilde{f}$.

The L_2 loss can be related to Hellinger loss as follows

$$E\|\widehat{f}-f\|_2^2 = E\int\left(\sqrt{\widehat{f}}-\sqrt{f}\right)^2\left(\sqrt{\widehat{f}}+\sqrt{f}\right)^2 \leq 2E\int\left(\sqrt{\widehat{f}}-\sqrt{f}\right)^2\left(\widehat{f}+f\right).$$

Since f is bounded by a constant C_0, we then have

$$E\|\widehat{f}-f\|_2^2 \leq 2\left(C+C_0\right)E\int\left(\sqrt{\widehat{f}}-\sqrt{f}\right)^2 + 2E\int\left(\sqrt{\widehat{f}}-\sqrt{f}\right)^2\widehat{f}I\left(\left\|\widehat{\sqrt{f}}\right\|_\infty > C\right).$$

where the constant C will be specified later. To prove the Theorem, it suffices to show the second term is negligible for an appropriate constant C. The Cauchy-Schwarz inequality implies

$$\begin{aligned}\left[E\int\left(\sqrt{\widehat{f}}-\sqrt{f}\right)^2\widehat{f}I\left(\left\|\widehat{\sqrt{f}}\right\|_\infty>C\right)\right]^2 &\leq P\left(\left\|\widehat{\sqrt{f}}\right\|_\infty>C\right)E\int\left(\sqrt{\widehat{f}}-\sqrt{f}\right)^4\widehat{f}^2\\ &\leq 2P\left(\left\|\widehat{\sqrt{f}}\right\|_\infty>C\right)E\int\left(\widehat{f}^4+f^2\widehat{f}^2\right).\end{aligned}$$

It then suffices to show that there exists a constant C such that

$$\sup_{\sqrt{f}\in F^\alpha_{p,q}(M,\epsilon)}P\left\{\left\|\widehat{\sqrt{f}}\right\|_\infty>C\right\}\leq C_l n^{-l},$$

for any $l>1$, since it is easy to see that a crude bound for $E\int\left(\widehat{f}^4+f^2\widehat{f}^2\right)$ is Cn^4.

Recall that we can write the discrete wavelet transform of the binned data as

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

where θ'_{jk} are the discrete wavelet transform of $(\frac{\sqrt{p_i}}{\sqrt{T}})$ which are approximately equal to the true wavelet coefficients θ_{jk} of $\sqrt{f}$. Note that $\left|\theta'_{jk}-\theta_{jk}\right|=O\left(2^{-j(d+1/2)}\right)$, for $d=\min\left(\alpha-1/p,1\right)$. Note also that a Besov Ball $B^\alpha_{p,q}\left(M\right)$ can be embedded in $B^d_{\infty,\infty}\left(M_1\right)$ for some $M_1>0$. (See, e.g., Meyer (1992)). From the equation above, we have

$$\sum_{k=1}^{2^{j_0}}\widetilde{\theta}_{j_0,k}\phi_{j_0,k}(t)+\sum_{j=j_0}^{J-1}\sum_{k=1}^{2^j}\theta'_{j,k}\psi_{j,k}(t)\in B^d_{\infty,\infty}\left(M_2\right)$$

for some $M_2>0$. Applying the Block thresholding approach, we have

$$\begin{aligned}\hat{\theta}_{jk} &= (1-\frac{\lambda L\sigma^2}{S^2_{(j,i)}})_+\theta'_{j,k}+(1-\frac{\lambda L\sigma^2}{S^2_{(j,i)}})_+\epsilon_{j,k}+(1-\frac{\lambda L\sigma^2}{S^2_{(j,i)}})_+\left(\frac{1}{2\sqrt{n}}z_{j,k}+\xi_{j,k}\right)\\ &= \hat{\theta}_{1,jk}+\hat{\theta}_{2,jk}+\hat{\theta}_{3,jk}\text{ , for }(j,k)\in B^i_j,\ j_0\leq j<J.\end{aligned}$$

