Near-Minimax Recursive Density Estimation on the Binary Hypercube

Maxim Raginsky
Duke University
Durham, NC 27708
m.raginsky@duke.edu

Svetlana Lazebnik
UNC Chapel Hill
Chapel Hill, NC 27599
lazebnik@cs.unc.edu

Rebecca Willett
Duke University
Durham, NC 27708
willett@duke.edu

Jorge Silva
Duke University
Durham, NC 27708
jg.silva@duke.edu

Abstract

This paper describes a recursive estimation procedure for multivariate binary densities using orthogonal expansions. For d covariates, there are 2^d basis coefficients to estimate, which renders conventional approaches computationally prohibitive when d is large. However, for a wide class of densities that satisfy a certain sparsity condition, our estimator runs in probabilistic polynomial time and adapts to the unknown sparsity of the underlying density in two key ways: (1) it attains near-minimax mean-squared error, and (2) the computational complexity is lower for sparser densities. Our method also allows for flexible control of the trade-off between mean-squared error and computational complexity.

1 Introduction

Multivariate binary data arise in a variety of fields, such as biostatistics [1], econometrics [2] or artificial intelligence [3]. In these and other settings, it is often necessary to estimate a probability density from a number of independent observations. Formally, we have n i.i.d. samples from a probability density f (with respect to the counting measure) on the d-dimensional *binary hypercube* $\mathcal{B}^d$, $\mathcal{B} \triangleq \{0,1\}$, and seek an estimate $\widehat{f}$ of f with a small mean-squared error $\mathrm{MSE}(f,\widehat{f}) = \mathbb{E}\left\{\sum_{x\in\mathcal{B}^d}(f(x)-\widehat{f}(x))^2\right\}$.

In many cases of practical interest, the number of covariates d is much larger than $\log n$, so direct estimation of f as a multinomial density with 2^d parameters is both unreliable and impractical. Thus, one has to resort to "nonparametric" methods and search for good estimators in a suitably defined class whose complexity grows with n. Some nonparametric methods proposed in the literature, such as kernels [4] and orthogonal expansions [5, 6], either have very slow rates of MSE convergence or are computationally prohibitive for large d. For example, the kernel method [4] requires $O(n^2 d)$ operations to compute the estimate at any $x \in \mathcal{B}^d$, yet its MSE decays as $O(n^{-4/(4+d)})$ [7], which is extremely slow when d is large. In contrast, orthogonal function methods generally have much better MSE decay rates, but rely on estimating 2^d coefficients in a fixed basis, which requires enormous computational resources for large d. For instance, using the Fast Hadamard Transform to estimate the coefficients in the so-called *Walsh basis* using n samples requires $O(nd2^d)$ operations [5].

In this paper we take up the problem of accurate, computationally tractable estimation of a density on the binary hypercube. We take the minimax point of view, where we assume that f comes from a particular function class $\mathcal{F}$ and seek an estimator that approximately attains the minimax MSE

$$R_n^*(\mathcal{F}) \triangleq \inf_{\widehat{f}} \sup_{f\in\mathcal{F}} \mathrm{MSE}(f,\widehat{f}),$$

where the infimum is over all estimators based on n i.i.d. samples. We will define our function class to reflect another feature often encountered in situations involving multivariate binary data: namely,

that the shape of the underlying density is strongly influenced by small constellations of the d covariates. For example, when working with panel data [2], it may be the case that the answers to some specific subset of questions are highly correlated among a particular group of the panel participants, and the responses of these participants to other questions are nearly random; moreover, there may be several such distinct groups in the panel. To model such "constellation effects" mathematically, we will consider classes of densities that satisfy a particular *sparsity condition*.

Our contribution consists in developing a thresholding density estimator that adapts to the unknown sparsity of the underlying density in two key ways: (1) it is near-minimax optimal, with the error decay rate depending upon the sparsity, and (2) it can be implemented using a recursive algorithm that runs in probabilistic polynomial time and whose computational complexity is lower for sparser densities. The algorithm entails recursively examining empirical estimates of whole *blocks* of the 2^d basis coefficients. At each stage of the algorithm, the weights of the coefficients estimated at previous stages are used to decide which remaining coefficients are most likely to be significant, and computing resources are allocated accordingly. We show that this decision is accurate with high probability. An additional attractive feature of our approach is that it gives us a principled way of trading off MSE against computational complexity by controlling the decay of the threshold as a function of the recursion depth.

2 Preliminaries

We first list some definitions and results needed in the sequel. Throughout the paper, C and c denote generic constants whose values may change from line to line. For two real numbers a and b, $a \wedge b$ and $a \vee b$ denote, respectively, the smaller and the larger of the two.