Note that $\left|\hat{\theta}_{1,jk}\right|\leq\left|\theta'_{j,k}\right|$ and so $\widehat{g}_1=\sum_{k=1}^{2^{j_0}}\widetilde{\theta}'_{j_0,k}\phi_{j_0,k}+\sum_{j=j_0}^{J-1}\sum_{k=1}^{2^j}\hat{\theta}_{1,j,k}\psi_{j,k}\in B^d_{\infty,\infty}\left(M_2\right)$. This implies $\widehat{g}_1$ is uniformly bounded. Note that $T^{\frac{1}{2}}\left(\sum_{j,k}\left(\epsilon^2_{j,k}\right)\right)^{1/2}=T^{\frac{1}{2}}\cdot O\left(m^{-2}\right)=o\left(1\right)$, so $W^{-1}\cdot T^{\frac{1}{2}}\left(\hat{\theta}_{2,jk}\right)$ is a uniformly bounded vector. For $0<\beta<1/6$ and a constant $a>0$ we have

$$\begin{aligned}P\left(\left|\hat{\theta}_{3,jk}\right|>a2^{-j(\beta+1/2)}\right) &\leq P\left(\left|\hat{\theta}_{3,jk}\right|>aT^{-(\beta+1/2)}\right)\\ &\leq P\left(\left|\frac{1}{2\sqrt{n}}z_{j,k}\right|>\frac{1}{2}aT^{-(\beta+1/2)}\right)+P\left(\left|\xi_{j,k}\right|>\frac{1}{2}aT^{-(\beta+1/2)}\right)\\ &\leq A_l n^{-l}\end{aligned}$$

for any $l > 1$ by Mill's ratio inequality and Inequality (31). Let $A = \underset{j,k}{\cup}\left\{\left|\hat{\theta}_{3,jk}\right| > a2^{-j(\beta+1/2)}\right\}$. Then $P\left(A\right) = C_l n^{-l}$. On the event A^c we have

$$\widehat{g}_3\left(t\right) = \sum_{j=j_0}^{J-1}\sum_{k=1}^{2^j}\hat{\theta}_{3,jk}\psi_{j,k}(t) \in B^{\beta}_{\infty,\infty}\left(M_3\right), \text{ for some } M_3 > 0$$

which is uniformly bounded. Combining these results we know that for C sufficiently large,

$$\sup_{\sqrt{f}\in F^{\alpha}_{p,q}(M,\epsilon)} P\left\{\left\|\widehat{\sqrt{f}}\right\|_{\infty} > C\right\} \le \sup_{\sqrt{f}\in F^{\alpha}_{p,q}(M,\epsilon)} P\left(A\right) = C_l n^{-l}. \quad \blacksquare \tag{50}$$

7.6 Normalization

We now show that the normalized estimator $\tilde{f}$ has the same properties as the estimator $\widehat{f}$.

Theorem 5

$$\sup_{\sqrt{f}\in F^{\alpha}_{p,q}(M,\epsilon)} E\|f-\widetilde{f}\|_2^2 \quad\le\quad (1+o\left(1\right)) \sup_{\sqrt{f}\in F^{\alpha}_{p,q}(M,\epsilon)} E\|f-\widehat{f}\|_2^2. \tag{51}$$

$$\sup_{\sqrt{f}\in F^{\alpha}_{p,q}(M,\epsilon)} E\|\sqrt{f}-\sqrt{\widetilde{f}}\|_2^2 \quad\le\quad (1+o\left(1\right)) \sup_{\sqrt{f}\in F^{\alpha}_{p,q}(M,\epsilon)} E\|\sqrt{f}-\sqrt{\widehat{f}}\|_2^2. \tag{52}$$

Proof of Theorem 5: We will only prove (51). The Cauchy-Schwarz inequality yields

$$E\|f-\widetilde{f}\|_2^2 \le \left(\sqrt{E\|f-\widehat{f}\|_2^2} + \sqrt{E\|\widehat{f}-\widetilde{f}\|_2^2}\right)^2.$$

We know $\sup_{\sqrt{f}\in F^{\alpha}_{p,q}(M,\epsilon)} E\|f-\widehat{f}\|_2^2 \ge cn^{-2\alpha/(2\alpha+1)}$ in Theorem 2. It thus suffices to show

$$\sup_{\sqrt{f}\in F^{\alpha}_{p,q}(M,\epsilon)} E\|\widehat{f}-\widetilde{f}\|_2^2 = o\left(n^{-2\alpha/(2\alpha+1)}\right).$$