Biased Walsh bases. Let μ_d denote the counting measure on the d-dimensional binary hypercube $\mathcal{B}^d$. Then the space of all real-valued functions on $\mathcal{B}^d$ is the real Hilbert space $L^2(\mu_d)$ with the standard inner product $\langle f, g\rangle \triangleq \sum_{x\in\mathcal{B}^d} f(x)g(x)$. Given any $\eta \in (0,1)$, we can construct an orthonormal system $\Phi_{d,\eta}$ in $L^2(\mu_d)$ as follows. Define two functions $\varphi_{0,\eta}, \varphi_{1,\eta} : \mathcal{B} \to \mathbb{R}$ by

$$\varphi_{0,\eta}(x) \triangleq (1-\eta)^{x/2}\eta^{(1-x)/2} \quad \text{and} \quad \varphi_{1,\eta}(x) \triangleq (-1)^x \eta^{x/2}(1-\eta)^{(1-x)/2}, \quad x \in \{0,1\}. \tag{1}$$

Now, for any $s = (s(1), \ldots, s(d)) \in \mathcal{B}^d$ define the function $\varphi_{s,\eta} : \mathcal{B}^d \to \mathbb{R}$ by

$$\varphi_{s,\eta}(x) \triangleq \prod_{i=1}^{d} \varphi_{s(i),\eta}(x(i)), \qquad \forall x = (x(1), \ldots, x(d)) \in \mathcal{B}^d \tag{2}$$

(this is written more succinctly as $\varphi_{s,\eta} = \varphi_{s(1),\eta} \otimes \ldots \otimes \varphi_{s(d),\eta}$, where $\otimes$ is the tensor product). The set $\Phi_{d,\eta} = \{\varphi_{s,\eta} : s \in \mathcal{B}^d\}$ is an orthonormal system in $L^2(\mu_d)$, which is referred to as the *Walsh system with bias* η [8, 9]. Any function $f \in L^2(\mu_d)$ can be uniquely represented as

$$f = \sum_{s\in\mathcal{B}^d} \theta_{s,\eta}\varphi_{s,\eta},$$

where $\theta_{s,\eta} = \langle f, \varphi_{s,\eta}\rangle$. When $\eta = 1/2$, we get the standard Walsh system used in [5, 6]; in that case, we shall omit the index $\eta = 1/2$ for simplicity. The product structure of the biased Walsh bases makes them especially convenient for statistical applications as it allows for a computationally efficient recursive method for computing accurate estimates of squared coefficients in certain hierarchically structured sets.

Sparsity and weak-ℓ_p balls. We are interested in densities whose representations in some biased Walsh basis satisfy a certain sparsity constraint. Given $\eta \in (0,1)$ and a function $f \in L^2(\mu_d)$, let $\theta(f)$ denote the list of its coefficients in $\Phi_{d,\eta}$. We are interested in cases when the components of $\theta(f)$ decay according to a power law. Formally, let $\theta_{(1)}, \ldots, \theta_{(M)}$, where $M = 2^d$, be the components of $\theta(f)$ arranged in decreasing order of magnitude: $|\theta_{(1)}| \geq |\theta_{(2)}| \geq \ldots \geq |\theta_{(M)}|$. Given some $0 < p < \infty$, we say that $\theta(f)$ belongs to the *weak-ℓ_p* ball of radius R [10], and write $\theta(f) \in w\ell_p(R)$, if

$$|\theta_{(m)}| \leq R \cdot m^{-1/p}, \qquad 1 \leq m \leq M. \tag{3}$$

It is not hard to show that the coefficients of any probability density on $\mathcal{B}^d$ in $\Phi_{d,\eta}$ are bounded by $R(\eta) = [\eta \vee (1-\eta)]^{d/2}$. With this in mind, let us define the class $\mathcal{F}_d(p,\eta)$ of all functions f on $\mathcal{B}^d$ satisfying $\theta(f) \in w\ell_p(R(\eta))$ in $\mathbb{R}^M$. We are particularly interested in the case $0 < p < 2$. When $\eta = 1/2$, with $R(\eta) = 2^{-d/2}$, we shall write simply $\mathcal{F}_d(p)$.

We will need approximation properties of weak-ℓ_p balls as listed, e.g., in [11]. The basic fact is that the power-law condition (3) is equivalent to the concentration estimate

$$\left|\left\{s \in \mathcal{B}^d : |\theta_s| \geq \lambda\right\}\right| \leq (R/\lambda)^p, \qquad \forall \lambda > 0. \tag{4}$$

For any $1 \leq k \leq M$, let $\theta_k(f)$ denote the vector $\theta(f)$ with $\theta_{(k+1)}, \ldots, \theta_{(M)}$ set to zero. Then it follows from (3) that $\|\theta(f) - \theta_k(f)\|_{\ell_M^2} \leq CRk^{-r}$, where $r \triangleq 1/p - 1/2$, and C is some constant that depends only on p. Given any $f \in \mathcal{F}_d(p,\eta)$ and denoting by f_k the function obtained from it by retaining only the k largest coefficients, we get from Parseval's identity that

$$\|f - f_k\|_{L^2(\mu_d)} \leq CRk^{-r}. \tag{5}$$

To get a feeling for what the classes $\mathcal{F}_d(p,\eta)$ could model in practice, we note that, for a fixed $\eta \in (0,1)$, the product of d $\mathrm{Bernoulli}(\eta^*)$ densities with $\eta^* \triangleq \sqrt{\eta}/(\sqrt{\eta} + \sqrt{1-\eta})$ is the unique sparsest density in the entire scale of $\mathcal{F}_d(p,\eta)$ spaces with $0 < p < 2$: all of its coefficients in $\mathcal{F}_{d,\eta}$ are zero, except for $\theta_{s,\eta}$ with $s = (0,\ldots,0)$, which is equal to $(\eta^*/\sqrt{\eta})^d$. Other densities in $\{\Phi_d(p,\eta) : 0 < p < 2\}$ include, for example, mixtures of components that, up to a permutation of $\{1,\ldots,d\}$, can be written as a tensor product of a large number of $\mathrm{Bernoulli}(\eta^*)$ densities and some other density. The parameter η can be interpreted either as the default noise level in measuring an individual covariate or as a smoothness parameter that interpolates between the point masses $\delta_{(0,\ldots,0)}$ and $\delta_{(1,\ldots,1)}$. We assume that η is known (e.g., from some preliminary exploration of the data or from domain-specific prior information) and fixed.