We write $\int\left(\widehat{f}-\widetilde{f}\right)^2 = \left(\int\widehat{f}-1\right)^2\int\widehat{f}^2/\left(\int\widehat{f}\right)^2$, where

$$\int\widehat{f} = \sum_{k=1}^{2^{j_0}}\widetilde{u}^2_{j_0,k} + \sum_{j=j_0}^{J-1}\sum_{k=1}^{2^j}(1-\frac{\lambda Ln^{-1}}{S^2_{(j,i)}})^2_+ u^2_{j,k}$$

and $S^2_{j,i} \equiv \sum_{(j,k)\in B^i_j} u^2_{j,k}$ with $B^i_j = \{(j,k) : (i-1)L+1 \le k \le iL\}$(see Section 4.1). Let $D = \left\{\underline{x} : \int\widehat{f}^2 \ge \epsilon_1^{-1} \text{ or } \int\widehat{f} \le \epsilon_1\right\}$ where ϵ_1 will be specified later, then we have

$$\begin{aligned} E\int\left(\widehat{f}-\widetilde{f}\right)^2 &= E\left[\left(\int\widehat{f}-1\right)^2\int\widehat{f}^2/\left(\int\widehat{f}\right)^2\mathbb{I}_{D^c}\right] + E\|\widehat{f}-\widetilde{f}\|_2^2\mathbb{I}_D \\ &\le \epsilon_1^{-3}E\left(\int\widehat{f}-1\right)^2 + 2(E\|\widehat{f}-\widetilde{f}\|_2^4)^{1/2}P^{1/2}\left(D\right). \end{aligned}$$

To prove the theorem, it suffices to show

(i). $\sup_{\sqrt{f}\in F^{\alpha}_{p,q}(M,\epsilon)} E\|\widehat{f}-\widetilde{f}\|_2^4 \le Cn^b$ for a fixed $b>0$ and $\sup_{\sqrt{f}\in F^{\alpha}_{p,q}(M,\epsilon)} P\left(D\right) \le C_l n^{-l}$ for all $l>0$.

(ii). $\sup_{\sqrt{f}\in F^{\alpha}_{p,q}(M,\epsilon)} E\left(\int \widehat{f}-1\right)^2 = o\left(1\right) \sup_{\sqrt{f}\in F^{\alpha}_{p,q}(M,\epsilon)} E\|f-\widehat{f}\|_2^2.$

The first part of (i) follows from the following crude bound,

$$\int \widehat{f}^2 \le Cn^4\left[\sum_{k=1}^{2^{j_0}}\widetilde{u}_{j_0,k}^4+\sum_{j=j_0}^{J-1}\sum_{k=1}^{2^j}(1-\frac{\lambda Ln^{-1}}{S^2_{(j,i)}})_+^4 u_{j,k}^4\right] \le Cn^4\left(\int \widehat{f}\right)^2$$

which implies $\int \widetilde{f}^2 = O\left(n^4\right)$. To establish the second part of (i), it is enough to show $\sup_{\sqrt{f}\in F^{\alpha}_{p,q}(M,\epsilon)} P\left(\int \widehat{f}\le \epsilon_1\right) \le C_l n^{-l}$ for all $l>0$ since $P\left(\int \widehat{f}^2 \ge 1/\epsilon_1\right)$ decays faster than any polynomial of n^{-1} from equation (50) for ϵ_1 sufficiently small. Let

$$A=\sum_{k=1}^{2^{j_0}}\widetilde{u}_{j_0,k}^2+\sum_{j=j_0}^{\log^{1/2}n}\sum_{k=1}^{2^j}(1-\frac{\lambda Ln^{-1}}{S^2_{(j,i)}})_+^2 u_{j,k}^2,\ B=\sum_{k=1}^{2^{j_0}}\widetilde{u}_{j,k}^2+\sum_{j=j_0}^{\log^{1/2}n}\sum_{k=1}^{2^j}u_{j,k}^2$$

Note that $\int \widehat{f}\ge A$, and $-2\lambda Ln^{-1}\le \sum_{(j,k)\in B^i_j}[(1-\frac{\lambda Ln^{-1}}{S^2_{(j,i)}})_+^2 u_{j,k}^2-u_{j,k}^2]\le 0$. By the Hoeffding's inequality we have $P\left(|A-B-E\left(A-B\right)|\ge t\right)$ decays faster than any polynomial of n^{-1} for a fixed $t>0$. It is easy to see $EA=EB=1+o\left(1\right)$, i.e., A and B are both consistent estimator of $\int f$. Write

$$\begin{aligned} P\left(A\le\epsilon_1\right) &\le P\left(A-B\le\epsilon_1-1/2,\ B\ge 1/2\right)+P\left(B\le 1/2\right)\\ &\le P\left(A-B\le\epsilon_1-1/2\right)+P\left(B-EB\le 1/2-EB\right). \end{aligned}$$

Since it is obvious to see $P\left(|B-EB|\ge 1/2-EB\right)$ decays faster than any polynomial of n^{-1} and so does $P\left(|A-B|\ge 1/2-\epsilon_1\right)$ for ϵ_1 sufficiently small, then $P\left(\int \widehat{f}\le\epsilon_1\right)\le P\left(A\le\epsilon_1\right)$ decays faster than any polynomial of n^{-1} uniformly over $F^{\alpha}_{p,q}(M,\epsilon)$.

We now turn to (ii). Let $J_1^-=\left[\left(\frac{1}{1+2\alpha}-\epsilon_2\right)\log_2 n\right]$ and $J_1^+=\left[\left(\frac{1}{1+2\alpha}+\epsilon_2\right)\log_2 n\right]$ for some $\epsilon_2>0$. Note that $E\left(\int \widehat{f}-1\right)^2=E\left(\int\left(\sqrt{\widehat{f}}\right)^2-\int\left(\sqrt{f}\right)^2\right)^2$. So

$$\begin{aligned} E\left(\int \widehat{f}-1\right)^2 &= E\left(\left(\sum_{j\le J_1^-}+\sum_{J_1^-<j<J_1^+}+\sum_{j\ge J_1^+}\right)\sum_k\left(\widehat{\theta}_{j,k}^2-\theta_{j,k}^2\right)\right)^2\\ &\le 2E\left(\left(\sum_{j\le J_1^-}+\sum_{j\ge J_1^+}\right)\sum_k\left(\widehat{\theta}_{j,k}^2-\theta_{j,k}^2\right)\right)^2+2E\left(\sum_{J_1^-<j<J_1^+}\sum_k\left(\widehat{\theta}_{j,k}^2-\theta_{j,k}^2\right)\right)^2\\ &= R_1+R_2. \end{aligned}$$

Let g_P denote the projection of a function g to a subspace which only contains functions whose coefficients vanish with resolutions between J_1^- and J_1^+. We write

$$R_1 = E\left(\int\left(\widehat{\sqrt{f}}_P\right)^2 - \int\left(\sqrt{f}_P\right)^2\right)^2.$$

Similar to equation (50) we have $\sup_{\sqrt{f}\in F_{p,q}^{\alpha}(M,\epsilon)} P\left(\underline{x} : \left\|\widehat{f}_P\right\|_\infty \geq M_1\right) \leq C_l n^{-l}$ for some $M_1 > 0$ and any $l > 1$, then R_1 is bounded by $CE\int\left(\widehat{\sqrt{f}}_P - \sqrt{f}_P\right)^2 + C/n$, where

$$\begin{aligned}
E\int\left(\widehat{\sqrt{f}}_P - \sqrt{f}_P\right)^2 &\leq \sum_j\sum_{k=1}^{2^j}\min\left\{8\sum_{(j,k)\in B_j^i}\theta_{j,k}^2,\ 8\lambda_* L n^{-1}\right\} + Cm^{-4} + Cn^{-1} + CT^{-2d} \\
&= o\left(n^{-2\alpha/(2\alpha+1)}\right)
\end{aligned}$$

uniformly over all f following from similar arguments for equations (46), (47) and (48) in the section of the proof of the main theorem. Now we show $R_2 = o\left(n^{-2\alpha/(2\alpha+1)}\right)$ uniformly over all f. Write $\hat{\theta}_{j,k} = \widehat{a}_{j,k}u_{j,k} = \widehat{a}_{jk}\left(\theta'_{j,k} + \epsilon_{j,k} + \frac{1}{2\sqrt{n}}z_{j,k} + \xi_{j,k}\right)$ where $0 \leq \widehat{a}_{j,k} \leq 1$ is the shrinkage factor, then R_2 is bounded by