In the following, we limit ourselves to the "noisiest" case $\eta = 1/2$ with $R(1/2) = 2^{-d/2}$. Our theory can be easily modified to cover any other $\eta \in (0,1)$: one would need to replace $R = 2^{-d/2}$ with the corresponding $R(\eta)$ and use the bound $\|\varphi_{s,\eta}\|_\infty \leq R(\eta)$ instead of $\|\varphi_s\|_\infty \leq 2^{-d/2}$ when estimating variances and higher moments.

3 Density estimation via recursive Walsh thresholding

We now turn to our problem of estimating a density f on $\mathcal{B}^d$ from a sample $\{X_i\}_{i=1}^n$ when $f \in \mathcal{F}_d(p)$ for some unknown $0 < p < 2$. The minimax theory for weak-ℓ_p balls [10] says that

$$R_n^*(\mathcal{F}_d(p)) \geq CM^{-p/2}n^{-2r/(2r+1)}, \qquad r = 1/p - 1/2$$

where $M = 2^d$. We shall construct an estimator that *adapts to unknown sparsity* of f in the sense that it achieves this minimax rate up to a logarithmic factor without prior knowledge of p and that its computational complexity improves as $p \to 0$.

Our method is based on the thresholding of empirical Walsh coefficients. A thresholding estimator is any estimator of the form

$$\widehat{f} = \sum_{s \in \mathcal{B}^d} I_{\{T(\widehat{\theta}_s) \geq \lambda_n\}} \widehat{\theta}_s \varphi_s,$$

where $\widehat{\theta}_s = (1/n)\sum_{i=1}^n \varphi_s(X_i)$ are empirical estimates of the Walsh coefficients of f, $T(\cdot)$ is some statistic, and $I_{\{\cdot\}}$ is an indicator function. The threshold λ_n depends on the sample size. For example, in [5, 6] the statistic $T(\widehat{\theta}_s) = \widehat{\theta}_s^2$ was used with the threshold $\lambda_n = 1/M(n+1)$. This choice was motivated by the considerations of bias-variance trade-off for each individual coefficient.

The main disadvantage of such direct methods is the need to estimate all $M = 2^d$ Walsh coefficients. While this is not an issue when $d \asymp \log n$, it is clearly impractical when $d \gg \log n$. To deal with this issue, we will consider a recursive thresholding approach that will allow us to reject whole *groups* of coefficients based on efficiently computable statistics. This approach is motivated as follows. For any $1 \leq k \leq d$, we can write any $f \in L^2(\mu_d)$ with the Walsh coefficients $\theta(f)$ as

$$f = \sum_{u \in \mathcal{B}^k} \sum_{v \in \mathcal{B}^{d-k}} \theta_{uv}\varphi_{uv} = \sum_{u \in \mathcal{B}^k} f_u \otimes \varphi_u,$$

where uv denotes the concatenation of $u \in \mathcal{B}^k$ and $v \in \mathcal{B}^{d-k}$ and, for each $u \in \mathcal{B}^k$, $f_u \triangleq \sum_{v \in \mathcal{B}^{d-k}} \theta_{uv} \varphi_v$ lies in $L^2(\mu_{d-k})$. By Parseval's identity, $W_u \triangleq \|f_u\|^2_{L^2(\mu_{d-k})} = \sum_{v \in \mathcal{B}^{d-k}} \theta^2_{uv}$. This means that if $W_u < \lambda$ for some $u \in \mathcal{B}^k$, then $\theta^2_{uv} < \lambda$ for every $v \in \mathcal{B}^{d-k}$. Thus, we could start at $u = 0$ and $u = 1$ and check whether $W_u \geq \lambda$. If not, then we would discard all θ_{uv} with $v \in \mathcal{B}^{d-1}$; otherwise, we would proceed on to $u0$ and $u1$. At the end of this process, we will be left only with those $s \in \mathcal{B}^d$ for which $\theta^2_s \geq \lambda$. Let f_λ denote the resulting function. If $f \in \mathcal{F}_d(p)$ for some p, then we will have $\|f - f_\lambda\|^2_{L^2(\mu_d)} \leq CM^{-1}(M\lambda)^{-2r/(2r+1)}$.