$$\begin{aligned}
& 2E\left[\sum_{J_1^-<j<J_1^+}\sum_i\left(\widehat{a}_{j,k}^2 - 1\right)\theta_{j,k}^2\right]^2 + 2E\left[\sum_{J_1^-<j<J_1^+}\sum_i\widehat{a}_{j,k}^2\left(u_{j,k}^2 - \theta_{j,k}^2\right)\right]^2 \\
\leq\ & 2E\left(\sum_{J_1^-<j<J_1^+}\sum_i\theta_{ij}^2\right)^2 + 2E\left[\sum_{J_1^-<j<J_1^+}\sum_i(u_{j,k}^2 - \theta_{j,k}'^2 + \theta_{j,k}'^2 - \theta_{j,k}^2)\right]^2 \\
=\ & 2E\left(\sum_{J_1^-<j<J_1^+}\sum_i\theta_{ij}^2\right)^2 + 2^{J_1^++2}E\sum_{J_1^-<j<J_1^+}\sum_i\left(u_{j,k}^2 - \theta_{j,k}'^2\right)^2 + 2\left(\sum_{J_1^-<j<J_1^+}\sum_i(\theta_{j,k}'^2 - \theta_{j,k}^2)\right)^2 \\
=\ & R_{21} + R_{22} + R_{23}
\end{aligned}$$

It is straightforward to see

$$R_{22} \leq C\frac{2^{J_1^+}}{n}\left(\sum_{J_1^-<j<J_1^+}\sum_i\theta_{j,k}'^2 + \frac{1}{n}\right) \leq C\frac{2^{J_1^+}}{n}\left(\sum_{J_1^-<j<J_1^+}\sum_i\theta_{j,k}^2 + CT^{-2d}\right),$$

and by the Cauchy-Schwarz inequality we have

$$R_{23} \leq \sum_{J_1^-<j<J_1^+}\sum_i\left(\theta'_{j,k} - \theta_{j,k}\right)^2 \cdot \sum_{J_1^-<j<J_1^+}\sum_i\left(\theta'_{j,k} + \theta_{j,k}\right)^2 < CT^{-2d}.$$

From equation (45) we have $\sum_{J_1^- < j < J_1^+} \sum_j \theta_{j,k}^2 \leq C \left(2^{J_1^-}\right)^{-2(\alpha - (1/p-1/2)_+)}$. It is easy to check $2\left(\alpha - (1/p-1/2)_+\right) > \alpha$ under the assumptions of Theorem 4 and so

$$R_2 \leq C \left(\left(2^{J_1^-}\right)^{-4(\alpha-(1/p-1/2)_+)} + \frac{2^{J_1^+}}{n} \left(2^{J_1^-}\right)^{-2(\alpha-(1/p-1/2)_+)} \right) = o\left(n^{-2\alpha/(1+2\alpha)}\right)$$

uniformly over all f, when ϵ_2 is sufficiently small. This proves (ii). ▮

7.7 Proof of Theorem 1

We have given a complete proof of Theorem 4, which gives asymptotic risk properties of our procedure for the Poissonized density estimation model,

$$F_n : N \sim Poi(n) \text{ and given } N,\, x_1, x_2, \ldots, x_N \text{ i.i.d. with density } f.$$

We shall now show that corresponding results hold for the density estimation problem,

$$E_n : x_1, x_2, \ldots, x_n \text{ i.i.d. with density } f.$$

Proof of Theorem 1: The Poisson experiment F_n can be generated from E_n as follows. Generate $N \sim \text{Poisson}(n)$. If $N > n$, generate $N - n$ i.i.d. additional observations with density f; otherwise, throw away $N - n$ observations.