We will follow this reasoning in constructing our estimator. We begin by developing an estimator for W_u. We will use the following fact, easily proved using the definitions (1) and (2) of the Walsh functions: for any density f on $\mathcal{B}^d$, any k and $u \in \mathcal{B}^k$, we have

$$f_u(y) = \mathbb{E}_f \left\{ \varphi_u(\pi_k(X)) I_{\{\sigma_k(X) = y\}} \right\}, \forall y \in \mathcal{B}^{d-k} \quad \text{and} \quad W_u = \mathbb{E}_f \left\{ \varphi_u(\pi_k(X)) f_u(\sigma_k(X)) \right\},$$

where $\pi_k(x) \triangleq (x(1), \ldots, x(k))$ and $\sigma_k(x) \triangleq (x(k+1), \ldots, x(d))$ for any $x \in \mathcal{B}^d$. This suggests that we can estimate W_u by

$$\widehat{W}_u = \frac{1}{n^2} \sum_{i_1=1}^n \sum_{i_2=1}^n \varphi_u(\pi_k(X_{i_1})) \varphi_u(\pi_k(X_{i_2})) I_{\{\sigma_k(X_{i_1}) = \sigma_k(X_{i_2})\}}. \tag{6}$$

Using induction and Eqs. (1) and (2), we can prove that $\widehat{W}_u = \sum_{v \in \mathcal{B}^{d-k}} \widehat{\theta}^2_{uv}$. An advantage of computing $\widehat{W}_u$ indirectly via (6) rather than as a sum of $\widehat{\theta}^2_{uv}$, $v \in \mathcal{B}^{d-k}$, is that, while the latter has $O(2^{d-k}n)$ complexity, the former has only $O(n^2 d)$ complexity. This can lead to significant computational savings for small k. When $k \geq d - \log(nd)$, it becomes more efficient to use the direct estimator.

Now we can define our density estimation procedure. Instead of using a single threshold for all $1 \leq k \leq d$, we consider a more flexible strategy: for every k, we shall compare each $\widehat{W}_u$ to a threshold that depends not only on n, but also on k. Specifically, we will let

$$\lambda_{k,n} = \frac{\alpha_k \log n}{n}, \qquad 1 \leq k \leq d \tag{7}$$

where $\boldsymbol{\alpha} = \{\alpha_k\}_{k=1}^d$ satisfies $\alpha_1 \geq \alpha_k \geq \alpha_d > 0$. (This k-dependent scaling will allow us to trade off MSE and computational complexity.) Given $\boldsymbol{\lambda} = \{\lambda_{k,n}\}_{k=1}^d$, define the set $A(\boldsymbol{\lambda}) \triangleq \{s \in \mathcal{B}^d : \widehat{W}_{\pi_k(s)} \geq \lambda_{k,n}, \forall 1 \leq k \leq d\}$ and the corresponding estimator

$$\widehat{f}_{\mathrm{RWT}} \triangleq \sum_{s \in \mathcal{B}^d} I_{\{s \in A(\boldsymbol{\lambda})\}} \widehat{\theta}_s \varphi_s, \tag{8}$$

where RWT stands for "recursive Walsh thresholding." To implement $\widehat{f}_{\mathrm{RWT}}$ on a computer, we adapt the algorithm of Goldreich and Levin [12], originally developed for cryptography and later applied to the problem of learning Boolean functions from membership queries [13]: we call the routine RECURSIVEWALSH, shown in Algorithm 1, with $u = \varnothing$ (the empty string) and with $\boldsymbol{\lambda}$ from (7).

Analysis of the estimator. We now turn to the asymptotic analysis of the MSE and the computational complexity of $\widehat{f}_{\mathrm{RWT}}$. We first prove that $\widehat{f}_{\mathrm{RWT}}$ adapts to unknown sparsity of f:

Theorem 3.1 *Suppose the threshold sequence $\boldsymbol{\lambda} = \{\lambda_k\}_{k=1}^d$ is such that $\alpha_d \geq (20d + 25)^2 / 2^d$. Then for all $0 < p < 2$ the estimator (8) satisfies*

$$\sup_{f \in \mathcal{F}_d(p)} \mathrm{MSE}(f, \widehat{f}_{\mathrm{RWT}}) = \sup_{f \in \mathcal{F}_d(p)} \mathbb{E}_f \|f - \widehat{f}_{\mathrm{RWT}}\|^2_{L^2(\mu_d)} \leq \frac{C}{2^d} \left(\frac{2^d \alpha_1 \log n}{n} \right)^{2r/(2r+1)}, \tag{9}$$

where the constant C depends only on p.

Proof: Let us decompose the squared L^2 error of $\widehat{f}_{\mathrm{RWT}}$ as

$$\|f - \widehat{f}_{\mathrm{RWT}}\|^2_{L^2(\mu_d)} = \sum_s I_{\{s \in A(\boldsymbol{\lambda})\}} (\theta_s - \widehat{\theta}_s)^2 + \sum_s I_{\{s \in A(\boldsymbol{\lambda})^c\}} \theta^2_s \equiv T_1 + T_2.$$

Algorithm 1 RECURSIVEWALSH$(u, \boldsymbol{\lambda})$

$k \leftarrow \mathrm{length}(u)$
if $k = d$ **then**
  compute $\widehat{\theta}_u \leftarrow \frac{1}{n}\sum_{i=1}^{n} \varphi_u(X_i)$; **if** $\widehat{\theta}_u^2 \geq \lambda_{d,n}$ **then output** $u, \widehat{\theta}_u$; **return**
end if
compute $\widehat{W}_{u0} \leftarrow \frac{1}{n^2}\sum_{i_1=1}^{n}\sum_{i_2=1}^{n} \varphi_{u0}(\pi_{k+1}(X_{i_1}))\varphi_{u0}(\pi_{k+1}(X_{i_2}))I_{\{\sigma_{k+1}(X_{i_1})=\sigma_{k+1}(X_{i_2})\}}$
compute $\widehat{W}_{u1} \leftarrow \frac{1}{n^2}\sum_{i_1=1}^{n}\sum_{i_2=1}^{n} \varphi_{u1}(\pi_{k+1}(X_{i_1}))\varphi_{u1}(\pi_{k+1}(X_{i_2}))I_{\{\sigma_{k+1}(X_{i_1})=\sigma_{k+1}(X_{i_2})\}}$
if $\widehat{W}_{u0} \leq \lambda_{k+1,n}$ **then return else** RECURSIVEWALSH$(u0, \boldsymbol{\lambda})$; **end if**
if $\widehat{W}_{u1} \leq \lambda_{k+1,n}$ **then return else** RECURSIVEWALSH$(u1, \boldsymbol{\lambda})$; **end if**

We start by observing that $s \in A(\boldsymbol{\lambda})$ only if $\widehat{\theta}_s^2 \geq \lambda_{d,n}$, while for any $s \in A(\boldsymbol{\lambda})^c$ there exists some $1 \leq k \leq d$ such that $\widehat{\theta}_s^2 < \lambda_{k,n} \leq \lambda_{1,n}$. Defining the sets $A_1 = \{s \in \mathcal{B}^d : \widehat{\theta}_s^2 \geq \lambda_{d,n}\}$ and $A_2 = \{s \in \mathcal{B}^d : \widehat{\theta}_s^2 < \lambda_{1,n}\}$, we get $T_1 \leq \sum_s I_{\{s\in A_1\}}(\theta_s - \widehat{\theta}_s)^2$ and $T_2 \leq \sum_s I_{\{s\in A_2\}}\theta_s^2$. Further, defining $B = \{s \in \mathcal{B}^d : \theta_s^2 < \lambda_{d,n}/2\}$ and $S = \{s \in \mathcal{B}^d : \theta_s^2 \geq 3\lambda_{1,n}/2\}$, we can write

$$T_1 = \sum_s I_{\{s\in A_1\cap B\}}(\theta_s - \widehat{\theta}_s)^2 + \sum_s I_{\{s\in A_1\cap B^c\}}(\theta_s - \widehat{\theta}_s)^2 \equiv T_{11} + T_{12},$$

$$T_2 = \sum_s I_{\{s\in A_2\cap S\}}\theta_s^2 + \sum_s I_{\{s\in A_2\cap S^c\}}\theta_s^2 \equiv T_{21} + T_{22}.$$

First we deal with the easy terms T_{12}, T_{22}. Applying (4), (5) and a bit of algebra, we get

$$\mathbb{E}\,T_{12} \leq \frac{1}{Mn}\left|\left\{s : \theta_s^2 \geq \lambda_{d,n}/2\right\}\right| \leq \frac{1}{Mn}\left(\frac{2}{M\lambda_{d,n}}\right)^{p/2} \leq \frac{1}{M}n^{-2r/(2r+1)}, \tag{10}$$

$$\mathbb{E}\,T_{22} \leq \sum_{s\in\mathcal{B}^d} I_{\{\theta_s^2 < (3\alpha_1/2)\log n/n\}}\theta_s^2 \leq \frac{C}{M}\left(\frac{M\alpha_1\log n}{n}\right)^{2r/(2r+1)}. \tag{11}$$

Next we deal with the large-deviation terms T_{11} and T_{21}. Using Cauchy–Schwarz, we get

$$\mathbb{E}\,T_{11} \leq \sum_s \left[\mathbb{E}(\theta_s - \widehat{\theta}_s)^4 \cdot \mathbb{P}(s \in A_1 \cap B)\right]^{1/2}. \tag{12}$$

To estimate the fourth moment in (12), we use Rosenthal's inequality [14] to get $\mathbb{E}(\theta_s - \widehat{\theta}_s)^4 \leq c/M^2n^2$. To bound the probability that $s \in A_1 \cap B$, we observe that $s \in A_1 \cap B$ implies that $|\widehat{\theta}_s - \theta_s| \geq (1/5)\sqrt{\lambda_{d,n}}$, and then use Bernstein's inequality [14] to get

$$\mathbb{P}\left(|\widehat{\theta}_s - \theta_s| \geq (1/5)\sqrt{\lambda_{d,n}}\right) \leq 2\exp\left(-\frac{\beta^2\log n}{2(1+2\beta/3)}\right) = 2n^{-\beta^2/[2(1+2\beta/3)]} \leq 2n^{-(\beta-1)/2}$$

with $\beta = (1/5)\sqrt{M\alpha_d} \geq 4d+5$. Since $n^{-(\beta-1)/2} \leq n^{-2(d+1)}$, we have

$$\mathbb{E}\,T_{11} \leq Cn^{-(d+1)} \leq C/(Mn). \tag{13}$$

Finally, $\mathbb{E}\,T_{21} \leq \sum_s \mathbb{P}(s \in A_2 \cap S)\theta_s^2$. Using the same argument as above, we get $\mathbb{P}(s \in A_2 \cap S) \leq 2n^{-(\gamma-1)/2}$, where $\gamma = (1/5)\sqrt{M\alpha_1}$. Since $\theta_s^2 \leq 1/M$ for all $s \in \mathcal{B}^d$ and since $\gamma \geq \beta$, this gives

$$\mathbb{E}\,T_{21} \leq 2n^{-2(d+1)} \leq 2/(Mn). \tag{14}$$

Putting together Eqs. (10), (11), (13), and (14), we get (9), and the theorem is proved. ■

Our second result concerns the running time of Algorithm 1. Let $K(\boldsymbol{\alpha}, p) \triangleq \sum_{k=1}^{d} \alpha_k^{-p/2}$.