Recall that we use N_i to denote the number of observations in the ith bin for F_n and Y_i denotes $\sqrt{N_i + 1/4}$ for F_n. Similarly, let N_i^* be the number of observations in the ith bin for E_n and $Y_i^* = \sqrt{N_i^* + 1/4}$. Apply the root-unroot procedure for both E_n and F_n and obtain two estimators of f for E_n and F_n respectively. Let $\widehat{h}$ denote the estimator of f for F_n. Following the notations in Section 4.1 $\widehat{\sqrt{f}}$ and $\widehat{\sqrt{h}}$ are given as follows

$$\begin{aligned}
\widehat{\sqrt{h}} &= \sum_{k=1}^{2^{j_0}} \hat{\hat{\theta}}_{j_0,k}\phi_{j_0,k}(t) + \sum_{j=j_0}^{J-1}\sum_{k=1}^{2^j} \hat{\theta}_{j,k}\psi_{j,k}(t),\ \hat{\theta}_{j,k} = (1 - \frac{\lambda_* L}{4nS_{j,i}^2})_+ u_{j,k} \\
\widehat{\sqrt{f}} &= \sum_{k=1}^{2^{j_0}} \hat{\hat{\theta}}^*_{j_0,k}\phi_{j_0,k}(t) + \sum_{j=j_0}^{J-1}\sum_{k=1}^{2^j} \hat{\theta}^*_{j,k}\psi_{j,k}(t),\ \hat{\theta}^*_{j,k} = (1 - \frac{\lambda_* L}{4nS_{j,i}^{*2}})_+ u^*_{j,k}.
\end{aligned}$$

where $(u^*_{j,k}) = W \cdot (n^{-\frac{1}{2}} Y_i^*)$, and $S_{j,i}^{*2} \equiv \sum_{(j,k)\in B_j^i} u_{j,k}^{*2}$ with $B_j^i = \{(j,k) : (i-1)L + 1 \leq k \leq iL\}$. Note that the Cauchy-Schwarz inequality yields

$$E\|\widehat{\sqrt{f}} - \sqrt{f}\|_2^2 \leq \left(\sqrt{E\|\widehat{\sqrt{f}} - \widehat{\sqrt{h}}\|_2^2} + \sqrt{E\|\widehat{\sqrt{h}} - \sqrt{f}\|_2^2}\right)^2.$$

It then suffices to show $\sup\limits_{\sqrt{f}\in F^{\alpha}_{p,q}(M,\epsilon)} E\|\widehat{\sqrt{f}-\sqrt{h}}\|_2^2 = O\left(n^{-1}\right)$ to establish the theorem. Note that $\|\widehat{\sqrt{f}-\sqrt{h}}\|_2^2 = \sum_{k=1}^{2^{j_0}}\left(\widetilde{u}_{j_0,k}-\widetilde{u}^*_{j_0,k}\right)^2 + \sum_{j=j_0}^{J-1}\sum_{k=1}^{2^j}\left(\hat{\theta}_{j,k}-\hat{\theta}^*_{j,k}\right)^2$. It is easy to check

$$\begin{aligned}\sum_{(j,k)\in B^i_j}\left(\hat{\theta}_{j,k}-\hat{\theta}^*_{j,k}\right)^2 &\le 2[(1-\frac{\lambda_* L}{4nS^{*2}_{j,i}})_+]^2\sum_{(j,k)\in B^i_j}\left(u_{j,k}-u^*_{j,k}\right)^2 + 2[(1-\frac{\lambda_* L}{4nS^2_{j,i}})_+ - (1-\frac{\lambda_* L}{4nS^{*2}_{j,i}})_+]^2 S^2_{j,i}\\ &\le 6\sum_{(j,k)\in B^i_j}\left(u_{j,k}-u^*_{j,k}\right)^2\end{aligned}$$

by applying the Cauchy-Schwarz inequality twice. Then we have

$$\|\widehat{\sqrt{f}-\sqrt{h}}\|_2^2 \le 6[\sum_{k=1}^{2^{j_0}}\left(\widetilde{u}_{j_0,k}-\widetilde{u}^*_{j_0,k}\right)^2 + \sum_{j=j_0}^{J-1}\sum_{k=1}^{2^j}\left(u_{j,k}-u^*_{j,k}\right)^2] = 6\frac{1}{n}\sum_{i=1}^{T}\left(Y_i-Y_i^*\right)^2. \quad (53)$$

Note that $Y_i = Y_i^* - \left(\sqrt{N_i^*+1/4}-\sqrt{N_i+1/4}\right) = Y_i^* - \frac{N_i^*-N_i}{\sqrt{N_i^*+1/4}+\sqrt{N_i+1/4}}$, and given N and N_i the distribution of $|N_i^*-N_i|$ is Binomial$(|N-n|, \int_{\frac{i-1}{T}}^{\frac{i}{T}} f(x)dx)$. It is then easy to check $E\left(Y_i-Y_i^*\right)^2 \le Cn^{-3/4}$. Thus $E\|\widehat{\sqrt{f}-\sqrt{h}}\|_2^2 \le 6C/n$ and the asymptotic optimality of $\widehat{f}$ under Hellinger loss is proved.