Theorem 3.2 *Given any $\delta \in (0,1)$, provided each α_k is chosen so that*

$$\sqrt{2^k\alpha_k n\log n} \geq 5\left[C_2\sqrt{n} + (\log(d/\delta) + k)/\log e\right], \tag{15}$$

Algorithm 1 runs in $O(n^2 d(n/M\log n)^{p/2}K(\boldsymbol{\alpha}, p))$ time with probability at least $1-\delta$.

Proof: The complexity is determined by the number of calls to RECURSIVEWALSH. For each k, a call to RECURSIVEWALSH is made at every $u \in \mathcal{B}^k$ with $\widehat{W}_u \geq \lambda_{k,n}$. Let us say that a call to RECURSIVEWALSH$(u, \boldsymbol{\lambda})$ is *correct* if $W_u \geq \lambda_{k,n}/2$. We will show that, with probability at least $1-\delta$, only the correct calls are made. The probability of making at least one incorrect call is

$$\mathbb{P}\left(\bigcup_{k=1}^{d}\bigcup_{u\in\mathcal{B}^k}\{\widehat{W}_u \geq \lambda_{k,n}, W_u < \lambda_{k,n}/2\}\right) \leq \sum_{k=1}^{d}\sum_{u\in\mathcal{B}^k}\mathbb{P}\left(\widehat{W}_u \geq \lambda_{k,n}, W_u < \lambda_{k,n}/2\right).$$

For a given $u \in \mathcal{B}^k$, $\widehat{W}_u \geq \lambda_{k,n}$ and $W_u < \lambda_{k,n}/2$ together imply that $\|f_u - \widehat{f}_u\|^2_{L^2(\mu_{d-k})} \geq (1/5)\sqrt{\lambda_{k,n}}$, where $\widehat{f}_u \triangleq \sum_{v\in\mathcal{B}^{d-k}}\widehat{\theta}_{uv}\varphi_v$. Now, it can be shown that, for every $u \in \mathcal{B}^k$, the norm $\|f_u - \widehat{f}_u\|_{L^2(\mu_{d-k})}$ can be expressed as a supremum of an empirical process [15] over a certain function class that depends on k (details are omitted for lack of space). We can then use Talagrand's concentration-of-measure inequality for empirical processes [16] to get

$$\mathbb{P}(\widehat{W}_u \geq \lambda_{k,n}, W_u < \lambda_{k,n}/2) \leq \exp\big\{-nC_1(2^k a^2_{k,n} \wedge 2^{k/2}a_{k,n})\big\},$$

where $a_{k,n} = (1/5)\sqrt{\alpha_k \log n/n} - C_2/\sqrt{2^k n}$, and C_1, C_2 are the absolute constants in Talagrand's bound. If we choose α_k as in (15), then $\mathbb{P}(\widehat{W}_u \geq \lambda_{k,n}, W_u < \lambda_{k,n}/2) \leq \delta/(d2^{d-k})$ for all $u \in \mathcal{B}^k$. Summing over $k, u \in \mathcal{B}^k$, we see that, with probability $\geq 1-\delta$, only the correct calls will be made.

It remains to bound the number of the correct calls. For each k, $W_u \geq \lambda_{k,n}/2$ implies that there exists at least one $v \in \mathcal{B}^{d-k}$ such that $\theta^2_{uv} \geq \lambda_{k,n}/2$. Since for every $1 \leq k \leq d$ each θ_s contributes to exactly one W_u, we have by the pigeonhole principle that

$$\big|\{u \in \mathcal{B}^k : W_u \geq \lambda_{k,n}/2\}\big| \leq \big|\{s \in \mathcal{B}^d : \theta^2_s \geq \lambda_{k,n}/2\}\big| \leq (2/M\lambda_{k,n})^{p/2},$$

where in the second inequality we used (4) with $R = 1/\sqrt{M}$. Hence, the number of correct recursive calls is bounded by $N = \sum_{k=1}^{d}(2/M\lambda_{k,n})^{p/2} = (2n/M\log n)^{p/2}K(\boldsymbol{\alpha}, p)$. At each call, we compute an estimate of the corresponding W_{u0} and W_{u1}, which requires $O(n^2 d)$ operations. Therefore, with probability at least $1-\delta$, the time complexity will be as stated in the theorem. ■

MSE vs. complexity. By controlling the rate at which the sequence α_k decays with k, we can trade off MSE against complexity. Consider the following two extreme cases: (1) $\alpha_1 = \ldots = \alpha_d \sim 1/M$ and (2) $\alpha_k \sim 2^{d-k}/M$. The first case, which reduces to term-by-term thresholding, achieves the best bias-variance trade-off with the MSE $O((\log n/n)^{2r/(2r+1)}(1/M))$. However, it has $K(\boldsymbol{\alpha}, p) = O(M^{p/2}d)$, resulting in $O(d^2n^2(n/\log n)^{p/2})$ complexity. The second case, which leads to a very severe estimator that will tend to reject a lot of coefficients, has MSE of $O((\log n/n)^{2r/(2r+1)}M^{-1/(2r+1)})$, but $K(\boldsymbol{\alpha}, p) = O(M^{p/2})$, leading to a considerably better $O(dn^2(n/\log n)^{p/2})$ complexity. From the computational viewpoint, it is preferable to use rapidly decaying thresholds. However, this reduction in complexity will be offset by a corresponding increase in MSE. In fact, using exponentially decaying α_k's in practice is not advisable as its low complexity is mainly due to the fact that it will tend to reject even the big coefficients very early on, especially when d is large. To achieve a good balance between complexity and MSE, a moderately decaying threshold sequence might be best, e.g., $\alpha_k \sim (d-k+1)^m/M$ for some $m \geq 1$. As $p \to 0$, the effect of $\boldsymbol{\lambda}$ on complexity becomes negligible, and the complexity tends to $O(n^2 d)$.