The asymptotic optimality of $\widehat{f}$ under L_2 loss and the parallel result for $\widetilde{f}$ can be proved by using the upper bound in equation (53) together with similar arguments given in Sections 7.5 and 7.6. ▮

References

[1] Anscombe, F. J. (1948). The transformation of Poisson, binomial and negative binomial data. *Biometrika* ***35***, 246-254.

[2] Bar-Lev, S. K. and Enis, P. (1990). On the construction of classes of variance stabilizing transformations. *Statist. Probab. Lett.* ***10***, 95-100.

[3] Bartlett, M. S. (1936). The square root transformation in analysis of variance. *J. Roy. Statist. Soc. Suppl.* ***3***, 68-78.

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

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

[6] Brown, L.D., Gans, N., Mandelbaum, A., Sakov, A., Shen, H., Zeltyn, S. and Zhao, L. H. (2005). Statistical analysis of a telephone call center: a queuing science perspective. *Jour. Amer. Statist. Assoc.* ***100***, 36-50.

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

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

[9] Cai, T. (2002). On block thresholding in wavelet regression: Adaptivity, block Size, and threshold level. *Statistica Sinica* ***12***, 1241-1273.

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

[11] Cai, T. and Silverman, B.W. (2001). Incorporating information on neighboring coefficients into wavelet estimation. *Sankhya Ser. B* ***63***, 127-148.

[12] Chicken, E. and Cai, T. (2005). Block thresholding for density estimation: Local and global adaptivity. *J. Multivariate Anal.* ***95***, 76-106.

[13] Cohen, A., Daubechies, I., Jawerth, B. and Vial, P. (1993). Multiresolution analysis, wavelets, and fast algorithms on an interval. *Comptes Rendus Acad. Sci. Paris* (A). **316**, 417-421.

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

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

[16] Donoho, D.L. (1993). Nonlinear wavelet methods for recovery of signals, densities, and spectra from indirect and noisy data. In *Different perspectives on Wavelets* (I. Daubechies Ed.), Vol. 47 of *Proc. Symp. Appl. Math.*, 173-205.

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

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

[19] Efron, B. (1982). Transformation theory: How normal is a family of a distributions? *Ann. Statist.* ***10***, 323-339.

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

[21] Hall, P., Kerkyacharian, G. and Picard, D. (1998). Block threshold rules for curve estimation using kernel and wavelet methods. *Ann. Statist.* ***26***, 922-942.

[22] Hall, P., Kerkyacharian, G. and Picard, D. (1999). On the minimax optimality of block thresholded wavelet estimators. *Statist. Sinica* ***9***, 33-50.

[23] Hall, P. and Patil, P. (1995). Formulae for mean integrated squared error of nonlinear wavelet-based density estimators. *Ann. Statist.* ***23***, 905-928.

[24] Hoyle, M. H. (1973). Transformations - an introduction and bibliography. *International Statistical Review* ***41***, 203-223.

[25] 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.

[26] Le Cam, L. (1974). On the information contained in additional observations. *Ann. Statist.* **2**, 630-649.

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

[28] Low, M. G. and Zhou, H. H. (2007). A complement to Le Cam's theorem. *Ann. Statist.* ***35***, 1146-1165.

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

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

[31] Runst, T. (1986). Mapping properties of non-linear operators in spaces of Triebel-Lizorkin and Besov type, *Anal. Math.* **12**, 313-346.

Silverman, B.W.(1986). *Density Estimation for Statistics and Data Analysis.* Chapman an Hall, New York.

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

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

[34] Weinberg, J., Brown, L.D. and Stroud, J. (2007). Bayesian forecasting of an inhomogeneous Poisson process, with applications to call center data. *Jour. Amer. Statist. Assoc.* ***102***, 1185-1198.

[35] Zhang, R. (2002). Nonparametric density estimation via wavelets. Ph.D. dissertation, Department of Statistics, University of Pennsylvania.

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