Positivity and normalization issues. As is the case with orthogonal series estimators, $\widehat{f}_{\mathrm{RWT}}$ may not necessarily be a bona fide density. In particular, there may be some $x \in \mathcal{B}^d$ such that $\widehat{f}_{\mathrm{RWT}}(x) < 0$, and it may happen that $\int \widehat{f}_{\mathrm{RWT}} d\mu_d \neq 1$. In principle, this can be handled by clipping the negative values at zero and renormalizing, which can only improve the MSE. In practice renormalization may be computationally expensive when d is very large. If the estimate is suitably sparse, however, the renormalization can be carried out approximately using Monte-Carlo methods.

4 Simulations

The focus of our work is theoretical, consisting in the derivation of a recursive thresholding procedure for estimating multivariate binary densities (Algorithm 1), with a proof of its near-minimaxity

and an asymptotic analysis of its complexity. Although an extensive empirical evaluation is outside the scope of this paper, we have implemented the proposed estimator, and now present some simulation results to demonstrate its small-sample performance. We generated synthetic observations from a mixture density f on a 15-dimensional binary hypercube. The mixture has 10 components, where each component is a product density with 12 randomly chosen covariates having $\mathrm{Bernoulli}(1/2)$ distributions, and the other three having $\mathrm{Bernoulli}(0.9)$ distributions. For $d = 15$, it is still feasible to quickly compute the ground truth, consisting of 32768 values of f and its Walsh coefficients. These values are shown in Fig. 1 (left). As can be seen from the coefficient profile in the bottom of the figure, this density is clearly sparse. Fig. 1 also shows the estimated probabilities and the Walsh coefficients for sample sizes $n = 5000$ (middle) and $n = 10000$ (right).

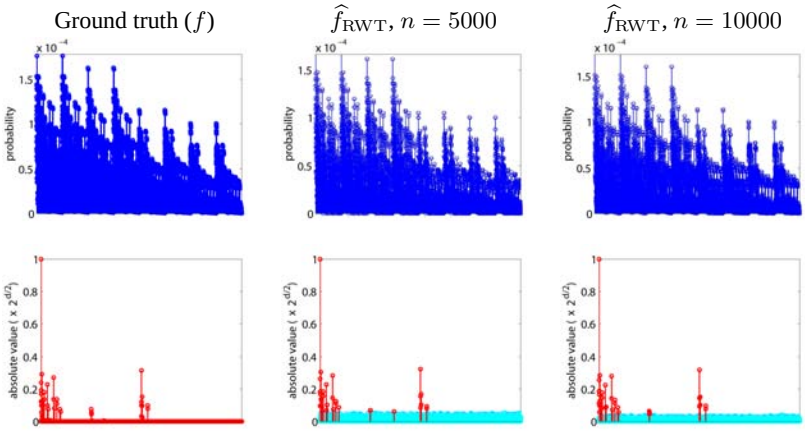

Figure 1: Ground truth (left) and estimated density for $n = 5000$ (middle) and $n = 10000$ (right) with constant thresholding. Top: true and estimated probabilities (clipped at zero and renormalized) arranged in lexicographic order. Bottom: absolute values of true and estimated Walsh coefficients arranged in lexicographic order. For the estimated densities, the coefficient plots also show the threshold level (dotted line) and absolute values of the rejected coefficients (lighter color).

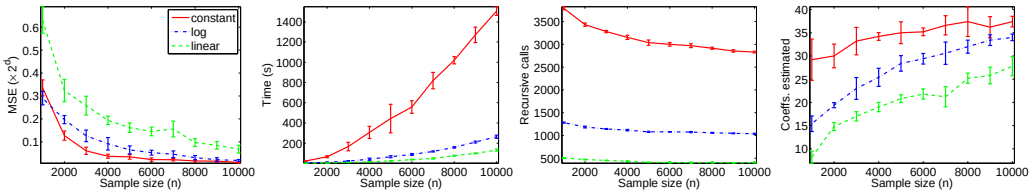

Figure 2: Small-sample performance of $\widehat{f}_{\mathrm{RWT}}$ in estimating f wth three different thresholding schemes: (a) MSE; (b) running time (in seconds); (c) number of recursive calls; (d) number of coefficients retained by the algorithm. All results are averaged over five independent runs for each sample size (the error bars show the standard deviations).

To study the trade-off between MSE and complexity, we implemented three different thresholding schemes: (1) constant, $\lambda_{k,n} = 2\log n/(2^d n)$, (2) logarithmic, $\lambda_{k,n} = 2\log(d-k+2)\log n/(2^d n)$, and (3) linear, $\lambda_{k,n} = 2(d-k+1)\log n/(2^d n)$. Up to the $\log n$ factor (dictated by the theory), the thresholds at $k = d$ are set to twice the variance of the empirical estimate of any coefficient whose value is zero; this forces the estimator to reject empirical coefficients whose values cannot be reliably distinguished from zero. Occasionally, spurious coefficients get retained, as can be seen in Fig. 1 (middle) for the estimate for $n = 5000$. Fig 2 shows the performance of $\widehat{f}_{\mathrm{RWT}}$. Fig. 2(a) is a plot of MSE vs. sample size. In agreement with the theory, MSE is the smallest for the constant thresholding scheme [which is simply an efficient recursive implementation of a term-by-term thresholding estimator with $\lambda_n \sim \log n/(Mn)$], and then it increases for the logarithmic and for the linear schemes. Fig. 2(b,c) shows the running time (in seconds) and the number of recursive

calls made to RECURSIVEWALSH vs. sample size. The number of recursive calls is a platform-independent way of gauging the computational complexity of the algorithm, although it should be kept in mind that each recursive call has $O(n^2 d)$ overhead. The running time increases polynomially with n, and is the largest for the constant scheme, followed by the logarithmic and the linear schemes. We see that, while the MSE of the logarithmic scheme is fairly close to that of the constant scheme, its complexity is considerably lower, in terms of both the number of recursive calls and the running time. In all three cases, the number of recursive calls decreases with n due to the fact that weight estimates become increasingly accurate with n, which causes the expected number of false discoveries (i.e., making a recursive call at an internal node of the tree only to reject its descendants later) to decrease. Finally, Fig. 2(d) shows the number of coefficients retained in the estimate. This number grows with n as a consequence of the fact that the threshold decreases with n, while the number of accurately estimated coefficients increases. The true density f has 40 parameters: 9 to specify the weights of the components, 3 per component to locate the indices of the nonuniform covariates, and the single Bernoulli parameter of the nonuniform covariates. It is interesting to note that the maximal number of coefficients returned by our algorithm approaches 40.

Overall, these preliminary simulation results show that our implemented estimator behaves in accordance with the theory even in the small-sample regime. The performance of the logarithmic thresholding scheme is especially encouraging, suggesting that it may be possible to trade off MSE against complexity in a way that will scale to large values of d. In the future, we plan to test our method on high-dimensional real data sets. Our particular interest is in social network data, e.g., records of meetings among large groups of individuals. These are represented by binary strings most of whose entries are zero (i.e., only a very small number of people are present at any given meeting). To model their densities, we plan to experiment with Walsh bases with η biased toward unity.

Acknowledgments

This work was supported by NSF CAREER Award No. CCF-06-43947 and DARPA Grant No. HR0011-07-1-003.

References

[1] I. Shmulevich and W. Zhang. Binary analysis and optimization-based normalization of gene expression data. *Bioinformatics* 18(4):555–565, 2002.

[2] J.M. Carro. Estimating dynamic panel data discrete choice models with fixed effects. *J. Econometrics* 140:503–528, 2007.

[3] Z. Ghahramani and K. Heller. Bayesian sets. *NIPS* 18:435–442, 2006.

[4] J. Aitchison and C.G.G. Aitken. Multivariate binary discrimination by the kernel method. *Biometrika* 63(3):413–420, 1976.

[5] J. Ott and R.A. Kronmal. Some classification procedures for multivariate binary data using orthogonal functions. *J. Amer. Stat. Assoc.* 71(354):391–399, 1976.

[6] W.-Q. Liang and P.R. Krishnaiah. Nonparametric iterative estimation of multivariate binary density. *J. Multivariate Anal.* 16:162–172, 1985.

[7] J.S. Simonoff. Smoothing categorical data. *J. Statist. Planning and Inference* 47:41–60, 1995.

[8] M. Talagrand. On Russo's approximate zero-one law. *Ann. Probab.* 22:1576–1587, 1994.

[9] I. Dinur, E. Friedgut, G. Kindler and R. O'Donnell. On the Fourier tails of bounded functions over the discrete cube. *Israel J. Math.* 160:389–421, 2007.

[10] I.M. Johnstone. Minimax Bayes, asymptotic minimax and sparse wavelet priors. In S.S. Gupta and J.O. Berger, eds., *Statistical Decision Theory and Related Topics V*, pp. 303–326, Springer, 1994.

[11] E.J. Candès and T. Tao. Near-optimal signal recovery from random projections: universal encoding strategies? *IEEE Trans. Inf. Theory* 52(12):5406–5425, 2006.

[12] O. Goldreich and L. Levin. A hard-core predicate for all one-way functions. *STOC*, pp. 25–32, 1989.

[13] E. Kushilevitz and Y. Mansour. Learning decision trees using the Fourier spectrum. *SIAM J. Comput.* 22(6):1331-1348, 1993.

[14] W. Härdle, G. Kerkyacharian, D. Picard and A.B. Tsybakov. *Wavelets, Approximation, and Statistical Applications*, Springer, 1998.

[15] S.A. van de Geer. *Empirical Processes in M-Estimation*, Cambridge Univ. Press, 2000.

[16] M. Talagrand. Sharper bounds for Gaussian and empirical processes. *Ann. Probab.* 22:28–76, 1994.